MSCI 718 : Assignment 2

March 3, 2020

Data

Global Health Epidemics is a major concern, particularly Tuberculosis. We use the estimates of TB burden provided by WHO for the Global Tuberculosis Report in accordances to the data given by 216 countries. Here, we analyse the dataset for **TB burden countries estimates** for 2018 and dataset on the **Expenditure and utilization** of health services in the fiscal year 2018.

TB Expenditure and utilization data contains records of health services from fiscal year 2017 onwards. This dataset gives us the data on the actual cost of drugs, individual expenditure on laborartory infrastructure, equipments and supplies, operational research and surveys, programme costs to treat different forms of TBs etc and total expenditure and total fundings received.

Variables Description:
exp_tot - Total actual expenditure (US Dollars) **rcvd_tot_sources**- Total funding received from all sources (US Dollars) **rcvd_tot_gf** - Funding received from the Global Fund to Fight AIDS, Tuberculosis and Malaria (US Dollars) **rcvd_tbhiv** - Funding received for collaborative TB/HIV activities (US Dollars) **exp_tbhiv** - Actual expenditure on collaborative TB/HIV activities (US Dollars)

Summary of the dataset - TB expenditure and utilization of health services:

```
##    exp_tot        rcvd_tot_sources    rcvd_tot_gf        rcvd_tbhiv
## Min.   :0.000e+00   Min.   :0.000e+00   Min.   :      0   Min.   :     0
## 1st Qu.:8.110e+05   1st Qu.:1.798e+06   1st Qu.:  701972   1st Qu.:   7426
## Median :3.993e+06   Median :6.511e+06   Median :  2586887   Median :  50000
## Mean   :2.985e+07   Mean   :3.634e+07   Mean   : 7335774   Mean   : 637603
## 3rd Qu.:1.463e+07   3rd Qu.:1.658e+07   3rd Qu.:  6513419   3rd Qu.: 370026
## Max.   :1.451e+09   Max.   :1.451e+09   Max.   :167198354   Max.   :26557836
## NA's   :154       NA's   :210       NA's   :227       NA's   :237
##    exp_tbhiv        year
## Min.   :     0   2017:216
## 1st Qu.:   5000   2018:216
## Median :  41430
## Mean   : 599509
## 3rd Qu.:  304972
## Max.   :26557836
## NA's   :238
```

Description for some of the variables in **country burden estimates dataset**:
cfr - Estimated fatality ratio in a country **e_inc_100k** - Estimated incidence (all forms) per 100 000 population **e_inc_num** - Total number of incidence cases(all froms) in a country **e_inc_tbhiv_num** - Estimated incidence of TB cases in a country who are HIV-positive **e_mort_100k** - Estimated mortality of TB cases (all forms) per 100,000 population **e_pop_num** - Estimated total population number

Summary of the dataset country burden estimates for 2018:

```
##      cfr         e_inc_100k       e_inc_num       e_inc_tbhiv_num
## Min.   :0.000   Min.   :  0.0   Min.   :     0   Min.   :    0
## 1st Qu.:0.070   1st Qu.:  13.0   1st Qu.:   220   1st Qu.:   16
## Median :0.100   Median :  48.0   Median :  2900   Median :  210
## Mean   :0.138   Mean   : 125.8   Mean   : 49963   Mean   : 7212
## 3rd Qu.:0.170   3rd Qu.: 163.0   3rd Qu.:  17000   3rd Qu.: 1800
## Max.   :0.740   Max.   :1280.0   Max.   :3200000   Max.   :332000
## NA's   :3832                       NA's   :613
```

```
##  e_mort_100k      e_pop_num          year
## Min.  :  0.00   Min.  :1.126e+03   2011  : 216
## 1st Qu.:  1.10   1st Qu.:7.187e+05   2012  : 216
## Median :  4.90   Median :5.663e+06   2013  : 216
## Mean  : 27.05   Mean  :3.211e+07   2014  : 216
## 3rd Qu.: 24.00   3rd Qu.:2.016e+07   2015  : 216
## Max.  :538.00   Max.  :1.428e+09   2016  : 216
##                                    (Other):2744
```

We are filtering the datasets for the year 2018 as we are analysing the report for most recent available data.

