Abstract:

Real Estate Valuation:

Many factors are taken into consideration when estimating the price of a house. Using our statistical knowledge, we create a model that uses necessary factors applying transformations to estimate the price of real estate in the Sindian District. Our findings concluded that the distance to the nearest MRT station, number of convenience stores in proximity, and latitude were the most important factors to the house price of unit area.

Concrete Compressive Strength:

There are seven concrete components as well as age, in our data set that influence the concrete compressive strength. We created a model through statistical algorithms that help us dictate which of these components are needed or not needed to estimate the concrete compressive strength.

Problem and Motivation:

Real Estate Valuation:

The data set was collected from the Sindian District, New Taipei City, Taiwan during the period of June 2012 to May 2013. Someone interested in purchasing or investing in a property in the Sindian District, New Taipei City, or Taiwan, can learn which factors are important to look at how important they are. Although this data set is only collected from a specific period, the conclusions we draw from our findings can probably be applied to different time periods. However, the conclusions we draw in this project are to show the significance of certain factors for this specific data set.

Concrete Compressive Strength:

The data set was collected from the paper, Modeling of strength of high performance concrete using artificial neural networks, by Dr. I-Cheng Yeh. This data can be useful to those who are interested in creating a level of concrete compressive strength they desire. One would be able to learn which factors are more important than others and be able to create their concrete at a price they want or create it at a quality they want.

Data:

Real Estate Valuation:

1. TDate (the transaction date)
2. Age (the house age)
3. Metro (the distance to the nearest MRT station)
4. Stores (the number of convenience stores in the living circle on foot)
5. Latitude (the geographic coordinate)
6. Longitude (the geographic coordinate)

We look at these variables to see how they affect the variable Price (the house price of unit area), which is the most important variable customers will look at.

Concrete Compressive Strength:

1. Cement (component 1)
2. Blast Furnace Slag (component 2)
3. Fly Ash (component 3)
4. Water (component 4)
5. Superplasticizer (component 5)
6. Coarse Aggregate (component 6)
7. Fine Aggregate (component 7)

8. Age

Concrete compressive strength is variable of interest and we will analyze the relationship it has with the eight variables above.

Questions of Interest:

Real Estate Valuation:

Which variables are significant to the house price of unit area? How influential are these variables to the house price of unit area? What variable transformations will improve the model?

Concrete Compressive Strength:

Which variables are significant to the concrete compressive strength? What algorithm will give us the best model? Are there any outliers that affect the model? Given that we know certain variables, can we estimate or predict a concrete compressive strength?

Regression Methods:

Real Estate Valuation:

1. We had preliminary expectations between the predictors and the response, using logic and our basic understanding of real estate.
2. Created a fitted linear model for the Price Per Unit Area of a house.
3. Tested for significance of predictors in a model using t-tests and hypothesis testing.
4. Tested to see if adding certain variables would be a significant addition to the model using ANOVA tests and hypothesis testing.
5. Created a second fitted linear model for the Price Per Unit Area of a house and compared it to the original using diagnostic plots and finding their AIC's and RSS's.
6. Transformed certain predictors logarithmically we found looking at a scatterplot matrix.
7. Tested to see if transforming the response has an effect using Box-Cox.
8. Tested a further transformation, a Box-Tidwell transformation, to attempt to create a better model.

Concrete Compressive Strength:

1. Applied forward selection using BIC as a criterion function to get a model.
2. Performed diagnostic checks: residual vs. fitted, QQ plot, and Durbin-Watson test, to determine if linear regression assumptions hold.
3. Checked for influential observations using Cook's Distance.
4. Estimated a mean response for our predictor values.
5. Predicted a new response for our predictor values.
6. Applied backward selection using BIC as a criterion function to get a model.

Real Estate Valuation Data Analysis

Preliminary Expectations

Before fitting the model, I expect the following relationships to occur within the model:

1. A positive linear relationship between Transaction Date and Price per Unit Area due to the effects of inflation.

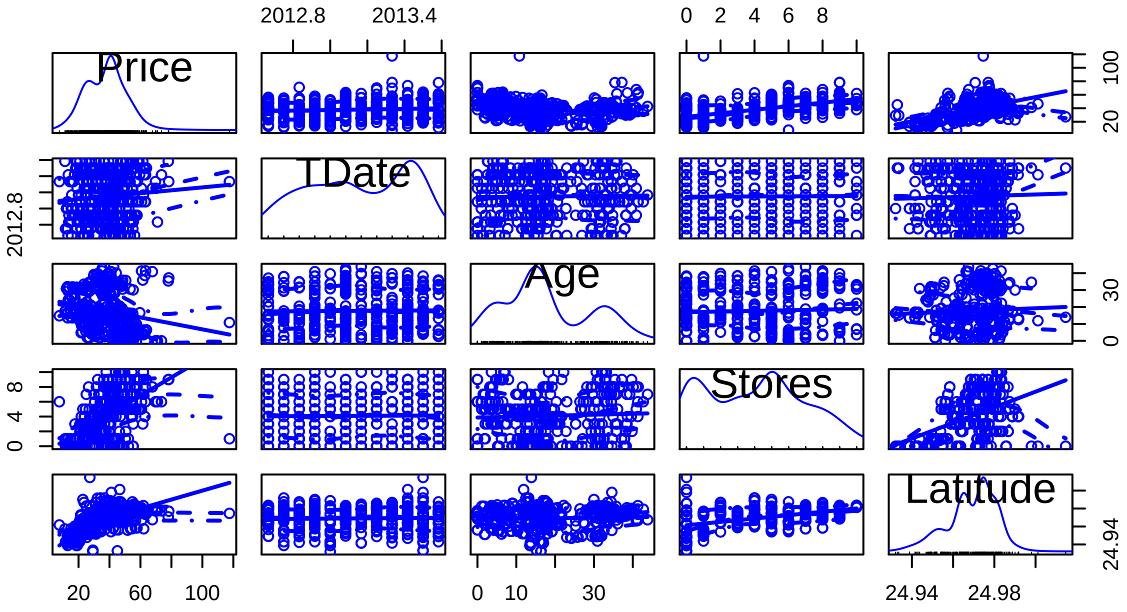
2. A negative linear relationship between House Age and Price per Unit Area due to the desirability of newly built homes.

3. A positive linear relationship between the Number of Convenience Stores in the Living Circle and Price per Unit Area due to the desirability of proximity to urban commercial centers as well as the finite extent of urban residential land.

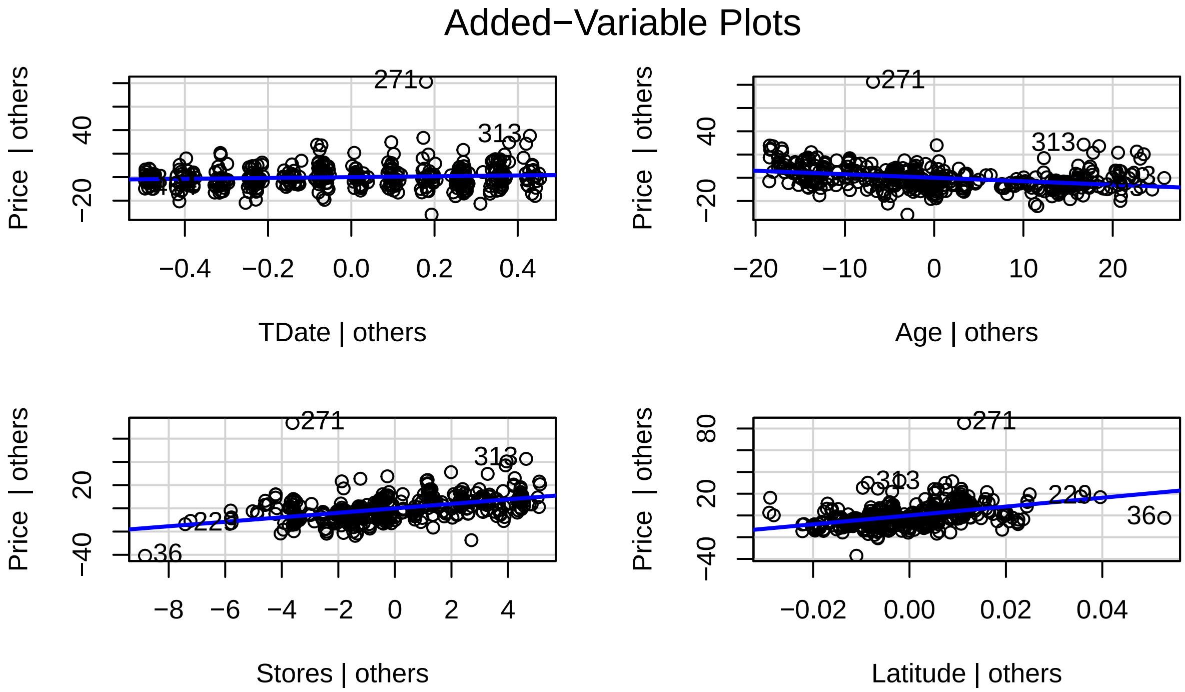
4. A positive linear relationship between the Latitude of a home and the Price per Unit Area as the Xindian District, the district from which data was collected, is located in southern New Taipei City, Taiwan. Therefore observations with larger latitude values are located farther north, closer to the center of New Taipei City. These observations should correlate to a higher Price per Unit Area due to the desirability of proximity to urban centers and the finite extent of urban residential land.

Exploratory Analysis

```
fit<-lm(Price~TDate+Age+Stores+Latitude)
scatterplotMatrix(~Price+TDate+Age+Stores+Latitude)
```



```
avPlots(fit)
```



Observation 271 seems to be an outlier.

```r
outlierTest(fit)
```

```
##     rstudent unadjusted p-value Bonferonni p
## 271 9.189094         2.0369e-18   8.4327e-16
```

Based on the calculated P-Values, observation 271 is a significant outlier.

```r
summary(fit)
```

```
##
## 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
```

The fitted linear model for the Price Per Unit Area of a house based on the transaction date, the house age, the number of convenience stores in the living circle, and the latitude of the house is:

Price=-17419.95+3.613(Transaction Date)-0.302(House Age)+1.929(Stores in Living Circle)+407.81(Latitude).

```r
datetest<-lm(Price~TDate)
agetest<-lm(Price~Age)
storestest<-lm(Price~Stores)
lattest<-lm(Price~Latitude)
dfTDate<-c("TDate SLR Coefficient",(datetest$coefficients[2]))
dfAge<-c("Age SLR Coefficient", agetest$coefficients[2])
dfStores<-c("Stores SLR Coefficient", storestest$coefficients[2])
dfLatitude<-c("Latitude SLR Coefficient", lattest$coefficients[2])
df<-data.frame(dfTDate, dfAge, dfStores, dfLatitude)
print.data.frame(df, row.names=FALSE, )
```

```
##                dfTDate                dfAge                dfStores
##  TDate SLR Coefficient Age SLR Coefficient Stores SLR Coefficient
##       4.22190839820424  -0.251488419085345       2.63765346340437
##                dfLatitude
##  Latitude SLR Coefficient
##          598.96833158964
```

As expected, TDate has a positive relationship with Price, Age has a negative relationship with Price, Stores has a positive relationship with Price, and Latitude has a positive relationship with Price.

Tests for Significance of Predictors in Model:

Transaction Date:

Null Hypothesis $H_0 : \beta_1 = 0$, Alternative Hypothesis $H_A : \beta_1 \neq 0$.

```r
#T-Value
T=fit$coefficients[2]/(1.686)
#Critical Value
CV=qt(0.995, 412)
#Test
abs(T)>CV
```

```
## TDate
## FALSE
```

Since the absolute value of the T value is less than the critical value at level $\alpha = 0.01$, the null hypothesis is not rejected and it is determined that $\beta_1 = 0$. Therefore Transaction Date is not a significant predictor in the model after accounting for the other predictors.

House Age:

Null Hypothesis $H_0 : \beta_2 = 0$, Alternative Hypothesis $H_A : \beta_2 \neq 0$.

```r
#T-Value
T=fit$coefficients[3]/(0.04178)
#Critical Value
CV=qt(0.995,412)
#Test
abs(T)>CV
```

```
##  Age
## TRUE
```

Since the absolute value of the T value is greater than the critical value at level $\alpha = 0.05$, the null hypothesis is rejected and it is determined that $\beta_2 \neq 0$. Therefore House Age is a significant predictor in the model after accounting for the other predictors.

