PSTAT 126 Project

How to Choose an Appropriate Model Selection Method

Inwoong Bae, Sunnie Oh
2019-6-9

Abstract

We will figure out how to select the best model in our interest. This process could include transforming variables and selecting the sort of variables itself. In the first part, we will try to set up a fixed model and transform variables in the fixed model and figure out if it is really meaningful. On the other hand, in the second part, we will construct the model by ourselves using two different methods and compare the results of each method from empty model or full model. We conclude that constructing the model by ourselves is more efficient to build the proper model.

Problem and Motivation

In data analysis process, model selection is one of the basic processes before we analyze data and conclude the results. Even though we select valid dataset and clean it up properly, the result could be totally different or even opposite from our expected result because of choosing wrong method and inappropriate transformations. So, we wonder how we are able to figure out which model would be the best to make a right conclusion. This question leads us to conduct two types of study about two different multiple linear model selection using several methods. The first study is to set up several abstract models, add or remove variables, and transform the variables. On the other hand, the second study is to run multiple model selecting method, which add or remove and order variables from the empty model or full model. There are two types of our dataset that we are going to use for the model selection: The first dataset is the market historical dataset of real estate valuation that were collected from Sindian District, New Taipei City, Taiwan during the period of June 2012 to May 2013. The second dataset is the dataset from the paper, Modeling of strength of high performance concrete using artificial neural networks, written by Dr. I-Cheng Yeh.

Data

We use two types of datasets for this project. The first dataset is about real estate valuation dataset from Sindian District, New Taipei City, Taiwan during the period of June 2012 to May 2013. This dataset consists of 8 variables:
TDate(the transaction date (e.g., 2013.250=2013 March, 2013.500=2013 June, etc.),
Age(the house age (unit: year)),
Metro(the distance to the nearest MRT station (unit: meter)),
Stores(the number of convenience stores in the living circle on foot (integer)),
Latitude(the geographic coordinate (unit: degree)),
Longitude (the geographic coordinate (unit: degree))
Price(the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local 7 unit, 1 Ping = 3.3 meter squared)).
We are going to use this dataset to find the best model by selecting variables and transforming them in the first part.
The second dataset was collected for the paper, Modeling of strength of high performance concrete using artificial neural networks, by Dr. I-Cheng Yeh. We will use this data set to

model the relationship between the concrete compressive strength and the 8 concrete components:
X1(Cement (component 1, unit kg=m3)),
X2(Blast Furnace Slag (component 2, unit kg=m3)),
X3(Fly Ash (component 3, unit kg=m3)),
X4(Water (component 4, unit kg=m3)),
X5(Superplasticizer (component 5, unit kg=m3)),
X6(Coarse Aggregate (component 6, unit kg=m3)),
X7(Fine Aggregate (component 7, unit kg=m3)),
X8 Age (Day, 1 365)
Y(Concrete compressive strength (MPa)).
We will use this dataset to find out the best model by multiple of variable selection methods.

Questions of Interest

In the first part, we think about which factors make price of real estate and select several factors. we set up a model with some variables that we think they are affecting the price of . After we build the model, we investigate the relationship between price and other factors such as location(latitude, longitude), distance from the nearest subway station, age of the house. Once we find the relationship, we figure out how much each variable affects the price by conducting some tests. Lastly, we transform the variables to get more clear relationship by using several methods.
In the second part, we select the most adequate model which identifies the relationship between concrete compressive strength and the 8 concrete components form X1 to X8. In this process, we will use two different methods. One is adding a variable continuously starting from the smallest model, and the other is deducting a variable continuously starting from the biggest model. After selecting which variables to contain in the model, we detect whether the selected model is really adequate model or not. In this process, we can remove some points. Next, we compare two different models and figure out if they are different from each other and the reason of the difference between each other.

Regression Methods

In the first part, we plot price and other predictors(TDate, Age, Stores, Latitude) and then see the relationship between price and the variables briefly. After that, we set up the linear model between price and other predictors and conduct global f-test and plots such as Residual and Fitted value and Normal Q-Q plot to know how much each predictor would be significant for the model. We then build other possible model and compare it with the previous model by ANOVA table. Once we find the possibly best predictors for the response, price, then we try to transform the predictors to get more clear and ideal results using powerTransformation, Box-Cox method and polynomial regression.

In the second part, we will use two different variable selection methods, forward selection and backward elimination, to choose variables in the model. After fixing the model, we will check if it fulfills the linear regression assumptions by residuals vs. fitted values plot to

check both non-linearity and non-constant variance, and also Normal Q-Q plot to check non-normality. And then, by using influential index plot, we will find any influential points to remove, particularly using leverages to detect data points with large influence. Finally, we will figure out if two models derived from two different selection methods are different from each other using ANOVA.

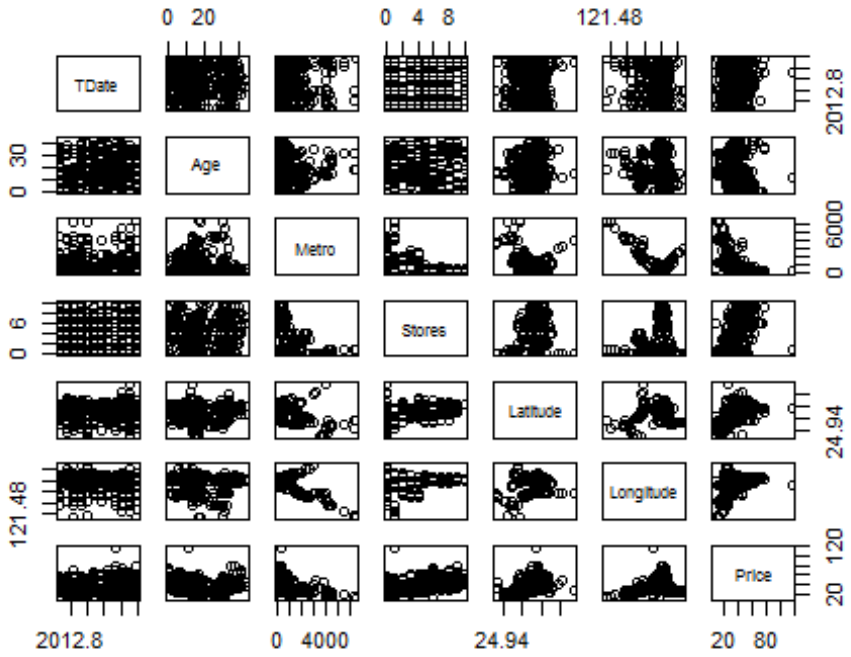
Regression Analysis, Results and Interpretation

part 1

```
MyData <- read.table("RealEstateValuation.txt", header = TRUE)
head(MyData)
```

```
##      TDate  Age      Metro Stores Latitude Longitude Price
## 1 2012.917 32.0   84.87882     10 24.98298  121.5402  37.9
## 2 2012.917 19.5  306.59470      9 24.98034  121.5395  42.2
## 3 2013.583 13.3  561.98450      5 24.98746  121.5439  47.3
## 4 2013.500 13.3  561.98450      5 24.98746  121.5439  54.8
## 5 2012.833  5.0  390.56840      5 24.97937  121.5425  43.1
## 6 2012.667  7.1 2175.03000      3 24.96305  121.5125  32.1
```

```
plot(MyData)
```




```
#According to the plot among variables in the dataset from Real Estate Valuat
ion, there seems no significant relationship between price and each factor.
```

```
mod <- lm(Price ~ TDate + Age + Stores + Latitude, data = MyData)
summary(mod)
```

```
##
## Call:
## lm(formula = Price ~ TDate + Age + Stores + Latitude, data = MyData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.620  -5.601  -0.714   4.207  80.465
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.742e+04  3.524e+03  -4.944 1.12e-06 ***
## TDate        3.613e+00  1.686e+00   2.143   0.0327 *
## Age         -3.020e-01  4.178e-02  -7.227 2.44e-12 ***
## Stores       1.929e+00  1.801e-01  10.712  < 2e-16 ***
## Latitude     4.078e+02  4.278e+01   9.534  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.654 on 409 degrees of freedom
## Multiple R-squared:  0.5015, Adjusted R-squared:  0.4966
## F-statistic: 102.8 on 4 and 409 DF,  p-value: < 2.2e-16
```

