Is world improving? A Data Analysis on TB Dataset by WHO

Dhaval Thakur, Rushi Bhuva, Tejas Pandit
06/04/2020

Introduction

In this report we analyzed the dataset collected by various countries and gathered by World Health Organization. We have been given task to identify factors that may be related to Tuberculosis. We selected two variables, one is incident numbers for TB, second is total expenditure by countries to fight the diseases for the year 2017 and 2018 and found the correlation between them. Apart from this, to check the effects of total population, we have found the partial correlation as well by considering total population as a controlling variable.

Data and Planning

Getting to know about the Datasets

The two datasets we would be deep diving into for our factors of TB are TB_burden dataset and TB_expenditure. In order to do analysis on TB, we need the incidence numbers (e_inc_num: double) and TB_burden_countries dataset aids us with that only alongside with each country. In order to furthermore simplify and see the world trends, we would be focusing on the regions present in the dataset namely **AFR, AMR, EMR, EUR, SEA and WPR** rather than all the 216 countries separately making the analysis and plots more complex and difficult to compare. In addition to that Our objective is to identify relation between these two variables, and furthermore find relations with other variables available in the dataset.

Why these variables only?

1. In order to analyses the Tuberculosis, we certainly required the estimated **number of incident population**, so that we can have some estimates of affected people to analyses with other variables too, without this estimate we will not be able to focus on the required affected population.
2. From the datasets provided by the WHO, we observed that it is an epidemic disease and each country has done some expenditures to tackle this disease. According to WHO, increased funding has shown improvement in number of TB cases over time for various countries and regions [4]. Therefore, we have chosen **total expenditure** as a second variable.
3. The last variable we took **population estimate** is to compare and analyses how the total estimated population stands with the incidence cases.

The TB burden dataset has **4040** observations for **50** variables (columns). On the other side, the expenditure dataset has **432** observations and **43** variables (columns). The variables taken into consideration are **incidence of TB, population estimate** and **total expenditure**. Mean of these variables are **$4.9962608*10^4$, $3.2109693*10^7$** and **$2.9854555*10^7$** respectively. All the variables taken into consideration are continuous variable.

Data Cleaning / Tidying Data: Getting the Data ready

The dataset in the consideration is not in the form in which we can start our analysis and thus requires some transformation since we are focusing on the 5 regions of the world and not the countries. Using various transformation functions, we created a new data frame with variables namely Regions, Year, mean estimate of incident cases and mean of Total estimated expenditure by those regions. Now after transforming the data and having a look at it we saw that the maximum incidence cases for TB was in SEA (South East Asia) region in 2017, 2018 in both time periods with estimated mean values 484,212 and 475,545 respectively. So, from the box plot we can observe the mean and deviation of incident cases and expenditure done for each region.