Number of Convenience Stores:

Null Hypothesis $H_0 : \beta_3 = 0$, Alternative Hypothesis $H_A : \beta_3 \neq 0$.

```r
#T-Value
T=fit$coefficients[4]/(0.1801)
#Critical Value
CV=qt(0.995,412)
#Test
abs(T)>CV
```

```
## Stores
##   TRUE
```

Since the absolute value of the T value is greater than the critical value at level $\alpha = 0.01$, the null hypothesis is rejected and it is determined that $\beta_3 \neq 0$. Therefore the Number of Convenience Stores is a significant predictor in the model after accounting for the other predictors.

Latitude:

Null Hypothesis $H_0 : \beta_4 = 0$, Alternative Hypothesis $H_A : \beta_4 \neq 0$.

```
#T-Value
T=fit$coefficients[5]/(42.78)
#Critical Value
CV=qt(0.995, 412)
#Test
abs(T)>CV
```

```
## Latitude
##     TRUE
```

Since the absolute value of the T value is greater than the critical value at level $\alpha = 0.01$, the null hypothesis is rejected and it is determined that $\beta_4 \neq 0$. Therefore Latitude is a significant predictor in the model after accounting for the other predictors.

Adding Predictors Metro/Longitude

Adding Metro

The following test will determine if the model is improved enough with the inclusion of the predictor Distance to the Nearest MRT Station to warrant the increased complextiy of an added predictor. Null Hypothesis H_0 : The model without Metro is sufficient, Alternative Hypothesis H_A : Metro is a significant addition to the model given the predictors.

```
submodel<-lm(Price~TDate+Age+Stores+Latitude)
fullmodel<-lm(Price~TDate+Age+Stores+Latitude+Metro)
anova(submodel, fullmodel)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ TDate + Age + Stores + Latitude
## Model 2: Price ~ TDate + Age + Stores + Latitude + Metro
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    409 38119
## 2    408 31938  1    6181.8 78.972 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The value of the test statistic is 78.972, the null distribution of the test statistic is $F_{1,408}$. The P-Value is less than $2.2 * 10^{-16}$, which is less than $\alpha = 0.05$. Therefore the null hypothesis is rejected and it is determined that Metro is a significant addition to the model.

Adding Longitude

The following test will determine if the model is improved enough with the inclusion of the predictor Longitude to warrant the increased complexity of an added predictor.

Null Hypothesis H_0 : The model without Longitude is sufficient, Alternative Hypothesis H_A : Longitude is a significant addition to the model given the predictors.

```
submodel<-lm(Price~TDate+Age+Stores+Latitude)
fullmodel<-lm(Price~TDate+Age+Stores+Latitude+Longitude)
anova(submodel, fullmodel)
```

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

The value of the test statistic is 36.402, the null distribution of the test statistic is $F_{1,408}$. The P-Value is $3.605 * 10^{-9}$, which is less than $\alpha = 0.05$. Therefore the null hypothesis is rejected and it is determined that Longitude is a signifcant addition to the model given the predictors.

Adding Metro Given Longitude is in the Model

The following test will determine if the model with the predictor Longtiude included is improved enough with the addition of Metro to warrant the increased complexity of an added predictor.

Null Hypothesis H_0 : The model with Longtidue and without Metro is suffcient, H_A : Metro is a significant addition to the model.

```
submodel<-lm(Price~TDate+Age+Stores+Latitude+Longitude)
fullmodel<-lm(Price~TDate+Age+Stores+Latitude+Longitude+Metro)
anova(submodel, fullmodel)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ TDate + Age + Stores + Latitude + Longitude
## Model 2: Price ~ TDate + Age + Stores + Latitude + Longitude + Metro
##   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
```

The value of the test statistic is 39.059, the null distribution of the test statistic is $F_{1,407}$. The P-Value is $1.039 * 10^{-09}$, which is less than $\alpha = 0.05$. Therefore the null hypothesis is rejected and it is determined that Metro is a significant addition to the model that includes Longitude given the predictors.

Adding Longitude Given Metro is in the Model

The following test will determie if the model with the predictor Metro included is improved enough with the addition of Longitude to warrant the increased complexity of an added predictor.

Null Hypothesis H_0 : The model with Metro and without Longitude is suffcient, H_A : Longitude is a significant addition to the model.

```
submodel<-lm(Price~TDate+Age+Stores+Latitude+Metro)
fullmodel<-lm(Price~TDate+Age+Stores+Latitude+Metro+Longitude)
anova(submodel, fullmodel)
```

```
## 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
```

The value of the test statistic is 0.0654, the null distribution of the test statistic is $F_{1,407}$. The P-Value is 0.7983, which is greater than $\alpha = 0.05$. Therefore the null hypothesis is not rejected and it is determined that Longitude is not a significant addition to the model after accounting for the predictors, including Metro.

Fitting a Second Model

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

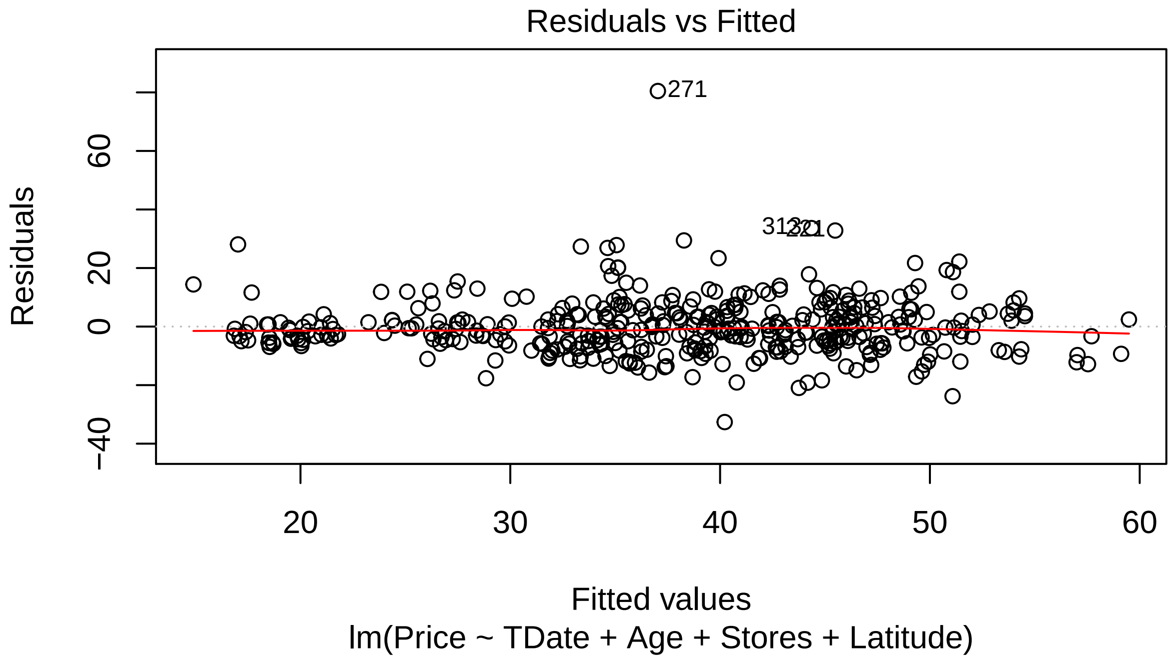
```
##
## Call:
## lm(formula = Price ~ TDate + Age + Metro + Latitude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.218  -5.269  -0.700   4.433  70.502
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.767e+04  3.359e+03  -5.262 2.30e-07 ***
## 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
```

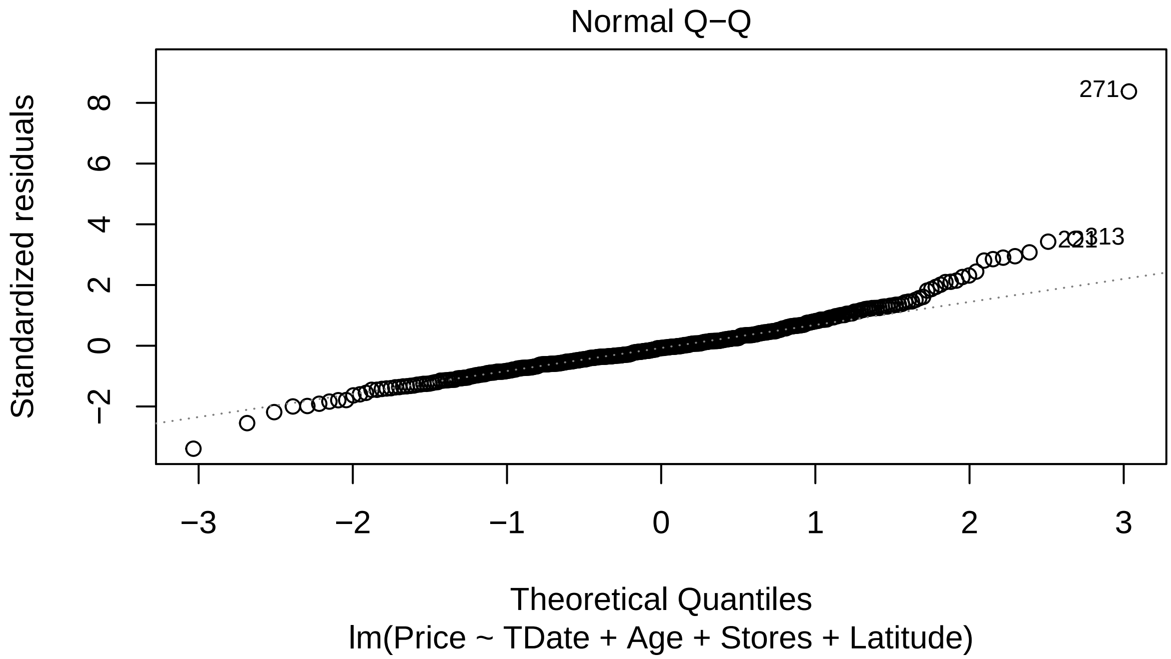
```
step(object=lm(Price~TDate+Age+Latitude), scope=(~TDate+Age+Latitude+Metro+Stores), direction="forward"
```

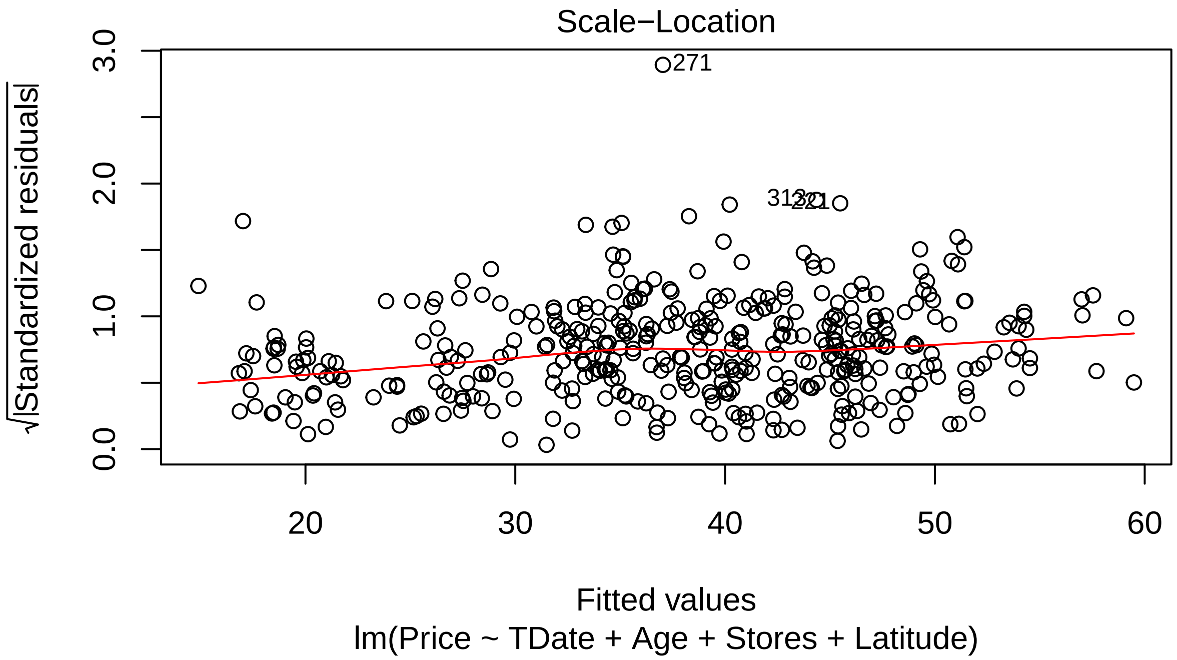
```
## Start:  AIC=1982.75
## Price ~ TDate + Age + Latitude
##
##          Df Sum of Sq   RSS    AIC
## + Metro   1     14007 34808 1844.7
## + Stores  1     10696 38119 1882.4
## <none>                48815 1982.8
##
## Step:  AIC=1844.74
## Price ~ TDate + Age + Latitude + Metro
##
##          Df Sum of Sq   RSS    AIC
## + Stores  1    2870.6 31938 1811.1
## <none>                34808 1844.7
##
## Step:  AIC=1811.11
## Price ~ TDate + Age + Latitude + Metro + Stores
```

