Assignment 2

Jyotsna, Pallav, Shashank

01/03/2020

1. Introduction to Data

The objective of this assignment is to understand the *global health epidemics* and to analyse the factors related to Tuberculosis (TB). There are different data sets which provide country-wise information about estimates of incidence & mortality, outcomes, community engagement,laboratories, budgets etc. over a period of time. We thoroughly checked all the data sets and interpreted the relation among different variables.

After scrutinizing the data, we decided to focus on the estimates of HIV patients with TB and Total Estimates of TB. To initiate our analysis, we chose only one data set, i.e., *TB_burden_countries_2020-02-29 data set*. The data set contains estimation of incidences, incidences per 100k, mortality rate, case fatality ratio along with other useful parameters. It contains 4040 observation and 50 variables out of which 5 are **"character"** variables and remaining are **"numeric"**.

The summary statistics of the two variables which are used for further analysis is as follows:

Variables of Interest:

a) e_inc_100k - This variable shows the estimate incidence of TB per 100k population. It ranges from 0 to 1280 with a mean of 125.8.

b) e_inc_tbhiv_100k - This variable shows the estimate incidence of TB who are HIV positive per 100k population. It ranges from 0 to 983 with a mean of 38.04.

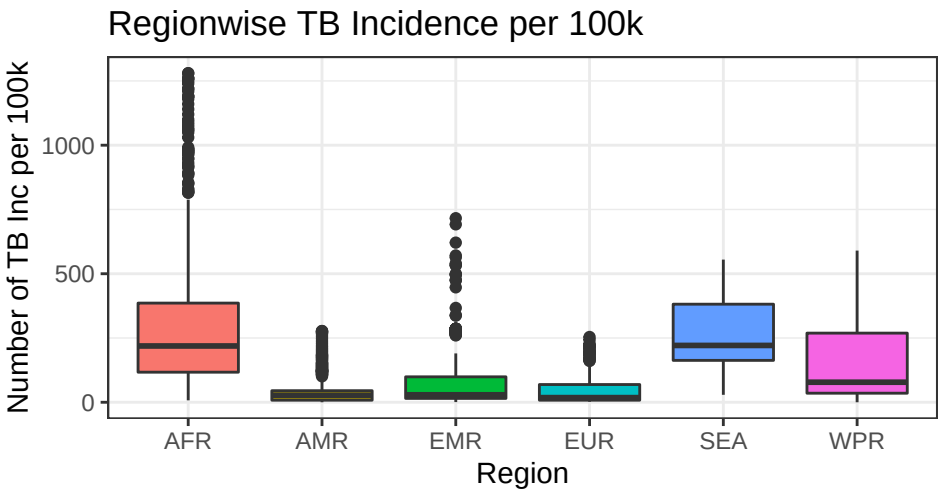
Data Wrangling

Tidy - The data set looks tidy as each variable has a column and each observation on those variables has its own row.

Filter - We will select specific variables of interest by using dplyr.

Clean - Next step is to clean the data, which involves identifying the missing values and removing them if required.There are 613 missing values in the "e_inc_tbhiv_100k" variable. Since there is no specific logic for imputation of the missing values, it is better to remove these observations.

The region-wise boxplot shows the variation in the TB incidence per 100k among different regions-

