PSTAT 126Final Project Option 2

Kelcie Opp and Emma Furr

6/2/2019

Abstract:

In this project, we are conducting a regression analysis of a real dataset. Project option 2 asks us to analyze two different datasets. The first data set is real estate valuation and it has a response variable, Price, and six predictors. This data is modelling the relationship between these predictors and how they effect on the house price of unit area. We are investigating the relationship between the response and all of the predictors as well as which combination of predictors will give us the best and most accurate response.The second data set models the relationship between the concrete compressive strength and eight concrete components. We are asked to apply forward and backward selection to find the best fitted model for the data. Once we have created the best fitted model, we will test whether there are any outliers or influential points. As well as estimate a mean response and predict a new response.

Problem and Motivation:

The Real Estate Valuation data set contains one response and six predictors. We are modeling to see how the transaction date, the house age, the distance to the nearest MRT station, the number of convenience stores, and the two geographic coordinates affect the house price of unit area. Readers should be interested in this data if they are planning on purchasing a house in Sindian District, New Taipei City, Taiwan. This regression model will give buyers an accurate price of a home based on the predictor values.The Concrete data contains one response and eight predictors. We are modeling to see how cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age affect the concrete compressive strength. Readers should be interested in this data if they are planning to purchase or use concrete for their own purposes or company. The strength of a concrete will be important for someone who needs it for different purposes. For example, a stronger concrete will be important for someone who is using it for intense building purposes. This project is meant to fit the best model for these two data sets.

Data:

The first data set, Real Estate Valuation, contains the variables: Price, Longitude, Latitude, Stores, Metro, Age, and TDate. We are looking for the best predictors for our response variable Price. When asked to decide between models 1 and 2 it is evident that the more precise model and the one that we prefer will be model 2 that contains the predictors TDate, Age, Metro, and Latitude. This is the model that will be analyzed for the project. All of the testing and analysis for this is show in Section 7(Regression Analysis, Results and Interpretation). However, at the end of this project after analyzing model 2, we tested the best overall model with the stepwise function and it gave the results of using the predictors: TDate, Age, Metro, Latitude, and Stores. The second data set, Concrete Compressive Strength, contains the variables: X1 (Cement), X2 (Blast Furnace Slag), X3 (Fly Ash), X4 (Water), X5 (Superplasticizer), X6 (Coarse Aggregate), X7 (Fine Aggregate), X8 (Age), Y (Concrete compressive strength). We are looking for the best predictors for our response variable Y. The relevant variables that we will use to predict our response are X1, X2, X3, X4, X7, X8. These variables that excluded X5 and X6 came from applying the stepwise function.

Questions of Interest:

In the first part, we are given a model from the dataset (Price ~ TDate + Age + Stores + Latitude) and asked to find the relationships between the response and each of the four predictors. We want to figure out what kind of associations are present. After we fit the model, we are asked to conduct tests on each predictor to interpret if it is significant to the model. We want to see if there is a stronger relationship between the response and predictor when Metro and/or Longitude is added to the plot. Then, we are given a new model (Price ~ TDate + Age + Metro + Latitude), and asked to compare it to the first model to see which is a better fit. After we have chosen the best model, we are asked to investigate whether we should transform some of the predictors and/or the response. We want to see whether transforming any variable improves the model and creates a better fit. We then want to summarize our overall analysis and findings. In the second part, we are given a full model (Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8) and asked to apply the forward selection algorithm, using BIC. Beginning with the smallest model, the forward selection will add variables one at a time until the chosen information criterion cannot be decreased anymore. After the best fitted model is produced, we begin diagnostic checks to assess if the assumptions hold. We also want to look for any influential points to possibly remove. We, then, are asked to estimate the mean response, as well as predict a new mean response for the predictor values. Then, we are asked to apply the backward selection algorithm, using BIC. Beginning with the largest model, we remove one at a time until the information cannot be decreased. We want to check the diagnostics and influential points again. Choosing the best final model, we want to analyze and comment on it.

Regression Methods:

Different testing and analysis methods that we used were: summary tables, avPlots, Anova tables, Residual tables (checking assumptions), AIC, Step-wise, BIC, Power Transformation, Cooks, boxCox (Special case for power transformation), Outlier test, influenceIndexPlot(To test outliers with Cook's distance and high leverage).

Regression Analysis, Results, and Interpretation:

Part 1:Real Estate Valuation

```
library(alr4)
```

```
## Loading required package: car
```

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

```
## Loading required package: effects
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
setwd("~/Desktop")
RealEstateValuation<- read.table("RealEstateValuation.txt", header=TRUE)
```

Question 1

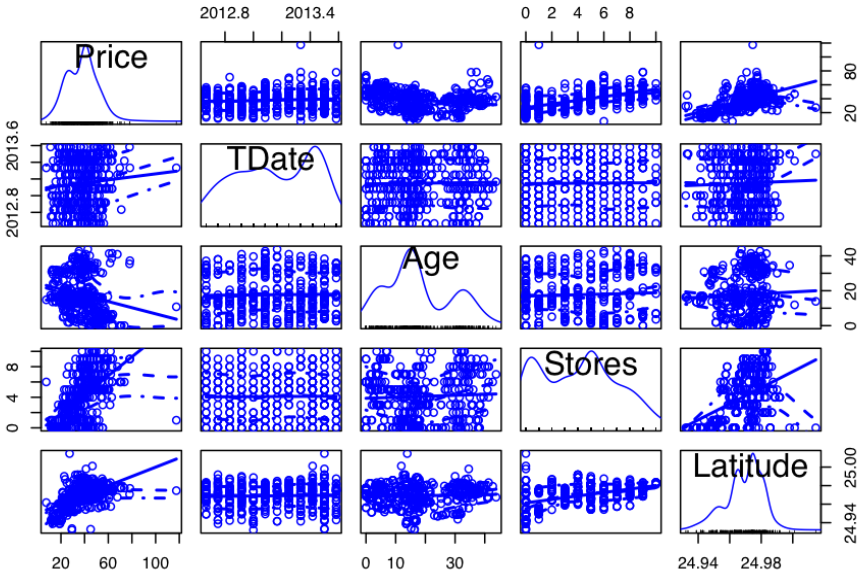
A

Looking at the anova table it seems like all the predictors will be statistically significant given that they all have a p-value less than 0.05. TDate and Stores are both categorical variables. We can observe that Price and Age have a negative correlation, Price and Latitude have a positive correlation, and Age and Latitude have a weak positive correlation. The partial regression plots are used to see the conditional relationship. Age, Stores, and Latitude seem like they will be useful predictors for Price because of their slopes. Age has a negative slope and Stores and Latitude have positive slopes. TDate seems to have no slope and this could be because it is a categorical variable, so it may turn out to be significant.

