Real Estate Valuation

Jenny Zhang Friday 10am

Annie Ma Thursday 2pm

Abstract (summary, questions of interest, findings)

Real estate in Taipei, Taiwan is influenced by a variety of factors. Convenience, proximity to transportation, wear and tear, and date of transaction are important predictors, just to name a few. We wanted to see which variables were particularly influential, and if it would be possible to accurately predict the housing price of property using a linear regression model using these predictors. We found that the best predictor to determine the value of a real estate is the distance to the nearest MRT station. This makes sense because people tend to chose to live near a MRT station for its convenient location and transportation. Although house values are highly influenced by the distance to the nearest MRT station, an exploratory analysis has shown that there are also other relevant features (e.g. house's age, the number of convenience stores nearby, house's transaction date). As a direct outcome of this research, more efficient housing values predictors can be developed, improving the estimation of the housing prices.

Introduction (relevant background of topic and motivation for the project)

Housing affordability remains a serious problem. Taipei is one of the world's most expensive cities to live in. The statistics shows that most young Taiwanese are unable to afford a home. For example, an average income in Taipei would need to spend two-thirds of their income to pay their mortgage. Many people will eventually need to buy homes, and in larger cities, real estate is becoming increasingly expensive. This will affect many people: those who are moving into a new home for their profession and those who are in the lucrative business of real estate. By studying real estate valuation of Taipei, we can analyze the influential factors that affect housing price and then effectively predict the potential cost of a property in a certain area. This will help many people choose where they want to live based on what their needs on, whether it be convenience, transportation, or location. In our case, we wanted to see if the house price of unit area would be affected by transaction date, house age, distance to the nearest MRT station, number of convenience stores, latitude and longitude. Taiwan is an urban city, meaning the inhabitants are likely busy people who rely on convenience and ease of travel. We wanted to see how prominent these factors were to these residents, if at all. If there is a correlation, then this study can be repeated to other metropolitan areas to determine influential factors in housing prices.

Attribute	Description (Domain)
X1 transaction date	house’s transaction date (integer: for example, 2013.250=2013 March, 2013.500=2013 June)
X2 house age	house’s age (integer: from 0 to 50 years)
X3 distance to the nearest MRT station	house’s distance to the nearest MRT station (integer: meters)
X4 number of convenience stores	the number of convenience stores in the living circle on foot (integer)
X5 latitude	the geographic coordinate, latitude
X6 longitude	the geographic coordinate, longitude
Y house price of unit area	house’s price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

Questions of interest

- What is the predicted mean unit house price for the averages of transaction date, house age, and distance to the nearest MRT station of an estate with no convenience stores?
- Which predictor would have the most effect on the housing price?
- Which predictor is the best way to access the value of a home?

Data and Regression Methods

We obtained our data from a study done on real estate valuation in Sindian district, New Taipei City, Taiwan. Our predictor values consist of transaction dates, house age in years, distance to the nearest MRT station in meters, and the number of convenience stores in the living circle on foot. We use these predictors to analyze the relationship and effects they have on house price of unit area: a unit measured in 10000 New Taiwan Dollar per ping (3.3 meter squared).

To test the predicted mean unit house price for the averages of transaction date, house age, and distance to the nearest MRT station of an estate with no convenience stores, we will write a confidence interval using the mean values of X1, X2, and X3 while assigning the value of X4 = 0. To test the most valuable predictor, we will write an anova table and find the the predictor with the lowest p-value.

To test which predictor would have the most effect on the housing price, we will produce a summary table. We will select the predictor with the lowest p-value to show how strongly (or weakly) the predictor effects the housing price.

To test which predictor is the best way to access the value of a home, we will use regsubsets to see which of the variables of our transformed regression model. We will select only one predictor.

Our first model with all 6 predictors did not follow the assumptions of linear regression. In the ANOVA table, the p-value for predictor X6 is 0.798203, greater than 0.05 and fails to be statistically significant. This is confirmed when looking at the Added-Variable plots, as the line for X6 is essentially horizontal meaning the slope is zero. The box plots for X5 and X6 had large variability in their outliers, suggesting that they would potentially skew our response variable. Although the quantile-quantile plot followed our assumptions, the residual vs fitted plot showed a skewed pattern suggesting that the data did not follow equal variance. Transforming a couple of the predictors would be helpful in fixing this violation of our assumption.

Transformation to our predictors will be needed. The variable X1, the transaction date is only in 2013. Transformations are unlikely to be much use here. X2 and X4 have wide ranges but we cannot use power transformations or logs because of the 0 values. We will need to add a small constant before transforming. X3 has a wide range so logarithms are likely to be appropriate for them. Each of the predictors seems to be at least modestly associated with Y (House Prices), as the mean function for each of the plots in the top row of the scatterplot matrix is not flat.

From our boxplots, it consistently shows an outlier when the House Age is 10.8 years, Distance to the Nearest MRT Station is 252.58220 meters, Number of Convenience Stores is 1, latitude and longitude are 24.97460 121.5305 respectively and our House Price is 117 house's price of unit area. To remedy this, we will remove this outlier. To check if this point is influential, we will produce a summary table for this reduced model to check whether or not the standard errors are relatively the same in comparison to the full model.

Regression Analysis, Results and Interpretation

To obtain the predicted mean unit house price for the averages of transaction date, house age, and distance to the nearest MRT station of an estate with no convenience stores, we determined a 95% confidence interval for the response utilizing a dataframe in which the predictor variables for transaction date, house age and distance to the nearest MRT station were all the mean values. We then set the value for number of convenience stores in this data frame to be equal to zero. The predicted mean unit house price is 29.55788 10000 Taiwan New Dollar per Ping. We are 95% confident that the predicted mean unit house price is between 27.82855 and 31.28721 10000 Taiwan New Dollar per ping.

To test the efficacy of predictors, we examined the summary table to the linear regression model of Y ~ X1 + sqrt(X2) + log(X3) + I(X4^(0.69)). According to the summary table, the p-values for predictors X1, X2, X3, and X4 are 8.82e-07, 1.94e-10, < 2e-16, and 0.000306 respectively. The predictor with the lowest p-value is likely the be the most valuable. In this case, < 2e-16 is the lowest value and corresponds to distance to the nearest MRT station. We conclude that the most valuable predictor to estimate real estate valuation is going to be the distance of the estate to the nearest MRT station.

To test which predictor would have the most effect on the housing price, we selected one predictor using summary table.

From our summary table it seems log(X3) is our best predictor to access the value of a home because the p-value is the smallest among the other predictors.