let Y be price, x1 be TDate, x2 be Age, x3 be Stores, x4 be Latitude. then the equation for the fitted regression line is Y = -1.742e+04 + 3.613e+00*x1* - *3.020e-01*x2 + 1.929e+00*x3 + 4.078e+02 + error. By summary, the r-value of each variable except TDate is lower than 0.01. This means that the all variables except TDate are significant for this model. TDate is not significant for this model, so it can be removed to make the model become better model.

```
#Suppose we add Metro or Longitude on the previous model.
mod2 = lm(Price ~ TDate + Age + Stores + Latitude + Metro, data = MyData )
mod3 = lm(Price ~ TDate + Age + Stores + Latitude + Longitude, data = MyData)

#The null hypothesis for the original model and the the model that adds Metro
 is betha of Metro equals zero and the alternative hypothesis is the betha of
 Metro is nonzero.
anova(mod,mod2) #anova table for the original model and the the model with Me
tro
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ TDate + Age + Stores + Latitude
## Model 2: Price ~ TDate + Age + Stores + Latitude + Metro
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    409 38119
## 2    408 31938  1    6181.8 78.972 < 2.2e-16 ***
```

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

#The null hypothesis for the original model and the the model that adds Longi
tude is betha of Longitude equals zero and the alternative hypothesis is the
betha of Longitude is nonzero.
anova(mod,mod3) #anova table for the original model and the the model with Lo
ngitude

## Analysis of Variance Table
##
## Model 1: Price ~ TDate + Age + Stores + Latitude
## Model 2: Price ~ TDate + Age + Stores + Latitude + Longitude
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    409 38119
## 2    408 34997  1    3122.5 36.402 3.605e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#by the anova tables, p-values of Metro and Longitude are much lower than sig
nificant level = 0.05. Therefore, both are available to be added on the origi
nal model.

#Suppose we have another possible model
modSec <- lm(Price ~ TDate + Age + Metro + Latitude, data = MyData)
summary(modSec)

##
## Call:
## lm(formula = Price ~ TDate + Age + Metro + Latitude, data = MyData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.218  -5.269  -0.700   4.433  70.502
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.767e+04  3.359e+03  -5.262 2.30e-07 ***
## TDate        5.570e+00  1.619e+00   3.440 0.000642 ***
## Age         -2.530e-01  4.001e-02  -6.323 6.71e-10 ***
## Metro       -5.764e-03  4.493e-04 -12.829  < 2e-16 ***
## Latitude     2.607e+02  4.569e+01   5.705 2.23e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.225 on 409 degrees of freedom
## Multiple R-squared:  0.5448, Adjusted R-squared:  0.5403
## F-statistic: 122.4 on 4 and 409 DF,  p-value: < 2.2e-16
```

```
#The equation for the regression line is Price = -1.767e+4 + 5.570e+00 * TDat
e - 2.530e-01 * Age - 5.764e-03 * Metro + 2.607e+02 * Latitude.
summary(mod)
```

```
##
## Call:
## lm(formula = Price ~ TDate + Age + Stores + Latitude, data = MyData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.620  -5.601  -0.714   4.207  80.465
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.742e+04  3.524e+03  -4.944 1.12e-06 ***
## TDate        3.613e+00  1.686e+00   2.143   0.0327 *
## Age         -3.020e-01  4.178e-02  -7.227 2.44e-12 ***
## Stores       1.929e+00  1.801e-01  10.712  < 2e-16 ***
## Latitude     4.078e+02  4.278e+01   9.534  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.654 on 409 degrees of freedom
## Multiple R-squared:  0.5015, Adjusted R-squared:  0.4966
## F-statistic: 102.8 on 4 and 409 DF,  p-value: < 2.2e-16
```

In both model, global p-values are same, but p-values for each individual variable are different. In the first model, p-value for TDate is relatively high and it results in being insignificant for the model by some significance levels. However, in the second model, all p-values are low enough to be significant for all significance levels. Therefore, we prefer the second model.

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.5.3
```

```
## Loading required package: carData
```

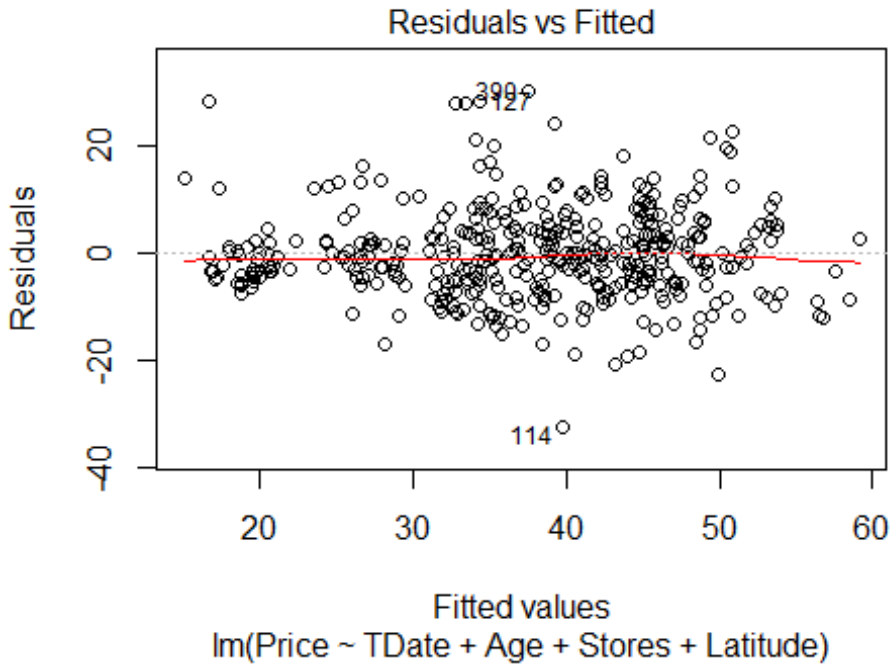
```
## Warning: package 'carData' was built under R version 3.5.2
```

```
attach(MyData)
outlierTest(modSec) #find outliers on original model
```

