STAT 425 Final Project

Jeff Massman

11/21/2021

Contributions

Jeff Massman (NetID = massman4) - Entire project (I decided to work alone on this project since I already had three other group projects to do simultaneously).

1. Introduction

The goal of this report is to build and analyze several different statistical models to predict the valuation of real estate property in Taiwan based on seven different predictor variables. The dataset on which this report is based originates from a publication titled "Building Real Estate Valuation Models with Comparative Approach Through Case-Based Reasoning" by authors I-Cheng Yeh and Tzu-Kuang Hsu. In their publication, they analyze contemporary methods of building real estate valuation models, while proposing their own new innovative approach which they call the "Quantitative Comparative Approach." The actual composition of the data will be explored in the next section.

This report will be organized into three sections: the first is a data exploratory section. As mentioned earlier, this is where the components of the data will be broken down and explained, as well as a preliminary analysis including graphics, summary statistics, etc.

The second section will consist of the methodology, where the actual models and procedures for selecting those models will be outlined.

Finally, there will be a brief conclusion, where the results of the analysis and other findings will be summarized.

2. Exploratory Data Analysis

The original dataset is composed of six predictors and one response, though a seventh predictor, labelled X_7, will be appended to the data.

2.1 Components

The data components are the following:

- X_1 -- the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)
- X_2 -- the house age (unit: year)
- X_3 -- the distance to the nearest MRT station (unit: meters)
- X_4 -- the number of convenience stores in the living circle on foot (integer)
- X_5 -- the geographic coordinate, latitude. (unit: degree)
- X_6 -- the geographic coordinate, longitude. (unit: degree)
- X_7 -- the transaction month
- Y -- response variable; house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 squared meters)

2.2 Summary Statistics

```
##             X1             X2                X3                 X4        
##  2013.4166667: 58   Min.   : 0.000   Min.   :  23.38   Min.   : 0.000  
##  2013.5      : 47   1st Qu.: 9.025   1st Qu.: 289.32   1st Qu.: 1.000  
##  2013.0833333: 46   Median :16.100   Median : 492.23   Median : 4.000  
##  2012.9166667: 38   Mean   :17.713   Mean   :1083.89   Mean   : 4.094  
##  2013.25     : 32   3rd Qu.:28.150   3rd Qu.:1454.28   3rd Qu.: 6.000  
##  2012.8333333: 31   Max.   :43.800   Max.   :6488.02   Max.   :10.000  
##  (Other)     :162                                                      
##        X5               X6                X7            Y         
##  Min.   :24.93   Min.   :121.5   May     : 58   Min.   :  7.60  
##  1st Qu.:24.96   1st Qu.:121.5   June    : 47   1st Qu.: 27.70  
##  Median :24.97   Median :121.5   January : 46   Median : 38.45  
##  Mean   :24.97   Mean   :121.5   November: 38   Mean   : 37.98  
##  3rd Qu.:24.98   3rd Qu.:121.5   March   : 32   3rd Qu.: 46.60  
##  Max.   :25.01   Max.   :121.6   October : 31   Max.   :117.50  
##                                  (Other) :162                   
```

A more descriptive display of X_1 and X_7, respectively:

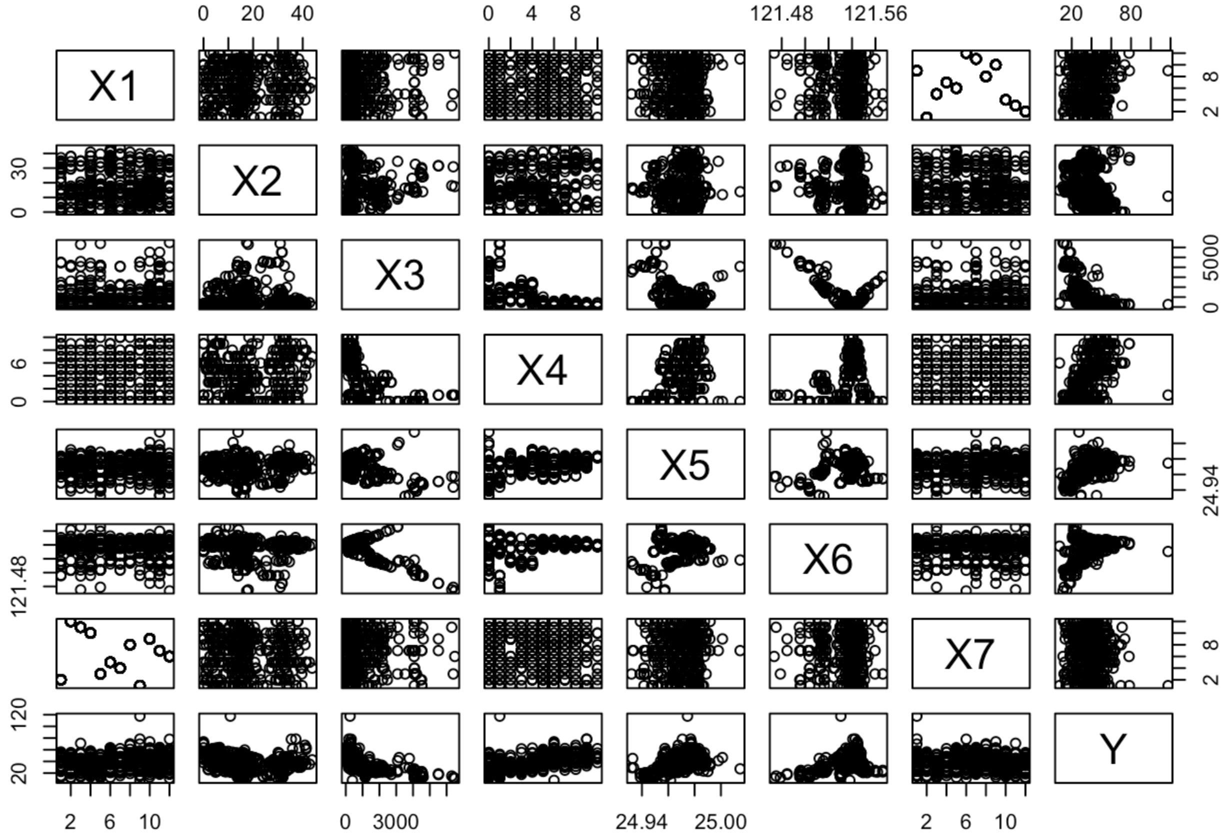
```
## 
## 2012.6666667      2012.75 2012.8333333 2012.9166667         2013 2013.0833333 
##           30           27           31           38           28           46 
## 2013.1666667      2013.25 2013.3333333 2013.4166667       2013.5 2013.5833333 
##           25           32           29           58           47           23 
```

```
## 
##     April    August  December  February   January      July      June     March 
##        29        30        28        25        46        23        47        32 
##       May  November   October September 
##        58        38        31        27 
```