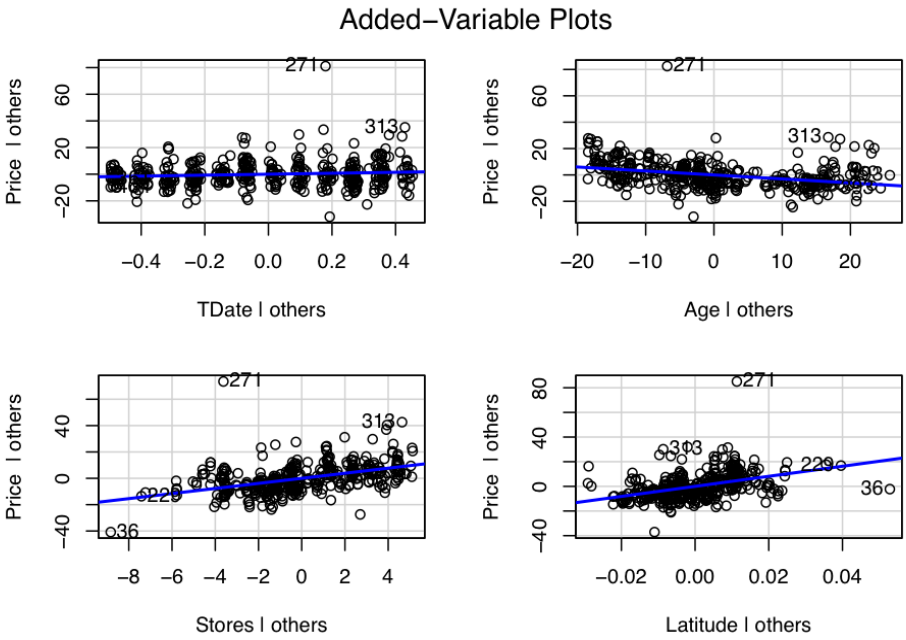
```
Real.lm<-lm(Price~TDate+Age+Stores+Latitude, data=RealEstateValuation)
anova(Real.lm)
```

```
## Analysis of Variance Table
##
## Response: Price
##             Df Sum Sq Mean Sq  F value    Pr(>F)
## TDate        1    585   585.3   6.2797    0.0126 *
## Age          1   3441  3440.9  36.9184 2.822e-09 ***
## Stores       1  25845 25844.7 277.2982 < 2.2e-16 ***
## Latitude     1   8471  8471.1  90.8899 < 2.2e-16 ***
## Residuals  409  38119    93.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
scatterplotMatrix(~Price+TDate+Age+Stores+Latitude, data=RealEstateValuation)
```




```
avPlots(Real.lm)
```



B

The equation for the fitted regression line is E(Price | TDate(X1), Age(X2), Stores(X3), Latitude(X4))= -17420 + 3.613(X1) - 0.3020(X2) + 1.929(X3) + 407.8(X4)

It can be seen that the coefficient of TDate is not significantly different from zero at any conventional level of significance because the p-value from the summary table is 0.0327 which is greater than alpha=0.01 The coefficient of Age, Stores, and Latitude are significantly different from zero at any conventional level of significance because the p value is very small. The expected change in Price is -0.3020 when Age increases by one unit and all else constant. The expected change in Price is 1.929 when Stores increases by one unit and all else constant. The expected change in Price is 407.8 when Latitude increases by one unit and all else constant.

```
Real.lm<-lm(Price~TDate+Age+Stores+Latitude, data=RealEstateValuation)
summary(Real.lm)
```

```
##
## Call:
## lm(formula = Price ~ TDate + Age + Stores + Latitude, data = RealEstateValuation)
##
## 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
```

C

Let B5 be the estimate for Metro and B6 be the estimate for Longitude. The null hypothysis is H0: B5=B6=0 such that the model without Metro and Longitude is significant versus H1: One of the Betas is not equal to zero such that the model with Metro and Longitude is significant. The p-value is significantly small with an F-statistic from the anova table of 39.428 on 2 degrees of freedom, so we reject the null and conclude that the fitted regression model including Metro and Longitude is significant. It is now necessary to check if we need both of these predictors, or if only one is sufficient. Using the anova table with the full model against a model without Longitude as well as an anova table with the full model against one without Metro, it is clear that with Metro in the model, adding Longitude is not useful. The p-value from the anova table with adding longitude with Metro already in the model is 0.7983 which is greater than alpha = 0.05.

```
Real.lm<-lm(Price~TDate+Age+Stores+Latitude, data=RealEstateValuation)
MetroLong.lm<-lm(Price~TDate+Age+Stores+Latitude+Metro+Longitude, data=RealEstateValuation)
summary(MetroLong.lm) #summary of the full model
```

```
##
## Call:
## lm(formula = Price ~ TDate + Age + Stores + Latitude + Metro +
##     Longitude, data = RealEstateValuation)
##
## 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
```

```r
anova(Real.lm,MetroLong.lm) #anova between full model and original model
```

```
## 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
```

```r
Metro.lm<-lm(Price~TDate+Age+Stores+Latitude+Metro, data=RealEstateValuation)
Long.lm<-lm(Price~TDate+Age+Stores+Latitude+Longitude, data=RealEstateValuation)
anova(Metro.lm,MetroLong.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ TDate + Age + Stores + Latitude + Metro
## Model 2: Price ~ TDate + Age + Stores + Latitude + Metro + Longitude
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    408 31938
## 2    407 31933  1    5.1308 0.0654 0.7983
```

```r
anova(Long.lm,MetroLong.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ TDate + Age + Stores + Latitude + Longitude
## Model 2: Price ~ TDate + Age + Stores + Latitude + Metro + Longitude
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    408 34997
## 2    407 31933  1    3064.5 39.059 1.039e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

D

The equation for the fitted regression line for the second model is E(Price | TDate(X1), Age(X2), Metro(X3), Latitude(X4)= -17670 + 5.57X1 - 0.253X2 - 0.00576X3 + 260.7X4.

We can use the anova table to compare the model with only the three predictors TDate, Age, and Latitude to the model with the added variable Metro and the model with the added variable Stores to see if both of these could work. Since both models seem reasonable, we can compare the models using the AIC function and choose the model with the smaller AIC. The model Real.lm has an AIC of 3059.244 and Real.lm2 has an AIC of 3021.623. This leads us to choose the model Real.lm2 that contains the predictors TDate, Age, Metro, and Latitude.

```r
Real.lm2<-lm(Price~TDate+Age+Metro+Latitude, data=RealEstateValuation)
#We can look at the overall summary of both
summary(Real.lm2)
```

```
##
## Call:
## lm(formula = Price ~ TDate + Age + Metro + Latitude, data = RealEstateValuation)
```

```
##
## 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(Real.lm)
```

```
##
## Call:
## lm(formula = Price ~ TDate + Age + Stores + Latitude, data = RealEstateValuation)
##
## 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
```

```
Real.lmpart<-lm(Price~TDate+Age+Latitude, data=RealEstateValuation)
anova(Real.lmpart, Real.lm2)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ TDate + Age + Latitude
## Model 2: Price ~ TDate + Age + Metro + Latitude
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    410 48815
## 2    409 34808  1     14007 164.58 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(Real.lmpart, Real.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ TDate + Age + Latitude
## Model 2: Price ~ TDate + Age + Stores + Latitude
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    410 48815
## 2    409 38119  1     10696 114.76 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Now we can compare the AIC of both models
AIC(Real.lm2) #AIC 3021.623
```

```
## [1] 3021.623
```

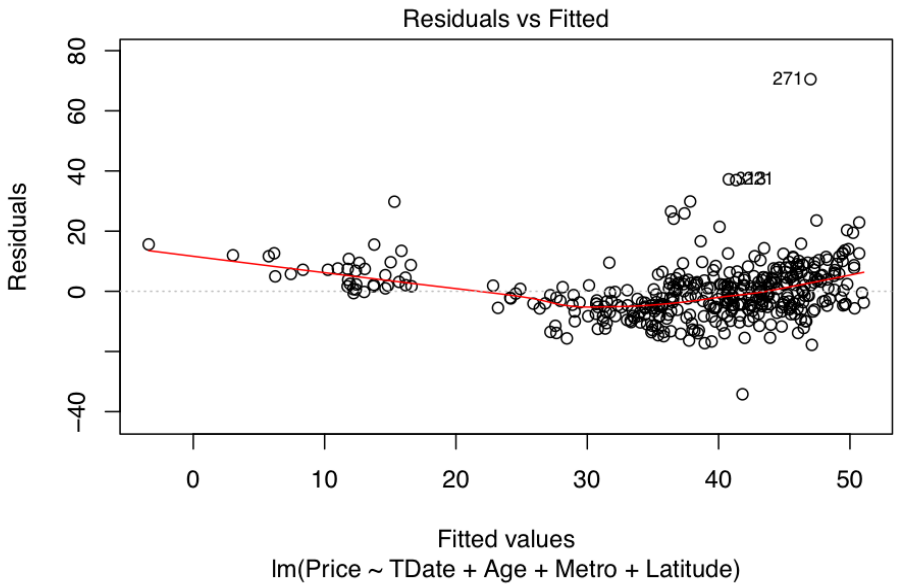
```r
AIC(Real.lm) #AIC 3059.244
```

