Lab 3: Model Creation

put the date, 2017

### Names:

##### What each person contributed to the lab:

**Please make sure to show all R code and output after each question so that I can see your work.** Write a sentence for each numerical value produced describing its meaning in context with the proper units.

Consider the data, pollution.csv, which measured air pollution in 41 U.S. cities. The type of air pollution under study is the annual mean concentration of sulfur dioxide (SO2). The values of six potential explanatory variables were recorded in order to examine the variation in the sulfur dioxide concentrations

SO2 = the annual mean concentration of sulfur dioxide (micrograms per cubic meter)  
Temp = the average annual temperature in ?F  
Comp = number of manufacturing enterprises with 20 or more workers  
Pop = population size in last census (thousands)  
Wind = average annual wind speed (mph)  
Precip = average annual precipitation (inches) Pdays = average number of day with precipitation per year

We want to develop a model relating sulfur dioxide to the six possible explanatory variables to determine which are able to predict air pollution levels within the cities. Please provide plots or R output to justify your assessments. Provide detailed description of what each plot or output displays or calculates.

Your goal is to determine the best model. To do so, be sure to provide the following information

## Part 1: Exploratory Data Analysis

### 1.a.

Define each variable (response/explanatory, quantitative/categorical)

In this study we have the annual mean concentration of sulfur dioxide being our response variable, which is measured as micrograms per cubic meter and is therefore a quantitative variable. The other variables in our dataset could be considered as explanatory variables, except for the "City" variable, since this is only a column to identify where the data was collected. For our explanatory variables we can observe that they are all quantitative variables. Our first being temperature, which is the average annual temperature in Farenheits, has a range of 43.5 degrees to 75.50 degrees, a mean of 55.76 degrees and a standard deviation equal to 7.2277 degrees. Another explanatory variable "Comp", is the number of manufacturing enterprises with 20 or more workers, it has a mean of 463.1 and a standard deviation of 563.47. We also have the variable "Pop", which is population size in last census in thousands, that has a mean of 608.6 thousand and a standard deviation of 579.113 thousand. The next explanatory variable is "Wind", which is the average annual wind speed in miles per hour, that has a mean of 9.444 mph and a standard deviation of 1.428644 mph. We also have "Precip", which is the average annual precipitation measured in inches, that has a mean of 36.77 inches and a standard deviation of 11.77 inches. Finally, we have the variable "Pdays", which is the average number of days with precipitation per year, it has a mean of 113.9 days and a standard deviation of 26.506 days.

### 1.b.

Check for Linear Relationships (Tip: if any extreme outliers, try removing them to see how things look.) Is there any evidence that higher-order terms should be included?

#Provide R code and output/graphs   
library(readr)  
library(mosaic)

## Loading required package: dplyr

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

## Loading required package: lattice

## Loading required package: ggformula

## Loading required package: ggplot2

##   
## New to ggformula? Try the tutorials:   
## learnr::run\_tutorial("introduction", package = "ggformula")  
## learnr::run\_tutorial("refining", package = "ggformula")

## Loading required package: mosaicData

## Loading required package: Matrix

##   
## The 'mosaic' package masks several functions from core packages in order to add   
## additional features. The original behavior of these functions should not be affected by this.  
##   
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.

##   
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':  
##   
## mean

## The following objects are masked from 'package:dplyr':  
##   
## count, do, tally

## The following objects are masked from 'package:stats':  
##   
## binom.test, cor, cor.test, cov, fivenum, IQR, median,  
## prop.test, quantile, sd, t.test, var

## The following objects are masked from 'package:base':  
##   
## max, mean, min, prod, range, sample, sum

pollution <- read\_csv("~/Desktop/ucsc/pollution.csv")

## Warning: Missing column names filled in: 'X9' [9]

## Parsed with column specification:  
## cols(  
## City = col\_integer(),  
## SO2 = col\_double(),  
## Temp = col\_double(),  
## Comp = col\_integer(),  
## Pop = col\_integer(),  
## Wind = col\_double(),  
## Precip = col\_double(),  
## Pdays = col\_integer(),  
## X9 = col\_character()  
## )

View(pollution)  
attach(pollution)  
summary(pollution)

