Regression

Shiyou Yan

2023-02-12

Linear regression is a statistical model and analysis technique used to predict the value of one variable based on the value of another variable. It is based on supervised learning and finds the best or optimal relationship between the explanatory variables and the dependent variable. First, linear regression analyzes the correlation between the variables in the data set, then it fits a line to the data points. This line shows the trend in the data and allows users or machine to make more accurate predictions in the future. Linear regression has several strengths, including its simplicity, ease of understanding, and speed. It can quickly provide results or predictions and is easily understandable to users.The weakness of linear regression is unable to perform the results of the non-linear relationships, not flexible.

#Read data set

```
data <- read.csv ("https://archive.ics.uci.edu/ml/machine-learning-databases/00381/PRSA_data_2010.
1.1-2014.12.31.csv")
```

1.Divide into 80/20 train/test

```
data <- read.csv ("https://archive.ics.uci.edu/ml/machine-learning-databases/00381/PRSA_data_2010.
1.1-2014.12.31.csv")
set.seed(1234)

sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE, prob=c(0.8,0.2))
train  <- data[sample, ]
test   <- data[!sample, ]
```

2.Use at least 5 R functions for data exploration, using the training data

```
data <- read.csv ("https://archive.ics.uci.edu/ml/machine-learning-databases/00381/PRSA_data_2010.
1.1-2014.12.31.csv")
set.seed(1234)
sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE, prob=c(0.8,0.2))
train  <- data[sample, ]
test   <- data[!sample, ]
dim(train)
```

```
## [1] 35120    13
```

```
str(train)
```

```
## 'data.frame':    35120 obs. of  13 variables:
##  $ No   : int  1 2 3 4 6 7 8 9 10 11 ...
##  $ year : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ month: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ day  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ hour : int  0 1 2 3 5 6 7 8 9 10 ...
##  $ pm2.5: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ DEWP : int  -21 -21 -21 -21 -19 -19 -19 -19 -20 -19 ...
##  $ TEMP : num  -11 -12 -11 -14 -10 -9 -9 -9 -8 -7 ...
##  $ PRES : num  1021 1020 1019 1019 1017 ...
##  $ cbwd : chr  "NW" "NW" "NW" "NW" ...
##  $ Iws  : num  1.79 4.92 6.71 9.84 16.1 ...
##  $ Is   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Ir   : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
head(train,5)
```

	No	**year**	**month**	**day**	**hour**	**pm2.5**	**DEWP**	**TEMP**	**PRES**
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<dbl>	<dbl>
1	1	2010	1	1	0	*NA*	-21	-11	1021
2	2	2010	1	1	1	*NA*	-21	-12	1020
3	3	2010	1	1	2	*NA*	-21	-11	1019
4	4	2010	1	1	3	*NA*	-21	-14	1019
6	6	2010	1	1	5	*NA*	-19	-10	1017

5 rows | 1-10 of 14 columns

```
tail(train,5)
```

	No	**year**	**month**	**day**	**hour**	**pm2.5**	**DEWP**	**TEMP**	**PRES**
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<dbl>	<dbl>
43820	43820	2014	12	31	19	8	-23	-2	1034
43821	43821	2014	12	31	20	10	-22	-3	1034
43822	43822	2014	12	31	21	10	-22	-3	1034
43823	43823	2014	12	31	22	8	-22	-4	1034
43824	43824	2014	12	31	23	12	-21	-3	1034