```
## [1] 3059.244
```

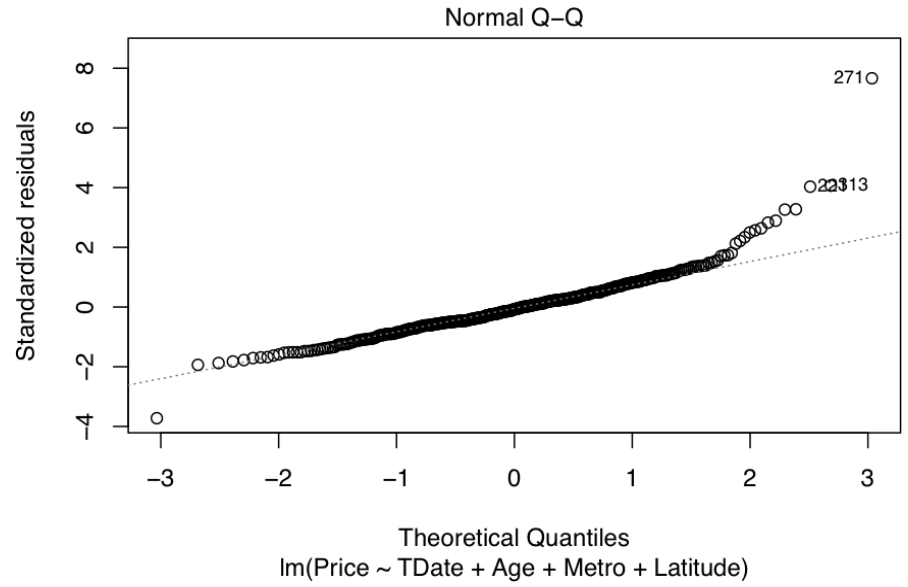
E

Looking at our model before making any transformations, we have heavy tails in the QQ-plot which show that there is non-normality and we have a violation of linearity and/or constant variance shown in the residuals versus fitted plot. In the avPlot, TDate may look insignificant because of its slope but it looks like this because it is a categorical variable. In order to find all of our transformations, we need to add one to Age and one to Metro so our data does not contain any zeros. Then, we can conduct a power transformation test. We want a log transformation for Metro1 because the rounded power from the power Transformation function shows a zero. We want a square root transformation for Age1 because the rounded power shows a 0.5. For TDate and Latitude we do not need a transformation because the rounded power is 1. The Box Cox function shows us that we want a log transformation for Price because 0 is extremely close to being in the interval. Looking at the model after we have made the transformations it is evident that linearity and constant variance have improved, but normality is still violated because the QQ-plot is heavy tailed.

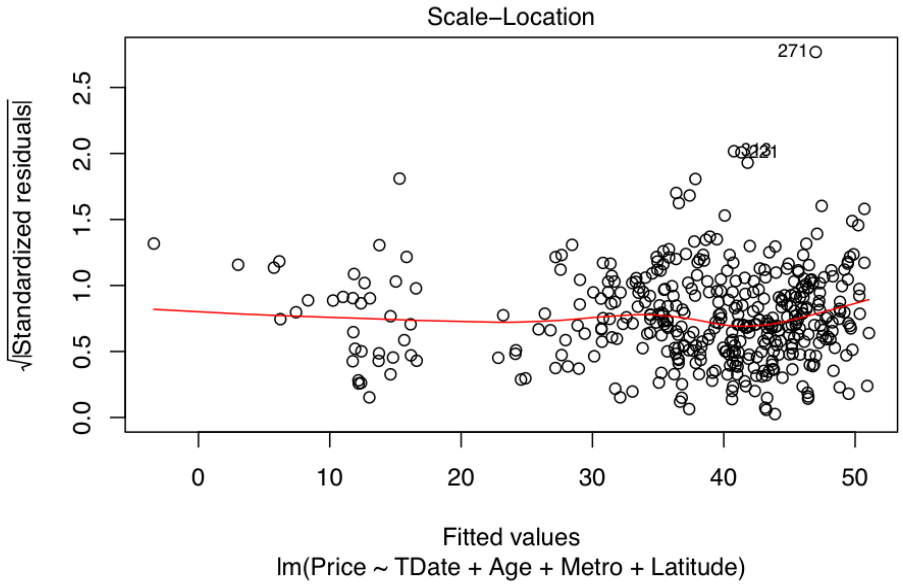
```r
plot(Real.lm2,which=1)
```



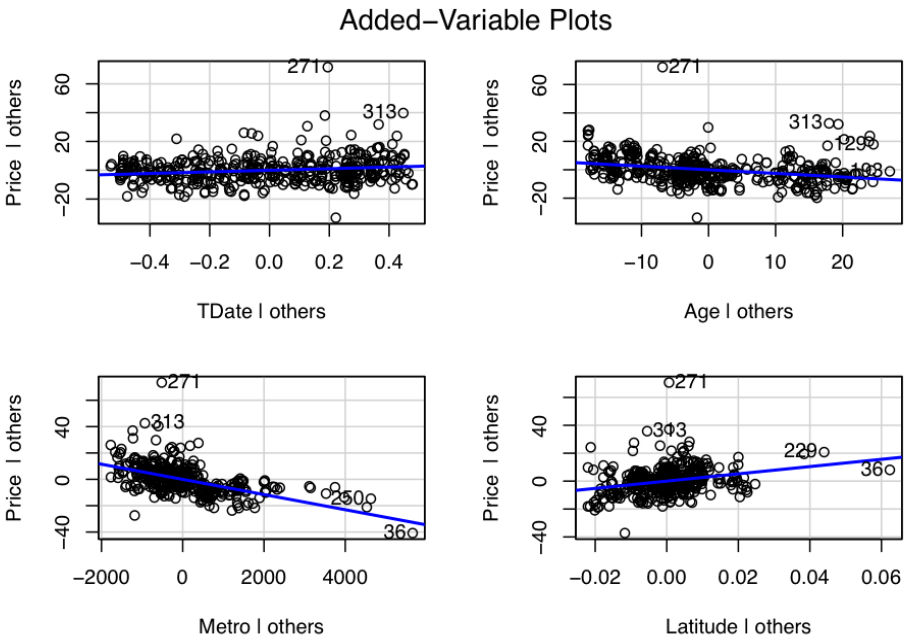

```
plot(Real.lm2,which=2)
```




```
plot(Real.lm2,which=3)
```




```
avPlots(Real.lm2)
```

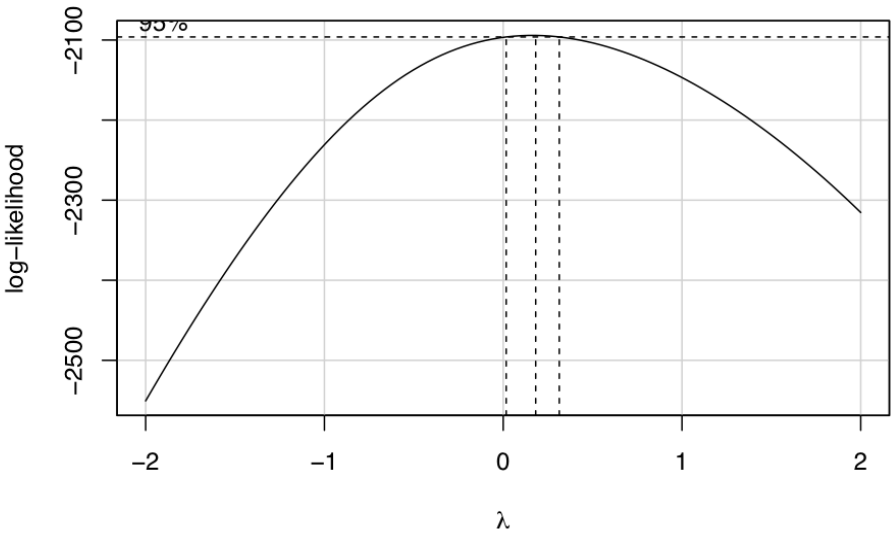



```
Age1=RealEstateValuation$Age+1
Metro1=RealEstateValuation$Metro+1
```

```
Real.pt=powerTransform(cbind(TDate,Age1,Metro1,Latitude)~1, RealEstateValuation)
summary(Real.pt)
```