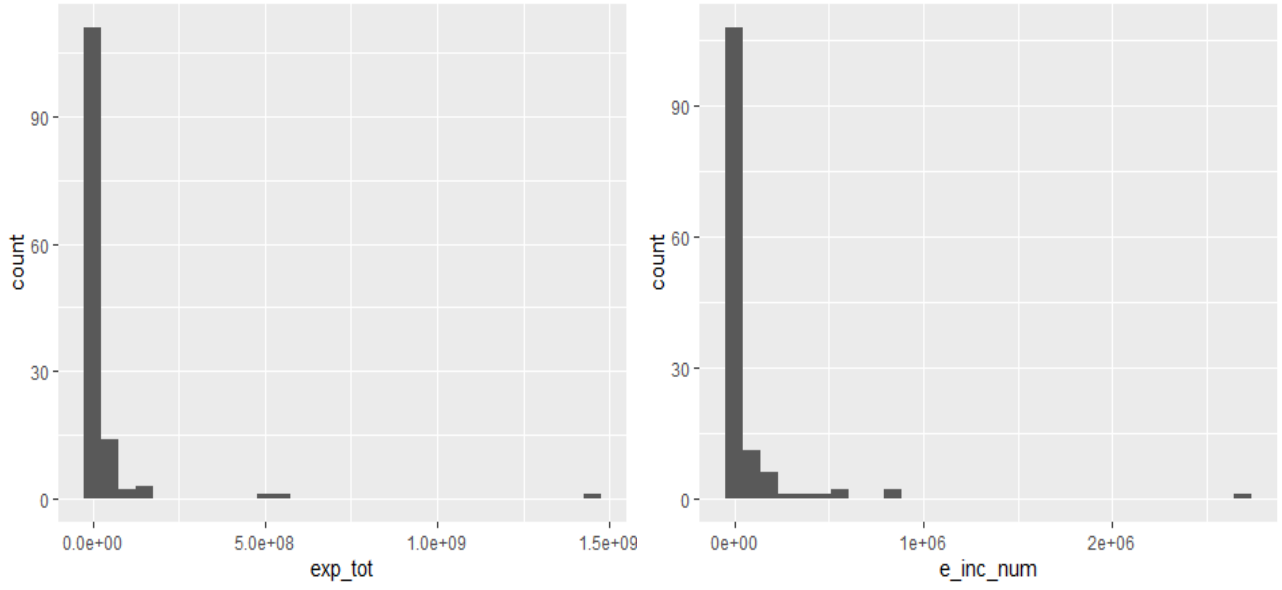
Planning

Worldwide, TB is one of the top 10 causes of death and to identify the major risk factors and **measures taken to tackle** the disease, we analyze the two datasets summarized above. Research on ending the Tuberculosis epidemic reveals that an estimated 58 million lives were saved through TB diagnosis and treatment between 2000 and 2018.[2.1] This can be attributed to the expenditure every country does based on the number of incidents a particular country has per year. Allocating funds for laboratories, drug researches, preventive measures and spreading awareness can account for reduction of mortality due to TB. So, based on the number of TB incidents logged in a country, the expenditure done by the same should be correlated. To further support this argument, our research shows that one of the targets that United Nations wishes to achieve is to treat 40 million people for TB disease in the 5-year period 2018-2022; and to do that they plan to mobilize at least US$ 13 billion annually for universal access to TB diagnosis, treatment and care by 2022 and mobilize at least US$ 2 billion annually for TB research.[2.2]

