CUNY 606 - Final Project

Kishore Prasad

May 8, 2016

Part 1 - Introduction:

The goal of this research is to find if there is a link between Pollution levels and incidence of Tuberculosis. There have been studies that indicate that there is a relation between indoor pollution / smoking to TB. However, in this proposal I am looking to study if PM2.5 Air Pollution has any effect on incidence of TB.

This is an observational study

Part 2 - Data:

Part of the data is sourced from the data respository maintained by World Bank. Hence it is secondary data. This secondary data is collected from the World Bank website.

World Bank in turn gets the TB data from the Global Tuberculosis Report of World Health Organization (WHO)

similarly, the pollution data (Proportion of population above acceptable PM25) is sourced by World Bank from Brauer, M. et al. 2015. "Ambient Air Pollution Exposure Estimation for the Global Burden of Disease 2013." (Paper submitted for publication.), Institute for Health Metrics and Evaluation, University of Washington, Seattle.

The data is a part of the World Bank data. The required data for the above 2 variables are available for download from the website.

http://data.worldbank.org/indicator/EN.ATM.PM25.MC.ZS?page=5&display=default

http://data.worldbank.org/indicator/SH.TBS.INCD?display=default

The Response Variable is Incidence of TB

In this study, I am trying to explain an increase / decrease in the occurance of TB based on the increase / decrease in the pollution variables. The Response variable is Incidence of TB.

It is a numeric variable.

Incidence of tuberculosis (per 100,000 people) is the estimated number of new and relapse tuberculosis cases arising in a given year, expressed as the rate per 100,000 population. All forms of TB are included, including cases in people living with HIV.

```
TB_data <-read.csv("https://raw.githubusercontent.com/kishkp/CUNY-StatsAndProb/master/DataProject/Incid

TB_data<-TB_data[,c("Country Name", "1990", "1995", "2000", "2005", "2010", "2011", "2013")]

TB_data <- gather(TB_data, "Country Name", "TB_Value")

names(TB_data) <- c("Country", "Year", "TB")
```

The Explanatory variables are related to Pollution data. The below are the 2 explanatory variables that I will explore independently.

1. Proportion of Population above 10 micrograms per cubic meter of PM2.5. - This measures the population exposed to levels exceeding WHO guideline value (% of total) is defined as the portion of a country's population living in places where mean annual concentrations of PM2.5 are greater than 10 micrograms per cubic meter, the guideline value recommended by the World Health Organization as the lower end of the range of concentrations over which adverse health effects due to PM2.5 exposure have been observed.

Please note: This data is not actual PM2.5 level. This is proportion of a country's population that is living in areas having PM2.5 greater than 10 micrograms per cubic meter.

2. Actual PM2.5 - This is the actual PM2.5 levels in various countries.This data is sourced from the below data source. This data has been cleansed and posted to github.

Data source: http://www.google.ae/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&cad=rja&uact=8&ved=0ahUKEwjZ5sS

```
Pop_above_PM25_data <-read.csv("https://raw.githubusercontent.com/kishkp/CUNY-StatsAndProb/master/DataP

Pop_above_PM25_data<-Pop_above_PM25_data[,c("Country Name", "1990", "1995", "2000", "2005", "2010", "20

Pop_above_PM25_data <- gather(Pop_above_PM25_data, "Country Name", "Pop_above_PM25_Value")

names(Pop_above_PM25_data) <- c("Country", "Year", "Pop_Above_PM25")

Actual_PM25_data <-read.csv("https://raw.githubusercontent.com/kishkp/CUNY-StatsAndProb/master/DataProj

names(Actual_PM25_data) <- c("Country", "Actual_PM25", "Year")
```

Lets now prepare the data to be used in the analysis from these three different data sources.

Since we will be doing 2 analysis one being TB against population proportion and the other being TB against actual PM2.5, we will be generating 2 analysis data sets.

```
# generate the 2 datasets
data_Pop_prop_PM25 <- full_join(TB_data, Pop_above_PM25_data, by = c("Country", "Year"))

TB_data$Year<- as.numeric(TB_data$Year)
data_Actual_PM25 <- full_join(TB_data, Actual_PM25_data, by = c("Country", "Year"))


# Remove all rows where there are NAs
data_Pop_prop_PM25<-subset(data_Pop_prop_PM25, !is.na(TB))
data_Pop_prop_PM25<-subset(data_Pop_prop_PM25, !is.na(Pop_Above_PM25))

data_Actual_PM25<-subset(data_Actual_PM25, !is.na(TB))
data_Actual_PM25<-subset(data_Actual_PM25, !is.na(Actual_PM25))
```

Part 3 - Exploratory data analysis:

Summaries

Lets first see some data statistics:

```
summary(data_Pop_prop_PM25)
```

```
##    Country               Year                 TB          Pop_Above_PM25
##  Length:1284        Length:1284        Min.   :   0.88   Min.   :  0.00
##  Class :character   Class :character   1st Qu.:  20.00   1st Qu.: 40.20
##  Mode  :character   Mode  :character   Median :  65.50   Median : 98.89
##                                        Mean   : 144.03   Mean   : 72.65
##                                        3rd Qu.: 195.25   3rd Qu.:100.00
##                                        Max.   :1292.00   Max.   :100.00
```

```
summary(data_Actual_PM25)
```

```
##    Country               Year            TB           Actual_PM25
##  Length:79          Min.   :2010   Min.   :  2.20   Min.   :  5.00
##  Class :character   1st Qu.:2010   1st Qu.:  7.05   1st Qu.: 13.50
##  Mode  :character   Median :2011   Median : 16.00   Median : 21.00
##                     Mean   :2011   Mean   : 66.22   Mean   : 25.42
##                     3rd Qu.:2011   3rd Qu.: 60.50   3rd Qu.: 28.00
##                     Max.   :2013   Max.   :922.00   Max.   :101.00
```