```
## bcPower Transformations to Multinormality
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## TDate        3.0000         1.0    -261.9137     267.9137
## Age1         0.5510         0.5       0.4389       0.6631
## Metro1       0.0755         0.0      -0.0050       0.1560
## Latitude     3.0000         1.0    -146.8659     152.8660
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                   LRT df       pval
## LR test, lambda = (0 0 0 0) 116.537  4 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                    LRT df       pval
## LR test, lambda = (1 1 1 1) 506.5595  4 < 2.22e-16
```

```
boxCox(lm(Price~., data=RealEstateValuation))
```

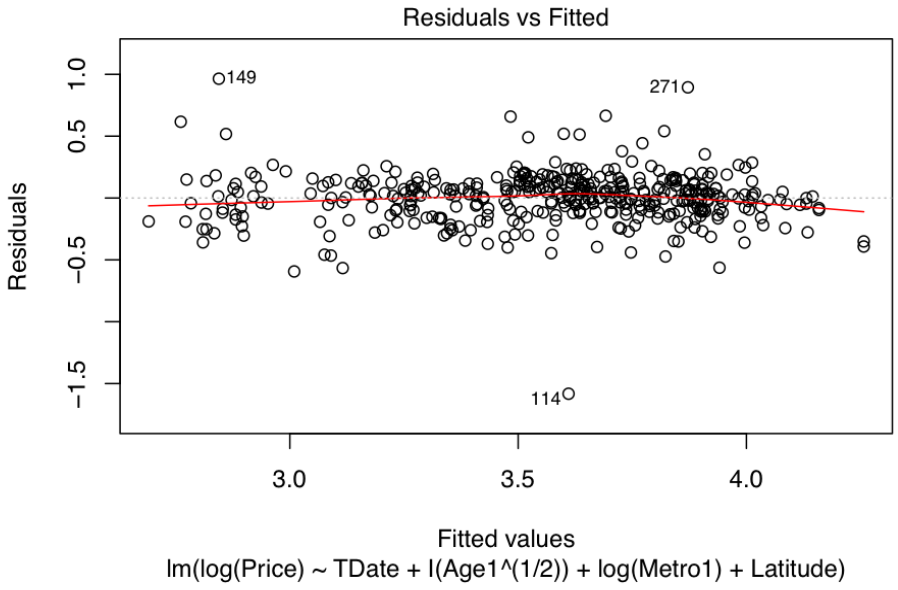



```
#looking at the model after we have made the transformations:
Real.lm2trans<-lm(log(Price)~TDate+I(Age1^(1/2))+log(Metro1)+Latitude, data=RealEstateValuation)
summary(Real.lm2trans)
```

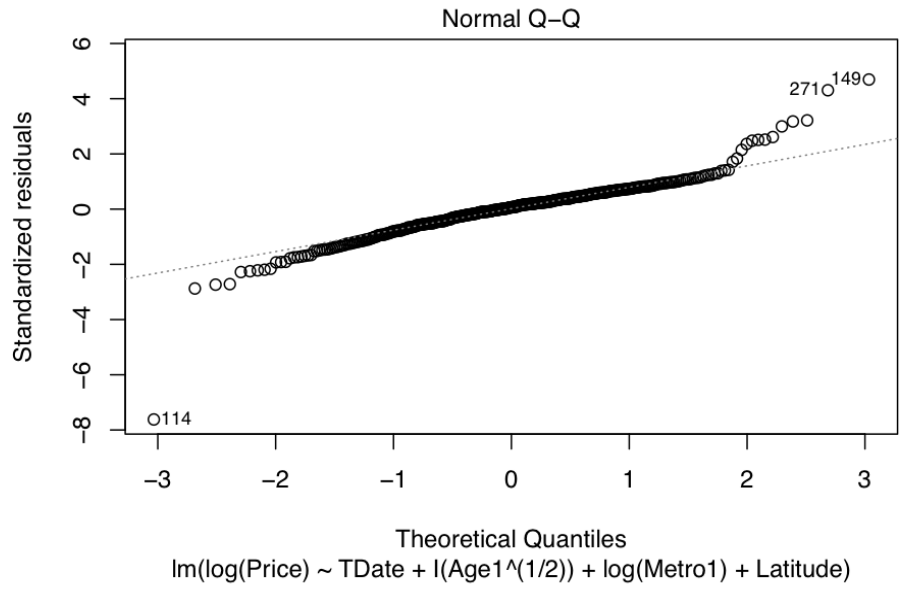
```
##
## Call:
## lm(formula = log(Price) ~ TDate + I(Age1^(1/2)) + log(Metro1) +
##     Latitude, data = RealEstateValuation)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58230 -0.10512  0.01351  0.11151  0.96447
```

```
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.280e+02  7.547e+01  -8.321 1.32e-15 ***
## TDate          1.774e-01  3.668e-02   4.837 1.87e-06 ***
## I(Age1^(1/2)) -5.137e-02  7.261e-03  -7.074 6.56e-12 ***
## log(Metro1)   -2.068e-01  1.055e-02 -19.592  < 2e-16 ***
## Latitude       1.105e+01  9.373e-01  11.790  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2083 on 409 degrees of freedom
## Multiple R-squared:  0.721,  Adjusted R-squared:  0.7183
## F-statistic: 264.3 on 4 and 409 DF,  p-value: < 2.2e-16
```

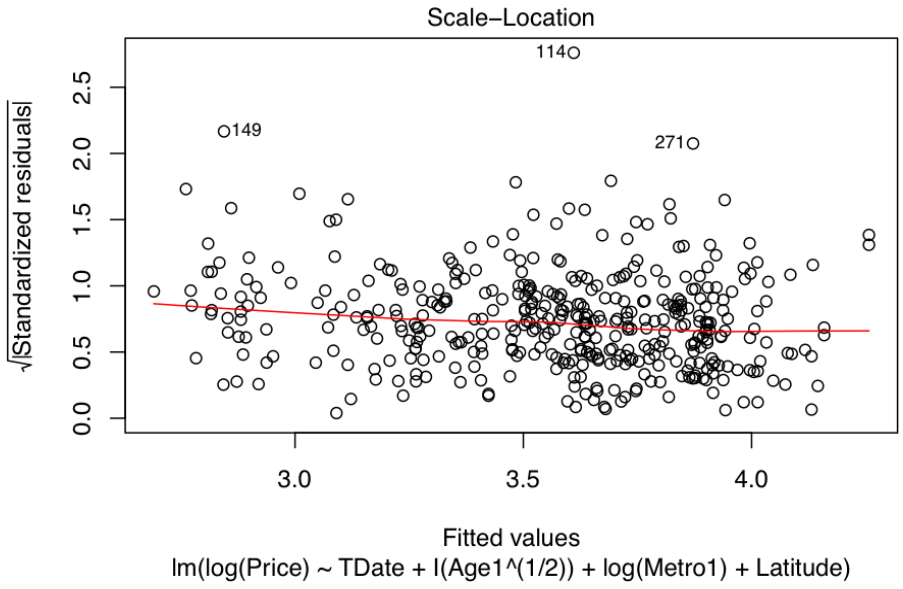
```
plot(Real.lm2trans, which = 1)
```



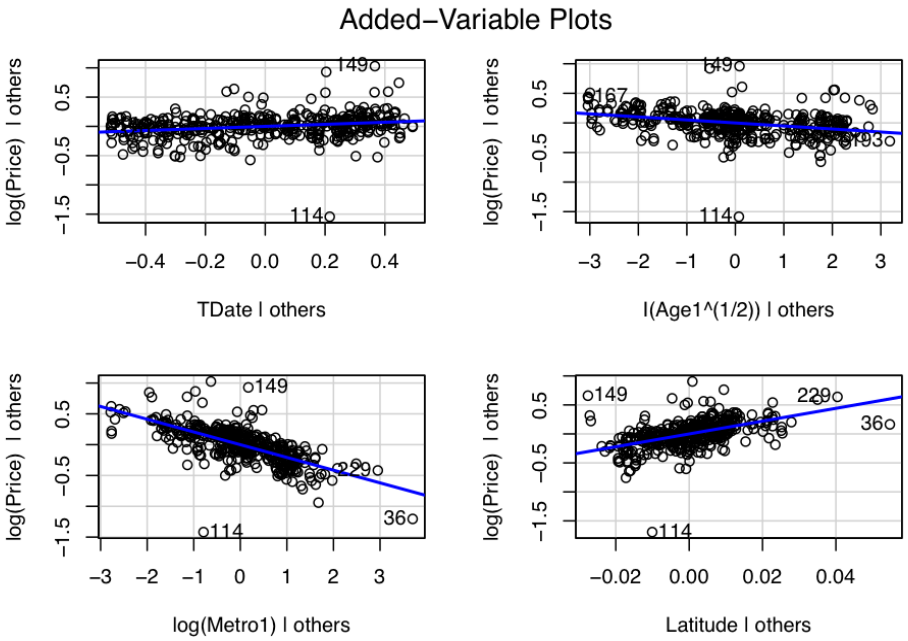

