Real Estate Valuation

Matthew Peters & Jalen Souksamlane

May 18, 2019

1.2 Project Components

a. We assume that the relationship between TDate and price will be relatively low (R^2 value will be low). Although the economic state of the area could have an influence on the transaction date we believe that their would be minimal influence compared to our others predictors. Thus, we assume that the association between TDate and price will be negative as well.

We assume that the relationship between the house age (years) and price will be high (R^2 value will be high) because the age of the real estate will determine how much money an investor will have to invest into the property for renovations and thus has a big effect on the price. Thus, we assume that the assocaition between house age and price will be positive.

We assume that the relationship between the number of conveince stores in the area and the price will be relatively high (R^2 value will be high) because more conveience stores could imply a more modern and urban area which would cause the price of real estate to jump higher. Thus, we assume that the association between the number of convenience stores and price will be positive.

We assume that the relationship between the latitude (geographic location) and the price will be relatively high (R^2 value will be high) because the price of a pice of real estate in the state of California vaires greatly from the price of real estate in the state of Colorado. Thus, we assume that the association between the latitude and price will positive.

```r
# Assigning the data points to their respective variable names
Price <- RealEstateValuation$Price
TDate <- RealEstateValuation$TDate
Age <- RealEstateValuation$Age
Metro <- RealEstateValuation$Metro
Stores <- RealEstateValuation$Stores
Latitude <- RealEstateValuation$Latitude
Longitude <- RealEstateValuation$Longitude

# Calculating the R^2 value for each variable individually
cor(Price, TDate)^2
```

```
## [1] 0.007654606
```

```r
cor(Price, Age)^2
```

```
## [1] 0.04433848
```

```r
cor(Price, Stores)^2
```

```
## [1] 0.3260466
```

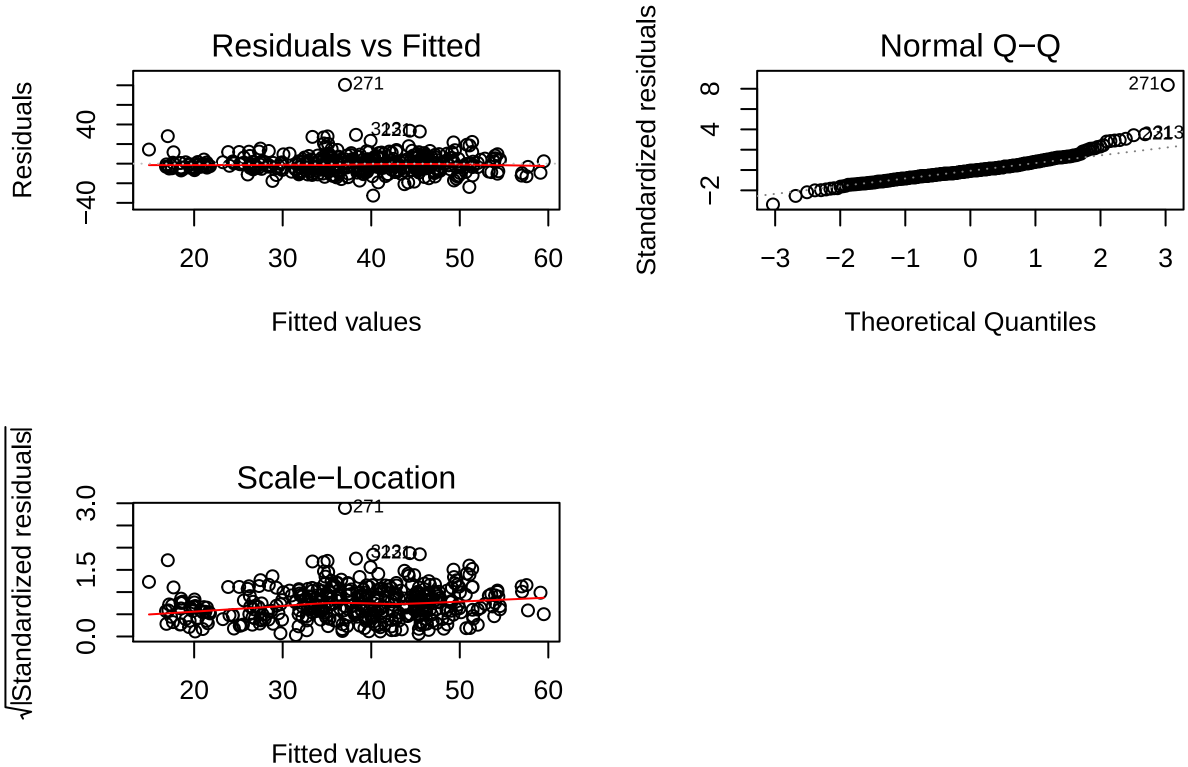
```r
cor(Price, Latitude)^2
```

```
## [1] 0.298451
```

```r
# Fitting the regression model
Model_1.lm <- lm(Price~TDate+Age+Stores+Latitude)
par(mfrow=c(2,2))
plot(Model_1.lm,which=1:3)

beta <- summary(Model_1.lm)$coefficients
beta
```

```
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -17419.9480668 3523.66616105 -4.943700 1.120613e-06
## TDate            3.6125826    1.68604620  2.142636 3.273204e-02
## Age             -0.3019689    0.04178232 -7.227194 2.436219e-12
## Stores           1.9291168    0.18008122 10.712482 9.059813e-24
```

```
## Latitude       407.8136729   42.77638010  9.533618 1.377128e-19
```



From the summary output we find that by conducting tests on individual regression coefficients with ?? = 0.01 and reading the summary output we find that, Age, Stores and Latitude have significant p-values. First off, the estimator coefficient of Age tells us that if the age of a house increases by 1 year then the price of a house decreases by -.301 Ping. The estimator coeffcient of Stores tells us that if 1 store is added to the area then the price of the house wull increase 1.929 Ping. The estimator coefficient of Latitude tells us that if the degree of latitude is increased by 1 unit then the price of the house increases by 407.814 Ping.