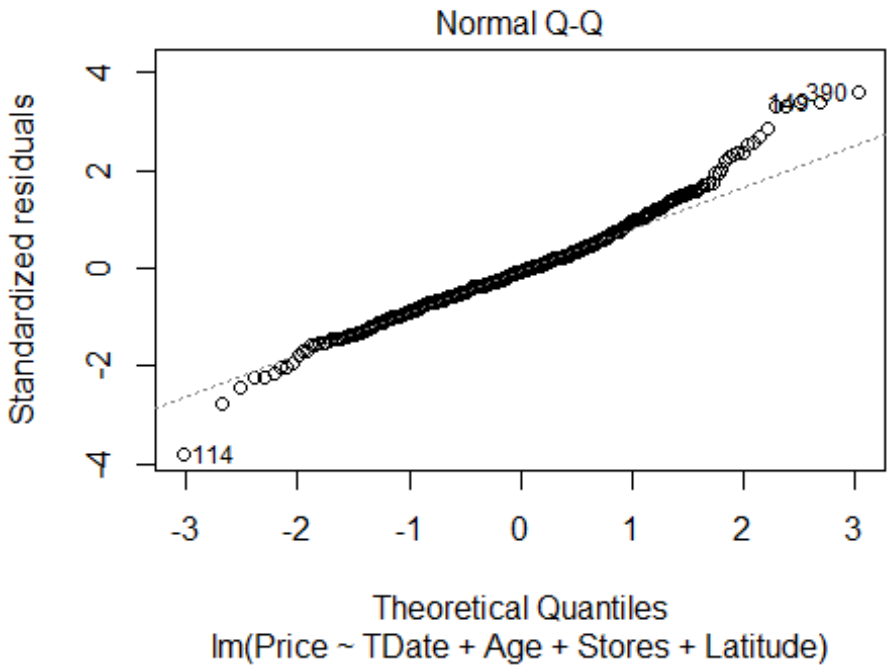
```
##     rstudent unadjusted p-value Bonferonni p
## 271 8.270876         1.8999e-15   7.8657e-13
## 313 4.147533         4.0910e-05   1.6937e-02
## 221 4.108887         4.8069e-05   1.9900e-02
```

```
newData <- MyData[c(1:220,222:270,272:312, 314:414),] #delete outliers
newmodSec <- lm(Price ~ TDate + Age + Stores + Latitude, data = newData) #new
```

```r
 dataset without outliers
plot(newmodSec, which = 1)
```



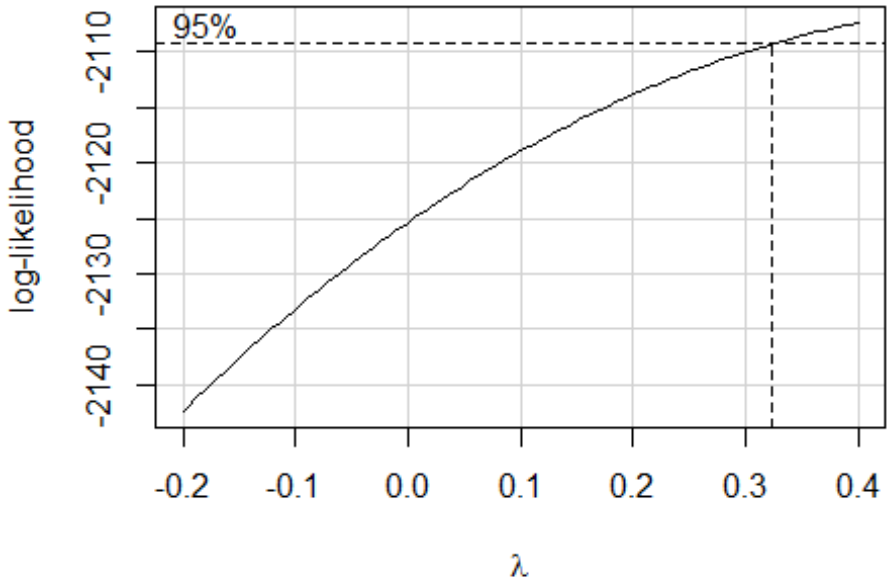

```r
plot(newmodSec, which = 2)
```




```
#When we see the Residuals vs Fitted, the line is not parallel at 0 and all p
oints in Normal Q-Q plot are not in the line
```

Using the Box-Cox method, we see that $\lambda = 0$ is both in the interval, and extremely close to the maximum, which suggests a transformation of the form log(Price)

```
boxCox(newmodSec, lambda = seq(-0.2, 0.4, by = 0.05), plotit = TRUE)
```




```
modSec_cox <- lm(log(Price) ~ TDate + Age + Stores + Latitude, data = newData)
summary(modSec_cox)
```

```
##
## Call:
## lm(formula = log(Price) ~ TDate + Age + Stores + Latitude, data = newData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57120 -0.13585  0.00979  0.14493  0.94143
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.770e+02  9.104e+01  -5.240 2.59e-07 ***
## TDate        6.565e-02  4.359e-02   1.506    0.133
## Age         -8.331e-03  1.083e-03  -7.693 1.10e-13 ***
## Stores       5.465e-02  4.679e-03  11.678  < 2e-16 ***
## Latitude     1.395e+01  1.104e+00  12.638  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2486 on 406 degrees of freedom
## Multiple R-squared:  0.588,  Adjusted R-squared:  0.5839
## F-statistic: 144.8 on 4 and 406 DF,  p-value: < 2.2e-16
```

```
#After modifying the model by Box-Cox method, we conclude that the model is n
ot needed to be changed by Box-Cox method because the modified model has insi
gnificant variable, TDate.
```

We now apply the powertransform method to the model.

```
summary(Price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.60   27.70   38.45   37.98   46.60  117.50
```

```
summary(TDate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2013    2013    2013    2013    2013    2014
```

```
summary(Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   9.025  16.100  17.713  28.150  43.800
```

```
summary(Metro)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.38  289.32  492.23 1083.89 1454.28 6488.02
```

```
summary(Latitude)
```

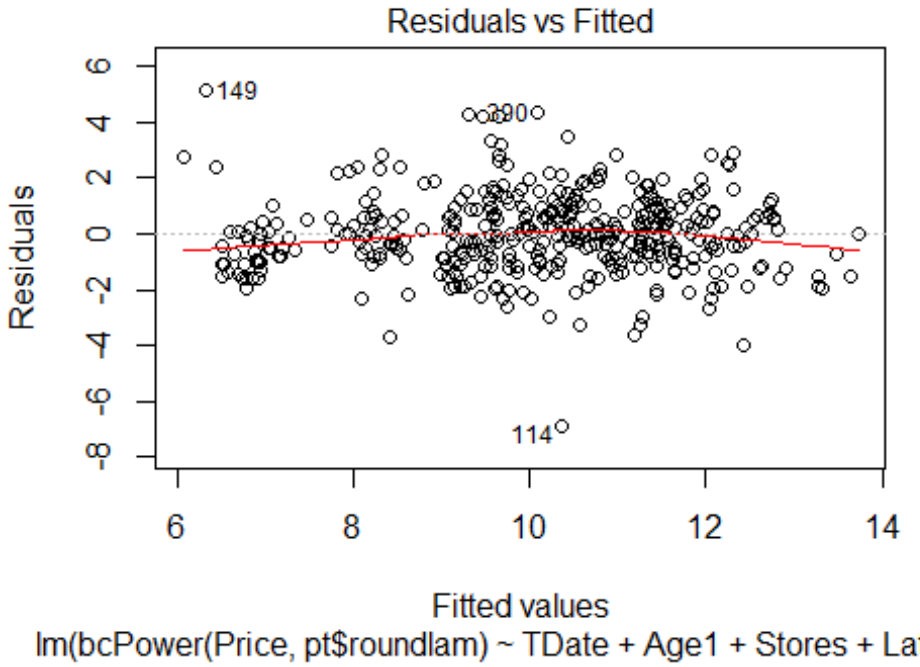
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   24.93   24.96   24.97   24.97   24.98   25.01
```

