Homework 4

Chimdi

5/3/2023

```
library(Sleuth3)
library(Sleuth3)
library(ggplot2)
library(Rmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: plyr
```

```
library(graphics)
```

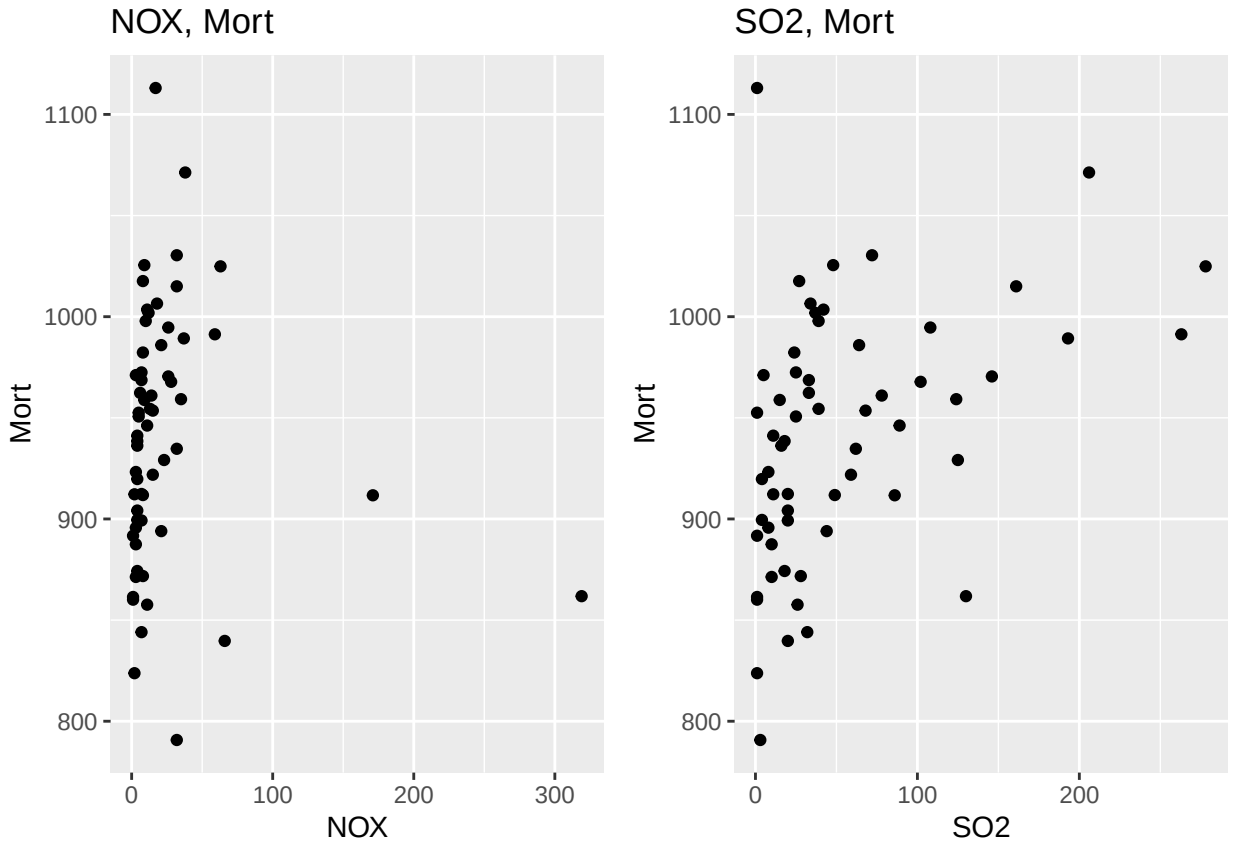
Problem 1: Data taken from Exercise 11.27