Multiple Regression Line of Price ~ TDate + Age + Stores + Latitude

Y = `-17419.948` + `3.613x1` + `-0.302x2` + `1.929x3` + `407.814x4`

```r
# Adding Metro and Longitude into the linear model
Model_4.lm <- lm(Price~TDate+Age+Stores+Latitude+Metro+Longitude)
model_41.lm<-lm(Price~TDate+Age+Stores+Latitude)
summary(Model_4.lm)
```

```
##
## Call:
## lm(formula = Price ~ TDate + Age + Stores + Latitude + Metro +
##     Longitude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.664  -5.410  -0.966   4.217  75.193
##
## Coefficients:
##                     Estimate    Std. Error t value Pr(>|t|)
```

```
## (Intercept) -14437.100802   6775.670673  -2.131  0.03371 *
## TDate            5.146228      1.557073   3.305  0.00103 **
## Age             -0.269695      0.038531  -7.000 1.06e-11 ***
## Stores           1.133277      0.188164   6.023 3.84e-09 ***
## Latitude       225.472976     44.566685   5.059 6.38e-07 ***
## Metro           -0.004488      0.000718  -6.250 1.04e-09 ***
## Longitude      -12.423601     48.581995  -0.256  0.79829
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.858 on 407 degrees of freedom
## Multiple R-squared:  0.5824, Adjusted R-squared:  0.5762
## F-statistic: 94.59 on 6 and 407 DF,  p-value: < 2.2e-16
```

```r
anova(model_41.lm,Model_4.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ TDate + Age + Stores + Latitude
## Model 2: Price ~ TDate + Age + Stores + Latitude + Metro + Longitude
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    409 38119
## 2    407 31933  2      6187 39.428 2.229e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

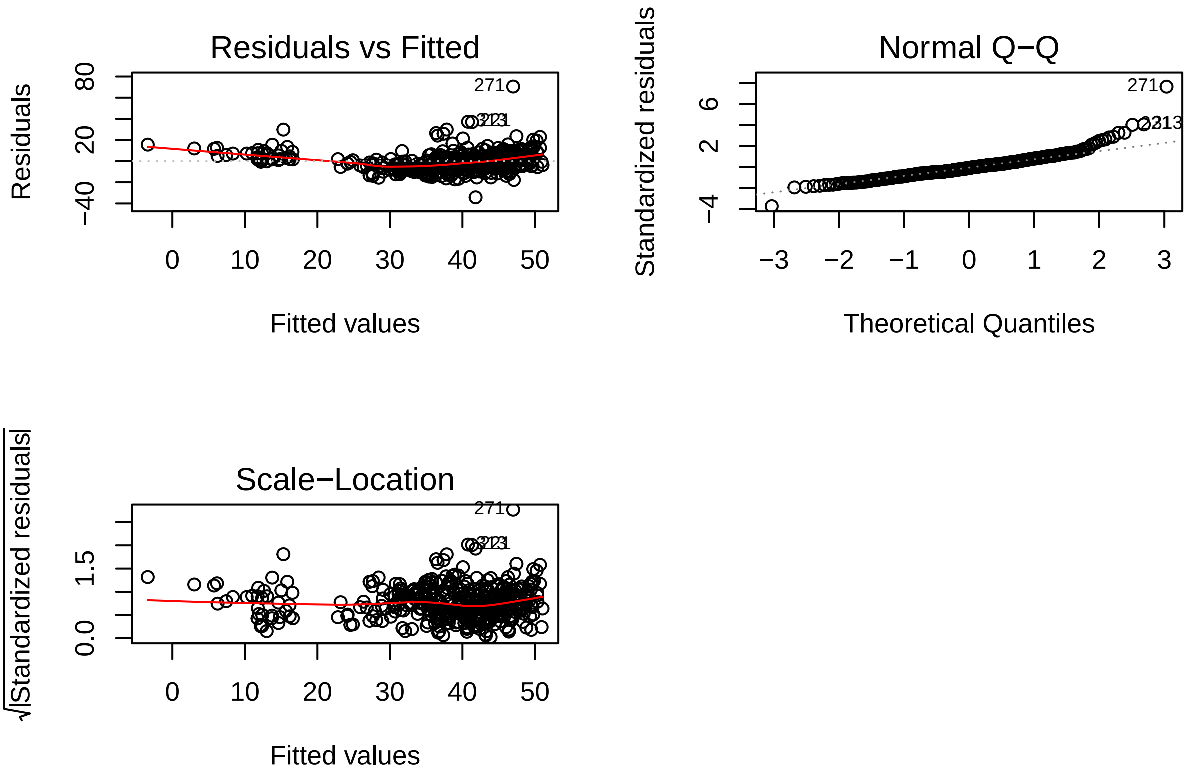
By using Partial F Test our null hypothesis would be that beta5=beta6=0 and the alternative hypothesis would be that either beta5 or beta6 doesn't equal 0. The value of the test statistic is 39.428 and our null dsitribution is 3.02. Since the test statistic is greater than the null distribution we would reject the null hypothesis. Additionally, from the summary of the model with Metro and Longitude added we can see that the p-value for Longitude is not significant and Metro is significant, thus we would keep Metro and disgard Longitude.

```r
Model_5.lm <- lm(Price~TDate+Age+Metro+Latitude)
summary(Model_5.lm)
```

```
##
## Call:
## lm(formula = Price ~ TDate + Age + Metro + Latitude)
##
## 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) -17673.0128549   3358.5720332  -5.262 2.30e-07 ***
## TDate            5.5698680      1.6192921   3.440 0.000642 ***
## Age             -0.2529986      0.0400098  -6.323 6.71e-10 ***
## Metro           -0.0057643      0.0004493 -12.829  < 2e-16 ***
## Latitude       260.6728430     45.6914324   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
```

```r
par(mfrow=c(2,2))
plot(Model_5.lm,which = 1:3)
```



We will now make added variable plots for the model Price~TDate+Age+Stores+Latitude and the model Price~TDate+Age+Metro+Latitude.

