Final Project: Option 2

Seth Marceno, Kevin Sandoval, Derek Yan

6/9/2019

Contents

Abstract

In this project, we examined real estate valuation from Sindian District, New Taipei City, Taiwan and looked to see how different predictors and different transformations would affect the price of real estate in this city in Taiwan. We found that the best regression we could do for price was TDate, sqrt(Age), log(Metro), and Latitude as our predictors. We also examined the eight different components of concrete strength, to determine which components were the most significant as predictors. We found that Cement(X1) , Blast Furnace Slag(X2), Fly Ash(X3), Water(X4), Superplasticizer(X5), and Age(X8) were the most significant predictors in determining concrete strength.

Problem and Motivation

For the first data set, the information provided is relevant because there are many people who may be impacted by the findings. For example, if someone in Taiwan wanted to purchase real estate in Taiwan, the data we collected and compiled can give an insight into what would cause this person to pay more, or less money for their real estate. In our testing, we found in our best regression model that for every unit (year) that TDate increases, each square meter in the house will increase in price by 1,794 Taiwanese dollars. This is interesting as this predictor on its own, implies that waiting to buy will cost you money. The concrete dataset is a similar representation. The strength of concrete is an integral part of the society we live in today, as sidewalks, roads, and bridges are often made using this material. It is important to distinguiush the important factors that contribute to the overall effect of the strength of concrete. As stated in the abstract, we found that certain predictors play a large role in determining the overall strength of the concrete, which would impact which parts of the construction process should be more heavily focused on and considered.

Description of Data

The real estate valuation data looks at different predictors and their effect on the price of the houses in that real estate market. In our scenario, the relevant variables that we have are TDate (transaction date), Age, Metro, Latitude. In the Concrete dataset, we examined the predictors effects on the strength of the concrete as a whole. We found are notable variables to be Cement(X1) , Blast Furnace Slag(X2), Fly Ash(X3), Water(X4), Superplasticizer(X5), and Age(X8).

Questions of Interest

For our first data set, real estate in Taiwan, we wanted to see if the price of the home was dependent on age of the house, distance to metro stations, number of convenience stores, as well as degree of latitude and longitude that the house was located at. Our first step is to test to see whether or not all of these predictors had any significant effect on the price of the house. Once we have found the best fit to predict the price of the house, we can look at our model to see which predictors cause an increase in housing price or a decrease in it. On top of this we can answer the mean price for a house given it is a certain distance from a metro, is a certain age, is located on the West coast ... etc. For our second data set concrete, we wanted to see if the concrete compressive strength was dependent on its makeup. Thus, once we have found the best components of concrete that predict its compressive strength, we are able to see which components increase or decrease its strength. Furthermore, given a certain makeup of the concrete, we will be able to see the mean compressive strength as wel as predict the compressive strenth of all concrete of the same makeup.

Regression Methods

To answer our questions of interest, for part 1 with the real estate data set, in order to fit our model we first looked at a scatterplot matrix in order to determine relationships between the predictors and the response, as well as relationships between the predictors themselves. Next, we fit a base model and tested whether or not we should add the variable Metro and/or Longitude. Once we had found our best model, the next step was to determine if the model needed to have its predictors and response transformed. We did this by doing power transform tests for the predictors and Box Cox estimations for the response. For part 2 with the concrete data set, in order to fit our model we used both forward and backward selection with BIC criterion. Once we found our best model we went on to run diagnostics to check if our model assumptions held. To do this we looked at added variable plots to test relations, residual vs. fitted plots to test linearity, scale-invariance plot to test constant variance, as well as Q-Q plot to test normality.

Regression Analysis

Part 1:

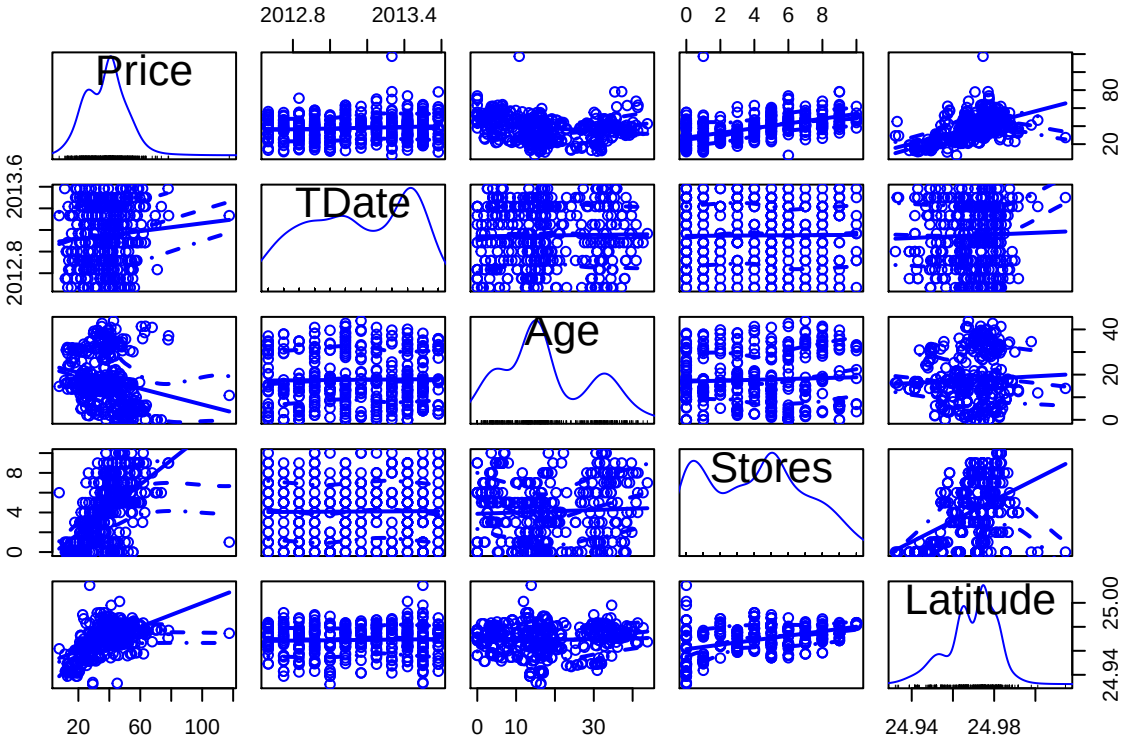
Here we will be looking at the Real Estate Valuation Data set:

```
RealEstate <- read.table('RealEstateValuation.txt')
Price <- RealEstate$Price
TDate <- RealEstate$TDate
Age <- RealEstate$Age
Stores <- RealEstate$Stores
Latitude <- RealEstate$Latitude
Metro <- RealEstate$Metro
Longitude <- RealEstate$Longitude
```

Checking Relationships Between Variables

The first step we will take is to explore explore any relationships between the predictors and response, or between the predictors using the model Price ~ TDate + Age + Stores + Latitude. Based off of intuition, we would expect to see a relationship between Price and Age, as well as Price and Stores. However, we do not expect to see any significant relationships between the predictors TDate, Age, Stores, and Latitude. Now, looking at the scatter plot matrix to see any relationships:

```
scatterplotMatrix(~Price + TDate + Age + Stores + Latitude)
```



Here we can see that Price vs. each predictor seems to have some sort of relationship as expected. Looking at the plots comparing our predictors to each other, we see that the relation between TDate and Age, Stores, or Latitude seems to have no significant relationship; simlarly between Age and Stores there seems to be no correlation. Now looking at the relationship between Stores and Latitude, it looks like a normal distribution when Latitude is the predictor and Stores is the response, and a flat line when Stores is the predictor and Latitude is the response. Thus, it seems that given a certain latitude, we may expect there to be more stores around.

Fitting a Model to Predict Price

Now we will run a regression to see if we can predcit house price using the predictors TDate, Age, Stores, and Latitude.

```
fit1 <- lm(Price ~ TDate + Age + Stores + Latitude)
summary(fit1)
```

```
##
## 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) -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
```

Using the summary we can see that the model for our data is Price = -1.742e+04(TDate) - 3.02e-1(Age) + 1.929(Stores) + 4.078e+02(Latitude). Also, we are given all of the P-values of the individual coefficients. We can see that all coefficients, expect for TDate's main effect are significant at the alpha = 0.01 level.

