16-17Quiz2

Lídia Montero

23rd December 2016

US Air

Data from a dataset of air pollution in US cities. Seven variables were recorded for 41 cities:

- SO2: Sulphur dioxide content of air in micrograms per cubic meter
- NegTemp: Average annual temperature less than -1 Fahrenheit degrees
- Manuf: Number of manufacturing enterprises employing 20 or more workers
- Pop: Population size (1970 census) in thousands
- Wind: Average annual wind speed in miles per hour
- Precip: Average annual precipitation in inches
- Days: Average number of days with precipitation per year.

Source Everitt, B.S. (2005), An R and S-PLUS Companion to Multivariate Analysis, Springer

Load usair.RData file in your current R or RStudio session

Pop contains the description of thousands of inhabitants for the cities included in the data set. Create a new factor variable consisting on an indicator for small, medium and large cities (named it f.size). Small cities are those with less than half million inhabitants, medium cities are those in the range from half medium to one millium and a half and large cities have a number of inhabitants greater than one million and a half. Our target is defined as SO2.

```
# Set properly the working directory: setwd("xxxx")
load("usair.RData")

# Point 1 - Quiz1
usair$f.size<-factor(cut(usair$Pop,breaks=c(0,500,1500,3500)),labels=c("Small","Medium","Large"))
summary(usair)
```

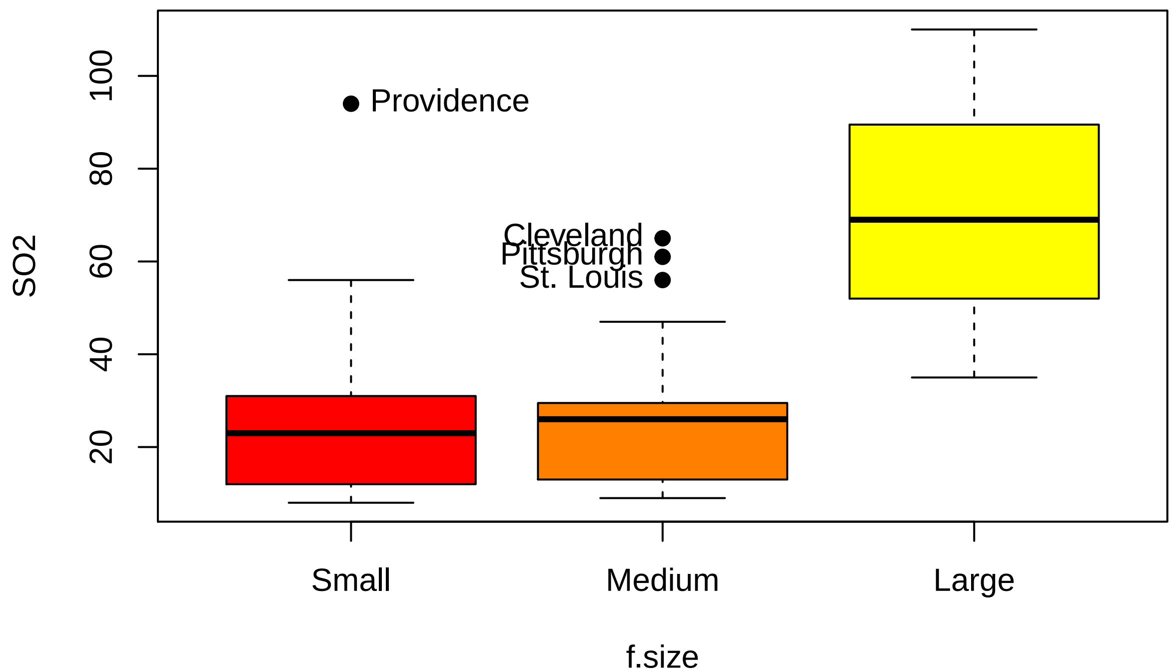
```
##       SO2           Neg.Temp          Manuf             Pop
##  Min.   :  8.00   Min.   :-75.50   Min.   :  35.0   Min.   :  71.0
##  1st Qu.: 13.00   1st Qu.:-59.30   1st Qu.: 181.0   1st Qu.: 299.0
##  Median : 26.00   Median :-54.60   Median : 347.0   Median : 515.0
##  Mean   : 30.05   Mean   :-55.76   Mean   : 463.1   Mean   : 608.6
##  3rd Qu.: 35.00   3rd Qu.:-50.60   3rd Qu.: 462.0   3rd Qu.: 717.0
##  Max.   :110.00   Max.   :-43.50   Max.   :3344.0   Max.   :3369.0
##       Wind            Precip           Days          f.size
##  Min.   : 6.000   Min.   : 7.05   Min.   : 36.0   Small :19
##  1st Qu.: 8.700   1st Qu.:30.96   1st Qu.:103.0   Medium:19
##  Median : 9.300   Median :38.74   Median :115.0   Large : 3
##  Mean   : 9.444   Mean   :36.77   Mean   :113.9
##  3rd Qu.:10.600   3rd Qu.:43.11   3rd Qu.:128.0
##  Max.   :12.700   Max.   :59.80   Max.   :166.0
```

1. The average SO2 in the cities can be argued to be the same for all city size levels (f.size)? Check the hypothesis by estimating one-way model/s with method lm() and using a suitable inferential tool.

By visual inspection using a boxplot tool for SO2 - Sulphur dioxide (microg/m3), the average contents and 50% central range of SO2 in air for large cities is clearly greater than the average contents and 50% central ranges for cities in small and medium size groups.

A null model with the constant and a one-way model is calculated using a general linear model method. A null hypothesis stating *H0: m0 = m01 or mu(Small)=mu(Medium)=mu(Large)=mu* is tested using Fisher Test and a pvalue of 0.004 is returned by R. The null H0 is not likely and can be rejected, thus there exists at least a city size group with mean of S02 air contents different from the rest of groups.

```
m0<-lm(SO2~1,data=usair)
library(car)
Boxplot(SO2~f.size,data=usair,col=heat.colors(3),pch=19)
```




```
## [1] "Providence" "St. Louis"  "Cleveland"  "Pittsburgh"
```

```
m01<-lm(SO2~f.size,data=usair)
anova(m0,m01)
```

