

```
# load data
```

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

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

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

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.

Cases

What are the cases, and how many are there?

There are 214 countries with Pollution data every 5 years between 1990 to 2010 and then in 2011 and in 2013. The data for TB is available from 1990 onwards to 2014.

Data collection

Describe the method of data collection.

The data is sourced from the data repository 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 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.

Type of study

What type of study is this (observational/experiment)?

This is an Observational study

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

The data is a part of the World Bank data. The required data for the 2 variables in question 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>

Response

What is the response variable, and what type is it (numerical/categorical)?

In this study, I am trying to explain an increase / decrease in the occurrence of TB based on the increase / decrease in the pollution rate. 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.

Explanatory

What is the explanatory variable, and what type is it (numerical/categorical)?

The Explanatory variable is the Pollution data

PM2.5 air pollution, 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.

Relevant summary statistics

Provide summary statistics relevant to your research question. For example, if you are comparing means across groups provide means, SDs, sample sizes of each group. This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

The summary for Incidence of TB are as below:

```
summary(TB_data_raw)
```

##	Country.Name	Country.Code	X1990	X1991
##	Length:214	Length:214	Min. : 0.00	Min. : 0.0
##	Class :character	Class :character	1st Qu.: 25.25	1st Qu.: 25.0
##	Mode :character	Mode :character	Median : 65.50	Median : 66.0
##			Mean :133.68	Mean :135.6
##			3rd Qu.:193.50	3rd Qu.:198.5
##			Max. :864.00	Max. :933.0
##			NA's :12	NA's :12
##	X1992	X1993	X1994	X1995
##	Min. : 2.00	Min. : 1.8	Min. : 1.6	Min. : 1.5
##	1st Qu.: 24.25	1st Qu.: 26.0	1st Qu.: 25.0	1st Qu.: 24.5
##	Median : 67.00	Median : 70.0	Median : 71.0	Median : 71.0
##	Mean : 137.84	Mean : 140.5	Mean : 144.1	Mean : 146.2
##	3rd Qu.: 198.00	3rd Qu.: 209.8	3rd Qu.: 219.0	3rd Qu.: 215.8
##	Max. :1013.00	Max. :1090.0	Max. :1156.0	Max. :1201.0
##	NA's :12	NA's :12	NA's :12	NA's :12
##	X1996	X1997	X1998	X1999
##	Min. : 1.5	Min. : 1.50	Min. : 1.60	Min. : 0.96
##	1st Qu.: 26.0	1st Qu.: 24.25	1st Qu.: 23.25	1st Qu.: 23.25
##	Median : 73.5	Median : 76.00	Median : 76.50	Median : 72.50

```

## Mean : 148.9 Mean : 151.01 Mean : 152.60 Mean : 153.37
## 3rd Qu.: 216.2 3rd Qu.: 214.25 3rd Qu.: 218.50 3rd Qu.: 215.50
## Max. :1219.0 Max. :1211.00 Max. :1181.00 Max. :1134.00
## NA's :12 NA's :12 NA's :12 NA's :12
## X2000 X2001 X2002 X2003
## Min. : 1.40 Min. : 1.4 Min. : 1.2 Min. : 1.1
## 1st Qu.: 22.25 1st Qu.: 21.0 1st Qu.: 20.0 1st Qu.: 18.5
## Median : 69.00 Median : 68.0 Median : 67.0 Median : 66.0
## Mean : 153.23 Mean : 152.9 Mean : 153.3 Mean : 152.2
## 3rd Qu.: 216.00 3rd Qu.: 215.5 3rd Qu.: 217.5 3rd Qu.: 209.5
## Max. :1073.00 Max. :1120.0 Max. :1162.0 Max. :1218.0
## NA's :12 NA's :12 NA's :11 NA's :11
## X2004 X2005 X2006 X2007
## Min. : 0.95 Min. : 0.74 Min. : 1.2 Min. : 1.6
## 1st Qu.: 17.50 1st Qu.: 17.00 1st Qu.: 17.0 1st Qu.: 17.0
## Median : 66.00 Median : 64.00 Median : 64.0 Median : 62.0
## Mean : 151.13 Mean : 147.47 Mean : 144.9 Mean : 141.2
## 3rd Qu.: 206.50 3rd Qu.: 195.00 3rd Qu.: 189.0 3rd Qu.: 186.0
## Max. :1259.00 Max. :1292.00 Max. :1332.0 Max. :1354.0
## NA's :11 NA's :9 NA's :9 NA's :9
## X2008 X2009 X2010 X2011
## Min. : 1.5 Min. : 1.4 Min. : 1.4 Min. : 0.76
## 1st Qu.: 17.0 1st Qu.: 16.0 1st Qu.: 14.5 1st Qu.: 14.75
## Median : 60.0 Median : 57.0 Median : 54.0 Median : 53.00
## Mean : 137.0 Mean : 133.1 Mean : 128.4 Mean : 124.16
## 3rd Qu.: 183.0 3rd Qu.: 169.0 3rd Qu.: 164.0 3rd Qu.: 154.50
## Max. :1347.0 Max. :1308.0 Max. :1246.0 Max. :1157.00
## NA's :9 NA's :9 NA's :7 NA's :6
## X2012 X2013 X2014 X2015
## Min. : 0.75 Min. : 0.9 Min. : 0.0 Mode:logical
## 1st Qu.: 14.00 1st Qu.: 13.0 1st Qu.: 12.0 NA's:214
## Median : 52.00 Median : 49.0 Median : 48.5
## Mean : 120.90 Mean :117.0 Mean :113.0
## 3rd Qu.: 160.25 3rd Qu.:159.0 3rd Qu.:159.5
## Max. :1042.00 Max. :916.0 Max. :852.0
## NA's :6 NA's :6 NA's :6