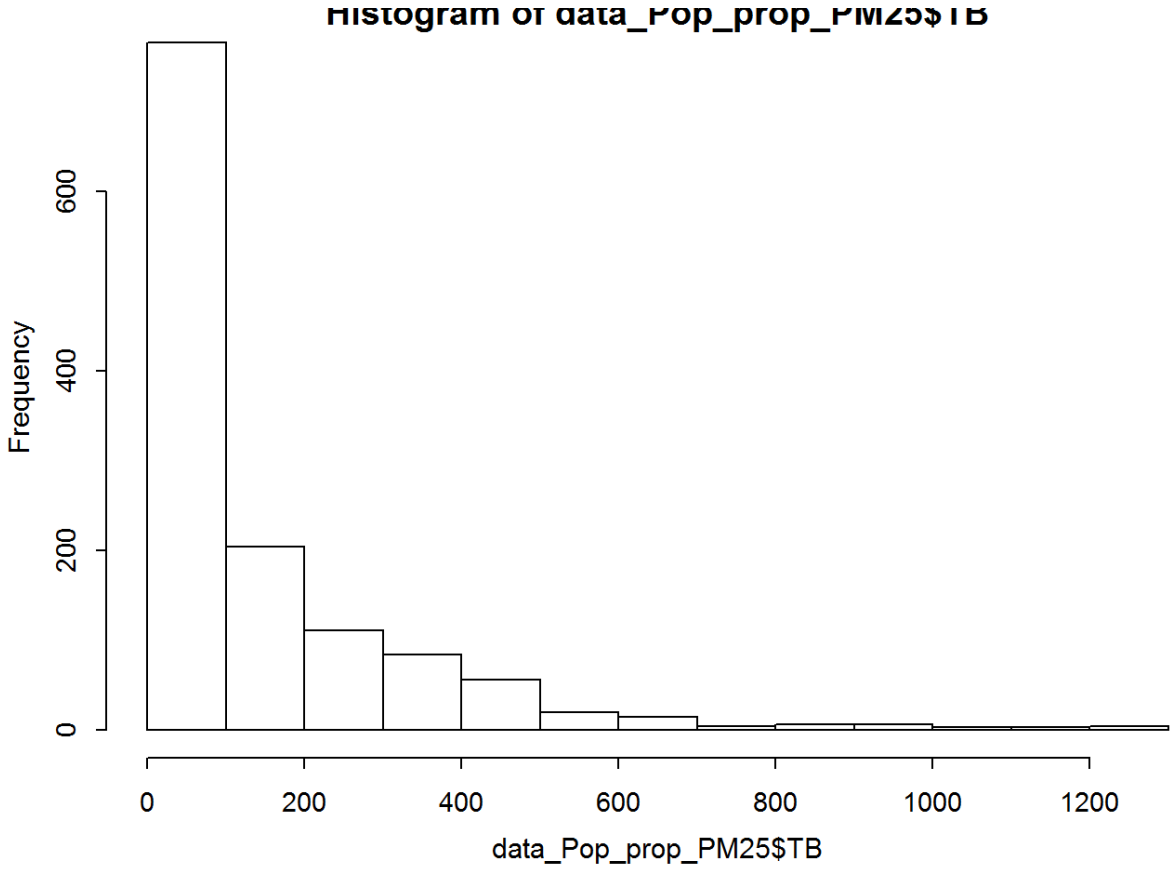
We have about 1284 observations in data_Pop_prop_PM25 dataset and 79 observations in data_Actual_PM25.The Summaries seem to indicate a skew for the TB counts variable in both the datasets.