## City SO2 Temp Comp   
## Min. : 1 Min. :2.079 Min. :43.50 Min. : 35.0   
## 1st Qu.:11 1st Qu.:2.565 1st Qu.:50.60 1st Qu.: 181.0   
## Median :21 Median :3.258 Median :54.60 Median : 347.0   
## Mean :21 Mean :3.153 Mean :55.76 Mean : 463.1   
## 3rd Qu.:31 3rd Qu.:3.555 3rd Qu.:59.30 3rd Qu.: 462.0   
## Max. :41 Max. :4.700 Max. :75.50 Max. :3344.0   
## Pop Wind Precip Pdays   
## Min. : 71.0 Min. : 6.000 Min. : 7.05 Min. : 36.0   
## 1st Qu.: 299.0 1st Qu.: 8.700 1st Qu.:30.96 1st Qu.:103.0   
## Median : 515.0 Median : 9.300 Median :38.74 Median :115.0   
## Mean : 608.6 Mean : 9.444 Mean :36.77 Mean :113.9   
## 3rd Qu.: 717.0 3rd Qu.:10.600 3rd Qu.:43.11 3rd Qu.:128.0   
## Max. :3369.0 Max. :12.700 Max. :59.80 Max. :166.0   
## X9   
## Length:41   
## Class :character   
## Mode :character   
##   
##   
##

names(pollution)

## [1] "City" "SO2" "Temp" "Comp" "Pop" "Wind" "Precip" "Pdays"   
## [9] "X9"

sd(City)

## [1] 11.97915

sd(SO2)

## [1] 0.7023011

sd(Temp)

## [1] 7.227716

sd(Comp)

## [1] 563.4739

sd(Pop)

## [1] 579.113

sd(Wind)

## [1] 1.428644

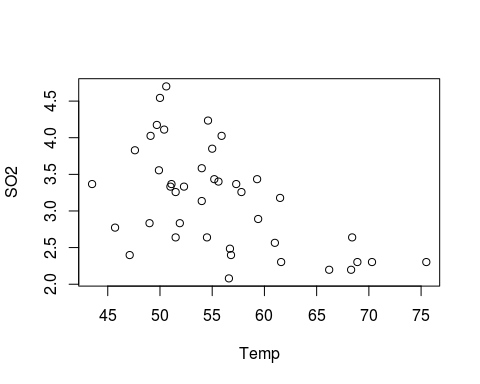
sd(Precip)

## [1] 11.77155

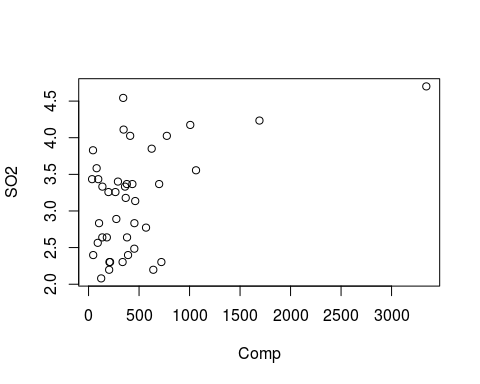
sd(Pdays)

## [1] 26.50642

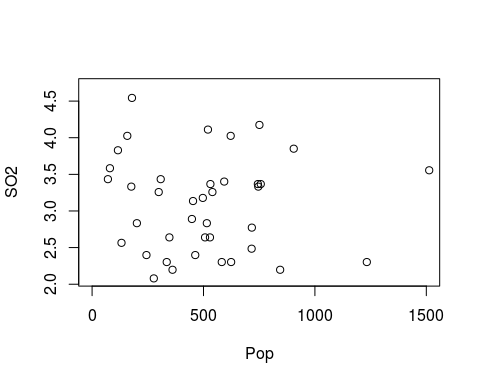
plot(Temp, SO2)



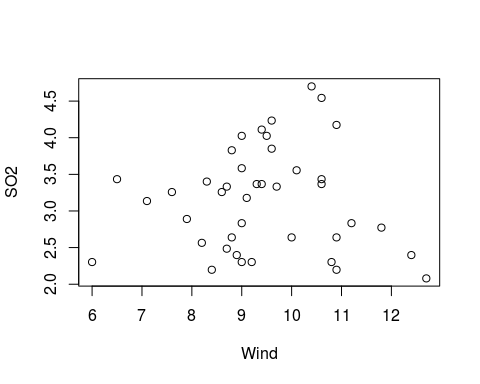
plot(Comp, SO2)