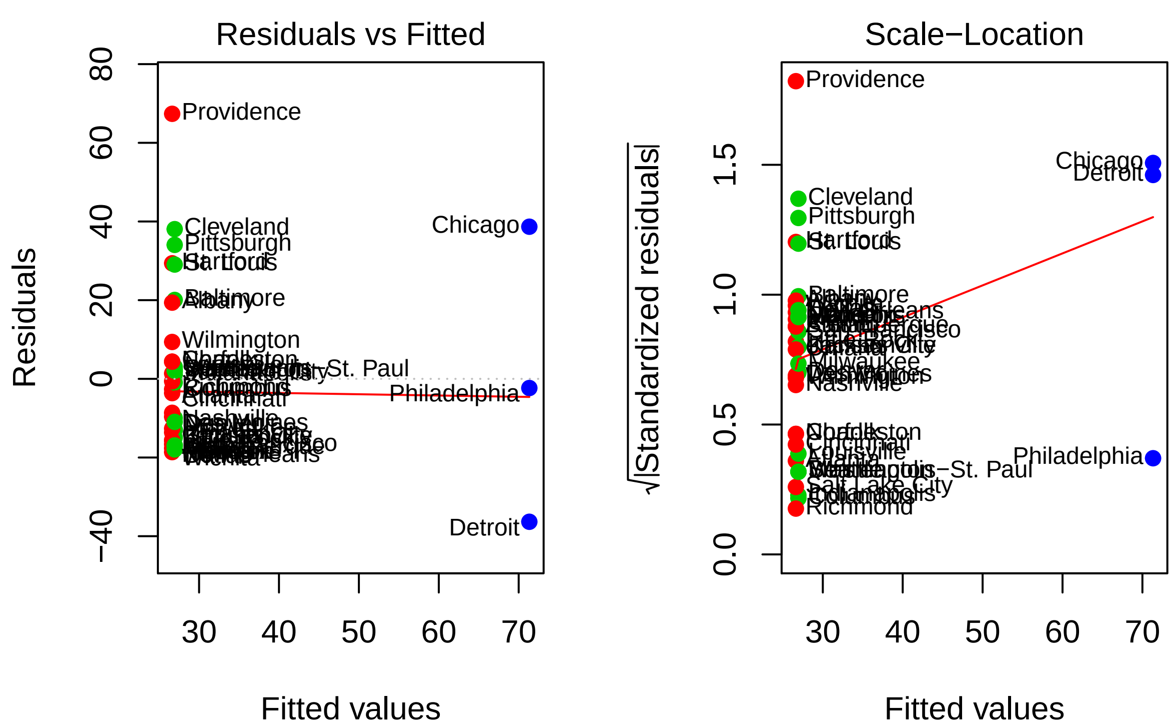
```
## Analysis of Variance Table
##
## Model 1: SO2 ~ 1
## Model 2: SO2 ~ f.size
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1     40 22038
## 2     38 16520  2    5517.9 6.3462 0.004188 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. The variance of SO2 in the cities can be argued to be the same for all city size levels (f.size)? Check and discuss residuals after estimating a suitable one-way model with method lm().

According to diagnostics in Scale-Location plot, also called Spread-Location or 'S-L' plot that takes the square root of the absolute residuals in Y axis across fitted values in X axis, the variance of in large city group is greater than the one in Small and Medium groups. Nevertheless, an outstanding outlier is observed for Providence (Small group).

```r
par(cex=0.5)
par(mfrow=c(1,2))
plot(m01,which=c(1,3),id.n=41,pch=19,col=(as.numeric(usair$f.size)+1))
```




```r
par(mfrow=c(1,1))
```

3. Consider a multiple regression model (m1) for target SO2 on all numeric variables in the dataset. Assess the quality of the model.

The model explains 67% of the variability of the target SO2 (sulphur dioxide) contents in air. All variables consume 1 degree of freedom, but only Negative Temperature, Manuf and Population have net effects statistically significant according to Fisher tests implemented by method Anova() in library car; Wind, Precip and Days are not significant.

```r
names(usair)
```

```
## [1] "SO2"      "Neg.Temp" "Manuf"    "Pop"      "Wind"     "Precip"
## [7] "Days"     "f.size"
```

```r
m1<-lm(SO2~.,data=usair[,1:7])
summary(m1)
```

```
##
## Call:
## lm(formula = SO2 ~ ., data = usair[, 1:7])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.004  -8.542  -0.991   5.758  48.758
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 111.72848   47.31810   2.361 0.024087 *
## Neg.Temp      1.26794    0.62118   2.041 0.049056 *
## Manuf         0.06492    0.01575   4.122 0.000228 ***
## Pop          -0.03928    0.01513  -2.595 0.013846 *
## Wind         -3.18137    1.81502  -1.753 0.088650 .
## Precip        0.51236    0.36276   1.412 0.166918
## Days         -0.05205    0.16201  -0.321 0.749972
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.64 on 34 degrees of freedom
## Multiple R-squared:  0.6695, Adjusted R-squared:  0.6112
## F-statistic: 11.48 on 6 and 34 DF,  p-value: 5.419e-07
```

```r
Anova(m1)
```

```
## Anova Table (Type II tests)
##
## Response: SO2
##           Sum Sq Df F value    Pr(>F)
## Neg.Temp   892.5  1  4.1664 0.0490557 *
## Manuf     3640.1  1 16.9929 0.0002278 ***
## Pop       1443.1  1  6.7365 0.0138462 *
## Wind       658.1  1  3.0723 0.0886504 .
## Precip     427.3  1  1.9949 0.1669176
## Days        22.1  1  0.1032 0.7499725
## Residuals 7283.3 34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Consider model (m1), check significance for all variables. Propose a new reduced model (m2), if non-significance variables are found in (m1). Discuss your proposal.

According to the AIC criteria searching for a reduced model with a lower Akaike statistic, Days should be removed from (m1) model reducing almost 2 units in AIC from the initial model. Explicability is reduced by 3 points

An statitiscal Fisher Test is applied between the initial (m1) and the reduced (m2) model, they are nested models. The null hypothesis of equivalence from the inferential point of view can not be rejected with a p value of 0.75 >> 0.05. Thus, (m1) and (m2) do the same work, so we also prefer a reduced and simple model (m2). Nevertheless, Minimum AIC and inferential criteria do not match since Precipitation is clearly non-significant (net effect) and Wind is in the borderline, but not significant also.

