

MVA_Ass_7.R

mihikagupta

2020-10-29

```
##### Assignment 7 #####
## Applying Multiple Regression Analysis

#Getting working directory
getwd()

## [1] "/Users/mihikagupta/Desktop/SEM_2/MVA"

#Setting directory to load data set
setwd("/Users/mihikagupta/Desktop/SEM_2/MVA")

#Reading the data into a data frame
df <- read.csv(file = 'US_Acc_June20.csv')
num <- read.csv(file = 'num.csv')
# Performing clustering on the first 500 records for now to achieve easy and quick results and test the
attach(num)
# Printing first few columns of data set for inference
#head(df)

## Setting random seed to shuffle data before splitting
set.seed(23)

#Checking number of rows
#rows<-sample(nrow(df))

#Shuffling the data
#mva<-df[rows,]

#Taking the required number of instances from the shuffled data to reduce any biases
#mva<-mva[950000:1000000,]

#Checking the structure of the data set
#str(mva)

# Checking the number of rows and columns in the current uncleaned dataset
#ncol(mva)
#nrow(mva)

# Printing all the column names to find and filter the relevant and irrelevant attributes
#names<-names(mva)
#names

## DATA CLEANING ##
```

```

#Dropping the surplus attributes which do not contribute to the analysis
#mva <- mva[-c(1:3,7:10,13,14,19,21:23,33,47:49)] 

#Checking for any null values in the present data set
# is.na(mva[,])

#Checking which rows have all the values filled and complete
# complete.cases(mva)

#Making a new dataframe with only the rows that have complete information and all values filled
#Mva<-na.omit(mva)
#Mva<-Mva[!(is.na(Mva$Sunrise_Sunset) | Mva$Sunrise_Sunset==""), ]
#Mva<- Mva[complete.cases(Mva),]
#Verifying for missing values in the new dataframe
#complete.cases(Mva)
#unique(Mva$Sunrise_Sunset)

# Creating new dataframe with only the numerical attributes to perform statistical functions
#num<-Mva[,c(1,4,11:15,17,18)]
#write.csv(num, "/Users/mihikagupta/Desktop/SEM_2/MVA/num.csv", row.names = FALSE)

# Scaling the new data set for better accuracies
# num<-scale(num)

# Checking the dimensions of the data
nrow(num)

## [1] 18250
ncol(num)

## [1] 9
names(num)

## [1] "Severity"          "Distance.mi."      "Temperature.F."
## [4] "Wind_Chill.F."     "Humidity..."       "Pressure.in."
## [7] "Visibility.mi."    "Wind_Speed.mph."   "Precipitation.in."
names(num)[names(num) == "Distance.mi."] <- "dist"
names(num)[names(num) == "Temperature.F."] <- "temp"
names(num)[names(num) == "Wind_Chill.F."] <- "windchill"
names(num)[names(num) == "Humidity..."] <- "humidity"
names(num)[names(num) == "Pressure.in."] <- "pressure"
names(num)[names(num) == "Visibility.mi."] <- "visibility"
names(num)[names(num) == "Wind_Speed.mph."] <- "windspeed"
names(num)[names(num) == "Precipitation.in."] <- "precip"
names(num)

## [1] "Severity"    "dist"        "temp"        "windchill"    "humidity"
## [6] "pressure"    "visibility"  "windspeed"   "precip"
# finding covariance,Covariance measures the linear relationship between two variables. ... The correlation
cov(num)

##                 Severity          dist         temp      windchill      humidity
## Severity  0.310236580  0.1670840167 -0.35171651 -0.47620943  0.6439822

```

```

## dist      0.167084017 2.4204304155 -0.46482491 -0.64706455 0.5613710
## temp     -0.351716514 -0.4648249146 354.39442593 397.07844765 -185.2055863
## windchill -0.476209435 -0.6470645543 397.07844765 450.54026468 -200.8257533
## humidity   0.643982246 0.5613709540 -185.20558632 -200.82575330 530.2655286
## pressure   -0.005748438 -0.0567534649 1.01969027 1.16018376 5.2382330
## visibility -0.049013351 -0.1056278936 16.89140979 19.68781198 -27.8803082
## windspeed   0.143762668 0.1759889303 -1.21123799 -7.64665928 -18.3282076
## precip      0.001178471 0.0004767836 -0.05005218 -0.05860904 0.1685233
##           pressure visibility windspeed precip
## Severity   -0.005748438 -0.04901335 0.14376267 0.0011784711
## dist       -0.056753465 -0.10562789 0.17598893 0.0004767836
## temp        1.019690266 16.89140979 -1.21123799 -0.0500521830
## windchill   1.160183765 19.68781198 -7.64665928 -0.0586090372
## humidity    5.238233012 -27.88030819 -18.32820757 0.1685233120
## pressure    1.319215237 -0.28537521 -0.29431416 0.0015056899
## visibility -0.285375212 8.02960442 -0.42228979 -0.0303542240
## windspeed   -0.294314160 -0.42228979 29.19965932 0.0161919228
## precip      0.001505690 -0.03035422 0.01619192 0.0070357365

# here we find that the highest covariances with severity in either directions, positive or negative are

# Performing multiple regression on dataset
fit<-lm(Severity~dist+temp+windchill+humidity,data = num)

# showing results
summary(fit)

## 
## Call:
## lm(formula = Severity ~ dist + temp + windchill + humidity, data = num)
## 
## Residuals:
##   Min     1Q Median     3Q    Max 
## -3.6912 -0.2745 -0.2413  0.5915  1.8145 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2.0659506  0.0292263 70.688 < 2e-16 ***
## dist        0.0679376  0.0025946 26.184 < 2e-16 ***
## temp        0.0160049  0.0019505  8.205 2.45e-16 ***
## windchill   -0.0145144  0.0017157 -8.460 < 2e-16 ***
## humidity    0.0012356  0.0001955  6.319 2.70e-10 ***
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5449 on 18245 degrees of freedom
## Multiple R-squared:  0.04329, Adjusted R-squared:  0.04308 
## F-statistic: 206.4 on 4 and 18245 DF, p-value: < 2.2e-16

coefficients(fit)

## (Intercept)          dist          temp         windchill        humidity
## 2.065950617 0.067937558 0.016004929 -0.014514415 0.001235566

# plotting the scatterplot matrix
library(ggplot2)

