MSCI 718 - Assignment2

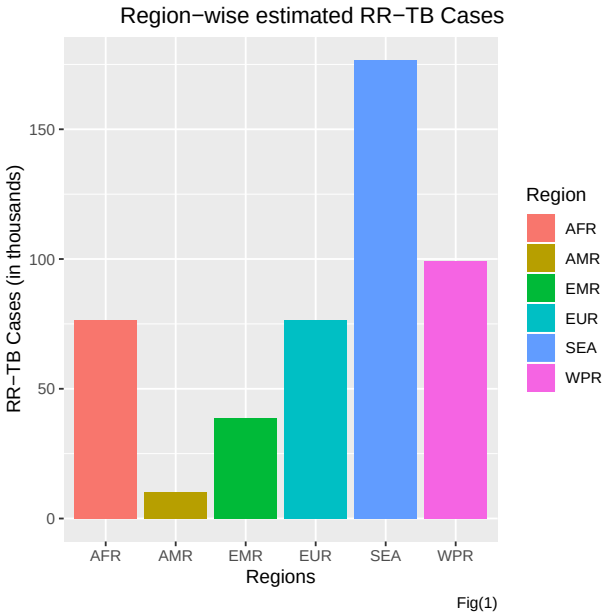
Karan Kohli and Rishabh Karwayun

28/02/2020

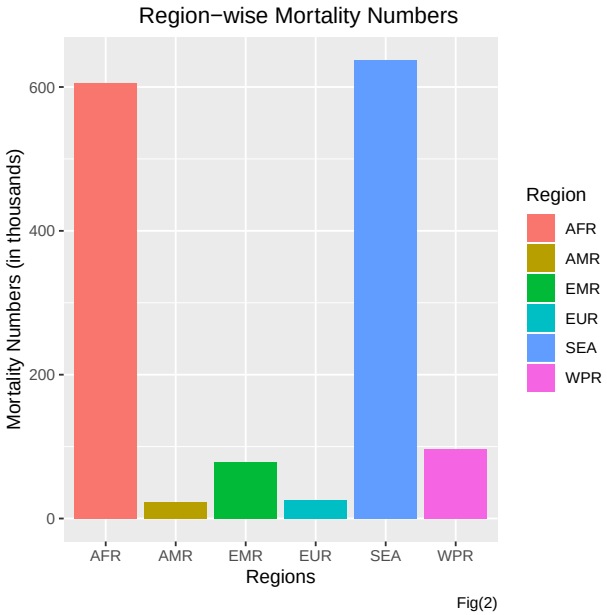
Tuberculosis Around the World

Tuberculosis is a bacterial disease. It is estimated that atleast one in four people in the world are carrying a form of TB bacteria. However, only after the bacteria is activated does the carrier becomes infected with TB. The datasets explored are released from WHO and contain WHO-generated estimates of TB mortality, incidence (including disaggregation by age and sex and incidence of TB/HIV), case fatality ratio, treatment coverage (previously called case detection rate), proportion of TB cases that have rifampicin-resistant TB (RR-TB, which includes cases with multidrug-resistant TB, MDR-TB), RR/MDR-TB among notified pulmonary TB cases etc.

WHO in its annual report states that globally more than 10 million people fell ill with TB. Approximately 1.5 million people died from the disease in 2018. Thus TB still remains to be a fatal disease globally. TB can be cured with appropriate medical assistance. One such drug used in TB treatments is Rifampicin. However, forms of TB bacterias have been found which are resitant to rifampicin. These are called RR-TB. TB can alse be resistant to multiple drugs that are used in the treatment. Such cases are called MDR-TB (Multiple Drug Resistant TB). Since these forms of bacteria are resistant to drugs used in treatments, it becomes much more difficult to monitor treatment and its outcomes than Drug-Susceptible TB. Hence, Drug-Resistant TB continues to be a public health threat.



Fig(1)



Fig(2)

In 2018, there were about half a million new cases of Rifampicin-Resistant TB, and close to 85000 cases of Multi Drug Resistant TB globally. It would be interesting to observe if number of cases of Drug Resistance

TB has some relation with Mortality Rate of TB. A positive correlation may help in securing more funding for research on treatment of Drug Resistance TB.