Air Pollution and Mortality. Does pollution kill people? Data in one early study designed to explore this issue came from five Standard Metropolitan Statistical Areas (SMSA) in the United States, obtained for the years 1959–1961. (Data from G. C. McDonald and J. A. Ayers, "Some Applications of the 'Chernoff Faces': A Technique for Graphically Representing Multivariate Data", in Graphical Representation of Multivariate Data, New York: Academic Press, 1978.) Total age-adjusted mortality from all causes, in deaths per 100,000 population, is the response variable. The explanatory variables include mean annual precipitation (in inches); median number of school years completed, for persons of age 25 years or older; percentage of 1960 population that is nonwhite; relative pollution potential of oxides of nitrogen, NOX ; and relative pollution potential of sulfur dioxide, SO2. (Note: Two cities–Lancaster and York–are heavily populated by members of the Amish religion, who prefer to teach their children at home. The lower years of education for these two cities do not indicate a social climate similar to other cities with similar years of education.)

```
Data <- ex1123
```

(a) Give two figures containing the scatterplot of Mortality vs. NOX and Mortality vs. SO2. Comment on the scatterplots (does it look like there is a linear relationship? are there outliers?).

```
p1 <- ggplot(Data, aes(NOX, Mort)) + geom_point() + ggtitle('NOX, Mort')
p2 <- ggplot(Data, aes(SO2, Mort)) + geom_point() + ggtitle('SO2, Mort')
multiplot(p1, p2, cols=2)
```



There seems to be violation of the linear relationship because there is no straight line pattern between the observations, especially in the second plot.

There are 2 distinct outliers in the first plot and not so distinct ones in the second plot.

b) Fit the linear regression model of mean mortality against the explanatory variables: Precipitation, Education, Non-white population percent and SO2. What are R2 and adjusted-R2 values? Fit the linear regression model of mean mortality against the explanatory variables: Precipitation, Education, Non-white population percent, NOX and SO2. How do the value of R2 and adjusted-R2 change? What conclusion can you draw?

```
lin_mod1 <- lm(Mort ~ Precip + Educ + NonWhite + SO2, data = Data)
summary(lin_mod1)
```

```
##
## Call:
## lm(formula = Mort ~ Precip + Educ + NonWhite + SO2, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.799 -20.580  -2.503  17.433  92.208
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 999.31666   92.07879  10.853 2.74e-15 ***
## Precip        1.61114    0.63358   2.543 0.013836 *
## Educ        -15.77352    6.99213  -2.256 0.028074 *
## NonWhite      3.06091    0.61401   4.985 6.53e-06 ***
## SO2           0.32719    0.08387   3.901 0.000263 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.24 on 55 degrees of freedom
## Multiple R-squared:  0.6659, Adjusted R-squared:  0.6416
## F-statistic: 27.41 on 4 and 55 DF,  p-value: 1.554e-12
```

R-squared: 0.6659, Adjusted R-squared: 0.6416

```
lin_mod2 <- lm(Mort ~ Precip + Educ + NonWhite + NOX + SO2, data = Data)
summary(lin_mod2)
```