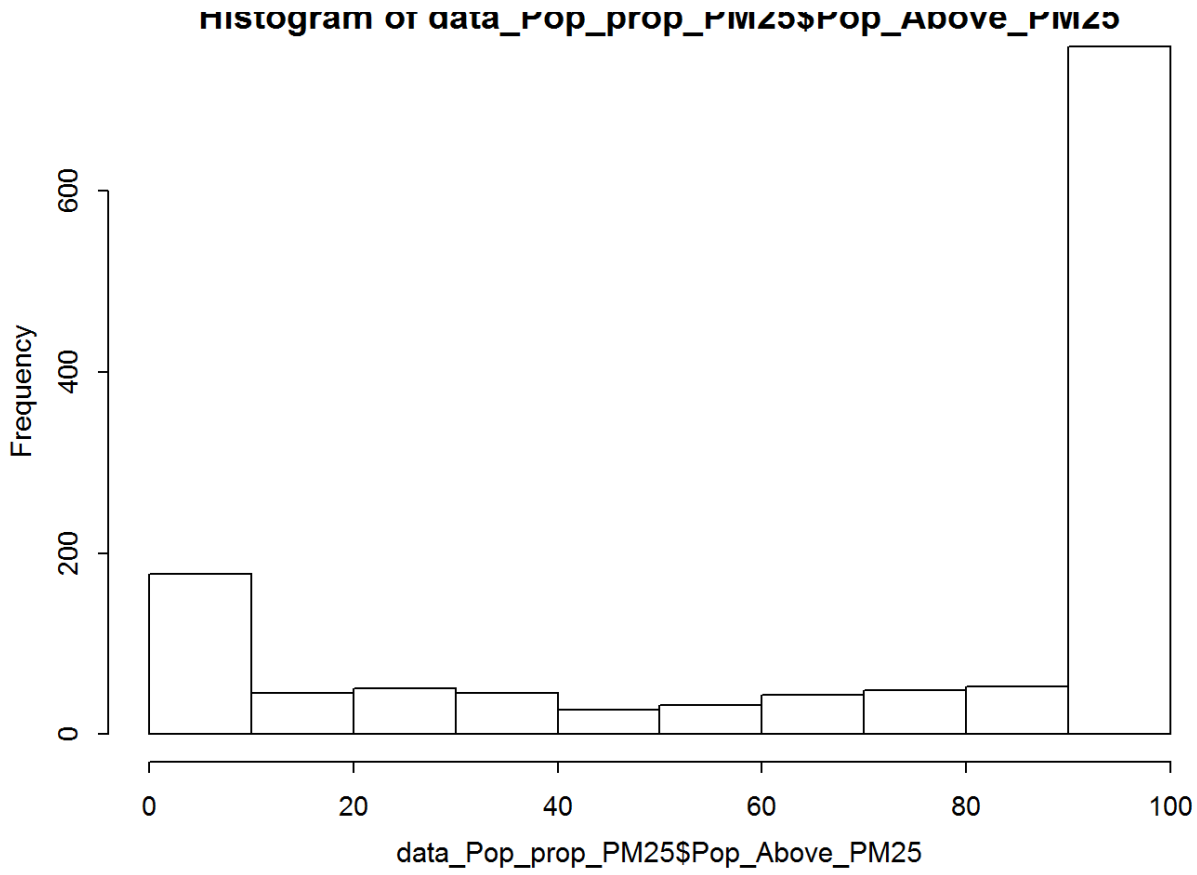
Histograms

Lets look at the distribution of the variables in both datasets. Lets start with data_Pop_prop_PM25.

```
hist(data_Pop_prop_PM25$TB)

hist(data_Pop_prop_PM25$Pop_Above_PM25)
```


Histogram of data_Pop_prop_PM25$TB
Frequency
0
200
400
600
0
200
400
600
800
1000
1200
data_Pop_prop_PM25$TB