Looking at all of the significant coefficients at alpha = 0.01 level, Age, Stores, and Latitude, we find that for every year the house ages, the price will drop 3,020 New Taiwan Dollars/Ping. For every store added to the living circle on foot, the price of the house will rise 19,200 New Taiwan Dollars/Ping. Lastly, for every degree of latitude added, the price of the house will rise 4,078,00 New Taiwan Dollars/Ping.

Testing To add Metro and/or Longitude

Now we will preform partial F-Tests to test out if Metro and Longitude should be added to the model.

H0:The coeffiencet for Longitude is 0 vs. Ha:The coefficient for Longitude is not 0 at level alpha = 0.05

```
fit2 <- lm(Price ~ TDate + Age + Stores + Latitude + Longitude)
summary(fit2)
```

```
##
## Call:
## lm(formula = Price ~ TDate + Age + Stores + Latitude + Longitude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.482  -5.722  -1.110   4.363  80.383
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.231e+04  5.333e+03  -7.933 2.08e-14 ***
## TDate        4.210e+00  1.621e+00   2.598  0.00972 **
## Age         -2.797e-01  4.025e-02  -6.949 1.46e-11 ***
## Stores       1.566e+00  1.830e-01   8.558 2.35e-16 ***
## Latitude     3.376e+02  4.265e+01   7.916 2.35e-14 ***
## Longitude    2.093e+02  3.470e+01   6.033 3.60e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.262 on 408 degrees of freedom
## Multiple R-squared:  0.5423, Adjusted R-squared:  0.5367
## F-statistic: 96.68 on 5 and 408 DF,  p-value: < 2.2e-16
```

Thus based off of our T-test, we find our test statistic = 6.033, with a null distribution of T_n-p-1 = T_408. Our P-value is 3.60e-09 which is less than our alpha = 0.05; thus, we reject our null hypothesis.

H0:The coeffiencet for Metro is 0 vs. Ha:The coefficient for Metro is not 0 at level alpha = 0.05

```
fit3 <- lm(Price ~ TDate + Age + Stores + Latitude + Metro)
summary(fit3)
```

```
##
## Call:
## lm(formula = Price ~ TDate + Age + Stores + Latitude + Metro)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.623  -5.371  -1.020   4.244  75.346
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.596e+04  3.233e+03  -4.936 1.17e-06 ***
## TDate        5.135e+00  1.555e+00   3.303  0.00104 **
## Age         -2.694e-01  3.847e-02  -7.003 1.04e-11 ***
## Stores       1.136e+00  1.876e-01   6.056 3.17e-09 ***
## Latitude     2.269e+02  4.417e+01   5.136 4.36e-07 ***
## Metro       -4.353e-03  4.899e-04  -8.887  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.848 on 408 degrees of freedom
## Multiple R-squared:  0.5823, Adjusted R-squared:  0.5772
## F-statistic: 113.8 on 5 and 408 DF,  p-value: < 2.2e-16
```

Thus based off of our T-test, we find our test statistic = -8.887, with a null distribution of T_n-p-1 = T_408. Our P-value is 2e-16 which is less than our alpha = 0.05; thus, we reject our null hypothesis.

H0:The coeffiencet for longitude is 0 with metro in our model vs. Ha: The coefficient for longitude is not 0 with metro is in our model at level alpha = 0.05

```
fit4 <- lm(Price ~ TDate + Age + Stores + Latitude + Metro + Longitude)
summary(fit4)
```

```
##
## 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) -1.444e+04  6.776e+03  -2.131  0.03371 *
## TDate        5.146e+00  1.557e+00   3.305  0.00103 **
## Age         -2.697e-01  3.853e-02  -7.000 1.06e-11 ***
## Stores       1.133e+00  1.882e-01   6.023 3.84e-09 ***
## Latitude     2.255e+02  4.457e+01   5.059 6.38e-07 ***
## Metro       -4.488e-03  7.180e-04  -6.250 1.04e-09 ***
## Longitude   -1.242e+01  4.858e+01  -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
```

Thus based off of our T-test, we find our test statistic = -0.256, with a null distribution of T_n-p-1 = T_408. Our P-value is 0.79829 which is not less than our alpha = 0.05; thus, we fail reject our null hypothesis.

Since we have found that the main effect for Metro is significant both with and without Longitude in the model, whereas Longitude is only significant when Metro is not in the model, we have determined our best model is Price ~ TDate + TDate + Age + Latitude + Metro.

Considering Another Model

Here we will compare two different models, Price ~ TDate + Age + Metro + Latitude and Price ~ TDate + Age + Stores + Latitude

```
fit5 <- lm(Price ~ TDate + Age + Metro + Latitude)
summary(fit1)
```

```
##
## 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) -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
```

```
summary(fit5)
```

```
##
## 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) -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
```

Based off of our summary table we can see that fit1 has a correlation coefficient of 0.5015 whereas fit5 has a correlation coefficient of 0.5448. Since our model fit5 (Price ~ TDate + Age + Metro + Latitude) explains more of the variance in house price given its predictors, we will use this model. On top of this, we see that the main effect of TDate is also significant at a high level in fit5 than in fit1.

Transforming the Predictors

Since the predictor Age contains values that are 0, and we cannot power transform or log transform these predictors. Therefore, we must do some manipulation to make sure our predictors are stricly non-zero.

```
RealEstate$Age <- with(RealEstate, (Age + 0.01))
```

Now we test to see if we need to power transform:

```
RE.pt <- powerTransform(cbind(TDate, Age, Metro, Latitude) ~ 1, RealEstate)
summary(RE.pt)
```

```
## bcPower Transformations to Multinormality
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## TDate       3.0000         1.0    -503.9213     509.9213
## Age         0.5469         0.5       0.4751       0.6187
## Metro       0.0780         0.0      -0.0020       0.1581
## Latitude    3.0000         1.0    -147.6445     153.6446
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                   LRT df       pval
## LR test, lambda = (0 0 0 0) 451.5795  4 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                   LRT df       pval
## LR test, lambda = (1 1 1 1) 557.2776  4 < 2.22e-16
```

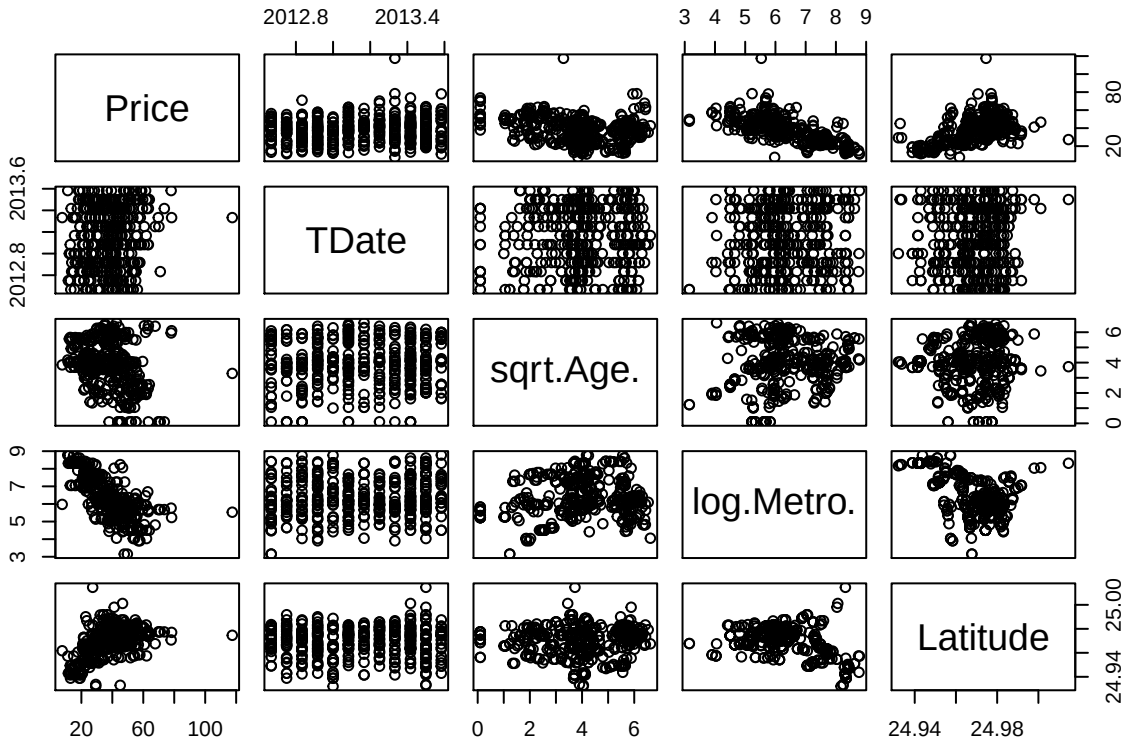
Here we can see all predictors except for Age and Metro contain 1, thus we must look to see what transformation of age and Metro we need.

```
testTransform(RE.pt, lambda = c(1, 0.5, 0, 1))
```

```
##                                     LRT df    pval
## LR test, lambda = (1 0.5 0 1) 5.530874  4 0.23703
```