If BIC criteria is used to monitor the step procedure a model (m2b) is obtained with Population and Manuf. Explicability is reduced by more than 10 points. But, we can not reject Fisher test between (m1) and (m2b), they are equivalence according to inferential rules.

```r
m2<-step(m1) # AIC can be used, there are few observations
```

```
## Start:  AIC=226.37
## SO2 ~ Neg.Temp + Manuf + Pop + Wind + Precip + Days
##
##            Df Sum of Sq     RSS    AIC
## - Days      1      22.1  7305.4 224.50
## <none>                   7283.3 226.37
## - Precip    1     427.3  7710.6 226.71
## - Wind      1     658.1  7941.4 227.92
## - Neg.Temp  1     892.5  8175.8 229.11
## - Pop       1    1443.1  8726.3 231.78
## - Manuf     1    3640.1 10923.4 240.99
##
## Step:  AIC=224.49
## SO2 ~ Neg.Temp + Manuf + Pop + Wind + Precip
##
##            Df Sum of Sq     RSS    AIC
## <none>                   7305.4 224.50
## - Wind      1     636.1  7941.5 225.92
## - Precip    1     785.4  8090.8 226.68
## - Pop       1    1447.5  8752.9 229.91
## - Neg.Temp  1    1517.4  8822.8 230.23
## - Manuf     1    3636.8 10942.1 239.06
```

```r
summary(m2)
```

```
##
## Call:
## lm(formula = SO2 ~ Neg.Temp + Manuf + Pop + Wind + Precip, data = usair[,
##     1:7])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.253  -7.655  -0.581   6.059  49.438
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 100.15245   30.27521   3.308 0.002182 **
## Neg.Temp      1.12129    0.41586   2.696 0.010707 *
```

```
## Manuf         0.06489    0.01554   4.174 0.000188 ***
## Pop          -0.03933    0.01494  -2.633 0.012499 *
## Wind         -3.08240    1.76562  -1.746 0.089622 .
## Precip        0.41947    0.21624   1.940 0.060498 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.45 on 35 degrees of freedom
## Multiple R-squared:  0.6685, Adjusted R-squared:  0.6212
## F-statistic: 14.12 on 5 and 35 DF,  p-value: 1.409e-07
```

```r
anova(m2,m1)
```

```
## Analysis of Variance Table
##
## Model 1: SO2 ~ Neg.Temp + Manuf + Pop + Wind + Precip
## Model 2: SO2 ~ Neg.Temp + Manuf + Pop + Wind + Precip + Days
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     35 7305.4
## 2     34 7283.3  1     22.11 0.1032   0.75
```

```r
m2b<-step(m1,k=log(nrow(usair)))
```

```
## Start:  AIC=238.37
## SO2 ~ Neg.Temp + Manuf + Pop + Wind + Precip + Days
##
##            Df Sum of Sq     RSS    AIC
## - Days      1      22.1  7305.4 234.78
## - Precip    1     427.3  7710.6 236.99
## - Wind      1     658.1  7941.4 238.20
## <none>                   7283.3 238.37
## - Neg.Temp  1     892.5  8175.8 239.39
## - Pop       1    1443.1  8726.3 242.06
## - Manuf     1    3640.1 10923.4 251.27
##
## Step:  AIC=234.78
## SO2 ~ Neg.Temp + Manuf + Pop + Wind + Precip
##
##            Df Sum of Sq     RSS    AIC
## - Wind      1     636.1  7941.5 234.49
## <none>                   7305.4 234.78
## - Precip    1     785.4  8090.8 235.25
## - Pop       1    1447.5  8752.9 238.47
## - Neg.Temp  1    1517.4  8822.8 238.80
## - Manuf     1    3636.8 10942.1 247.63
##
## Step:  AIC=234.49
## SO2 ~ Neg.Temp + Manuf + Pop + Precip
##
##            Df Sum of Sq     RSS    AIC
## - Precip    1     597.1  8538.7 233.75
## <none>                   7941.5 234.49
## - Neg.Temp  1    1026.9  8968.4 235.76
```

```
## - Pop       1    1706.2  9647.7 238.75
## - Manuf     1    3851.8 11793.3 246.99
##
## Step:  AIC=233.74
## SO2 ~ Neg.Temp + Manuf + Pop
##
##            Df Sum of Sq     RSS    AIC
## - Neg.Temp  1     578.0  9116.6 232.72
## <none>                   8538.7 233.75
## - Pop       1    2125.2 10663.8 239.14
## - Manuf     1    4539.0 13077.6 247.51
##
## Step:  AIC=232.72
## SO2 ~ Manuf + Pop
##
##         Df Sum of Sq     RSS    AIC
## <none>                9116.6 232.72
## - Pop    1    3759.5 12876.2 243.16
## - Manuf  1    7548.0 16664.7 253.73
```

```r
summary(m2b)
```

```
##
## Call:
## lm(formula = SO2 ~ Manuf + Pop, data = usair[, 1:7])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.389 -12.831  -1.277   7.609  49.533
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.32508    3.84044   6.855 3.87e-08 ***
## Manuf        0.08243    0.01470   5.609 1.96e-06 ***
## Pop         -0.05661    0.01430  -3.959 0.000319 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.49 on 38 degrees of freedom
## Multiple R-squared:  0.5863, Adjusted R-squared:  0.5645
## F-statistic: 26.93 on 2 and 38 DF,  p-value: 5.207e-08
```

```r
anova(m2b,m1)
```

```
## Analysis of Variance Table
##
## Model 1: SO2 ~ Manuf + Pop
## Model 2: SO2 ~ Neg.Temp + Manuf + Pop + Wind + Precip + Days
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     38 9116.6
## 2     34 7283.3  4    1833.4 2.1396 0.0972 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. Write the equation for the resulting model (m2).

Both models show colinearity between Manuf and Pop variables, we have to select manually one of them, the one most related with SO2 target and remove the other from the model, in this case Population is removed. It is convenient to repeat procedure in Point 4. Now, both monitoring criteria for step() lead to the same resulting model (m2), the one including Negative Temperature, Manuf, Wind and Precipitation. A 60% explicability of the target variance is obtained. And the equation of model is:

SO2= 123.11 + 1.61 Neg.Temp + 0.02 Manuf - 3.63 Wind + 0.52 Precip

```r
vif(m2)
```

```
##  Neg.Temp     Manuf       Pop      Wind    Precip
##  1.731366 14.703099 14.338797  1.219354  1.241777
```

```r
vif(m2b)
```

```
##    Manuf      Pop
## 11.43374 11.43374
```

```r
round(cor(usair[,1:7]),dig=2)
```

```
##           SO2 Neg.Temp Manuf   Pop  Wind Precip Days
## SO2      1.00     0.43  0.64  0.49  0.09   0.05 0.37
## Neg.Temp 0.43     1.00  0.19  0.06  0.35  -0.39 0.43
## Manuf    0.64     0.19  1.00  0.96  0.24  -0.03 0.13
## Pop      0.49     0.06  0.96  1.00  0.21  -0.03 0.04
## Wind     0.09     0.35  0.24  0.21  1.00  -0.01 0.16
## Precip   0.05    -0.39 -0.03 -0.03 -0.01   1.00 0.50
## Days     0.37     0.43  0.13  0.04  0.16   0.50 1.00
```