```r
model10<-lm(Price~TDate+Age+Stores+Latitude)
model11<-lm(Price~TDate+Age+Metro+Latitude)
summary(model10)
```

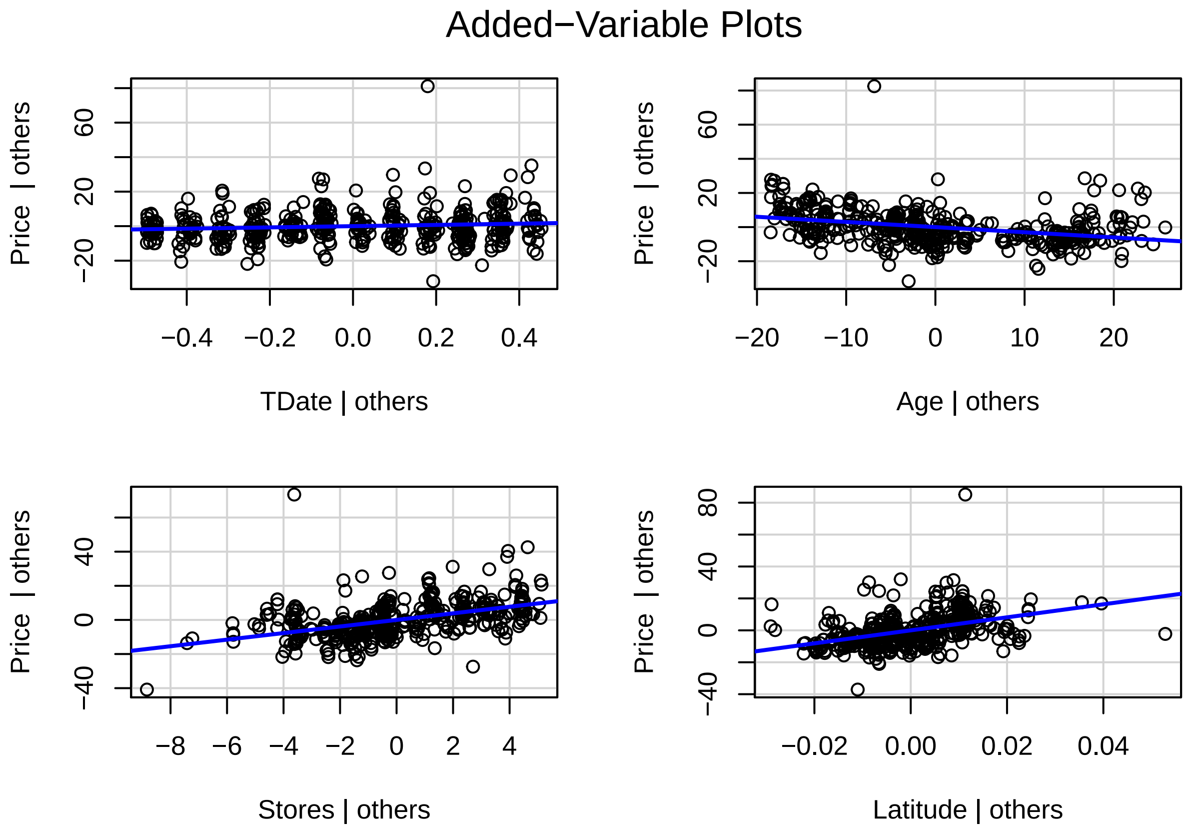
```
##
## Call:
## lm(formula = Price ~ TDate + Age + Stores + Latitude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.620  -5.601  -0.714   4.207  80.465
##
## Coefficients:
##                 Estimate   Std. Error t value Pr(>|t|)
## (Intercept) -17419.94807   3523.66616  -4.944 1.12e-06 ***
## TDate            3.61258      1.68605   2.143   0.0327 *
## Age             -0.30197      0.04178  -7.227 2.44e-12 ***
## Stores           1.92912      0.18008  10.712  < 2e-16 ***
## Latitude       407.81367     42.77638   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
```

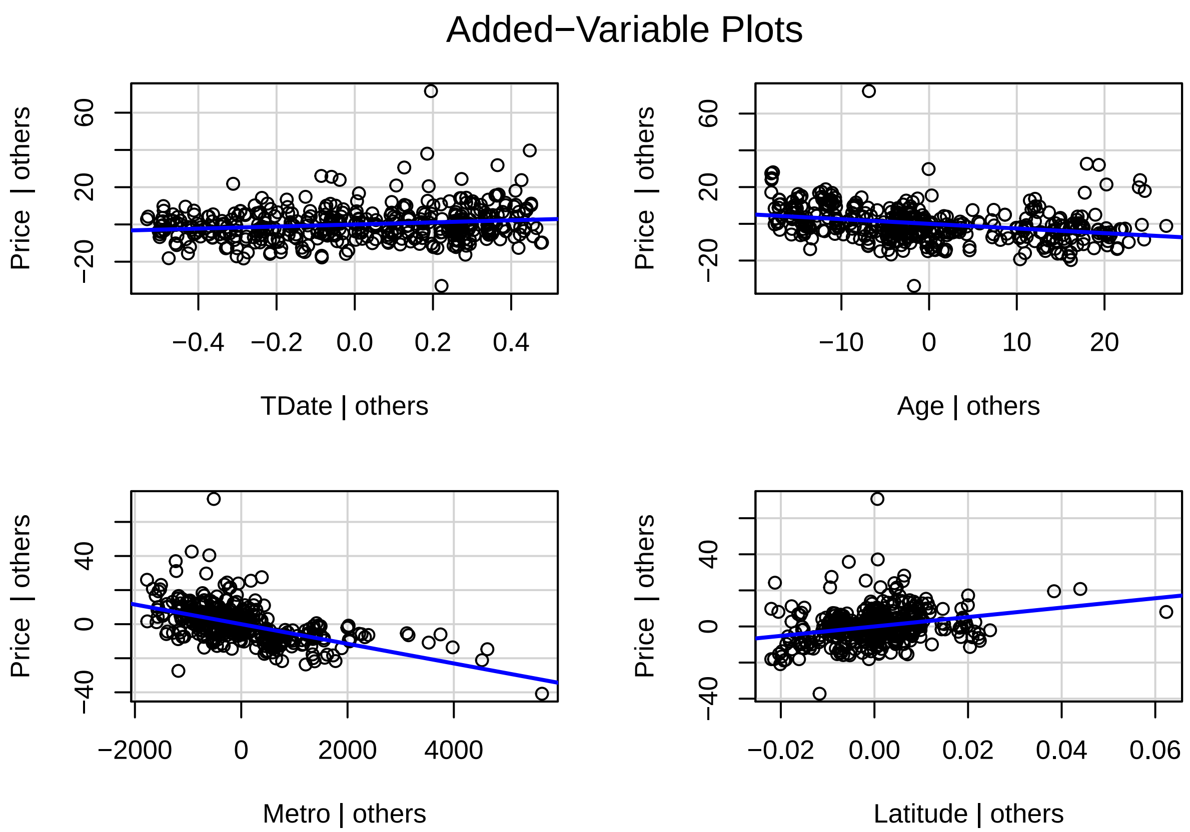
```
summary(model11)
```

```
##
## Call:
## lm(formula = Price ~ TDate + Age + Metro + Latitude)
##
## 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) -17673.0128549   3358.5720332  -5.262 2.30e-07 ***
## TDate            5.5698680      1.6192921   3.440 0.000642 ***
## Age             -0.2529986      0.0400098  -6.323 6.71e-10 ***
## Metro           -0.0057643      0.0004493 -12.829  < 2e-16 ***
## Latitude       260.6728430     45.6914324   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
```

```
avPlots(model10,id=FALSE)
```




```r
avPlots(model11,id=FALSE)
```



From the Added Variable Plots we can see that the slope of Metro is greater than the slope of stores thus showing that Metro is a more significant variable to have in the model than Stores.Additionally, we can see from the summary of both models that the adjusted R^2 value for the model containing Metro is greater than the Adjuested R^2 value for the model containing Stores thus Metro has a greater effect on Price. Therefore the preferred model is Price~TDate+Age+Metro+Latitude.

Price~Metro+Age+Latitude+TDate

```
library(car)
age <- ifelse(Age==0,Age +.01,Age)
pt<-powerTransform(cbind(Metro, age, Latitude,TDate)~-1, data = RealEstateValuation)
```

```
## Warning in sqrt(diag(solve(res$hessian))): NaNs produced
```

```
summary(pt)
```

```
## Warning in sqrt(diag(object$invHess)): NaNs produced
```

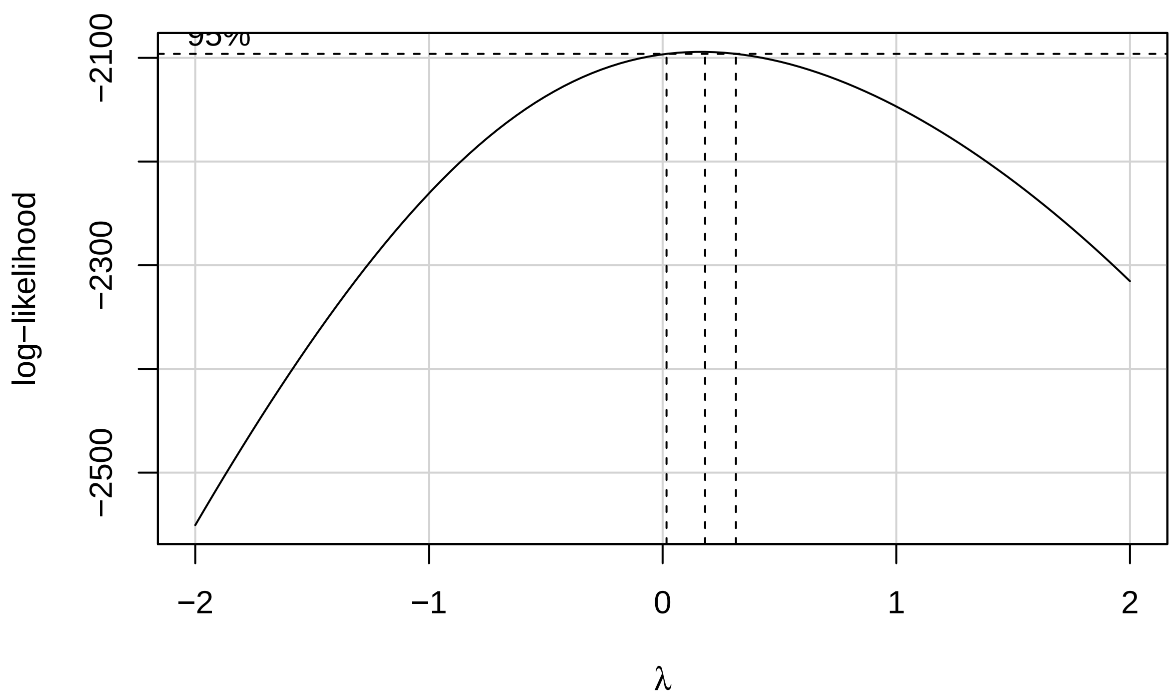
```
## bcPower Transformations to Multinormality
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Metro        0.0780         0.0      -0.0020       0.1581
## age          0.5467         0.5       0.4749       0.6185
## Latitude     3.0000         1.0    -147.1913     153.1914
## TDate        3.0000         3.0          NaN          NaN
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
```