To test which predictor is the best way to access the value of a home, we selected one predictor using regsubsets.

In our best subset regression, X3 (the house's distance to the nearest MRT station) should be chosen if we can only choose one predictor. This makes sense because Taipen is fully equipped with a completed metro public transit system. It provides better travel experience and reduces the traffic congestion. This transport system is a fundamental factor to human activities because it is connected to the nearly all destination in Taipei. wE found that the shorter the distance to the MRT station, the higher the property values.

The model is valid and follows the assumptions of linearity, independence, normality and equal variance. The residual vs fitted plot shows that the model has a horizontal line with residuals scattered randomly and evenly around the line. There is no funneling or quadratic pattern, so the model follows linearity. We know that the model follows independence because there is no pattern or clustering between the residuals. The plot follows normality: the quantile-quantile plot shows that the residuals follow the 45 degree line. There is no positive skew nor negative skew, and there are no heavy tails or outliers. We also know our model follows equal variance. When looking at the residual vs fitted plot, the residuals are scattered randomly across the horizontal line, and the vertical width doesn't increase or decrease throughout. We can then assume that the mean of the error is zero and follows constant and equal variance.

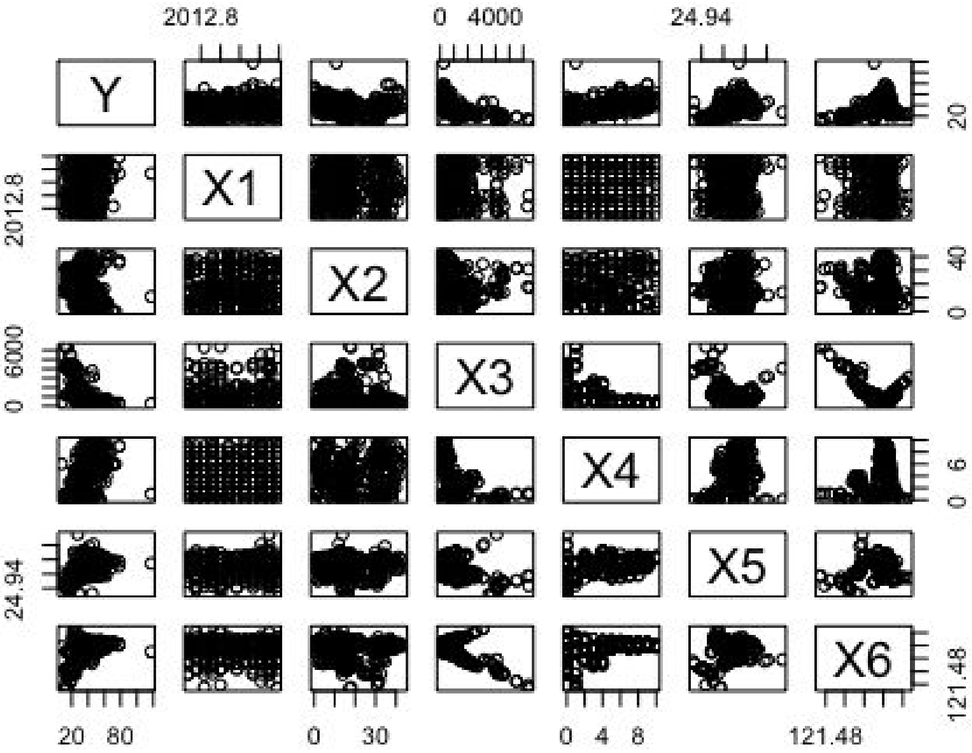
Conclusion

Accessibility to transport infrastructure is a main factor that affects property value. We used transformed our predictors then used multiple model selection to yield a final model. However, our final model is not guaranteed to be optimal in any specified sense. Our project shows it can be simple to build an accurate linear regression model to predict the house’s price, provided that the distance to the nearest MRT distance is available. However accurate the model may be, it lacks to provide insights that can allow one to intervene before purchasing an overpriced house. Home price listings are not necessarily the best estimate. Houses can be undervalued or overvalued. You cannot truly take into account the quality and condition of the real estate until you personally visited the property. For example, you cannot tell from a listing price if the house has poor plumbing or missing a shingle on the roof. A more complicated regression model would be required to understand and consider all sorts of special cases. For example, we can also take into account the number of good schooling nearby, or the real estate’s proximity to the financial or business district. Although the predictors used in the model generated a very well fitting model that follows the assumptions of linear regression, that’s not to say that there aren’t any more influential predictors that may be used in place of the predictors used in our model. If this regression was redone on another location, we would definitely want to use predictors that are influential and useful for that particular location. Overall however, the model produced very in-tune results with what the demands of the location in which the experiment was performed.

