Individual Assignment 1

Ali Rafieepouralavialavijeh, 20900871

Data Summary:

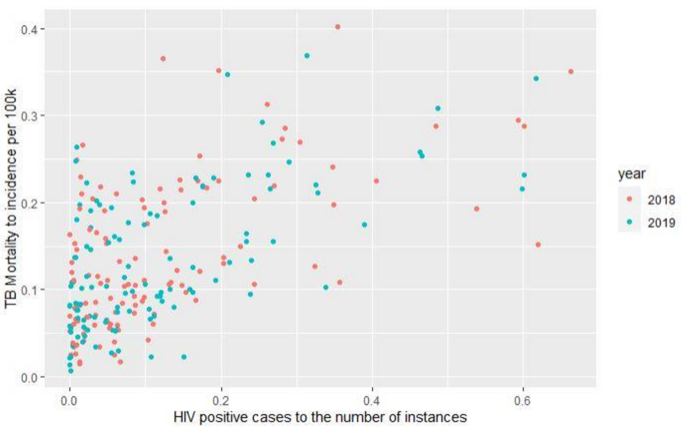
WHO has published a global tuberculosis (TB) report every year since 1997. The report provides a comprehensive and up-to-date assessment of the TB epidemic, and of progress in prevention, diagnosis and treatment of the disease at global, regional and country levels. The main research question in this report is ***how ratio of expenditure to budget is correlated to mortality ratio while controlling ratio of HIV infection to TB mortality?*** To figure out the answer and for making grounds for further investigation certain datasets are chosen including -*TB_Budget* (because it contains data about budgets of each country throughout years 2018-2019), -*TB_Expenditure* (because it contains data about expenditure of each country throughout years 2017-2019), and- *TB_burden_countries* (because it contains estimates of infection and death of each country in years 2017-2019). Based on the -*TB Dictionary* (the dataset including the definitions of all columns and variables) and the question we are concerned with, definition of selected variables for this study is as follows: **-** *budget_tot*: Total budget required (US Dollars, numeric) **-** *exp_tot*: Total actual expenditure (US Dollars, numeric) **-** *e_mort_100k*: Estimated mortality of TB cases (all forms) per 100,000 population (numeric) **-** *e_inc_100k*: Estimated incidence (all forms) per 100,000 population (numeric), and **-** *e_inc_tbhiv_100k*: Estimated incidence of TB cases who are HIV-positive per 100 000 population (numeric).

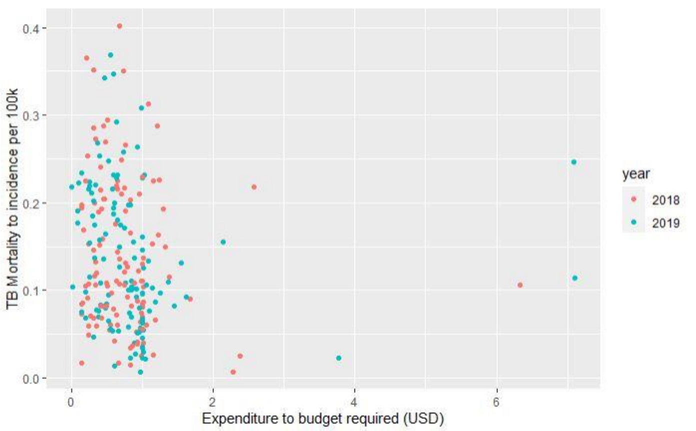
Further looking into the datasets, it is now noticeable that studies are done in different years and therefore the resulting dataset lacks data for the year 2017 for certain countries. Here, I decided to remove the year 2017 to make a consistent dataset. The type of each column perfectly matches the tests that we are going to hold and therefore need no change (since they are ratios and of "double" type). The next step is to check whether there are certain rows that are abnormal or in other words outlier, and as a result highly affect the result of our study. As such, I decided to remove the outliers to reduce their effect on our result [please see details of this step at Appendix 1, since the purpose of this assignment has not been on data exploration more focus is put on expected deliverables].