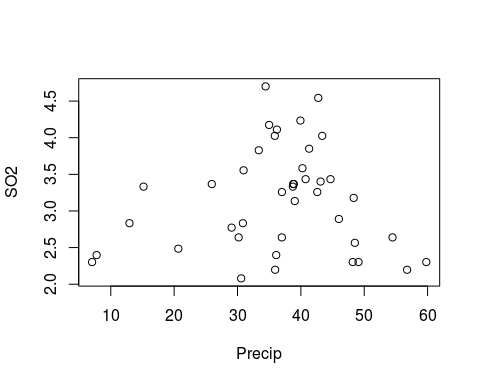
plot(Pop, SO2, xlim = c(0,1500))



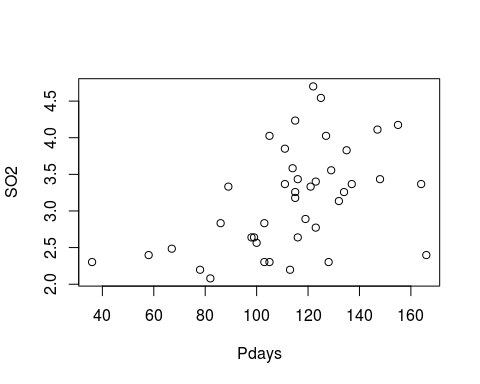
plot(Wind, SO2)



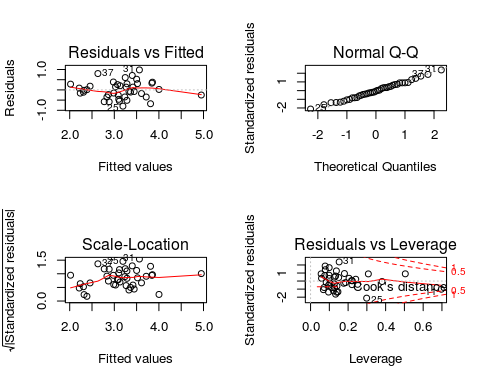
plot(Precip, SO2)



plot(Pdays, SO2)



pol.model<-lm(SO2~Temp+Comp+Pop+Wind+Precip+Pdays)  
par(mfrow=c(2,2))  
plot((pol.model))



#Here is a hint on how to remove one row from a data set, say row 2  
#when we use [] brackets, we can reference certain rows and columns of the data  
#so this code says to remove row 11 (rows are before the comma) but keep all columns (Columns are after the comma)  
#pollution[-11,]

There appears to be a negative linear relationship with the average annual temperature in Farenheits. It appears that as the the temperature increases, the SO2 levels decreases. From my second plot comparing Comp variable with SO2 levels, it appears there is a small positive linear relationship with Comp, so as the number of companies with twenty or more people in the city increases, we also have an increase in SO2 levels. From this plot, we can also observe that there might be an outlier for this data, which would need to be investigated. There does not appear to be a linear relationship between the population and the SO2 levels because there is no visible pattern in the scatterplot. From our scatterplot with Wind and SO2 levels it appears there is a negative linear relationship between these two variables. From the scatterplot comparing precipitation with SO2 levels, it does not appear as there is a linear relationship. Instead, there is evidence that a higher order term might be needed, which appears to be a quadratic term. Finally, between the variables Pdays and SO2, there does appear to be a positive linear relationship in this case. We can see from the scatterplot that as the number of Pdays increases, the SO2 level also increases.

### 1.c.

Check of collinearity (remember to check your 3b notes - how do you procede in the presence of collinearity?)

#Provide R code and output/graphs   
library(car)

##   
## Attaching package: 'car'

## The following objects are masked from 'package:mosaic':  
##   
## deltaMethod, logit

## The following object is masked from 'package:dplyr':  
##   
## recode

vif(pol.model)

## Temp Comp Pop Wind Precip Pdays   
## 3.763996 14.703652 14.340833 1.255519 3.404921 3.443651

We do appear to have collinearity with the variables Comp and Pop, since their VIF scores are 14.703 and 14.34, respectively. These scores are greater than five and more importantly greater than ten, which implies collinearity. In the presence of collinearity, we would create two separate models, which each has one of the variables but not the other. We would then follow up by comparing these two models, by the assumptions and the Adjusted R-squared to decide which one is better and what variable should be removed.