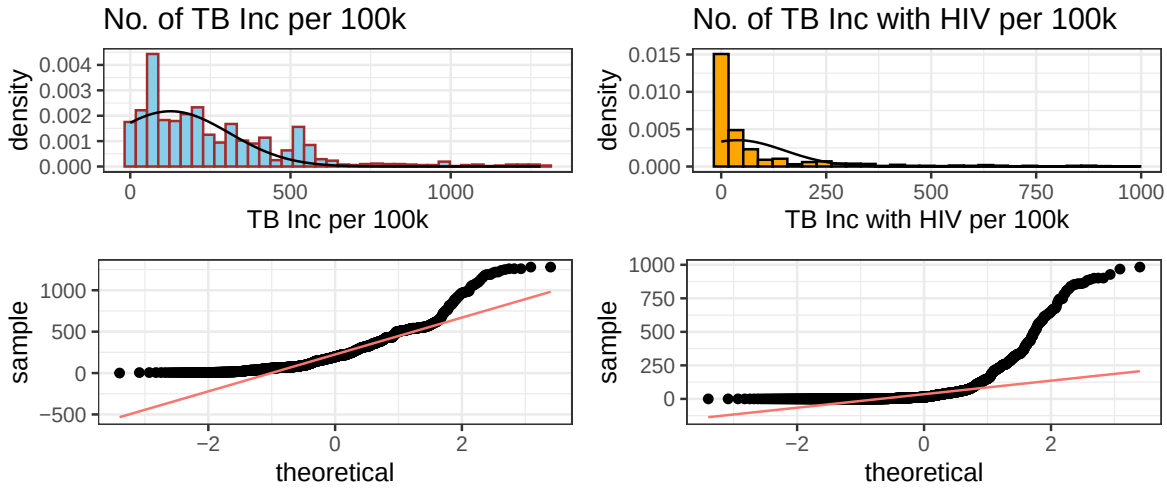


The boxplot indicates that Africa (AFR) and South East Asia (SEA) have most number of TB incidence cases per 100k followed by Western Pacific Region (WPR). There are few data points which fall outside the upper bound of **inter-quartile range (IQR)**. A value is suspected to be a potential outlier if it is less than (1.5)(IQR) below the first quartile or more than (1.5)(IQR) above the third quartile. On further investigation, it was observed that these values correspond to the countries with low population and high number of TB incidence per 100k so they can not be considered as **outliers**.

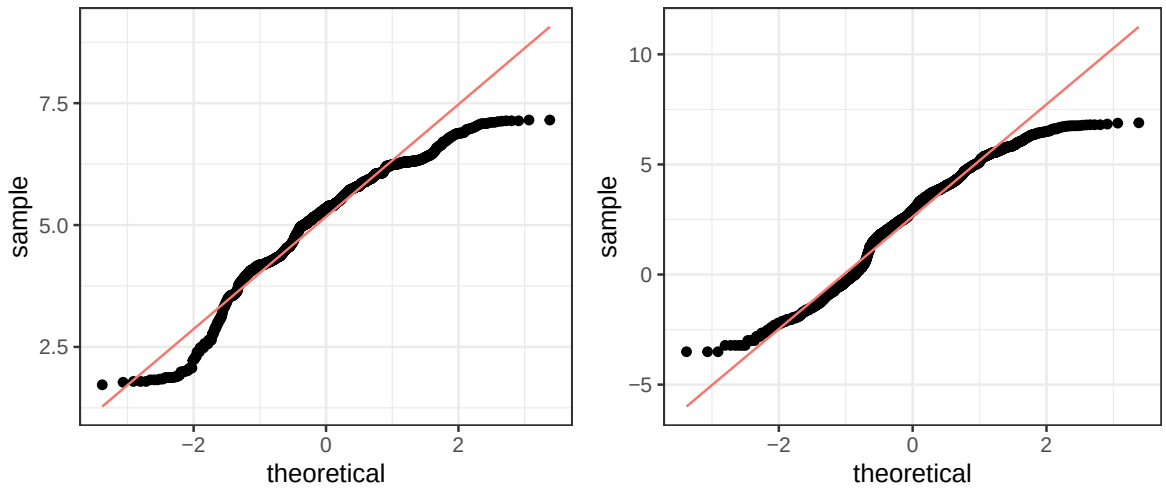
2. Planning

AFR, SEA and WPR regions encounter almost 80% of the all the TB incidence cases, so, further data analysis is done by focusing majorly on these three regions. **Our hypotheses is - People who have HIV are at a higher risk of developing TB as compared with non HIV population**.

In this section, we will check for *assumptions for normality and homoscedacity*. The basic visual inspection of normality can be done by plotting histograms. Let us see the **histograms** for our variables-



QQ plots and histograms are positively skewed and does not look normal. Let's try transforming the data using **log transformation** and compare the QQ plots.



The plots look better after transformation but are still skewed.