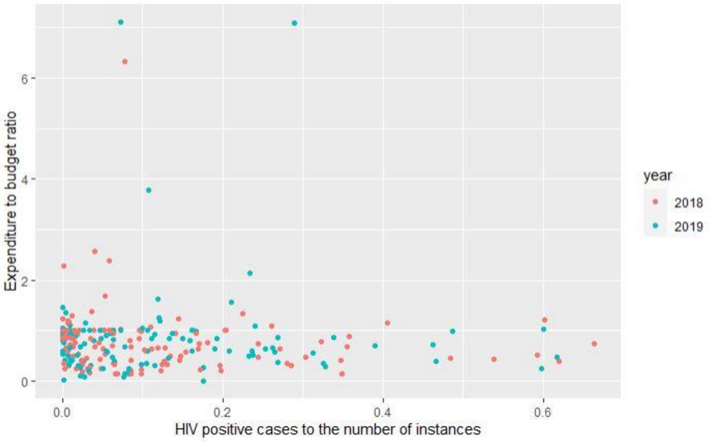
Planning:

To start with our analysis, we need to create three new columns. The first is the ratio of expenditure to budget (*exp_budget*), this shows how much more or less in comparison to the budget allocated to each country is spent on this disease. We need this column because making a comparison between different countries which come with different populations and budgets would otherwise be impossible due to the difference in the scale. The second variable is the ratio between number of deaths in 100,000, and number of incidents in 100,000 cases (*mort_inc_100k*). This variable is important to our analysis because the number of instances in each country loses its significance if we do not consider the size and population of that country. The third and the final column is the ratio of TB cases which are also HIV positive to the number of incidents in 100,000 cases (*hiv_inc_100k*). This column is important to consider because otherwise we might fail to consider all factors affecting the deaths of cases and HIV could have an impact on this ratio Figures 1, 2, and 3 illustrate the distribution of these variables in different years two by two.

Since the number of rows in our newly modified dataset is well beyond 30 (n>30), it is reasonable to assume that this dataset follows a normal distribution based on the central theorem limit. This assumption is constructive in making a decision about the type of correlation test we use because Pearson test for example requires normality of the data.

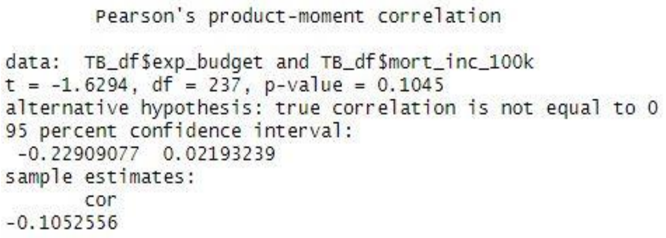






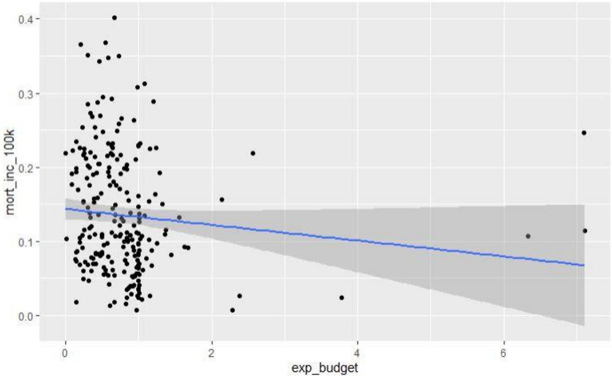
Analysis:

As the main question of this report suggests, we are going to analyze the correlation between the ratio of expenditure to budget and the ratio of mortality controlling for the ratio of HIV positive cases to total incidents. The study of these ratios is important because HIV positivity could potentially affect the result of our analysis. Thus, first we are going to calculate the Pearson correlation test (based on the results of the previous section) for the two columns:

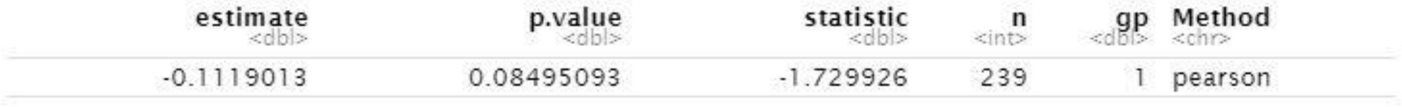


```
	Pearson's product-moment correlation

data:  TB_df$exp_budget and TB_df$mort_inc_100k
t = -1.6294, df = 237, p-value = 0.1045
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.22909077  0.02193239
sample estimates:
       cor
-0.1052556
```



It appears these two columns are negativaly correlated but this correlation is not strong. So with 95 percent confidence we have found enough evidence to reject the null hypothesis (which is that these two columns are not correlated). Now lets study the partial correlatoin of these two columns controling for the thrid one:



estimate <dbl>	p.value <dbl>	statistic <dbl>	n <int>	gp <dbl>	Method <chr>
-0.1119013	0.08495093	-1.729926	239	1	pearson