## Part 2: Use two variable selection methods

(One out of R2 and Cp, and one out of Forward and Backward):

### 2.a.

Choose one: Adjusted R2 or Mallow's Cp to select a subset of 2 best models

#Provide R code and output/graphs   
library(leaps)  
pol2\_model <- leaps(x=pollution[,3:8], y = SO2, method = "adjr2", nbest = 2, names=names(pollution[3:8]))  
pol2\_model

## $which  
## Temp Comp Pop Wind Precip Pdays  
## 1 TRUE FALSE FALSE FALSE FALSE FALSE  
## 1 FALSE TRUE FALSE FALSE FALSE FALSE  
## 2 TRUE TRUE FALSE FALSE FALSE FALSE  
## 2 FALSE TRUE FALSE FALSE FALSE TRUE  
## 3 TRUE TRUE FALSE TRUE FALSE FALSE  
## 3 TRUE TRUE FALSE FALSE TRUE FALSE  
## 4 TRUE TRUE FALSE TRUE TRUE FALSE  
## 4 TRUE TRUE FALSE TRUE FALSE TRUE  
## 5 TRUE TRUE TRUE TRUE TRUE FALSE  
## 5 TRUE TRUE FALSE TRUE TRUE TRUE  
## 6 TRUE TRUE TRUE TRUE TRUE TRUE  
##   
## $label  
## [1] "(Intercept)" "Temp" "Comp" "Pop" "Wind"   
## [6] "Precip" "Pdays"   
##   
## $size  
## [1] 2 2 3 3 4 4 5 5 6 6 7  
##   
## $adjr2  
## [1] 0.2806286 0.2149533 0.4197955 0.3785409 0.4987603 0.4792731 0.5892461  
## [8] 0.5542219 0.6046195 0.5775664 0.5930827

From this analysis, we would choose the models with adjusted R-squared of 0.6046 and 0.5930 since these are the largest two values. These are the model without Pdays variable and the full model.

### 2.b.

Choose one: Forward Selection or Backward Elimination

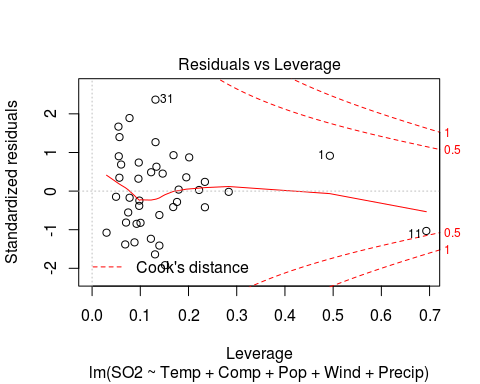
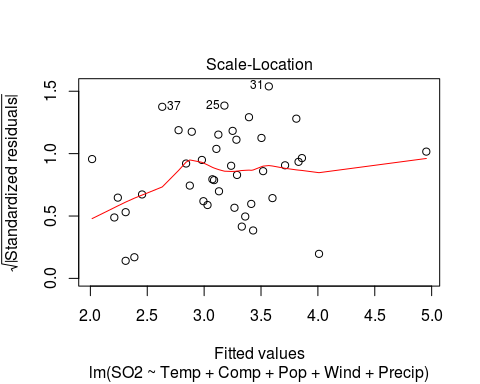
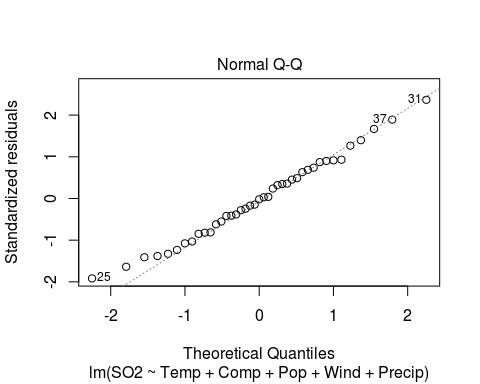
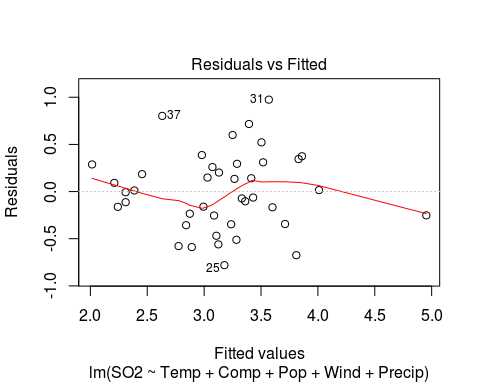
#Provide R code and output/graphs   
back.elem.lm <-step(pol.model,scope=list(lower=~1,upper=pol.model),direction = "backward",trace=T)