```
##
## Call:
## lm(formula = Price ~ TDate + Age + Latitude + Metro + Stores)
##
## Coefficients:
## (Intercept)        TDate          Age     Latitude        Metro
##  -1.596e+04    5.135e+00   -2.694e-01    2.269e+02   -4.353e-03
##      Stores
##    1.136e+00
```

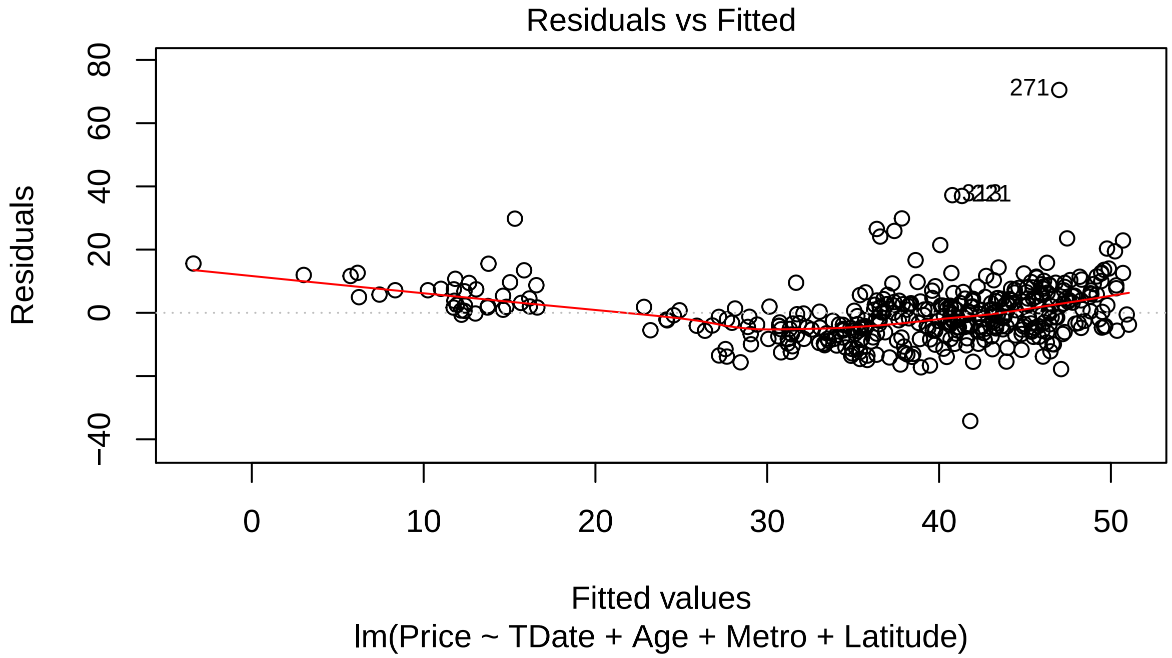
```
plot(fit, which=1:3)
```

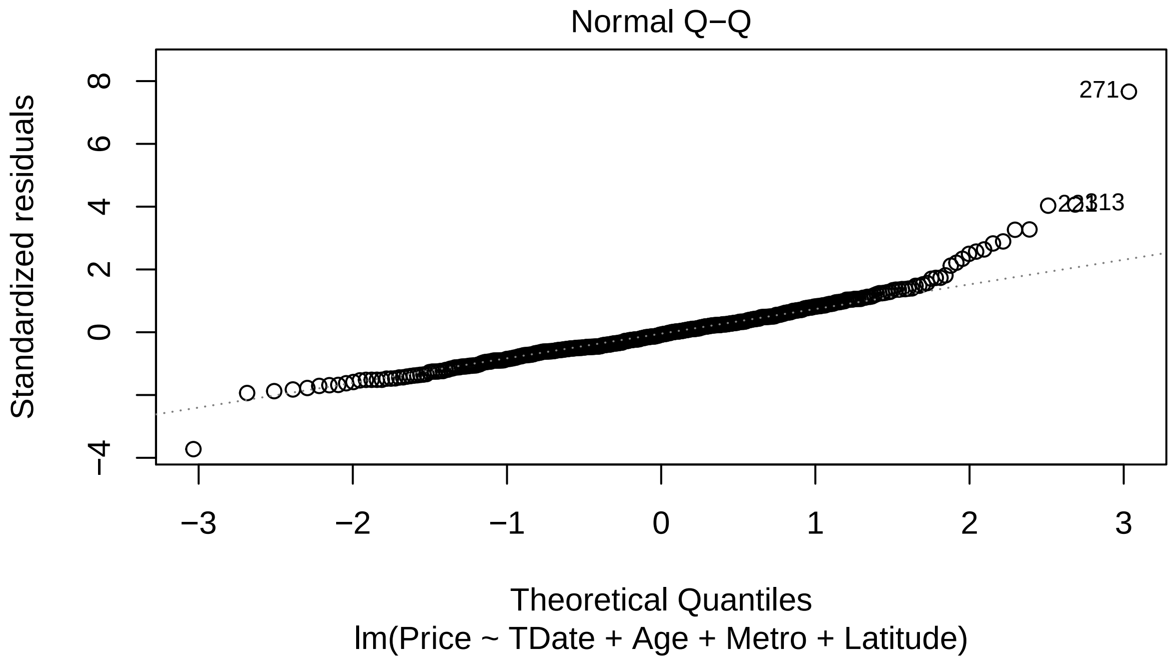


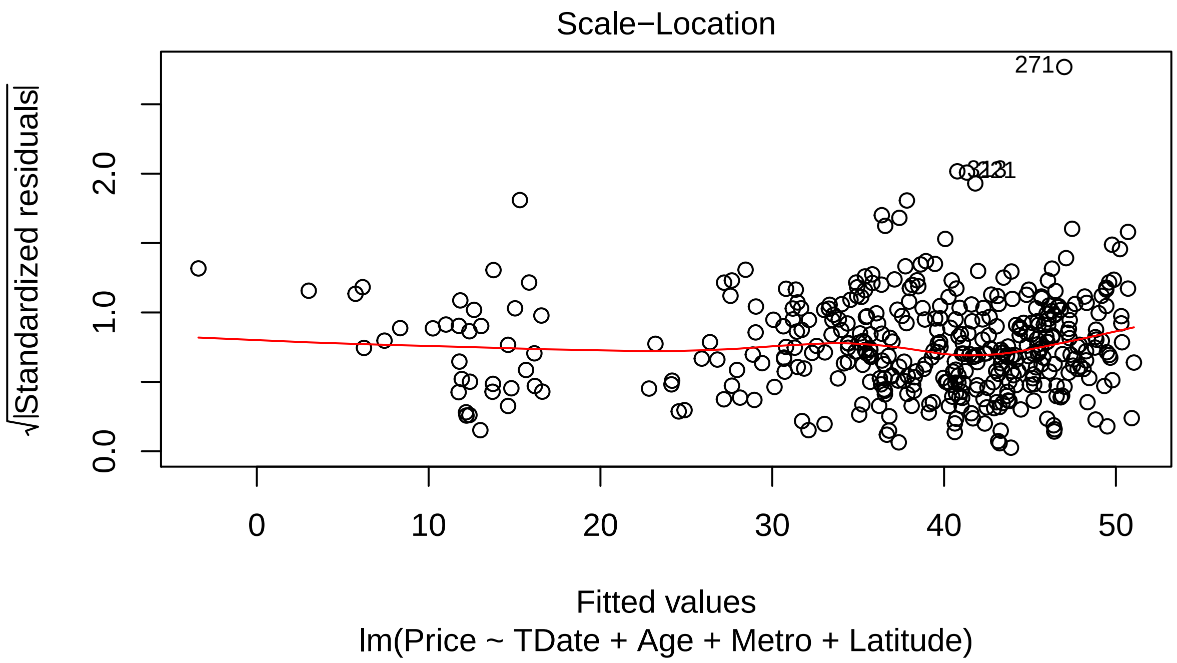