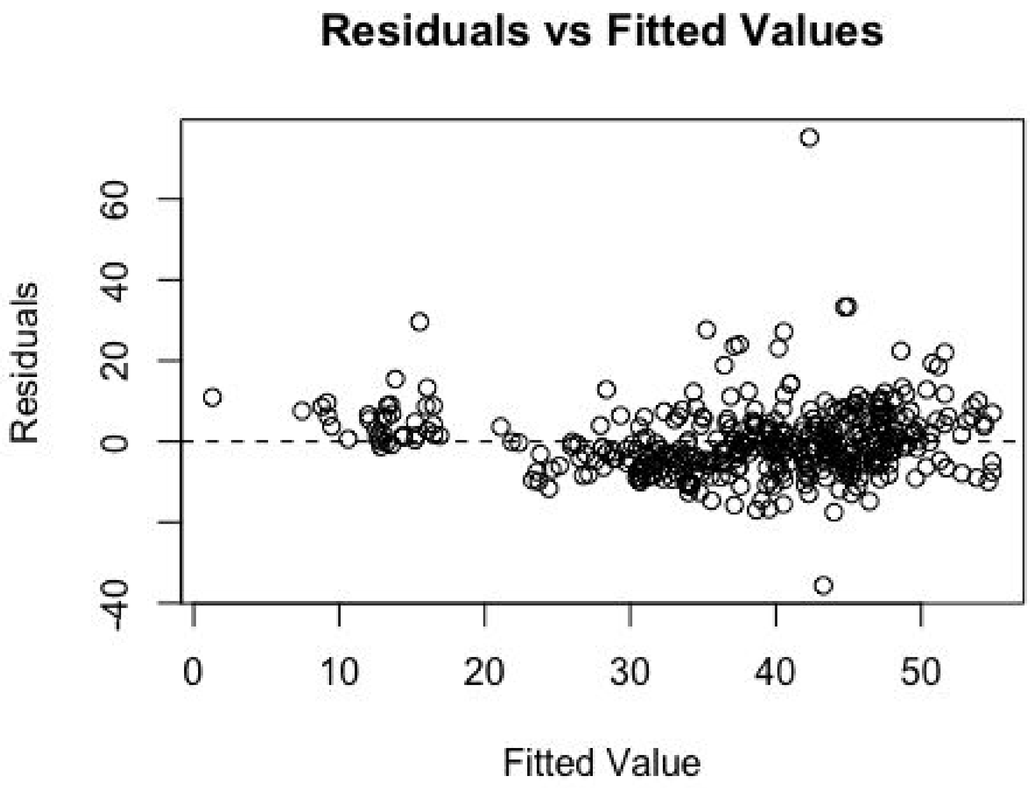
Real Estate Valuation

```
library(alr4)

## Loading required package: car

## Loading required package: carData

## Loading required package: effects

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

library(readxl)
Real_estate_valuation_data_set <-
read_excel("/Users/jennyzhang/Downloads/Real estate valuation data
set.xlsx")

## readxl works best with a newer version of the tibble package.
## You currently have tibble v1.4.2.
## Falling back to column name repair from tibble <= v1.4.2.
## Message displays once per session.

data(Real_estate_valuation_data_set)

## Warning in data(Real_estate_valuation_data_set): data set
## 'Real_estate_valuation_data_set' not found

attach(Real_estate_valuation_data_set)
Y<-Real_estate_valuation_data_set$`Y house price of unit area`
X1<-Real_estate_valuation_data_set$`X1 transaction date`
X2<-Real_estate_valuation_data_set$`X2 house age`
X3<-Real_estate_valuation_data_set$`X3 distance to the nearest MRT
station`
X4<-Real_estate_valuation_data_set$`X4 number of convenience stores`
X5<-Real_estate_valuation_data_set$`X5 latitude`
X6<-Real_estate_valuation_data_set$`X6 longitude`
realestate.lm <- lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6)
summary(realestate.lm)

##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.667  -5.412  -0.967   4.217  75.190
##
## Coefficients:
```

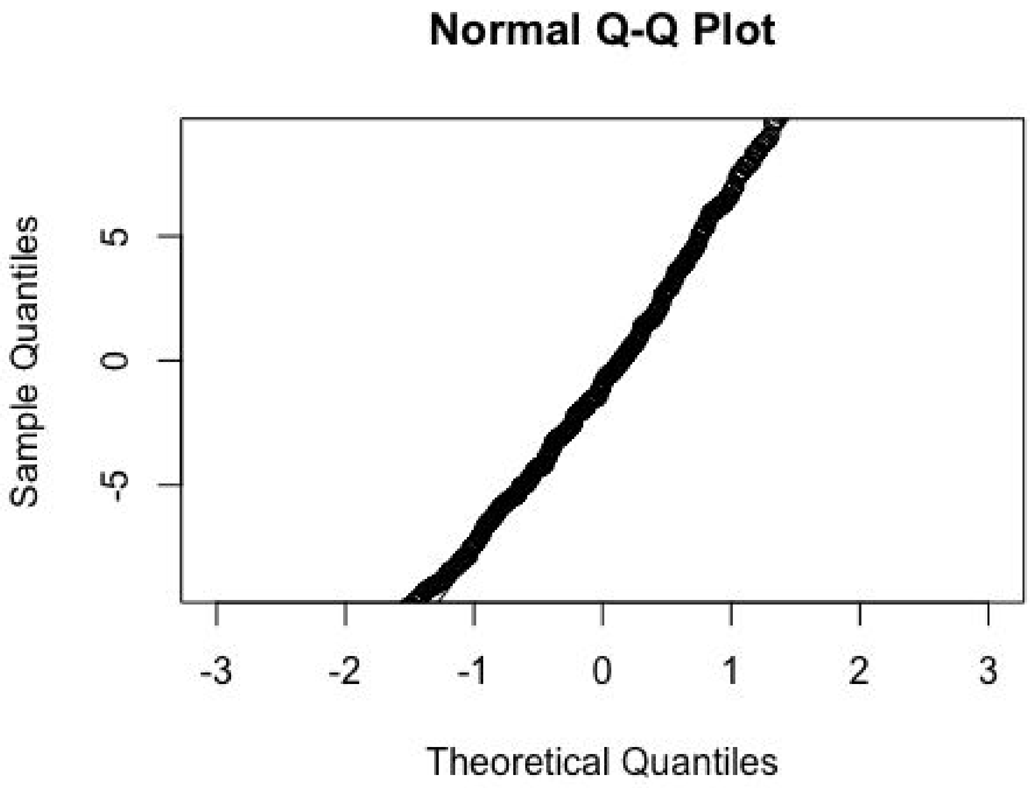
```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.444e+04  6.775e+03  -2.132  0.03364 *
## X1           5.149e+00  1.557e+00   3.307  0.00103 **
## X2          -2.697e-01  3.853e-02  -7.000 1.06e-11 ***
## X3          -4.488e-03  7.180e-04  -6.250 1.04e-09 ***
## X4           1.133e+00  1.882e-01   6.023 3.83e-09 ***
## X5           2.255e+02  4.457e+01   5.059 6.38e-07 ***
## X6          -1.243e+01  4.858e+01  -0.256  0.79820
## ---
## 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:  94.6 on 6 and 407 DF,  p-value: < 2.2e-16

pairs(Y ~ X1 + X2 + X3 + X4 + X5 + X6)
```