## Start: AIC=-59.52  
## SO2 ~ Temp + Comp + Pop + Wind + Precip + Pdays  
##   
## Df Sum of Sq RSS AIC  
## - Pdays 1 0.00154 6.8254 -61.510  
## <none> 6.8239 -59.519  
## - Pop 1 0.46856 7.2924 -58.796  
## - Precip 1 0.49128 7.3152 -58.669  
## - Comp 1 1.37989 8.2038 -53.968  
## - Wind 1 1.87301 8.6969 -51.575  
## - Temp 1 1.99194 8.8158 -51.018  
##   
## Step: AIC=-61.51  
## SO2 ~ Temp + Comp + Pop + Wind + Precip  
##   
## Df Sum of Sq RSS AIC  
## <none> 6.8254 -61.510  
## - Pop 1 0.4680 7.2934 -60.791  
## - Comp 1 1.3805 8.2059 -55.957  
## - Precip 1 1.4701 8.2956 -55.512  
## - Wind 1 1.9474 8.7728 -53.218  
## - Temp 1 4.5095 11.3349 -42.713

back.elem.lm

##   
## Call:  
## lm(formula = SO2 ~ Temp + Comp + Pop + Wind + Precip)  
##   
## Coefficients:  
## (Intercept) Temp Comp Pop Wind   
## 7.3499052 -0.0611259 0.0012642 -0.0007073 -0.1705448   
## Precip   
## 0.0181483

plot(back.elem.lm)



summary(back.elem.lm)

##   
## Call:  
## lm(formula = SO2 ~ Temp + Comp + Pop + Wind + Precip)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.78040 -0.25410 -0.00742 0.28752 0.97460   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.3499052 0.9254028 7.942 2.43e-09 \*\*\*  
## Temp -0.0611259 0.0127114 -4.809 2.86e-05 \*\*\*  
## Comp 0.0012642 0.0004752 2.661 0.01169 \*   
## Pop -0.0007073 0.0004566 -1.549 0.13035   
## Wind -0.1705448 0.0539687 -3.160 0.00325 \*\*   
## Precip 0.0181483 0.0066098 2.746 0.00947 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4416 on 35 degrees of freedom  
## Multiple R-squared: 0.654, Adjusted R-squared: 0.6046   
## F-statistic: 13.23 on 5 and 35 DF, p-value: 2.888e-07

vif(back.elem.lm)

## Temp Comp Pop Wind Precip   
## 1.731366 14.703099 14.338797 1.219354 1.241777

From the backwards elimination, we can see that the only variable that gets discarded from the full model is Pdays, until we reach the lowest possible AIC score of -61.51. This leaves us with a linear model that is better than the full, but does not seem that great since the Adjusted-R squared is only 0.6046. After having the best model, we would still need to preform follow up analysis like significance testing and testing the assuptions. From the plots and summary output of this back.elem.lm model, we can see that it does have an F-statistic of 13.23 and a p-value of 2.888e-07, which tells us that this model have some predictive power over the response variable. It also passes the first five assumptions, but fails to pass the VIF assumption since the VIF score for Comp and Pop are still about 14 to 15.

#Provide R code and output/graphs   
base <- lm(SO2~Temp)  
for.lm <- step(base, scope = list(lower=~1, upper=pol.model), direction = "forward",trace = T)

## Start: AIC=-40.53  
## SO2 ~ Temp  
##   
## Df Sum of Sq RSS AIC  
## + Comp 1 2.96317 10.875 -48.413  
## + Pop 1 1.89361 11.944 -44.567  
## + Precip 1 1.55717 12.281 -43.428  
## + Pdays 1 1.43505 12.403 -43.022  
## + Wind 1 0.94142 12.896 -41.422  
## <none> 13.838 -40.533  
##   
## Step: AIC=-48.41  
## SO2 ~ Temp + Comp  
##   
## Df Sum of Sq RSS AIC  
## + Wind 1 1.7272 9.1473 -53.505  
## + Precip 1 1.3716 9.5030 -51.941  
## + Pdays 1 1.2154 9.6591 -51.272  
## + Pop 1 1.0819 9.7927 -50.709  
## <none> 10.8746 -48.413  
##   
## Step: AIC=-53.5  
## SO2 ~ Temp + Comp + Wind  
##   
## Df Sum of Sq RSS AIC  
## + Precip 1 1.85390 7.2934 -60.791  
## + Pdays 1 1.23201 7.9153 -57.436  
## + Pop 1 0.85177 8.2956 -55.512  
## <none> 9.1473 -53.505  
##   
## Step: AIC=-60.79  
## SO2 ~ Temp + Comp + Wind + Precip  
##   
## Df Sum of Sq RSS AIC  
## + Pop 1 0.46799 6.8254 -61.510  
## <none> 7.2934 -60.791  
## + Pdays 1 0.00097 7.2924 -58.796  
##   
## Step: AIC=-61.51  
## SO2 ~ Temp + Comp + Wind + Precip + Pop  
##   
## Df Sum of Sq RSS AIC  
## <none> 6.8254 -61.510  
## + Pdays 1 0.0015434 6.8239 -59.519