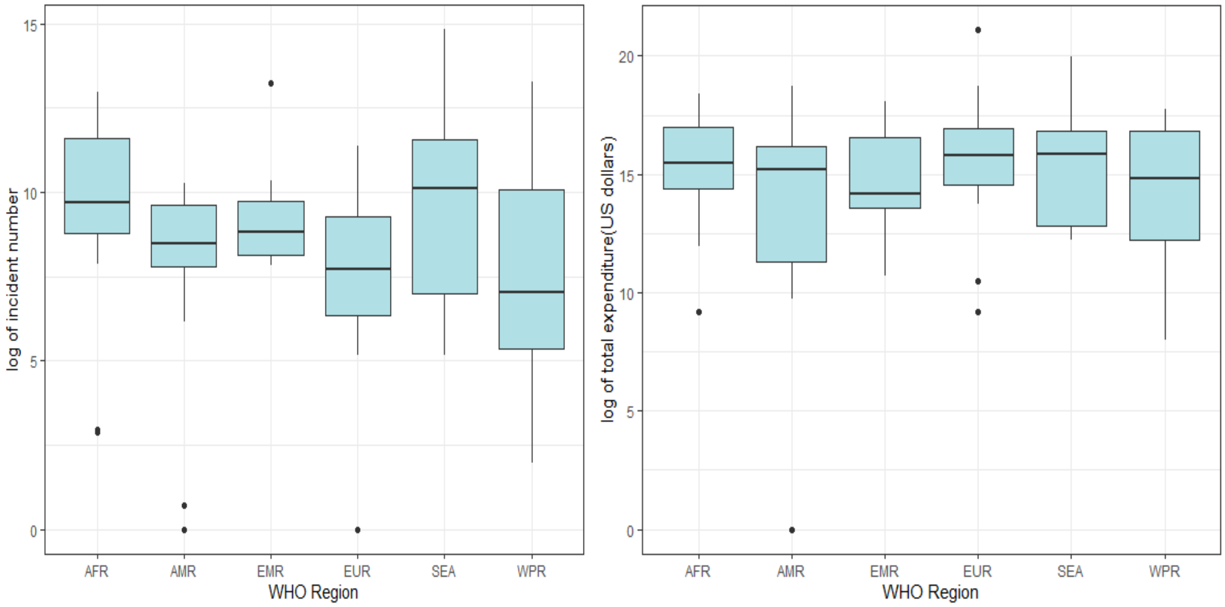

Figure 1 Box plots for TB incident number and Total expenditure for 2017-2018

As shown in Fig.1, we observe that Europe spends more money (mean is 15.43) for TB even it has less incidence numbers (mean is 7.42). SEA and WPR has Inter-quantile range of 5 but the mean for incidence numbers is high in SEA compared to WPR that suggests that SEA region has more countries with a greater number of TB incidences compare to WPR. From the boxplot of expenditure, we observe that SEA has spent more money compared to WPR as SEA has higher incidences of TB.

As we are observing that the variables of our focus are from two different datasets, there is a requirement of joining of two datasets with the focus variables. We also further did the log transformation of both the variables to scale down the data and for better representations in plots. Transformed dataset has 216 observations and 93 variables. This dataset has 86 missing values for total expenditure, so after removing it, new dataset has 130 observations.

In order to proceed with the analysis and tests on the dataset, we first analyzed the data and observed that the **estimates incident numbers** present in the **TB Burden Countries** does not follow normal distribution from using Q-Q plot and Shapiro's test.

Testing for Normality

Just to countercheck and to do the test for the presence of normality on the variables (incident number and total expenditure) we ran the Shapiro Wilk's Normality test.

We observed that the test gives **p-value = $8.375*10^{-6}$ (~0.00008375)** which is very less as compared to our significance level which is 0.05. Thus, it is failing our Normality test for incident number at the 5% significance level. In the same way, we tested normality for total expenditure that failed the normality test. Moreover, from Fig.2 we can observe that the data is not normal.

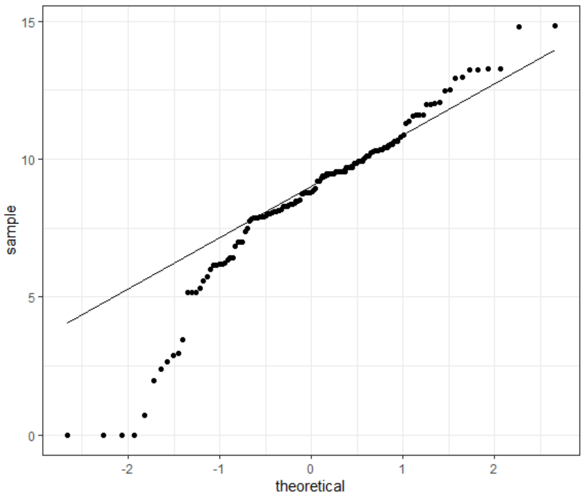

Fig.2 Q-Q Plot for incidence numbers

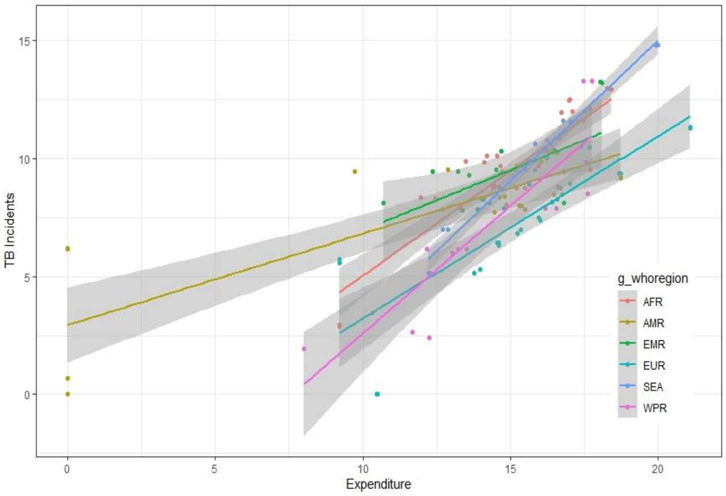

Fig.3 Correlation between incidence number and total expenditure

Setting up Hypothesis and Data Analysis

Research Question: Is there a statistically significant relationship between incident number and total expenditure, while controlling for estimated mean of population. [5](reference for two null hypotheses)
H01: There is no significant relationship between incident number and the expenditure of various country government in 6 regions.
H02: There is no statistically significant relationship between incident number and the expenditure done by various country government in 6 regions, while controlling for estimated of population.
H1: There is statistically significant relationship between incident number and the expenditure done by various country government in 6 regions, while controlling for the estimated mean of population.

Kendall Tau's Test

The Kendall rank coefficient is non-parametric test often used as a test statistic in a statistical hypothesis test to establish whether two variables may be regarded as statistically dependent & does not rely on any assumptions on the distributions of X or Y.