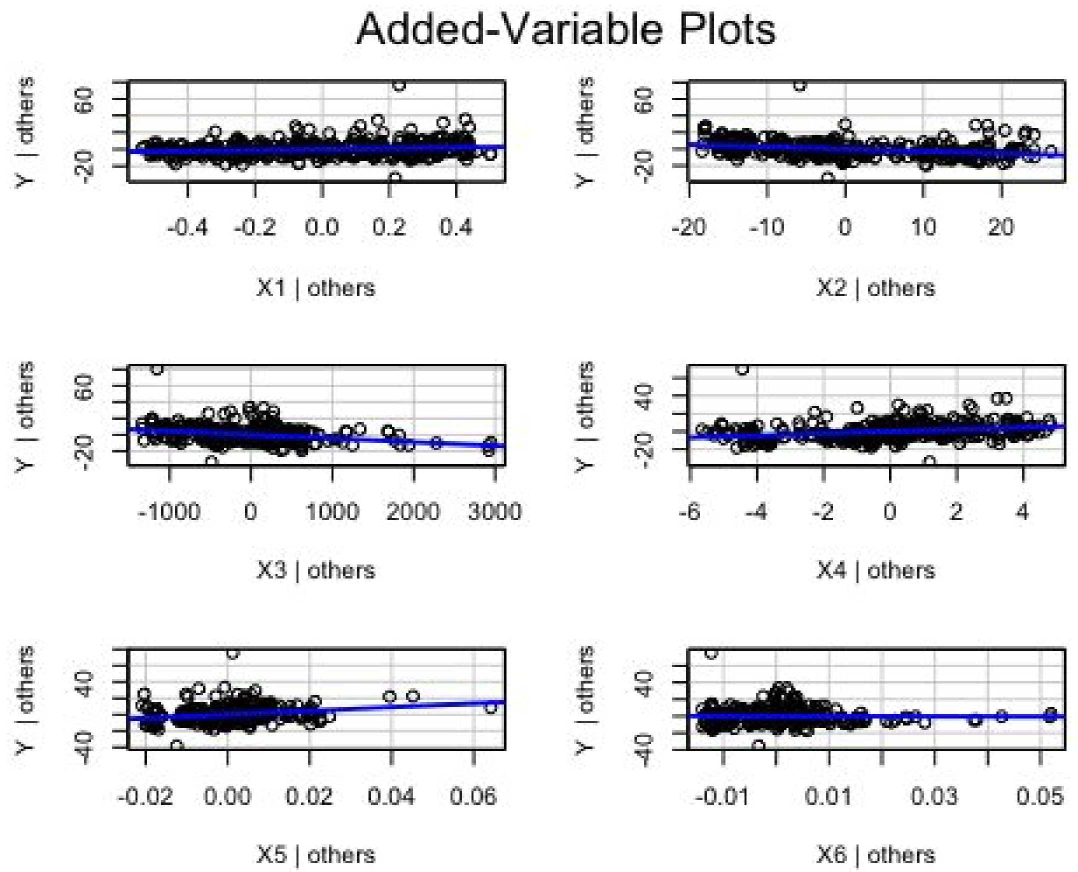
```
e <- resid(realestate.lm)
yhat2 <- fitted(realestate.lm)
plot(yhat2, e, xlab = 'Fitted Value', ylab = 'Residuals', main =
'Residuals vs Fitted Values')
abline(h = 0, lty = 2)
```



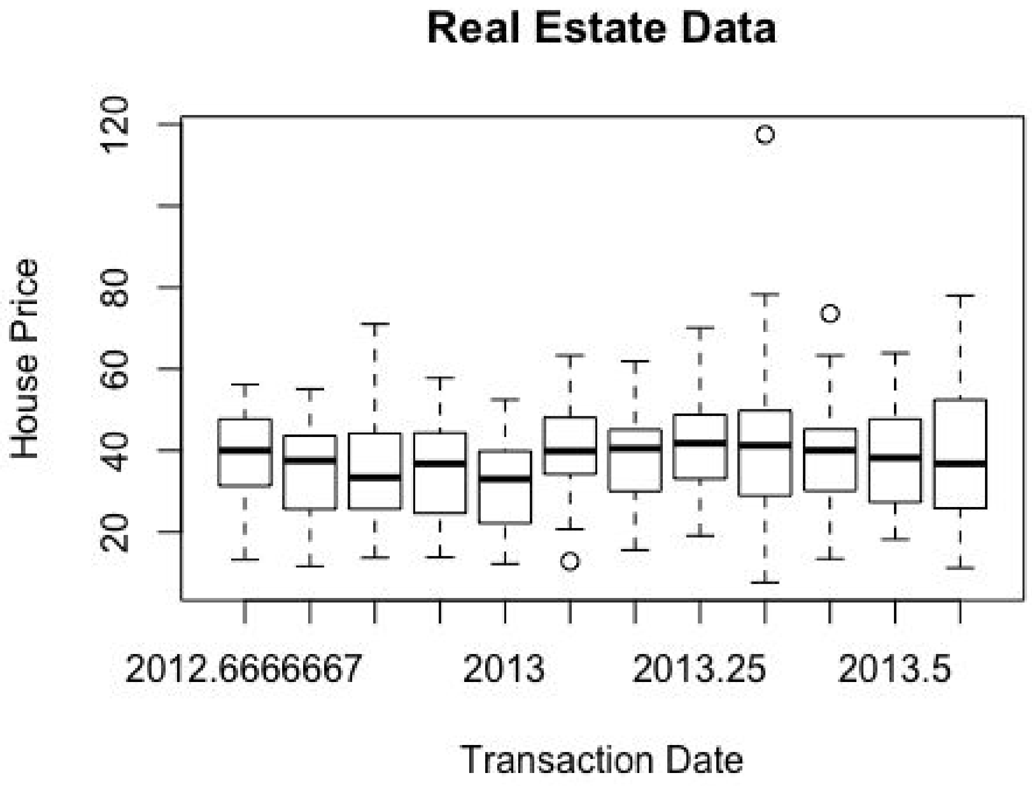
```
qqnorm(e, ylim = c(-9, 9))
qqline(e)
```