for.lm

##   
## Call:  
## lm(formula = SO2 ~ Temp + Comp + Wind + Precip + Pop)  
##   
## Coefficients:  
## (Intercept) Temp Comp Wind Precip   
## 7.3499052 -0.0611259 0.0012642 -0.1705448 0.0181483   
## Pop   
## -0.0007073

summary(for.lm)

##   
## Call:  
## lm(formula = SO2 ~ Temp + Comp + Wind + Precip + Pop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.78040 -0.25410 -0.00742 0.28752 0.97460   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.3499052 0.9254028 7.942 2.43e-09 \*\*\*  
## Temp -0.0611259 0.0127114 -4.809 2.86e-05 \*\*\*  
## Comp 0.0012642 0.0004752 2.661 0.01169 \*   
## Wind -0.1705448 0.0539687 -3.160 0.00325 \*\*   
## Precip 0.0181483 0.0066098 2.746 0.00947 \*\*   
## Pop -0.0007073 0.0004566 -1.549 0.13035   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4416 on 35 degrees of freedom  
## Multiple R-squared: 0.654, Adjusted R-squared: 0.6046   
## F-statistic: 13.23 on 5 and 35 DF, p-value: 2.888e-07

We observe the same pattern from the forward selection process as the backwards elimination. Also we attain the same model at the end.

### 2.c.

Compare and contrast the final model options between the two methods you used

Judging from the two methods I used, the final model given by the backwards elimination was the best option due to the reduction of collinearity by removing Pdays variable. In the full model, we had an adjusted R-squared of one for all models except one model, which was most likely caused by the collinearity between Pdays and precipitation. Also, we can see that the assumptions were met for both models, except for the VIF assumption that was not met by either model.

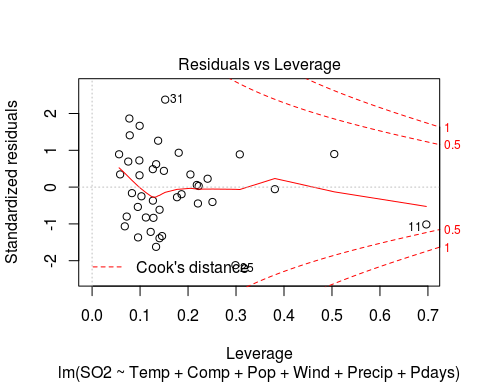
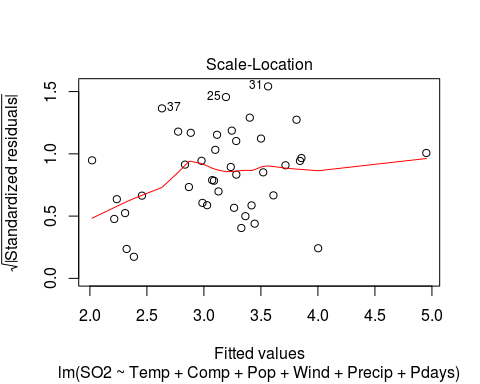
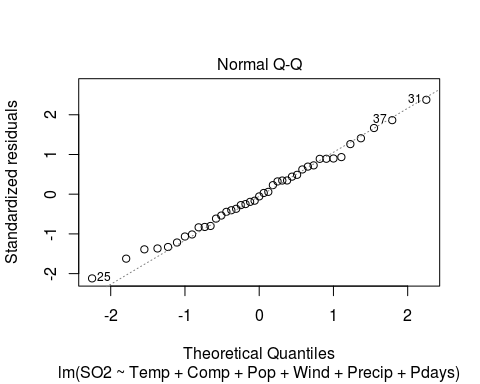
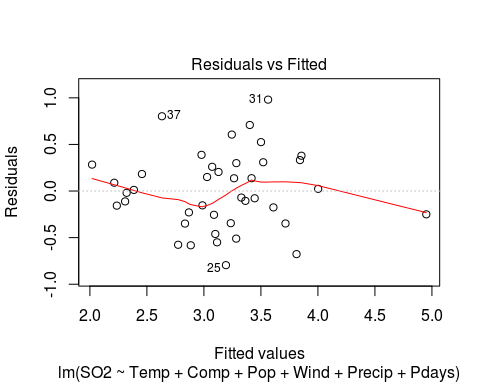
## Part 3: Detailed description of the model formulation