```
##                                   LRT df       pval
## LR test, lambda = (0 0 0 0) 451.1085  4 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                   LRT df       pval
## LR test, lambda = (1 1 1 1) 557.4249  4 < 2.22e-16
```

We can summarize that the null hypothesis is the lambda value of Metro,Age,Latitude, and TDate equal to 0 and the alternative hypothesis is that atleast one of the values of lambda is not equal to 0. Since from the power transformation the p-values of all the variabels are 0 so we would reject the null hypothesis thus we will log transofrm Metro since from the summary it's raised the power is 0 and we will square root the variable "Age" since it has a power of .5.

Next we will perform a Box-Cox transformation to see if the response varibale needs to be transformed.

```
#Box Cox Method
RElm<-lm(Price~.,data = RealEstateValuation)
boxCox(RElm)
```



From he Box Cox Transformation, since the interval is relatively close to 0 we can conclude that to use a log transformation for the response variable "Price".

```
#Box Cox Method, univariate
summary(l1<-powerTransform(Price~Metro+Age+Latitude+TDate,RealEstateValuation))
```

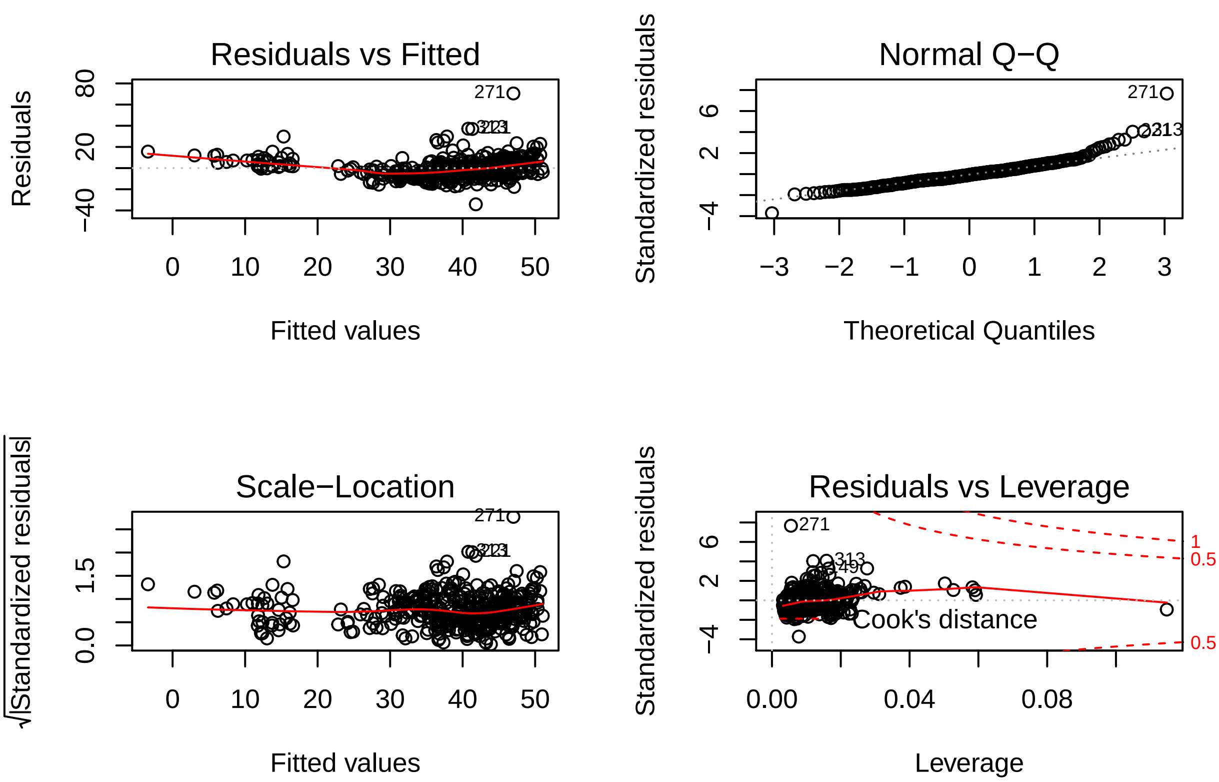
```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.1304           0      -0.0223       0.2832
##
```

```
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                              LRT df     pval
## LR test, lambda = (0) 2.877549  1 0.089823
##
## Likelihood ratio test that no transformation is needed
##                              LRT df       pval
## LR test, lambda = (1) 105.1433  1 < 2.22e-16
```

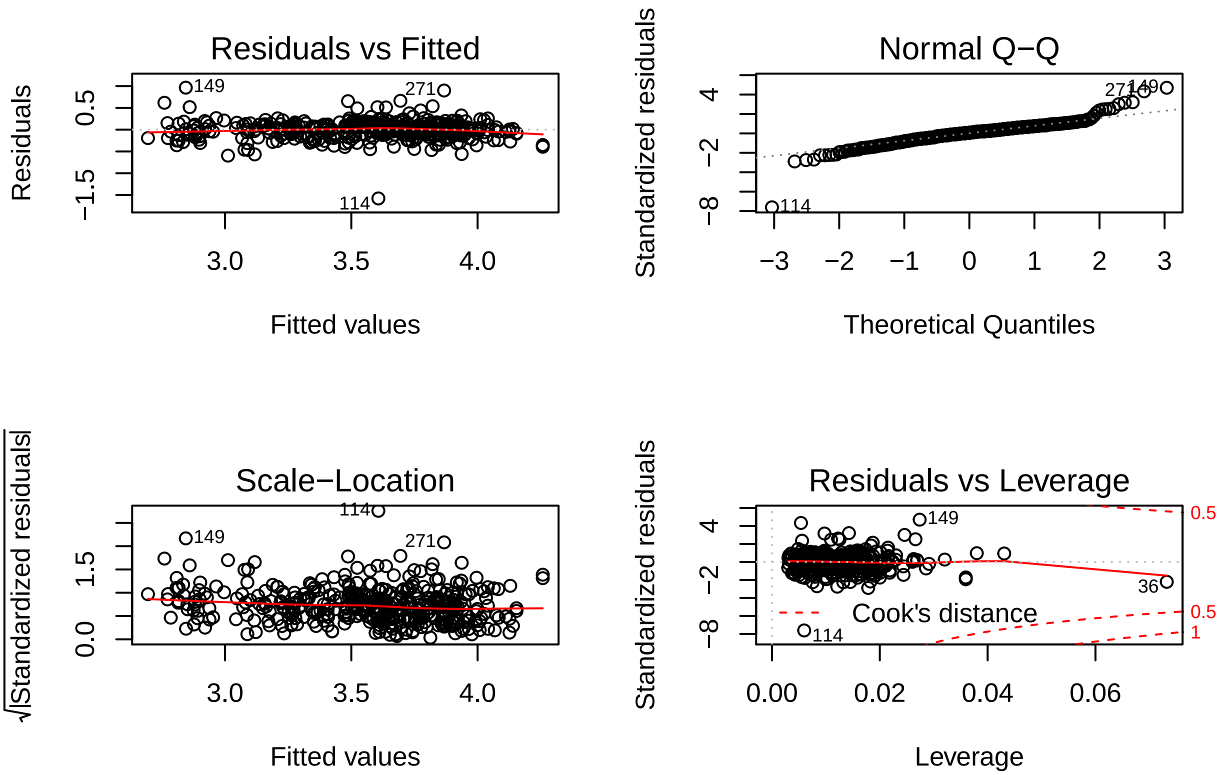
We can see from the Power Transformation that this supports our statement of performing a log transformation onto the response variable "Price".

Now we will fit the transformed predictors into a model and check to see if there were any improvements