```r
m1<-lm(SO2~.,data=usair[,c(1:3,5:7)])
m2<-step(m1)
```

```
## Start:  AIC=231.78
## SO2 ~ Neg.Temp + Manuf + Wind + Precip + Days
##
##            Df Sum of Sq     RSS    AIC
## - Days      1      26.6  8752.9 229.91
## <none>                   8726.3 231.78
## - Precip    1     647.1  9373.4 232.72
## - Wind      1     921.4  9647.7 233.90
## - Neg.Temp  1    1930.3 10656.6 237.97
## - Manuf     1    7692.0 16418.4 255.70
##
## Step:  AIC=229.91
## SO2 ~ Neg.Temp + Manuf + Wind + Precip
##
##            Df Sum of Sq     RSS    AIC
## <none>                   8752.9 229.91
## - Wind      1     894.8  9647.7 231.90
## - Precip    1    1269.7 10022.6 233.46
## - Neg.Temp  1    3919.0 12671.9 243.08
## - Manuf     1    7665.8 16418.7 253.70
```

```r
m2b<-step(m1,k=log(nrow(usair)))
```

```
## Start:  AIC=242.06
## SO2 ~ Neg.Temp + Manuf + Wind + Precip + Days
##
##            Df Sum of Sq     RSS    AIC
## - Days      1      26.6  8752.9 238.47
## - Precip    1     647.1  9373.4 241.28
## <none>                   8726.3 242.06
## - Wind      1     921.4  9647.7 242.47
## - Neg.Temp  1    1930.3 10656.6 246.54
## - Manuf     1    7692.0 16418.4 264.26
##
## Step:  AIC=238.47
## SO2 ~ Neg.Temp + Manuf + Wind + Precip
##
##            Df Sum of Sq     RSS    AIC
## <none>                   8752.9 238.47
## - Wind      1     894.8  9647.7 238.75
## - Precip    1    1269.7 10022.6 240.31
## - Neg.Temp  1    3919.0 12671.9 249.93
## - Manuf     1    7665.8 16418.7 260.55
```

```r
anova(m2,m1)
```

```
## Analysis of Variance Table
##
## Model 1: SO2 ~ Neg.Temp + Manuf + Wind + Precip
## Model 2: SO2 ~ Neg.Temp + Manuf + Wind + Precip + Days
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     36 8752.9
## 2     35 8726.3  1    26.575 0.1066  0.746
```

```r
summary(m2)
```

```
##
## Call:
## lm(formula = SO2 ~ Neg.Temp + Manuf + Wind + Precip, data = usair[,
##     c(1:3, 5:7)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.374  -9.088  -3.042   7.205  58.785
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 123.118333  31.290702   3.935 0.000365 ***
## Neg.Temp      1.611436   0.401373   4.015 0.000289 ***
## Manuf         0.025476   0.004537   5.615 2.27e-06 ***
## Wind         -3.630245   1.892342  -1.918 0.063020 .
## Precip        0.524235   0.229407   2.285 0.028297 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.59 on 36 degrees of freedom
## Multiple R-squared:  0.6028, Adjusted R-squared:  0.5587
## F-statistic: 13.66 on 4 and 36 DF,  p-value: 7.168e-07
```

6. Add f.size factor to obtain a new model including interactions with 2 numeric variables in (m2) and write the equation of the resulting model (m3).

f.size factor has a non-significant net effect according to Fisher test implemented in Anova() for the additive model. Explicability is almost 62% for (m3a) additive. And the equations are parallel and differ in the intercept

For Small group cities: SO2= 124.25 + 1.57 Neg.Temp + 0.031 Manuf - 3.70 Wind + 0.47 Precip

For Medium group cities: SO2= 124.25-5.69 + 1.57 Neg.Temp + 0.031 Manuf - 3.70 Wind + 0.47 Precip

For Large group cities: SO2= 124.25-14.94 + 1.57 Neg.Temp + 0.031 Manuf - 3.70 Wind + 0.47 Precip

Using f.size factor interacting as indicated in the question, one can see that it interactions have non-significant net effects according to Fisher test implemented in Anova() for the model (I selected interactions with the most rellevant variables according to m2). Explicability is 64% for (m3) . And the equations are not parallel and differ in the intercept and in the slope:

For Small group cities: SO2= 160.47 + 2.13 Neg.Temp + 0.01 Manuf - 4.23 Wind + 0.59 Precip

For Medium group cities: SO2= (160.47 - 63.74) + (2.13-0.88) Neg.Temp + (0.01+0.032) Manuf - 4.23 Wind + 0.59 Precip

For Large group cities: SO2= (160.47 - 203.32) + (2.13+3.50) Neg.Temp + (0.01+0.021) Manuf - 4.23 Wind + 0.59 Precip

```r
Anova(m2)
```

```
## Anova Table (Type II tests)
##
## Response: SO2
##           Sum Sq Df F value    Pr(>F)
## Neg.Temp  3919.0  1 16.1187 0.0002887 ***
## Manuf     7665.8  1 31.5287 2.273e-06 ***
## Wind       894.8  1  3.6802 0.0630196 .
## Precip    1269.7  1  5.2220 0.0282974 *
## Residuals 8752.9 36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
m3a <-lm(SO2 ~ (Neg.Temp + Manuf)+f.size + Wind + Precip,data=usair[,c(1:3,5:8)])
Anova(m3a)
```

```
## Anova Table (Type II tests)
##
## Response: SO2
##           Sum Sq Df F value    Pr(>F)
## Neg.Temp  3640.1  1 14.6296 0.0005333 ***
## Manuf     3645.5  1 14.6510 0.0005291 ***
## f.size     293.1  2  0.5889 0.5605049
## Wind       915.3  1  3.6785 0.0635495 .
```

```
## Precip     998.8  1  4.0141 0.0531433 .
## Residuals 8459.8 34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(m3a)
```