5 rows | 1-10 of 14 columns

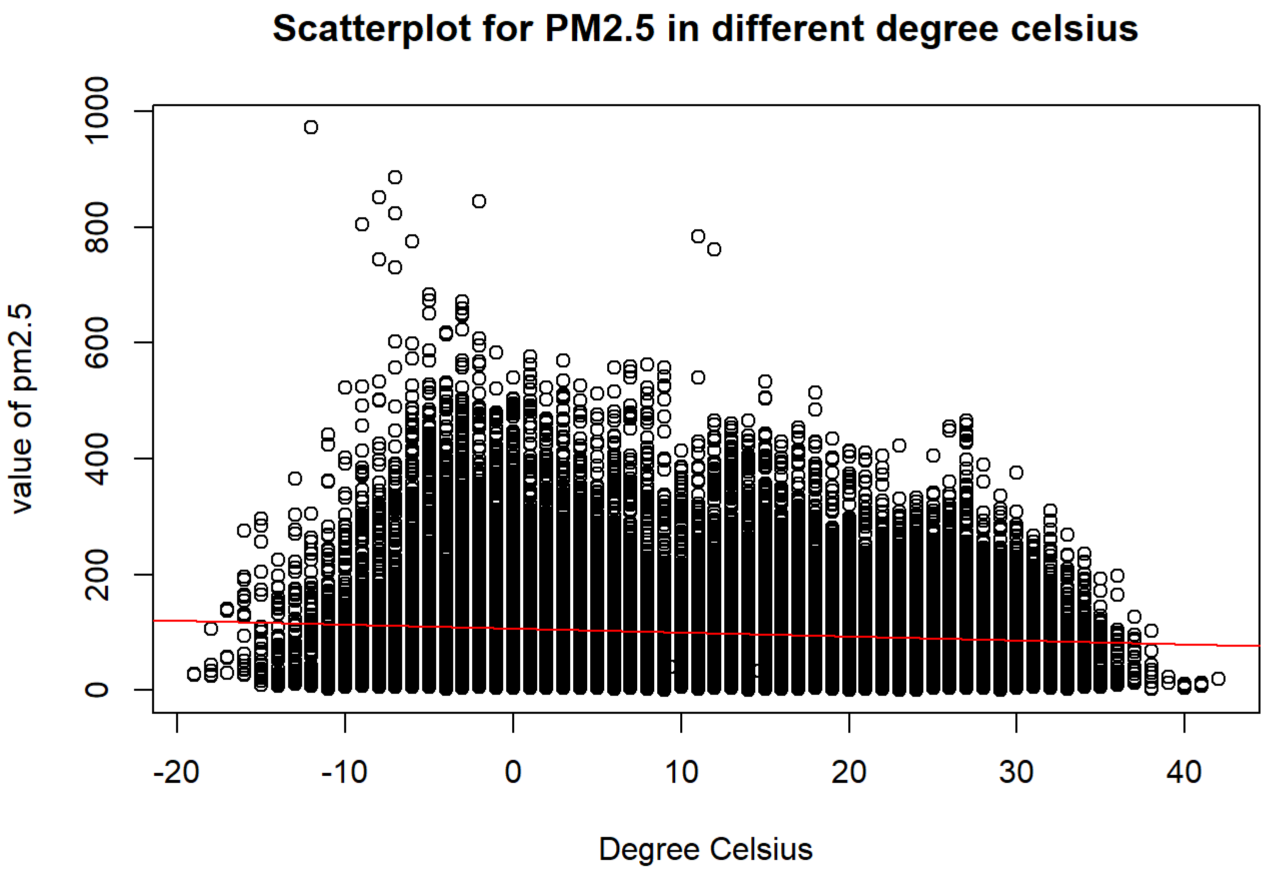
```
summary(train[6])
```

```
##      pm2.5
##  Min.   :  0.00
##  1st Qu.: 29.00
##  Median : 72.00
##  Mean   : 98.44
##  3rd Qu.:137.00
##  Max.   :972.00
##  NA's   :1672
```

3.Create at least 2 informative graphs, using the training data a. This is scatter plot for pm2.5 and temperature degree