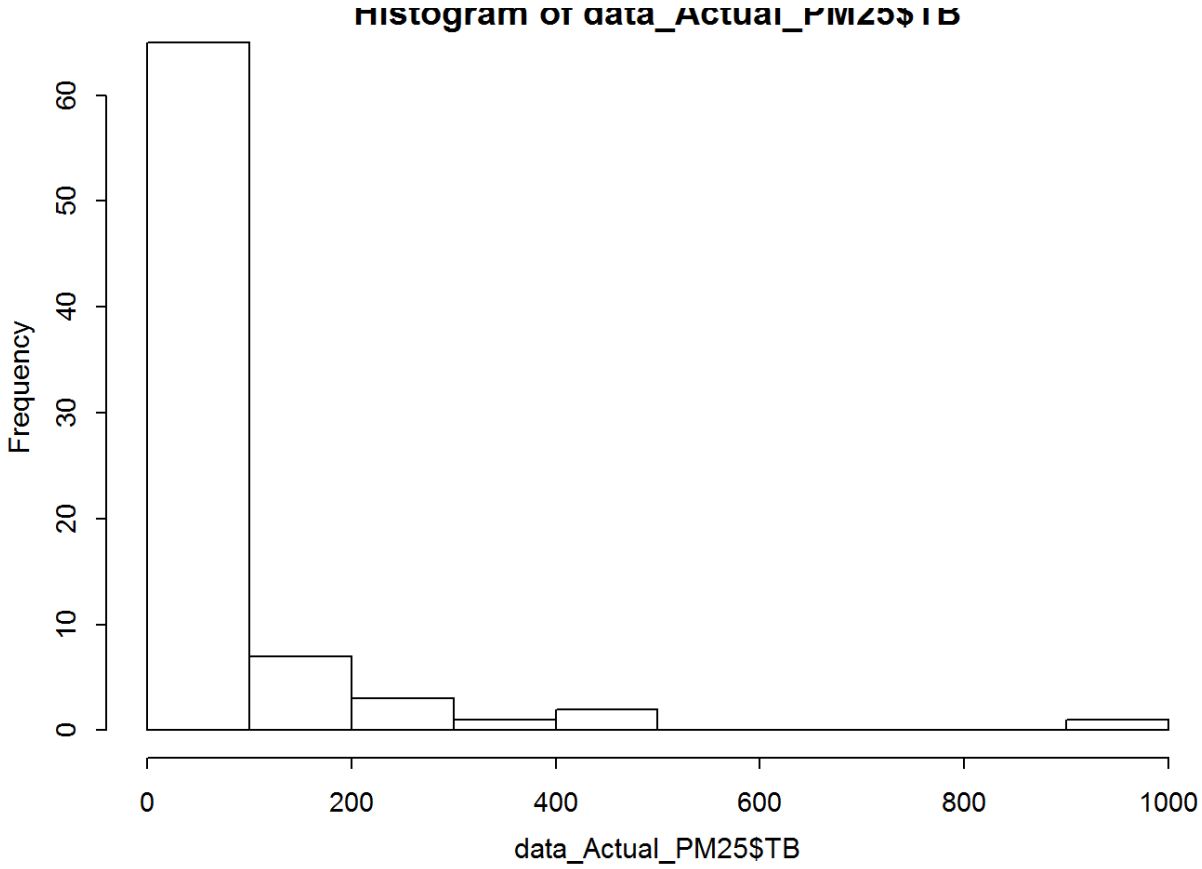


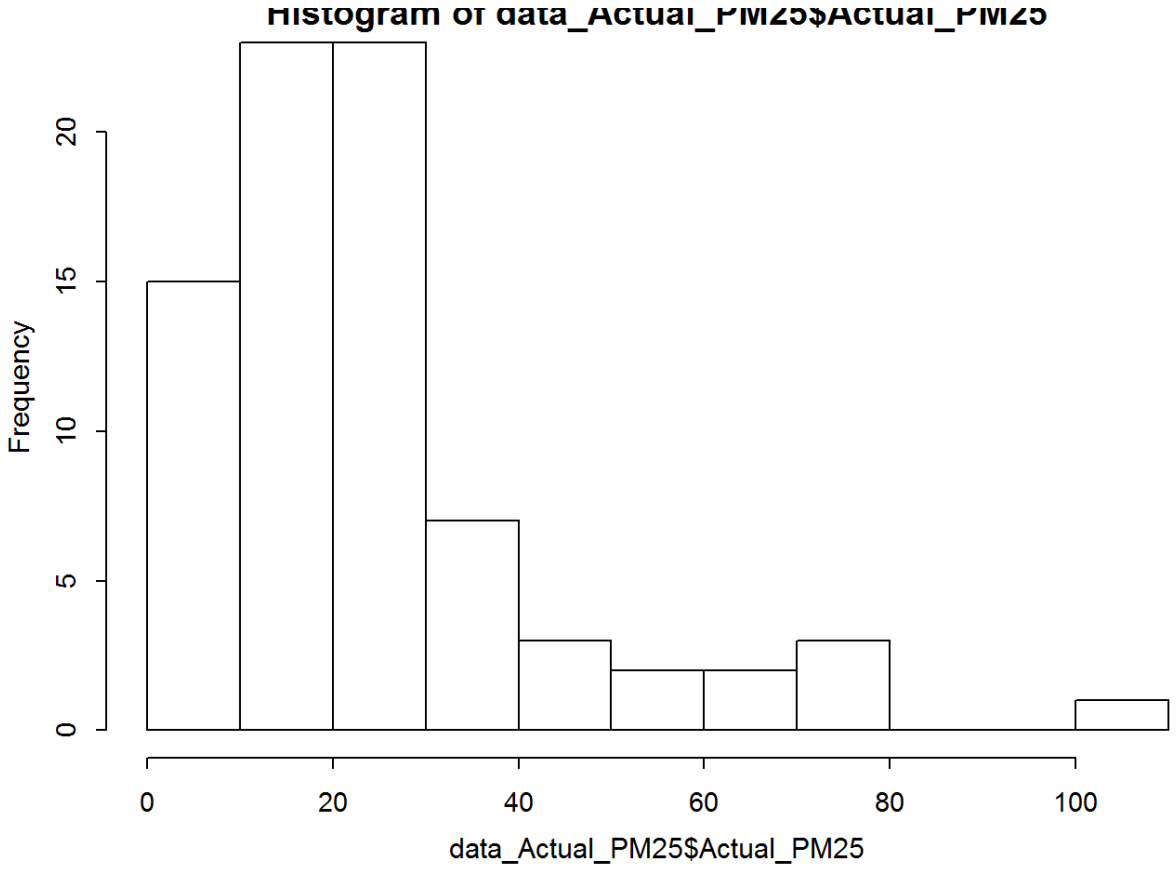
There is a strong right skew in TB values. The population above PM25 also does not have a normal distribution. However, since the number of observations are more, we can go ahead with the analysis.

Similarly, lets look at the other dataset data_Actual_PM25.

```
hist(data_Actual_PM25$TB)

hist(data_Actual_PM25$Actual_PM25)
```


Histogram of data_Actual_PM25$TB
Frequency
0
10
20
30
40
50
60
0
200
400
600
800
1000
data_Actual_PM25$TB

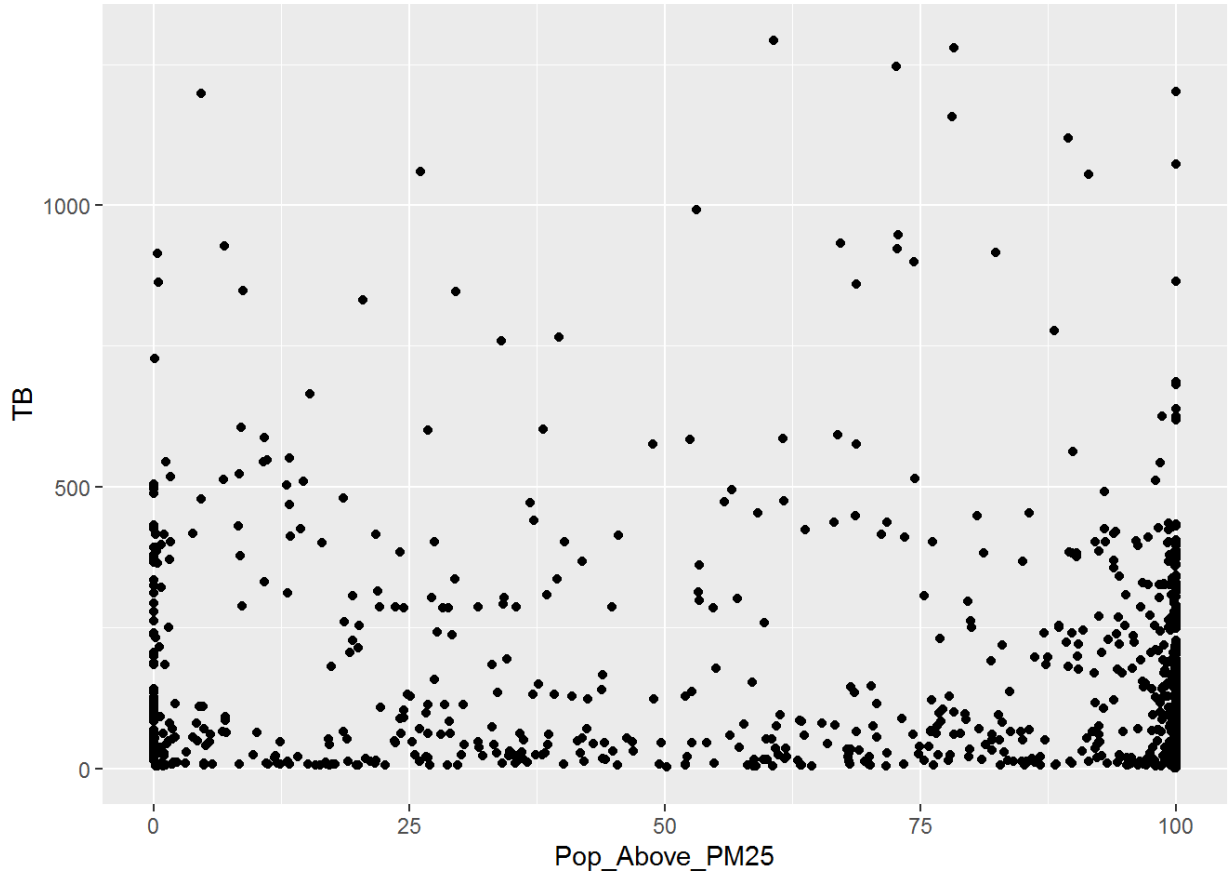


Here again, there is a strong right skew in TB values. The actual PM25 also seems to have a right skew.