```
##
## Call:
## lm(formula = SO2 ~ (Neg.Temp + Manuf) + f.size + Wind + Precip,
##     data = usair[, c(1:3, 5:8)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.271  -9.121  -2.497   8.134  56.378
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  124.247641  31.717616   3.917 0.000410 ***
## Neg.Temp       1.568034   0.409959   3.825 0.000533 ***
## Manuf          0.031263   0.008168   3.828 0.000529 ***
## f.sizeMedium  -5.691381   5.702032  -0.998 0.325266
## f.sizeLarge  -14.935303  17.712344  -0.843 0.405002
## Wind          -3.704002   1.931241  -1.918 0.063549 .
## Precip         0.475210   0.237188   2.004 0.053143 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.77 on 34 degrees of freedom
## Multiple R-squared:  0.6161, Adjusted R-squared:  0.5484
## F-statistic: 9.095 on 6 and 34 DF,  p-value: 5.937e-06
```

```r
# Requested model (m3)
m3<-lm(SO2 ~ (Neg.Temp + Manuf)*f.size + Wind + Precip,data=usair[,c(1:3,5:8)])
summary(m3)
```

```
##
## Call:
## lm(formula = SO2 ~ (Neg.Temp + Manuf) * f.size + Wind + Precip,
##     data = usair[, c(1:3, 5:8)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.945  -9.942   0.000   4.162  56.205
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             160.47358   42.80713   3.749 0.000758 ***
## Neg.Temp                  2.13108    0.62804   3.393 0.001958 **
## Manuf                     0.01039    0.02946   0.353 0.726762
## f.sizeMedium            -63.74435   47.22187  -1.350 0.187149
## f.sizeLarge            -203.32383  240.73328  -0.845 0.405018
## Wind                     -4.23253    2.10893  -2.007 0.053839 .
```

```
## Precip                     0.58892    0.26418   2.229 0.033433 *
## Neg.Temp:f.sizeMedium   -0.87783    0.77156  -1.138 0.264241
## Neg.Temp:f.sizeLarge    -3.49733    4.58597  -0.763 0.451647
## Manuf:f.sizeMedium       0.03174    0.03546   0.895 0.377958
## Manuf:f.sizeLarge        0.02174    0.03098   0.702 0.488235
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.2 on 30 degrees of freedom
## Multiple R-squared:  0.6425, Adjusted R-squared:  0.5234
## F-statistic: 5.392 on 10 and 30 DF,  p-value: 0.0001516
```

```r
Anova(m3)
```

```
## Anova Table (Type II tests)
##
## Response: SO2
##                 Sum Sq Df F value    Pr(>F)
## Neg.Temp        3682.4  1 14.0230 0.0007662 ***
## Manuf           3827.2  1 14.5744 0.0006287 ***
## f.size           293.1  2  0.5580 0.5781879
## Wind            1057.7  1  4.0279 0.0538390 .
## Precip          1305.0  1  4.9696 0.0334329 *
## Neg.Temp:f.size  465.5  2  0.8864 0.4226477
## Manuf:f.size     210.4  2  0.4006 0.6734724
## Residuals       7878.0 30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7. Use step() procedure with BIC criteria to simplify (m3) model. Call the new model (m4). Is the new model equivalent to (m3) from an inferential point of view?

The stepwise procedure based on BIC minimization removes the main effect and interactions with factor f.size. The Fisher Test between the nested models (m3) big and (m4) reduced are equivalent, according to the pvalue 0.76 >> 0.05.

```r
m4<-step(m3,k=log(nrow(usair)))
```

```
## Start:  AIC=256.44
## SO2 ~ (Neg.Temp + Manuf) * f.size + Wind + Precip
##
##                   Df Sum of Sq    RSS    AIC
## - Manuf:f.size     2    210.38 8088.4 250.09
## - Neg.Temp:f.size  2    465.55 8343.5 251.37
## <none>                         7878.0 256.44
## - Wind             1   1057.72 8935.7 257.89
## - Precip           1   1305.03 9183.0 259.01
##
## Step:  AIC=250.09
## SO2 ~ Neg.Temp + Manuf + f.size + Wind + Precip + Neg.Temp:f.size
##
##                   Df Sum of Sq     RSS    AIC
## - Neg.Temp:f.size  2     371.5  8459.8 244.50
```

```
## <none>                          8088.4 250.09
## - Wind             1     950.0  9038.4 250.93
## - Precip           1    1127.9  9216.2 251.73
## - Manuf            1    3827.2 11915.6 262.26
##
## Step:  AIC=244.51
## SO2 ~ Neg.Temp + Manuf + f.size + Wind + Precip
##
##            Df Sum of Sq     RSS    AIC
## - f.size    2     293.1  8752.9 238.47
## <none>                   8459.8 244.50
## - Wind      1     915.3  9375.1 245.00
## - Precip    1     998.8  9458.6 245.37
## - Neg.Temp  1    3640.1 12100.0 255.46
## - Manuf     1    3645.5 12105.3 255.48
##
## Step:  AIC=238.47
## SO2 ~ Neg.Temp + Manuf + Wind + Precip
##
##            Df Sum of Sq     RSS    AIC
## <none>                   8752.9 238.47
## - Wind      1     894.8  9647.7 238.75
## - Precip    1    1269.7 10022.6 240.31
## - Neg.Temp  1    3919.0 12671.9 249.93
## - Manuf     1    7665.8 16418.7 260.55
```

```r
anova(m4,m3)
```