Another way is to do quantitative normality tests for **skewness and kurtosis**. We will perform these tests on the transformed data and interpret the results.

```
##                   skew.2SE  kurt.2SE
## TB Inc           -6.417865  2.371959
## TB Inc with HIV -3.127947 -2.593082
```

Because the absolute value of skew.2SE is greater than 1, we conclude that the **skewness for both variables is different from 0**. For Kurtosis, values are greater than 1, therefore we can say that **kurtosis is different from 0**. The results of skewness & kurtosis were quite high before data transformation.

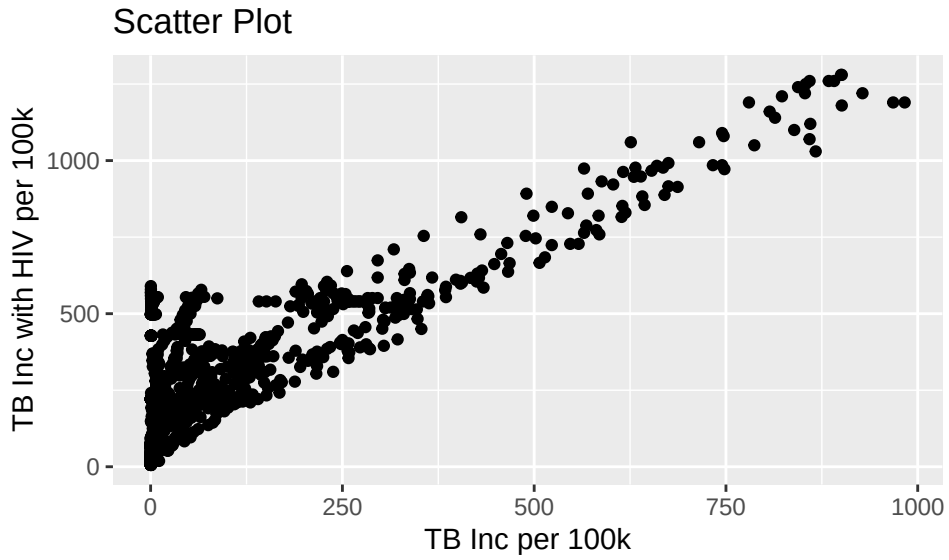
To further test for normality, we use Shapiro-Wilk normality test and the test shows that the *TB incidence per 100k is significantly non normal at 5% level of significance (p value < .05)*

So we can conclude that the normality assumption does not hold true for both the variables based on visual as well as quantitative inspection. Next is the **homoscedacity** assumption which can be checked through levene test.

Based on Levene Test, it is found that *homogeneity of variance is significant at 5% level of significance (F=1.5886, p < 0.05), so we can thus infer that the assumption of homoscedasity is not met.*

3. Analysis

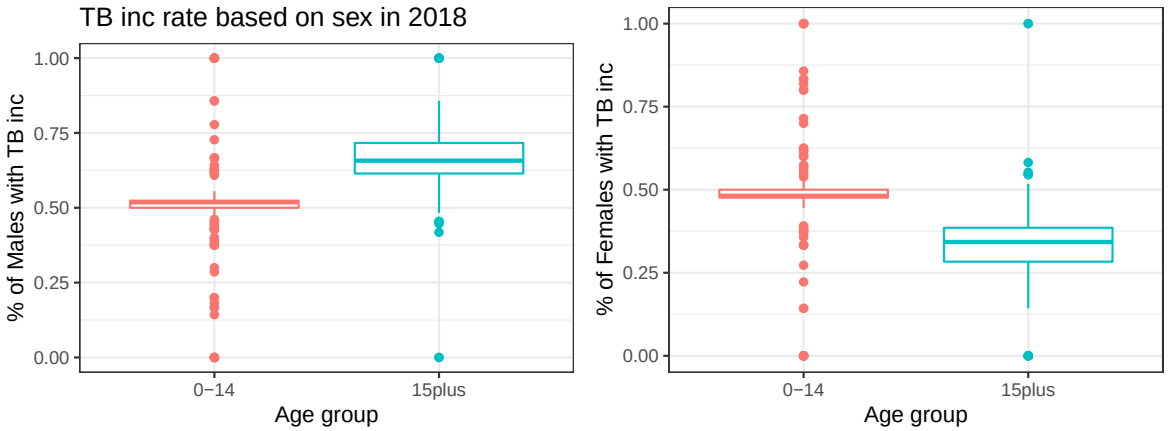
Now let's look at the correlation between our variables of interest. We will do a quick visual check by using a scatter plot.



