Association between levels of pollution and mortality in US cities

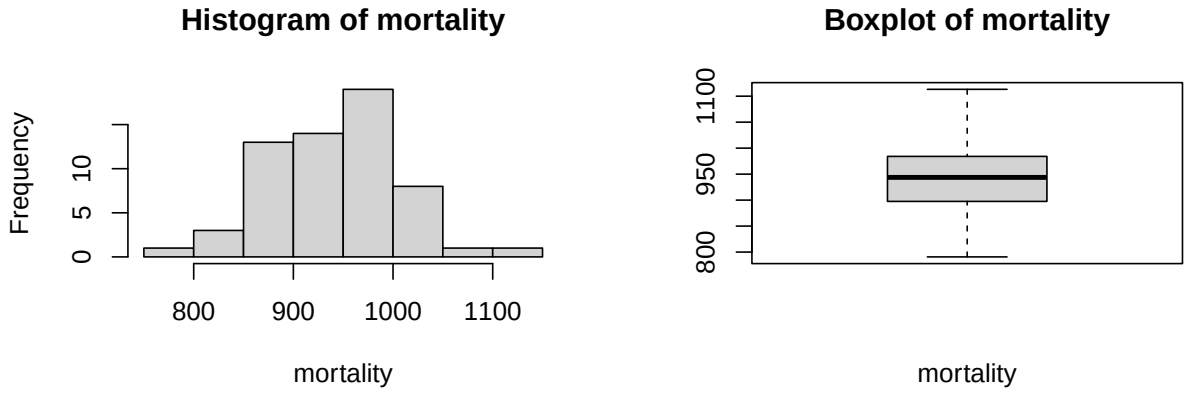
In this report, we investigate the possible association between levels of pollution and mortality rates in US cities. The specific aim is to model mortality rates based on the given variables, including levels of Nitrogen oxides (NOX), levels of Sulphur dioxides (SOX), mean annual precipitation (MP) and income level.

```
##   mortality Nox Sox Prec Income
## 1   921.870  15  59   36   High
## 2   997.875  10  39   35    Low
## 3   962.354   6  33   44   High
## 4   982.291   8  24   47    Low
## 5  1071.289  38 206   43    Low
## 6  1030.380  32  72   53    Low
```

Exploratory analyses

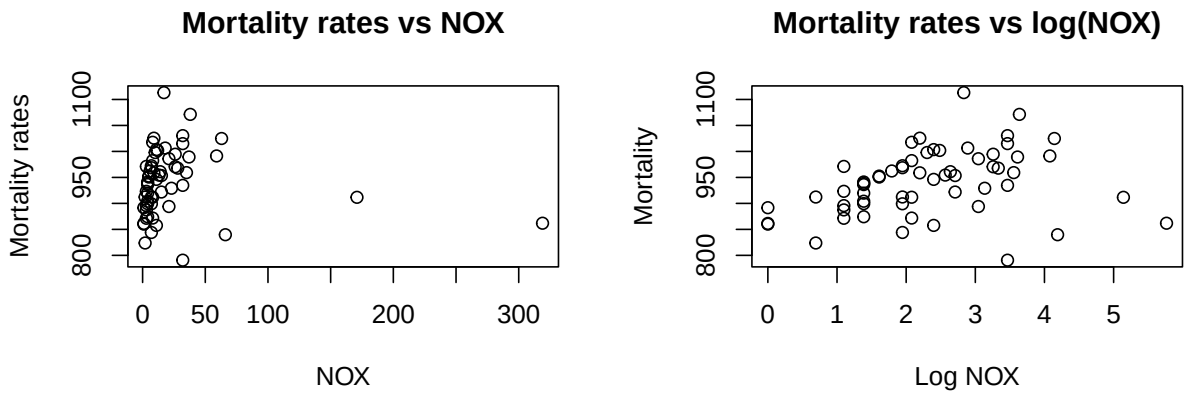
We started by looking at the distribution of mortality as shown in the figure below.

```
layout(t(1:2))
hist(Pol$mortality, main = "Histogram of mortality", xlab = "mortality")
boxplot(Pol$mortality, main = "Boxplot of mortality", xlab = "mortality")
```



Both plots show that mortality rates follows a normal distribution with no obvious outlier. Figure below shows the relationship between Nox and mortality rates before and after log transformation of Nox.

