Exploratory Data Analysis on TB

First of all, we downloaded different datasets like “TB_burden_age_sex”, “TB_laboratories”, “TB_outcomes”, “TB_burden_countries” from the WHO website. We explored all these datasets by looking at the data dictionary. Among these datasets, we found “TB_burden_countries” as an interesting dataset to make a value added report for a business analyst.This dataset explains the TB burden from year 2000 to 2018 by considering the TB cases globally.

Data

In this report, we are going to analyze the data set“TB_burden_countries” by doing exploratory data anlaysis followed by finding interesting correlations.This dataset contains 4040 observations and 50 variables which gives information mainly about estimated incidence of TB cases with HIV or without HIV, estimated number of deaths due to TB or HIV, estimated mortality rate per 100k populaion due to TB or HIV for all countries.

Analysis

After exploring the dataset, we are considering the estimated incidence of TB cases, estimated incidence of TB cases with HIV positive and estimated number of deaths over the years globally.so, we have created a new dataset “TB_Mortality”, by conisdering the variables country, year, e_inc_num, e_inc_tbhiv_num, e_mort_num for our further analysis. The country column is a categorical variable wich contains different countries all over the world, year column which is a nominal variable, e_mort_100k, e_inc_tbhiv_num, e_mort_num are numerical variables.

Loading Datasets

```
## # A tibble: 4,040 x 5
##    country      year e_inc_num e_inc_tbhiv_num e_mort_num
##    <chr>       <dbl>     <dbl>          <dbl>      <dbl>
##  1 Afghanistan  2000     39000            130      14000
##  2 Afghanistan  2001     41000            140      14000
##  3 Afghanistan  2002     43000             92      13000
## # ... with 4,030 more rows
```

Data Cleaning and filtering

By looking at our new dataset “TB_mortality” we can say that there are no missing values in our data. It is clean. In our data, every variable and observation has its own column and own row and every value has its own cell. So, it is evident that our data is tidy. Here we want to estimate the effect of number of deaths due to the estimated incident TB cases over the span of last 19 years for various countries.Accordinly, we filtered the data by doing the summation of estimated incident TB cases, estimated incident TB cases with HIV positive, estimated number of deaths w.r.t to each year for all the countries (Check rmd file)

Summary Of the data

To get a quick glance over the statistics of our varaibles, we will look at the summary of our data. As estimated number of incident TB cases, estimated number of incident TB cases with HIV positive, estimated number of deaths are continuous variables, they give sensible summary statistics.

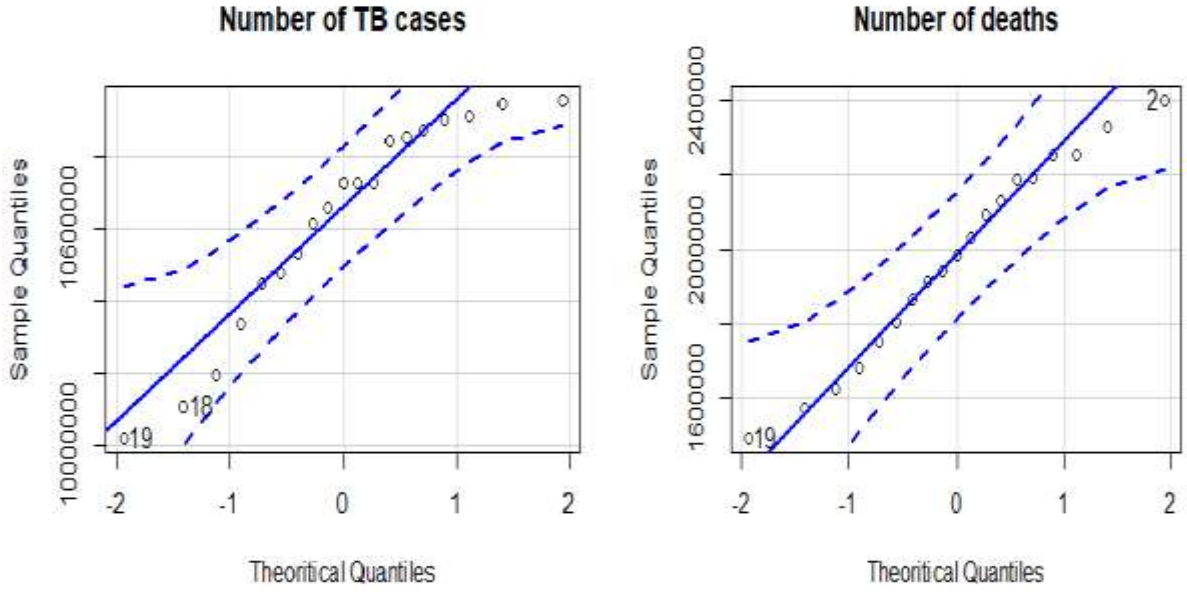
Summary of estimated number of incident TB cases

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 10018345 10461080 10725368 10623628 10862685 10956014
```

Summary of estimated number of deaths

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 1492863 1779518 1985944 1973670 2189240 2400244
```

Q-Q plots for checking normality



Normality Test

```
## Shapiro-Wilk normality test
## data: TB_df$Est_num_of_TB_cases
## W = 0.90574, p-value = 0.06187


## Shapiro-Wilk normality test
## data: TB_df$Est_num_of_deaths
## W = 0.96794, p-value = 0.7347
```

Correlation between number of incident cases and number of deaths

Null Hypothesis: There is no significant relationship between incident number of cases and number of deaths.