```
#Since Age has 0 value and it cannot be transformed by powerTransform. We the
refore need to add a small constant to transform
newData$Age1 <- with(newData, (Age*TDate + 1)/TDate)
pt <- powerTransform(Price ~ cbind(TDate, Age1, Stores, Latitude), newData)
summary(pt)
```

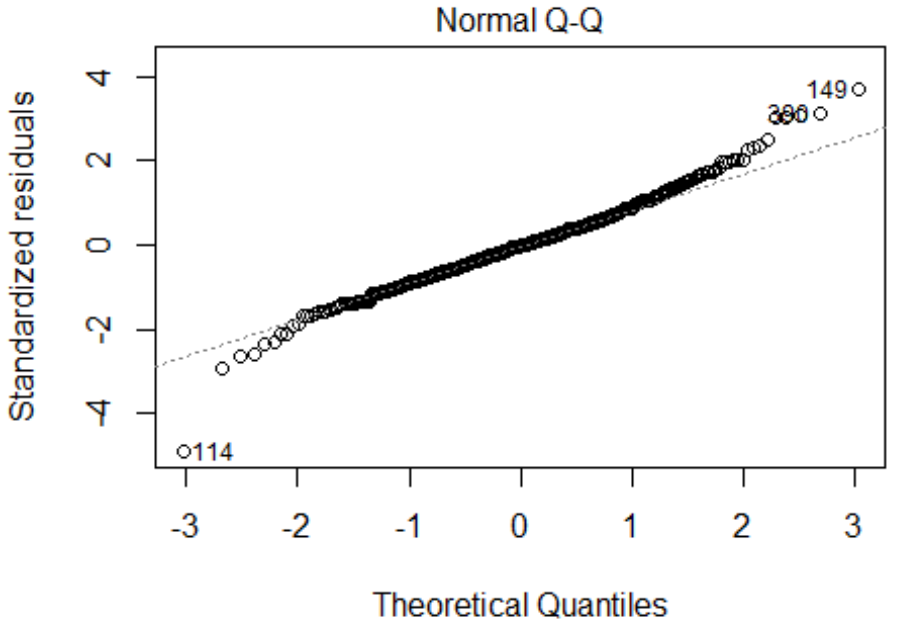
```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.5854         0.5       0.3941       0.7768
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df       pval
## LR test, lambda = (0) 39.79588  1 2.8194e-10
##
## Likelihood ratio test that no transformation is needed
##                            LRT df       pval
## LR test, lambda = (1) 16.71345  1 4.3472e-05
```

```
pt$roundlam
```

```
##  Y1
## 0.5

summary(newMod2 <- lm(bcPower(Price, pt$roundlam) ~ TDate + Age1 + Stores + L
atitude, data = newData))

##
## Call:
## lm(formula = bcPower(Price, pt$roundlam) ~ TDate + Age1 + Stores +
##     Latitude, data = newData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8564 -0.8779 -0.0252  0.7632  5.1160
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.626e+03  5.148e+02  -5.100 5.21e-07 ***
## TDate        3.983e-01  2.465e-01   1.616    0.107
## Age1        -5.040e-02  6.124e-03  -8.230 2.57e-15 ***
## Stores       3.211e-01  2.646e-02  12.135  < 2e-16 ***
## Latitude     7.343e+01  6.242e+00  11.763  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.406 on 406 degrees of freedom
## Multiple R-squared:  0.5836, Adjusted R-squared:  0.5795
## F-statistic: 142.2 on 4 and 406 DF,  p-value: < 2.2e-16

plot(newMod2, which = 1)
```




```
plot(newMod2, which = 2)
```



We apply polynomial fits

```r
quadratic.lm1 <- lm(Price ~ TDate  + Age + I(Age^2) + Stores + Latitude, data
 = newData)
summary(quadratic.lm1)
```

```
##
## Call:
## lm(formula = Price ~ TDate + Age + I(Age^2) + Stores + Latitude,
##     data = newData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.1291  -4.6118  -0.5942   4.7345  29.6598
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.629e+04  2.874e+03  -5.666 2.78e-08 ***
## TDate        3.425e+00  1.378e+00   2.485   0.0133 *
## Age         -1.367e+00  1.284e-01 -10.646  < 2e-16 ***
## I(Age^2)     2.663e-02  3.130e-03   8.506 3.51e-16 ***
## Stores       1.688e+00  1.502e-01  11.234  < 2e-16 ***
## Latitude     3.778e+02  3.488e+01  10.832  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.838 on 405 degrees of freedom
## Multiple R-squared:  0.6278, Adjusted R-squared:  0.6232
## F-statistic: 136.6 on 5 and 405 DF,  p-value: < 2.2e-16
```

```r
quadratic.lm2 <- lm(Price ~ TDate  + Age + I(Age^2) +I(Age^3)+ Stores + Latit
ude, data = newData)
summary(quadratic.lm2)
```

```
##
## Call:
## lm(formula = Price ~ TDate + Age + I(Age^2) + I(Age^3) + Stores +
##     Latitude, data = newData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.3748  -4.5603  -0.4862   4.5653  29.5234
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.607e+04  2.880e+03  -5.581 4.39e-08 ***
## TDate        3.345e+00  1.380e+00   2.424   0.0158 *
## Age         -1.099e+00  2.754e-01  -3.990 7.86e-05 ***
## I(Age^2)     8.544e-03  1.672e-02   0.511   0.6096
## I(Age^3)     3.117e-04  2.831e-04   1.101   0.2715
## Stores       1.695e+00  1.503e-01  11.275  < 2e-16 ***
```

```
## Latitude     3.758e+02  3.492e+01  10.763  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.836 on 404 degrees of freedom
## Multiple R-squared:  0.6289, Adjusted R-squared:  0.6234
## F-statistic: 114.1 on 6 and 404 DF,  p-value: < 2.2e-16
```

```
quadratic.lm3 <- lm(Price ~ TDate  + Age  + Stores + Latitude + I(Latitude^2),
 data = newData)
summary(quadratic.lm3)
```

```
##
## Call:
## lm(formula = Price ~ TDate + Age + Stores + Latitude + I(Latitude^2),
##     data = newData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.557  -5.488  -0.608   4.170  32.646
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.878e+06  1.237e+06  -2.327   0.0205 *
## TDate          2.824e+00  1.487e+00   1.898   0.0584 .
## Age           -3.120e-01  3.684e-02  -8.469 4.63e-16 ***
## Stores         1.813e+00  1.668e-01  10.870  < 2e-16 ***
## Latitude       2.297e+05  9.907e+04   2.319   0.0209 *
## I(Latitude^2) -4.592e+03  1.984e+03  -2.315   0.0211 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.454 on 405 degrees of freedom
## Multiple R-squared:  0.567,  Adjusted R-squared:  0.5617
## F-statistic: 106.1 on 5 and 405 DF,  p-value: < 2.2e-16
```

```
#All polynomial fits does not get better than original one.
```

After modifying the model by transform methods, we cannot see any significantly positive changes and improvements.There are some influencial points such as outliers in this models and the influencial points distract us when we find the appropriate model. We conclude that transformation is unnecessary to apply because none of method affects to make it clear.

Part 2