```
plot(Real.lm2trans, which = 2)
```




```
plot(Real.lm2trans, which = 3)
```




```
avPlots(Real.lm2trans)
```



F

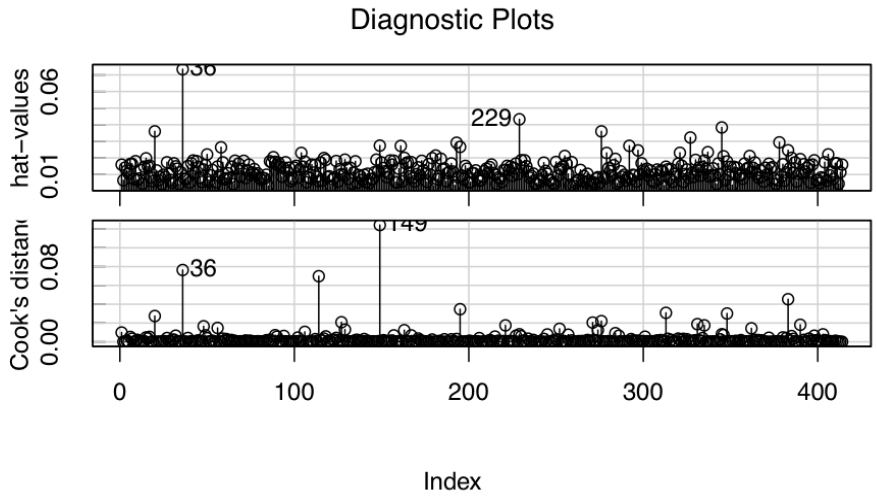
Looking at the summary of the transformed model, 72.1% of the variability in Price is explained by TDate, Age, Metro, and Latitude. This percentage of variability is good. To interpret our coefficients, we need to calculate them based off of the transformations we have made. Our response (Price) is log transformed, so for the predictors that are not transformed (TDate and Latitude) we will use the equation 100((e^Bj)-1). So when TDate increases by one unit, the expected value of Price increases by 19.41%. Interpreting latitude is difficult because latitude only changes by decimals and not a full unit. Metro is also log transformed so we use the equation 100((1+p)^(Bj)-1). When Metro increases by 1%, we can expect the expected value of Price to decrease by 20.5%. It would be very difficult to interpret Age since the variable is a square root transformation. Now, we will use the influence Index plot to look at the high leverage points and the outliers from the cook's distance. The high leverage points are 36 and 229 while the outliers shown from the cook's distance are 36 and 149. These may be outliers because of their high distance from the nearest Metro station, their low number of stores, and a resulting price that is higher than others with similar numbers for Metro and Stores. We can try to remove the 2 outliers from the cook's distance and see if there are any improvements. We need to create a new data set that does not contain the 2 outliers and replot the data. No improvements are found with normality. Lastly, we can test to see what the overall best model would be. It may be that a combination of model 1 and 2 together will be the best. The actual best model would include TDate, Age, Metro, Stores, and Latitude. The model with these predictors has a lower AIC than the previous model we selected in part E.

```
summary(Real.lm2trans)
```

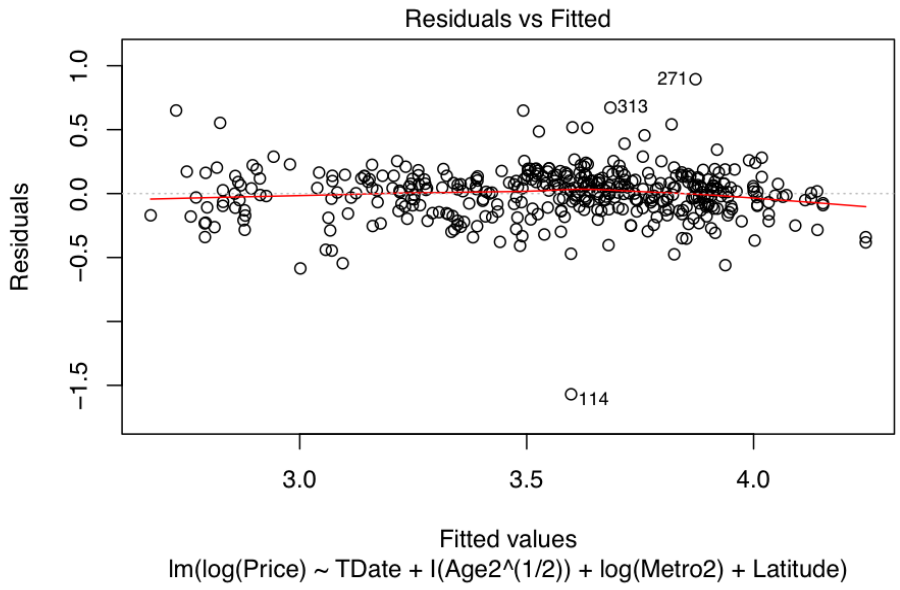
```
##
## Call:
```