Alternate Hypothesis : There is statistically significant relationship incident number of cases and number of deaths.

In order to conduct a pearson correlation test the data needs to meet two assumptions i.e.the data needs to have continuous variables and it needs to follow a normal distribution ("Discovering Statistics Using R").From the above qqplots and the results of normality test states that the data is normally distributed and the variables are continuous variables. As, no assumptions of pearson correlation test has been voilated we will conduct pearson correlation test on our data.

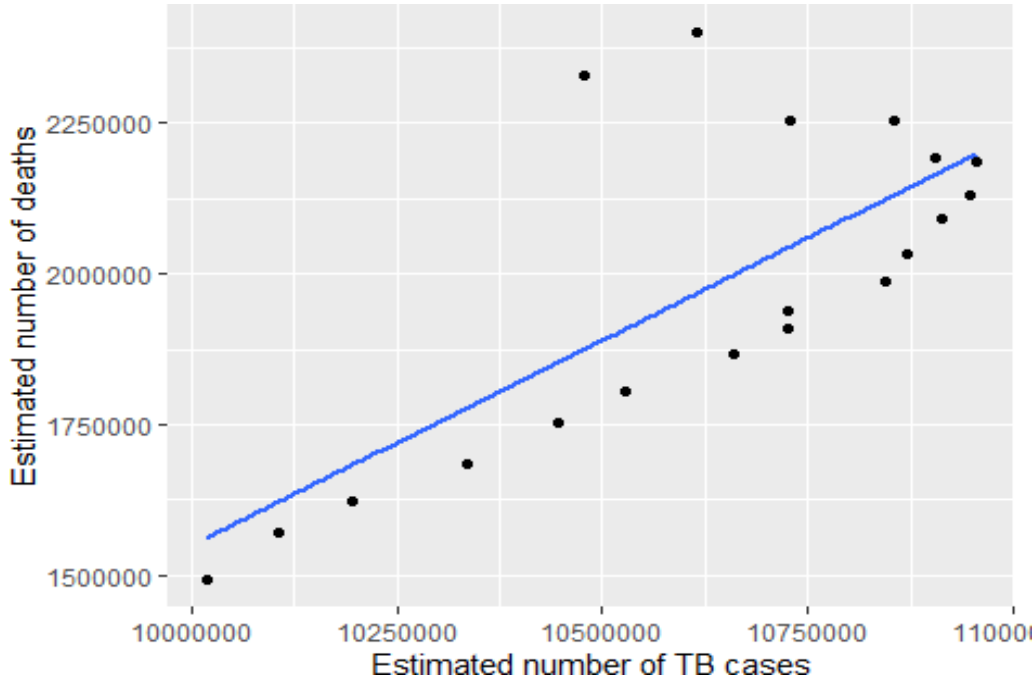
Running Pearson's Test:

Running the Pearson's Test, we got r = 0.7409046 and p−value =0.0002847. As $p < 0.05$, we can reject the null hypothesis and can accept the alternative hypothesis, that the variables are significantly correlated. The size of the data is small (after aggregating the data) but as we are getting large effect size for correlation and the Pearson's test rejects the null hypothesis, we can say that the two variables are positively correlated.

Outcome:

We observe that the two variables namely e_inc_num (Estimated number of incident cases) and e_mort_num (Estimated number of deaths) are positively correlated with large effect (>=0.5). (Topic 2.6.4 Effect sizes. DSUR)

Correlation Graph



From the above graph, we can conclude that these two variables are linearly related but this relationship is not casual.

Are there any other variables effecting our derived correlation?

From the above, 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 variables affecting our correlation. i.e. there might be presence of a control variable. Now, we think that number of TB cases with HIV positive might also affect cause of death. Therefore, we conducted Partial Correlation test where we consider number of HIV positive cases as our control variable.

We see that using Partial correlation under pearson method,the coefficient comes out to be r=0.5814653.Thus we see that there is partial correlation between incident number of TB cases and number of deaths, controlling for “number of TB cases with HIV positive” variable and the variance that is shared is 0.338025, or 33.80%. This is considerably less than the full correlation 0.7409046 , which explained 54.76% of the variance, when number of incident TB cases with HIV positive was not controlled/considered.This is a truer estimate of the unique relationship between Number of deaths and Number of incident TB cases.

There is a significant relationship between Number of deaths and Number of incident TB cases (all forms) (r=0.5814653,p=0.01137124) when controlling for the effect of Number of Incident TB cases with HIV positive.

Conclusion

With the data given, we found that number of deaths and number of incident TB cases have a large correlation which explains 54.76% variance.From partial correlation test we also observed that there is some effect of our controlling variable “Number of incident TB cases with HIV positive” which explains 33.80% variance which is less than the full correlation obtained when “number of incident TB cases with HIV positive” was not controlled. Therefore, we can conclude that number of incident TB cases and number of deaths have a large correlation, but there is also a complex relationship between number of incident cases, number of deaths, and number of TB cases with HIV positive.