```r
# the transformed model comapred with the original model
original_model<-lm(Price~Metro+Age+Latitude+TDate)
final_model<-lm(log(Price)~log(Metro)+sqrt(Age)+Latitude+TDate)
par(mfrow=c(2,2))
plot(original_model)
```




```r
par(mfrow=c(2,2))
plot(final_model)
```




```r
summary(original_model)
```

```
##
## Call:
## lm(formula = Price ~ Metro + Age + Latitude + TDate)
##
## 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) -17673.0128549  3358.5720332  -5.262 2.30e-07 ***
## Metro           -0.0057643     0.0004493 -12.829  < 2e-16 ***
## Age             -0.2529986     0.0400098  -6.323 6.71e-10 ***
## Latitude       260.6728430    45.6914324   5.705 2.23e-08 ***
## TDate            5.5698680     1.6192921   3.440 0.000642 ***
## ---
## 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
```

```r
summary(final_model)
```

```
##
```

```
## Call:
## lm(formula = log(Price) ~ log(Metro) + sqrt(Age) + Latitude +
##     TDate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57902 -0.10462  0.01289  0.11008  0.96421
##
## Coefficients:
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept) -632.721677   75.394641  -8.392 7.87e-16 ***
## log(Metro)    -0.204963    0.010526 -19.472  < 2e-16 ***
## sqrt(Age)     -0.047799    0.006657  -7.180 3.31e-12 ***
## Latitude      11.077076    0.935566  11.840  < 2e-16 ***
## TDate          0.179421    0.036637   4.897 1.40e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.208 on 409 degrees of freedom
## Multiple R-squared:  0.7218, Adjusted R-squared:  0.719
## F-statistic: 265.2 on 4 and 409 DF,  p-value: < 2.2e-16
```

From plotting the residual vs. fitted, Q-Q, and scale location plots, we were able to notice significant improvements from the transformations. For example, from the Residual vs. Fitted plots of the origianl model(Price~Metro+Age+Latitude+TDate) we can see an improvement in the transformed model(log(Price)~log(Metro)+sqrt(Age)+Latitude+TDate) because the residual vs. fitted of the original model does not hold linearity and constant variance while the residual vs. fitted plot of the transformed model holds both linearity and constant variance. Although there is slight improvement of normality in the Q-Q plot the difference in the Scale Location plot of the transformed model is more significant because we see the points more spreadout and having a constant variance along the line.

log(Price)~log(Metro)+sqrt(Age)+latitude+TDate

```r
summary(final_model)
```


```
##
## Call:
## lm(formula = log(Price) ~ log(Metro) + sqrt(Age) + Latitude +
##     TDate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57902 -0.10462  0.01289  0.11008  0.96421
##
## Coefficients:
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept) -632.721677   75.394641  -8.392 7.87e-16 ***
## log(Metro)    -0.204963    0.010526 -19.472  < 2e-16 ***
## sqrt(Age)     -0.047799    0.006657  -7.180 3.31e-12 ***
## Latitude      11.077076    0.935566  11.840  < 2e-16 ***
## TDate          0.179421    0.036637   4.897 1.40e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

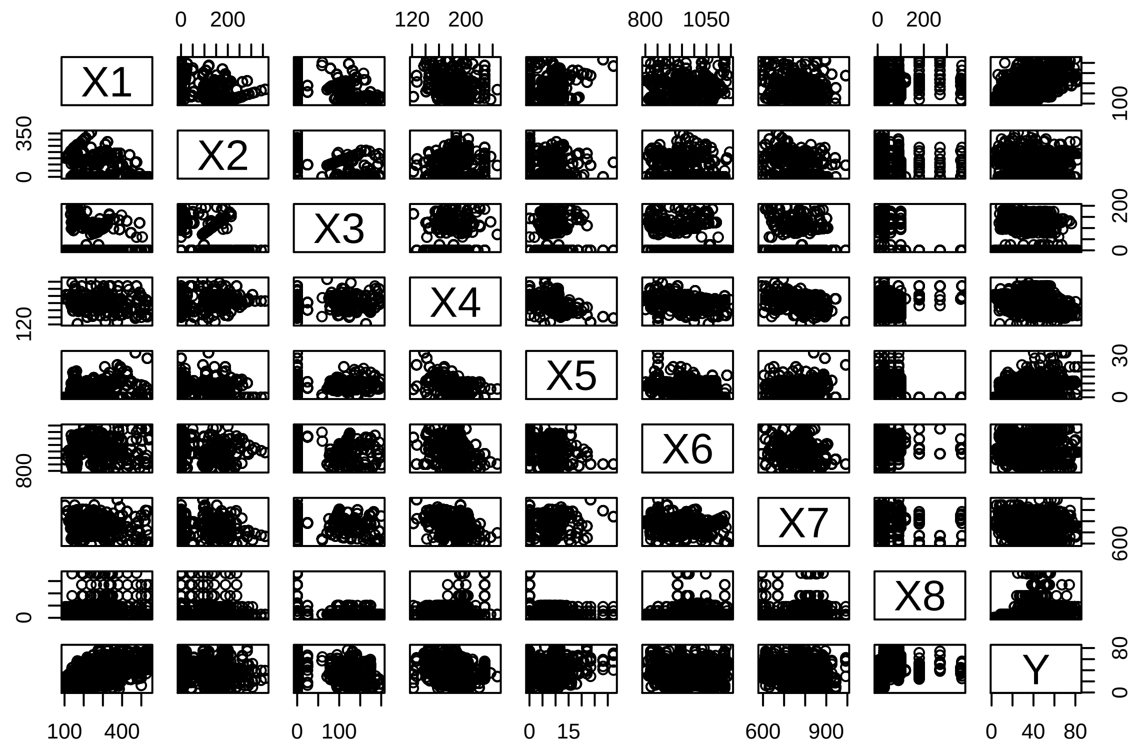
```
## Residual standard error: 0.208 on 409 degrees of freedom
## Multiple R-squared:  0.7218, Adjusted R-squared:  0.719
## F-statistic: 265.2 on 4 and 409 DF,  p-value: < 2.2e-16
```

From this analysis an interesting point we found out was that Latitude had the biggest effect on Price at the end. From the summary of our Final model we could interpret that if Latitude increases by 1 degree then the price of a house will increase by 11.08 Ping in terms of the Sindian District. It is also interesting to see that the distance to the nearest Metro station has a greater effect on the house price than the Age of the house becaause initially we thought that Age would have a large effeect rather it didn't in this case.

