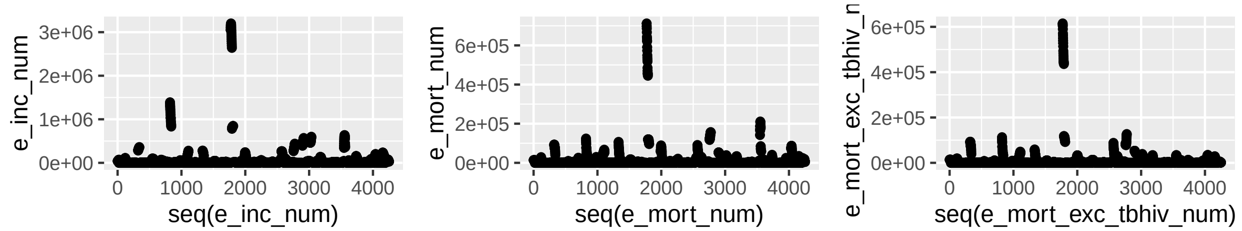
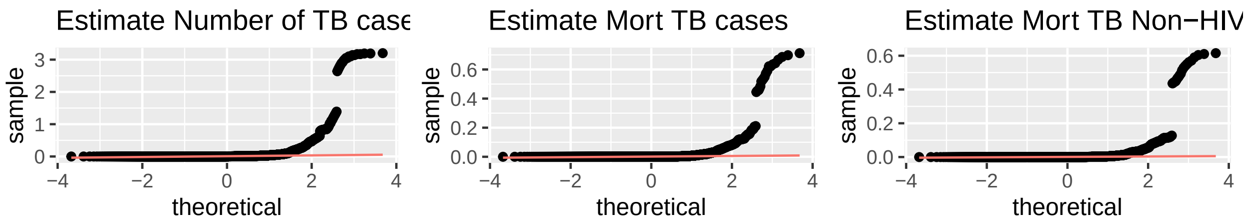
Individual Assignment 2

The features were selected from the dataset“TB_burden_countries”, which has 4772 variables and 50 features. In, Individual Assignment 1, it was observed that features Estimate Number of TB cases and Estimate Mortality Rate of TB have positive relation, even Estimate Mortality Rate of TB Non-Hiv cases is positively related to Estimate Number of TB cases. In this, assignment we will check the partial correlation between these features.

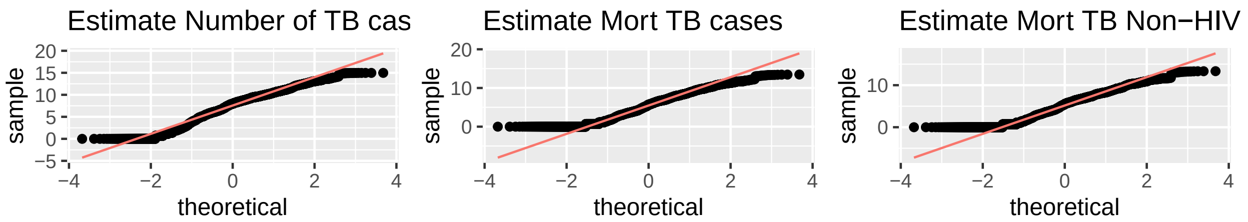
The dataset consists of 4252 observations and 5 variables.Owing to the large size of dataset 20 missing values in the dataset were omitted.The plot below shows the range of values in the features.As we are checking for partial correlation between features we need to satisfy the assumptions of the test method being used.Let’s first check if the data is normal or not, which will decide the partial correlation method to be used.



According to Shapiro-Wilk normality test(can be found in Appendix Calculation 1) and below shown qq plot,Estimate Number of TB cases (W = 0.19285, p-value < 2.2e-16), Estimate Mortality of TB cases (W = 0.19005, p-value < 2.2e-16) and Estimate Mortality of TB Non-HIV cases (W = 0.15205, p-value < 2.2e-16) is significantly non-normal at the 5% level of significance. So, I transformed the data with log transformation because it is positively skewed.