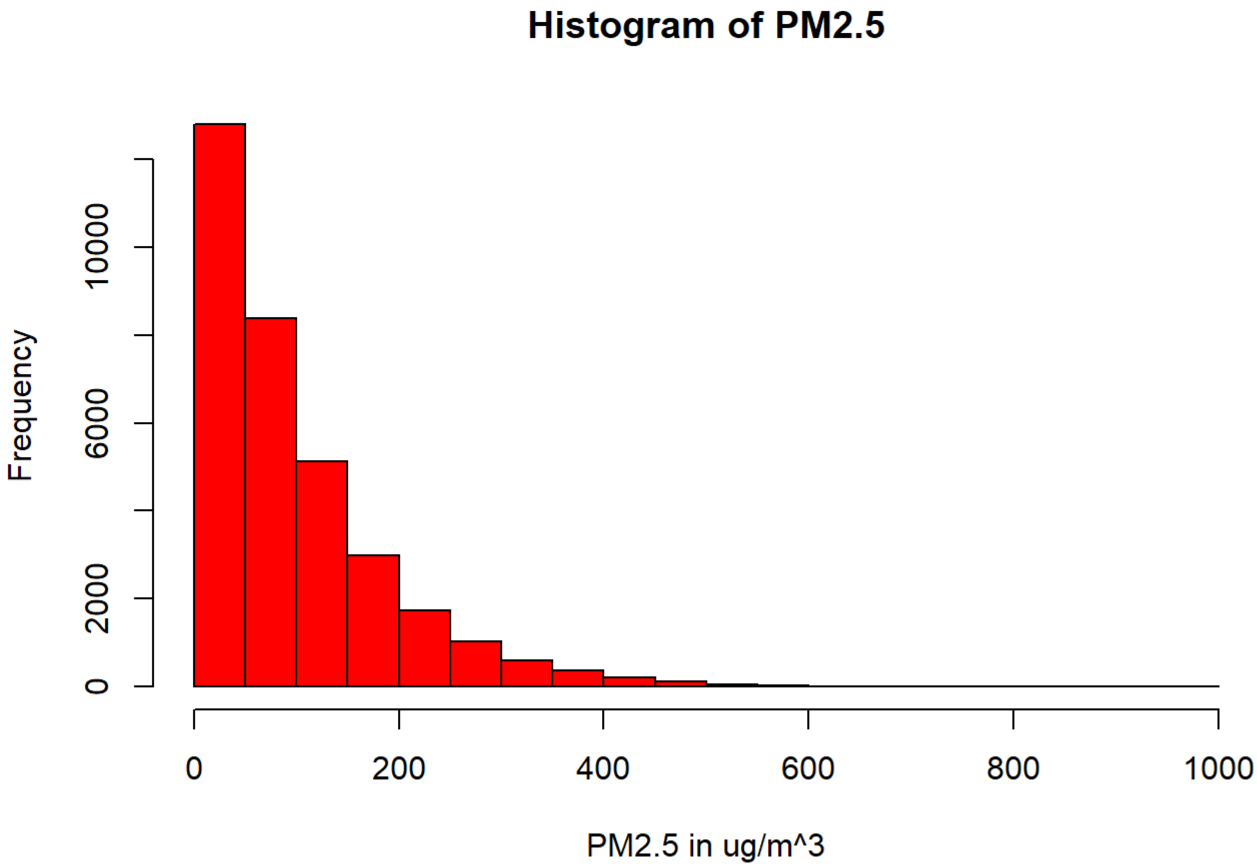
```
data <- read.csv ("https://archive.ics.uci.edu/ml/machine-learning-databases/00381/PRSA_data_2010.1.1-2014.12.31.csv")
set.seed(1234)

sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE, prob=c(0.8,0.2))
train  <- data[sample, ]
test   <- data[!sample, ]
plot(x = train$TEMP , y = train$pm2.5, xlab="Degree Celsius", ylab="value of pm2.5",
     main="Scatterplot for PM2.5 in different degree celsius")
abline(lm(pm2.5 ~ TEMP,data = train),col='red')
```



b. This is histogram for pm2.5

```
data <- read.csv ("https://archive.ics.uci.edu/ml/machine-learning-databases/00381/PRSA_data_2010.
1.1-2014.12.31.csv")
set.seed(1234)
sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE, prob=c(0.8,0.2))
train  <- data[sample, ]
test   <- data[!sample, ]
hist(train$pm2.5, main="Histogram of PM2.5", xlab="PM2.5 in ug/m^3", xlim=c(0,1000), col="red")
```



4.Build a simple linear regression model one predictor and output the summary. Write a thorough explanation of the information in the model summary.