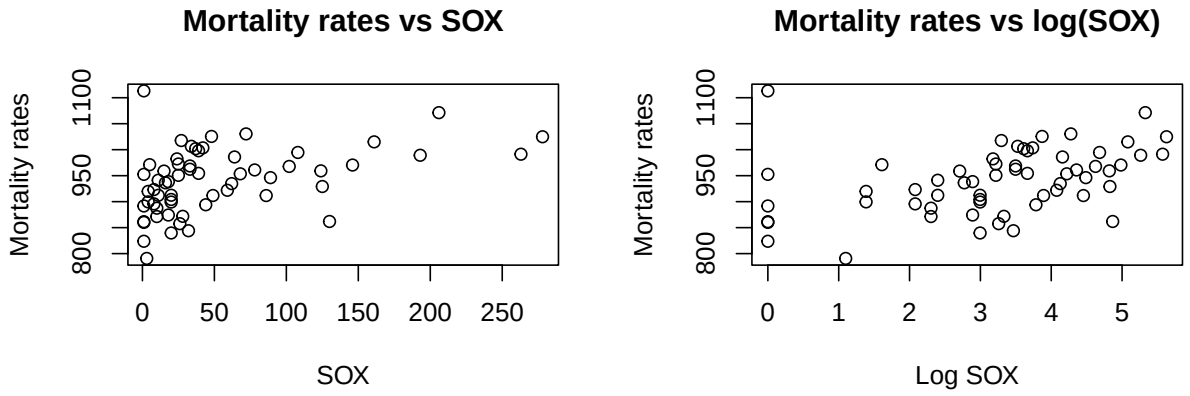
```
layout(t(1:2))
plot(Pol$Nox, Pol$mortality, main = "Mortality rates vs NOX", xlab = "NOX", ylab = "Mortality rates" )
plot(log(Pol$Nox), Pol$mortality, main = "Mortality rates vs log(NOX)", xlab = "Log NOX", ylab = "Morta
```



There appear to be a non-linear relationship between mortality rates and NOX (left panel), log transformation of NOX has corrected the non-linear relationship (right panel).

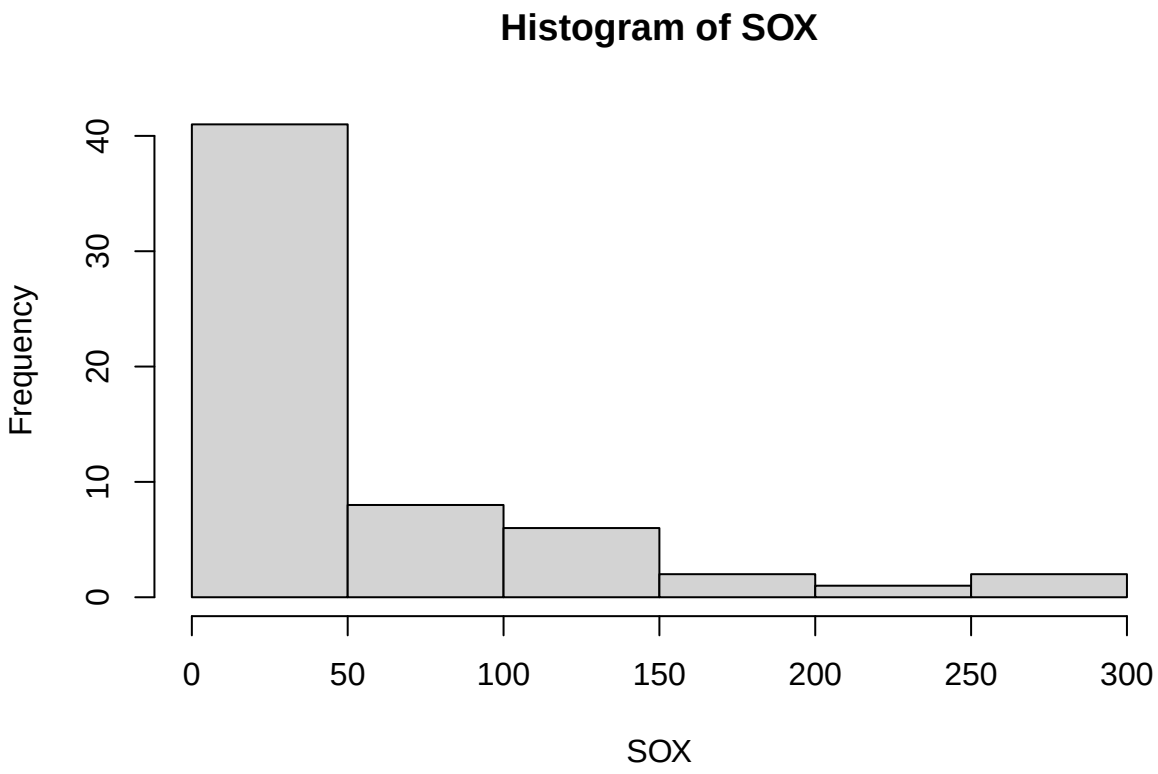
The next figure is a similar figure but for SOX. We observe a similar pattern in SOX as in NOX, though the relationship between SOX and mortality rates appears to be more linear compared to NOX, but the distribution of SOX is skewed (see histogram below). Hence for subsequent analyses, we will use log transformed NOX (Log-NOX) and log-transformed SOX (Log-SOX).

```
layout(t(1:2))
plot(Pol$Sox, Pol$mortality, main = "Mortality rates vs SOX", xlab = "SOX", ylab = "Mortality rates")
plot(log(Pol$Sox), Pol$mortality, main = "Mortality rates vs log(SOX)", xlab = "Log SOX", ylab = "Morta
```