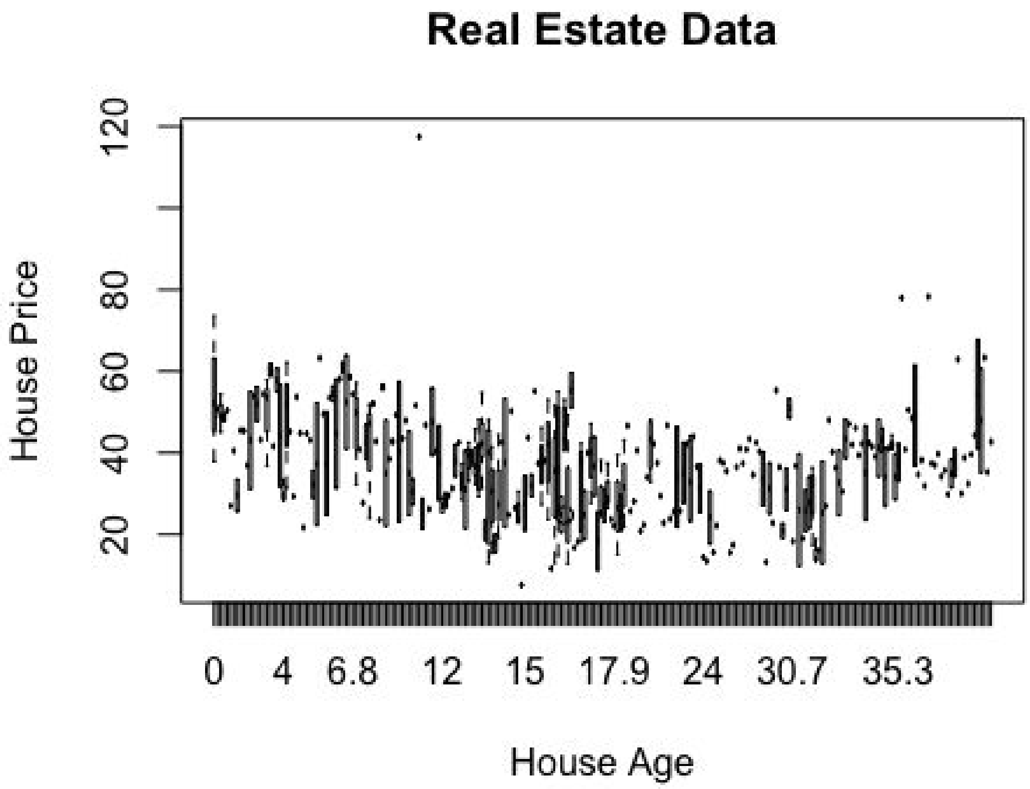
```r
avPlots(realestate.lm, id = FALSE)
```



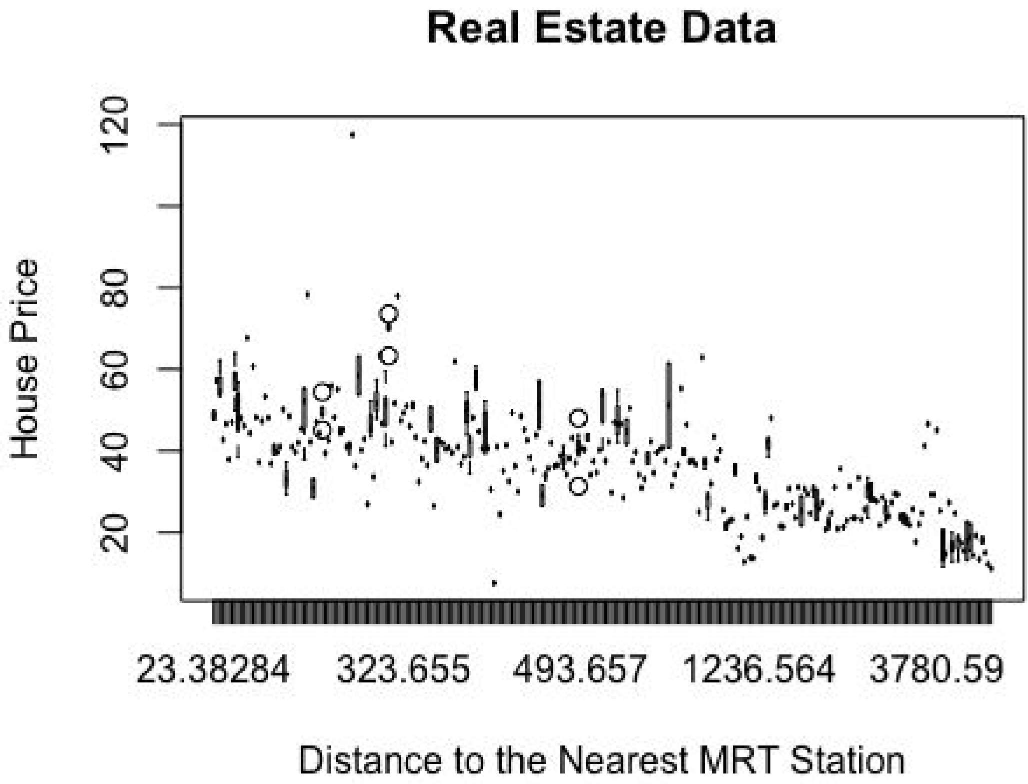
```
anova(realestate.lm)

## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## X1          1    586     586   7.4666  0.006559 **
## X2          1   3441    3441  43.8575 1.119e-10 ***
## X3          1  34857   34857 444.2919 < 2.2e-16 ***
## X4          1   3576    3576  45.5812 5.064e-11 ***
## X5          1   2065    2065  26.3192 4.488e-07 ***
## X6          1      5       5   0.0655  0.798203
## Residuals 407  31931      78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

boxplot(Y~X1, data = Real_estate_valuation_data_set, main = "Real
Estate Data", xlab = "Transaction Date", ylab = "House Price")
```



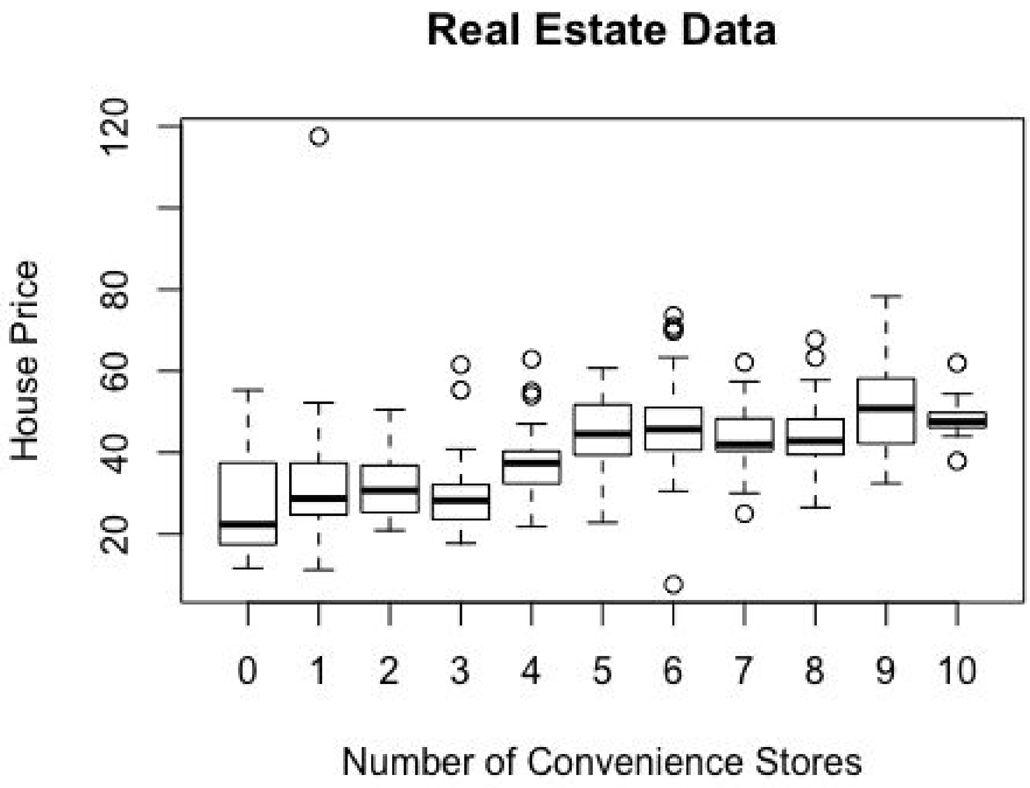
```
boxplot(Y~X2, data = Real_estate_valuation_data_set, main = "Real
Estate Data", xlab = "House Age", ylab = "House Price")
```