```

```

require(GGally)

## Loading required package: GGally

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

# ggpairs(data=num,title = "US Accidents Data")

# confidence levels
confint(fit,level=0.95)

##          2.5 %      97.5 %
## (Intercept) 2.008664227 2.123237007
## dist         0.062851860 0.073023256
## temp         0.012181739 0.019828119
## windchill    -0.017877309 -0.011151520
## humidity     0.000852287 0.001618845

# Predicted Values
fit_values<-fitted(fit)
res<-residuals(fit)

# ANOVA Table
anova(fit)

## Analysis of Variance Table
##
## Response: Severity
##             Df Sum Sq Mean Sq F value    Pr(>F)
## dist         1  210.5  210.483 709.001 < 2.2e-16 ***
## temp         1      5.3      5.262  17.725 2.565e-05 ***
## windchill    1     17.5     17.480  58.882 1.758e-14 ***
## humidity     1     11.9     11.853  39.926 2.699e-10 ***
## Residuals 18245 5416.4     0.297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

vcov(fit)

##          (Intercept)       dist        temp      windchill
## (Intercept) 8.541794e-04 -3.435445e-07 -3.888329e-05 3.028434e-05
## dist        -3.435445e-07  6.732050e-06 -1.738325e-07 1.595721e-07
## temp        -3.888329e-05 -1.738325e-07  3.804504e-06 -3.321667e-06
## windchill    3.028434e-05  1.595721e-07 -3.321667e-06 2.943559e-06
## humidity    -4.117036e-06 -7.407106e-09  7.097681e-08 -4.552161e-08
##          humidity
## (Intercept) -4.117036e-06
## dist        -7.407106e-09
## temp        7.097681e-08
## windchill   -4.552161e-08
## humidity    3.823631e-08

cov2cor(vcov(fit))

##          (Intercept)       dist        temp      windchill      humidity
## (Intercept) 1.0000000000 -0.004530382 -0.68208648  0.60395855 -0.72039741

```

```

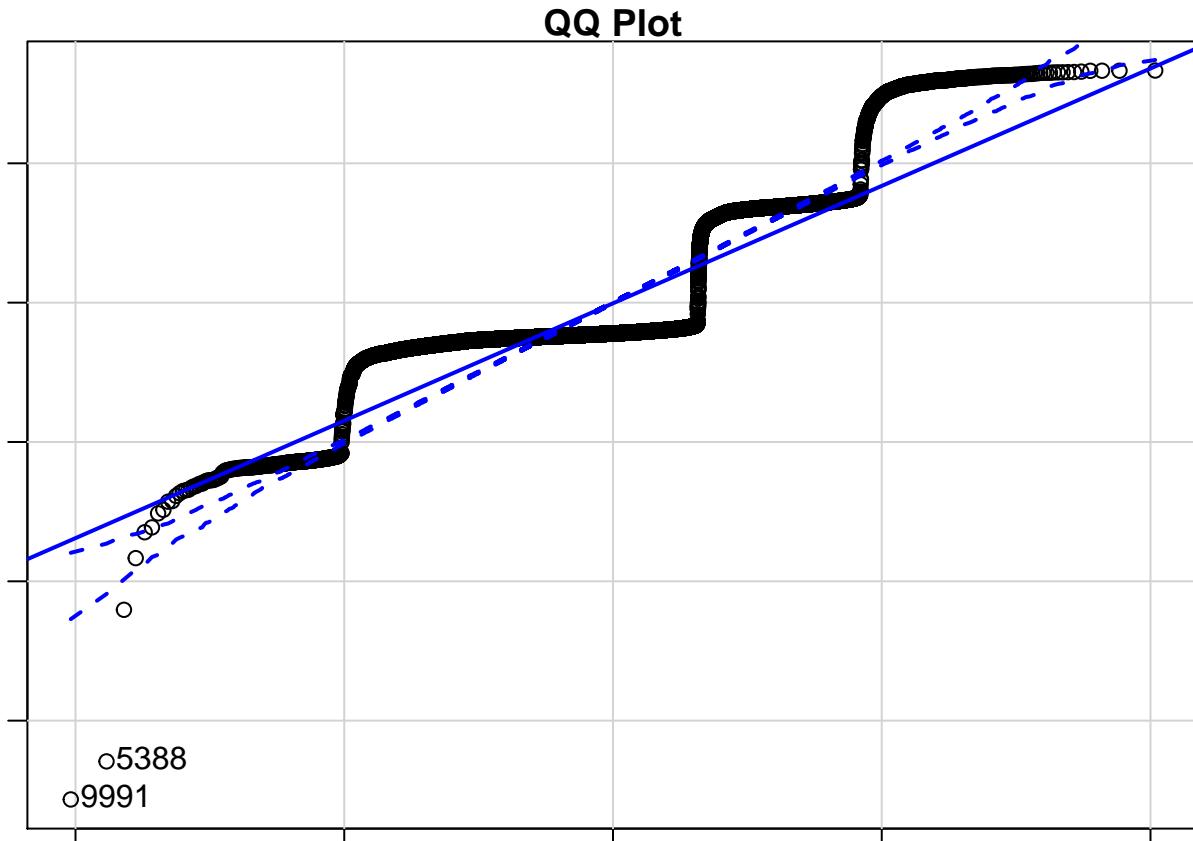
## dist      -0.004530382  1.000000000 -0.03434854  0.03584652 -0.01459946
## temp     -0.682086482 -0.034348542  1.000000000 -0.99259204  0.18609276
## windchill 0.603958553  0.035846523 -0.99259204  1.000000000 -0.13568857
## humidity -0.720397413 -0.014599464  0.18609276 -0.13568857  1.000000000
# Measure influence
temp <- influence.measures(fit)
# View(temp)

# Assessing Outliers
library(car)

## Loading required package: carData
outlierTest(fit)

##          rstudent unadjusted p-value Bonferroni p
## 9991    -7.131187      1.0323e-12   1.8840e-08
## 5388    -6.585259      4.6648e-11   8.5132e-07
par(mar=c(1,1,1,1))
qqPlot(fit, main="QQ Plot")