Part 2

```
# Reading in Concrete data file
Concrete <- read.csv("C://Users//Jalen//Desktop//PSTAT126//Concrete.txt", sep="")
# Creating variables for each predictors and response
X1 <- Concrete$X1
X2 <- Concrete$X2
X3 <- Concrete$X3
X4 <- Concrete$X4
X5 <- Concrete$X5
X6 <- Concrete$X6
X7 <- Concrete$X7
X8 <- Concrete$X8
Y <- Concrete$Y
```

```
# Sample size of the dataset
n <- length(Concrete$Y)

pairs(Concrete)
```



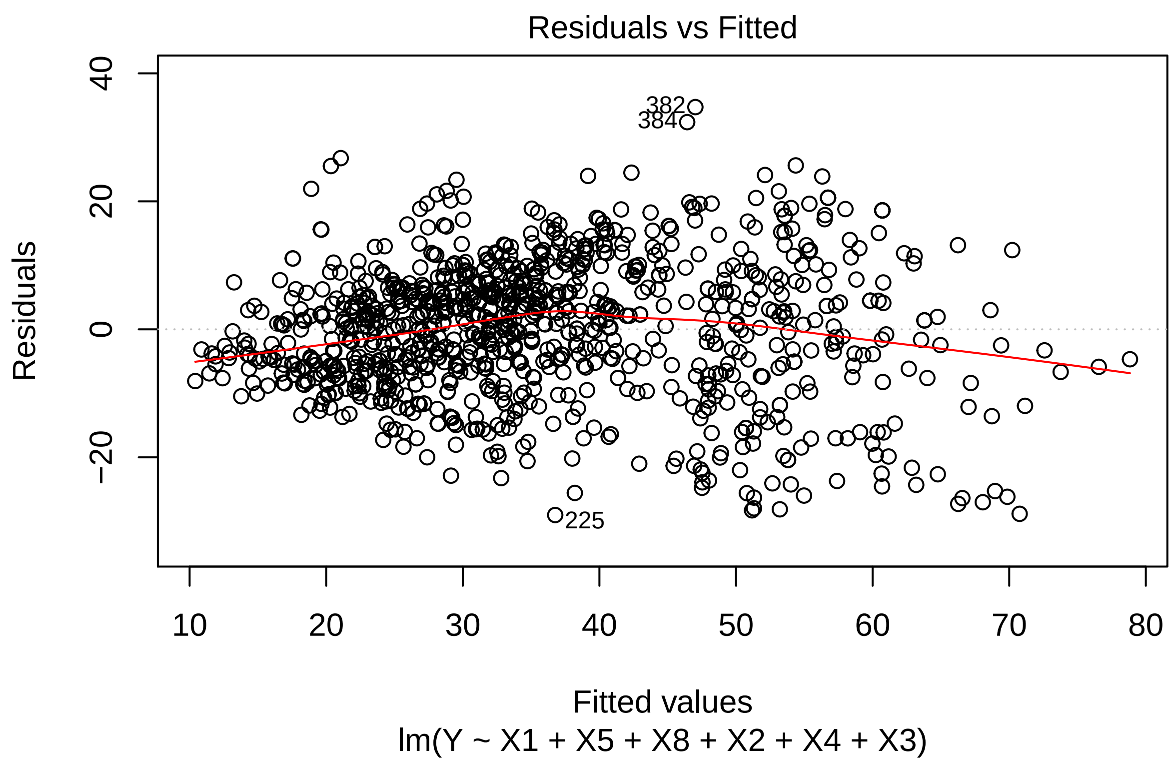

```r
# Smallest model for the dataset
mod.0 <- lm(Y~X1)

# Largest model for the dataset
mod.full <- lm(Y~X1+X2+X3+X4+X5+X6+X7+X8)

# Forward BIC test on the model
step(mod.0, scope = list(lower = mod.0, upper = mod.full), direction = 'forward', k = log(n), trace= 0)
```

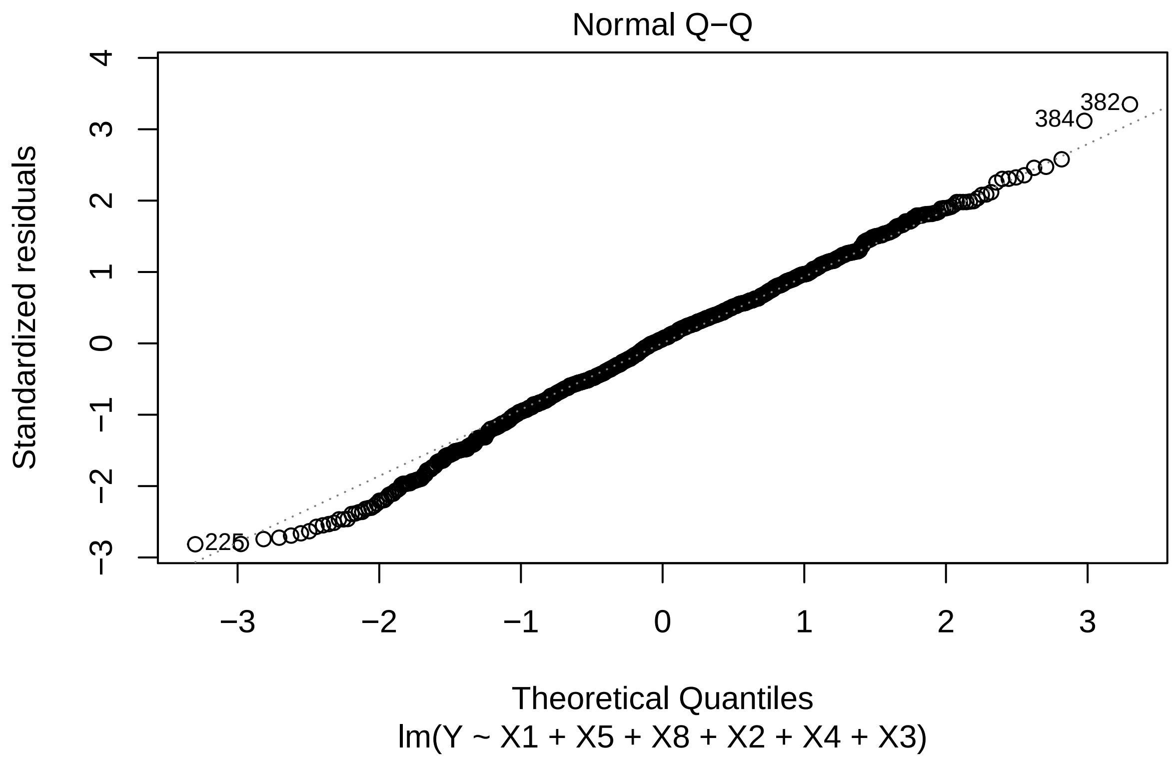
```
##
## Call:
## lm(formula = Y ~ X1 + X5 + X8 + X2 + X4 + X3)
##
## Coefficients:
## (Intercept)           X1           X5           X8           X2
##    29.03022      0.10543      0.23900      0.11349      0.08649
##          X4           X3
##    -0.21829      0.06871
```