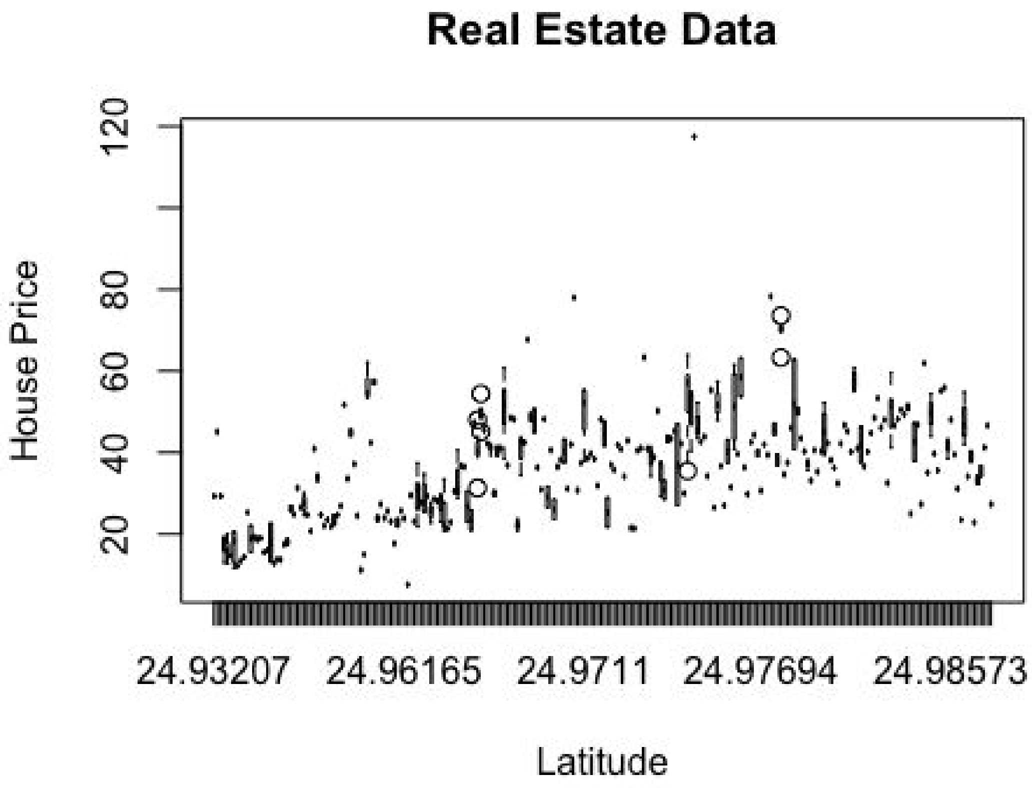
```
boxplot(Y~X3, data = Real_estate_valuation_data_set, main = "Real
Estate Data", xlab = "Distance to the Nearest MRT Station", ylab =
"House Price")
```



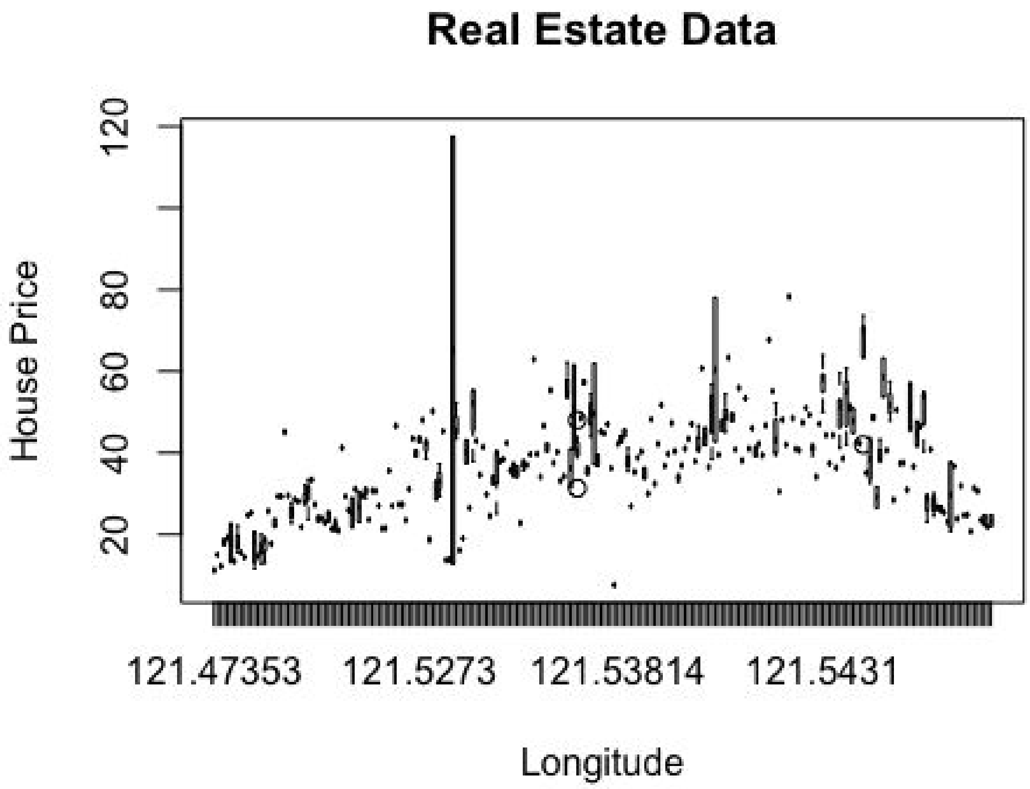
```
boxplot(Y~X4, data = Real_estate_valuation_data_set, main = "Real
Estate Data", xlab = "Number of Convenience Stores", ylab = "House
Price")
```



```
boxplot(Y~X5, data = Real_estate_valuation_data_set, main = "Real
Estate Data", xlab = "Latitude", ylab = "House Price")
```



```
boxplot(Y~X6, data = Real_estate_valuation_data_set, main = "Real
Estate Data", xlab = "Longitude", ylab = "House Price")
```