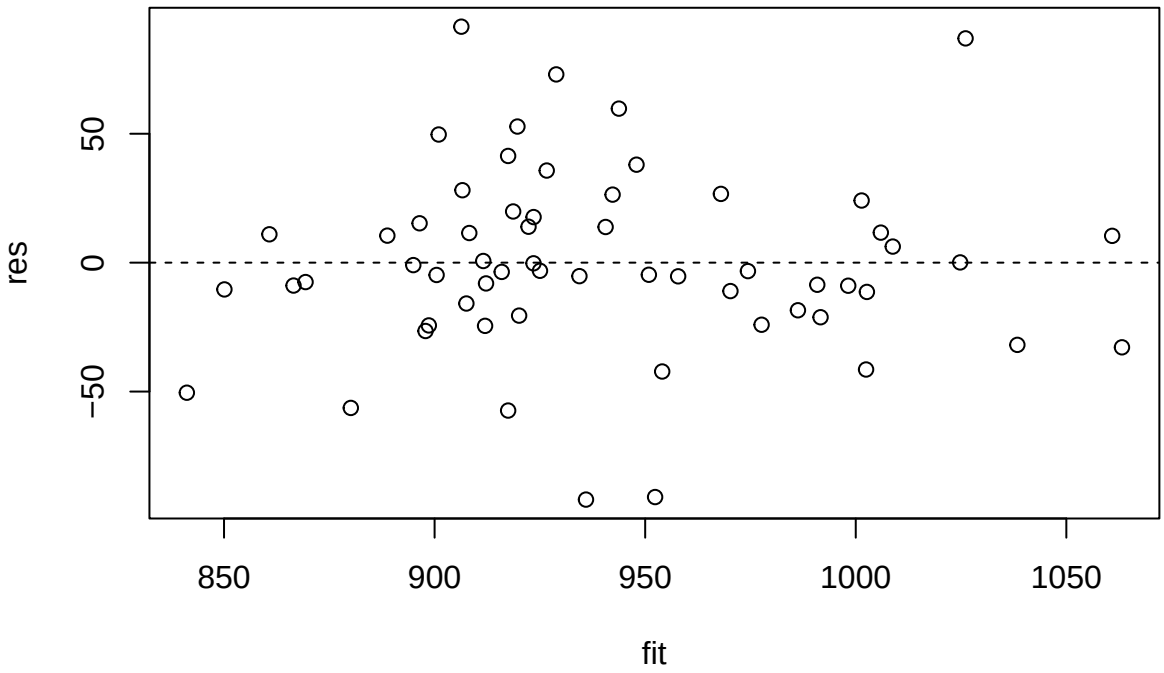
```
##
## Call:
## lm(formula = Mort ~ Precip + Educ + NonWhite + NOX + SO2, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.893 -18.986  -3.433  15.872  91.528
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1000.1026    92.3982  10.824 3.85e-15 ***
## Precip         1.3792     0.7000   1.970 0.053943 .
## Educ         -15.0791     7.0706  -2.133 0.037518 *
## NonWhite       3.1602     0.6287   5.026 5.84e-06 ***
## NOX           -0.1076     0.1359  -0.792 0.432030
## SO2            0.3554     0.0914   3.889 0.000278 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.36 on 54 degrees of freedom
## Multiple R-squared:  0.6698, Adjusted R-squared:  0.6392
## F-statistic:  21.9 on 5 and 54 DF,  p-value: 6.478e-12
```

R-squared: 0.6698, Adjusted R-squared: 0.6392

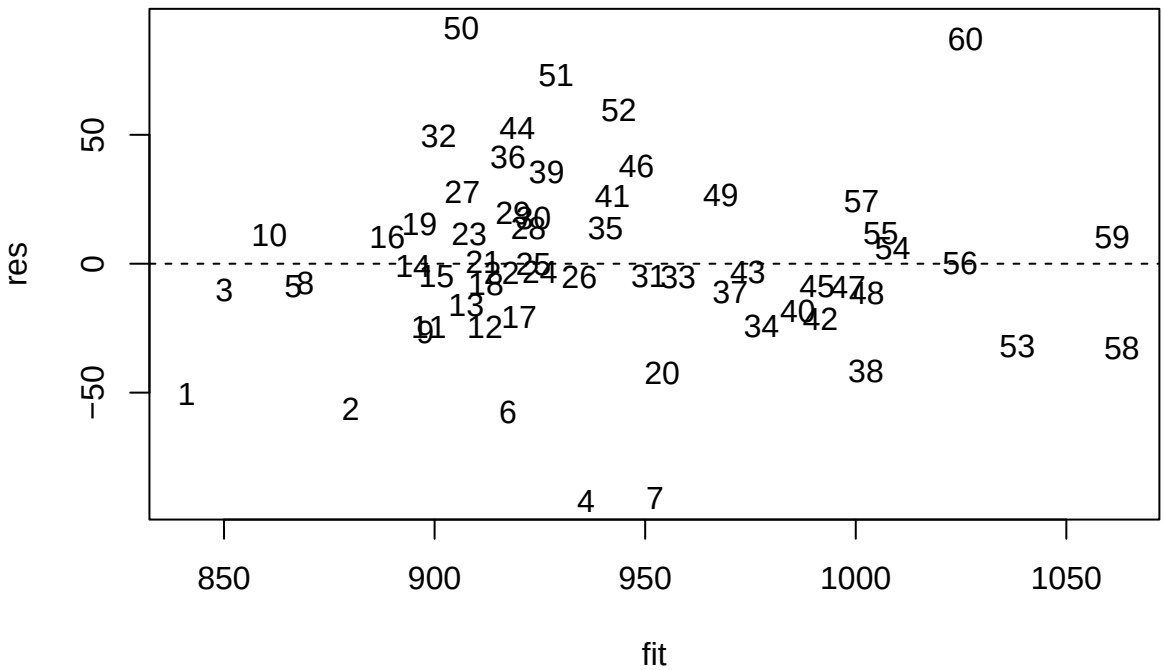
There is a slight increase in R-squared and a slight decrease in Adjusted R-squared. This means that the addition NOX to the regression model has not improved the model's ability to explain the variance in the dependent variable (total age-adjusted mortality from all causes).