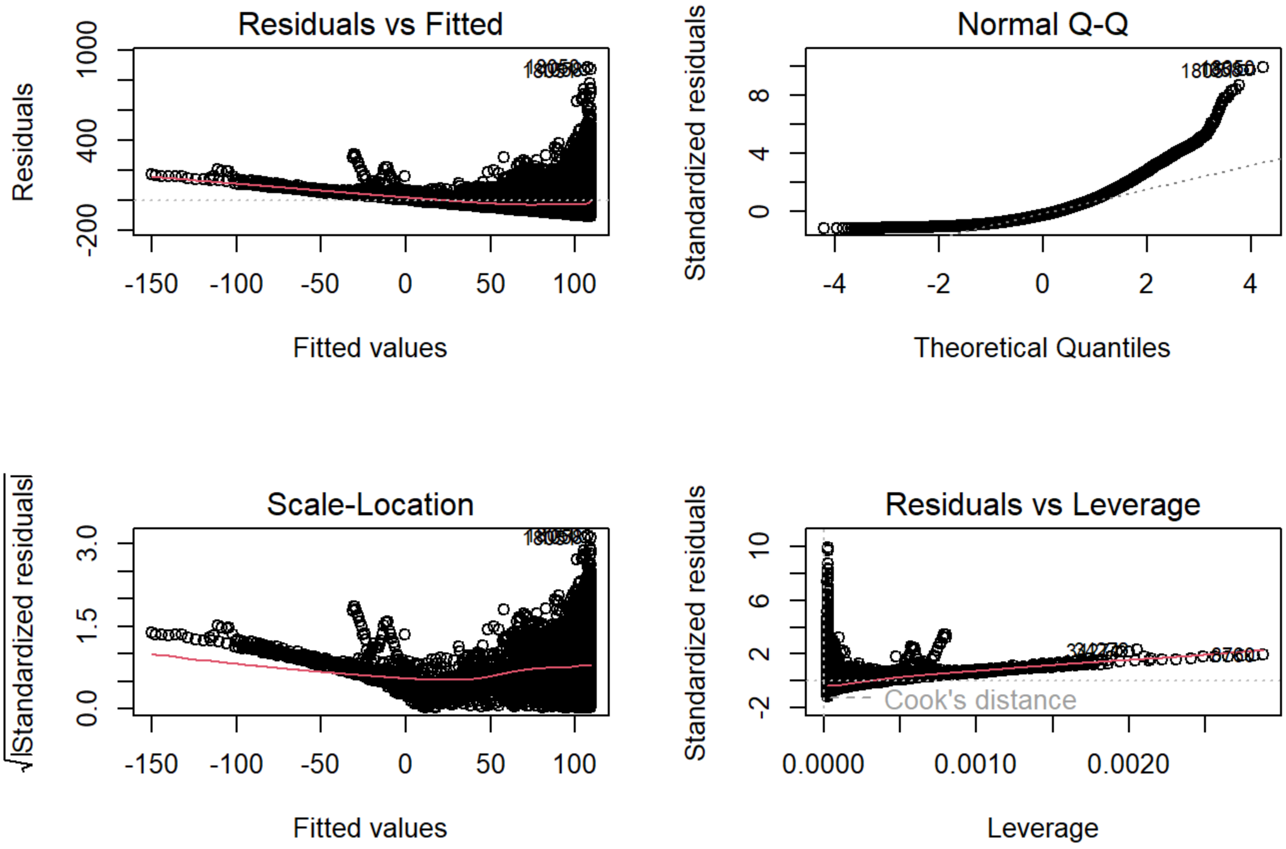
```
data <- read.csv ("https://archive.ics.uci.edu/ml/machine-learning-databases/00381/PRSA_data_2010.
1.1-2014.12.31.csv")
res <- lm(data$pm2.5 ~ data$Iws, data)
summary(res)
```

```
##
## Call:
## lm(formula = data$pm2.5 ~ data$Iws, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -107.15  -63.53  -24.54   36.62  886.68
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 109.584514   0.484288  226.28   <2e-16 ***
## data$Iws     -0.459690   0.008796  -52.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 89.18 on 41755 degrees of freedom
##   (因为不存在，2067个观察量被删除了)
## Multiple R-squared:  0.0614, Adjusted R-squared:  0.06137
## F-statistic:  2731 on 1 and 41755 DF,  p-value: < 2.2e-16
```

a. The call section shows pm2.5 is our dependent variable and Iws is our predictor.
b. The residuals section shows the distribution is not symmetrical and right-skewed.
c. The coefficients section means accumulated wind speed more fast, the value of pm2.5 less. y = -0.459690x+109.584514. Std.Error is 0.008796.
d. Residual standard error shows us that if our predictions be off on average by $89.18 won’t give accurate result.
e. F value larger, p value smaller. It means that they are strong relationship.