```

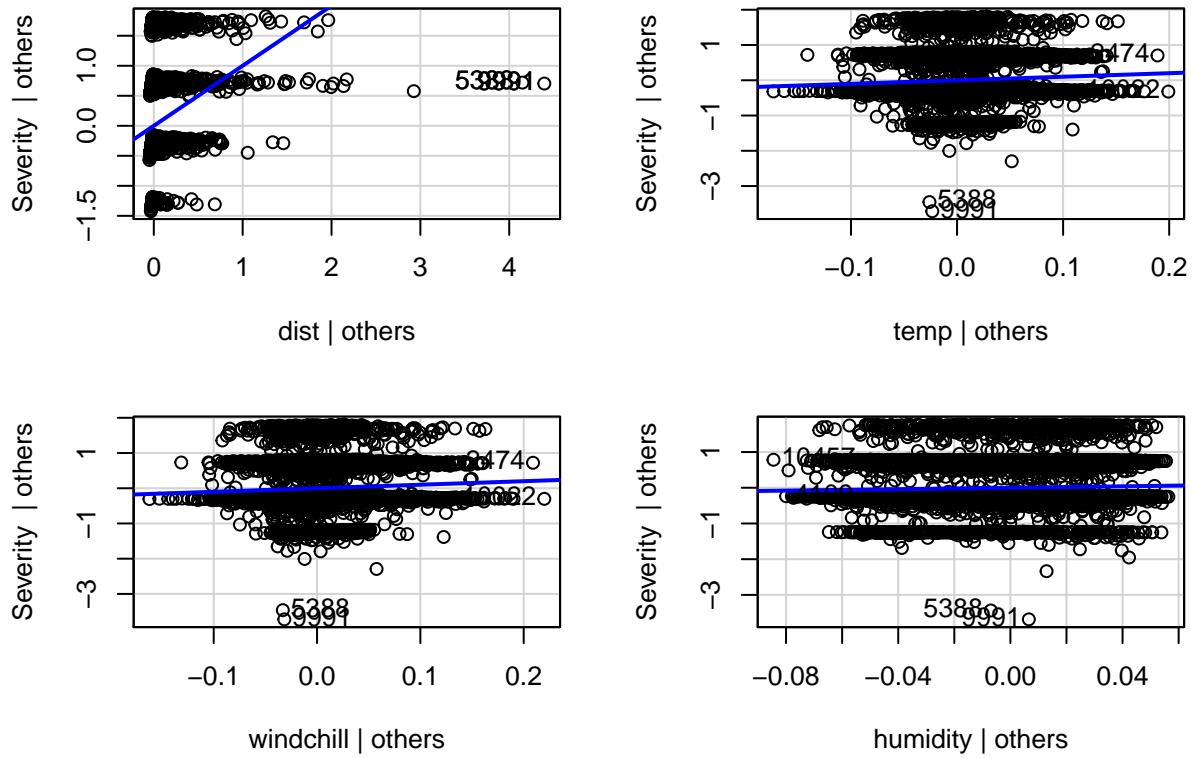


```

## [1] 5388 9991
leveragePlots(fit) # leverage plots

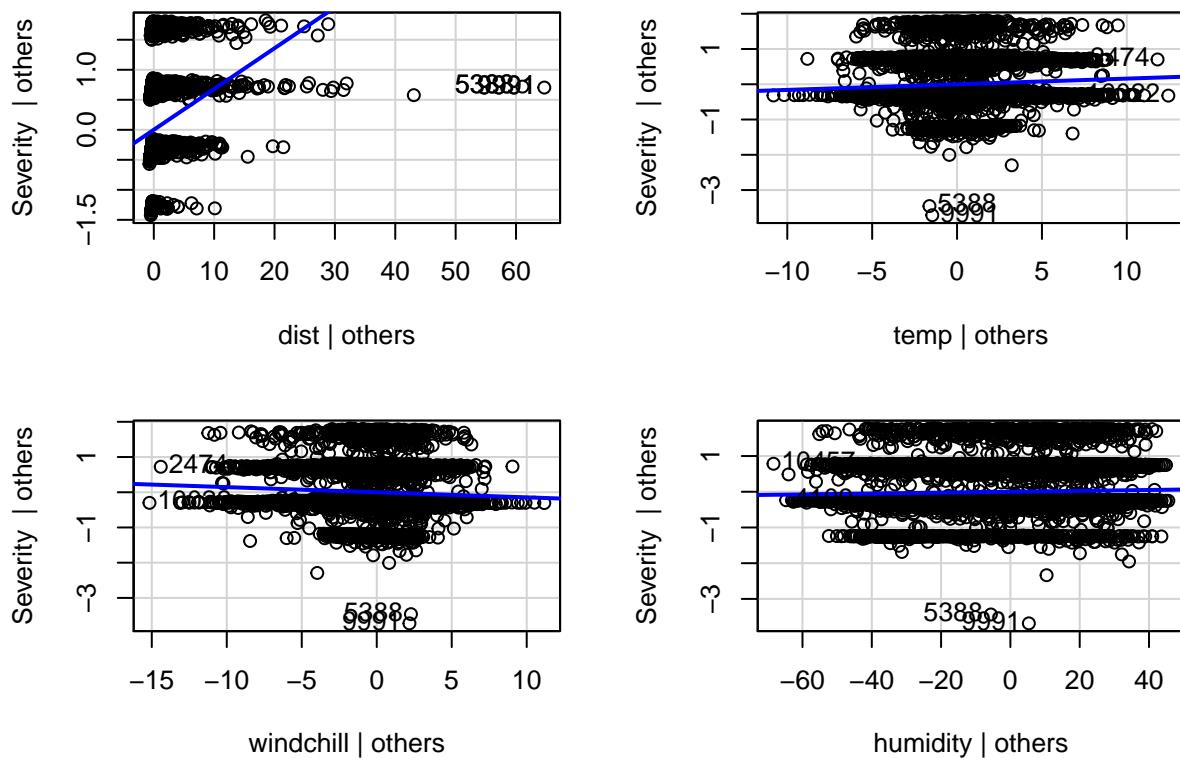
```

