MSCI 718 - Assignment 2

The whole dataset is all about the **TB estimates** in **216 countries** over a period of **18 years**. The data covers different forms of TB cases and the patients who have been affected by HIV as well. There are in total **4040 observations** collected for **50 different aspects**. From the four data sets available on TB related data, we are working on the foremost dataset to understand the effect of TB on population in general.

The other datasets do not have tidy data and have lots of missing values. In some of the cases, 0 values represent null values, which makes it difficult to distinguish between the absolute value and the unknown value. In other cases, we observe that columns have been combined into one which makes the data difficult to interpret.

In the following dataset, each country has been represented by **iso2(char), iso3(char), isonumeric (int) and g_whoregion(char)** (**all being factor values**). The g_whoregion has no specification in the dictionary, therefore the interpretation of the variable remains unknown. For each country, the data is observed for 18 years (int) and the corresponding values such as population, number of incidents occurred(with HIV and without HIV), number of mortal cases(in different scenarios such as cases with TB and TB-HIV), number of fatal cases, number of new TB cases and the detection ratio (all being numerical values). Each of the major categorizations as described above have an upper bound and lower bound defining the range of the data.

Different representations of the same category such as the **total number, number of cases in 100k and percentage** makes the data to be analyzed deeply. Also the values for **fatal ratio** is given for 2018 only.

For the following report we are focusing on the **detection ratio** (level_of_measurement = interval, data type = double), **number of incident cases** (level_of_measurement = interval, data type = integer) and **the number of mortal cases** (level_of_measurement = interval, data type = integer).

GOAL : The goal is to find how the detection of the infected patients affects the mortality rate of the population.

HYPOTHESIS : An increase in the detection ratio shall result in a decrease in the number of mortal cases.

The following observations were made about the dataset:

1. The dataset was observed to have some null values along with a few outliers which were cleaned during processing the data.
2. There were a few sets of columns that were repetitive and were ignored during the analysis.
3. There were also observed null values for the detection ratio, mortality cases and total incident cases for some countries and were removed.

ASSUMPTIONS :

1) There are 52 countries whose population is decreasing over the years, which can be told due to the migration of the people from those countries.
2) We are assuming the e_inc_num (Estimated number of incident cases) takes in consideration all the 3 types of TB cases only. No other kind of infected patient's count is calculated.
3) We assume that the data is complete and has been collected for people belonging to different age groups and gender.

The variables that we are taking into consideration are the **detection ratio** and **number of mortality cases** caused due to all forms of TB. Since the number of people died because of TB are directly dependent on the total number of people being infected with TB, therefore we create a new variable named **Mortality Rate** which is calculated as follows:

Mortality Rate=(Number of Mortal Cases/ Total Number of Incidents)*100

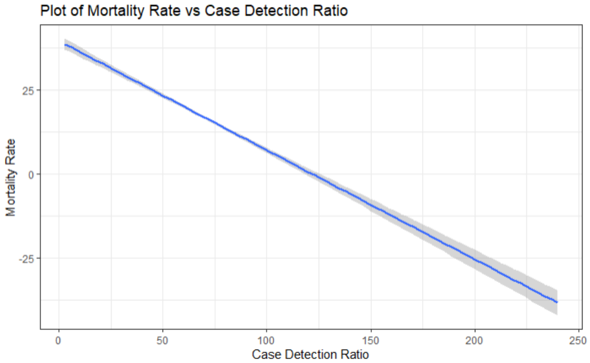
This above formula gives us the range of mortality between 0-100, thereby standardizing the data.

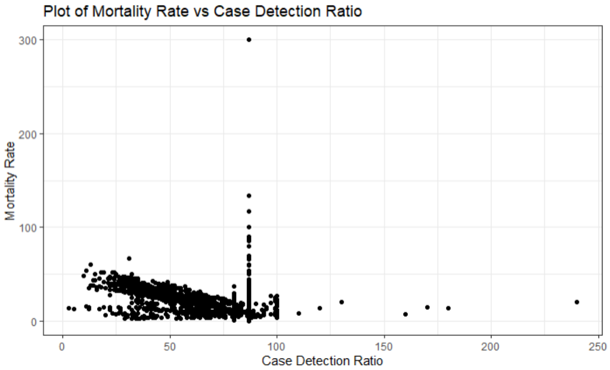
The data has been observed to have a lot of null values. Since this data increases uncertainty to our results, they have been removed. After this step we were left with 205 countries removing 654 observations.

For the tidy data obtained, the relation between the variables is calculated using **Pearson's correlation**. The data type of the variables under consideration is of the interval type, both ranging between 0 and 100 [percent]. Along with that, to test the assumptions, **qqplots** were used to check the **normality of the data**. To verify the normality, **Shapiro's wilk test** and **Levene's test** were executed both resulting in

p-values<0.05. **Shapiro's wilk test** resulted in **2.2e-16 p-value** while **LeveneTest** resulted in **2.396e-13 p-value**, both of them testifying that the distribution is normal.