Histogram of SOX which provides evidence to why we need to use log transformed NOX (Log-NOX).

```
hist(Pol$Sox, main = "Histogram of SOX", xlab = "SOX")
```



The following figure shows (left): the relationship between mortality rates and mean annual precipitation (MP) and (right): the distribution of mortality rates by income level

The relationship between MP and mortality rates appears to be linear. Although the two boxplots on the right panel overlapped, cities with low income seems to have higher mortality rates compared to cities with high income level.

```
layout(t(1:2))
plot(Pol$Prec, Pol$mortality, main = "Mortality vs MP", xlab = "MP", ylab = "Mortality")
boxplot(Pol$mortality~Pol$Income, main = "Mortality vs Income", xlab = "Income", ylab = "Mortality")
```

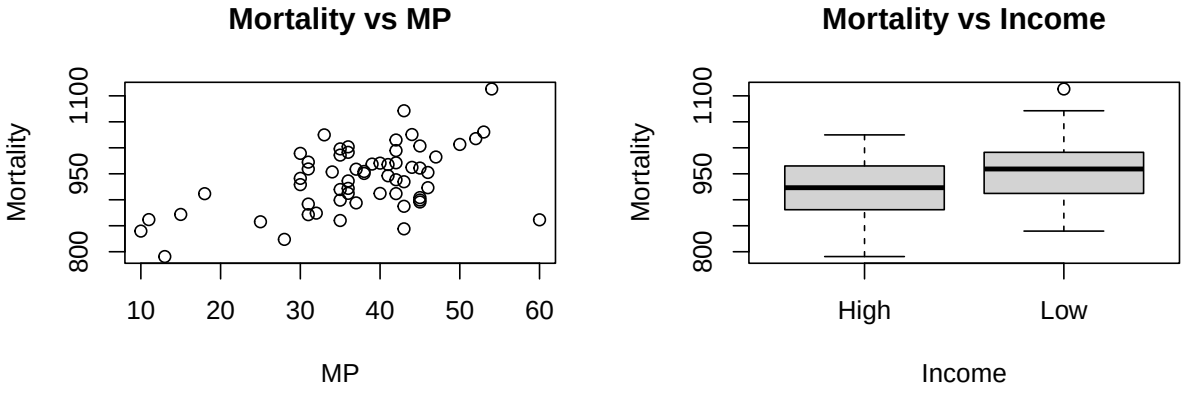
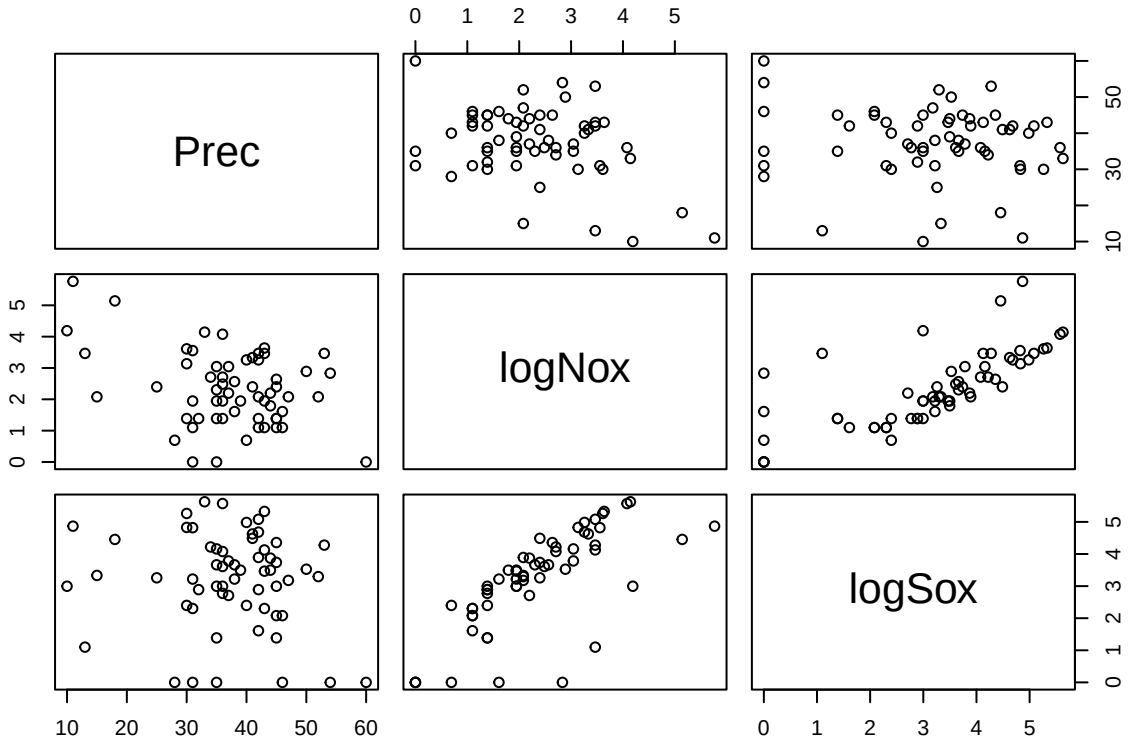