(c) Fit the linear regression model of mean mortality against the explanatory variables: Precipitation, Education, Non-white population percent, NOX and SO2. Plot the residual plot for fitted vs. residual values. What conclusions can you draw from the plot? (look for evidence of violations of the model assumptions, identify unusual observations)

```
res <- residuals(lin_mod2)
fit <- fitted(lin_mod2)
plot(fit, res)
abline(0, 0, lty=2)
```




```
#Overlay the row number of observations
plot(fit, res, cex=0)
abline(0, 0, lty=2)
text(fit, res)
```



The conclusion is that the observations have a fairly equal variance and there are no unusual observations as the points are close to each other. The observations are roughly Normaly distributed as they as they are almost evenly lying on the straight horizontal line.

(d) Fit the linear regression model of mean mortality against the explanatory variables: Precipitation, Education, Non-white population percent, log(NOX) and log(SO2) and the linear regression model of mean mortality against the explanatory variables: Precipitation, Education and Non-white population percent. Compare the two models using ANOVA F-test.

```r
lin_mod3 <- lm(Mort ~ Precip + Educ + NonWhite + log(NOX) + log(SO2), data = Data)
summary(lin_mod3)
```

```
##
## Call:
## lm(formula = Mort ~ Precip + Educ + NonWhite + log(NOX) + log(SO2),
##     data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.222  -19.547    0.239   20.084   95.386
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 940.6584    94.0551  10.001 6.81e-14 ***
## Precip        1.9467     0.7007   2.778   0.0075 **
## Educ        -14.6645     6.9379  -2.114   0.0392 *
## NonWhite      3.0289     0.6685   4.531 3.29e-05 ***
## log(NOX)      6.7164     7.3990   0.908   0.3680
## log(SO2)     11.3578     5.2955   2.145   0.0365 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.3 on 54 degrees of freedom
## Multiple R-squared:  0.6883, Adjusted R-squared:  0.6594
## F-statistic: 23.85 on 5 and 54 DF,  p-value: 1.418e-12
```

```r