```r
plot(fit2, which=1:3)
```







```
lmtest::dwtest(fit)[4]
```

```
## $p.value
## [1] 0.9130782
```

```
lmtest::dwtest(fit2)[4]
```

```
## $p.value
## [1] 0.9217555
```

The new linear model for the Price Per Unit Area of a house based on Transaction Date, House Age, Distance to the Nearest MRT Station, and Latitude is:

Price=-17673.010+5.570(Transaction Date)-0.253(House Age)-0.0058(Distance to Nearest MRT Station)+260.673(Latitude)

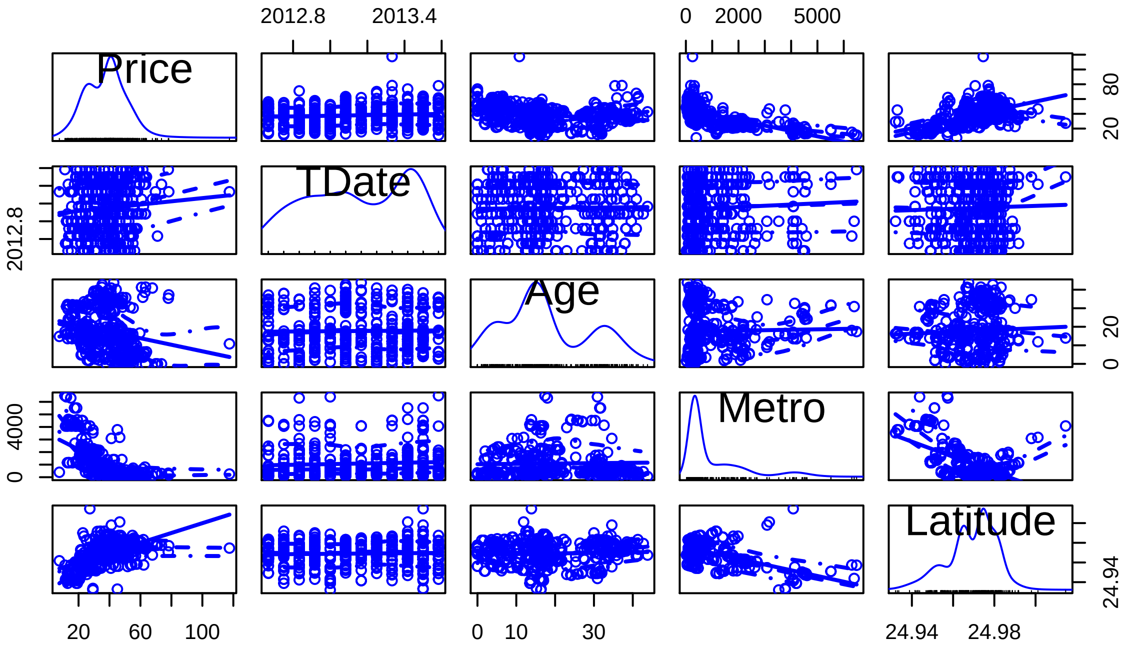
Based on the diagnostic plot for Residuals vs. Fitted values, the second model seems to violate the assumption of linearity and possibly constant variance. The QQ Plots for the first model is slightly heavy-tailed, and the second is slightly right skewed, with observation 271 notably affecting the plots. This could indicate a possible violation of the normality of errors assumption. Based on the Durbin-Watson test, both models satisfy the independence model assumption. The new model has an AIC of 1844.7 and RSS of 34808, as opposed to the first model which has an AIC of 1882.4 and RSS of 38119.

While the second model does not seem to satisfy the assumption of linearity, it will be chosen for further calculations do to its improvement through transformations and better fit.

Transforming the Predictors

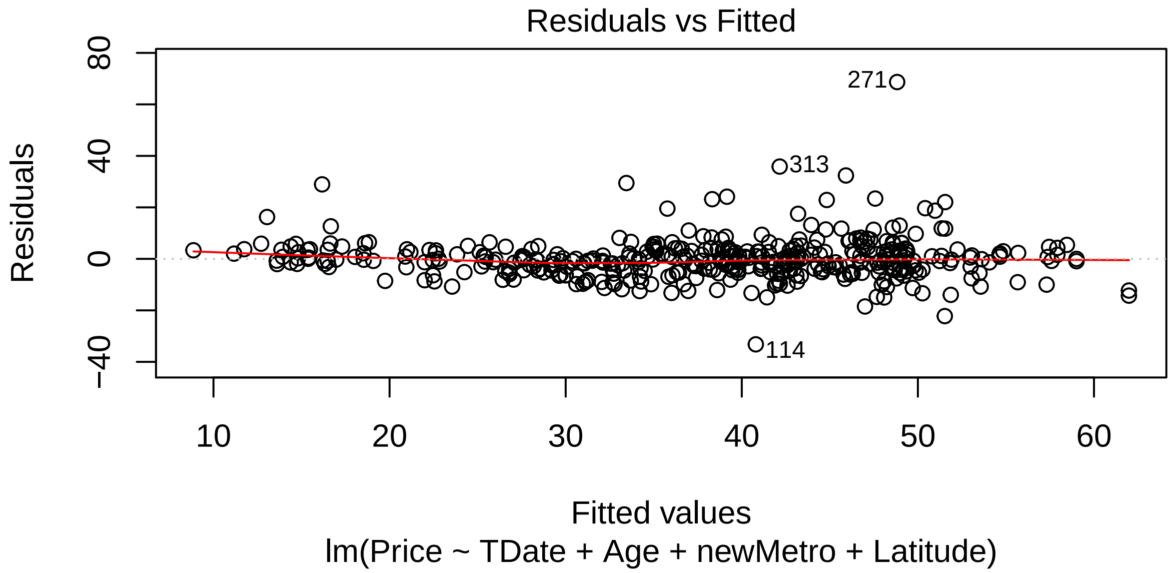
Logarithmic transformations are easily interpretable and are preferred over other transformations when they are appropriate to the model. Using a scatterplot matrix, one can see relationships between predictors and the response that may indicate the benefit of a logarithmic transformation of the preedictor.

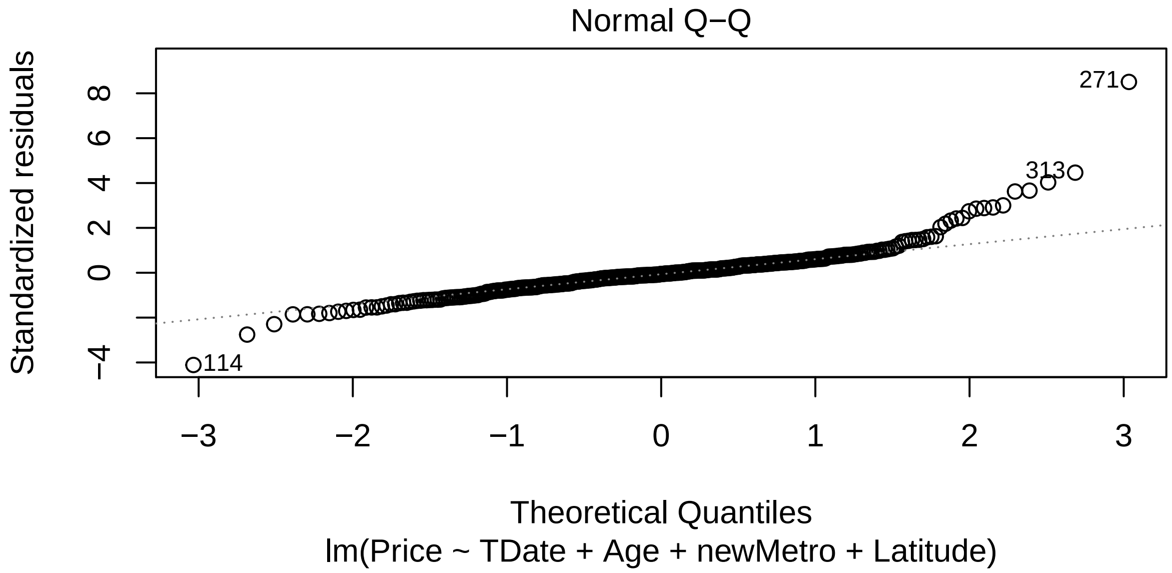
```
scatterplotMatrix(~Price+TDate+Age+Metro+Latitude)
```



Based on the scatterplot matrix, Metro clearly has an exponential relationship with Price, so applying the natural logarithm to Metro is an appropriate consideration.

```
newMetro=log(Metro)
transform<-lm(Price~TDate+Age+newMetro+Latitude)

plot(transform, which=1:2)
```





```
c(summary(fit2)$r.squared, summary(transform)$r.squared)
```

```
## [1] 0.5447599 0.6492648
```

```r
c(AIC(fit2), AIC(transform))
```

```
## [1] 3021.623 2913.655
```

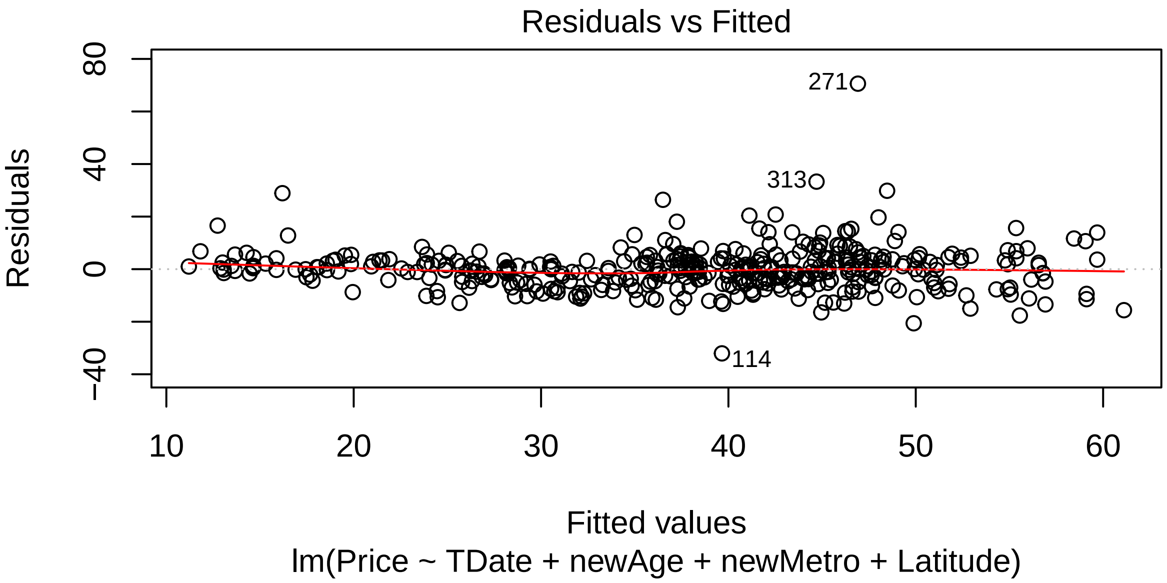
The transformation notably improved the models adherence to the linearity assumption. However, the model still produces a heavy-tailed QQ plot.

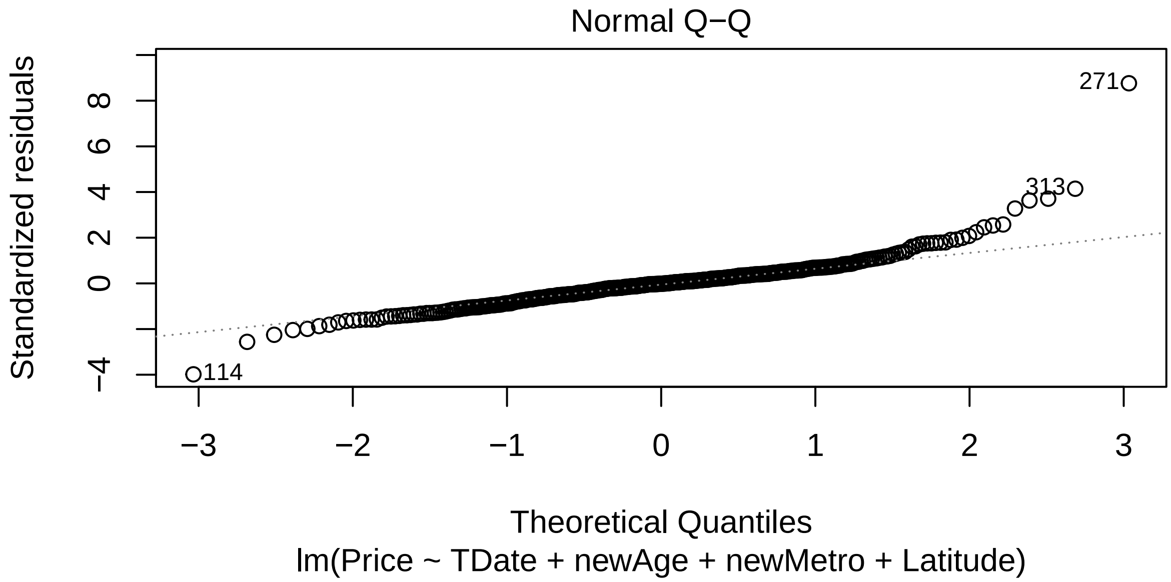
The new model with the logarithm transformed predictor Metro has a larger R-Squared value, namely 0.6493 as opposed to the R-Squared value of 0.5448, as well as a lesser AIC, indicating that the transformation did improve the model fit.

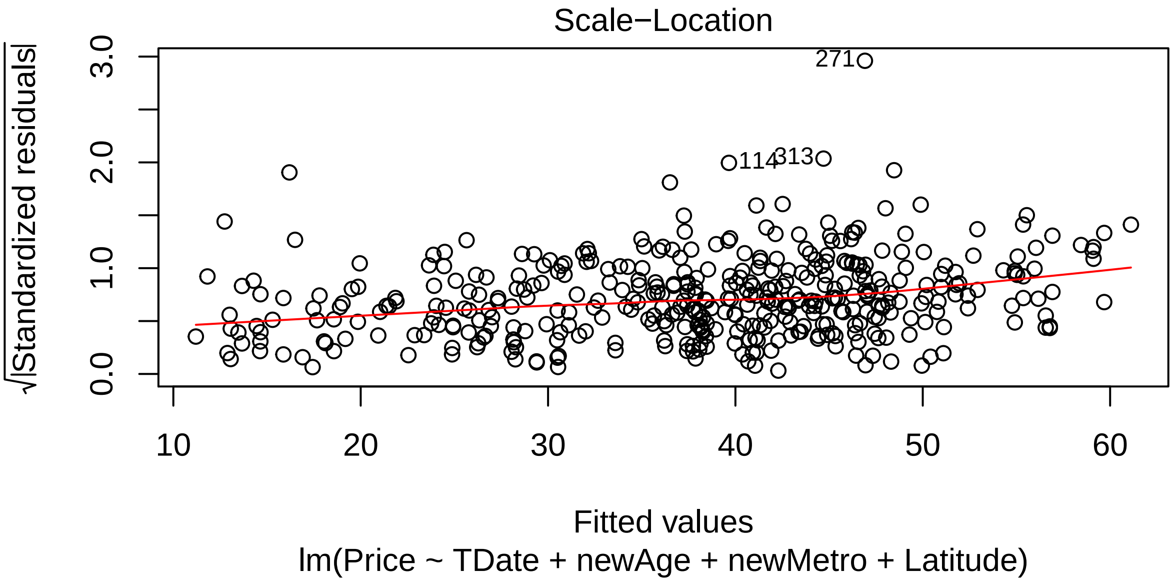
This transformation is appropriate to the model and will be used in the later transformation of the response.

Another logarithmic transformation to consider based on the scatterplot matrix is the transformation of the House Age predictor. Since the Age variable contains observations with values equal to 0, the variable must be modified before the logarithm can be applied:

```r
newAge=log(Age+0.0001)
newTransform<-lm(Price~TDate+newAge+newMetro+Latitude)
plot(newTransform, which=1:3)
```







```
c(summary(transform)$r.squared, summary(newTransform)$r.squared)
```

```
## [1] 0.6492648 0.6511630
```

```
c(AIC(transform), AIC(newTransform))
```

```
## [1] 2913.655 2911.408
```

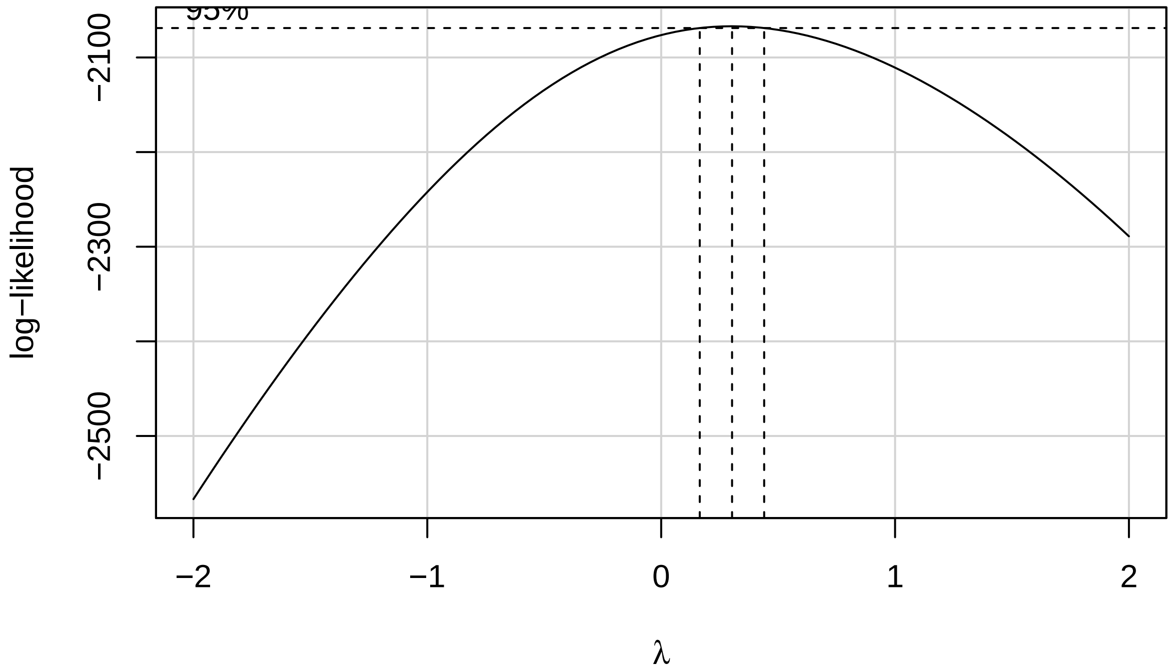
The transformation slightly improved the model's adherence to the normality of errors assumption, but also introduced a possible violation of the constant variance assumption, as there does seem to be a positive relationship between the residuals and the fitted values. The transformation did not improve the model's adherence to the independence of errors assumption.

While the R-Squared value increased from 0.649 to 0.651 and the AIC decreased from 2913.655 to 2911.408 with the application of the logarithm to Age, the transformation did not improve adherence to model assumptions enough to be considered signifcantly consequential to warrant the increased complexity of the transformation.

Therefore, this transformation is not neccesarily appropriate for the model, and will not be included in the following transformation of the reponse.

Transforming the Response

```
boxCox(transform)
```



While $\lambda = 0$ is not contained within the 95% confidence interval of the Box-Cox transformation, it is close enough to the interval to warrant consideration as a possible transformation. The following will compare the transformation using the maximum likelihood estimation of λ to the logarithmic transformation of the response Price:

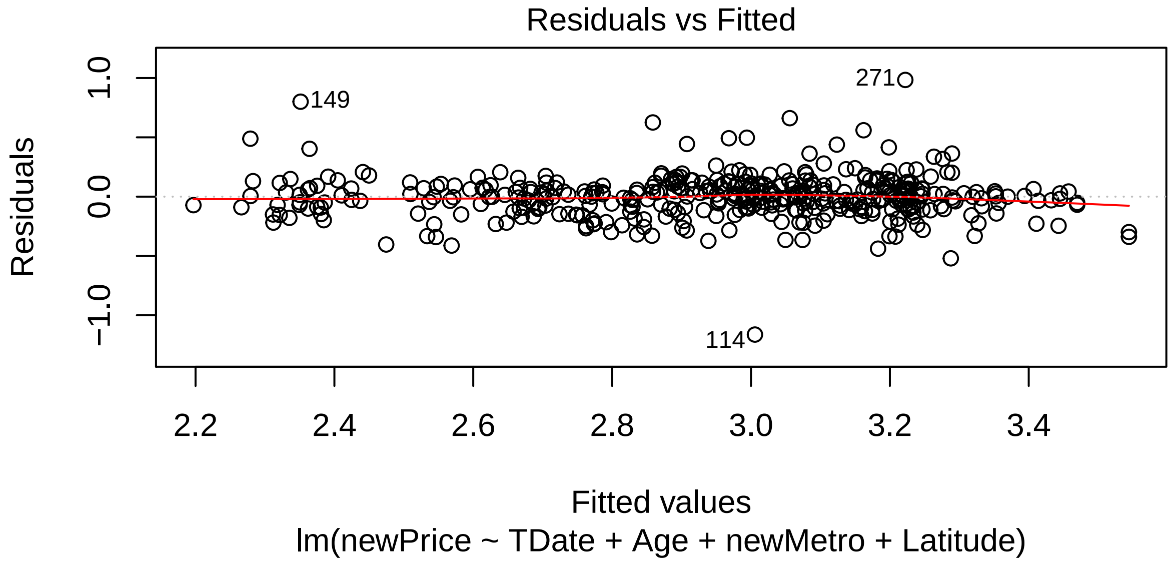
```
powerTransform(transform)
```

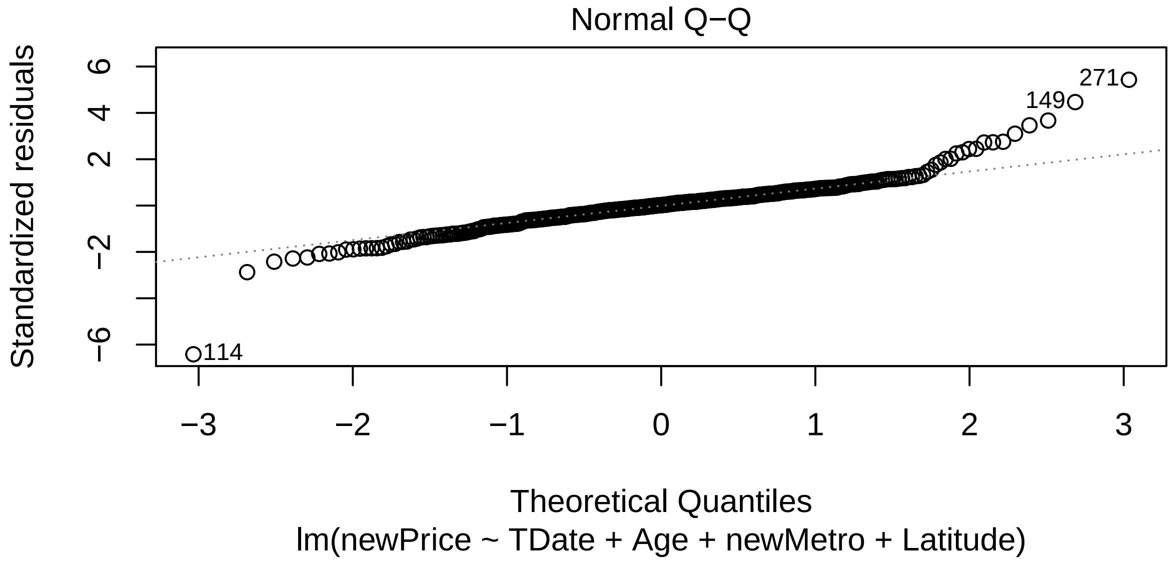
```
## Estimated transformation parameter
##        Y1
## 0.3013868
```

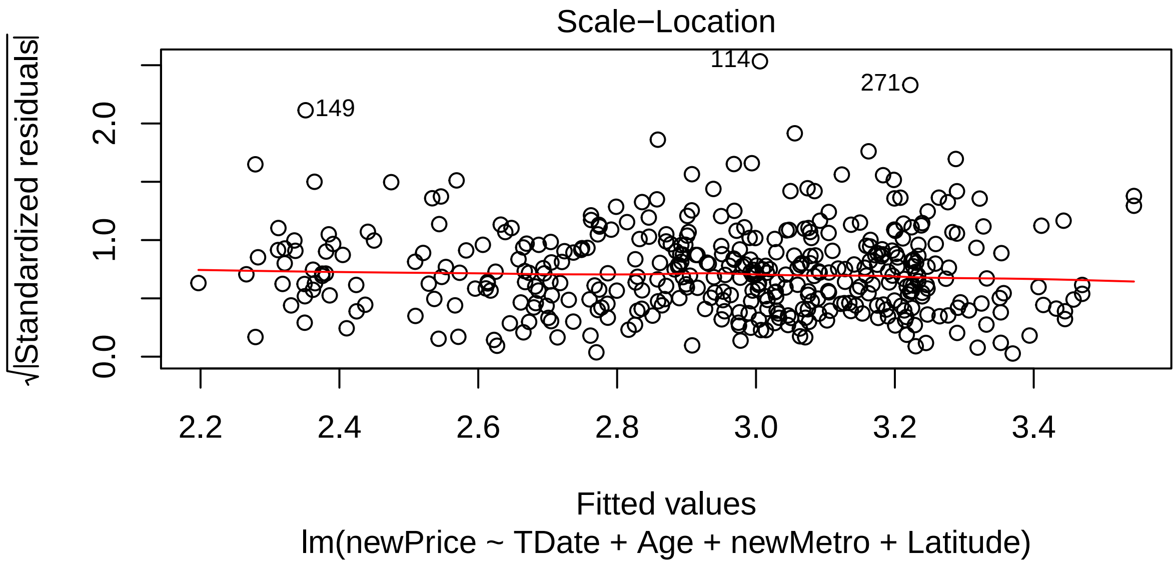
The maximum likelihood estimator for λ is 0.3013868.