Leverage Plots



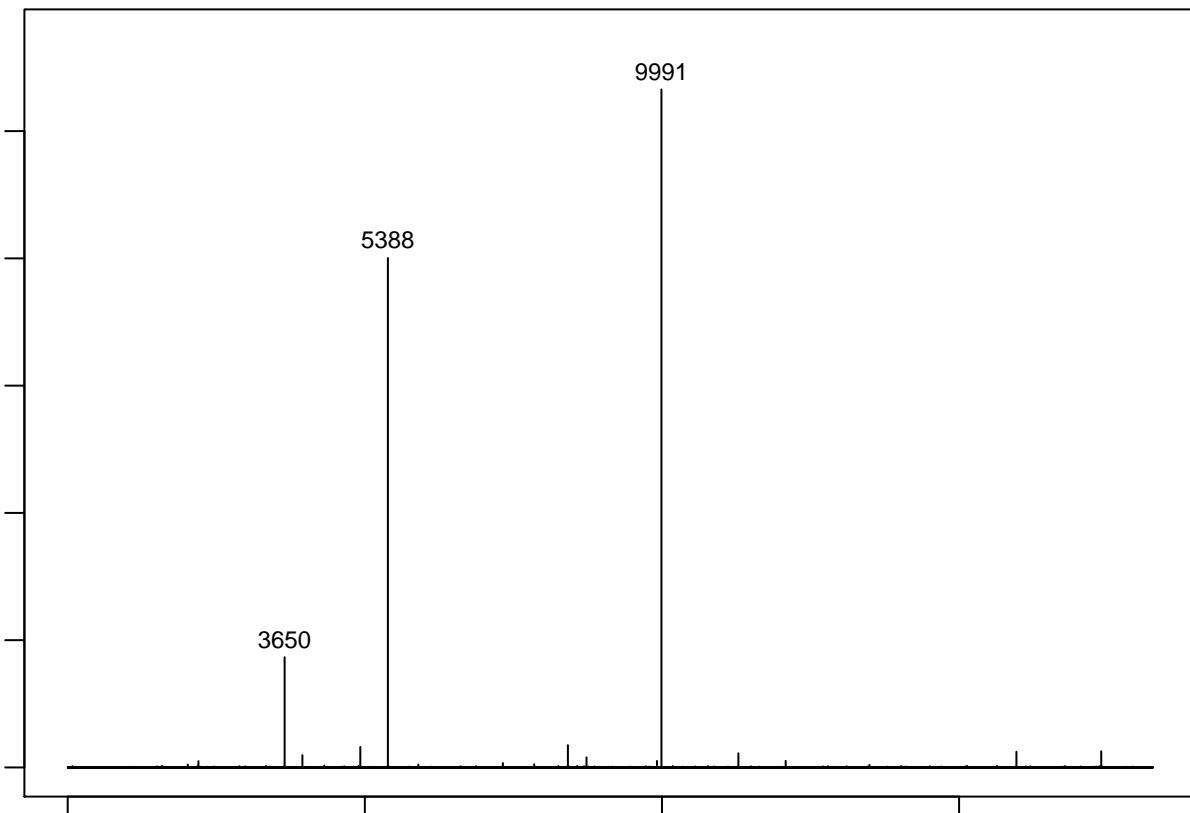
```
# Influential Observations  
# added variable plots  
avPlots(fit)
```

Added-Variable Plots



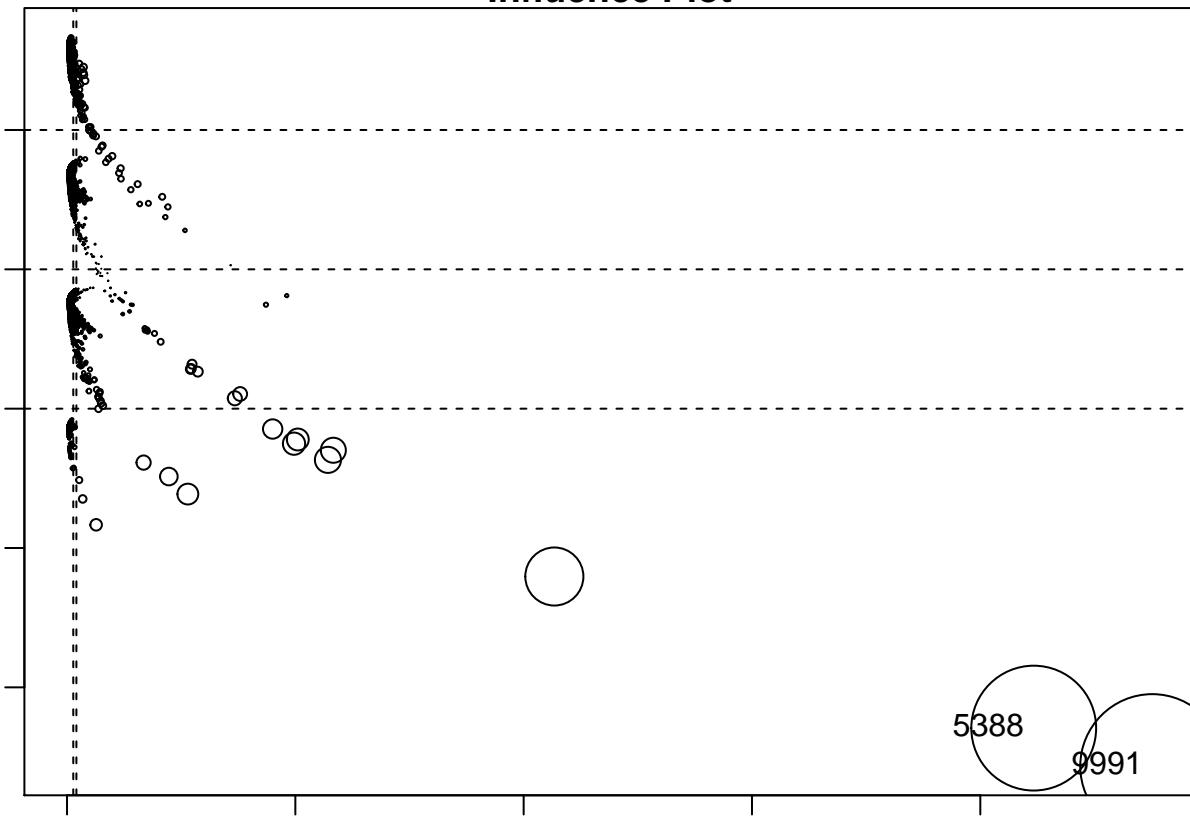
```
# Cook's D plot
# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(num)-length(fit$coefficients)-2))
plot(fit, which=4, cook.levels=cutoff)
```