For data visualizations, the data was first plotted between the Detection Ratio and Mortality Rate. The results were as follows:

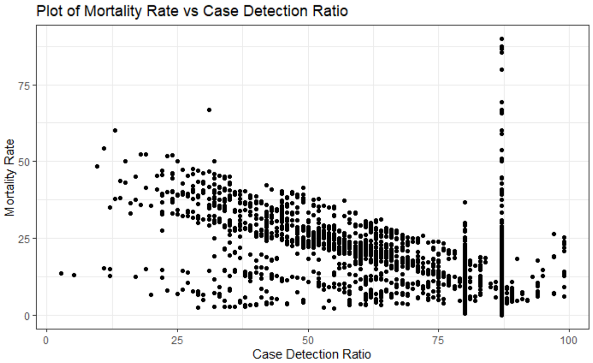


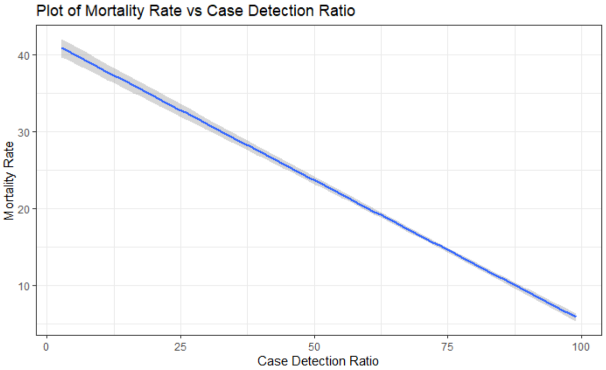


The slope shows that there is a **negative correlation** between the two variables under consideration. The correlation calculated for these data elements was observed to be **~-0.44**. But looking at the scatter plot, we could see a lot of outliers. Mortality rate and detection rate are percentages that shall not exceed 100%. But for some cases we see uncertainty. There could be multiple reasons for the same:
1. People infected moving from one country to another.
2. No records for infected patients.

These could be the probable reasons for the same. The number of cases for detection rate exceeding 100% are 25 and the the number of cases for mortality rate exceeding 100% are 12. Therefore these cases were removed.

Once the outliers were removed, the correlation between the data variables was calculated again and the trend was observed.



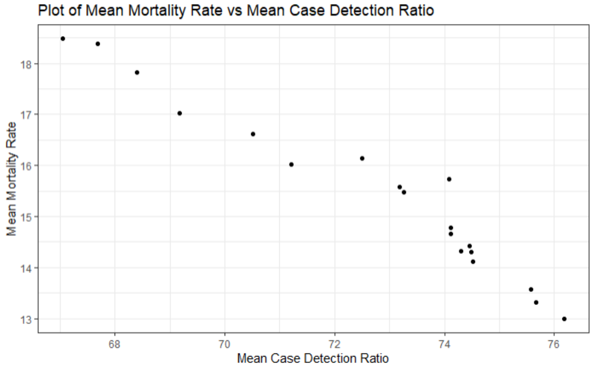


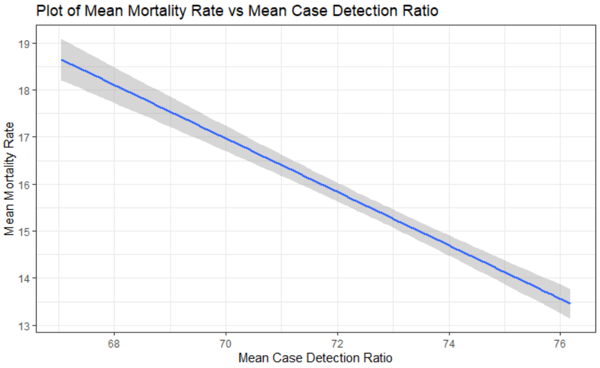
The removal of the outliers confines the data within the expected range. Along with that, we observe an increase in the correlation between detection rate and the mortality rate to **~-0.61**. This value portraits strong correlation between the variables.

For the countries that have an increasing population have shown to have correlation **> -0.9** (concluded based on results tested on 50 countries sample). But for the above dataset, the correlation is **-0.61**. The reasons for the same could be the following:

1. Decreasing population over the years
2. No change in the detection rate or the mortality(combined with the effects of decreasing population)

To understand the general trend, we find the mean values for all the countries combined for each year and then understand the results:





On the average, we see drastic changes in the results. The correlation between the variables rose to **-0.97**. From this result and the plots, we can say that, in general, the increase in the detection ratio directly impacts the mortality rate.

From the above results, we can conclude, upon analyzing, visualizing and finding correlation between variables, that it is found from this particular dataset that the detection ratio is **inversely proportional** to the mortality rate. So, as the detection ratio increased, the rate of mortality decreased. And also, the increase and decrease in the population has a huge impact on the mortality rate. Thus, the correlation test has been performed on a pair of variables in the chosen dataset and an inference on the dataset is obtained.