```
newPrice=(Price)^(0.3013868)
exptransform<-lm(newPrice~TDate+Age+newMetro+Latitude)
```

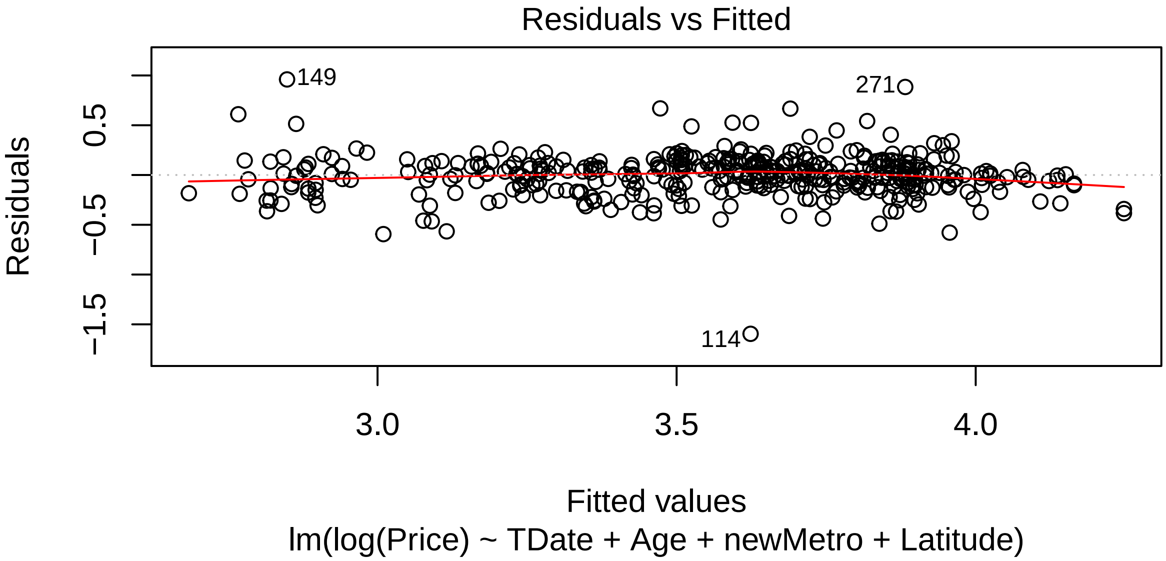
```
logtransform<-lm(log(Price)~TDate+Age+newMetro+Latitude)

plot(exptransform, which=1:3)
```

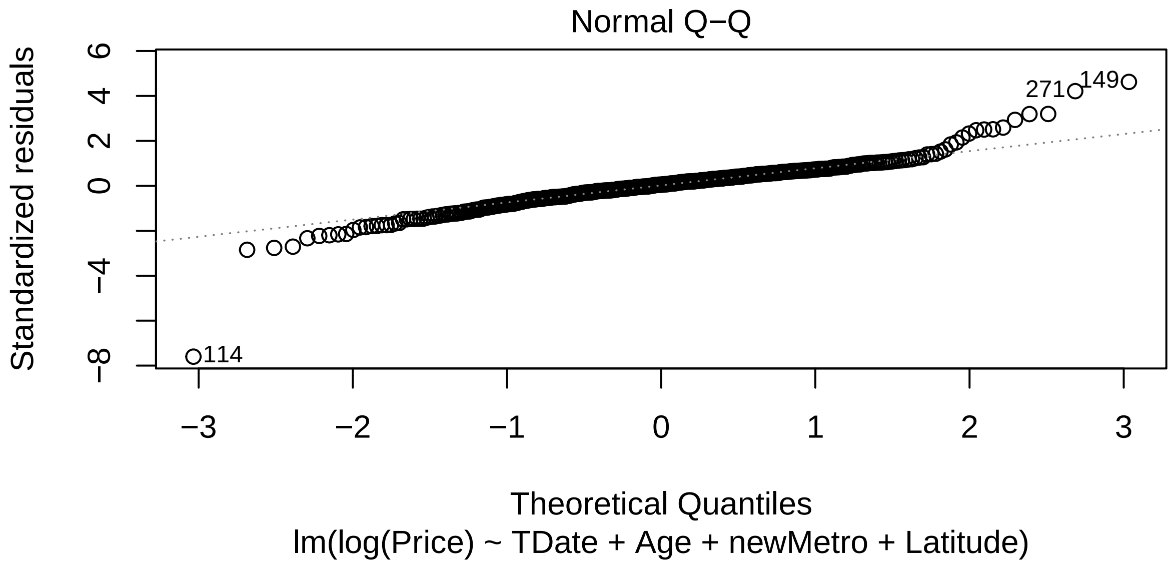


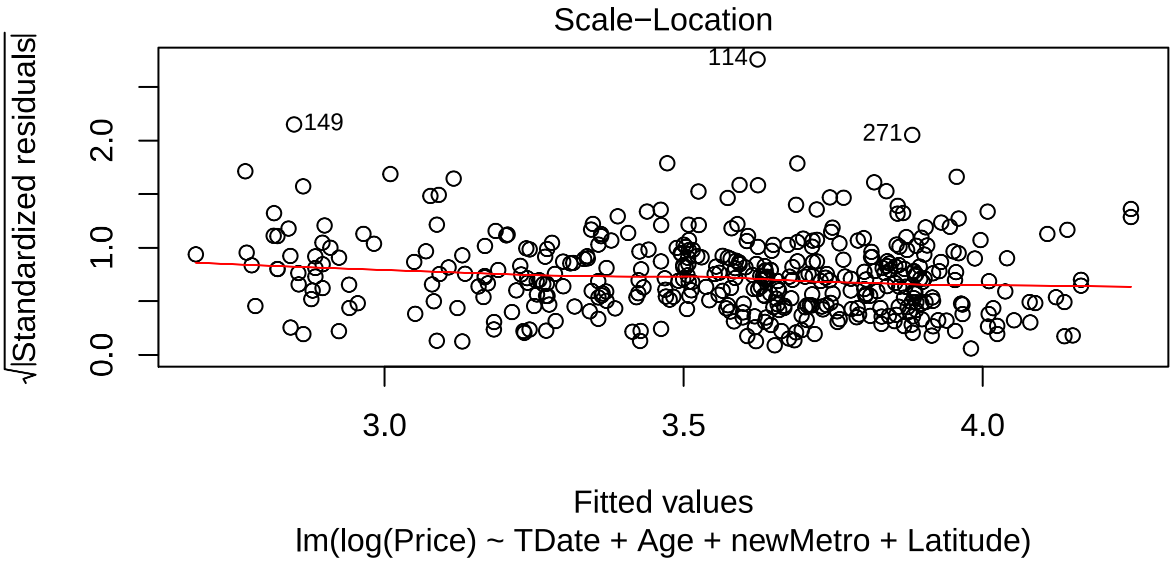




```
plot(logtransform, which=1:3)
```







```
c(summary(exptransform)$r.squared, summary(logtransform)$r.squared)
```

```
## [1] 0.7103601 0.7151100
```

```
c(AIC(exptransform), AIC(logtransform))
```

```
## [1] -230.0261 -108.3801
```

The logarithmically transformed response marginally improved the model's adherence to the normality of errors assumption while the maximum likeliehood power transformation had little effect on the assumption.

The exponentially transformed response model has an AIC of -230.0261, as compareed to the AIC of the logarithmically transformed response model with an AIC of -108.3801. However, the R-Squared value of the exponentially transformed model is 0.710 as opposed to the R-Squared value of 0.715 for the logarithmically transformed model. Due to the ease of interpretation of the logarithmic transformation and the slight difference in model fit, it would then be preferable to use the following model:

```
logtransform$coefficients
```

```
##   (Intercept)         TDate           Age      newMetro      Latitude 
## -6.202587e+02  1.743104e-01 -5.848415e-03 -2.113611e-01  1.098830e+01
```

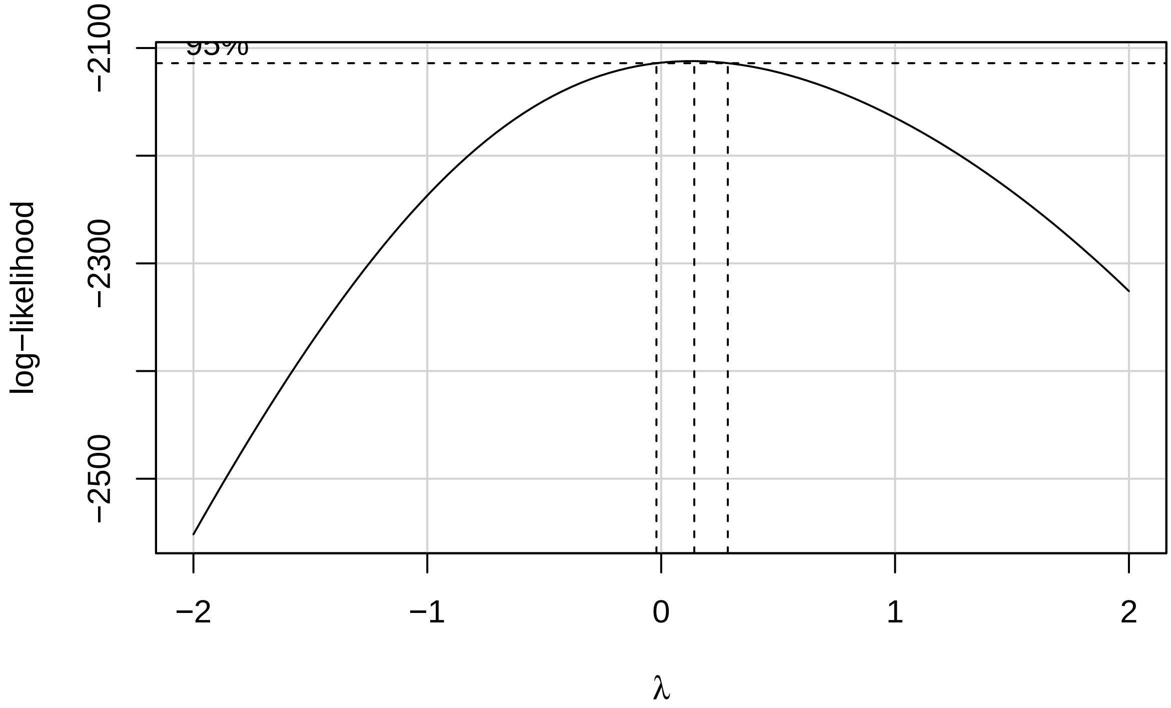
log(Price)=-620.259+0.174(Transaction Date)-0.00585(House Age)-0.211(log(Metro))+10.988(Latitude).

Box-Tidwell Transformations

Another set of predictor transformations to consider is the set of Box-Tidwell power transformations. While they may not present the same ease of interpretation as logarithmic transformations, they may provide a better model fit.

This transformation will use the logarithmic transformation of the response Price, but the non-transformed values of Metro, the following Box-Cox plot shows that the logarithm is still an appropriate transformation for Price given that Metro is not transformed.

```
boxCox(fit2)
```



The Box-Tidwell transformations require that all predictors be strictly positive. As previously mentioned, the variable Age contains values equal to 0, so it will have to be modified once again. The variable Latitude will neccesarily be modified so that the estimates of λ are calculable by the boxTidwell function.

```
modifiedAge=Age+0.001
modifiedLatitude=Latitude/25
```

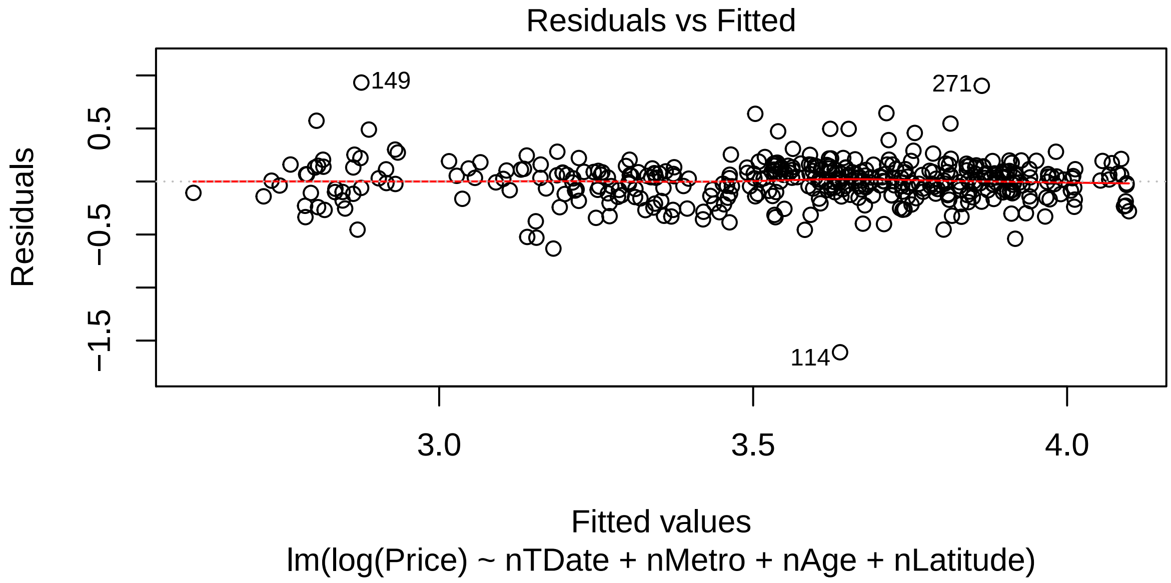
```
BT<-boxTidwell(formula=log(Price)~ TDate + modifiedAge + Metro + modifiedLatitude)