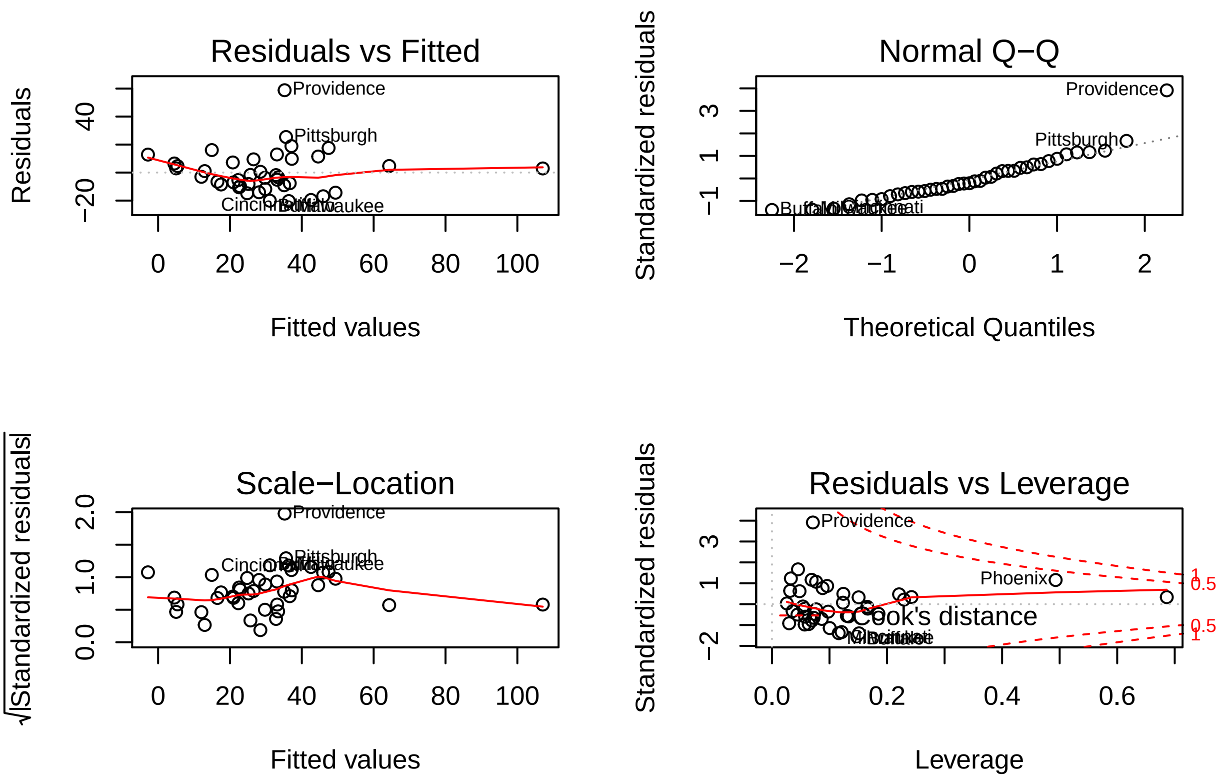
```
## Analysis of Variance Table
##
## Model 1: SO2 ~ Neg.Temp + Manuf + Wind + Precip
## Model 2: SO2 ~ (Neg.Temp + Manuf) * f.size + Wind + Precip
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     36 8752.9
## 2     30 7878.0  6    874.91 0.5553  0.762
```

8. Assess default residual plots in R for model (m4): are there any atypical residuals? Which one/s?

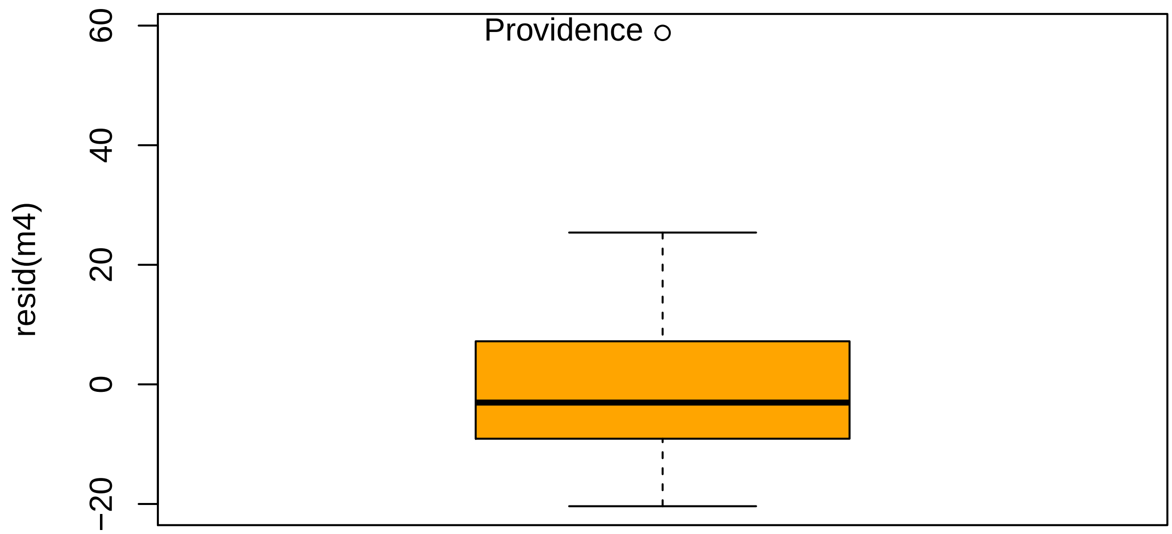
The first plot depicts the raw residuals vs fitted values according to the model, a noise pattern has to be shown for valid models. In this case no patter is present, but some large residuals on the positive Y axis (Providence, Pittsburgh). Normality of the residuals is checked with a QQPlot showing 2 cities that are not on the QQline, again these cities are Providence and Pittsburgh; residuals are too large to follow a normal distribution. On the scale-location plot, the smoother line is not flat, indicating non constant variance, but with 41 observations one has to be cautelous. The last plot on the right-down part shows an atypical city according to its leverage, so far away from the multidimensional cloud of points included in the design matrix, that does not seem relevant because the residual is close to 0.

Atypical residuals appear for Providence and Pittsburgh, so lack of fit for these 2 cities are remarkable: the observed SO2 is much, much greater than the predicted value according to m1 model. Provindece is a small city with observed SO2 of 94 micrograms per cubic meter while the model predicts 35.98, so a large lack of fit is found for this city.

```r
par(mfrow=c(2,2))
plot(m4,id.n=5)
```




```r
par(mfrow=c(1,1))
llist<-Boxplot(resid(m4),labels=row.names(usair),col="orange")   # For assessing atypical residuals
```




```r
llist
```

```
## [1] "Providence"
```

```r
predict(m1)[llist]
```

```
## Providence
##    35.9758
```

```r
usair[llist,]
```

```
##            SO2 Neg.Temp Manuf Pop Wind Precip Days f.size
## Providence  94      -50   343 179 10.6  42.75  125  Small
```

```r
sort(rstudent(m4),decreasing=T)
```