Figure below shows the pairwise scatterplot between the three continuous variables. There appear to be no correlation between MP and the two measures of pollution. However, there is a strong correlation between Log-Nox and Log-Sox as shown in the scatterplot as well as the coefficient of determination (r=0.73).

```
plot(Pol[,c(4,6,7)]) # col 4 is Prec, col 6 is logNox, col 7 is logSox
```




```
cor(Pol$logNox, Pol$logSox)
```

```
## [1] 0.7328074
```

Model development

Exploratory analyses showed that a linear model may be sufficient to model the relationship between mortality rates and the considered input variables since none of the plots shows obvious non-linear relationship that suggests the need of higher order term. However, given the high correlation between SOX and NOX, to avoid multicolinearity (ie. having two independent variables that are strongly correlated with each other in a regression model), one of these two variables shall be dropped from the analyses. We started by fitting a linear model with all the variables included.

```
fit1 = lm(Pol$mortality~Pol$logNox+Pol$logSox+Pol$Prec+Pol$Income)
summary(fit1)
```

```
##
```

```
## Call:
## lm(formula = Pol$mortality ~ Pol$logNox + Pol$logSox + Pol$Prec +
##     Pol$Income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -103.85  -25.50   -4.76   31.43  117.19
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   709.3200    29.2576  24.244  < 2e-16 ***
## Pol$logNox     19.5733     7.6032   2.574   0.0128 *
## Pol$logSox      9.2968     5.6483   1.646   0.1055
## Pol$Prec        3.9133     0.6429   6.086 1.17e-07 ***
## Pol$IncomeLow  19.8754    11.5823   1.716   0.0918 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.51 on 55 degrees of freedom
## Multiple R-squared:  0.5648, Adjusted R-squared:  0.5331
## F-statistic: 17.84 on 4 and 55 DF,  p-value: 1.92e-09
```

The summary table shows that levels of log(SOX) is not significantly associated with mortality rates after controlling for other variables, since its p-value of $0.1055 > 0.05$. We refit the model with log(SOX) removed.

```
fit2 = lm(Pol$mortality~Pol$logNox+Pol$Prec+Pol$Income)
summary(fit2)
```

```
##
## Call:
## lm(formula = Pol$mortality ~ Pol$logNox + Pol$Prec + Pol$Income)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -115.384  -21.046   -1.878   36.548   79.445
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   708.0657    29.6907  23.848  < 2e-16 ***
## Pol$logNox     28.9624     5.1030   5.676 5.09e-07 ***
## Pol$Prec        4.1947     0.6292   6.667 1.23e-08 ***
## Pol$IncomeLow  17.0738    11.6300   1.468    0.148
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.15 on 56 degrees of freedom
## Multiple R-squared:  0.5433, Adjusted R-squared:  0.5189
## F-statistic: 22.21 on 3 and 56 DF,  p-value: 1.332e-09
```