Thus if we take the square root of Age and a log transformation of Metro,

```
sqrt_Age <- sqrt(Age)
RE_trsf = with(RealEstate, data.frame(Price, TDate, sqrt(Age), log(Metro), Latitude))
pairs(RE_trsf)
```

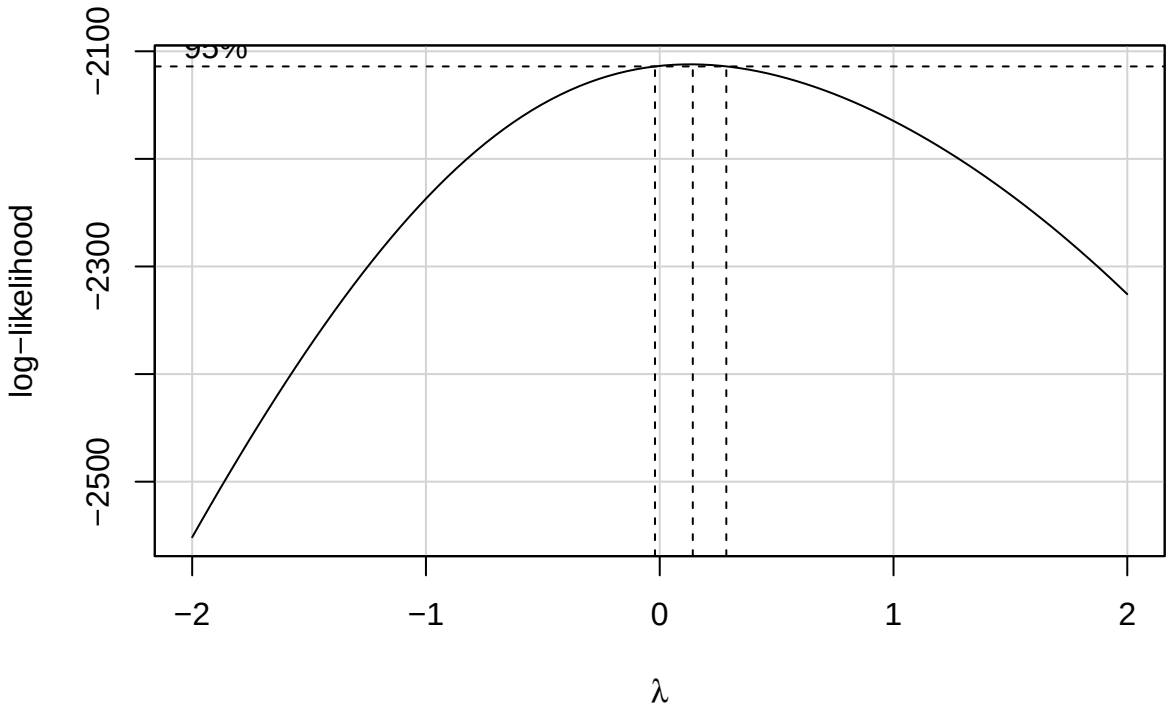


Now we can see linear relationships of all of the predictors with the response.

Transforming the Response

Now that we have found transformations for our predictors, we will look to see if we need to tranform our response.

```
boxCox(fit5)
```



Since we see that 0 is within our interval, we choose lambda = 0 and log transform our response Price. Hence, our model looks like log(Price) ~ TDate + sqrt(Age) + log(Metro) + Latitude.

```
final.fit <- lm(log(Price) ~ TDate  + sqrt(Age) + log(Metro) + Latitude)
summary(fit5)
```

```
##
## 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) -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
```

```
summary(final.fit)
```

```
##
## Call:
## lm(formula = log(Price) ~ TDate + sqrt(Age) + log(Metro) + Latitude)
##
## 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) -6.327e+02  7.539e+01  -8.392 7.87e-16 ***
## TDate        1.794e-01  3.664e-02   4.897 1.40e-06 ***
## sqrt(Age)   -4.780e-02  6.657e-03  -7.180 3.31e-12 ***
## log(Metro)  -2.050e-01  1.053e-02 -19.472  < 2e-16 ***
## Latitude     1.108e+01  9.356e-01  11.840  < 2e-16 ***
## ---
## 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
```

Comparing our fit from before our transformation on Price, Age, and Metro and after, we can see a large increase in our correlation coefficient. Before it was 0.5448, and after taking log(Price), sqrt(Age), and log(Metro), it is 0.7218. Hence we saw a large imporovemnet in our model after doing our transformations.

Summary

Based off of our regressions, we have found that, within our data set, the active regressors to calculate the expected log(price) of a house in Sindian District, New Taipei City, Taiwan, are best predicted by TDate, sqrt(Age), log(Metro), and Latitude. Some interesting results that we found in fitting our model was the fact that longitude was not significant in our model, given that metro. This could be becuase when you account for metro stations in the area, given the layout of Taiwan, on the Western shore, there are a lot of cities/urban areas that span the entire length of the country North to South. Since this entire Urban area is connected and runs longitude wise of the country, once we account for availability in transportation by metro in these areas, difference in longitude becomes less of an issue. Therefore, since movement is accounted for, difference in price does not seem to be determined by longitude location.

Part 2:

Here we will be looking at the data set 'Concrete' consisting of 8 predictors X1(Cement), X2(Blast Furnace Slag), X3(Fly Ash), X4(Water), X5(Superplasticizer), X6(Coarse Aggregate), X7(Fine Aggregate), X8(Age) and 1 response Y(Concrete Compressive Strength).

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

Forward Model Selection

To find our model we will use forward selection using BIC as a criterion function:

```
m0 <- lm(Y ~ 1)
full <- ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
n <- length(Y)

fit6 <- step(m0, full, direction = 'forward', k = log(n))
```

```
## 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
```