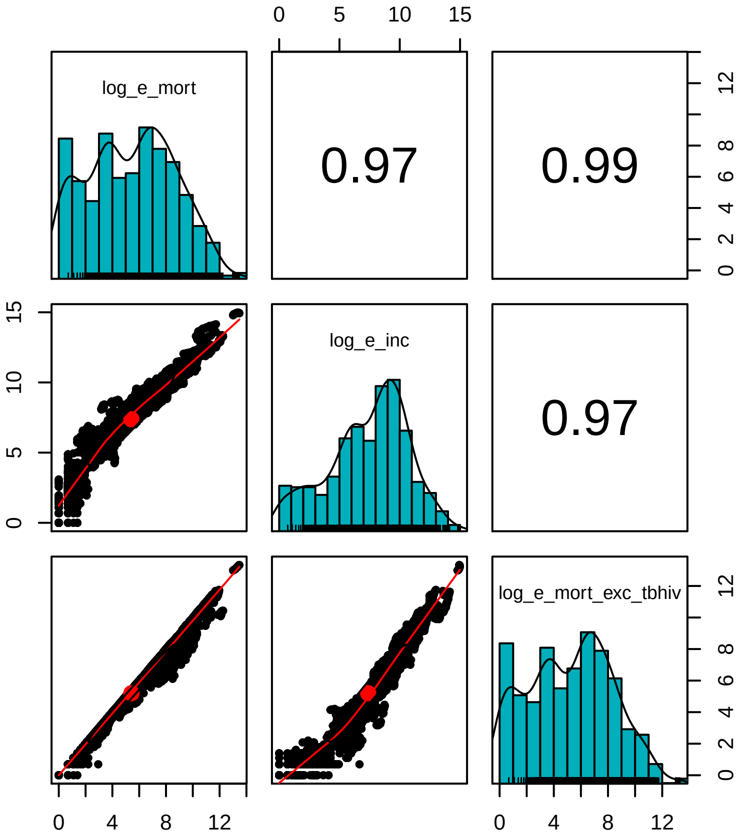


From the qq plot below after log transformation, it is observed that the features are close to normal so I will use Pearson partial correlation test on the 3 selected features.As the data is interval data Pearson correlation test can be used with any directly on transformed data. In Individual Assignment 1 limited number of samples were selected to test for partial correlation and that’s why the data after transformation was non-normal but in Individual Assignment 2, whole dataset was used and that’s why the data after transformation is close to normal.



Let's look at full correlation between variable Estimate Number of cases and Estimate Mortality rate.The Pearson correlation coefficient(can be found in Appendix Calculation 2) between Estimate Number TB cases and Estimate Mortality of TB cases is 0.9686306 with 93.8 % variance, which means both the features are positively correlated with 5% significance level. The Pearson correlation coefficient of Estimate number is TB cases and Estimate Mortality of TB Non-HIV cases is 0.9697551 with 94% variance, which means there exists a strong positive relation between the variables at 5% significance level as well.The figure below shows the partial correlation between variables and it's coefficient,along with it's distribution.

Now, let's check the partial correlation between Estimate Number of TB cases and Estimate Mortality Rate while controlling Estimate Mortality of TB Non-HIV cases, and how does it affect the relation between Estimate Number of TB cases and Estimate Mortality Rate.



The partial correlation using pearson method (can be found in Appendix Calculation 3) between Estimate number of TB cases, Estimate Mortality of TB cases,while controlling Estimate Mortality TB Non-HiV cases is 0.1700833, and the variance that is shared is 0.02892, or 2.892% which shows small relationship between the variables at 5% significance level.

It is considerably less than full correlation between Estimate number of TB cases and Estimate Mortality of TB cases 0.9686306 with 93.8% variance, when Estimate Mortality of TB Non-HIV cases was not controlled. Thus, it can be concluded that there is very small correlation between Estimate Number of TB cases and Estimate Mortality Of TB cases, but there exists a complex correlation between Estimate number of TB cases, Estimate Mortality of TB cases and Estimate Mortality TB Non-HiV cases.