The results show that income is not significantly associated with mortality rates, after controlling for other variables, since its p-value of $0.148 > 0.05$. We fit a model with income variable removed.

```
fit3 = lm(Pol$mortality~Pol$logNox+Pol$Prec)
summary(fit3)
```

```
##
## Call:
## lm(formula = Pol$mortality ~ Pol$logNox + Pol$Prec)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -111.862  -23.874    1.111   28.551   83.968
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 706.4473    29.9693  23.572  < 2e-16 ***
## Pol$logNox   29.1444     5.1529   5.656 5.23e-07 ***
## Pol$Prec      4.4476     0.6113   7.276 1.10e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.58 on 57 degrees of freedom
## Multiple R-squared:  0.5257, Adjusted R-squared:  0.5091
## F-statistic: 31.59 on 2 and 57 DF,  p-value: 5.838e-10
```

Both variables that remain in the model are highly significant (p-value < 0.001). Hence the most parsimonious model is *Mortality = 706.45 + 29.14 x log(NOX) + 4.45 x Prec*

Intepretation:

- On average, the mortality rates is 706.45.
- Every inch increase in mean annual precipitation will increase the expected mortality rate by 4.45, after controlling for nitrogen oxides
- Every 10% increase in nitrogen oxides will increase the expected mortality rate by 1.21, after controlling for mean annual precipitation.

[Note: Scaling log-NOX back to NOX The coefficient corresponding to log nitrogen oxides is 29.14. This means for every one unit increase in log NOX, the expected mortality rates is to increase by 29.14. However, we should interpret the coefficient of nitrogen oxides in the original unit than in log scale. The expected mean difference in mortality rate at x_1 and x_2 is: [recall $y_2 - y_1 = m(x_2 - x_1)$] mortality(x_2)-mortality(x_1) = 29.14 x $(\log x_2 - \log x_1)$ = 29.14 x $\log(\frac{x_2}{x_1})$ This means that as long as the percent increase in NOX is fixed, we will see the same difference in mortality rate, regardless of where the baseline NOX is.

For example, for a k% increase in NOX, Increase in mortality = 29.14 x $\log(\frac{100+k}{100})$

Hence, for a 10% increase in NOX, Increase in mortality = 29.14 x $\log(\frac{100+10}{100}) = 1.21$]

Model diagnostic

We also check if there is any outlier or influential point by producing the following *influence plot*.

[Note: What is an influential point? An observation or data point that has a significant effect on the results of a statistical analysis or modeling, e.g.

- one that has a high leverage, meaning it has a large effect on the estimated regression line
- an outlier that has a large residual, meaning it deviates significantly from the predicted values

What is an influence plot? A graphical tool used to visually identify influential observations or data points in a statistical analysis or modeling. It is a type of diagnostic plot that helps researchers evaluate the impact of individual data points on the overall analysis results.