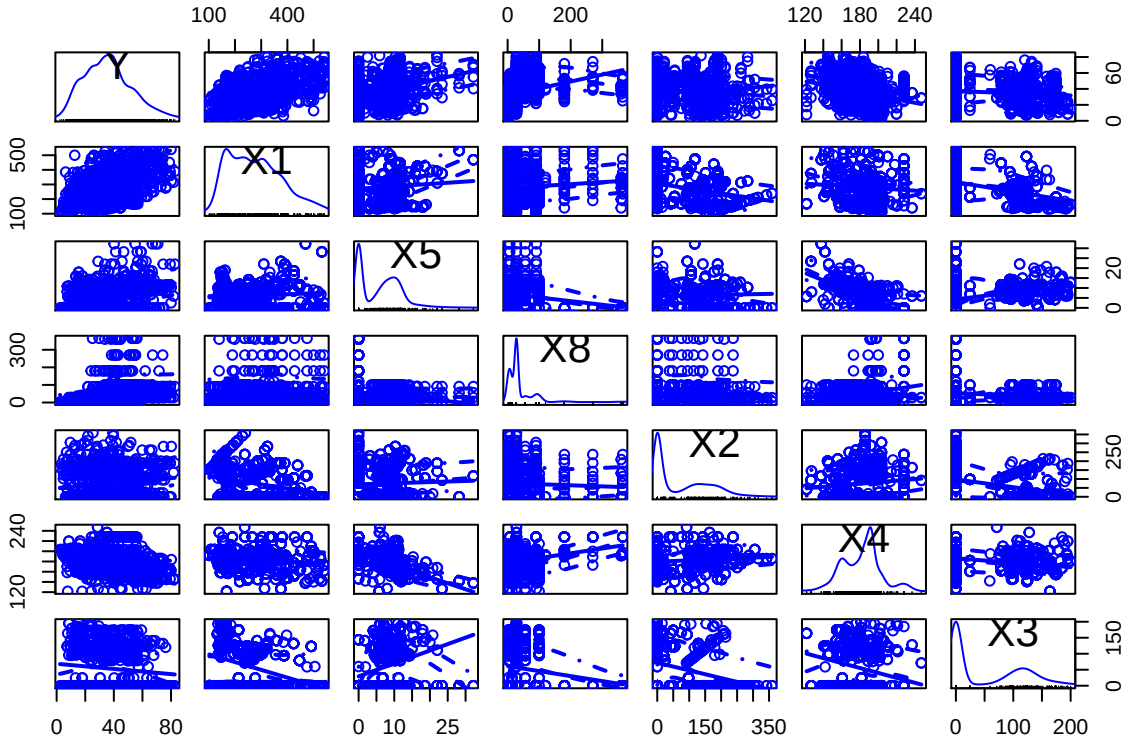
```
summary(fit6)
```

```
##
## 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
```

The result of forwards BIC model selection is the model: Y ~ X1 + X5 + X8 + X2 + X4 + X3.

Diagnostic Checks

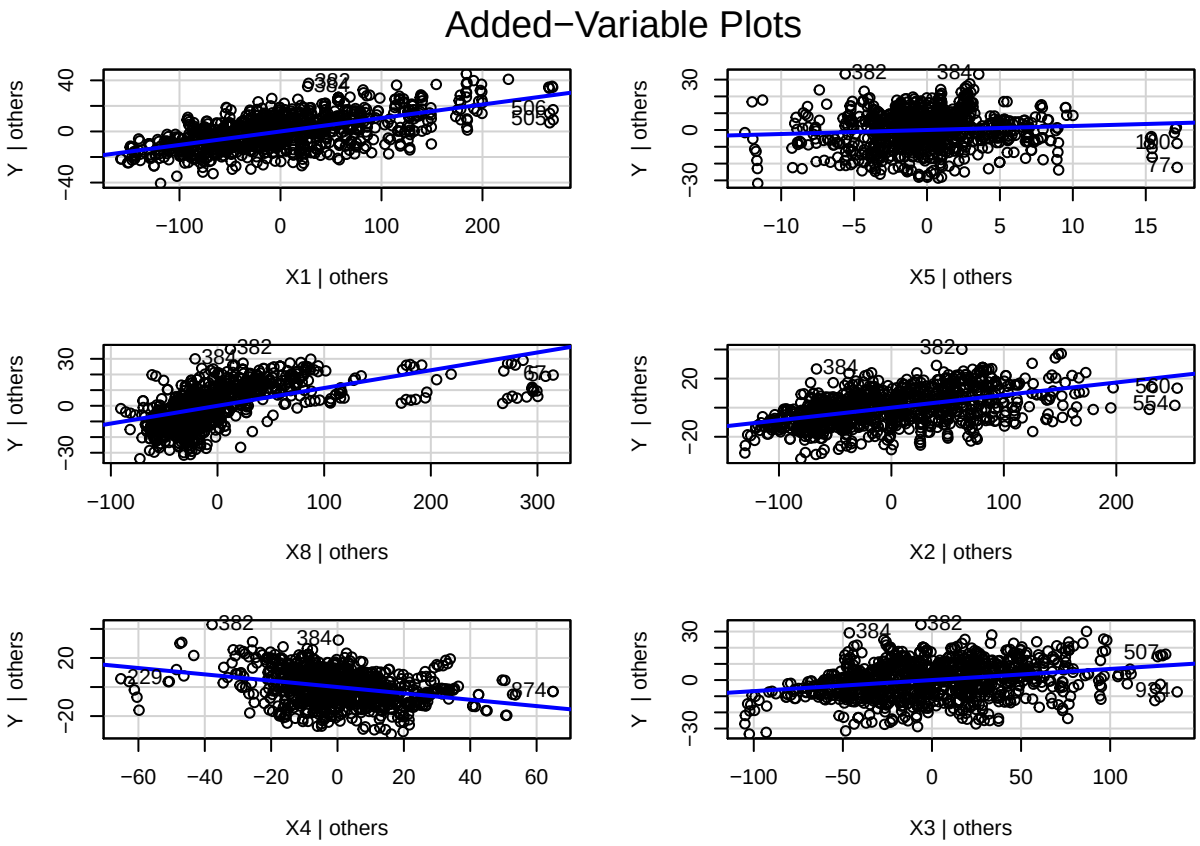
```
scatterplotMatrix(~ Y + X1 + X5 + X8 + X2 + X4 + X3)
```



Here is seems that the predictors all have a linear relationship with the response variable. Hence, we will continue with diagnostics to make sure our model assumptions hold.

Added Variable Plot

```
avPlots(fit6)
```

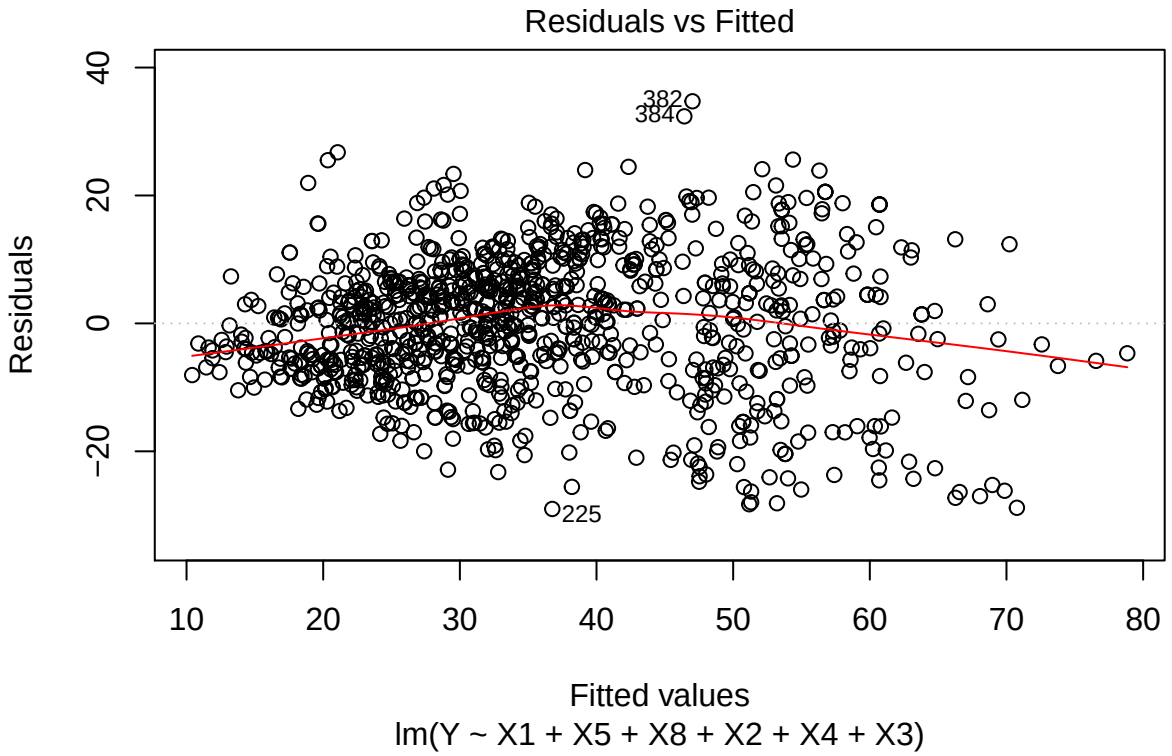


The marginal relationships are shown in the Added Variable plots: the plots regarding X1, X3, X2 seem like better linear fits, whilst the plot regarding X5 has a good looking linear, it may not tell us anything useful because the fit is around the the line Y|Others = 0. The plots regarding X4 and X8 seem to have a quadratic curve to them. We may or may not take a closer look at this later.