Data Exploration and Summary

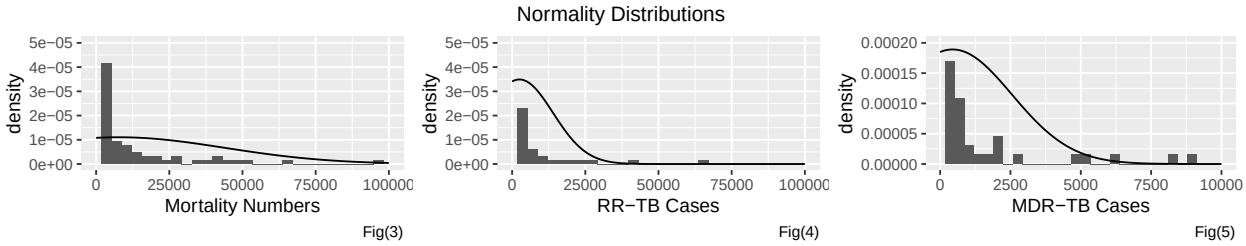
To start with, datasets were analysed to find respective variables. Variable e_mort_num which signifies countrywise mortality numbers was chosen from Table 1 (TB_burden_countries) which had 4040 observations of 50 variables. Then we selected feature e_inc_rr_num signifying Estimated number of RR-TB cases from Table 2 (MDR_RR_TB_burden) containing 216 observations of 26 variables.Finally we selected variable conf_mdr which signifies Laboratory Confirmed MDR-TB cases from Table 3 (TB_notifications) which had 8286 observations of 164 variables.

Since we only had data for RR-TB of the year 2018, we have considered all the data for this year only. The following steps were followed to get final data from above datasets. First we filtered the data in Tables 1 and 3 based on year (2018). Then, we selected the appropriate features from all three tables. After that, we joined filtered data from Table 1 with Table 2 using 'left-join'. Finally, we joined the result of the previous step with filtered data from Table 3 using 'left join'. There are 26 N/A values for MDR-TB cases which were removed. Hence we obtained our final dataset which is in Tidy format.

In our data, we have 216 observations of 5 variables. Two of the variables: Country and WHO-Region are factor variables, where as rest of them (Mortality numbers, RR-TB cases, MDR-TB cases) are integer variables having level of measurement as ratio. From Appendix 1, we can see that Mortality Numbers (e_mort_num) has a really diverse range of 449000 with standard deviation of 36130.5 and confidence interval ranging from 12895.044 to 2553.977. Similarly, Number of Incidence Cases of RR-TB (e_inc_rr_num) have a range of 130000 with a mean of 2515.4 and confidence interval ranging from 4150.5321 to 880.3521. Number of MDR Cases (conf_mdr) have a lower range value(24733) than the other 2 variables with confidence interval ranging from 744.7328 to 141.0040. All the other statistics of these varibales can be found in Appendix 1.

Plotting histograms and looking at the description of all the three ratio variables show that they all are poisitively skewed and leptokurtic.

From the graphs in Appendix 2 we observe that there are some outliers for mortality rate, RR-TB estimates and MDR-TB incident cases. We cannot remove the outliers because they represent countries having a very high mortality rate;RR-TB and/or MDR cases. In this case we have to consider these outliers as part of our analysis since they contain important information.



Fig(3) Fig(4) Fig(5)

To find correlation, we first check the assumptions for parametric models.

```
##
##  Shapiro-Wilk normality test
##
## data:  df$e_mort_num
## W = 0.19544, p-value < 2.2e-16
```

The Shapiro-Wilk test found that the mortality numbers were significantly non-normal at the 5% level of significance (W=0.19544, p<0.05). The non-normality can also be seen visually in the histogram plotted in

Fig (3). Similarly, conf_mdr and e_inc_rr_num were also found to be non-normal as seen in Fig(4) and Fig(5) and confirmed with Shapiro-Wilk normality test.

Hypothesis

We want to test correlation between Mortality Numbers and Estimated Cases of Drug Resistant TB.

Null Hypothesis: There is 0 correlation between Mortality Numbers and Estimated Cases of RR-TB.

Alternate Hypothesis: There is non-zero correlation between Mortality Numbers and Estimated Cases of RR-TB.