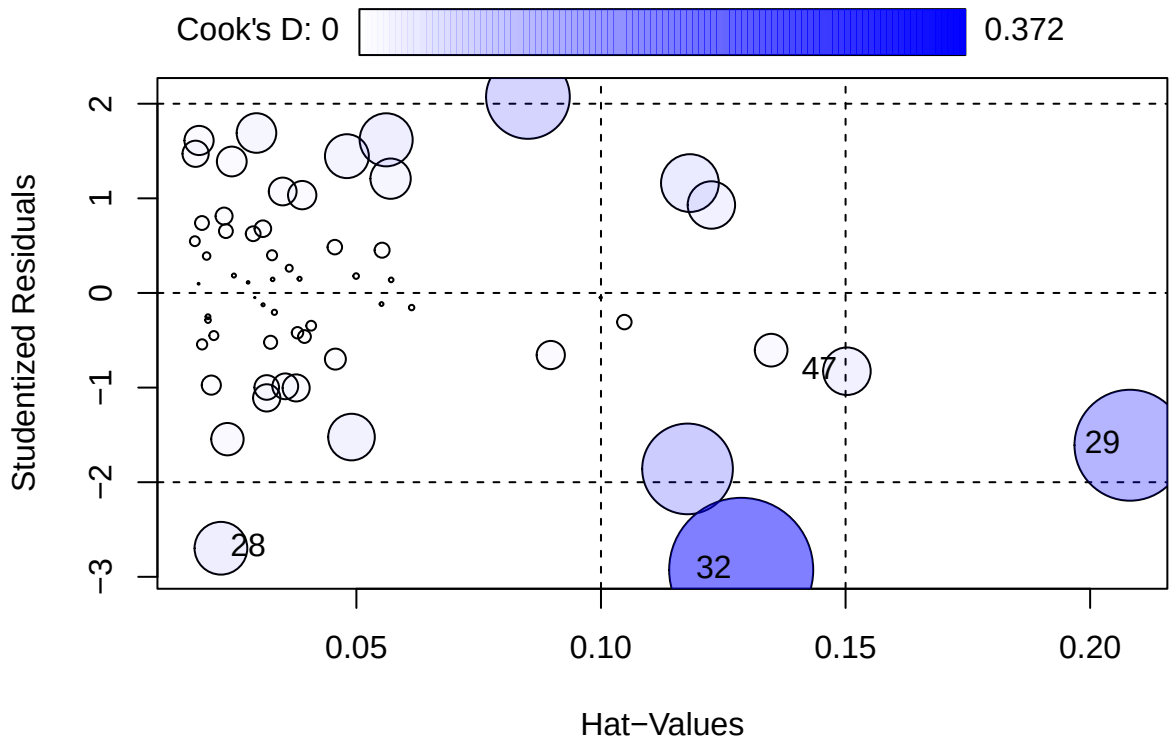
An influence plot typically shows each observation as a point on a scatterplot, with the vertical axis representing the residuals (the difference between the observed value and the predicted value), and the horizontal axis representing some measure of influence or leverage, such as Cook's distance or studentized residuals. The size of each point may also be scaled to represent the weight or importance of the observation.

By examining an influence plot, researchers can identify which data points have the greatest impact on the analysis results, as these will typically be the points with the highest residuals or leverage. These influential points may be outliers or simply observations that have a large effect on the estimated regression coefficients.
]

```
library(car)
```

```
## Loading required package: carData
```

```
influencePlot(fit3)
```




```
##       StudRes         Hat       CookD
```

```
## 28 -2.6980026 0.02230399 0.04986042
## 29 -1.6092382 0.20813822 0.22073736
## 32 -2.9262835 0.12866296 0.37210812
## 47 -0.8270149 0.15027690 0.04054478
```

```
Pol[c(29,32),] # print row 29 and 32 of original dataset to analyse what's unusual with these two data
```

```
##    mortality Nox Sox Prec Income   logNox   logSox
## 29   861.833 319 130   11   High 5.765191 4.867534
## 32   861.439   1   1   60    Low 0.000000 0.000000
```

The plot shows that:

- City 29 is a high leverage point because of its relatively high levels of NOX (ie. 319).
- City 32 is the most influential observation because it has relatively high mean annual precipitation (ie. 60) but relatively low mortality rate.

For sensitivity analysis, we removed these two observations and refit the model.