```

Similarly, the summary for Pollution are as below:

```
summary(Pollution_data_raw)
```

```

## Country.Name Country.Code X1990 X1991
## Length:214 Length:214 Min. : 0.00 Mode:logical
## Class :character Class :character 1st Qu.: 40.55 NA's:214
## Mode :character Mode :character Median : 99.75
## Mean : 72.90
## 3rd Qu.:100.00
## Max. :100.00
## NA's :28
## X1992 X1993 X1994 X1995
## Mode:logical Mode:logical Mode:logical Min. : 0.00
## NA's:214 NA's:214 NA's:214 1st Qu.: 35.74
## Median : 99.27

```

```

##                                     Mean   : 72.10
##                                     3rd Qu.:100.00
##                                     Max.   :100.00
##                                     NA's    :29
##      X1996      X1997      X1998      X1999
## Mode:logical Mode:logical Mode:logical Mode:logical
## NA's:214      NA's:214      NA's:214      NA's:214
##
##
##
##
##      X2000      X2001      X2002      X2003
## Min.   : 0.00 Mode:logical Mode:logical Mode:logical
## 1st Qu.: 36.10 NA's:214      NA's:214      NA's:214
## Median : 98.76
## Mean   : 71.88
## 3rd Qu.:100.00
## Max.   :100.00
## NA's    :29
##      X2004      X2005      X2006      X2007
## Mode:logical Min.   : 0.00 Mode:logical Mode:logical
## NA's:214      1st Qu.: 41.70 NA's:214      NA's:214
##                                     Median : 99.40
##                                     Mean   : 73.62
##                                     3rd Qu.:100.00
##                                     Max.   :100.00
##                                     NA's    :28
##      X2008      X2009      X2010      X2011
## Mode:logical Mode:logical Min.   : 0.00 Min.   : 0.00
## NA's:214      NA's:214      1st Qu.: 45.50 1st Qu.: 43.00
##                                     Median : 99.13 Median : 99.02
##                                     Mean   : 72.71 Mean   : 72.73
##                                     3rd Qu.:100.00 3rd Qu.:100.00
##                                     Max.   :100.00 Max.   :100.00
##                                     NA's    :28 NA's    :29
##      X2012      X2013      X2014      X2015
## Mode:logical Min.   : 0.00 Mode:logical Mode:logical
## NA's:214      1st Qu.: 49.44 NA's:214      NA's:214
##                                     Median : 95.84
##                                     Mean   : 72.97
##                                     3rd Qu.: 99.96
##                                     Max.   :100.00
##                                     NA's    :29

```