nTDate=TDate^(BT$result[1])
nAge=modifiedAge^(BT$result[2])
nMetro=Metro^(BT$result[3])
nLatitude=modifiedLatitude^(BT$result[4])
c(BT$result[1], BT$result[2], BT$result[3], BT$result[4])
```

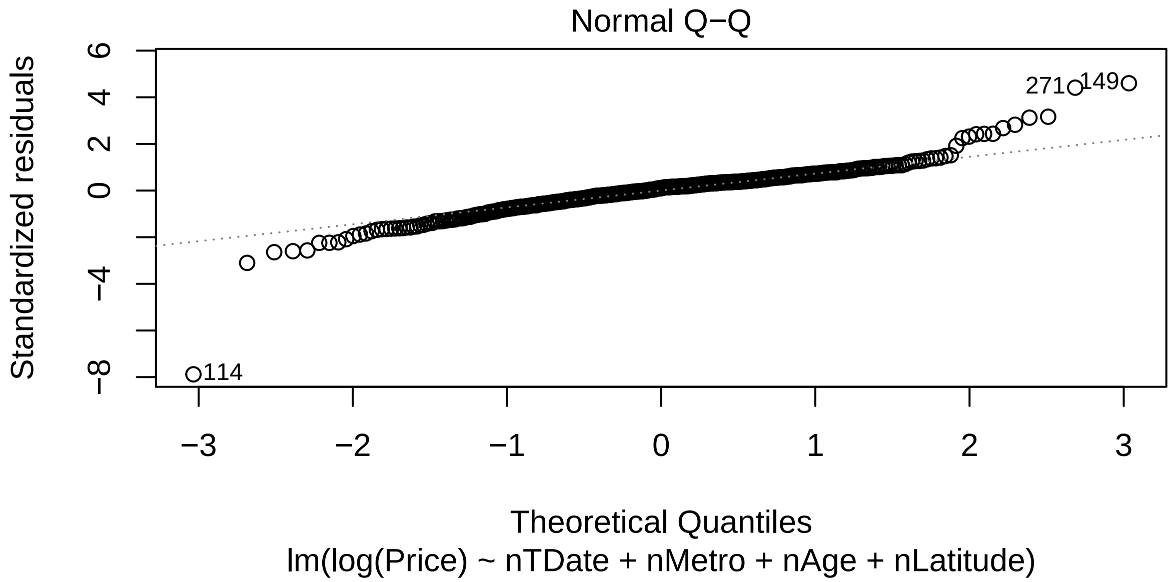
```
## [1]   4.3757998   0.3751708   0.2457525 286.8337133
```

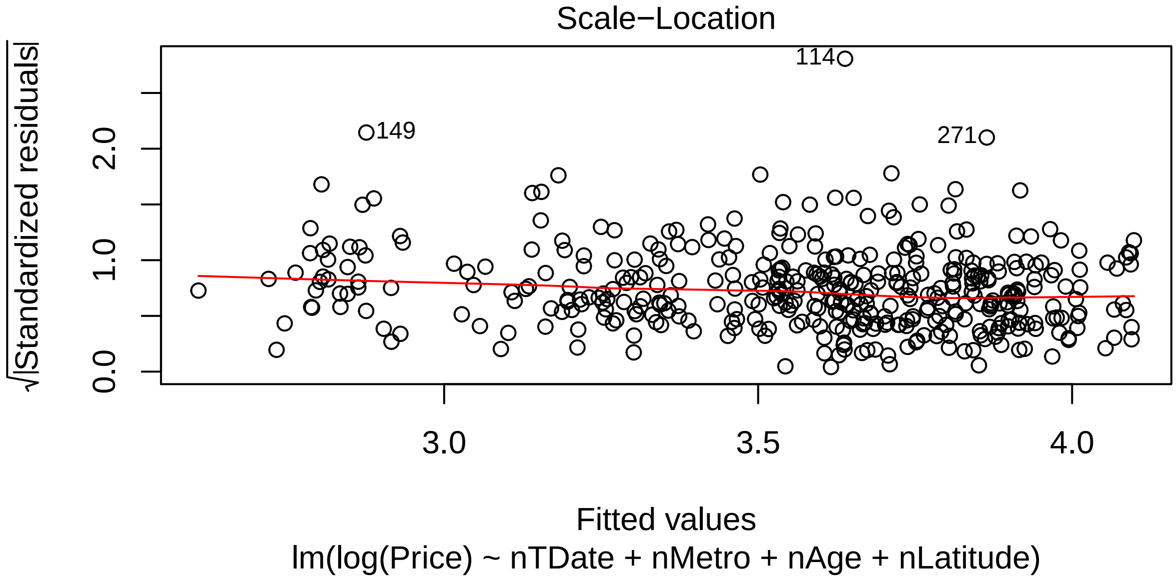
```
BTfit<-lm(log(Price)~nTDate + nMetro + nAge +nLatitude)
BTfit$coefficients
```

```
##   (Intercept)        nTDate        nMetro          nAge     nLatitude
## -7.800790e+01  2.857458e-13 -1.764955e-01 -8.600606e-02  1.171455e+00
```

```
plot(BTfit, which=1:3)
```







```
lmtest::dwtest(BTfit)[4]
```

```
## $p.value
## [1] 0.856475
```

```
nAIC<-AIC(BTfit)
nRSQ<-summary(BTfit)$r.squared

c(nAIC, nRSQ)
```

```
## [1] -130.0668385    0.7296493
```

Therefore, the model found using the Box-Tidwell maximum likelihood estimators is:

$\log(\text{Price}) = -78.008 + 2.857 * 10^{-13}(\text{TDate})^{4.376} - 0.176(\text{Metro})^{0.375} - 0.0086(\text{Age}+0.001)^{0.246} + 1.171(\text{Latitude}/25)^{286.834}$.

This model has the closest adherence to the noramlity of errors assumption from the transformations seen. It also satisfies the independence, constant variance, and linearity assumptions.

This model has an AIC of -130.067, and an R-Squared value of 0.730. Due to the complexity of the model, it is not practical as compared to the model found through the Box-Cox transformation.

Real Estate Valuation Conclusions

The intial linear models both exhbited the expected linear relationships between the individual predictors and the response, even after accounting for the other predictors in the model.

Interestingly, the Transaction Date predictor had a P-Value of 0.0327 in the first linear model and a P-Value of 0.000642 in the second linear model. The removal of Stores and addition of Metro substantially increased the significance of Transaction Date.

The most significant predictors in the fitted models were Distance to the Nearest MRT Station, Number of Convenience Stores in Proximity, and Latitude. All three of these predictors relate to the location of the observed home. In particular, all three predictors and their relationships found in the models show that proximity to urban centers is the most important factor impacting Price per Unit Area.

Concrete Compressive Strength Data Analysis

Forward Model Selection

```r
hpc <- ~X1+X2+X3+X4+X5+X6+X7+X8
mod.0 <- lm(Y ~ 1, data = concrete)
mod.full <- lm(Y ~ X1+X2+X3+X4+X5+X6+X7+X8, data = concrete)
model<-step(mod.0, hpc, direction = "forward", k=log(1030))
```

```
## Start:  AIC=5806.38
## Y ~ 1
##
##        Df Sum of Sq    RSS    AIC
## + X1    1     71172 216001 5520.0
## + X5    1     38490 248683 5665.1
## + X8    1     31061 256112 5695.4
## + X4    1     24087 263086 5723.1
## + X7    1      8033 279140 5784.1
## + X6    1      7811 279362 5784.9
## + X2    1      5220 281953 5794.4
## + X3    1      3212 283961 5801.7
## <none>              287173 5806.4
##
## Step:  AIC=5519.97
## Y ~ X1
##
##        Df Sum of Sq    RSS    AIC
## + X5    1   29646.5 186354 5374.8
## + X8    1   23993.8 192007 5405.6
## + X2    1   22957.4 193043 5411.2
## + X4    1   17926.8 198074 5437.7
## + X6    1    3548.0 212453 5509.8
## + X3    1    2894.4 213106 5513.0
## <none>              216001 5520.0
## + X7    1     960.2 215041 5522.3
##
## Step:  AIC=5374.85
## Y ~ X1 + X5
##
##        Df Sum of Sq    RSS    AIC
## + X8    1     37498 148857 5150.4
## + X2    1     19456 166898 5268.2
## + X7    1      5862 180493 5348.9
## <none>              186354 5374.8
## + X4    1       782 185572 5377.5
## + X3    1       741 185613 5377.7
## + X6    1       241 186113 5380.4
##
## Step:  AIC=5150.38
## Y ~ X1 + X5 + X8
##
##        Df Sum of Sq    RSS    AIC
## + X2    1   19908.5 128948 5009.4
## + X4    1    4868.8 143988 5123.1
## + X7    1    3385.5 145471 5133.6
## <none>              148857 5150.4
## + X3    1     323.9 148533 5155.1
## + X6    1      36.9 148820 5157.1
##
## Step:  AIC=5009.43
## Y ~ X1 + X5 + X8 + X2
##
##        Df Sum of Sq    RSS    AIC
## + X4    1    9544.7 119403 4937.2
## + X3    1    6524.7 122423 4962.9
## + X6    1    1737.0 127211 5002.4
## <none>              128948 5009.4
## + X7    1       3.5 128945 5016.3
##
## Step:  AIC=4937.16
## Y ~ X1 + X5 + X8 + X2 + X4
##
##        Df Sum of Sq    RSS    AIC
## + X3    1    8547.4 110856 4867.6
## + X7    1    1895.7 117508 4927.6
## <none>              119403 4937.2
## + X6    1      24.1 119379 4943.9
##
## Step:  AIC=4867.59
## Y ~ X1 + X5 + X8 + X2 + X4 + X3
##
##        Df Sum of Sq    RSS    AIC
## <none>              110856 4867.6
## + X6    1    44.271 110812 4874.1
## + X7    1    29.398 110827 4874.3
```

The forward selection algorithm with BIC as a criterion function gives the following linear model:

Y=29.03022+0.10543(X1)+0.08649(X2)+0.06871(X3)-0.21829(X4)+0.239(X5)+0.11349(X8)

Model Diagnostics

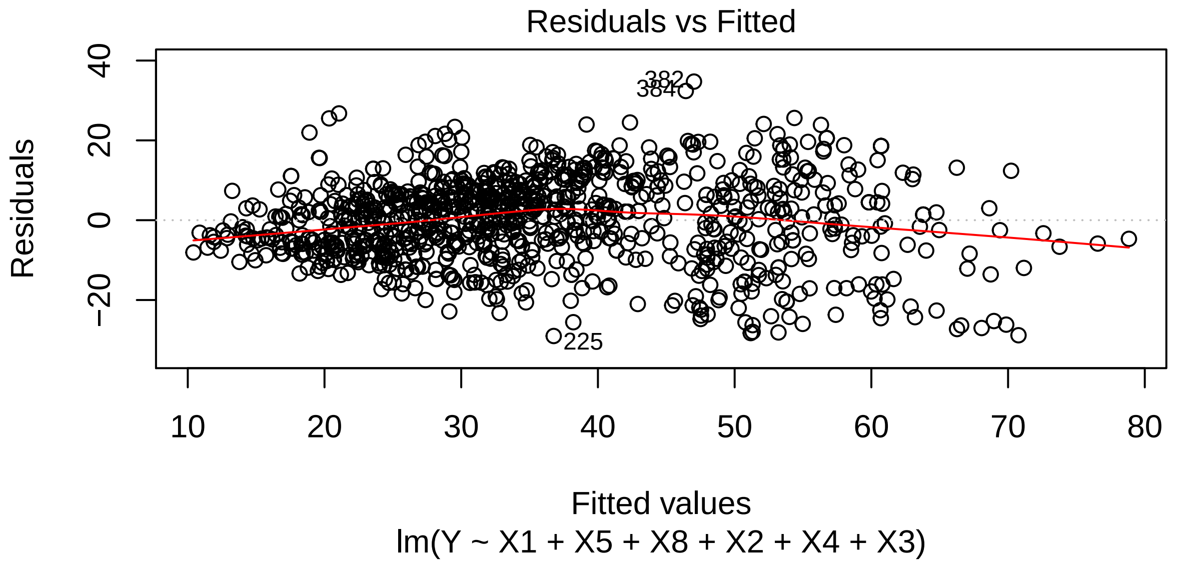
Linearity/Constant Variance

The first model assumption is that the mean of the response

$E[Y_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_8 x_{8i}$

is a linear function of \$X1, X2, X3, X4, X5, \$ and $X8$. This assumption can be verified if $E[e_i]$ appears to be 0.

The second model assumption is that the errors ϵ_i, and therefore the responses Y_i, have equal variances σ^2.

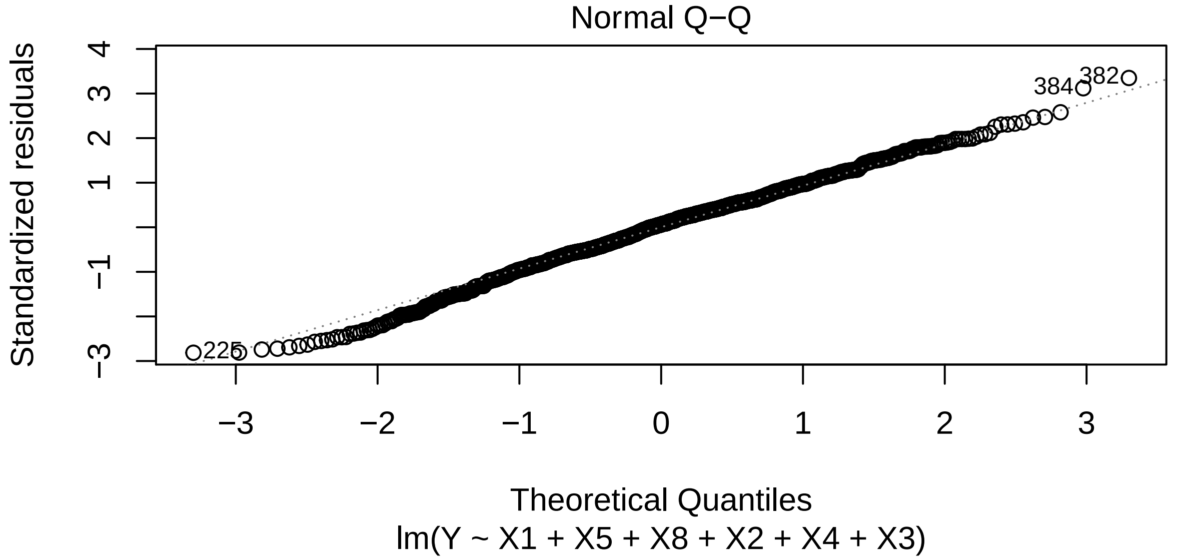


The residuals vs. fitted values plot shows the residuals tend to 0. Therefore it can be concluded that $E[Y_i] = 0$ and that the model satisfies the linearity assumption.

The plot also shows that the residuals are constantly varied in a horizontal band as the fitted values change. Therefore it can be concluded that the errors ϵ_i have constant variance and that the model satisfies the constant variance assumption.

Normality of Errors

The third assumption is that the errors are normally distributed: $\epsilon_i \sim N(0, \sigma^2)$. This assumption can be verified by a QQ Plot.



The plot forms a line that is roughly straight, indicating both sets of quantiles come from the same distribution. Therefore it is implied that $\epsilon_i \sim N(0, \sigma^2)$ and that the normality of errors assumption is satisfied.

Independence of Errors

The fourth model assumption is that the errors are independent of each other. This assumption can be verified through the Durbin-Watson Test:

```
lmtest::dwtest(model)
```

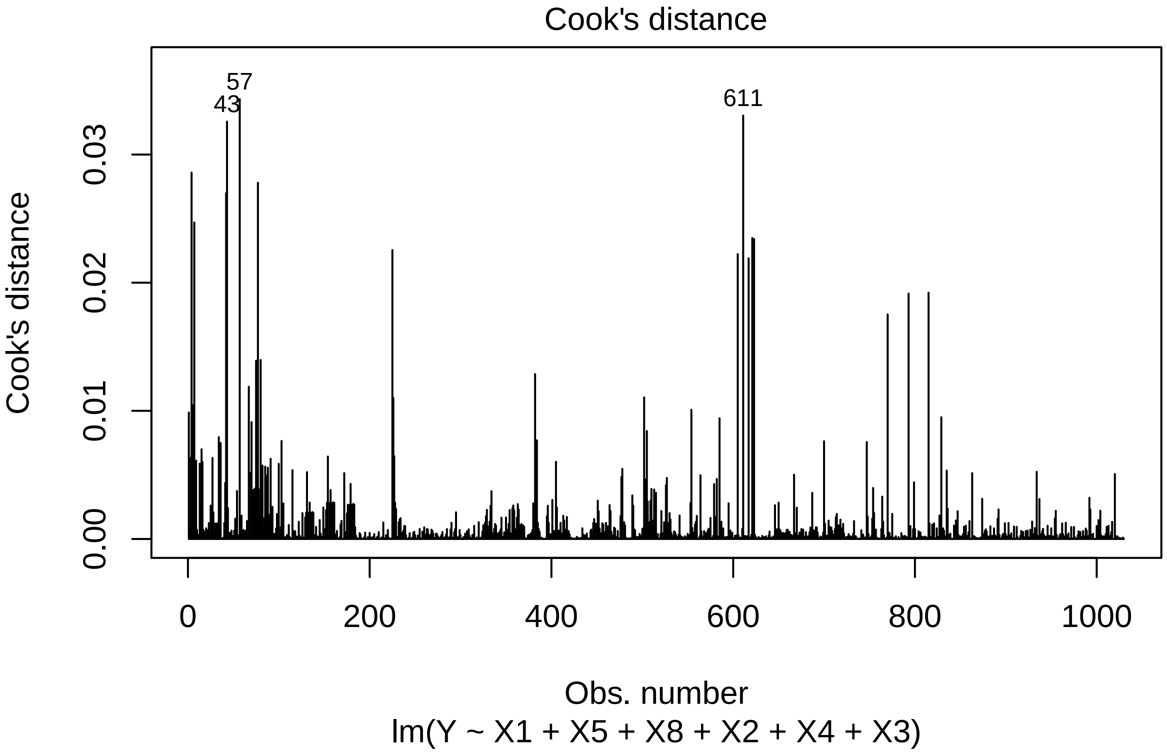
```
##
##  Durbin-Watson test
##
## data:  model
## DW = 1.2859, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

The P-Value is less than $2.2 * 10^{-16}$, implying that there is autocorrelation in the errors and that the independence assumption is not satisfied.

Influential Observations

Cook's Distance can be used as a measure of an observation's influence, the following plot will aim to identify influential observations in the model:

```
cutoff<-4/(1030-7-1)
plot(model, which=4, cook.levels=cutoff)
```



From the Cook's Distance plot, observations 43, 57, 225, and 611 stand out as influential observations relative to the other observations in the data.

```
hii<-hatvalues(model)
ei<-model$residuals
s.hat<-10.4
p=7

ri <- ei/(s.hat*sqrt(1-hii))
di <- (1/(p))*ri^2*(hii/(1-hii))
ti <- ri*sqrt((1030-p-1)/(1030-p-ri^2))
df<-data.frame(ei, hii, ri, di, ti)
subset(df, di>0.03)
```

```
##            ei        hii        ri         di        ti
## 43  -27.27104 0.03118460 -2.664084 0.03263599 -2.672067
## 57  -28.82860 0.02951080 -2.813810 0.03439393 -2.823382
## 611 -28.12314 0.02982989 -2.745406 0.03310693 -2.754229
```

```
c(mean(hii), mean(ei))
```

```
## [1] 6.796117e-03 8.727301e-17
```

The mean leverage is 0.006796, the mean residual is $8.727 * 10^{-17}$.

Observation 43 has a leverage of 0.0312 and a residual of -27.271,

Observation 57 has a leverage of 0.0295 and a residual of -28.829,

Observation 611 has a leverage of 0.0298 and a residual of -28.123

Based on the Cook's Distance, leverage, and residuals, these observations are clearly influential.

Removing Influential Points

To improve model fit and satisfaction of assumptions, one may consider removing the influential points previously identified.

```
concrete1=concrete[-43, ]
concrete1=concrete1[-57, ]
concrete1=concrete1[-611, ]
newmodel<-lm(Y~X1+X2+X3+X4+X5+X8, data=concrete1)
lmtest::dwtest(newmodel)
```

```
##
##  Durbin-Watson test
##
## data:  newmodel
## DW = 1.3033, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

```
c(AIC(model), AIC(newmodel))
```

```
## [1] 7758.047 7729.915
```

The new model with the influential points removed still fails to satisfy the independence of errors assumption. The model has a lesser AIC of 7729.915 compared to the AIC of 7758.047 from the original model. While this model is a slightly better fit, the original model will be used in further calculations.

Confidence Interval for Mean Response

Let N denote a new observation with the following predictor values:

```
##    x1 x2 x3  x4 x5 x8
## 1  X1 X2 X3  X4 X4 X8
## 2 280 74 54 182  6 46
```

```
predict(model, newdata=data.frame(X1=280, X2=74, X3=54, X4=182, X5=6, X8=46), interval = "confidence", 
```

```
##        fit      lwr      upr
## 1 35.58627 34.94936 36.22318
```

The estimated mean response for the N-th observation is 35.5862. A 95% confidence interval for the mean N-th response is (34.949, 36.223). Therefore we are 95% confident that the interval (34.949, 36.223) contains the true mean response for the N-th observation.

Prediction Interval for Response

Consider the N-th observation from the confidence interval calculation:

```
predict(model, newdata=data.frame(X1=280, X2=74, X3=54, X4=182, X5=6, X8=46), interval = "prediction", 
```

```
##        fit      lwr      upr
## 1 35.58627 15.14937 56.02317
```

The fitted response for this observation is 35.586. A 95% prediction interval for the N-th response is (15.149, 56.023). Therefore we are 95% confident that the N-th response will fall in the interval (15.149, 56.023).

Backward Model Selection

A new model will be found using the backward selection algorithm and BIC as a criterion function.

```
n=1030
model2<-step(mod.full, scope = c(lower=~1), direction = "backward", k = log(n), data=concrete)
```

```
## 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
```

The backward model selection algorithm using BIC as a criterion function outputs the same model as was found using the forward selection algorithm. The model is as follows:

Y=29.03022+0.10543(X1)+0.08649(X2)+0.06871(X3)-0.21829(X4)+0.239(X5)+0.11349(X8)

As the backward model selection and the forward mode selection models are the same, the model diagnostics, influential points, and model quality will be the same.

Concrete Data Analysis Conclusions

From the linear model found through the model selection process, we found a positive linear relationship between the predictors included and the compressive strength of the concrete, with the water content predictor being the only exception.

Interestingly, both the forward and backward model selection algorithms select the same model when using BIC as a criterion function. This can be attributed to the small number of predictors in the model.

Conclusions:

Part I:

Our findings led us to believe that the best model to use is the Box-Cox transformation applied second model which has a logarithmically transformed response variable with the predictors: transaction date, house age, logarithmically transformed distance to the nearest MRT station, and latitude; because of its satisfaction of model assumptions, relatively low AIC (-108.3801) high R-Squared value (0.715), and the simplicity of the model. The logarithmically transformed predictor and response variable significantly improved the results of the test for normality of errors. If simplicity did not influence our decision to pick the best model, the model we found using the Box-Tidwell would have been the best since it has the lower AIC (-130.067) and higher R-Squared value (0.730). From our analysis, we conclude that the proximity to urban centers is the most important factor impacting Price per Unit Area, since the most significant predictors in the fitted models were Distance to the Nearest MRT Station, Number of Convenience Stores in Proximity, and Latitude.

Part II:

The model we got from using the forward model selection algorithm using BIC as a criterion function outputs the same model as was what we found using the backward selection algorithm, which made the model diagnostics, influential points, and model quality be the same. This is not typically the case with most data sets, however it was the case with this dataset. The relatively small number of predictors in the model can be a possible explanation for this. The model satisfied all model assumptions except for the independence of errors assumption. Even when highly influential points were removed, the model failed to satisfy this assumption.