```
## lm(formula = log(Price) ~ TDate + I(Age1^(1/2)) + log(Metro1) +
##     Latitude, data = RealEstateValuation)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58230 -0.10512  0.01351  0.11151  0.96447
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.280e+02  7.547e+01  -8.321 1.32e-15 ***
## TDate          1.774e-01  3.668e-02   4.837 1.87e-06 ***
## I(Age1^(1/2)) -5.137e-02  7.261e-03  -7.074 6.56e-12 ***
## log(Metro1)   -2.068e-01  1.055e-02 -19.592  < 2e-16 ***
## Latitude       1.105e+01  9.373e-01  11.790  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2083 on 409 degrees of freedom
## Multiple R-squared:  0.721,  Adjusted R-squared:  0.7183
## F-statistic: 264.3 on 4 and 409 DF,  p-value: < 2.2e-16
```

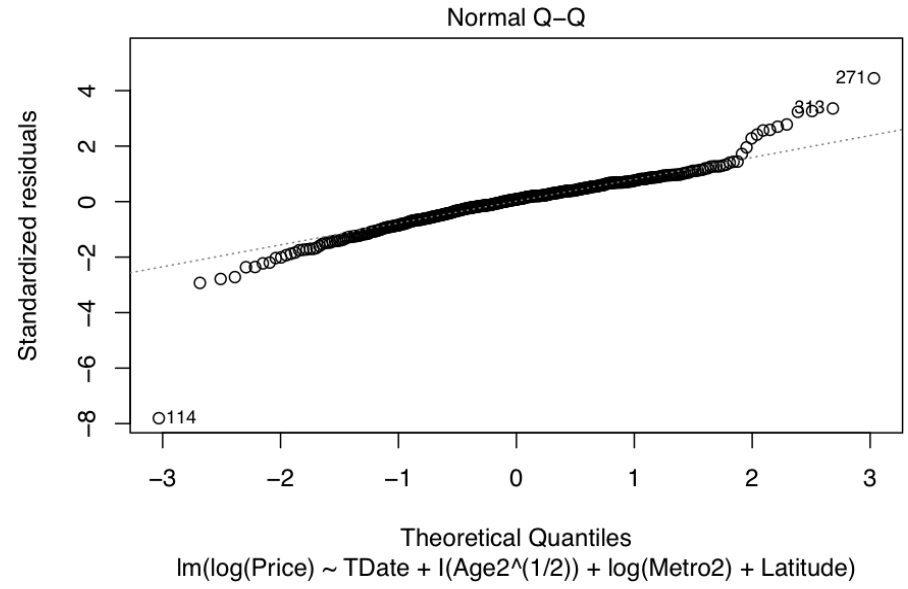
```
influenceIndexPlot(Real.lm2trans, vars= c('hat','Cook'), id.n=5, id.cex=0.7)
```



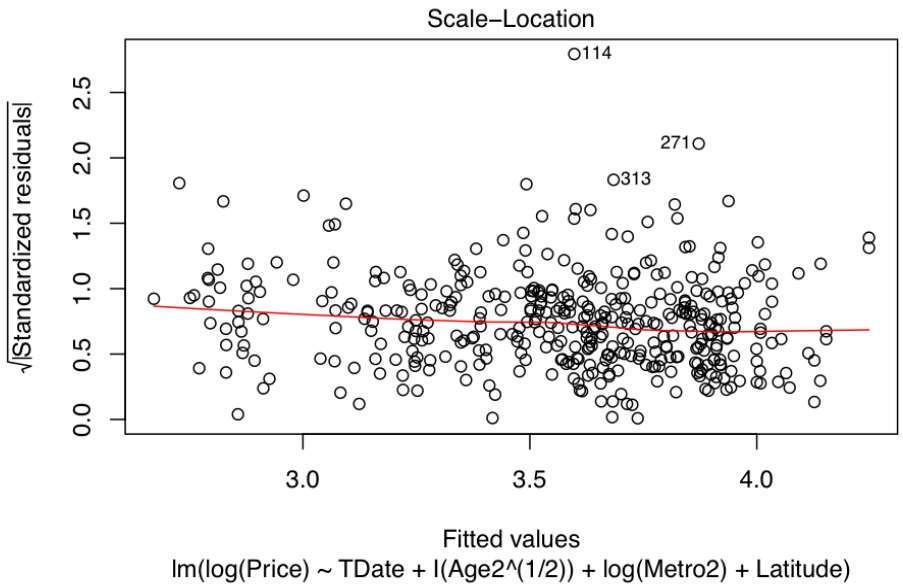

```
newRealEstate<-RealEstateValuation[-c(36,149),]
Age2=newRealEstate$Age+1
Metro2=newRealEstate$Metro+1
NEWFITREAL<- lm(log(Price)~TDate+I(Age2^(1/2))+log(Metro2)+Latitude, data=newRealEstate)
plot(NEWFITREAL, which = 1)
```




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




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




```
AIC(Real.lm2)
```

```
## [1] 3021.623
```

```
AIC(lm(Price~TDate+Age+Metro+Stores+Latitude, data=RealEstateValuation))
```

```
## [1] 2987.991
```

Part 2- Concrete Compressive Strength Data Set:

Question 2

```
library(alr4)
setwd("~/Desktop")
Concrete<- read.table("Concrete.txt", header=TRUE)
```

A

Applying Forward Selection (BIC), we begin with the smallest model and forward selection will add the variables one at a time until the chosen information criterion cannot be decreased anymore. From the forward selection procedure (using BIC), we select the final model (Y ~ X1 + X5 + X8 + X2 + X4 + X3). Now it is suitable to run tests on this model to see if it is a good fit. We can check the diagnostics for the linear regression assumptions for this new model. From the ANOVA table, we can see that each predictor has a p-value that is less than an alpha level of 0.05. X1 (Cement), is clearly correlated with the Y (Concrete compressive strength). X5 (Superplasticizer) is a significant predictor when accounting for X1, X8 (Age) is a significant predictor when accounting for X1 and X5, and so on. Each predictor is significant and clearly correlated with the response, even when accounting for the previous predictors.

This model could be a good fit. From the Added Variable Plots, we can see that each predictor is significant to the model because they have a relationship with the response. The predictors X1, X2, X3, and X8 have a positive correlation with the response, X5 has a weak positive correlation with the response, and X4 has a negative correlation with the response.

```
data(Concrete)
```

```
## Warning in data(Concrete): data set 'Concrete' not found
```

```
# The Full Model
confull.lm <- lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8, data = Concrete)
# The Empty Model
con0.lm <- lm(Y ~ 1, data = Concrete)
```

```
n = 1030
conforward.lm <- step(con0.lm, scope = list(lower = con0.lm, upper = confull.lm), direction = 'forward'
```

```
## 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
##
```

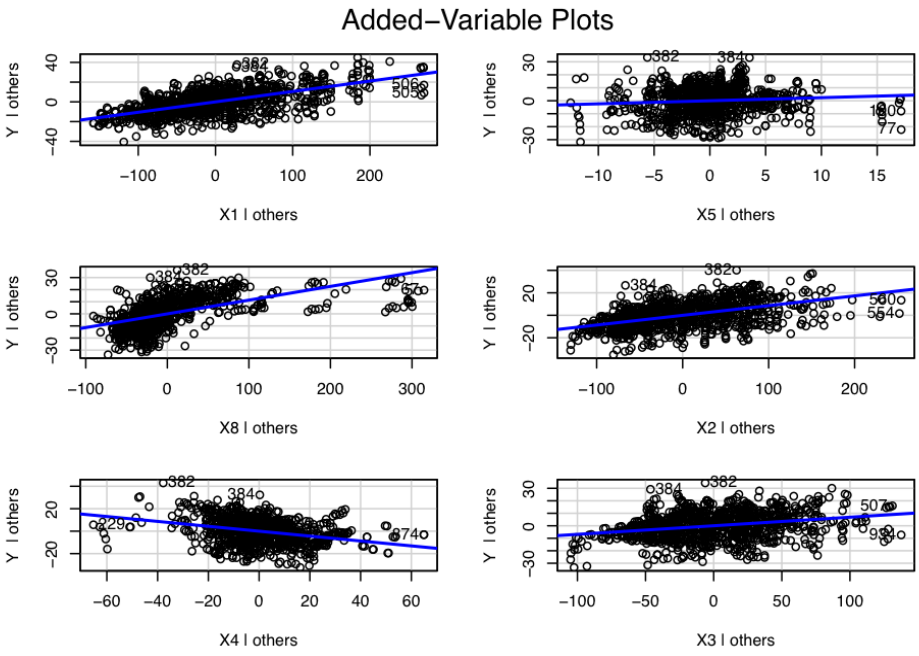
```
## Step:  AIC=5150.38
## Y ~ X1 + X5 + X8
##
##        Df Sum of Sq    RSS    AIC
## + X2    1   19908.5 128948 5009.4
## + X4    1    4868.8 143988 5123.1
## + X7    1    3385.5 145471 5133.6
## <none>               148857 5150.4
## + X3    1     323.9 148533 5155.1
## + X6    1      36.9 148820 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
```

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