Now in order to test whole model assumptions, we will be looking at a Residual vs. Fitted plot to test linearly, scale-location to test constant variance, and QQ plots for normality of our data.

Residual Vs. Fitted for Linearity

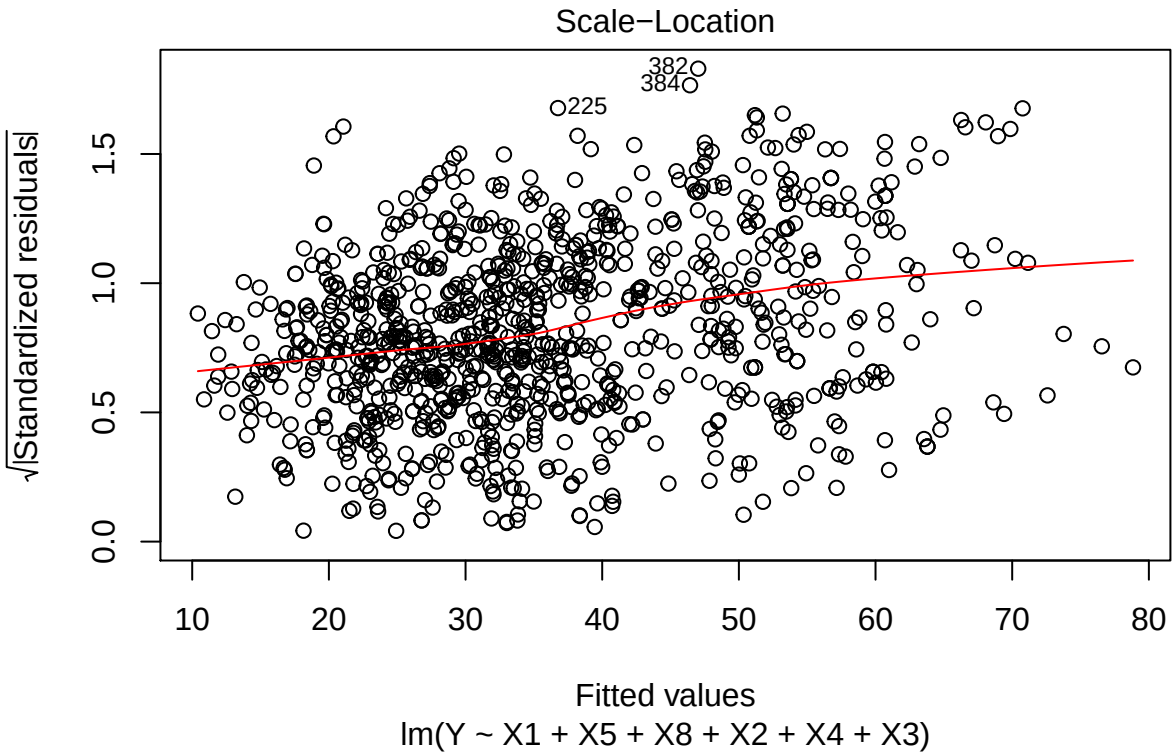
```
plot(fit6, which = 1)
```



From the Residual Vs. Fitted Values, we observe some clustering on the left side which leads to the conclusion that this does voilate the linearity assumption.

Scale-Location for constant variance

```
plot(fit6, which = 3)
```



Again we see some clustering on the left, thus we may see a voilation of constant variance in our model. To test to see if there is an effect of mean on the variance we will run a non-constant variance test.

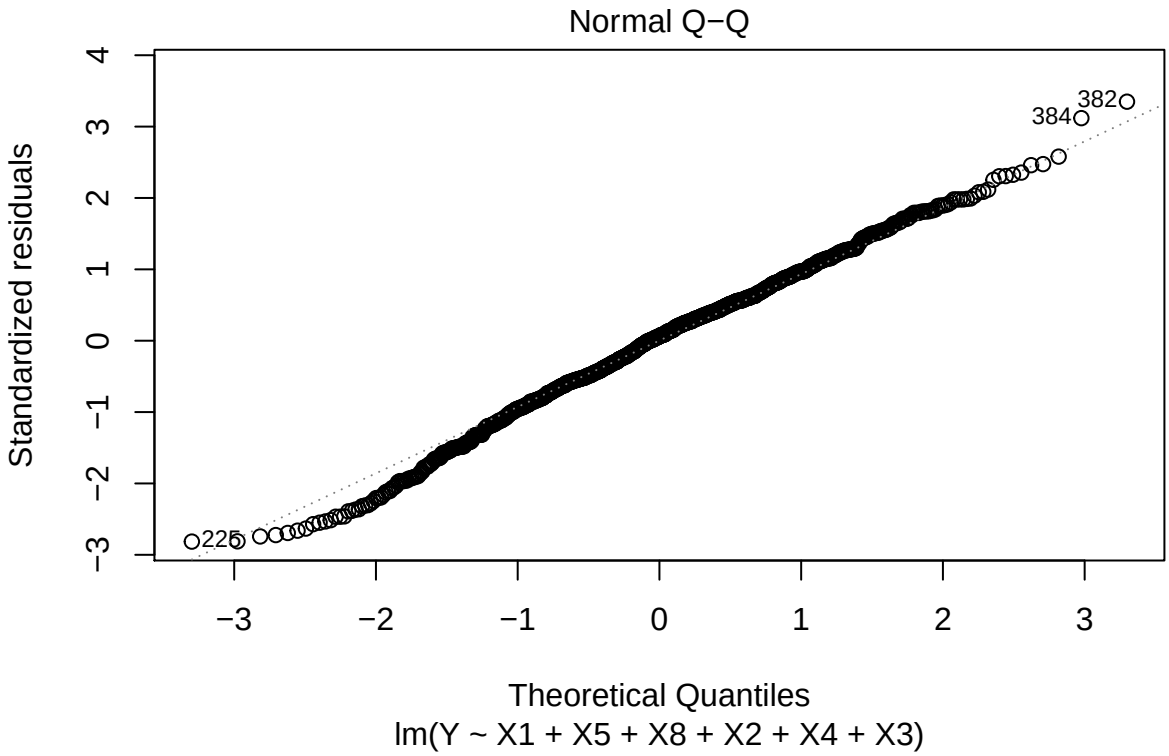
```
ncvTest(fit6)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 120.0133, Df = 1, p = < 2.22e-16
```

Thus, since we see a significant P-Value, we determine that variance increases as a function of the fitted values. Therefore in order to fix this issue of non-constant variance, we would have to put weights on the fitted values as 1/fitted values.