5.Plot the residuals and write a thorough explanation of what the residual plot tells you.

```
data <- read.csv ("https://archive.ics.uci.edu/ml/machine-learning-databases/00381/PRSA_data_2010.
1.1-2014.12.31.csv")
fit = lm(data$pm2.5 ~ data$Iws, data)
par(mfrow=c(2,2))
plot(fit)
```



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

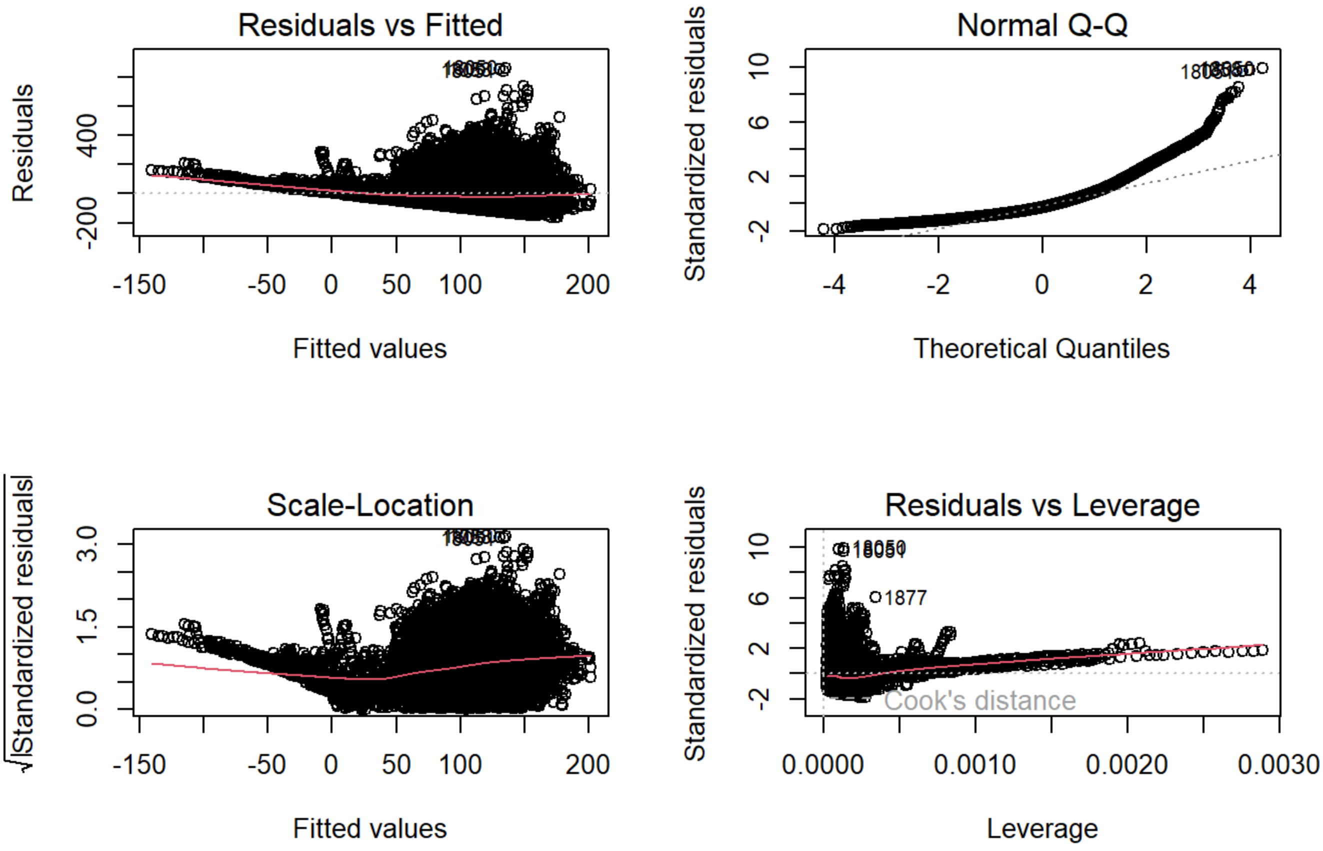
Explanation: The first picture (Residuals vs Fitted) shows that it doesn’t have non-linear relationship between the values of fitted 0 - 100.It is a good model during that range. But in -50 to -150, residuals increase as the fitted values decrease. It has non-linear relationship at that time. The second picture– Normal Q-Q means that residuals are not normally distributed The third picture–Scale-Location shows that the residuals almost appear randomly spread. The fourth picture–Residuals vs Leverage means that there are not any influential points.