```
anova(mod.better) #analysis of variance
```

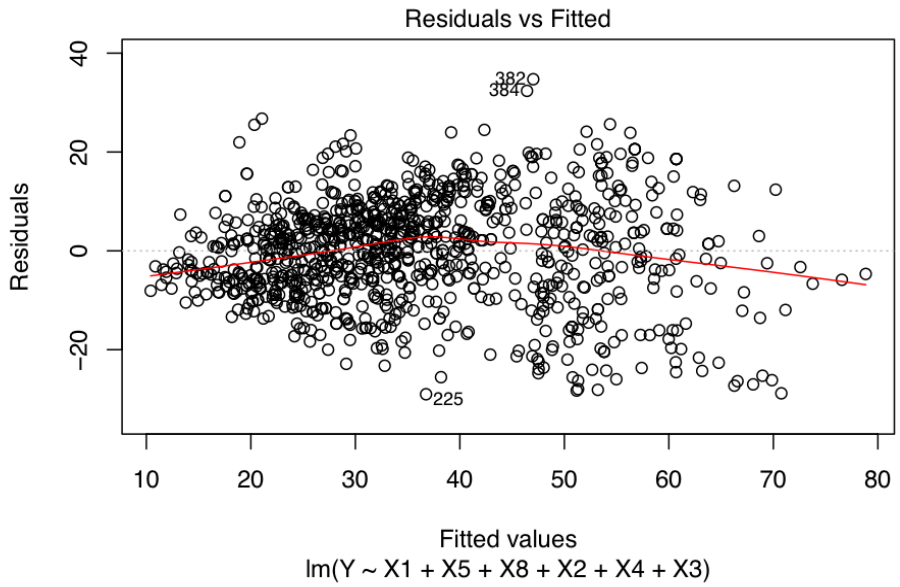
```
## Analysis of Variance Table
##
## Response: Y
##             Df Sum Sq Mean Sq F value    Pr(>F)
## X1           1  71172   71172 656.791 < 2.2e-16 ***
## X5           1  29647   29647 273.584 < 2.2e-16 ***
## X8           1  37498   37498 346.036 < 2.2e-16 ***
## X2           1  19908   19908 183.719 < 2.2e-16 ***
## X4           1   9545    9545  88.080 < 2.2e-16 ***
## X3           1   8547    8547  78.877 < 2.2e-16 ***
## Residuals 1023 110856     108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
avPlots(mod.better) #avPlots
```

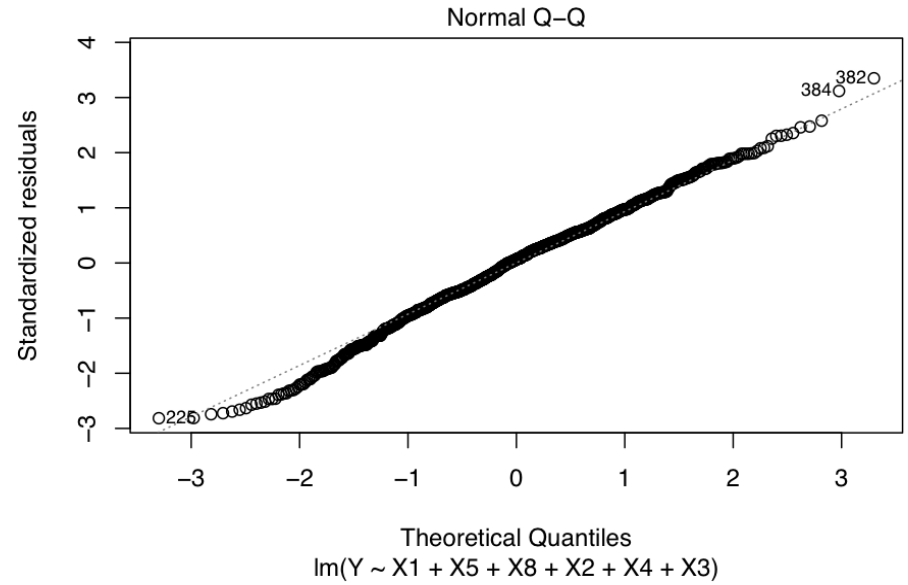


From the Residuals vs. Fitted plot, we can assume the linearity assumption holds for the most part. From the QQ-Plot, we can assume that the normality assumption holds for the most part because there are no outliers and it is not heavy-tailed, but it is slightly right-skewed. From the Scale-Location Plot, we can see that the constant variance assumption is slightly violated.

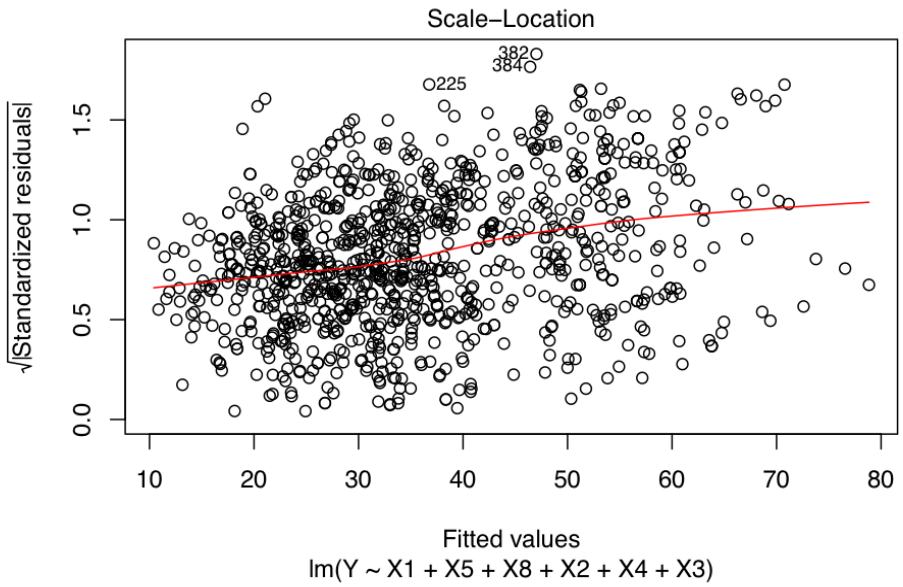
```
plot(mod.better, which = 1) #residual vs fitted
```




```
plot(mod.better, which = 2) #QQPlot
```




```
plot(mod.better, which = 3) # Scale-Location vs. Residuals
```



We can see from the Non-Constant Variance test that the p-value is statistically signficant, which means that there is non-constant variance so we should weight the residuals to get a better fitted model. From the outlier Test, we can see that the Bonferonni p is not significant, so the point, 382, may have high leverage and could potentially be an outlier. According to the Diagnostics Plot, there are two influential points, 611 and 57, which have large Cook's distance. These influential points are the outliers. 43 and 67 are points with high leverage.

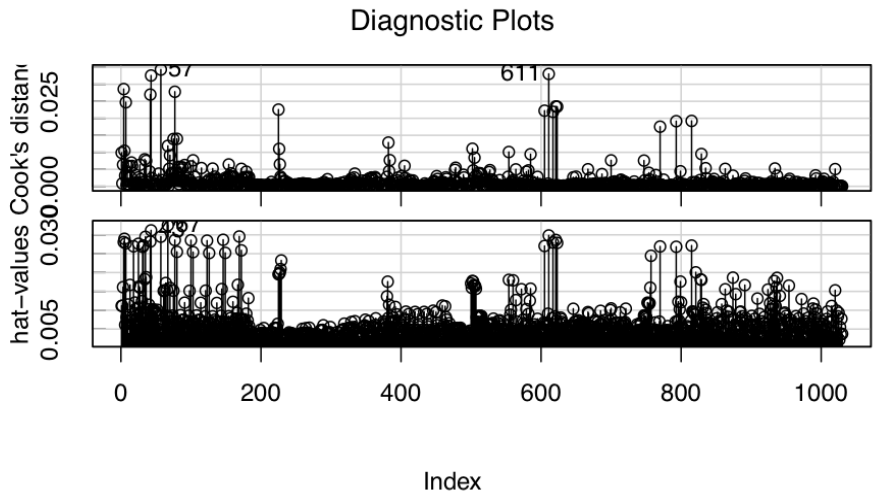
```
ncvTest(mod.better)
```

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

```
outlierTest(mod.better)
```

```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 382 3.366138         0.00079066      0.81438
```

```
# We will use influence Index Plot to see what predictors have more effect
influenceIndexPlot(mod.better, vars = c('Cook', 'hat'))
```



For estimating a new response, we can make a data frame of predictor values of our interest and compute a 95% confidence interval and prediction interval. So for our selected predictors X1 = 387, X2 = 20, X3 = 94, X4 = 157, X5 = 13.93, X8 = 56, we are 95% confident that the mean response will be between 52.12559 and 54.73889. Our 95% prediction interval is between 32.96352 and 73.90095. For our data, the expected Concrete compressive strength is 53.43224.