```
concrete<-read.table('Concrete.txt')
summary(concrete)
```

```
##        X1              X2              X3              X4
##  Min.   :102.0   Min.   :  0.0   Min.   :  0.00   Min.   :121.8
##  1st Qu.:192.4   1st Qu.:  0.0   1st Qu.:  0.00   1st Qu.:164.9
##  Median :272.9   Median : 22.0   Median :  0.00   Median :185.0
##  Mean   :281.2   Mean   : 73.9   Mean   : 54.19   Mean   :181.6
##  3rd Qu.:350.0   3rd Qu.:142.9   3rd Qu.:118.27   3rd Qu.:192.0
##  Max.   :540.0   Max.   :359.4   Max.   :200.10   Max.   :247.0
##        X5              X6              X7              X8
##  Min.   : 0.000   Min.   : 801.0   Min.   :594.0   Min.   :  1.00
##  1st Qu.: 0.000   1st Qu.: 932.0   1st Qu.:731.0   1st Qu.:  7.00
##  Median : 6.350   Median : 968.0   Median :779.5   Median : 28.00
##  Mean   : 6.203   Mean   : 972.9   Mean   :773.6   Mean   : 45.66
##  3rd Qu.:10.160   3rd Qu.:1029.4   3rd Qu.:824.0   3rd Qu.: 56.00
##  Max.   :32.200   Max.   :1145.0   Max.   :992.6   Max.   :365.00
##        Y
##  Min.   : 2.332
##  1st Qu.:23.707
##  Median :34.443
##  Mean   :35.818
##  3rd Qu.:46.136
##  Max.   :82.599
```

```
names(concrete)
```

```
## [1] "X1" "X2" "X3" "X4" "X5" "X6" "X7" "X8" "Y"
```

```
head(concrete)
```

```
##      X1    X2 X3  X4  X5     X6    X7  X8        Y
## 1 540.0   0.0  0 162 2.5 1040.0 676.0  28 79.98611
## 2 540.0   0.0  0 162 2.5 1055.0 676.0  28 61.88737
## 3 332.5 142.5  0 228 0.0  932.0 594.0 270 40.26954
## 4 332.5 142.5  0 228 0.0  932.0 594.0 365 41.05278
## 5 198.6 132.4  0 192 0.0  978.4 825.5 360 44.29608
## 6 266.0 114.0  0 228 0.0  932.0 670.0  90 47.02985
```

```
mod.full<-lm(Y~.,data=concrete)
anova(mod.full)
```

```
## Analysis of Variance Table
##
## Response: Y
##             Df Sum Sq Mean Sq  F value    Pr(>F)
## X1           1  71172   71172 658.0463 < 2.2e-16 ***
## X2           1  22957   22957 212.2606 < 2.2e-16 ***
## X3           1  21636   21636 200.0464 < 2.2e-16 ***
## X4           1  11459   11459 105.9488 < 2.2e-16 ***
## X5           1   1360    1360  12.5785 0.0004079 ***
## X6           1    253     253   2.3435 0.1261178
## X7           1      1       1   0.0058 0.9393393
```

```
## X8           1  47905   47905 442.9232 < 2.2e-16 ***
## Residuals 1021 110428     108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the anova table above, 1021=n-8-1. So n=1030

a) forward selection with BIC

```
mod.0<-lm(Y~1, data=concrete) #linear model with only intercepts
mod.full<-lm(Y~.,data=concrete) #full model
step(mod.0, scope = list(lower = mod.0, upper = mod.full), direction = 'forwa
rd', k = log(1030))
```

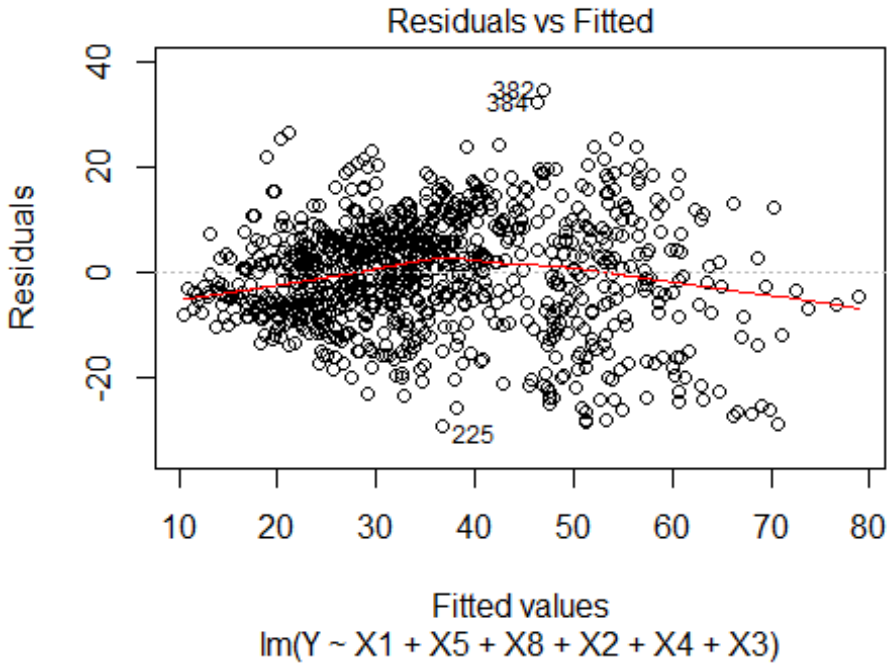
```
## Start:  AIC=5806.38
## Y ~ 1
##
##        Df Sum of Sq    RSS    AIC
## + X1    1     71172 216001 5520.0
## + X5    1     38490 248683 5665.1
## + X8    1     31061 256112 5695.4
## + X4    1     24087 263086 5723.1
## + X7    1      8033 279140 5784.1
## + X6    1      7811 279362 5784.9
## + X2    1      5220 281953 5794.4
## + X3    1      3212 283961 5801.7
## <none>              287173 5806.4
##
## Step:  AIC=5519.97
## Y ~ X1
##
##        Df Sum of Sq    RSS    AIC
## + X5    1   29646.5 186354 5374.8
## + X8    1   23993.8 192007 5405.6
## + X2    1   22957.4 193043 5411.2
## + X4    1   17926.8 198074 5437.7
## + X6    1    3548.0 212453 5509.8
## + X3    1    2894.4 213106 5513.0
## <none>              216001 5520.0
## + X7    1     960.2 215041 5522.3
##
## Step:  AIC=5374.85
## Y ~ X1 + X5
##
##        Df Sum of Sq    RSS    AIC
## + X8    1     37498 148857 5150.4
## + X2    1     19456 166898 5268.2
## + X7    1      5862 180493 5348.9
## <none>              186354 5374.8
## + X4    1       782 185572 5377.5
```