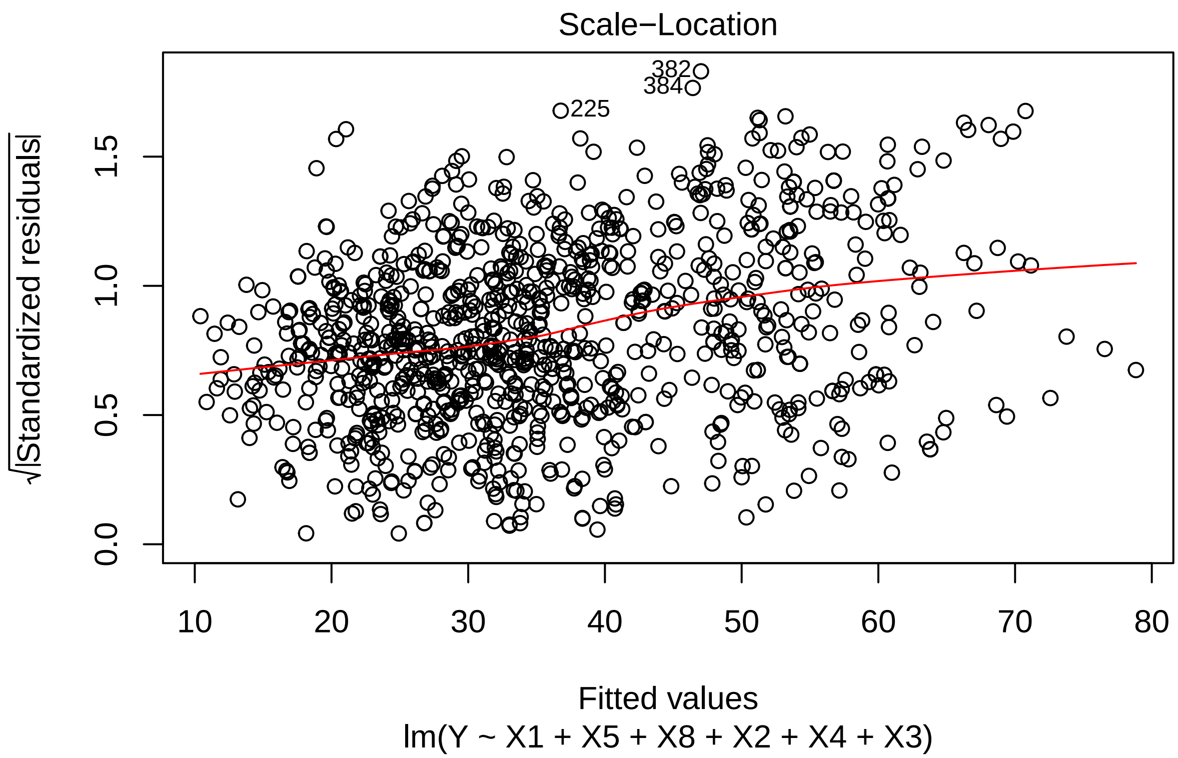
```r
# New model found through forward BIC method
mod.better <- lm(Y ~ X1 + X5 + X8 + X2 + X4 + X3)
```

```r
# Mean Response response of the model
Yhat <- fitted(mod.better)

# Calculation for the residuals of the model
e <- Y - Yhat
```

```r
# Diagnostic Checks to asses linear regression assumptions
#par(mfrow=c(1,1))
plot(mod.better, which=1:3)
```

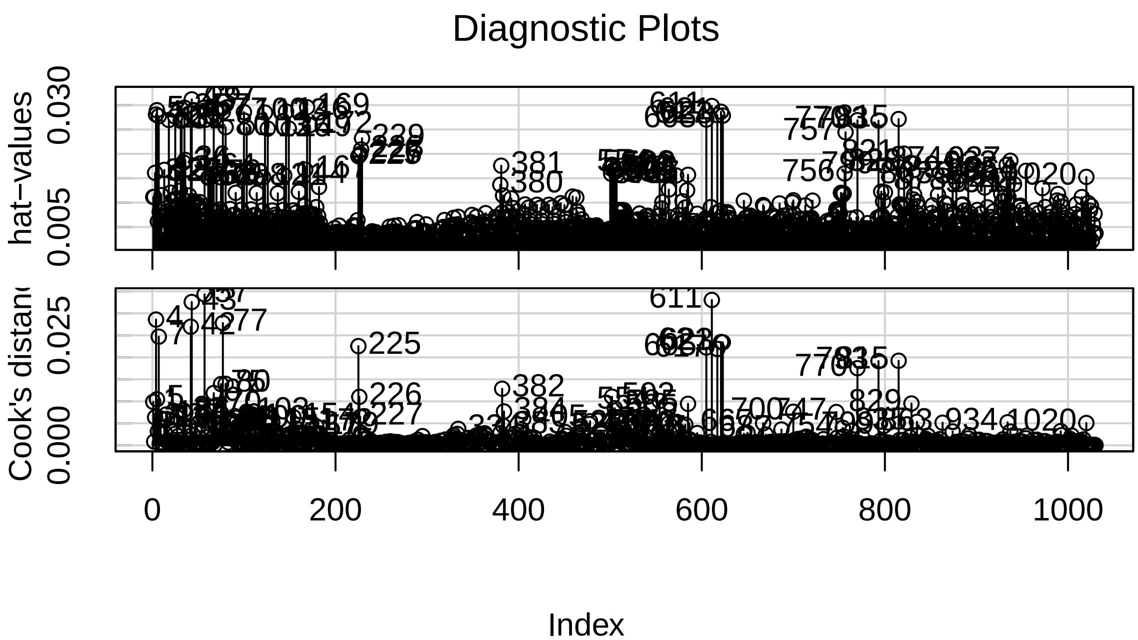



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Y ~ X1 + X5 + X8 + X2 + X4 + X3)
382
384
225
−3
−2
−1
0
1
2
3
4