```
x_new = data.frame(X1 = 387, X2 = 20, X3 = 94, X4 = 157, X5 = 13.93, X8 = 56)
predict(mod.better, newdata = x_new, interval = 'confidence', level = 0.95)
```

```
##        fit      lwr      upr
## 1 53.43224 52.12559 54.73889
```

```
predict(mod.better, newdata = x_new, interval = 'predict', level = 0.95)
```

```
##        fit      lwr      upr
## 1 53.43224 32.96352 73.90095
```

B

Now applying backward selection with BIC, we will begin with the largest model and remove one predictor at a time until the information criterion cannot be decreased. The final model presented is (Y~X1+X2+X3+X4+X5+X8). This is the same model that we are given in part a) using forward selection. However, the order of the predictors are different. The Diagnostics check is the same as in part a).

```
n = 1030
conbackward.lm <- step(confull.lm, scope = list(lower = con0.lm, upper = confull.lm), direction = 'backw
```

```
## 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
```

```
mod.betterb<-(lm(Y~X1+X2+X3+X4+X5+X8, data=Concrete))
anova(mod.betterb)
```

```
## Analysis of Variance Table
##
## Response: Y
##             Df Sum Sq Mean Sq F value    Pr(>F)
## X1           1  71172   71172 656.791 < 2.2e-16 ***
## X2           1  22957   22957 211.856 < 2.2e-16 ***
## X3           1  21636   21636 199.665 < 2.2e-16 ***
## X4           1  11459   11459 105.747 < 2.2e-16 ***
## X5           1   1360    1360  12.555 0.0004131 ***
## X8           1  47731   47731 440.475 < 2.2e-16 ***
## Residuals 1023 110856     108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

C

According to forward selection AIC and BIC, as well as backward selection BIC, we are given the same model, which is Y ~ X1 + X2 + X3 + X4 + X5 + X8. The diagnostics check for the selected final model shows us that the linearity assumption, and normality assumption hold. We can see that there is a violation to constant variance. To fix this violation, we can weight the residuals and create a newer model. We tested for outliers, and saw that there are two strong outliers in Cook's distance, as well as two points with high leverage.

```
conforwardAIC <- step(con0.lm, scope = list(lower = con0.lm, upper = confull.lm), direction = 'forward'
```

```
## Start:  AIC=5801.44
## Y ~ 1
##
##        Df Sum of Sq    RSS    AIC
## + X1    1     71172 216001 5510.1
## + X5    1     38490 248683 5655.2
## + X8    1     31061 256112 5685.5
## + X4    1     24087 263086 5713.2
## + X7    1      8033 279140 5774.2
## + X6    1      7811 279362 5775.0
## + X2    1      5220 281953 5784.5
## + X3    1      3212 283961 5791.9
## <none>              287173 5801.4
##
## Step:  AIC=5510.1
## Y ~ X1
##
##        Df Sum of Sq    RSS    AIC
## + X5    1   29646.5 186354 5360.0
## + X8    1   23993.8 192007 5390.8
## + X2    1   22957.4 193043 5396.4
## + X4    1   17926.8 198074 5422.9
## + X6    1    3548.0 212453 5495.0
## + X3    1    2894.4 213106 5498.2
## + X7    1     960.2 215041 5507.5
## <none>              216001 5510.1
##
## Step:  AIC=5360.03
## Y ~ X1 + X5
##
##        Df Sum of Sq    RSS    AIC
## + X8    1     37498 148857 5130.6
## + X2    1     19456 166898 5248.5
## + X7    1      5862 180493 5329.1
## + X4    1       782 185572 5357.7
## + X3    1       741 185613 5357.9
## <none>              186354 5360.0
## + X6    1       241 186113 5360.7
##
## Step:  AIC=5130.63
## Y ~ X1 + X5 + X8
##
##        Df Sum of Sq    RSS    AIC
## + X2    1   19908.5 128948 4984.7
```

```
## + X4    1    4868.8 143988 5098.4
## + X7    1    3385.5 145471 5108.9
## + X3    1     323.9 148533 5130.4
## <none>              148857 5130.6
## + X6    1      36.9 148820 5132.4
##
## Step:  AIC=4984.75
## Y ~ X1 + X5 + X8 + X2
##
##        Df Sum of Sq    RSS    AIC
## + X4    1    9544.7 119403 4907.5
## + X3    1    6524.7 122423 4933.3
## + X6    1    1737.0 127211 4972.8
## <none>              128948 4984.7
## + X7    1       3.5 128945 4986.7
##
## Step:  AIC=4907.54
## Y ~ X1 + X5 + X8 + X2 + X4
##
##        Df Sum of Sq    RSS    AIC
## + X3    1    8547.4 110856 4833.0
## + X7    1    1895.7 117508 4893.1
## <none>              119403 4907.5
## + X6    1      24.1 119379 4909.3
##
## Step:  AIC=4833.03
## Y ~ X1 + X5 + X8 + X2 + X4 + X3
##
##        Df Sum of Sq    RSS    AIC
## <none>              110856 4833.0
## + X6    1    44.271 110812 4834.6
## + X7    1    29.398 110827 4834.8
```

```
conbackwardAIC <- step(confull.lm, scope = list(lower = con0.lm, upper = confull.lm), direction = 'back
```

```
## Start:  AIC=4833.05
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##
##        Df Sum of Sq    RSS    AIC
## <none>              110428 4833.1
## - X7    1       384 110812 4834.6
## - X6    1       398 110827 4834.8
## - X5    1      1046 111474 4840.8
## - X4    1      1513 111942 4845.1
## - X3    1      5281 115709 4879.2
## - X2    1     11353 121781 4931.8
## - X1    1     21533 131961 5014.5
## - X8    1     47905 158333 5202.2
```

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

```
##
## Call:
## lm(formula = Y ~ X1 + X5 + X8 + X2 + X4 + X3, data = Concrete)
##
```

```
## 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
```

Conclusion:

In this project, we analyzed two different sets of data. The first set of a data was on Real Estate Valuation. We tested to see whether six variables had a strong relationship with the price of a house and which combination of variables were accurate in predicting the price. We came to the conclusion that an overall good model would include TDate, Age, Metro, Stores, and Latitude. This means that the transaction date of the house, the age of the house, the distance to the nearest MRT station, the number of convenience stores in the area, and the geographic coordinate of latitude all affect the price of a house. They not only affect the price but the combination of them in a regression model can result in an accurate estimated price. This is important when properties or houses are put on the market because realtors have to take all of these variables into consideration. There were more than one models that could be accurate for estimating price, so the preferance of the model could also depend on the buyers interensts and what predictors they want to take into account the most when finding an accurate price for a house.

The second set of data was on Concrete compressive strength. We tested whether eight concrete components had a strong relationship with the concrete compressive strength. We came to the conclusion that an overall good model would include Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, and Age. This is important for making strong concrete because compressive strength of concrete is the most common measure used by engineers when designing buildings. It was interesting to notice that the only predictor that decreased the overall strength as it was added was water. Water had a negative correlation with Concrete compressive strength; however, it is a necessary ingredient in the overall composition of the Concrete. It is necessary to have an accurate and appropriate strength depending on ones intended use of the Concrete.