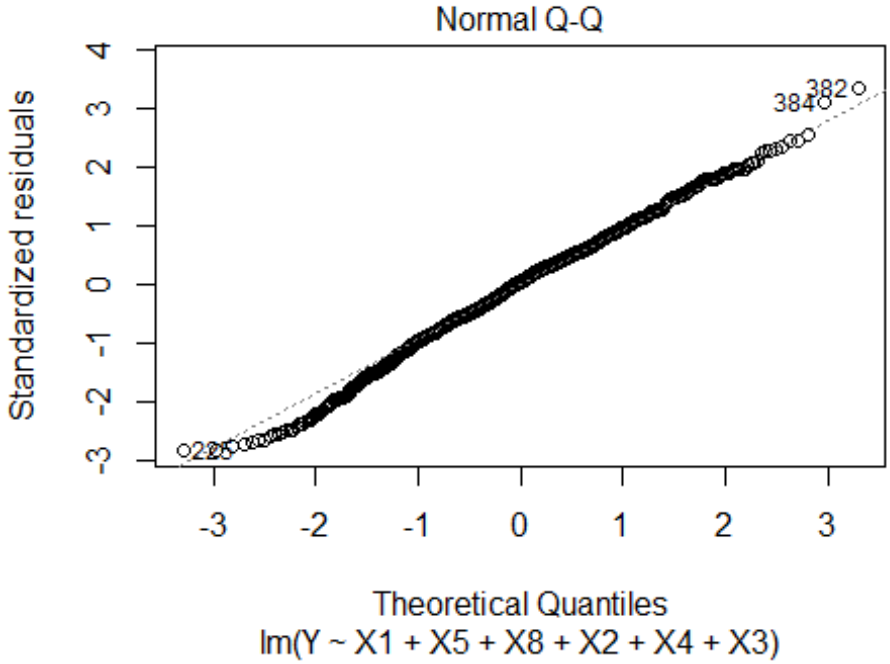
```
## + X3    1       741 185613 5377.7
## + X6    1       241 186113 5380.4
##
## Step:  AIC=5150.38
## Y ~ X1 + X5 + X8
##
##        Df Sum of Sq    RSS    AIC
## + X2    1   19908.5 128948 5009.4
## + X4    1    4868.8 143988 5123.1
## + X7    1    3385.5 145471 5133.6
## <none>              148857 5150.4
## + X3    1     323.9 148533 5155.1
## + X6    1      36.9 148820 5157.1
##
## Step:  AIC=5009.43
## Y ~ X1 + X5 + X8 + X2
##
##        Df Sum of Sq    RSS    AIC
## + X4    1    9544.7 119403 4937.2
## + X3    1    6524.7 122423 4962.9
## + X6    1    1737.0 127211 5002.4
## <none>              128948 5009.4
## + X7    1       3.5 128945 5016.3
##
## Step:  AIC=4937.16
## Y ~ X1 + X5 + X8 + X2 + X4
##
##        Df Sum of Sq    RSS    AIC
## + X3    1    8547.4 110856 4867.6
## + X7    1    1895.7 117508 4927.6
## <none>              119403 4937.2
## + X6    1      24.1 119379 4943.9
##
## Step:  AIC=4867.59
## Y ~ X1 + X5 + X8 + X2 + X4 + X3
##
##        Df Sum of Sq    RSS    AIC
## <none>              110856 4867.6
## + X6    1    44.271 110812 4874.1
## + X7    1    29.398 110827 4874.3

##
## Call:
## lm(formula = Y ~ X1 + X5 + X8 + X2 + X4 + X3, data = concrete)
##
## Coefficients:
## (Intercept)           X1           X5           X8           X2
##    29.03022      0.10543      0.23900      0.11349      0.08649
```

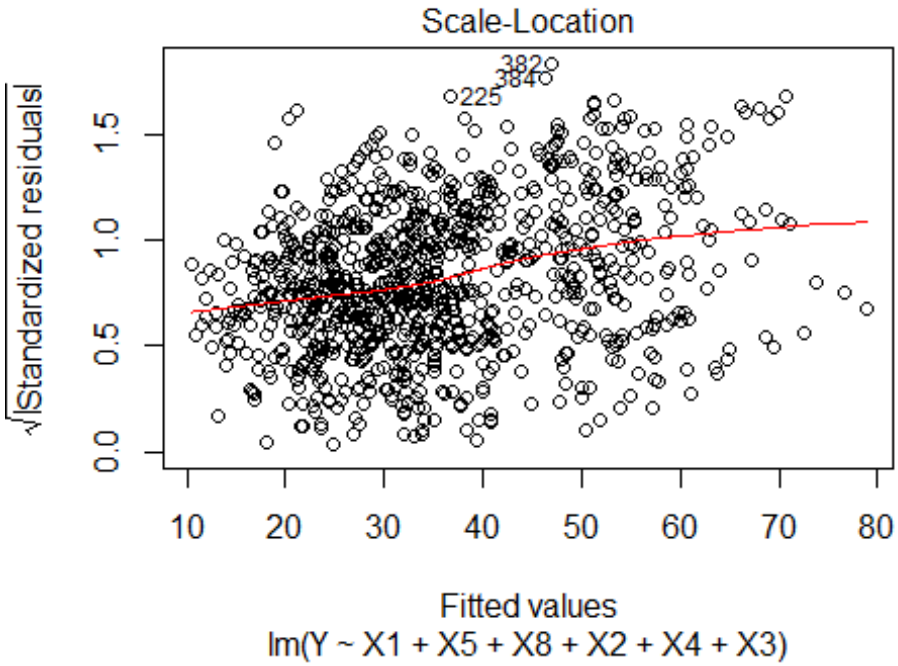
```
##          X4          X3
##    -0.21829     0.06871
```

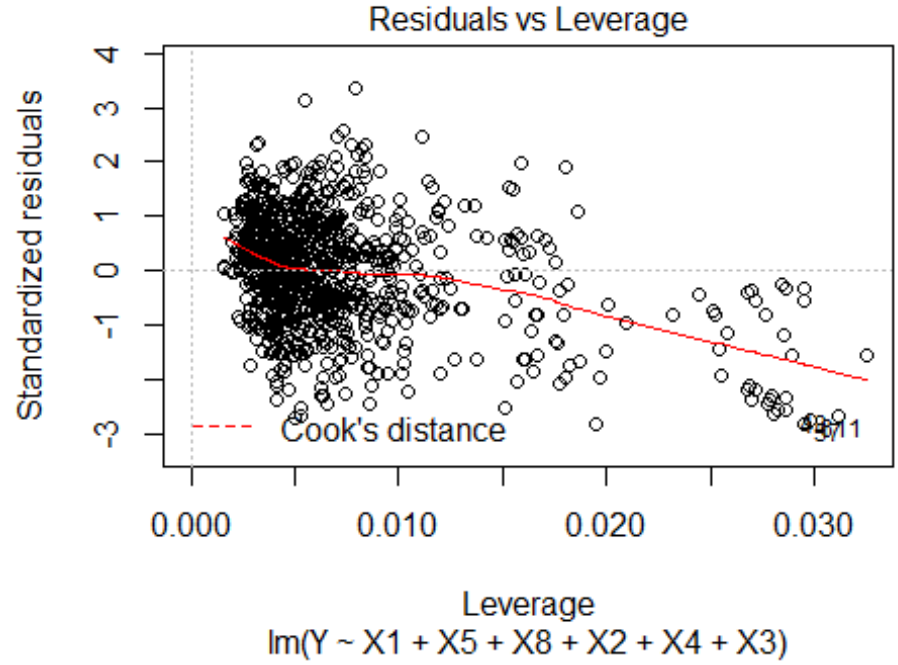
Do diagnostic checks

```
mod.for<-lm(Y ~ X1 + X5 + X8 + X2 + X4 + X3, data = concrete)
plot(mod.for)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Y ~ X1 + X5 + X8 + X2 + X4 + X3)
382
384
225
40
20
0
-20
10
20
30
40
50
60
70
80


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Y ~ X1 + X5 + X8 + X2 + X4 + X3)
382
384
225
4
3
2
1
0
-1
-2
-3

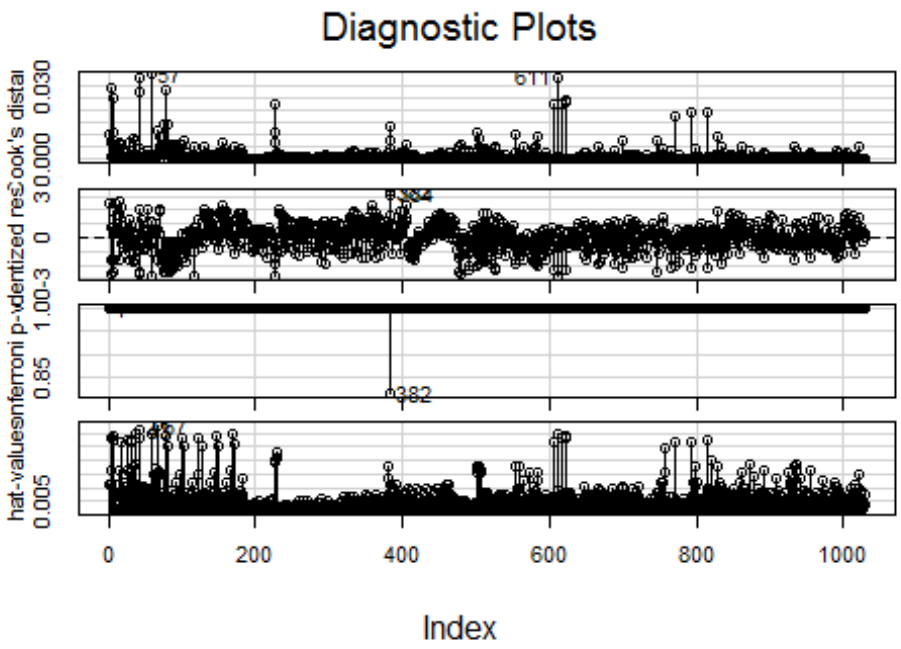




According to the plots, model by forward selection pretty fufills noramlity, linearity, and constant variance.

Use influenceIndexPlot to find influential points