Why did you choose the models you chose - provide statistical and practical evidence. Provide necessary code and output to support your decisions.

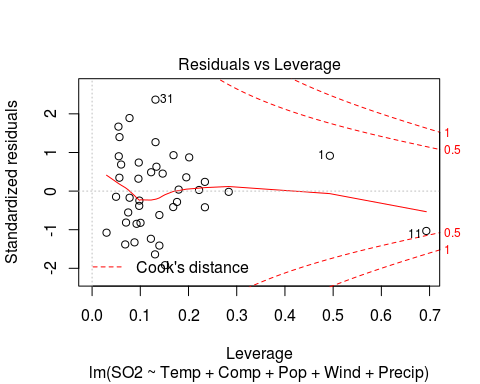
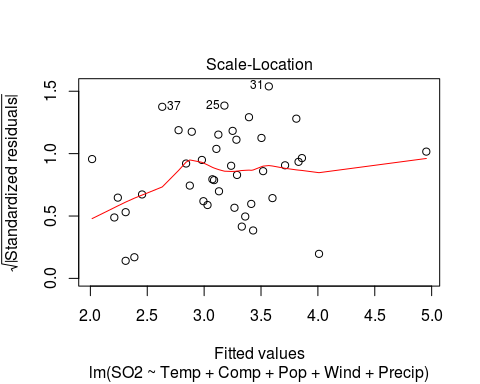
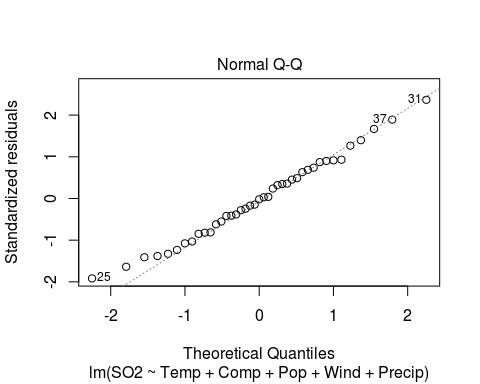
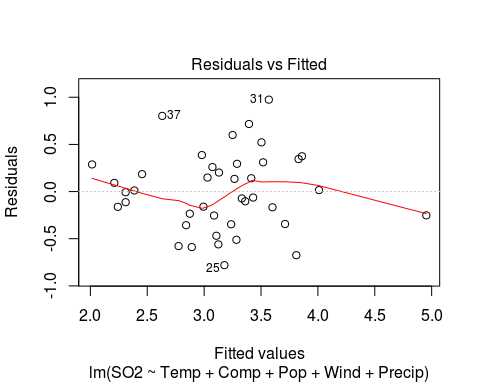
### 3.a.

Detailed description of assessment of model assumptions

#Provide R code and output/graphs   
plot(pol.model)



plot(back.elem.lm)



vif(pol.model)

## Temp Comp Pop Wind Precip Pdays   
## 3.763996 14.703652 14.340833 1.255519 3.404921 3.443651

vif(back.elem.lm)

## Temp Comp Pop Wind Precip   
## 1.731366 14.703099 14.338797 1.219354 1.241777

In the full model, we appear to fail the assumption of independence of residuals, since there appears to be an obvious pattern and could imply a higher order term. The other assumptions for the full model are met, except the VIF assumption as stated before, since the variables "Comp" and "Pop" have a significantly high VIF score. In the model that resulted from the backwards elimination, we appear to fail the assumption of independence of residuals since there is also a pattern with curvature. We also once again, fail the VIF assumption, due to the variables "Comp" and "Pop" having a significantly high VIF score.