Cook's distance

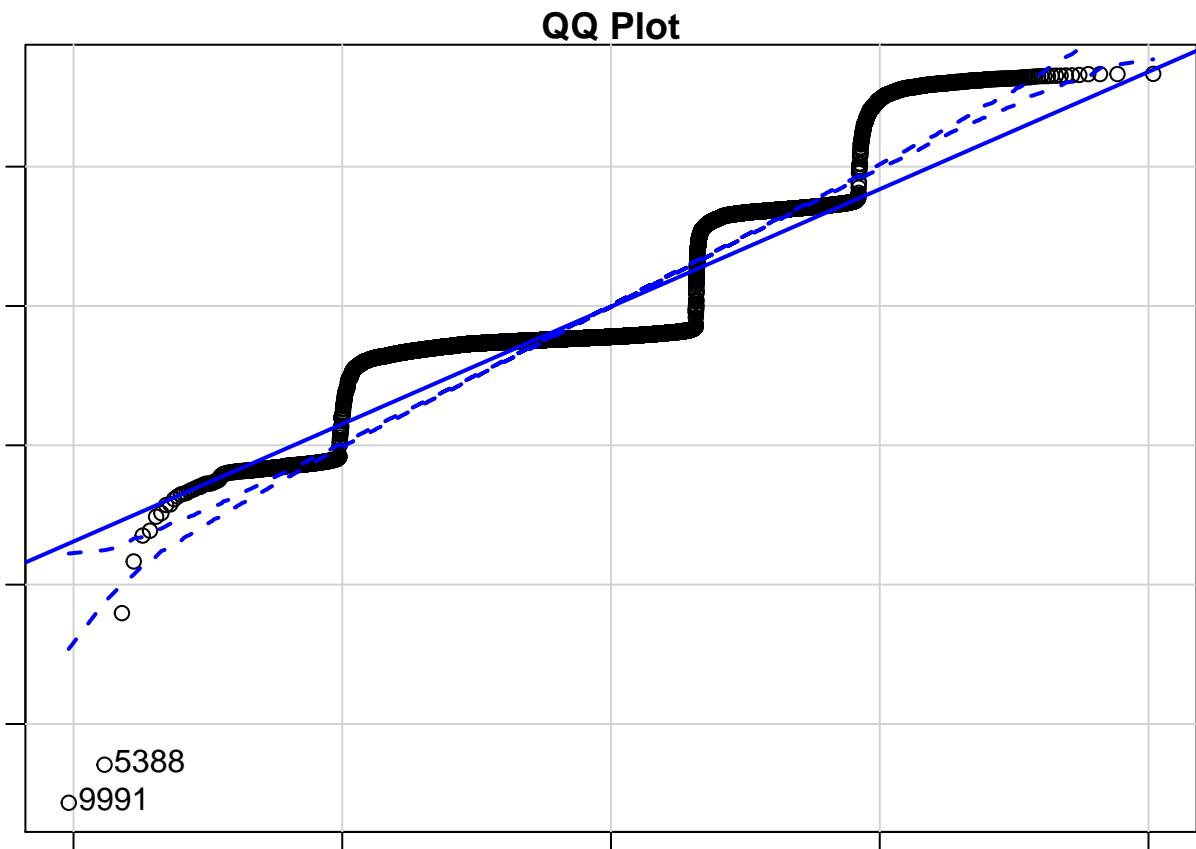


```
# Influence Plot
influencePlot(fit, id.method="identify", main="Influence Plot", sub="Circle size is proportional to Cook's
## Warning in plot.window(...): "id.method" is not a graphical parameter
## Warning in plot.xy(xy, type, ...): "id.method" is not a graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "id.method" is not
## a graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "id.method" is not
## a graphical parameter
## Warning in box(...): "id.method" is not a graphical parameter
## Warning in title(...): "id.method" is not a graphical parameter
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.method" is not a
## graphical parameter
```