Analyzing the Selected Variables

As inferred above, since the features are not normal, we cannot proceed with Pearson's correlation method. Therefore, to test our hypothesis, we use Kendall-Tau's method to find correlation.

```
##
##  Kendall's rank correlation tau
##
## data:  df$e_mort_num and df$e_inc_rr_num
## z = 15.659, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.7758461
```

From the correlation test between Mortality Numbers and RR-TB estimated cases, we infer that Mortality cases are significantly correlated with RR-TB estimated cases, with correlation coefficient = 0.7758 ($p <$ 2.2e-16). A correlation of 0.7758 represents large effect explaing 60.18% of the variance. Since we have a substantially small p-value, we reject the Null Hypothesis and proceed with the alternate hypothesis stating that there is some correlation between the two stated variables.

In the data, there are two types of Drug Resistant TB: RR-TB and MDR-TB. Multi Drug Resistant TB (MDR-TB) is a wider sector and encompasses drugs other than Rifampicin. So, we wanted to observe how number of MDR-TB cases affect the correlation between Mortality Number and RR-TB cases and how strong is the correlation between Rifampicin Drug Resistant TB alone with Mortality numbers. Hence, we do some further tests.

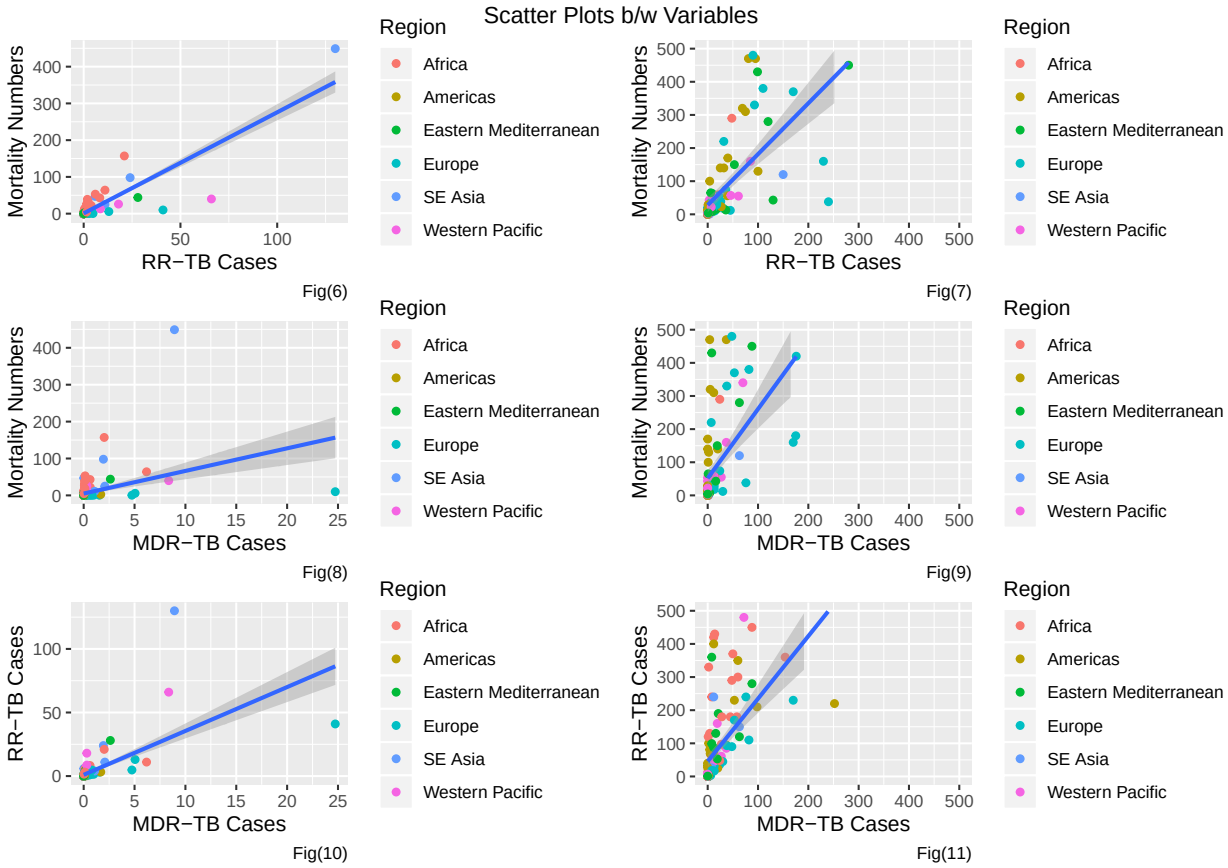
```
##              e_mort_num e_inc_rr_num  conf_mdr
## e_mort_num    1.0000000    0.7758461 0.5959234
## e_inc_rr_num  0.7758461    1.0000000 0.7391981
## conf_mdr      0.5959234    0.7391981 1.0000000
```