```
library(car)
infIndexPlot(mod.for)
```




```
predict(mod.for, data.frame(X1= 200, X5=10, X8=100, X2=150, X4=180, X3=85),
interval = 'confidence', level = 0.95)

##        fit      lwr      upr
## 1 43.37686 42.12569 44.62803
```

We are 95% confident that true mean response is between 42.126 and 44.628

```
predict(mod.for, data.frame(X1= 200, X5=10, X8=100, X2=150, X4=180, X3=85),
interval = 'prediction', level = 0.95)

##        fit      lwr      upr
## 1 43.37686 22.91161 63.84211
```

We are 95% confident that concrete compressive strength for and individual value of each predictor values is between 22.912 and 63.842

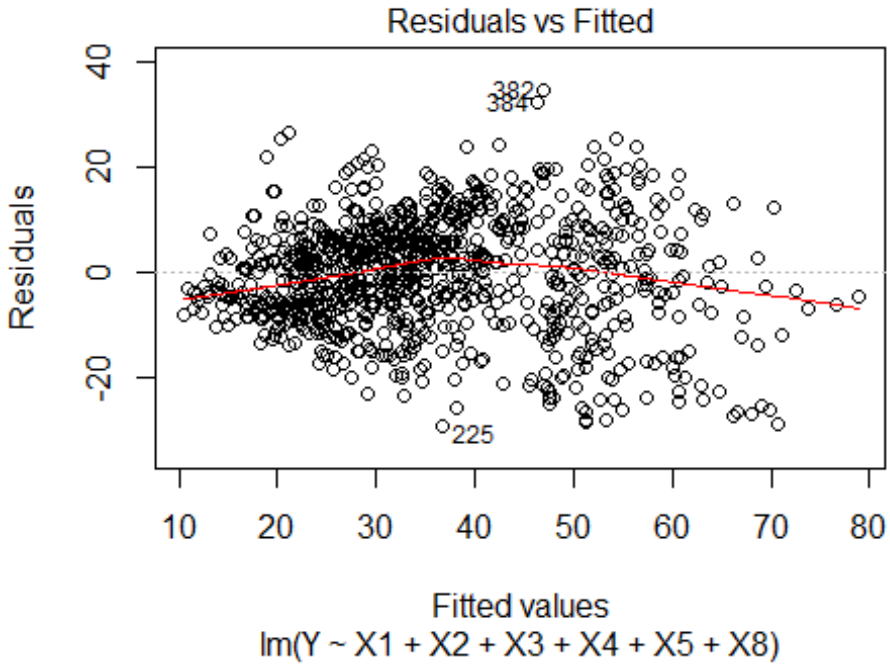
(b) Backward elimination with BIC

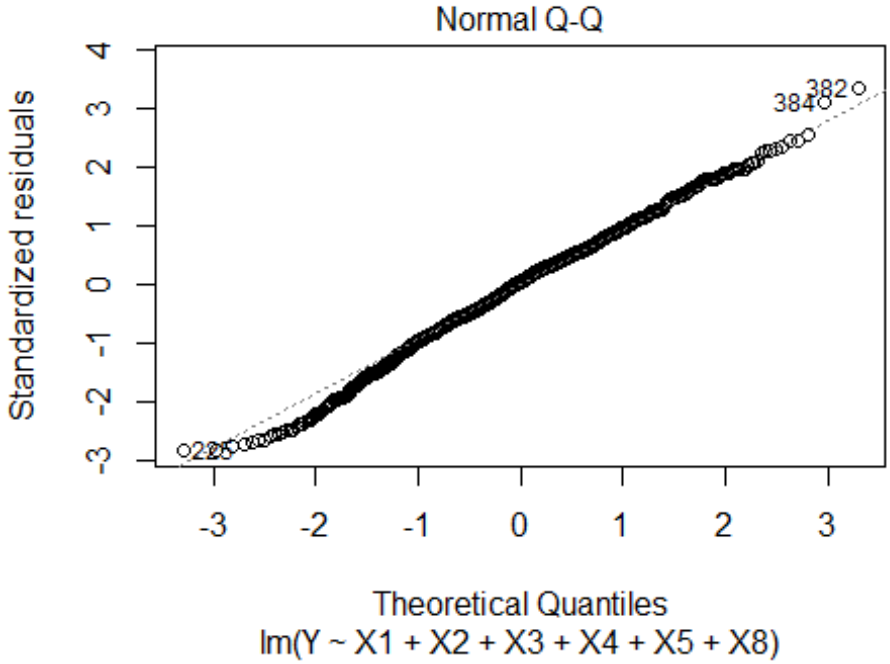
```
step(mod.full, scope = list(lower = mod.0, upper = mod.full), direction = 'ba
ckward', k = log(1030))
```

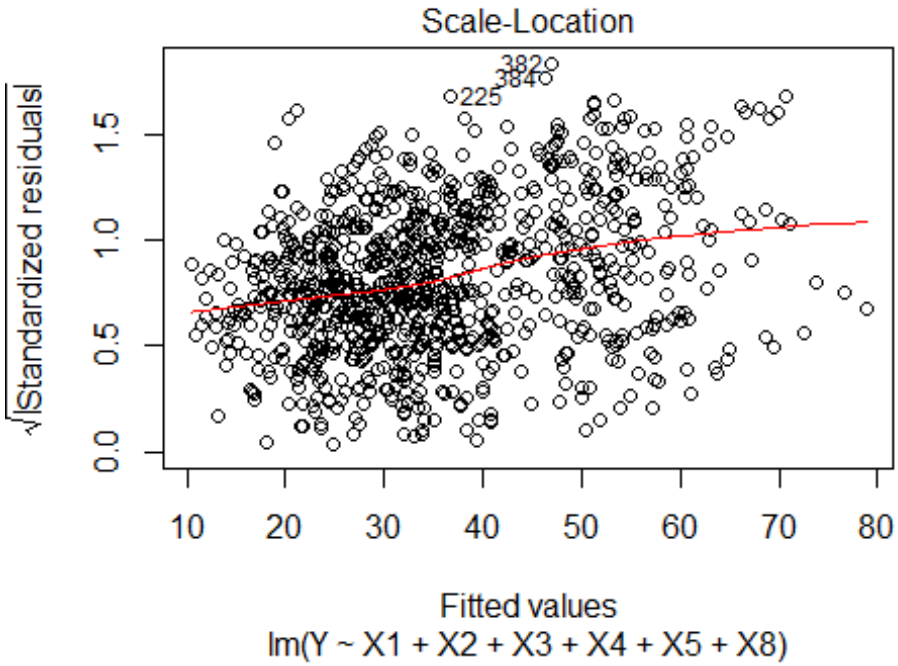
```
## Start:  AIC=4877.49
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##
##        Df Sum of Sq    RSS    AIC
## - X7    1       384 110812 4874.1
## - X6    1       398 110827 4874.3
## <none>              110428 4877.5
## - X5    1      1046 111474 4880.3
## - X4    1      1513 111942 4884.6
## - X3    1      5281 115709 4918.7
## - X2    1     11353 121781 4971.3
## - X1    1     21533 131961 5054.0
## - X8    1     47905 158333 5241.7
##
## Step:  AIC=4874.12
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X8
##
##        Df Sum of Sq    RSS    AIC
## - X6    1        44 110856 4867.6
## <none>              110812 4874.1
## - X5    1       877 111688 4875.3
## - X4    1      8526 119338 4943.5
## - X3    1      8568 119379 4943.9
## - X2    1     30693 141505 5119.0
## - X8    1     47522 158334 5234.8
## - X1    1     64008 174819 5336.8
##
## Step:  AIC=4867.59
## Y ~ X1 + X2 + X3 + X4 + X5 + X8
##
##        Df Sum of Sq    RSS    AIC
## <none>              110856 4867.6
## - X5    1       865 111721 4868.7
## - X3    1      8547 119403 4937.2
## - X4    1     11567 122423 4962.9
## - X2    1     32757 143613 5127.3
## - X8    1     47731 158587 5229.5
## - X1    1     66760 177616 5346.2