Influence Plot

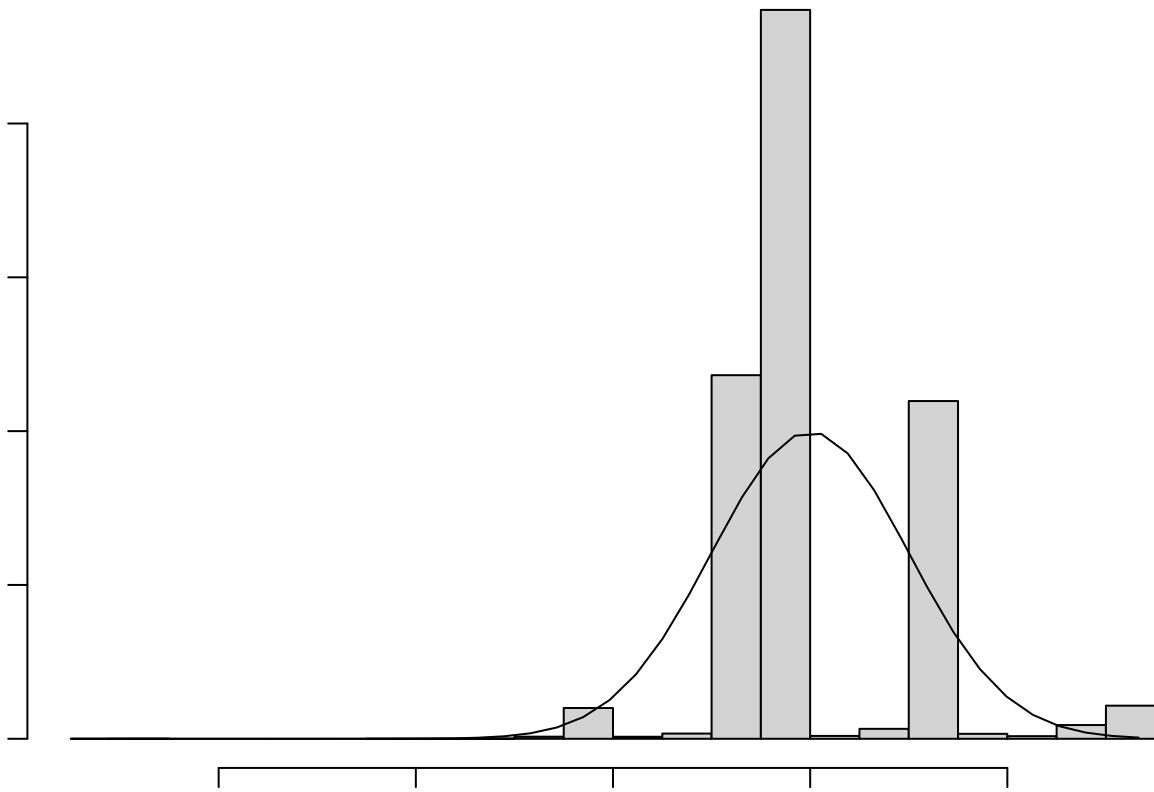


```
##          StudRes      Hat     CookD
## 5388 -6.585259 0.08468002 0.8005278
## 9991 -7.131187 0.09506396 1.0655322
# Normality of Residuals
# qq plot for studentized resid
qqPlot(fit, main="QQ Plot")
```



```
## [1] 5388 9991
# distribution of studentized residuals
library(MASS)
sresid <- studres(fit)
hist(sresid, freq=FALSE,
     main="Distribution of Studentized Residuals")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```

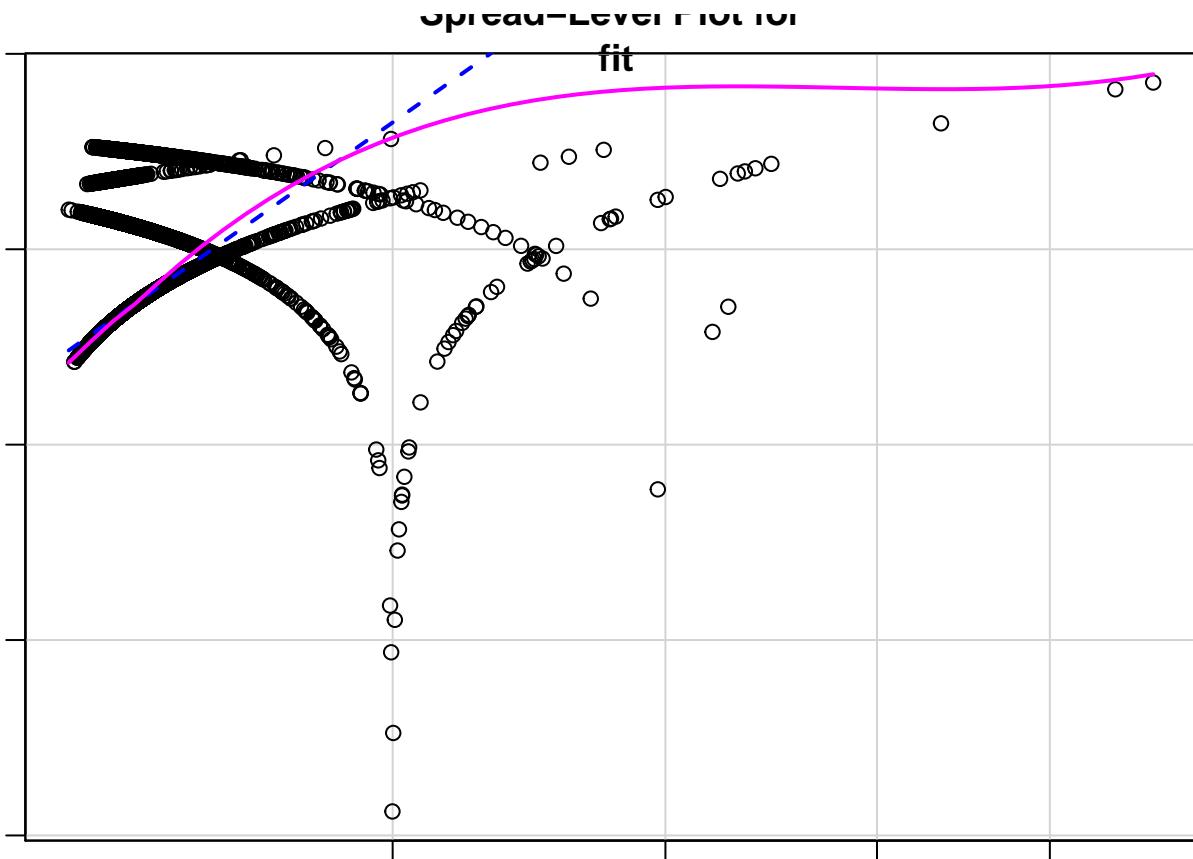
Distribution of Studentized Residuals



```
#Non-constant Error Variance
# Evaluate homoscedasticity
# non-constant error variance test
ncvTest(fit)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1688.63, Df = 1, p = < 2.22e-16
# plot studentized residuals vs. fitted values
spreadLevelPlot(fit)

## Warning in rlm.default(x, y, weights, method = method, wt.method = wt.method, :
## 'rlm' failed to converge in 20 steps
```