Q-Q Plot to Test Normality

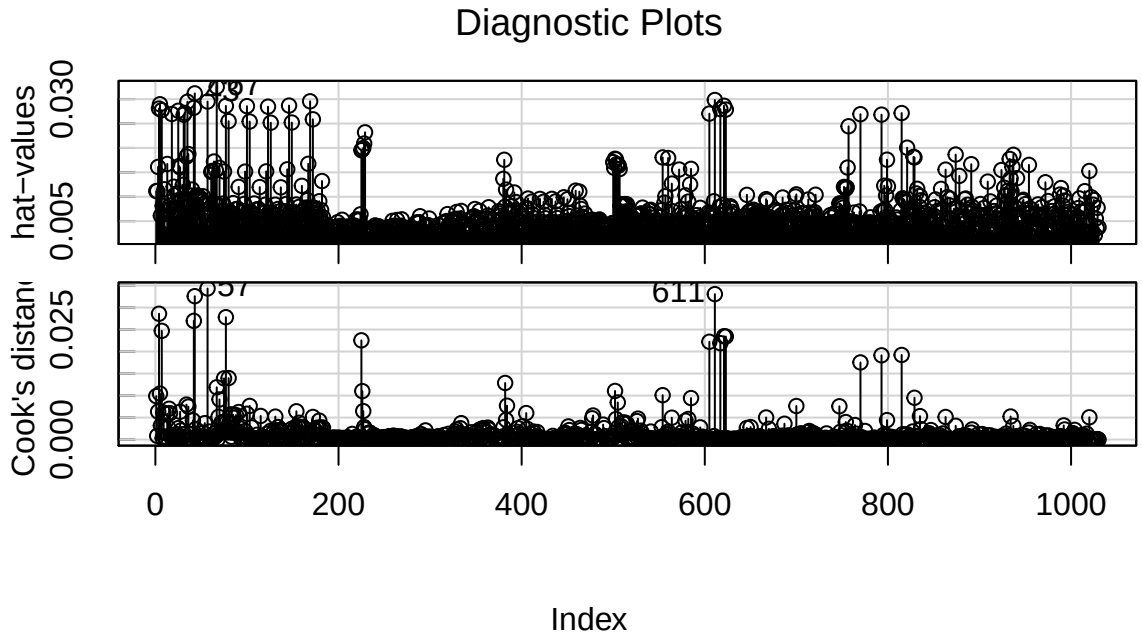
```
plot(fit6, which = 2)
```



It looks like observations 225, 384, and 382 may cause this to be a slightly heavy-tailed plot. Nonetheless, it violates the normality assumption since the plot is right-skewed.

Finding Influential Points

```
influenceIndexPlot(fit6, vars = c('hat', 'cook'))
```



Here we see the data point 57 has a high cooks distance as well as a high hat value/leverage, thus it is the

most influential point. Similarly point 611 has a large cooks distance. Since these points have high influence, we may consider either adding predictors in our model to account for these points, or even removing them altogether.

Estimating the Mean Response

Suppose that we want to predict the mean compressive strength of concrete with the makeup of 149kg per meter cubed of cement, 118kg per meter cubed of blast furnace slag, 92kg per meter cubed of fly ash, 183kg per meter cubed of water, 7kg per meter cubed of superplasticizer, and 28 days old.

```
x.new <- data.frame(X1 = 149, X2 = 118, X3 = 92, X4 = 183, X5 = 7, X8 = 28)
predict(fit6, newdata = x.new, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 26.16971 25.11649 27.22294
```

We see an estimated mean concrete compressive strength of 26.17 MPa with a lower bound of 25.12 MPa and an upper bound of 27.22 MPa with 95% confidence.

Predicting New Responses

Now looking at the same concrete as above, we want to predict that compressive strength that this makeup gives us for all concrete of this makeup.

```
predict(fit6, newdata = x.new, interval = "predict", level = 0.95)
```

```
##        fit      lwr      upr
## 1 26.16971 5.715607 46.62381
```

Thus we can predict that the compressive strength for all concrete with this makeup will lie between 5.72Mpa and 46.62 MPa with 95% confidence.

Backward Model Selection

Now we will apply the backward model selection using the BIC criterion. We start with our full model then compare to see if a submodel would be a better fit.

```
m1<-update(m0,full)

fit7 <- step(m1, scope = c(lower= ~X1), direction = 'backward', k = log(n))
```

```
## 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
## - X8    1     47905 158333 5241.7
##
## Step:  AIC=4874.12
```

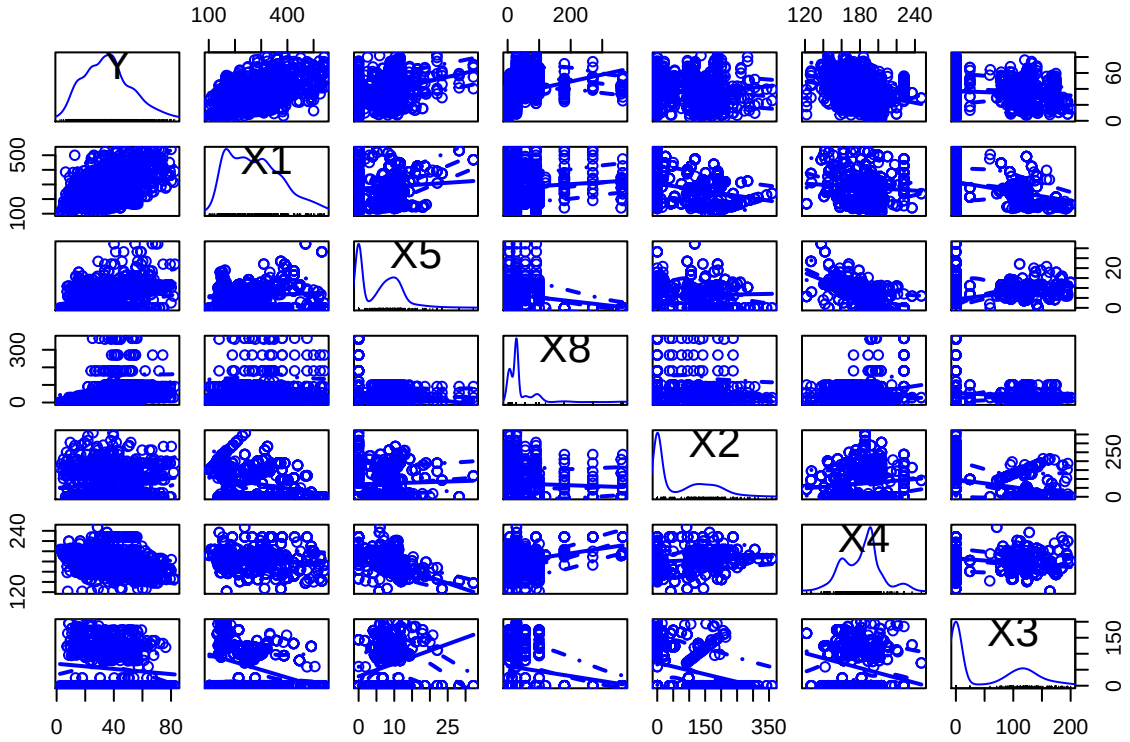
```
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X8
##
##        Df Sum of Sq    RSS    AIC
## - X6    1        44 110856 4867.6
## <none>               110812 4874.1
## - X5    1       877 111688 4875.3
## - X4    1      8526 119338 4943.5
## - X3    1      8568 119379 4943.9
## - X2    1     30693 141505 5119.0
## - X8    1     47522 158334 5234.8
##
## 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
```

The result of the backwards BIC model selection is the model: Y ~ X1 + X5 + X8 + X2 + X4 + X3. Through the backwards selection process, we have removed the predictors "Course Aggregate" (X6) and "Fine Aggregate" (X7) and have found that a smaller model is a better fit with lower BIC.