From the BoxPlots, it seems as though Latitude and Longitude are very random.

```
summary(X1)

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

summary(X2)

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

summary(X3)

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

summary(X4)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   4.000   4.094   6.000  10.000

Real_estate_valuation_data_set$X21 <-
with(Real_estate_valuation_data_set, (X2  + 1))
```

```
Real_estate_valuation_data_set$X41 <-
with(Real_estate_valuation_data_set, (X4 + 1))
Trans.Real_estate_valuation_data_set <- powerTransform(cbind(X1, X21,
X3, X41) ~ 1, Real_estate_valuation_data_set)
summary(Trans.Real_estate_valuation_data_set)
```

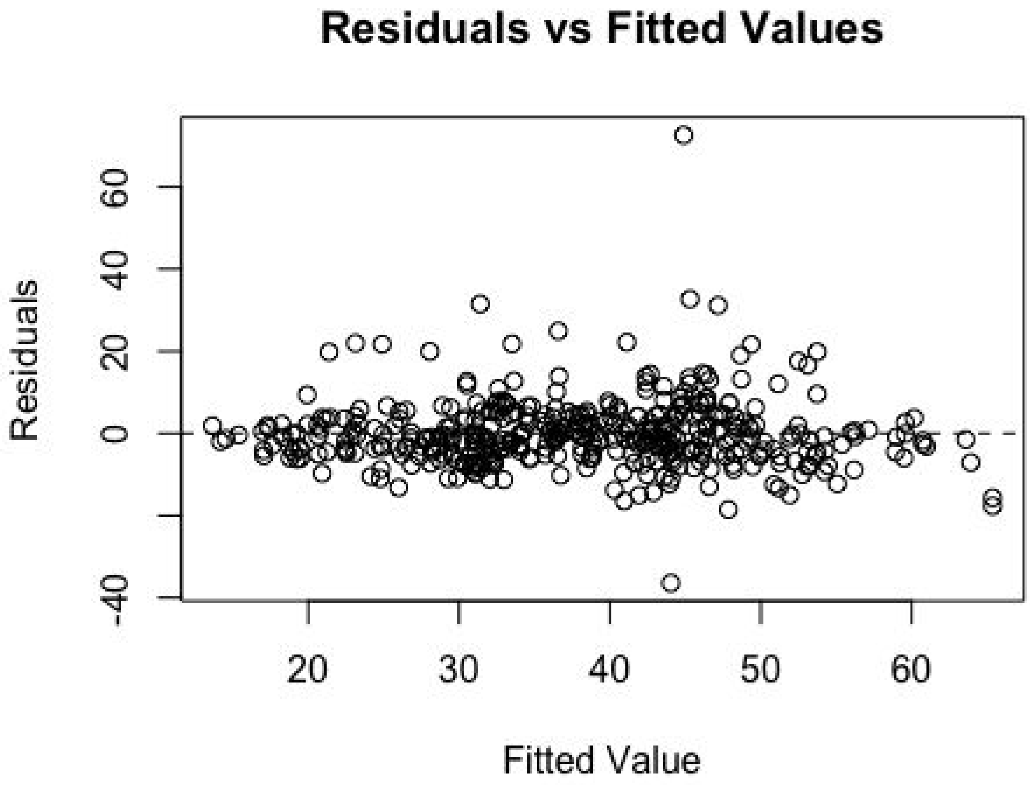
```
## bcPower Transformations to Multinormality
##     Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## X1     3.0000        1.00    -450.9223     456.9223
## X21    0.5470        0.50       0.4348       0.6593
## X3     0.0073        0.00      -0.0619       0.0766
## X41    0.6898        0.69       0.5551       0.8246
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                LRT df       pval
## LR test, lambda = (0 0 0 0) 214.643  4 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                 LRT df       pval
## LR test, lambda = (1 1 1 1) 687.2312  4 < 2.22e-16
```

```
realestate.lm <- lm(Y ~ X1 + sqrt(X2) + log(X3) + I(X4^(0.69)))
summary(realestate.lm)
```

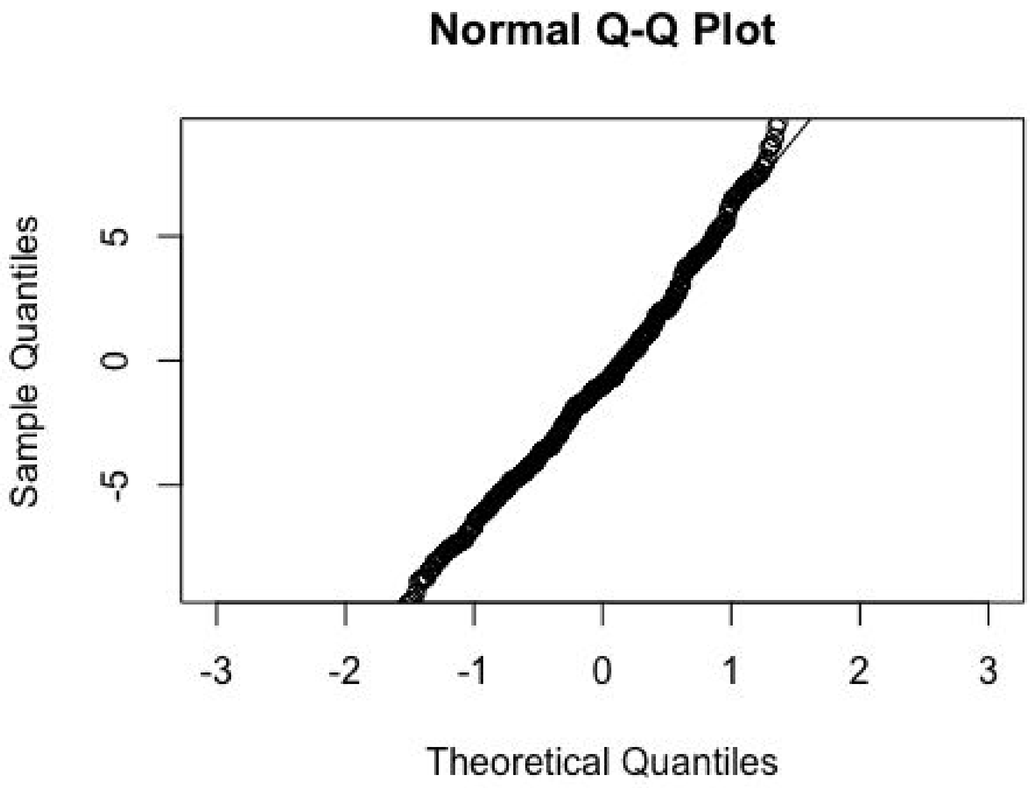
```
##
## Call:
## lm(formula = Y ~ X1 + sqrt(X2) + log(X3) + I(X4^(0.69)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.437  -4.739  -0.971   3.769  72.612
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.499e+04  3.020e+03  -4.964 1.01e-06 ***
## X1            7.491e+00  1.500e+00   4.993 8.82e-07 ***
## sqrt(X2)     -1.786e+00  2.735e-01  -6.531 1.94e-10 ***
## log(X3)      -7.474e+00  5.212e-01 -14.339  < 2e-16 ***
## I(X4^(0.69))  1.416e+00  3.889e-01   3.641 0.000306 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.51 on 409 degrees of freedom
## Multiple R-squared:  0.6126, Adjusted R-squared:  0.6089
## F-statistic: 161.7 on 4 and 409 DF,  p-value: < 2.2e-16
```

```
e <- resid(realestate.lm)
yhat2 <- fitted(realestate.lm)
```