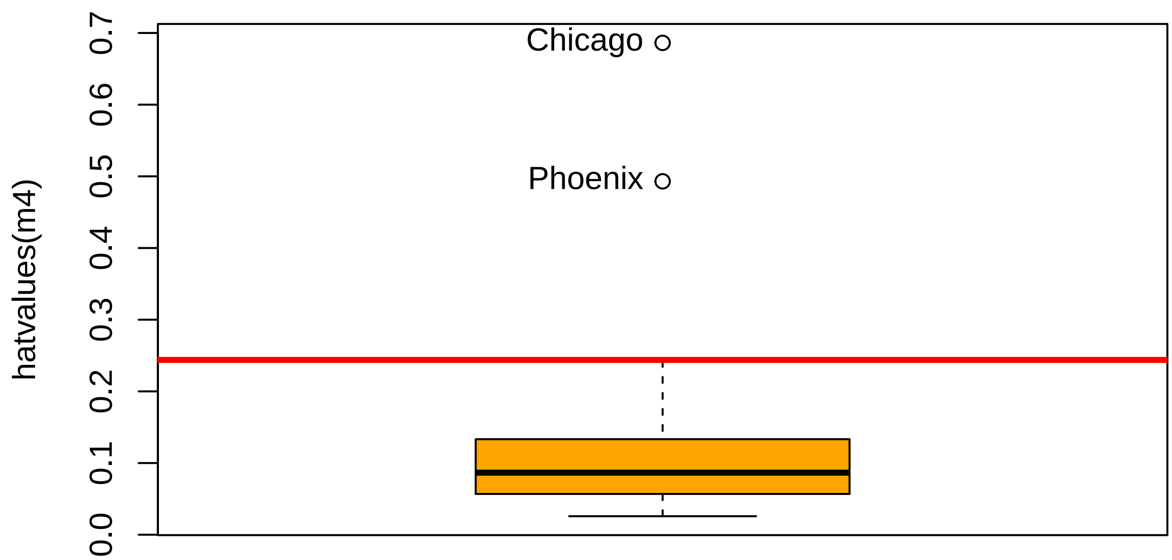
```
##          Providence          Pittsburgh           St. Louis
##          5.08645640          1.71079133          1.24103855
##           Cleveland             Phoenix             Norfolk
##          1.17199860          1.16070717          1.07366117
##              Albany            Hartford           Baltimore
##          0.86863330          0.76404315          0.63220235
##          Wilmington      Salt Lake City         Alburquerque
```

```
##           0.61461085           0.48848513           0.46534525
##                Miami              Chicago         Philadelphia
##           0.33437129           0.33120468           0.32150284
##              Wichita         Jacksonville           Washington
##           0.21463877           0.06959118           0.03423369
##              Atlanta           Charleston               Dallas
##          -0.11028928          -0.12634404          -0.21085870
##           Louisville             Richmond         Indianapolis
##          -0.22183666          -0.24594091          -0.33106286
##           Des Moines              Houston                Omaha
##          -0.35488290          -0.45770901          -0.46073975
##              Seattle               Denver          New Orleans
##          -0.49593263          -0.55620737          -0.58157410
##             Columbus          Little Rock        San Francisco
##          -0.59497892          -0.64775703          -0.70210314
##            Nashville          Kansas City              Detroit
##          -0.78299359          -0.91572614          -0.95183579
##              Memphis Minneapolis-St. Paul           Cincinnati
##          -0.97544480          -1.15100035          -1.35360262
##            Milwaukee              Buffalo
##          -1.40860700          -1.41847043
```

** 9. For your model (m4), determine the presence of observations with remarkable leverage. Specify city names, selected criteria and behavioral discrepancy.**

According to the threshold 2p/n=0.24, Chicago and Phoenix have a large leverage, since Chicago has a low residual it should not become a influent data. Phoenix is not clear.

```
llist<-Boxplot(hatvalues(m4),labels=row.names(usair),col="orange")  # For assessing atypical leverage o
abline(h=2*5/41,col="red",lwd=3)
```




```r
llist
```

```
## [1] "Phoenix" "Chicago"
```

```r
predict(m4)[llist]
```

```
##   Phoenix   Chicago
##  -2.82494 107.07042
```

```r
usair[llist,]
```

```
##         SO2 Neg.Temp Manuf  Pop Wind Precip Days f.size
## Phoenix  10    -70.3   213  582  6.0   7.05   36 Medium
## Chicago 110    -50.6  3344 3369 10.4  34.44  122  Large
```

```r
sort(hatvalues(m4),decreasing=T)
```

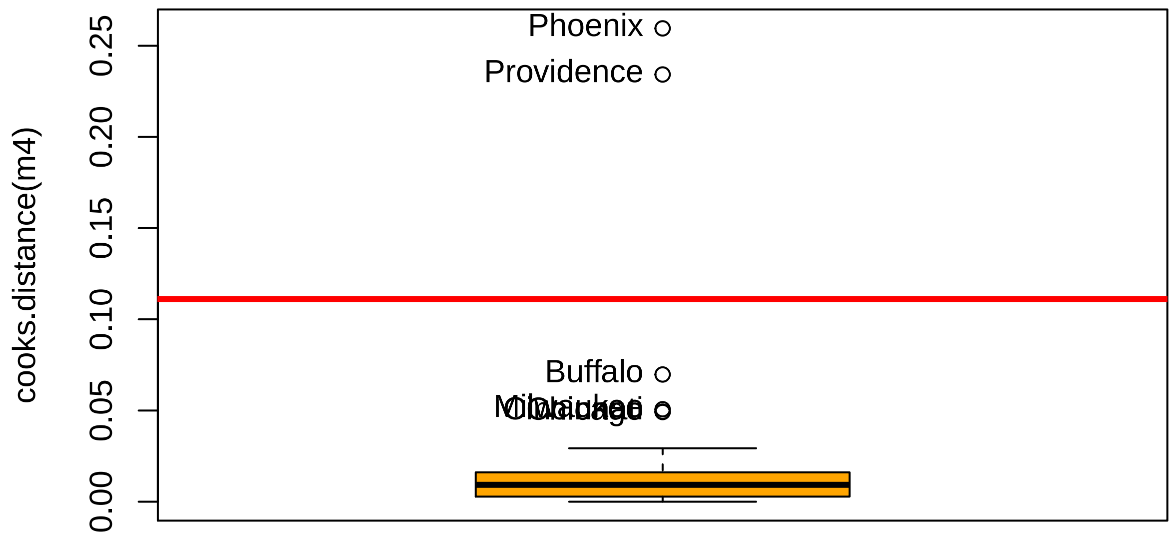
```
##              Chicago              Phoenix                Miami
##           0.68617086           0.49302683           0.24256960
##              Wichita          Alburquerque              Houston
##           0.22948521           0.22096445           0.18511110
##               Dallas           Charleston              Buffalo
##           0.16643477           0.16560793           0.15129041
```

```
##        Philadelphia          New Orleans               Denver
##          0.15059308           0.13330760           0.13017892
##      Salt Lake City         Jacksonville           Cincinnati
##          0.12408507           0.12338272           0.12114620
##           Milwaukee Minneapolis-St. Paul           Des Moines
##          0.11603424           0.10038751           0.09771408
##              Albany             Hartford        San Francisco
##          0.09589422           0.08804903           0.08651997
##            Richmond              Norfolk                Omaha
##          0.07701864           0.07676602           0.07347462
##         Little Rock           Providence            Nashville
##          0.07247181           0.07110453           0.06954632
##           Cleveland              Detroit             Columbus
##          0.06894536           0.06389416           0.05769325
##             Memphis           Louisville              Atlanta
##          0.05690168           0.05681707           0.05372839
##          Wilmington           Pittsburgh              Seattle
##          0.04792362           0.04527012           0.04435627
##        Indianapolis            St. Louis            Baltimore
##          0.03596018           0.03282401           0.03162696
##         Kansas City           Washington
##          0.02987110           0.02585207
```

10. For your final model (m4), determine the presence of actual influent data. Specify city names, selected criteria and behavior.