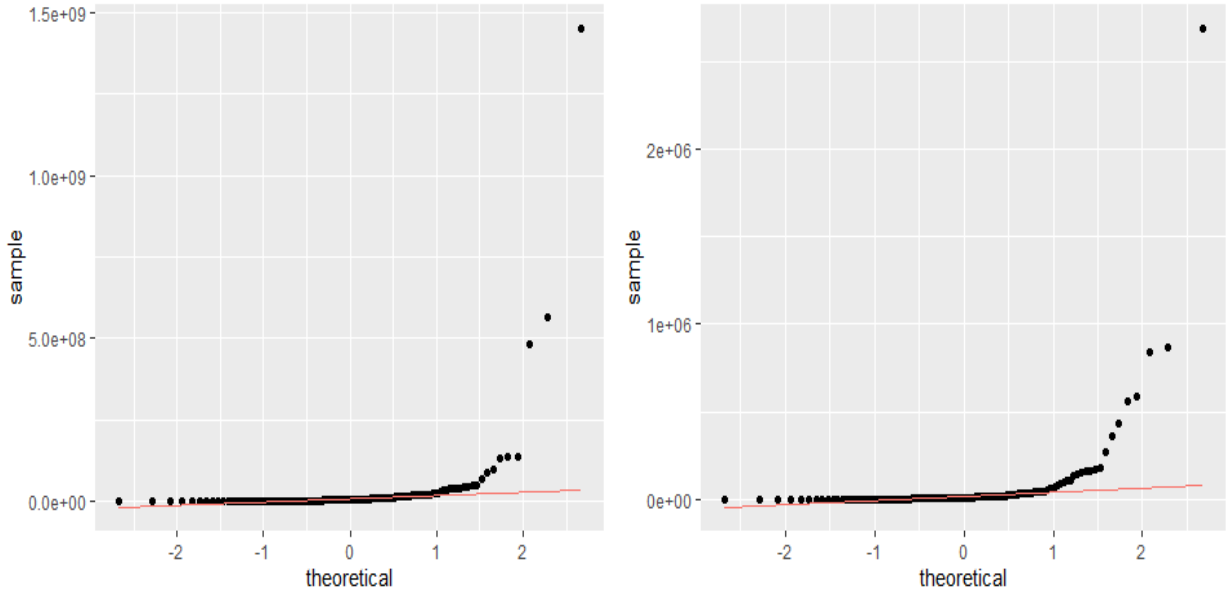
Hence we select the total number of incidences logged (e_inc_num from tb_burden_country dataset) per country for 2018 and find its relation with Expenditure done by each country (exp_tot from Expenditure and Utilization). We can test the correlation between Incidences and Expenditures by either using Pearson, Spearman or Kendall correlation test.

For either of the mentioned tests we need our data to be **interval data**, which is evident from the values we have for expenditure and incidences.

We also check whether our data distribution is normal. From the basic histogram we see both our variables are **highly right skewed and has outliers as well**.



On doing **a log transformation,** and checking it with qq plots, we still find that our data does not follow a normal distribution.



Owing to those factors, **we use Spearman correlation**. To test our Spearman correlation we postulate our **null hypothesis** that there is no significant correlation between the incidences of TB cases per country and expenditure done by each country; and **alternate hypothesis** as that there is a significant correlation between the chosen pair of variable.