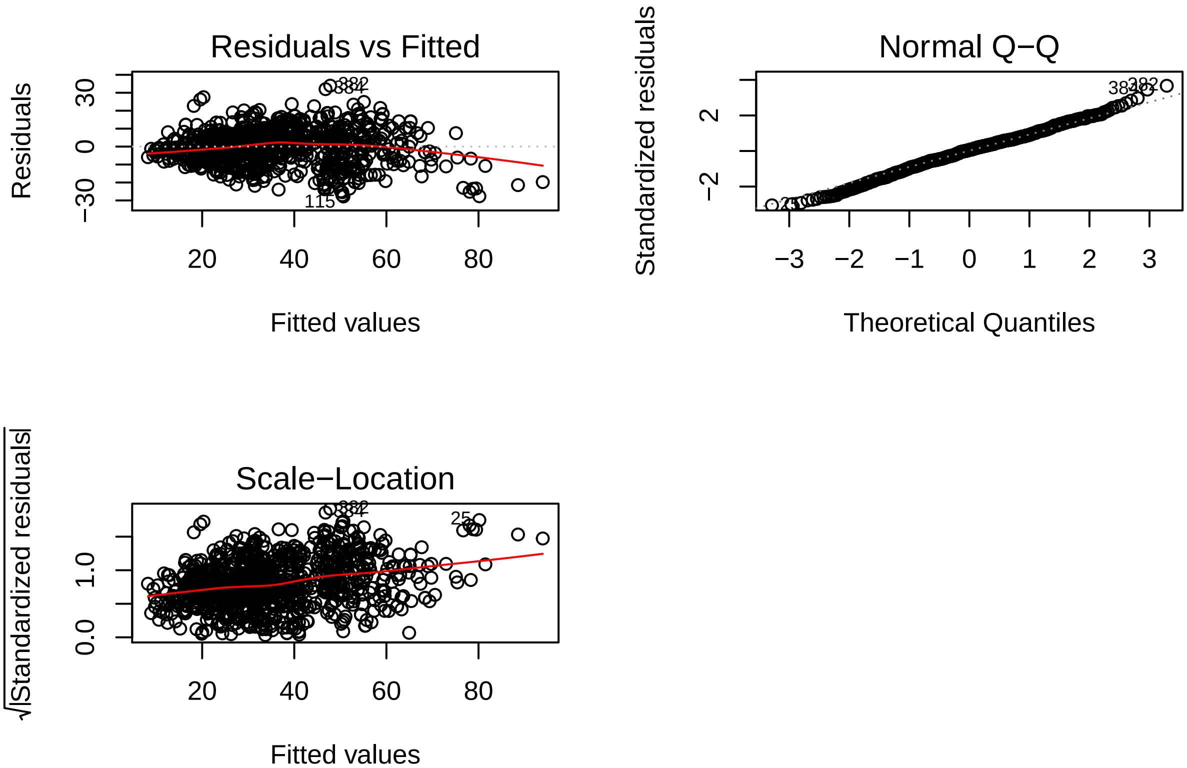


From running a Diagnostic Check on the new model (Y ~ X1 + X5 + X8 + X2 + X4 + X3) we can see from the Residual vs. Fitted plot that linearity does hold although the spread of residuals seems to be decreasing as the fitted values change. Thus, there is a slight variation for the constant variance assumption. For the Normal Q-Q plot we can see that the Normality asssumption does hold. With the Scale Location plot we can see that as the fitted values get larger the spread of the data points decrease although for the majority of data points constant variancs does hold.

```
# Leverage
h <- hatvalues(mod.better)

p <-sum(h)

high.leverage <- which(h > (2*p) / n )

# Cook's Statistic
cd <- cooks.distance(mod.better)

cooks <- which(cd > (4 / (n-p-1)))

# Taking out points with high leverage and high influence
Concrete2 <- Concrete[c(1:2,6,8:12,14:26,28:33,35,37:41,44:56,58:66,68:69,71:74,76,78:79,81:99,101:102,

# Finding influential points
influenceIndexPlot(mod.better, vars = c('hat', 'Cook'), id=list(n=83))
```




```
mod.better1 <- lm(Y ~ X1 + X5 + X8 + X2 + X4 + X3, data=Concrete2)
par(mfrow=c(2,2))
plot(mod.better1,which = 1:3)
```



So in order to test for influential points we checked which data points had a high leverage and which data points had a high Cook's statistic and we compared the two vectors to see where the data points intersected in order to determine the points that had both high leverage and high influence. Next, we created a new data set without the influential points and ran diagnostic checks to see if there were any improvements. From the Diagnostic Plots we can see a visual representation of the data points with high leverage veresus a high Cook's statisitc. From the Q-Q plot we can see that the Normality assumption holds as well. Additionally, we see the the normality assumption and constant variance assumption hold also in the Residual versus fitted and Scale Location Plots.

```
# 95% Condidence Interval
# Estimated Values
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
```

```
new <- data.frame(X1=mean(X1), X2=107, X3=100, X4=mean(X4), X5=7, X8=mean(X8))
ans <- predict(mod.better,new,se.fit=TRUE,interval='confidence',level=0.95,type='response')
ans$fit
```

```
##        fit      lwr      upr
## 1 42.01935 40.94628 43.09242
```

In order to create the mean response we compiled a data frame with the means of each predictor value although we changed the mean values of the predictors who's median was at 0 in order to present the data better. Then we computed a 95% confidence interval whcih tells us that, we are 95% confident that with the variabels presented in the model (Xi where i=1 through 8) the concrete comprehensive strength is between (40.95,43.09) MPa.

```
# 95% Prediction Interval
ans2 <- predict(mod.better,new,se.fit=TRUE,interval='prediction',level=0.95,type='response')
ans2$fit
```

```
##        fit      lwr      upr
## 1 42.01935 21.56422 62.47449
```

By calculating the prediction interval we are 95% confident that next new observation of concrete comprehensive strength will fall within the range of (21.56,62.47).

```
# Backward BIC test on the model
step(mod.full, scope = list(lower = mod.0, upper = mod.full), direction = 'backward', k = log(n), trace=
```

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

```
# New model found through forward BIC method
mod.better2 <- lm(Y ~ X1 + X2 + X3 + X4 + X5 + X8)
```

```
summary(mod.better)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X5 + X8 + X2 + X4 + X3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.014  -6.474   0.650   6.546  34.726
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 29.030224   4.212476   6.891 9.64e-12 ***
## X1           0.105427   0.004248  24.821  < 2e-16 ***
## X5           0.239003   0.084586   2.826  0.00481 **
## X8           0.113495   0.005408  20.987  < 2e-16 ***
## X2           0.086494   0.004975  17.386  < 2e-16 ***
## X4          -0.218292   0.021128 -10.332  < 2e-16 ***
## X3           0.068708   0.007736   8.881  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.41 on 1023 degrees of freedom
## Multiple R-squared:  0.614,  Adjusted R-squared:  0.6117
## F-statistic: 271.2 on 6 and 1023 DF,  p-value: < 2.2e-16
```

```r
summary(mod.better2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X8)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.014  -6.474   0.650   6.546  34.726
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.030224   4.212476   6.891 9.64e-12 ***
## X1           0.105427   0.004248  24.821  < 2e-16 ***
## X2           0.086494   0.004975  17.386  < 2e-16 ***
## X3           0.068708   0.007736   8.881  < 2e-16 ***
## X4          -0.218292   0.021128 -10.332  < 2e-16 ***
## X5           0.239003   0.084586   2.826  0.00481 **
## X8           0.113495   0.005408  20.987  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.41 on 1023 degrees of freedom
## Multiple R-squared:  0.614,  Adjusted R-squared:  0.6117
## F-statistic: 271.2 on 6 and 1023 DF,  p-value: < 2.2e-16
```

As we can see the model from backwards BIC is the same as the model from forwards BIC thus, the influential points will be the same for both models and we conclude that our final model will be:

Y ~ X1 + X2 + X3 + X4 + X5 + X8

An interesting point from this analysis is that forward and backward BIC both had the same model at the end and. In terms of our final model we found out that Superplasticizer (X5) had the largest effect on Concrete comprehesive strength. In fact, we found out that if superplasticizer increases by 1 (kg/m^3) then the concrete comprehensive strength increases by .24 Mpa.