6.Build a multiple linear regression model (multiple predictors), output the summary and residual plots

```
data <- read.csv ("https://archive.ics.uci.edu/ml/machine-learning-databases/00381/PRSA_data_2010.
1.1-2014.12.31.csv")
model <- lm(data$pm2.5 ~ data$Iws + data$TEMP + data$PRES, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = data$pm2.5 ~ data$Iws + data$TEMP + data$PRES, data = data)
##
## Residuals:
##      Min      1Q  Median      3Q     Max
## -162.04  -59.54  -22.19   36.70  858.48
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.264e+03  7.535e+01   43.31   <2e-16 ***
## data$Iws    -4.600e-01  8.682e-03  -52.99   <2e-16 ***
## data$TEMP   -3.109e+00  6.190e-02  -50.23   <2e-16 ***
## data$PRES   -3.065e+00  7.353e-02  -41.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 86.6 on 41753 degrees of freedom
##   (因为不存在，2067个观察量被删除了)
## Multiple R-squared:  0.1149, Adjusted R-squared:  0.1148
## F-statistic:  1807 on 3 and 41753 DF,  p-value: < 2.2e-16
```

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



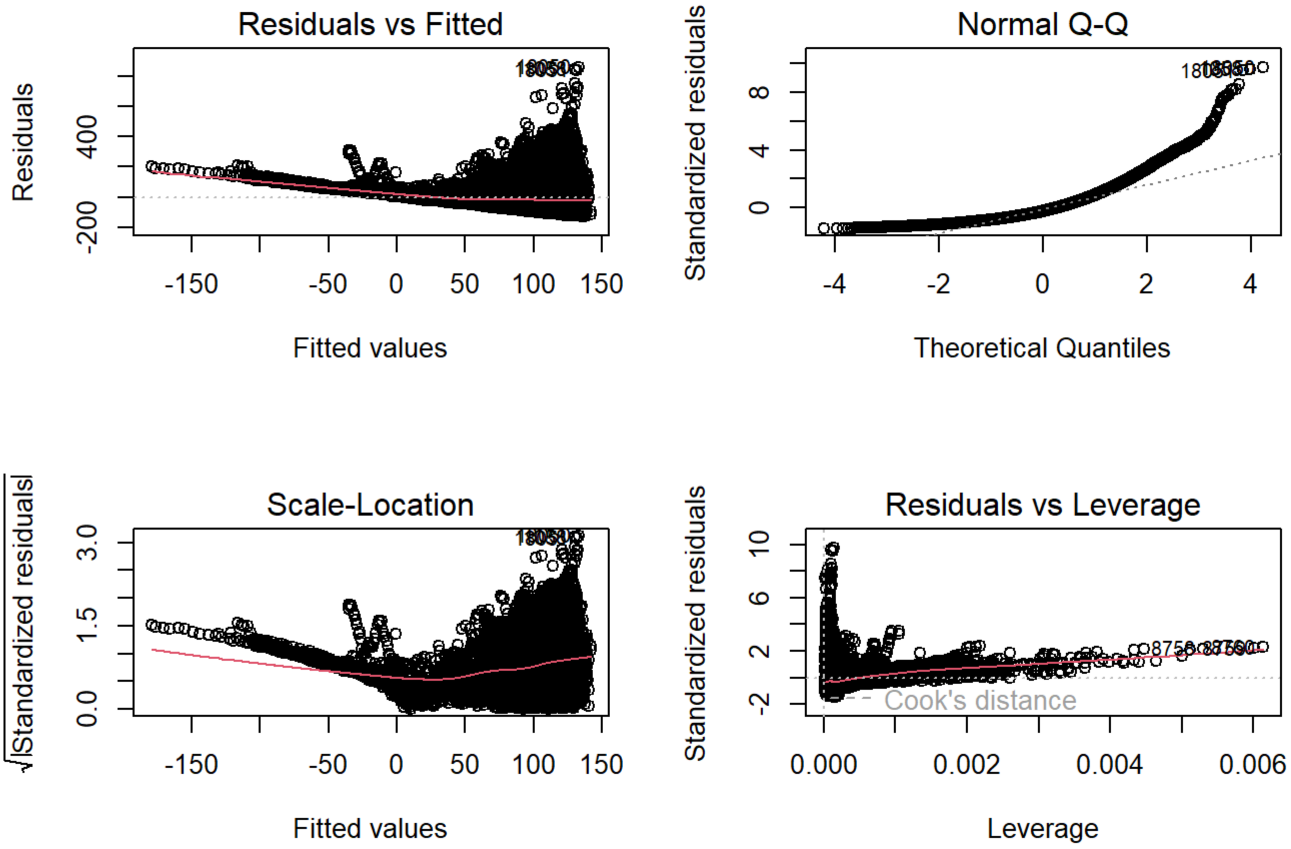
```
par(mfrow=c(1,1))
```

7.Build a third linear regression model using a different combination of predictors, interaction effects, polynomial regression, or any combination to try to improve the results. Output the summary and residual plots.

```
data <- read.csv ("https://archive.ics.uci.edu/ml/machine-learning-databases/00381/PRSA_data_2010.
1.1-2014.12.31.csv")
model <- lm(data$pm2.5 ~ data$Iws * data$TEMP, data = data)
summary(model)
```

```
## 
## Call:
## lm(formula = data$pm2.5 ~ data$Iws * data$TEMP, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -130.21  -62.73  -23.10   38.37  860.44 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         1.234e+02  6.866e-01 179.658  < 2e-16 ***
## data$Iws           -5.146e-01  9.409e-03 -54.693  < 2e-16 ***
## data$TEMP          -1.086e+00  3.993e-02 -27.195  < 2e-16 ***
## data$Iws:data$TEMP  4.902e-03  8.636e-04   5.676 1.39e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 88.35 on 41753 degrees of freedom
##   (因为不存在，2067个观察量被删除了)
## Multiple R-squared:  0.07876,    Adjusted R-squared:  0.0787 
## F-statistic:  1190 on 3 and 41753 DF,  p-value: < 2.2e-16
```

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



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

8.Write a paragraph or more comparing the results. Indicate which model is better and why you think that is the case. The results of these three modeling are most same. They have similar residuals and curves. I think third linear regression model is better for this data set because the value of pm2.5 is not supported by only one predictor variables, but it also supported by other predictor variables, including temperature, wind speed, and so on. The third linear regression model can be more appropriate for the relationship between the predictors and the response is non-linear.

9.Using your 3 models, predict and evaluate on the test data using metrics correlation and MSE. Compare the results and indicate why you think these results happened.

```
data <- read.csv ("https://archive.ics.uci.edu/ml/machine-learning-databases/00381/PRSA_data_2010.
1.1-2014.12.31.csv")
res1 <- lm(data$pm2.5 ~ data$Iws, data)
mean(res1$residuals^2)
```

```
## [1] 7952.849
```

```
res2 <- lm(data$pm2.5 ~ data$Iws + data$TEMP + data$PRES, data = data)
mean(res2$residuals^2)
```

```
## [1] 7499.582
```

```
res3 <- lm(data$pm2.5 ~ data$Iws * data$TEMP, data = data)
mean(res3$residuals^2)
```

```
## [1] 7805.697
```

The results are almost same, just have a little deviation because using different predictor variables or combinations of predictor variables.