Scatter Plots

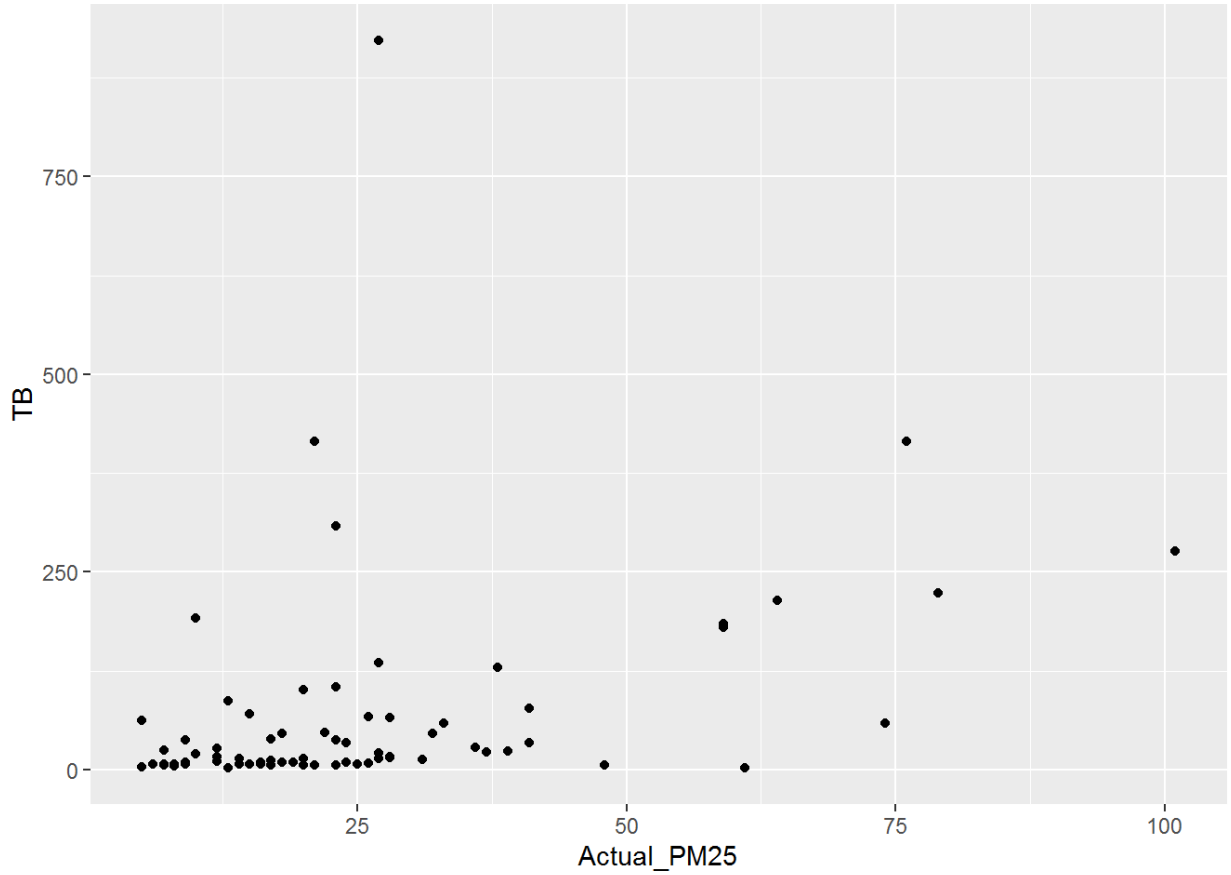
Lets look at a scatter plot of the 2 variables in both the datasets.

```
ggplot(data=data_Pop_prop_PM25, aes(x=Pop_Above_PM25, y=TB)) + geom_point()
```



data_Pop_prop_PM25 does not seem to provide a trend as such in the data.

```
ggplot(data=data_Actual_PM25, aes(x=Actual_PM25, y=TB)) + geom_point()
```



There seems to be a trend between the variables of interest in data_Actual_PM25.

Correlations

Lets next look if we can quantity the correlations.

```
cor(x = data_Pop_prop_PM25$Pop_Above_PM25, y=data_Pop_prop_PM25$TB)
```

```
## [1] -0.111472
```

This seems to indicate a negative correlation !!! Seems a bit counter intutive. Basically what it means is that more the proportion of population in areas with unacceptabe PM25 lesser is the incidence of TB. However, the correlation does not seem to be particularly strong since it is only about 0.11 in strength.

I don't think we have a basis for proceeding with this dataset (data_Pop_prop_PM25).

Next lets look at the correlation between Tb and actual PM25 values.

```
cor(x = data_Actual_PM25$Actual_PM25, y=data_Actual_PM25$TB)
```