lin_mod4 <- lm(Mort ~ Precip + Educ + NonWhite, data = Data)
summary(lin_mod4)
```

```
##
## Call:
## lm(formula = Mort ~ Precip + Educ + NonWhite, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -105.107  -27.721    4.334   24.982   90.419
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1141.7357    94.6593  12.062  < 2e-16 ***
## Precip         0.7981     0.6700   1.191  0.23859
## Educ         -24.9906     7.3692  -3.391  0.00128 **
## NonWhite       3.6249     0.6682   5.425 1.28e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.7 on 56 degrees of freedom
## Multiple R-squared:  0.5735, Adjusted R-squared:  0.5506
## F-statistic:  25.1 on 3 and 56 DF,  p-value: 2.017e-10
```

```r
anova(lin_mod3, lin_mod4)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: Mort ~ Precip + Educ + NonWhite + log(NOX) + log(SO2)
## Model 2: Mort ~ Precip + Educ + NonWhite
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     54 71159
## 2     56 97363 -2    -26204 9.9427 0.0002106 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

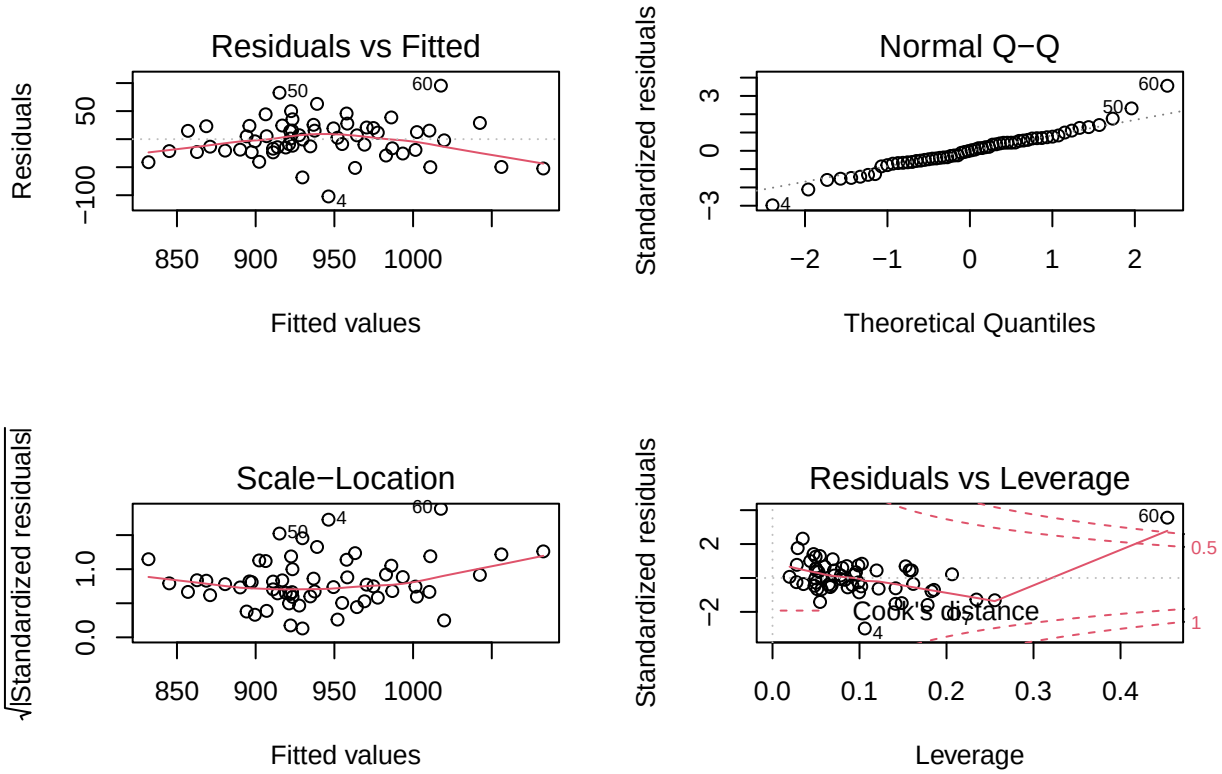
The comparison from the Anova test indicates that the more complex model is better at explaining the variance in the dependent variable (total age-adjusted mortality from all causes). This is because, the p-value(0.0002106) from the Anova test is less than $\alpha = 0.05$, making the more complex model statistically significant.

(e) Fit the linear regression model of mean mortality against the explanatory variables: Precipitation, Education, Non-white population percent, log(NOX) and log(SO2). Plot the Case Influence statistics (leverage, studentized residual and Cook's distance) for each of the observations for the model. Are there any unusual observations? Use the Case Influence Statistics plot to find unusually influential (high Cook's distance) observations. Fit the model again without the influential observations. Does the inference on coefficient parameters (whether the coefficient parameters are significantly different from zero) change?