Diagnostic Checks

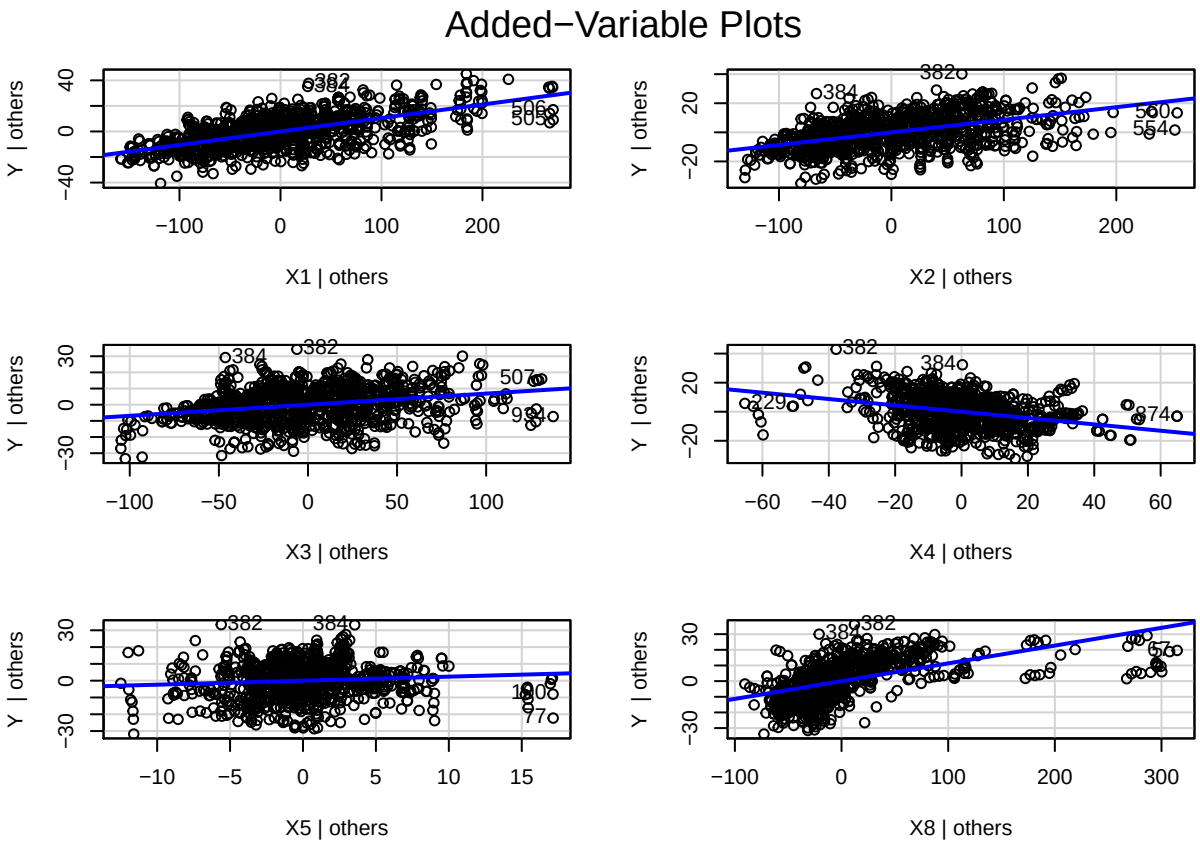
First, we will look at the scatterplot matrix for our model to determine whether it is reasonable to use linear regression to analyze the data.

```
scatterplotMatrix(~ Y + X1 + X5 + X8 + X2 + X4 + X3)
```



The scatterplot matrix shows that all the predictors have some sort of relationship with Cement Compressive Strength (MPa), therefore, it is reasonable to analyze this data with regression. We will continue to run diagnostics on our model by looking at the Added Variable Plots.

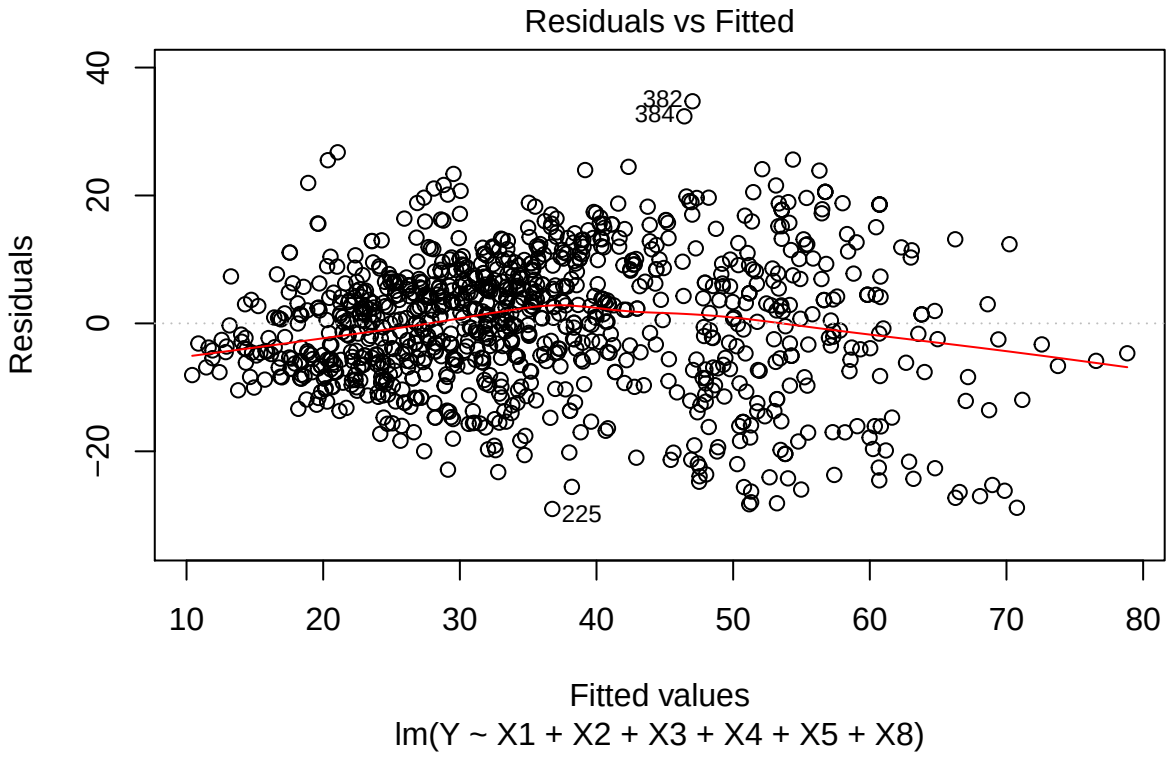
```
avPlots(fit7)
```



We can see marginal relationships from the added variable plots for each predictor, given all others are held constant. Clearly, they all seem to have a linear relationship of some sort. However, let it be noted that Y~X8|others and Y~X4|others appear to have a slight quadratic curve to it. The slopes of these graphs gives the estimated coefficient for their respective predictors (given all others are held constant).

We will continue to run diagnostics by using residual plots to check our assumptions of the model. First, we will look at our Residual vs. Fitted plot to assess our assumption of linearity.

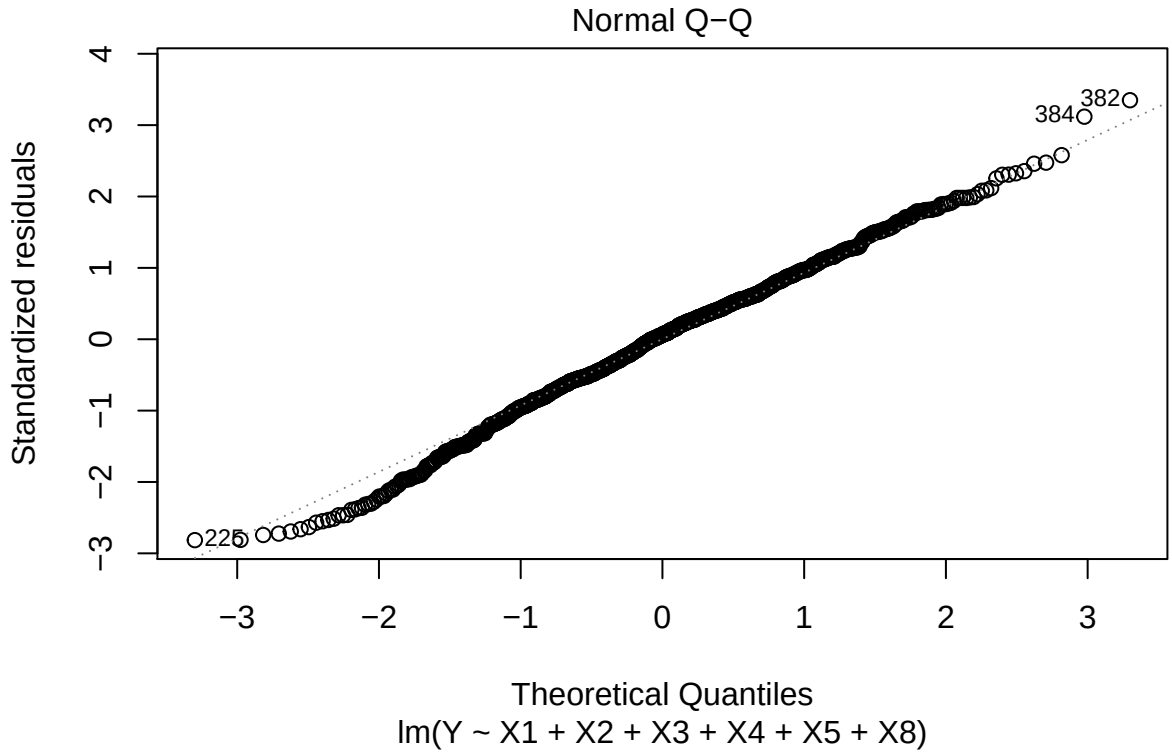
```
plot(fit7, which = 1)
```



The Residual vs. Fitted plot shows slight clustering on the left which leads us to believe that it is not as linear as we would like it to be.