It looks like Total TB incidences might be positively correlated with TB incidences in people who have HIV. But correlation does not necessarily implies causation. We can test for correlation by using cor.test() and Pearson coefficient as we have interval data.

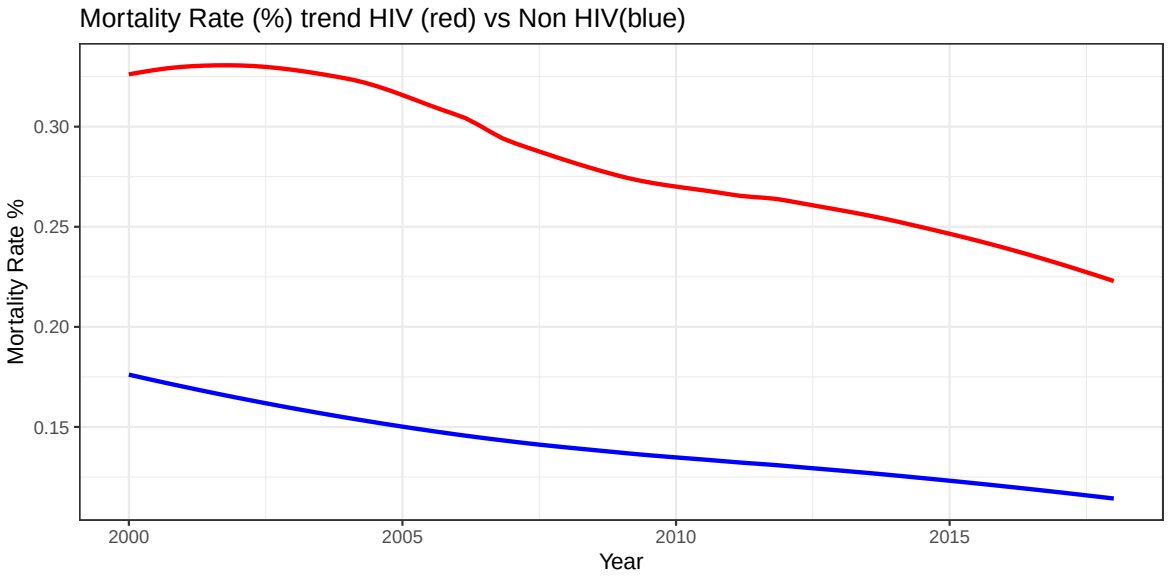
Total TB Incidences are significantly correlated with TB Incidences in people with HIV, r=0.84 ($p < 0.001$, 95% CI 0.827 to 0.858). A correlation of 0.84 represents a large effect explaining 71.11% of the variance.

Let's get more insights on TB incidence with respect to age and sex for the year 2018. After doing data wrangling, we see some interesting results.



From the boxplots of TB incidence rate, we observe that higher proportion of males is affected by TB as compared with females. If we look at 15 plus age group, males have 65% incidence rate whereas females are at about 35%.

If we observe the TB mortality rates of HIV and Non HIV population, there is significant difference between them. The percentage of people who had HIV before death is almost 2.5 times higher than the people who did not have HIV. The graph indicates that the mortality rates are declining over the last 18 years which may be a result of better healthcare and increased awareness about the disease but the difference in the HIV and Non HIV TB morality rates is more or less constant.



One of the reasons for this gap could be the *disproportionate budget allocation* for TB HIV patients. For example, Africa region got 4.5% of the total budget for TB HIV patients in 2018 and it had the highest proportion (37%) of TB HIV patients.

4. Conclusion

We got quite wonderful insights through our analysis which supports the hypothesis that people with HIV are at a higher risk of developing TB as compared with Non HIV population. Additionally, males have higher incidence of TB than females. Although the government is allocating resources to control this disease but still there are a few gaps in terms of financial allocation and health facilities which vary across countries. Moreover, the mortality rates have declined over the years which is a positive sign. There are other factors as well which directly or in-directly affect the incidence of TB but we can surely conclude that HIV is one of the important factors which causes TB.