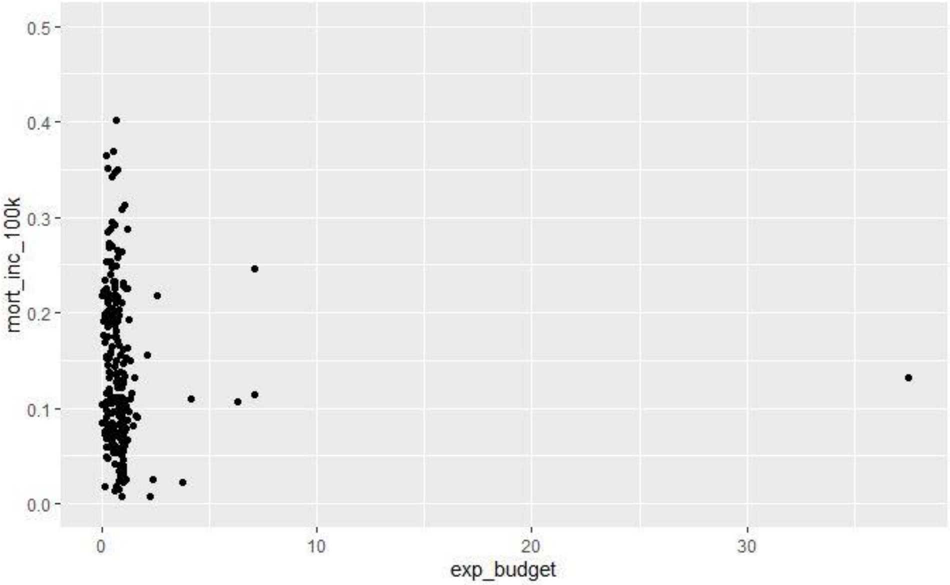
As such the partial correlation between *exp_budget* and *mort_inc_100k* when controlling for the *hiv_inc_100k* is not significant at the 5 percent level of significance. We thus do conclude that the two columns have a correlation, but the effect of the controlling variable suppress the effect of the other two variables in an opposite direction.

Conclusion:

Based the analysis and the results of other sections, we can conclude that the effect of the third column on the result of the correlation test between columns *exp_budget* and *mort_inc_100k* is significant and should be considered. The effect of HIV infection on the mortality ratio is intuitive and can obviously be noticed. This can be easily verified by checking the values of the estimates of the tables of the results of the tests that are provided in previous sections. Appendix 2 identifies this positive correlation in details.

Appendix 1:

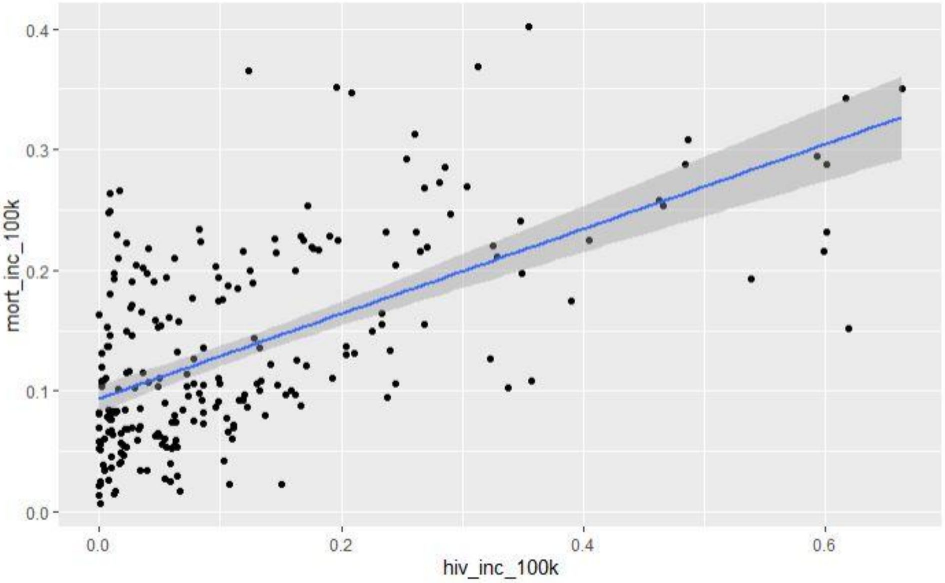
Figure below shows that there is a country whose data might be mistaken because it is too out of scale. So I tried removing it from my dataset.



"Vanuatu" country has value of 37.47 expenditure to budget rate which seems out of scale.

Appendix 2:

Figure below shows the positive correlation of HIV infection ratio with the mortality ratio.


mort_inc_100k
hiv_inc_100k