Next, we will look at a QQ-plot to assess our assumption of normality.

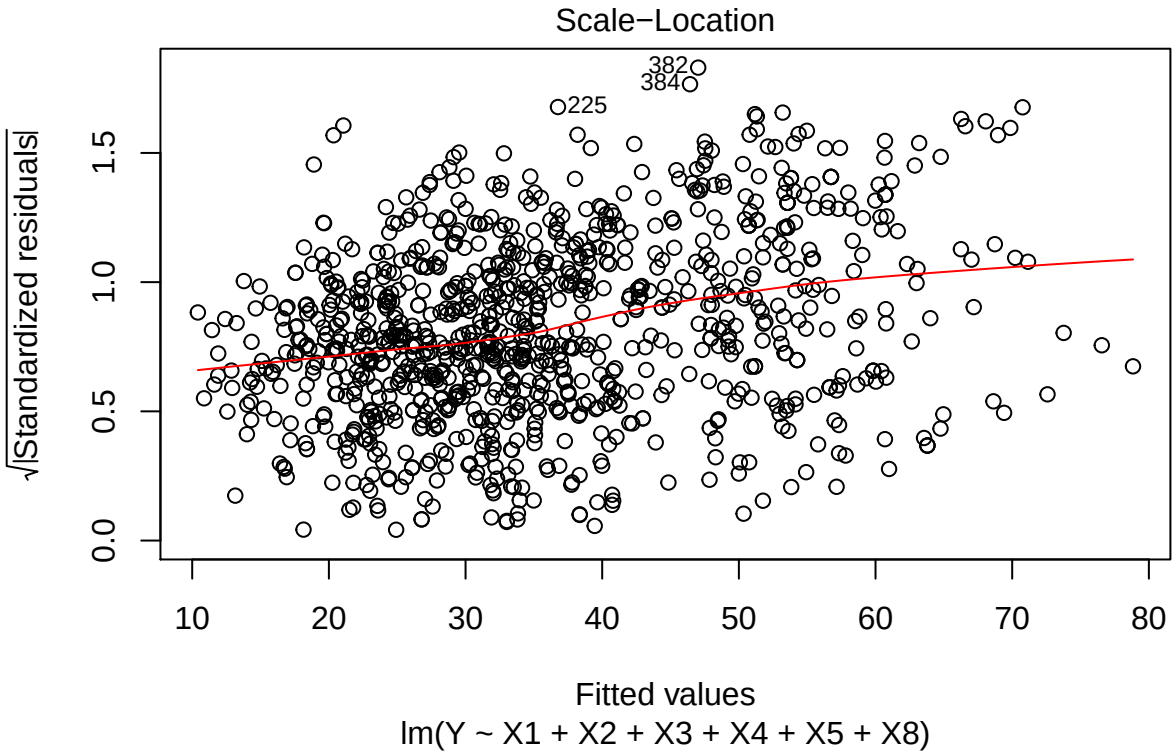
```
plot(fit7, which = 2)
```



There does not seem to be any heavy tails in our QQ plot but the plot appears to be right-skewed. This violates our assumption of normality.

Finally, we will take a look at a Scale-Location plot to determine constant variance.

```
plot(fit7, which = 3)
```

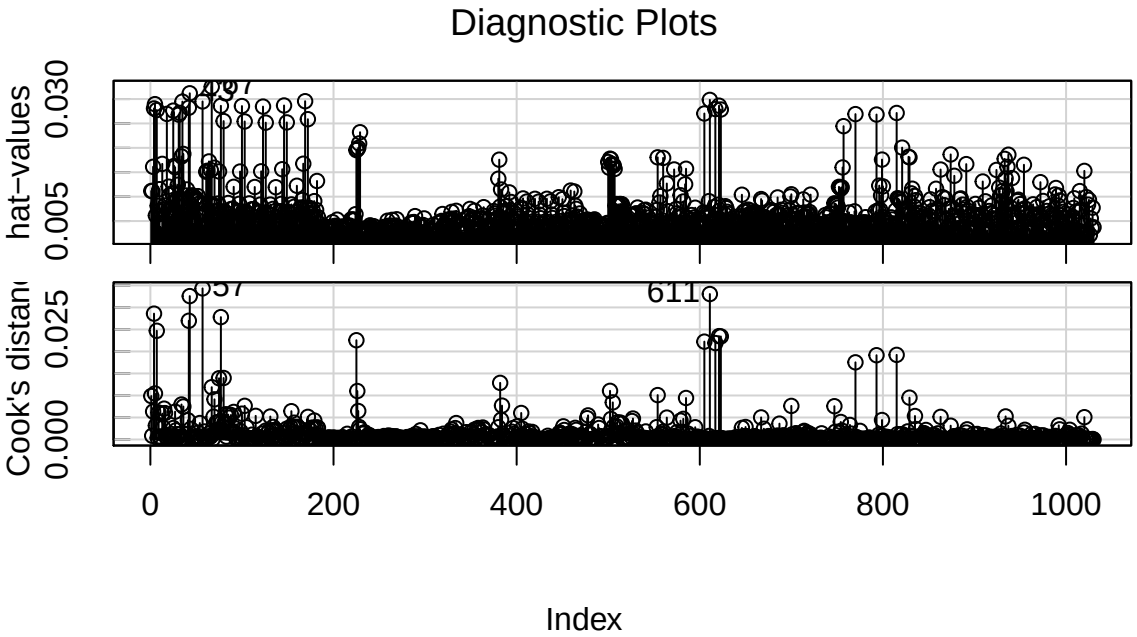


The scale-location plot also shows slightly more clustering on the left than right and seems to slowly spread out from left to right; this violates the constant variance assumption.

Finding Influential Points:

We will look at influence by examining the hat-values and Cook's distance:

```
influenceIndexPlot(fit7, vars = c('hat', 'Cook'))
```



From the two diagnostic plots, we can see that observation 57 has large leverage and Cook's distance, thus it is the most influential point. Likewise, Observation 611 also has a fairly high Cook's distance. Because these points are significantly influential, we may want to consider adding predictor(s). We may even want to consider removing those observations.

When using BIC as the criteion function, both the forward selection and backwards elination algorithm yielded the same model.

Summary

We took a model using BIC criterion and applied both forwards and backwards selection algorithms to find the best model for predicting Concrete Compressive Strength (MPa). After finding the best model, Y ~ X1 + X5 + X8 + X2 + X4 + X3, which was the same for forwards and backwards selection (but had different ordering), we checked linearity, constant variance, and normality with residual and QQ plots. We found some leverages and influences and decided what data to remove to better the model. We then calculated a confidence and prediction interval which we interpretated directly after.

Conclusion

In summary, for the real estate data set, the best model fit we found was log(Price) = -6.327e02 + 1.794e-01(TDate) – 4.78e-02(sqrt(Age)) – 2.05e-01(log(Metro)) + 1.108e01(Latitude). Thus, we can see a negative association with the age of the house, meaning that the price of the house declines with age, as expected. However, a more interesting finding is that closeness to a metro station also declines the price of the house. The generalizability of this model may be questionable due to the fact that the price is solely determined by location of the house. This model does not take into account for the size or characteristics (number of beds/baths and square footage) of the house. For the concrete data set, the best model fit we found was Y = 29.03 + 0.105(X1) + 0.068(X2) + 0.069(X3) - 0.218(X4) + 0.239(X5) + 0.113(X8). Thus we see that concrete compressive strength decreases for every 1kg per meter cubed of water. Therefore, for every kg per meter cubed of cement, blast furnace slag, fly ash, superplasticizer adds compressive strength, well as number of days old. The reliability of this model may be questionable because the relationship between concrete compressive strength and the amount of fly ash and blast furnace slag should be negative, however, we found a positive association.

Figure 1: Thanks For Reading