```
lin_mod5 <- lm(Mort ~ Precip + Educ + NonWhite + log(NOX) + log(SO2), data = Data)
summary(lin_mod5)
```

```
##
## Call:
## lm(formula = Mort ~ Precip + Educ + NonWhite + log(NOX) + log(SO2),
##     data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.222  -19.547    0.239   20.084   95.386
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 940.6584    94.0551  10.001 6.81e-14 ***
## Precip        1.9467     0.7007   2.778   0.0075 **
## Educ        -14.6645     6.9379  -2.114   0.0392 *
## NonWhite      3.0289     0.6685   4.531 3.29e-05 ***
## log(NOX)      6.7164     7.3990   0.908   0.3680
## log(SO2)     11.3578     5.2955   2.145   0.0365 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.3 on 54 degrees of freedom
## Multiple R-squared:  0.6883, Adjusted R-squared:  0.6594
## F-statistic: 23.85 on 5 and 54 DF,  p-value: 1.418e-12
```

```
par(mfrow=c(2,2))
plot(lin_mod5)
```



From the plots, the unusual observation is 60. This is because, it has a Cooks distance > 1(the threshold to determine an unusual observation.

```
Data2 <- Data[-c(60),]
lin_mod6 <- lm(Mort ~ Precip + Educ + NonWhite + log(NOX) + log(SO2), data = Data2)
summary(lin_mod6)
```

```
##
## Call:
## lm(formula = Mort ~ Precip + Educ + NonWhite + log(NOX) + log(SO2),
##     data = Data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -92.889 -20.610  -1.421  20.586  76.905
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 852.3781    85.9332   9.919 1.12e-13 ***
## Precip        1.3633     0.6357   2.144   0.0366 *
## Educ         -5.6672     6.5238  -0.869   0.3889
## NonWhite      3.0397     0.5906   5.147 3.95e-06 ***
## log(NOX)     -9.8984     7.7307  -1.280   0.2060
## log(SO2)     26.0327     5.9311   4.389 5.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.07 on 53 degrees of freedom
## Multiple R-squared:  0.7247, Adjusted R-squared:  0.6987
## F-statistic:  27.9 on 5 and 53 DF,  p-value: 9.929e-14
```

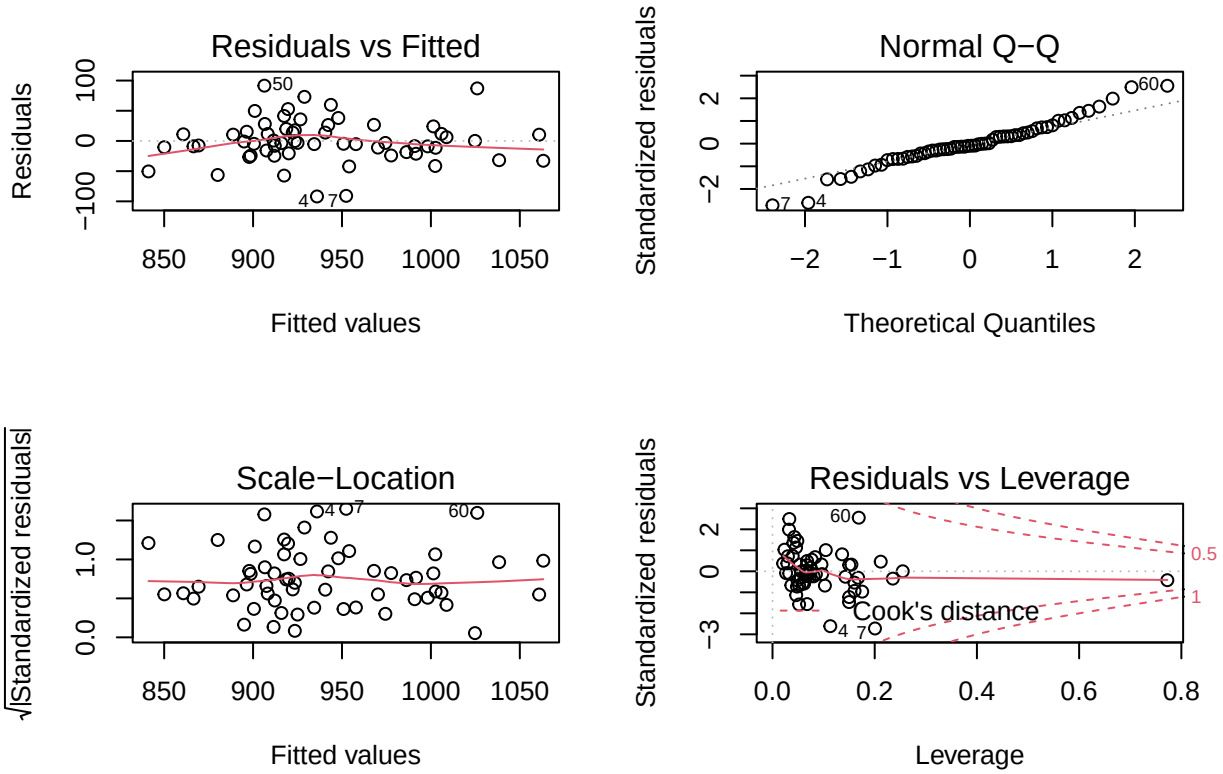
There is a change in the inference on coefficient parameters. β_2 was significant in the previous model but isn't the new model which is as a result of the unusual observation that was removed. This means that removing β_2 and using the simpler model would be better in explaining the variation in the independent variable.

(f) Fit the linear regression model of mean mortality against the explanatory variables: Precipitation, Education, Non-white population percent, NOX and SO2. Plot the Case Influence statistics (leverage, studentized residual and Cook's distance) for each of the observations for the model. Are there any unusual observations? Use the Case Influence Statistics plot to find unusually influential (high Cook's distance) observations. Fit the model again without the influential observations. Does the inference on coefficient parameters (whether the coefficient parameters are significantly different from zero) change?