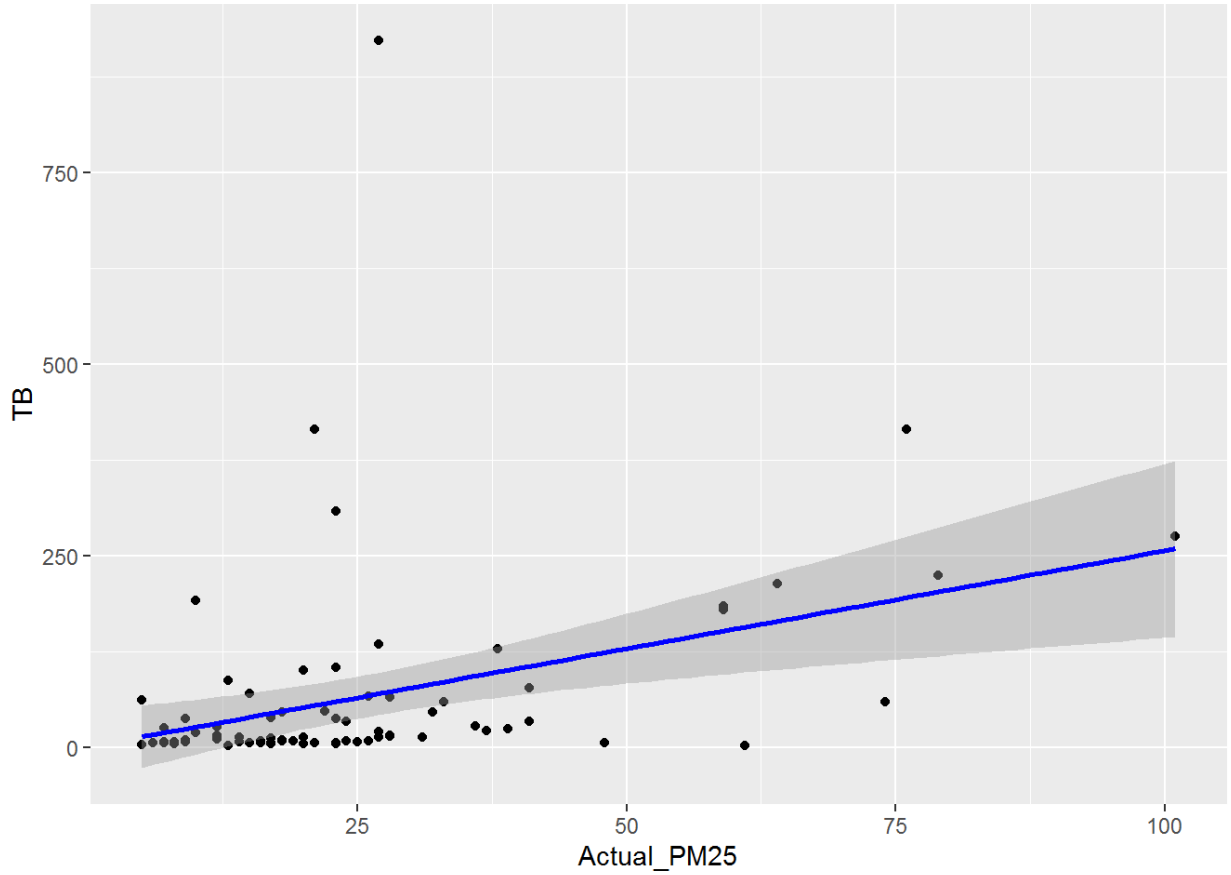
```
## [1] 0.3655598
```

This shows a positive correlation, though it is not very strong. What it means is that more the actual PM25 more is the instances of TB.

We will now explore further on this dataset (data_Actual_PM25).

We will now try to fit a trend line to this dataset to visually see the trend.

```
ggplot(data=data_Actual_PM25, aes(x=Actual_PM25, y=TB)) + geom_point() +   geom_smooth(method = "lm", c
```



The above plot does seem to suggest the trend and is consistent with the correlation output. There are however many outliers that impact the slope.

Lets fit a linear regression model on this data.

```
m1 <- lm(TB ~ Actual_PM25, data = data_Actual_PM25)
summary(m1)
```

```
##
## Call:
## lm(formula = TB ~ Actual_PM25, data = data_Actual_PM25)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -154.97  -47.79  -23.03   -0.98  851.74
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.2434    23.3577   0.053 0.957683
## Actual_PM25   2.5562     0.7417   3.446 0.000924 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.6 on 77 degrees of freedom
## Multiple R-squared:  0.1336, Adjusted R-squared:  0.1224
## F-statistic: 11.88 on 1 and 77 DF,  p-value: 0.0009236
```

Using the summary above, we can write down the least squares regression line for the linear model:

$$\hat{TB} = 1.2434384 + 2.5561595 * ActualPM25$$

Model Diagnostics

To assess whether the linear model is reliable, we need to check for (1) Linearity, (2) Nearly normal residuals, and (3) Constant variability.

Lets evaluate each one below:

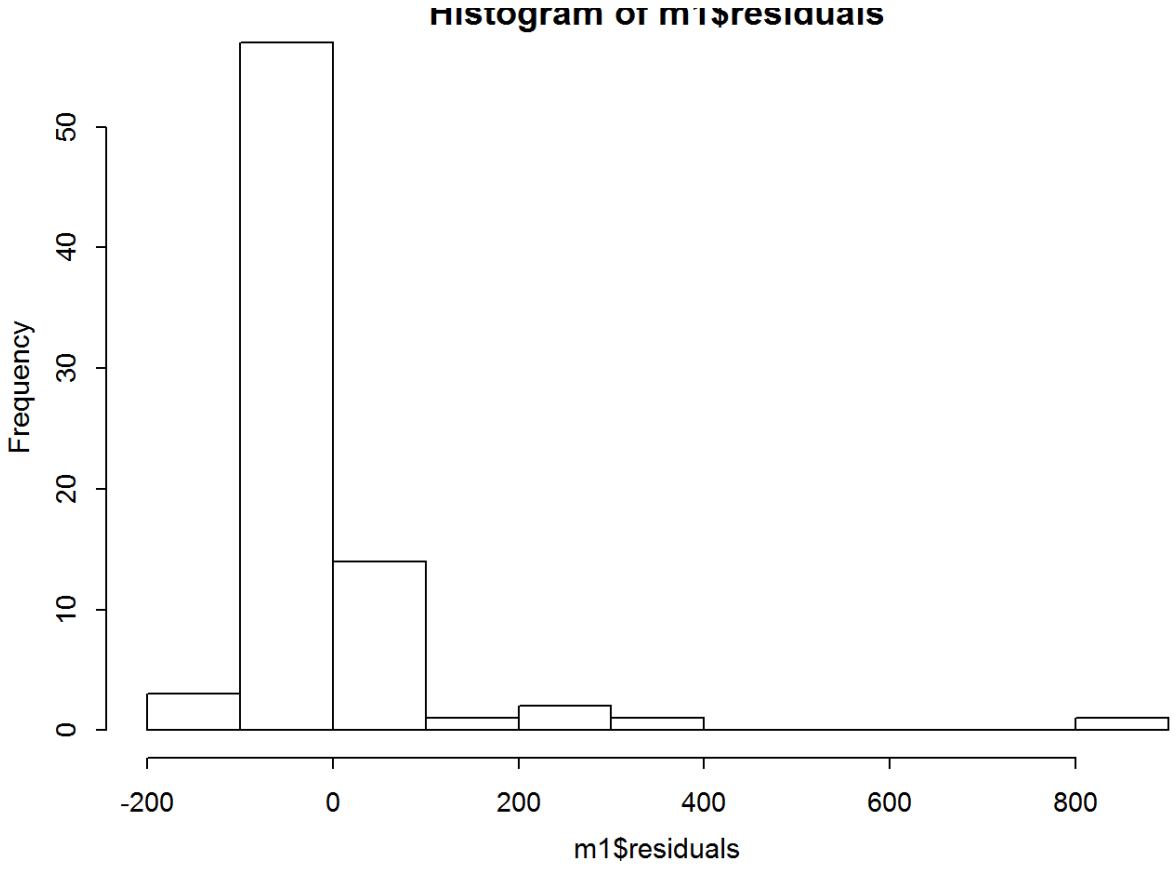
Linearity:

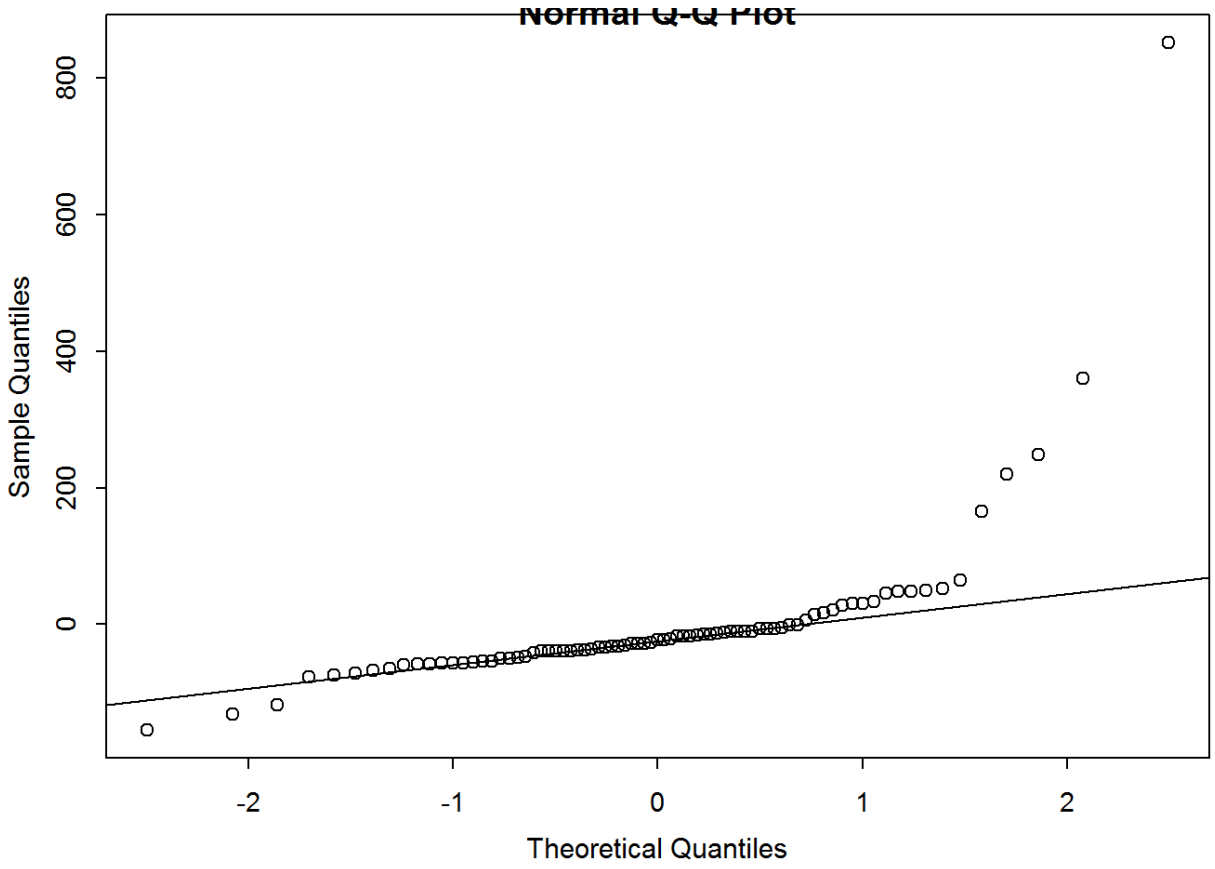
The data points in the scatter plot does indicate a trend even though it is influenced by the outliers. I would consider this to satisfy the linearlity requirement.