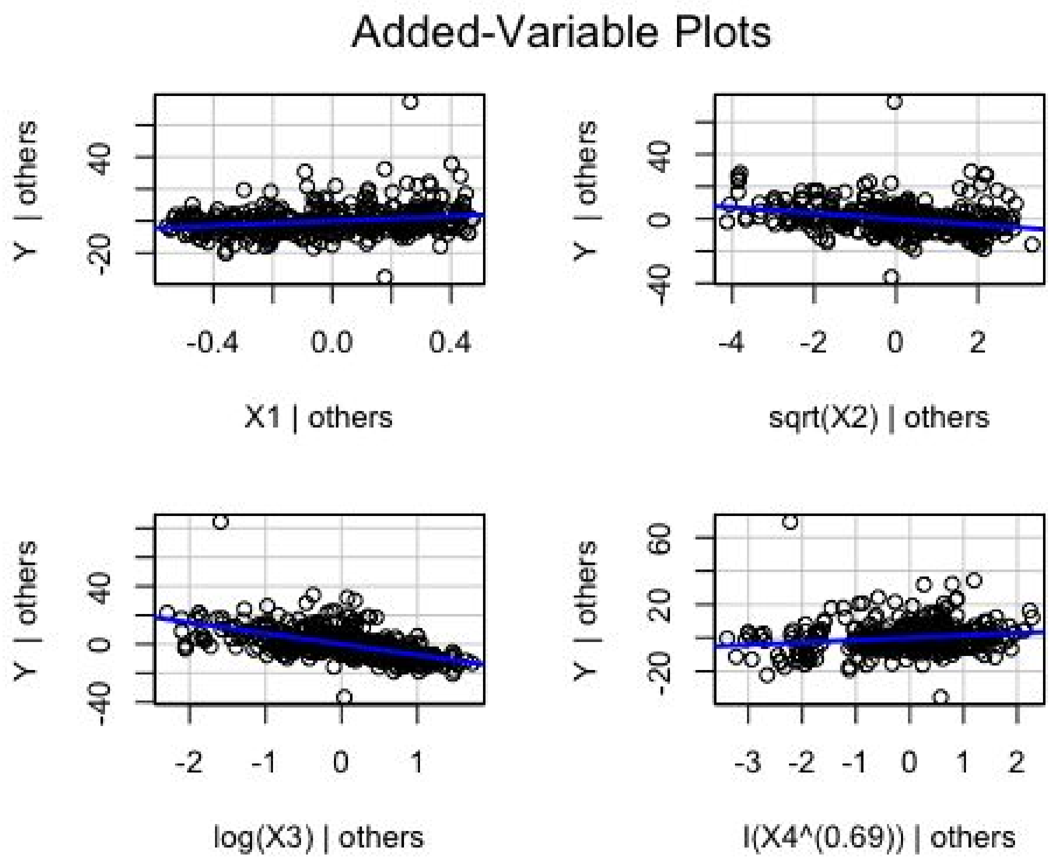
```
plot(yhat2, e, xlab = 'Fitted Value', ylab = 'Residuals', main =
'Residuals vs Fitted Values')
abline(h = 0, lty = 2)
```



```
qqnorm(e, ylim = c(-9, 9))
qqline(e)
```



```r
avPlots(realestate.lm, id = FALSE)
```



```
anova(realestate.lm)

## Analysis of Variance Table
##
## Response: Y
##                 Df Sum Sq Mean Sq  F value    Pr(>F)
## X1               1    586     586   8.0894  0.004676 **
## sqrt(X2)         1   6684    6684  92.2944 < 2.2e-16 ***
## log(X3)          1  38614   38614 533.2217 < 2.2e-16 ***
## I(X4^(0.69))     1    960     960  13.2599  0.000306 ***
## Residuals      409  29618      72
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

predict(realestate.lm , data.frame(X1 = mean(X1), X2 = mean(X2), X3 =
mean(X3), X4 = 0), interval = 'confidence', level = 0.95)

##        fit      lwr      upr
## 1 29.55788 27.82855 31.28721

summary(realestate.lm)
```

```
##
## Call:
## lm(formula = Y ~ X1 + sqrt(X2) + log(X3) + I(X4^(0.69)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.437  -4.739  -0.971   3.769  72.612
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.499e+04  3.020e+03  -4.964 1.01e-06 ***
## X1            7.491e+00  1.500e+00   4.993 8.82e-07 ***
## sqrt(X2)     -1.786e+00  2.735e-01  -6.531 1.94e-10 ***
## log(X3)      -7.474e+00  5.212e-01 -14.339  < 2e-16 ***
## I(X4^(0.69))  1.416e+00  3.889e-01   3.641 0.000306 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.51 on 409 degrees of freedom
## Multiple R-squared:  0.6126, Adjusted R-squared:  0.6089
## F-statistic: 161.7 on 4 and 409 DF,  p-value: < 2.2e-16
```