```

##  

## Suggested power transformation: -6.861469  

#Multi-collinearity  

# Evaluate Collinearity  

vif(fit) # variance inflation factors

##      dist      temp windchill  humidity
## 1.001636 82.880956 81.522254  1.246348

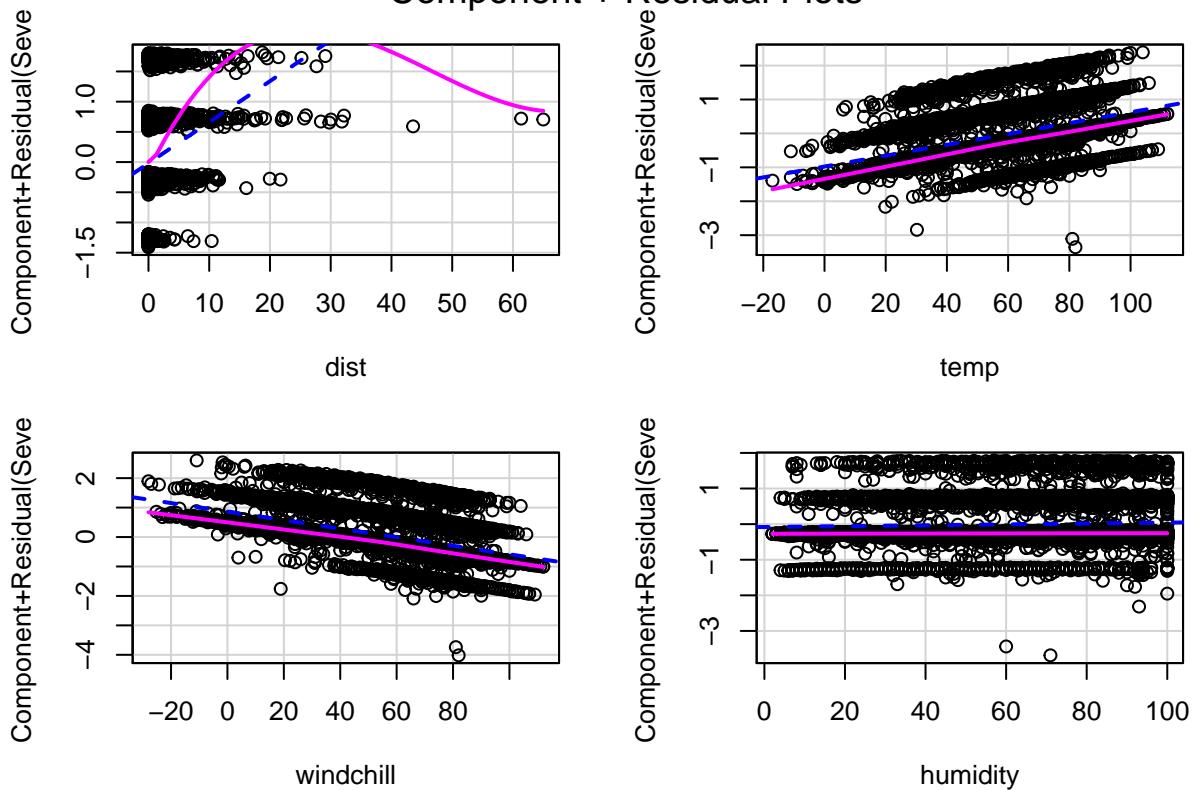
sqrt(vif(fit)) > 2 # problem?

##      dist      temp windchill  humidity
## FALSE    TRUE     TRUE    FALSE

#Nonlinearity
# component + residual plot
crPlots(fit)

```

Component + Residual Plots



```

# Ceres plots
#ceresPlots(fit)

#Non-independence of Errors
# Test for Autocorrelated Errors
durbinWatsonTest(fit)

## lag Autocorrelation D-W Statistic p-value
## 1 -0.01955319 2.039079 0.006
## Alternative hypothesis: rho != 0

# Global test of model assumptions
library(gvlma)
gvmmodel <- gvlma(fit)
summary(gvmmodel)

## 
## Call:
## lm(formula = Severity ~ dist + temp + windchill + humidity, data = num)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -3.6912 -0.2745 -0.2413  0.5915  1.8145 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2.0659506  0.0292263 70.688 < 2e-16 ***
## dist        0.0679376  0.0025946 26.184 < 2e-16 ***
## temp        0.0160049  0.0019505  8.205 2.45e-16 ***

```

```

## windchill   -0.0145144  0.0017157  -8.460  < 2e-16 ***
## humidity      0.0012356  0.0001955    6.319  2.70e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5449 on 18245 degrees of freedom
## Multiple R-squared:  0.04329,   Adjusted R-squared:  0.04308
## F-statistic: 206.4 on 4 and 18245 DF,  p-value: < 2.2e-16
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
## gvlma(x = fit)
##
##          Value p-value           Decision
## Global Stat      5090.7088  0.000 Assumptions NOT satisfied!
## Skewness         3260.7930  0.000 Assumptions NOT satisfied!
## Kurtosis        1512.2197  0.000 Assumptions NOT satisfied!
## Link Function    317.1380  0.000 Assumptions NOT satisfied!
## Heteroscedasticity  0.5581  0.455   Assumptions acceptable.

fit

##
## Call:
## lm(formula = Severity ~ dist + temp + windchill + humidity, data = num)
##
## Coefficients:
## (Intercept)       dist        temp     windchill     humidity
## 2.065951      0.067938     0.016005   -0.014514      0.001236

summary(fit)

##
## Call:
## lm(formula = Severity ~ dist + temp + windchill + humidity, data = num)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.6912 -0.2745 -0.2413  0.5915  1.8145
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.0659506  0.0292263 70.688  < 2e-16 ***
## dist        0.0679376  0.0025946 26.184  < 2e-16 ***
## temp        0.0160049  0.0019505  8.205 2.45e-16 ***
## windchill   -0.0145144  0.0017157 -8.460  < 2e-16 ***
## humidity     0.0012356  0.0001955  6.319  2.70e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5449 on 18245 degrees of freedom
## Multiple R-squared:  0.04329,   Adjusted R-squared:  0.04308

```

```

## F-statistic: 206.4 on 4 and 18245 DF, p-value: < 2.2e-16
fit1 <- fit
fit2 <- lm(Severity ~ dist+temp+windchill, data = num)

# compare models
anova(fit1, fit2)

## Analysis of Variance Table
##
## Model 1: Severity ~ dist + temp + windchill + humidity
## Model 2: Severity ~ dist + temp + windchill
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1 18245 5416.4
## 2 18246 5428.3 -1   -11.853 39.926 2.699e-10 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
step <- stepAIC(fit, direction="both")

## Start: AIC=-22158.79
## Severity ~ dist + temp + windchill + humidity
##
##           Df Sum of Sq   RSS   AIC
## <none>          5416.4 -22159
## - humidity     1    11.853 5428.3 -22121
## - temp         1    19.988 5436.4 -22094
## - windchill    1    21.247 5437.7 -22089
## - dist         1   203.536 5620.0 -21488
step$anova # display results

## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Severity ~ dist + temp + windchill + humidity
##
## Final Model:
## Severity ~ dist + temp + windchill + humidity
##
##           Step Df Deviance Resid. Df Resid. Dev      AIC
## 1                 18245    5416.429 -22158.79
library(leaps)
leaps<-regsubsets(Severity~dist+temp+windchill+humidity+precip,data=num,nbest=10)