From the correlation matrix, we observe that all the variables under observation are strongly correlated to one another indicating that both RR-TB and MDR-TB highly affect the mortality cases. Moreover we see that the cases of RR-TB and MDR-TB (correlation coeff = 0.7391) are also highly correlated.

It would be interesting to observe how MDR-TB cases affect the correlation of RR-TB Cases with Mortality Numbers. And to analyse that, we perform a partial correlation test between RR-TB cases and Mortality Numbers with MDR-TB cases being the control variable.

```
##    estimate      p.value statistic   n gp  Method
## 1 0.6200382 8.806177e-37  12.66879 190  1 kendall
```

The partial correlation between Mortality Numbers, RR-TB cases and MDR-TB cases is 0.6200382, and the variance shared between them is 38.44%. This is a truer estimate of unique correlation between Mortality Numbers and RR-TB cases.



For Fig(6), Fig(8) and Fig(10), all the variables are in thousands.
Fig(6), Fig(8), Fig(10) are the original graphs while Fig(7), Fig(9), Fig(11)
are the scaled down versions of original graphs respectively, to show the correlation more accurately.

From the test we see that there is a large correlation between Mortality Numbers and RR-TB cases (correlation coeff = 0.6200382, p =8.806177e-37) when controlling the effect of MDR-TB cases.

This is less than the full correlation of 0.7758, which explained 60.18% of the variance when MDR-TB cases were not controlled for. We thus conclude that Mortality Numbers and RR-TB cases have a strong correlation and there is also a complex relationship between Mortality Numbers, RR-TB cases and MDR-TB cases.

Conclusion

From the analysis, we observe that Drug Resistant TBs are a major factor affecting Mortality numbers worldwide. It can be seen that RR-TB cases are strongly correlated to Mortality numbers. Looking at Multi Drug Resistant TB cases for further concrete analysis, we see that MDR-TB also plays a significant role in this correlation amongst Drug Resistance in TBs and Mortality numbers. A strong correlation indicates that as number of RR-TB cases increase we see an increase in Mortality numbers too. Whether there exists a causal relationship between the two or not is subject to further research.