Providence and Phoenix are influent data, outliers of Cook's distance and over the threshold defined by Chatterjee-Hadi cut-off equal to 0.11. Providence is an outlier of residuals and Phoenix combines a medium residual with a large leverage, becoming an influent data.

```r
llist<-Boxplot(cooks.distance(m4),labels=row.names(usair),col="orange")  # For assessing atypical lever
abline(h=4/(41-5),col="red",lwd=3)
```




```
llist
```

```
## [1] "Phoenix"    "Chicago"    "Buffalo"    "Cincinnati" "Providence"
## [6] "Milwaukee"
```

```
predict(m4)[llist]
```

```
##    Phoenix    Chicago    Buffalo Cincinnati Providence  Milwaukee
##   -2.82494  107.07042   31.09575   42.56191   35.21511   36.37396
```

```
usair[llist,]
```

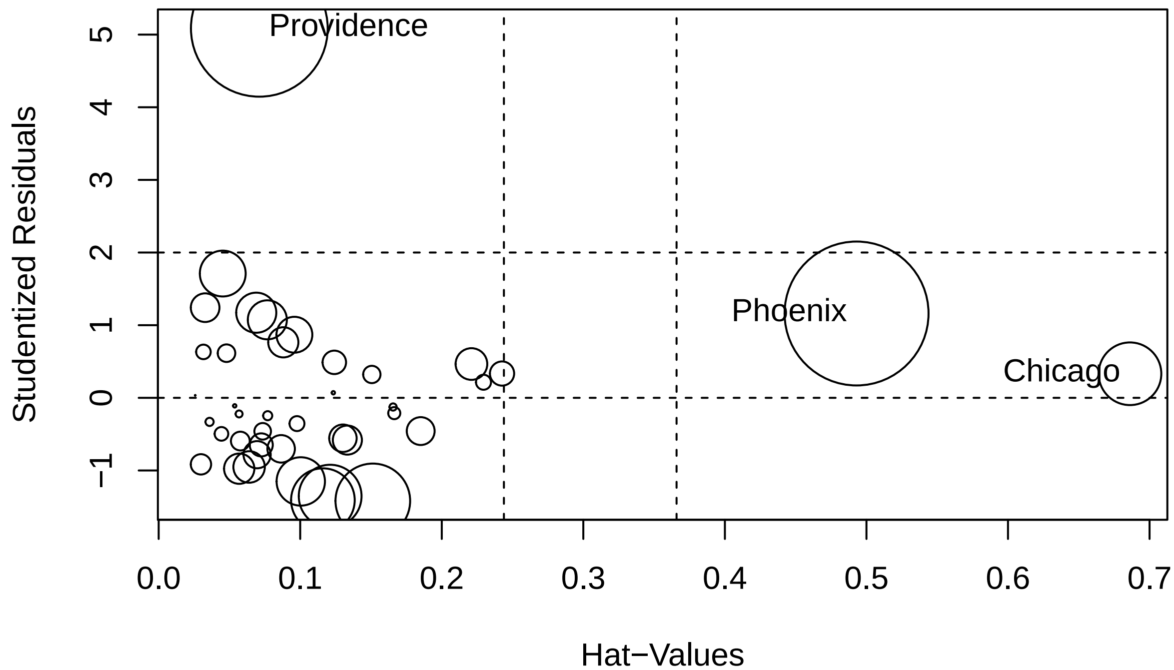
```
##             SO2 Neg.Temp Manuf  Pop Wind Precip Days f.size
## Phoenix      10    -70.3   213  582  6.0   7.05   36 Medium
## Chicago     110    -50.6  3344 3369 10.4  34.44  122  Large
## Buffalo      11    -47.1   391  463 12.4  36.11  166  Small
## Cincinnati  23    -54.0   462  453  7.1  39.04  132  Small
## Providence  94    -50.0   343  179 10.6  42.75  125  Small
## Milwaukee   16    -45.7   569  717 11.8  29.07  123 Medium
```

```
sort(cooks.distance(m4),decreasing=T)
```

```
##               Phoenix            Providence                Buffalo
```

```
##         2.595326e-01         2.342479e-01         6.977215e-02
##            Milwaukee           Cincinnati              Chicago
##         5.070456e-02         4.937191e-02         4.918553e-02
## Minneapolis-St. Paul           Pittsburgh            Cleveland
##         2.930248e-02         2.634580e-02         2.013402e-02
##              Norfolk               Albany          Alburquerque
##         1.908899e-02         1.611564e-02         1.255745e-02
##              Detroit              Memphis             Hartford
##         1.240012e-02         1.149712e-02         1.140432e-02
##          New Orleans            St. Louis              Houston
##         1.059956e-02         1.029956e-02         9.731632e-03
##        San Francisco               Denver            Nashville
##         9.471299e-03         9.441191e-03         9.264432e-03
##                Miami       Salt Lake City          Little Rock
##         7.342280e-03         6.906753e-03         6.664311e-03
##          Kansas City             Columbus           Wilmington
##         5.187223e-03         4.413974e-03         3.869730e-03
##         Philadelphia                Omaha              Wichita
##         3.758743e-03         3.442150e-03         2.818923e-03
##           Des Moines            Baltimore              Seattle
##         2.795681e-03         2.654973e-03         2.331995e-03
##               Dallas             Richmond         Indianapolis
##         1.823898e-03         1.036523e-03         8.384061e-04
##           Charleston           Louisville              Atlanta
##         6.514588e-04         6.089811e-04         1.420261e-04
##         Jacksonville           Washington
##         1.402029e-04         6.397759e-06
```

```
influencePlot(m4)
```




```
##               StudRes        Hat      CookD
## Phoenix    1.1607072 0.49302683 0.25953262
## Chicago    0.3312047 0.68617086 0.04918553
## Providence 5.0864564 0.07110453 0.23424790
```