# view results
summary(leaps)

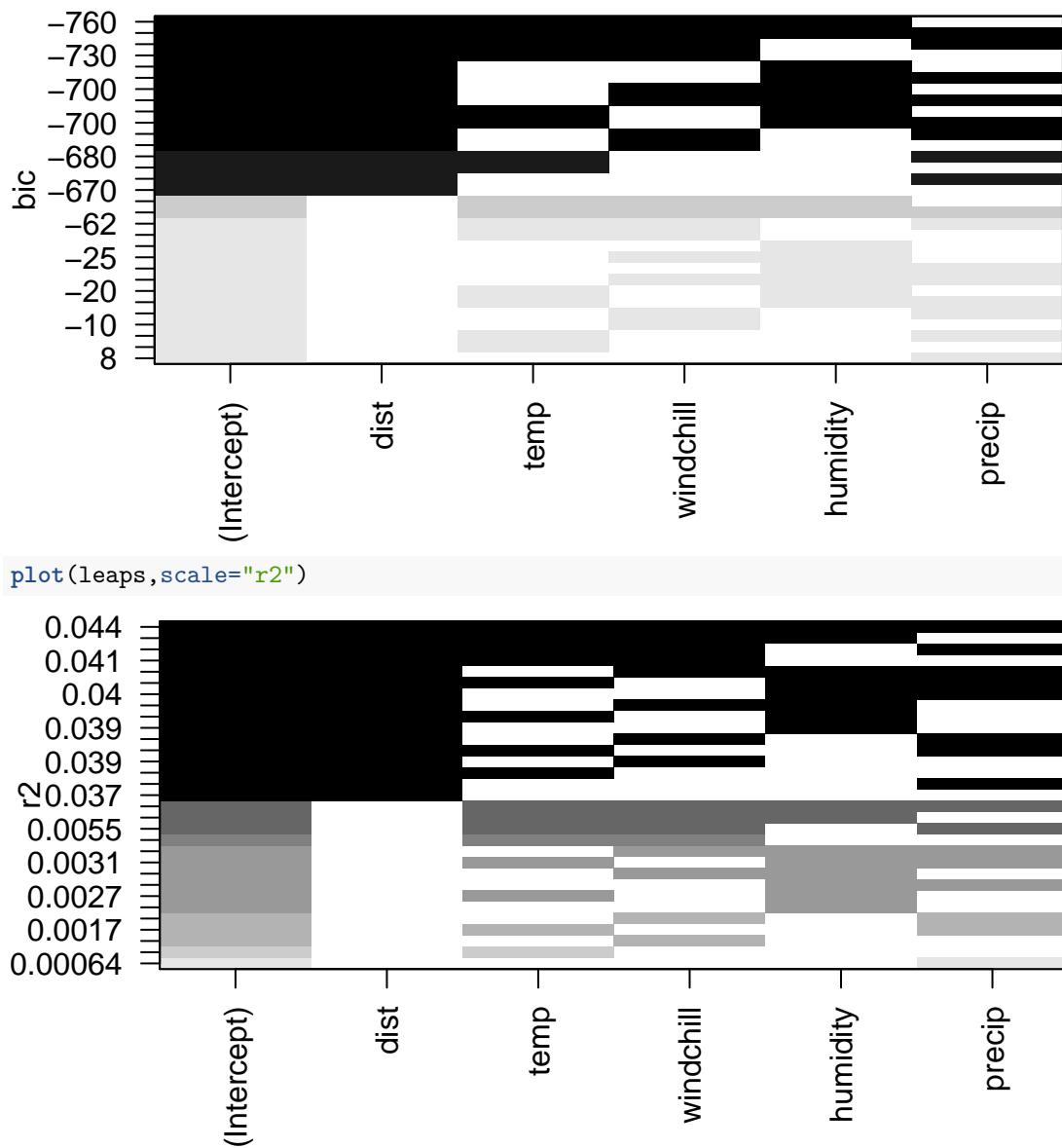
## Subset selection object
## Call: regsubsets.formula(Severity ~ dist + temp + windchill + humidity +
##   precip, data = num, nbest = 10)
## 5 Variables (and intercept)
##   Forced in Forced out
## dist      FALSE      FALSE

```

```

## temp      FALSE    FALSE
## windchill FALSE    FALSE
## humidity FALSE    FALSE
## precip   FALSE    FALSE
## 10 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          dist temp windchill humidity precip
## 1 ( 1 ) "*" " " " " " "
## 1 ( 2 ) " " " " " " *
## 1 ( 3 ) " " " " *" " " " "
## 1 ( 4 ) " " " *" " " " " "
## 1 ( 5 ) " " " " " " " "
## 2 ( 1 ) "*" " " " " " "
## 2 ( 2 ) "*" " " *" " " " "
## 2 ( 3 ) "*" " *" " " " "
## 2 ( 4 ) "*" " " " " " "
## 2 ( 5 ) " " " *" " *" " "
## 2 ( 6 ) " " " " " *" " "
## 2 ( 7 ) " " " " " " " "
## 2 ( 8 ) " " " *" " " " "
## 2 ( 9 ) " " " " " *" " "
## 2 ( 10 ) " " " *" " " " "
## 3 ( 1 ) "*" " *" " *" " "
## 3 ( 2 ) "*" " " " " " "
## 3 ( 3 ) "*" " " " *" " "
## 3 ( 4 ) "*" " *" " " " "
## 3 ( 5 ) "*" " " " *" " "
## 3 ( 6 ) "*" " *" " " " "
## 3 ( 7 ) " " " *" " *" " "
## 3 ( 8 ) " " " *" " *" " "
## 3 ( 9 ) " " " " " *" " "
## 3 ( 10 ) " " " *" " " " "
## 4 ( 1 ) "*" " *" " *" " "
## 4 ( 2 ) "*" " *" " *" " "
## 4 ( 3 ) "*" " " " *" " "
## 4 ( 4 ) "*" " *" " " " "
## 4 ( 5 ) " " " *" " *" " "
## 5 ( 1 ) "*" " *" " *" " "
# plot a table of models showing variables in each model.
# models are ordered by the selection statistic.
plot(leaps)

```



```

plot(leaps,scale="r2")

# subsets(leaps, statistic="rsq")

# All Subsets Regression
#plot(leaps,scale="bic")
#summary(leaps)
#?regsubsets
#summary(leaps)
#View(leaps)
#leaps
#coef(leaps,1:5)

```