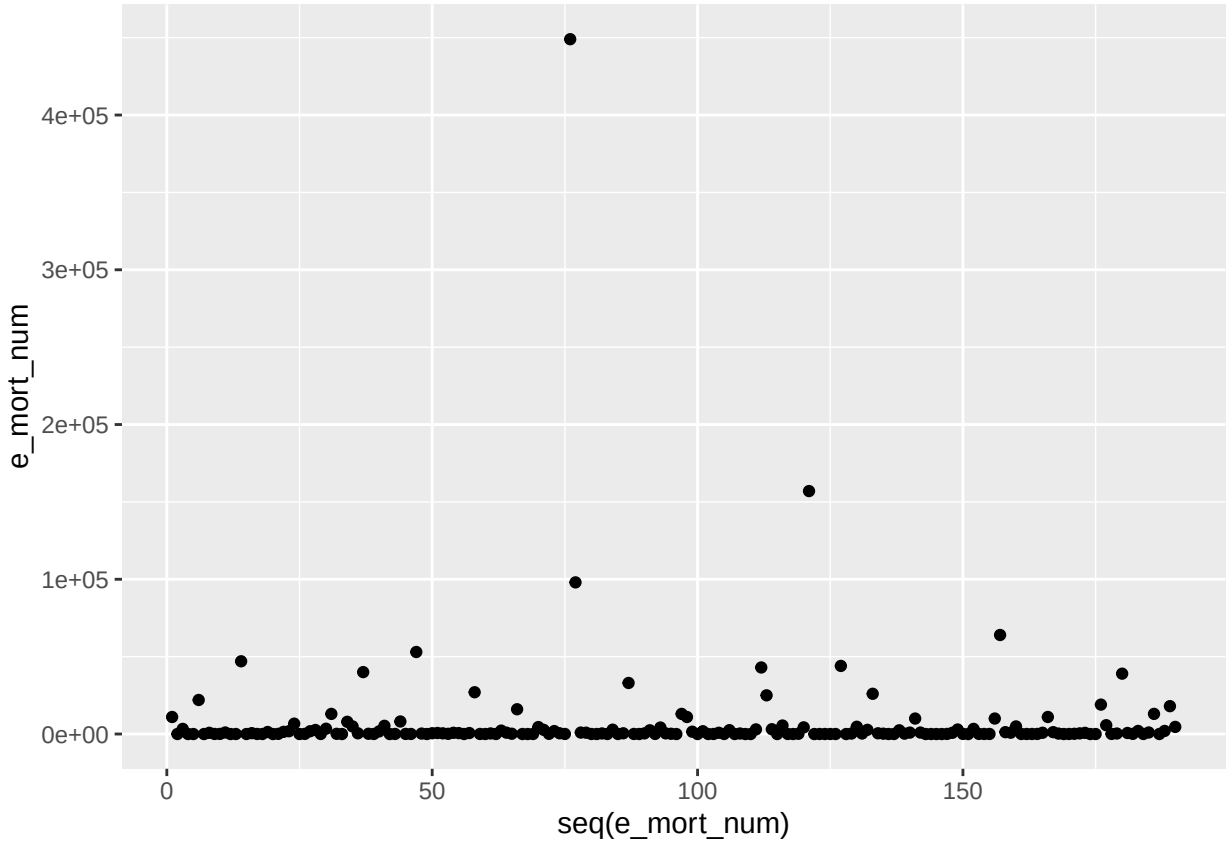
Appendix

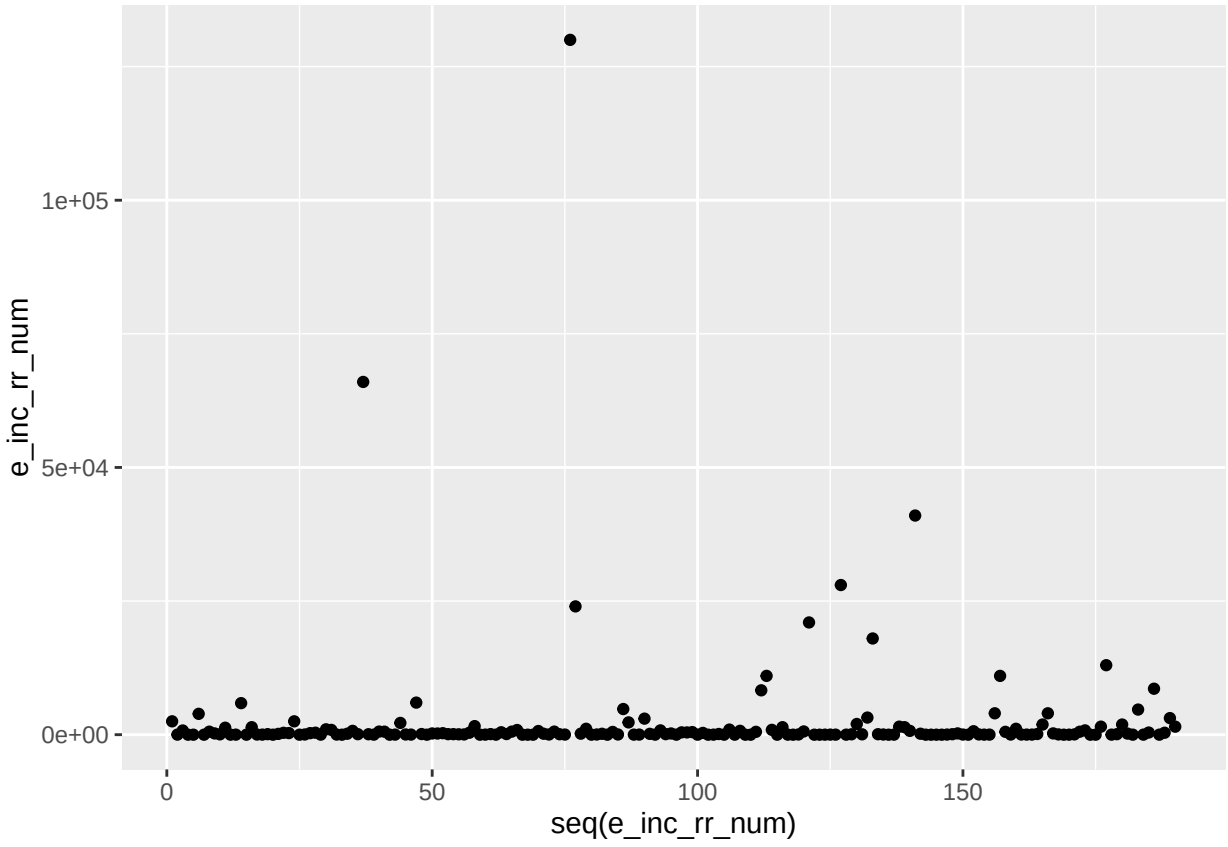
Appendix 1

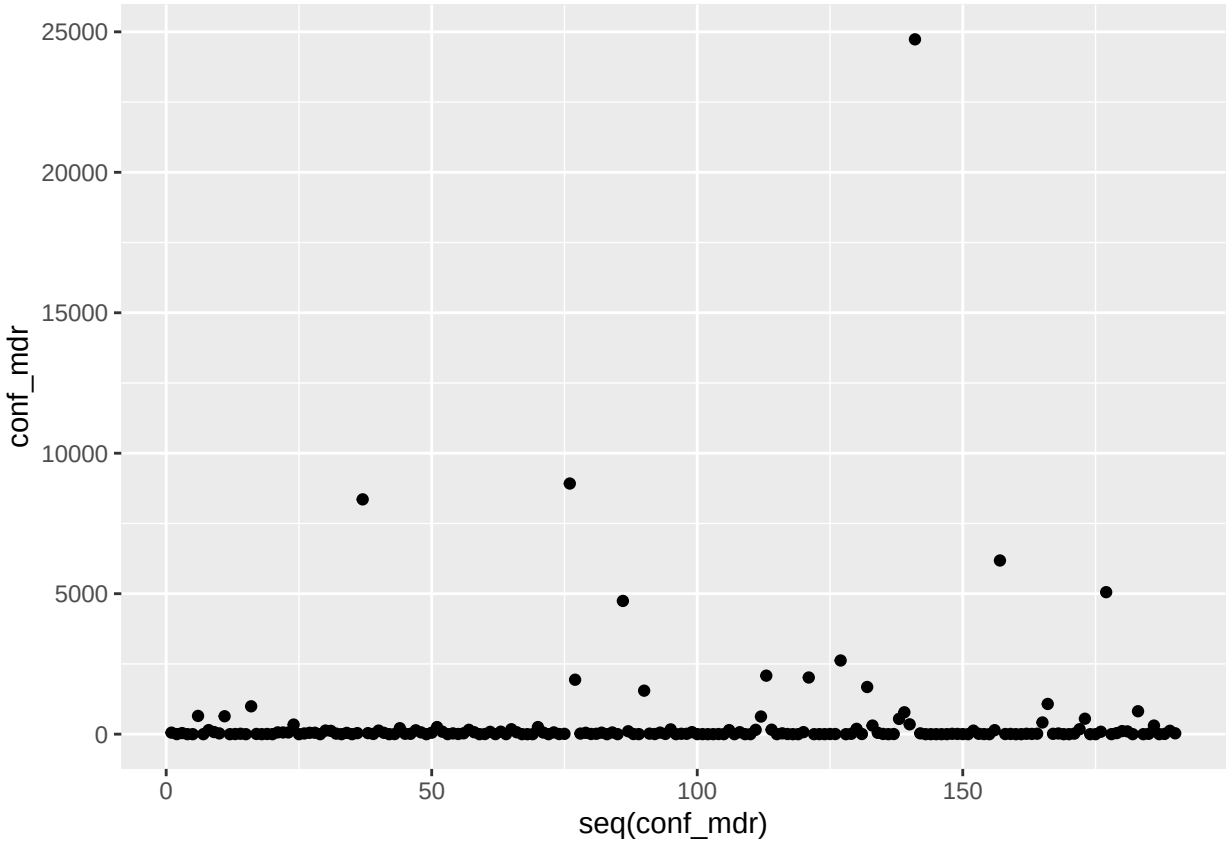
```
##            country      e_mort_num     g_whoregion  e_inc_rr_num
##  Afghanistan   :  1   Min.   :     0   AFR:45      Min.   :     0.00
##  Albania       :  1   1st Qu.:    20   AMR:39      1st Qu.:     8.25
##  Algeria       :  1   Median :   355   EMR:22      Median :   100.00
##  American Samoa:  1   Mean   :  7724   EUR:42      Mean   :  2515.44
##  Andorra       :  1   3rd Qu.:  2675   SEA:10      3rd Qu.:   717.50
##  Angola        :  1   Max.   :449000   WPR:32      Max.   :130000.00
##  (Other)       :184
##     conf_mdr
##  Min.   :    0.0
##  1st Qu.:    1.0
##  Median :   13.5
##  Mean   :  442.9
##  3rd Qu.:   95.5
##  Max.   :24733.0
##

## 'data.frame':    190 obs. of  5 variables:
##  $ country     : Factor w/ 216 levels "Afghanistan",..: 1 2 3 4 5 6 8 9 10 12 ...
##  $ e_mort_num  : int  11000 10 3300 0 0 22000 1 790 38 55 ...