### 3.b.

Detailed description of assessment of significance tests

#Provide R code and output/graphs   
summary(pol.model)

##   
## Call:  
## lm(formula = SO2 ~ Temp + Comp + Pop + Wind + Precip + Pdays)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.79548 -0.25540 -0.01969 0.28329 0.98029   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.2531874 1.4483719 5.008 1.68e-05 \*\*\*  
## Temp -0.0599007 0.0190138 -3.150 0.00339 \*\*   
## Comp 0.0012640 0.0004820 2.622 0.01298 \*   
## Pop -0.0007077 0.0004632 -1.528 0.13578   
## Wind -0.1697180 0.0555564 -3.055 0.00436 \*\*   
## Precip 0.0173722 0.0111037 1.565 0.12695   
## Pdays 0.0004349 0.0049591 0.088 0.93064   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.448 on 34 degrees of freedom  
## Multiple R-squared: 0.6541, Adjusted R-squared: 0.5931   
## F-statistic: 10.72 on 6 and 34 DF, p-value: 1.126e-06

summary(back.elem.lm)

##   
## Call:  
## lm(formula = SO2 ~ Temp + Comp + Pop + Wind + Precip)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.78040 -0.25410 -0.00742 0.28752 0.97460   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.3499052 0.9254028 7.942 2.43e-09 \*\*\*  
## Temp -0.0611259 0.0127114 -4.809 2.86e-05 \*\*\*  
## Comp 0.0012642 0.0004752 2.661 0.01169 \*   
## Pop -0.0007073 0.0004566 -1.549 0.13035   
## Wind -0.1705448 0.0539687 -3.160 0.00325 \*\*   
## Precip 0.0181483 0.0066098 2.746 0.00947 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4416 on 35 degrees of freedom  
## Multiple R-squared: 0.654, Adjusted R-squared: 0.6046   
## F-statistic: 13.23 on 5 and 35 DF, p-value: 2.888e-07

For the full model, we have an F-statistic of 10.72 and a p-value of 1.126e-06, which implies that there is some predicitve power behind the model. However, we can also see that the variables Pop, Precip and Pdays appear not to be significant due to their p-value being greater than our significance level of 0.05. We can see from the full mode that the adjusted R-squared being 0.5931 is quite small. For the model without Pdays, we have an F-statistic of 13.23 and a p-value of 2.888e-07, which implies that there is some predicitve power behind the model. We can see that the only variable that is not significant because of a large p-value is Pop. Finally, the adjusted R-squared is 0.6046, which is still small, but it is larger than that of the full model. That is why this model

### 3.c.

Detailed description of final model including interpretation of the parameter values in the context of the problem.

#Provide R code and output/graphs   
final.lm <- lm(SO2 ~ Temp + Comp + Pop + Wind + Precip )  
summary(final.lm)

##   
## Call:  
## lm(formula = SO2 ~ Temp + Comp + Pop + Wind + Precip)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.78040 -0.25410 -0.00742 0.28752 0.97460   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.3499052 0.9254028 7.942 2.43e-09 \*\*\*  
## Temp -0.0611259 0.0127114 -4.809 2.86e-05 \*\*\*  
## Comp 0.0012642 0.0004752 2.661 0.01169 \*   
## Pop -0.0007073 0.0004566 -1.549 0.13035   
## Wind -0.1705448 0.0539687 -3.160 0.00325 \*\*   
## Precip 0.0181483 0.0066098 2.746 0.00947 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4416 on 35 degrees of freedom  
## Multiple R-squared: 0.654, Adjusted R-squared: 0.6046   
## F-statistic: 13.23 on 5 and 35 DF, p-value: 2.888e-07

the final model will be . The intercept coefficient tells us that when all variables are equal to zero, we will have a SO2 level of 7.349, which is not practical in this context because it is an extremely large value for SO2, according to our data. Also when Comp, Pop, Wind and Prec variables are held constant, the SO2 levels will decrease by 0.061 units, when Temperature decreases by one Farenheit. When the variables Temp, Pop, Wind and Precip are held constant, the SO2 will increase by 0.00126 when the number of companies with over 20 employees increases by one. These other interpretations are similar for Pop, Wind and Precip variables.

### 3.d.

General conclusion based on the model - what factors seem to contribute to SO2 output. To what population/situations can we apply our findings?

The factors that seem to contribute to SO2 output are Temp, Comp, Pop, Wind and Precip according to our test and final model. We can apply these findings to the population being the United States, if this dataset of forty one cities was chosen randomly and is representative of the population. If this is not the case, we would apply these findings mainly for the forty one cities or the states where these cities were selected from.