##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X8, data = concrete)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4
##    29.03022      0.10543      0.08649      0.06871     -0.21829
##          X5           X8
##     0.23900      0.11349
```

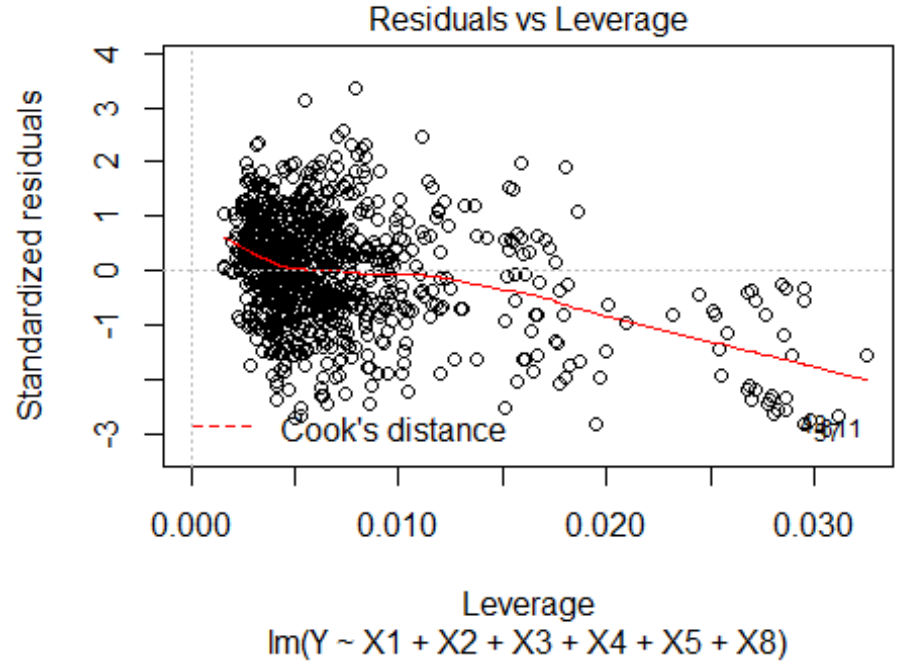
Do diagnostic checks

```
mod.back<-lm(Y ~ X1 + X2 + X3 + X4 + X5 + X8, data=concrete)
plot(mod.back)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Y ~ X1 + X2 + X3 + X4 + X5 + X8)
382
384
225


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Y ~ X1 + X2 + X3 + X4 + X5 + X8)
382
384
225

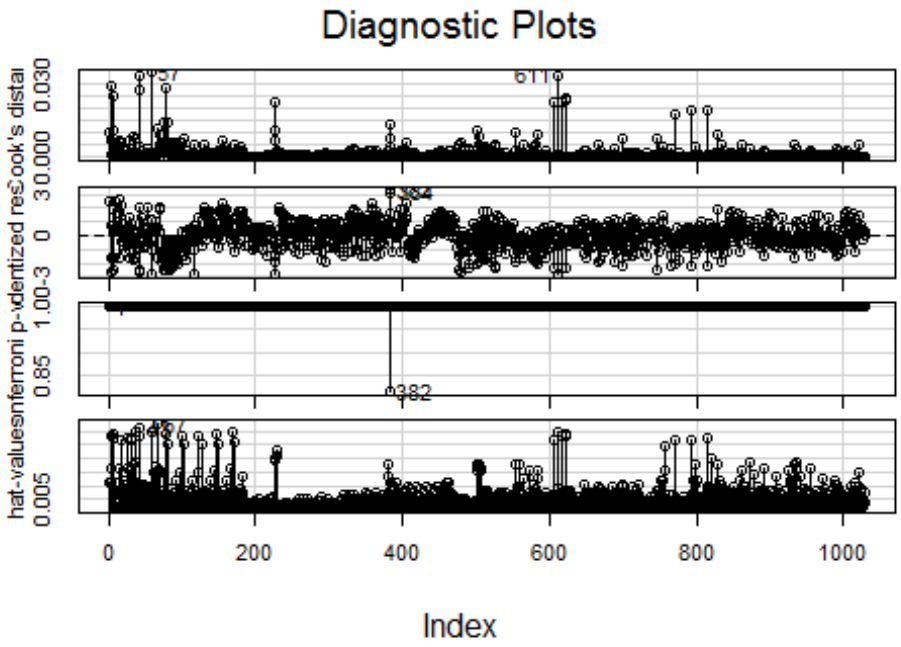




According to the plots, model by forward selection pretty fufills noramlity, linearity, and constant variance.

Use influenceIndexPlot to find influential points to remove

```
library(car)
infIndexPlot(mod.back)
```




```
anova(mod.for, mod.back)

## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X5 + X8 + X2 + X4 + X3
## Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X8
##   Res.Df    RSS Df  Sum of Sq F Pr(>F)
## 1   1023 110856
## 2   1023 110856  0 5.8208e-11
```

The models derived by two differnt methods are same except for the order of predictors. And the result of ANOVA also same.

Conclusions

We conduct two types of study to find the best way of model selection and both part 1 and part 2 are sort of choosing models, but the way they approach how to select the model is different. In the case of part 1, we set an abstract model first and then we try to find an adequate model by adding and deducting some variables. Lastly, we transform some relatively insignificant variables to get more precise model. On the other hand, in part 2, we

used two different model selecting methods: forward selection and backward elimination. When we use these methods, we do not have to go through all the processes of adding and deducting variables as we tried in part 1. Therefore, we conclude that even though we intend to choose the most adequate model in both Part1 and Part2, the method we use in the Part 2 is more efficient to choose the model. However, we believe that this method does not guarantee that the model we have as a result of the method in part 2 is absolutely right to make a conclusion because there is a possible way to get better model such as transformation and applying to nonlinear model.