##  $ g_whoregion : Factor w/ 6 levels "AFR","AMR","EMR",..: 3 4 1 6 4 1 2 2 4 6 ...
##  $ e_inc_rr_num: int  2500 14 780 0 0 3900 0 560 240 61 ...
##  $ conf_mdr    : int  54 2 39 1 0 649 0 144 76 27 ...

##          country   e_mort_num g_whoregion e_inc_rr_num     conf_mdr
## nbr.val       NA 1.900000e+02          NA 1.900000e+02 1.900000e+02
## nbr.null      NA 1.100000e+01          NA 2.900000e+01 4.000000e+01
## nbr.na        NA 0.000000e+00          NA 0.000000e+00 0.000000e+00
## min           NA 0.000000e+00          NA 0.000000e+00 0.000000e+00
## max           NA 4.490000e+05          NA 1.300000e+05 2.473300e+04
## range         NA 4.490000e+05          NA 1.300000e+05 2.473300e+04
## sum           NA 1.467657e+06          NA 4.779340e+05 8.414500e+04
## median        NA 3.550000e+02          NA 1.000000e+02 1.350000e+01
## mean          NA 7.724511e+03          NA 2.515442e+03 4.428684e+02
## SE.mean       NA 2.621183e+03          NA 8.289030e+02 1.530291e+02
## CI.mean       NA 5.170534e+03          NA 1.635090e+03 3.018644e+02
## var           NA 1.305415e+09          NA 1.305452e+08 4.449400e+06
## std.dev       NA 3.613052e+04          NA 1.142564e+04 2.109360e+03
## coef.var      NA 4.677386e+00          NA 4.542199e+00 4.762950e+00

##              min    max  range median   mean std.dev
## e_mort_num     0 449000 449000  355.0 7724.5 36130.5
## e_inc_rr_num   0 130000 130000  100.0 2515.4 11425.6
## conf_mdr       0  24733  24733   13.5  442.9  2109.4

##     upper      mean     lower
## 12895.044  7724.511  2553.977

##     upper      mean     lower
## 4150.5321 2515.4421  880.3521
```

```
##    upper     mean    lower
## 744.7328 442.8684 141.0040
```

Appendix 2


4e+05
3e+05
2e+05
1e+05
0e+00
e_mort_num
0
50
100
150
seq(e_mort_num)


1e+05
5e+04
0e+00
e_inc_rr_num
0
50
100
150
seq(e_inc_rr_num)



References

https://apps.who.int/iris/bitstream/handle/10665/329368/9789241565714-eng.pdf?ua=1

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression3.html