Nearly normal residuals:

Let's look at the below plots to determine this requirement.

```
hist(m1$residuals)
qqnorm(m1$residuals)
qqline(m1$residuals)  # adds diagonal line to the normal prob plot
```


Histogram of m1$residuals
Frequency
0
10
20
30
40
50
-200
0
200
400
600
800
m1$residuals

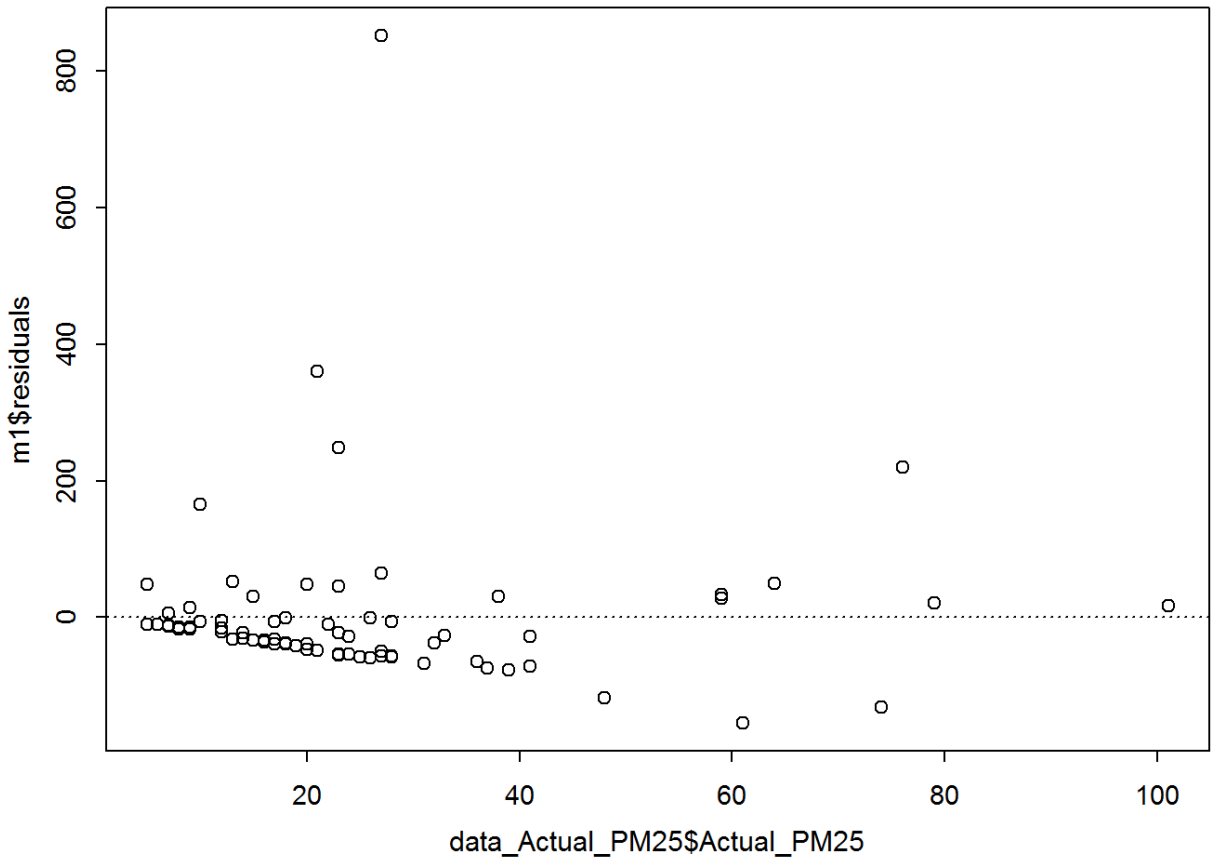


The above histogram indicate a slight right skew in the data influenced by a couple of data points. The Q-Q plot also validates this. I would be a bit cautious of accepting this requirement as being satisfied.

Constant Variability

We will use the below plot to determine the constant variability

```
plot(m1$residuals ~ data_Actual_PM25$Actual_PM25)
abline(h = 0, lty = 3)  # adds a horizontal dashed line at y = 0
```



Constant variability: There are quite a few outliers. Visually, it seems that there are more points below the line towards the lower end of PM25. Again here, I would be a bit cautious of accepting this requirement as being satisfied.

Independent Observations

Given that there are 196 countries in the world and the dataset spans across 3 different years (from 2011, 2012 and 2013) the population in this case would have been 588 observations. Out of these we now have 79 observations in the study which is slightly above 10% of the population. since this is close to the 10% threshold, we can consider this condition to be satisfied.

Part 4 - Inference:

Interpreting the Regression equation

```
summary(m1)
```

```
##
## Call:
## lm(formula = TB ~ Actual_PM25, data = data_Actual_PM25)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -154.97  -47.79  -23.03   -0.98  851.74
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.2434    23.3577   0.053 0.957683
## Actual_PM25   2.5562     0.7417   3.446 0.000924 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.6 on 77 degrees of freedom
## Multiple R-squared:  0.1336, Adjusted R-squared:  0.1224
## F-statistic: 11.88 on 1 and 77 DF,  p-value: 0.0009236
```

The intercept of 1.2434384 is the number of TB cases when the actual PM25 is 0. The Slope of 2.5561595 indicates that number of TB cases increases by 2.5561595 cases for every 1 unit increase in actual PM25.

The other piece of information from the summary output is the Multiple R-squared, or more simply, R^2. The R^2 value represents the proportion of variability in the response variable that is explained by the explanatory variable. For this model, 13.36% of the variability in TB cases is explained by Actual PM25.

The p-value is very low (well below 0.05). We can say that Actual PM25 is a good predictor of TB cases.

Part 5 - Conclusion:

PM25 is one of the measures of pollution across the world along with others like CO2, PM10 etc. It has been observed from this study that a rise in actual PM2 value is a predictor of occurances of TB cases. Even though the correlation is not very strong, it is present nonetheless. However, I am not very confident of the data and the conditions for reliability of the linear model.

On the other hand the proportion of population staying in areas with unacceptable PM25 values does not seem to be a good indicator of occurances of TB cases.

References:

Data sources:

http://data.worldbank.org/indicator/EN.ATM.PM25.MC.ZS?page=5&display=default

http://data.worldbank.org/indicator/SH.TBS.INCD?display=default

http://www.google.ae/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&cad=rja&uact=8&ved=0ahUKEwjZ5sSI4tnMAhXJ8