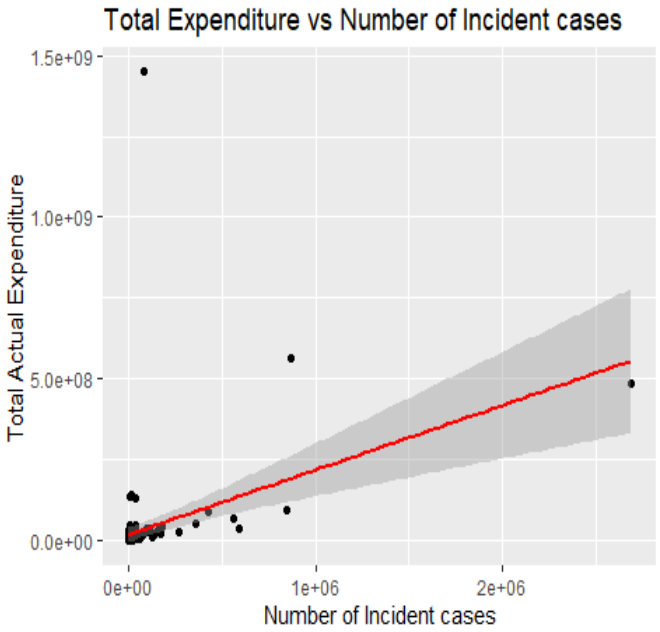
Analysis

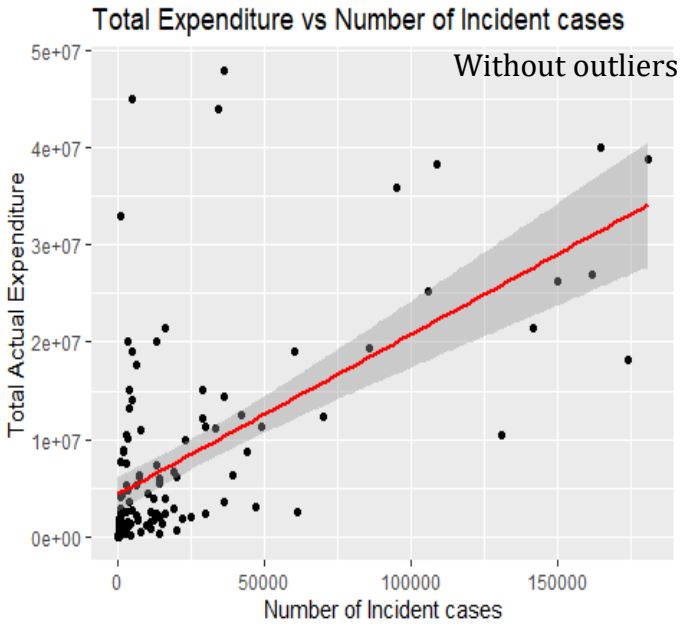
Upon performing Spearman's rho correlation test on Total Actual Expenditure variable from the Expenditure and Utilisation dataset and Number of Incident cases variable from the estimate's dataset, we get **a correlation coefficient of 0.7266404** and **p-value of 2.2 e-16**. Since the correlation coefficient is **greater than 0.5, there is a large effect**. Since the p-value is less than 0.05 (alpha- value), we must **reject the null hypothesis in favour of the alternative hypothesis.** In other words, we can say that correlation coefficient is not equal to 0 with 95% confidence. The results of the correlation test are given below.

```
##  Spearman's rank correlation rho
##
## data:  merge2$e_inc_num and merge2$exp_tot
## S = 107180, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.7266404
```

To graphically show the relation, we plot the chosen pair. The first graph given below is a scatter plot between expenditure and number of incident cases of TB. The outliers in number of incidents are for the under developed countries with high population.The outliers for expenditure are for India, China and Russia. India and China have very high population which explains the high values for expenditure. On the contrary, Russia has less incidence but very high expenditure and this is something which must be investigated. The hospitilization policies and lengthy case management periods account for the high values.

To better visualize the correlation between expenditure and number of incident cases, we set the range for no. of incidents between (0,250000) and for expenditure, between (0, 100000000)





Upon further investigation, we found out that some of the variance in Total Actual Expenditure is accounted for by the Number of Incident cases who are HIV-positive. To confirm this, we decided to find the partial correlation between Total actual Expenditure and Number of incident cases whilst controlling for the effect of Number of incident cases who are HIV positive. We get a partial correlation coefficient of 0.5401039 and p-value of 2.3314e-11. We can see that this correlation coefficient is different from what we got before controlling for the effect of HIV cases. This confirms that the number of HIV cases has an influence on the Total Expenditure and the value of 0.5401039 is a truer estimate of the unique relationship between Total Expenditure and Number of Incident cases.

Conclusion

As we have seen in our analysis, there is a positive correlation between expenditure and number of incidences. This means that countries tend to spend more on TB treatment with increasing number of incidences. However, we cannot conclude that expenditure on treatment is the sole contributing factor towards eradication of TB as there are various other factors that we must consider. A prime example of this is Russia, their TB incidence is 79000 for the year 2018 but their expenditure on TB treatment is twice that of China which has a TB incidence of 866000 for the same year. This goes on to tell us that funds allocated towards treatment of TB is wasted unless utilized effectively.