```
lin_mod7 <- lm(Mort ~ Precip + Educ + NonWhite + NOX + SO2, data = Data)
summary(lin_mod7)
```

```
##
## Call:
## lm(formula = Mort ~ Precip + Educ + NonWhite + NOX + SO2, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.893 -18.986  -3.433  15.872  91.528
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1000.1026    92.3982  10.824 3.85e-15 ***
## Precip         1.3792     0.7000   1.970 0.053943 .
## Educ         -15.0791     7.0706  -2.133 0.037518 *
## NonWhite       3.1602     0.6287   5.026 5.84e-06 ***
## NOX           -0.1076     0.1359  -0.792 0.432030
## SO2            0.3554     0.0914   3.889 0.000278 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.36 on 54 degrees of freedom
## Multiple R-squared:  0.6698, Adjusted R-squared:  0.6392
## F-statistic:  21.9 on 5 and 54 DF,  p-value: 6.478e-12
```

```
par(mfrow=c(2,2))
plot(lin_mod7)
```



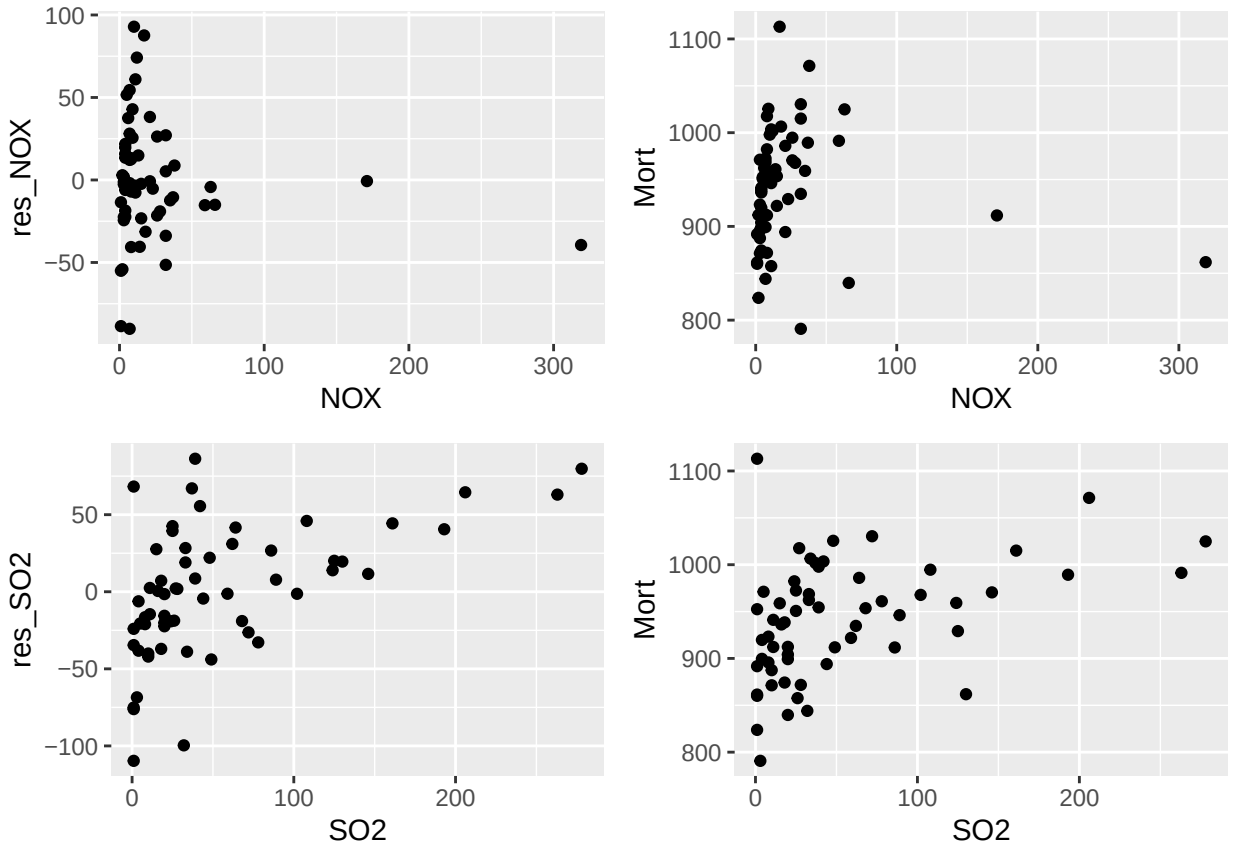
From the plots, there are no unusual observations, because no observation has a Cook's distance > 1.

Therefore, we don't need to exclude any observation.

(g) Construct partial residuals plots for the NOX and SO2 variables after accounting for the variables Precipitation, Education and Non-white population percent. Comment on whether the relationship between the mortality and the pollution variables seems to change after accounting for the weather and socio-economic variables (Precipitation, Education, Non-white population percent).