Result and analysis

```
##  Kendall's rank correlation tau
## data:  joined_cleaned$log_e_inc_num and joined_cleaned$log_exp_tot
## z = 9.4755, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.5642557
```

Test we get the Tau's value equal to **0.5642557** and p value comes as $\mathbf{2.2 * 10^{-16}}$. As p < 0.05 we can reject the null hypothesis and can accept the alternative hypothesis that the variables are correlated. A correlation of **0.5642557** represents medium effect size explaining **31.83%** variance. Furthermore, from Fig.3 we can observe that for all the regions total expenditure and incidence number is positively correlated.

The size of the data is small (after joining two datasets and removing missing values) but as we are getting large effect size for correlation and the Kendall's test rejects the null hypothesis, we can say that the two variables are positively correlated.

Outcome: We observe that incidence number and total expenditure are positively correlated with large effect (>=0.5). *(Topic 2.6.4 Effect sizes. DSUR)*

Are there any other features effecting our derived correlation?

In the above section we found out that there is large effect of correlation. But there can be slight or maybe more chances that there is one or more variable effecting our correlation. i.e. there might be presence of an effective control variable. Now, we think that population might be related to incidence of TB. Thus, we ran Partial Correlation test where we would take **population** as our control variable.
We see that using Partial correlation under Kendall's method, the coefficient comes out to be **r'=0.3371381** explaining **11.36% variance.** Thus, we see that there is moderate partial correlation between incident number and government expenditure, controlling for "estimate of population" variable. Results of zero-order correlation yielded that there was correlation between our focus variables with large effect of **0.54**, but r' being around **0.33** indicates that controlling for "estimates of population" has medium effect on the strength of relationship between two variables.

Conclusion

In our Analysis we divided the whole dataset into 6 regions and concentrated our analysis on "incidence number", "total expenditure" and "estimated population" variables. We observed that every region increased their expenditure in all 6 regions through exploratory data analysis. We established Hypothesis for correlation (**between incidence number and total expenditure**) and by running the Kendall's Test we got the correlation value of **0.5642557** representing medium to large effect size explaining **31.83% variance**. To check the presence of an effective control variable, we set up Null and Alternative Hypothesis consisting of the controlling variable **estimated population**. Through Kendall's method, the correlation coefficient comes out to be **r'=0.3371381** explaining **11.36% variance**. Thus, through above tests, we can **safely reject** our both **Null hypothesis** and can accept our alternative hypothesis that there is a statistically significant relationship between incident number and the expenditure done by various country governments in 6 regions, while controlling for the estimated mean of population at **5%** level of significance. Furthermore, tuberculosis is a communicable disease thus if population is more there are more chances that it would spread more rapidly, but we cannot imply from the correlation (*It does not imply causality*).

References

[1] "WHO | Global tuberculosis report 2019," *WHO*, 2020.

[2] Discovering Statistics using R textbook, 2012

[3] "Covariance and Correlation Part 1: Covariance - YouTube." [Online]. Available: https://www.youtube.com/watch?v=qtaqvPAeEJY. [Accessed: 28-Feb-2020].

[4] "WHO | Financing tuberculosis control: the role of a global financial monitoring system," *WHO*, 2011.

[5] "Partial Correlation Tutorial - YouTube." [Online]. Available: https://www.youtube.com/watch?v=LF0WAVBIhNA. [Accessed: 28-feb-2020].

Summary of the Changes done in Assignment 2 Report

In this revised report, we reworked upon the feedbacks given by the TA through feedback section and in addition to that we also focussed and improved from the suggestions given through the PEER review website. Furthermore, we contacted TA's through email to get further insights on our report in order to improve our report as we wanted to know the areas where we can furthermore enhance the quality of report.

Rework from the Learn Feedback

- Data summary for the new dataset as well as old dataset has been updated and added with relevant information about variables in consideration and furthermore added more statistical insights such as central tendencies, giving us more insight to the reader of our new dataset
- Sizes and statistics of the new dataset has been added along with any missing values (if present)
- For 2 Null Hypothesis, cited the source which depicts the logic behind it and added the source in reference section.

Rework from the feedback got from email from TA's

- We furthermore enhanced the readability of the report and as according to the feedback we reduced the occurrences of variable names in our report from 8 to 2, thus reducing redundancy of variable names throughout the report.
- In order to be more precise as what we are trying to do, we added objective of our assignment/report in second section of data so that readers can get an idea what they can expect in upcoming sections of the report.
- For the last suggestion, we rewrote the conclusion section, which is now more readable for any person, has connections with above sections using various quantitative results to conclude which supports our stated hypothesis.

Rework from the feedback from PEER evaluation

- Out of the 4 reviews we got, all were positive, but we got one feedback on cutting paragraphs into smaller paragraphs, which we implemented in this revised report to enhance the readability and the ease of reading the report.

Extra rework

- We furthermore proofread the report and enhance the quality of report by implementing a proper format of references and citations present throughout our analysis. And lastly, we added refined explanation for all the figures.