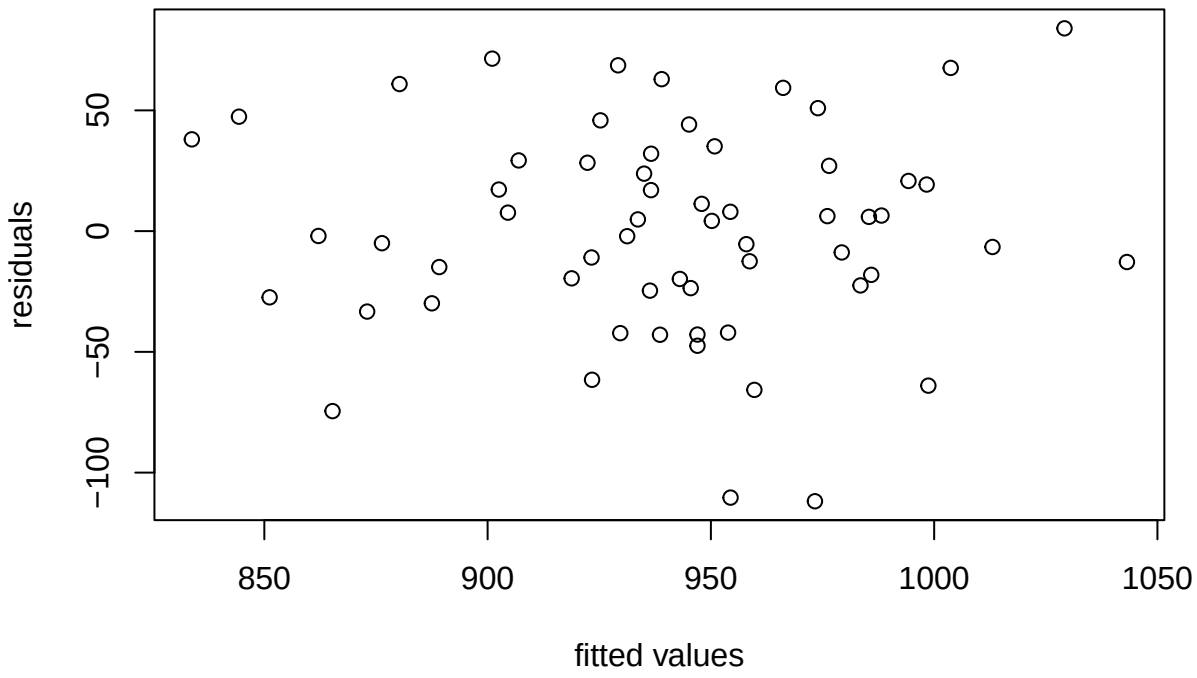
```
fit4=(update(fit3, subset = Pol$Nox!=319&Pol$Prec!=60))
compareCoefs(fit3,fit4)
```

```
## Calls:
## 1: lm(formula = Pol$mortality ~ Pol$logNox + Pol$Prec)
## 2: lm(formula = Pol$mortality ~ Pol$logNox + Pol$Prec, subset = Pol$Nox !=
##   319 & Pol$Prec != 60)
##
##             Model 1 Model 2
## (Intercept)   706.4   703.4
## SE             30.0    28.3
##
## Pol$logNox    29.14   28.78
## SE             5.15    5.14
##
## Pol$Prec      4.448   4.631
## SE            0.611   0.609
##
```

It appears that the estimated coefficients are very similar with or without using these two observations in model fitting.

The *residual plot* of the final model is shown in the plot below. There is no obvious relationship between residuals and fitted values, indicating that the model is a good fit to the data.

```
plot(fit3$fitted.values, fit3$residuals, xlab = "fitted values", ylab = "residuals")
```



Advice for the agency on City 48

```
City48 = Pol[48,]
City48
```

```
##    mortality Nox Sox Prec Income   logNox   logSox
## 48   911.701 171  86   18   High 5.141664 4.454347
```

The current nitrogen oxides level in City 48 is 171, our fitted model predict that if the nitrogen oxides level is reduced to 40, the expected mortality rates is 899.044, approximately 1.9% reduction compared to the current mortality rate simultaneously as sulphur dioxides is not included in our final model. However, from our data analysis, we have shown that pollution levels of sulphur dioxides does not seem to have significant impact on mortality rate. From exploratory analysis, we have shown that nitrogen oxides and sulphur dioxides are highly correlated, ie. city with lower nitrogen oxides tends to have lower sulphur dioxides too. Taken together, we would recommend the Agency to reduce the pollution level of Nitrogen oxides to 40.

Summary

In this report, we investigated the relationship between mortality rates and pollution levels in 60 US cities using linear regression analysis. Our findings show that levels of Nitrogen oxides and Sulphur dioxides are positively correlated with each other, that is, city with higher level of Nitrogen oxides also has higher level of Sulphur dioxides. Although both of these pollution levels are associated with increase in mortality rates

individually, after controlling for each other as well as mean annual precipitation, Sulphur dioxide is no longer significantly associated with mortality rates.

We also observed no significant difference in mortality rate between city with high income and city with low income.

In conclusion, the variables that best describe the mortality rates are Nitrogen oxides and mean annual precipitation. Our results showed that every unit change in the mean annual precipitation will increase mortality rate by 4.45 while every 10% increase in levels of Nitrogen Oxides will increase mortality rate by 2.78.

It is however noteworthy that only approximately 51% (adj $R^2 = 0.5091$) of the variability in mortality rates is explained by mean annual precipitation and levels of Nitrogen oxides, this indicates that there are other important factors of mortality that has not been considered in this study.