```
p_res <- residuals(lin_mod2, type = "partial")
colnames(p_res) <- paste("res_", colnames(p_res), sep = "")
Data_res <- cbind(Data, p_res)

plot1 <- ggplot(data = Data_res, aes(NOX, res_NOX)) + geom_point()
plot2 <- ggplot(data = Data_res, aes(NOX, Mort)) + geom_point()
plot3 <- ggplot(data = Data_res, aes(SO2, res_SO2)) + geom_point()
plot4 <- ggplot(data = Data_res, aes(SO2, Mort)) + geom_point()
multiplot(plot1, plot3, plot2, plot4, cols=2)
```



The shapes look about the same on a different scale. Therefore, there is no need for the interaction term between NOX and SO2. This implies that the socio-economic variables (Precipitation, Education, Non-white population percent) in the model has little impact on the relationship between pollution variables (NOX and SO2) on Mortality.

(h) Looking at the data analysis you have done in part (a)-(g), write two summarizing sentences about the relationship between mortality and the pollution variables (NOX and SO2), after the effects of the weather and socioeconomic variables have been accounted for in the study.

From the data analysis, between NOX and SO2, it was only SO2 that has the ability to explain the variance in the dependent variable (total age-adjusted mortality from all causes). And that the socio-economic variables (Precipitation, Education, Non-white population percent) in the model has little impact on the relationship between pollution variables (NOX and SO2) on Mortality