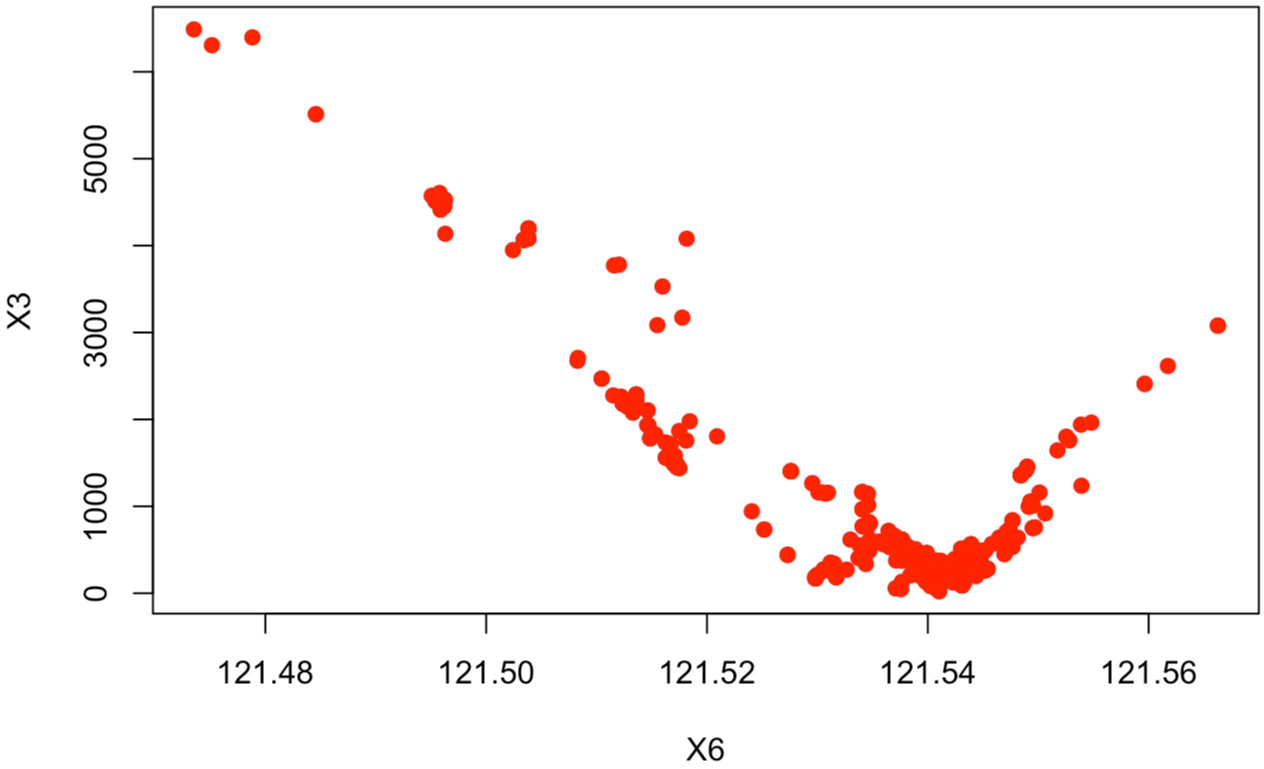
There is, roughly, a uniform distribution of months and dates. All the properties are within 0.08° longitude and 0.1° latitude of each other.

All variables are quantitative except for X_1 and X_7, which are categorical. X_4 is discrete, since it can only take on nonnegative integral values. There are no missing observations in the data, so imputation is not necessary.

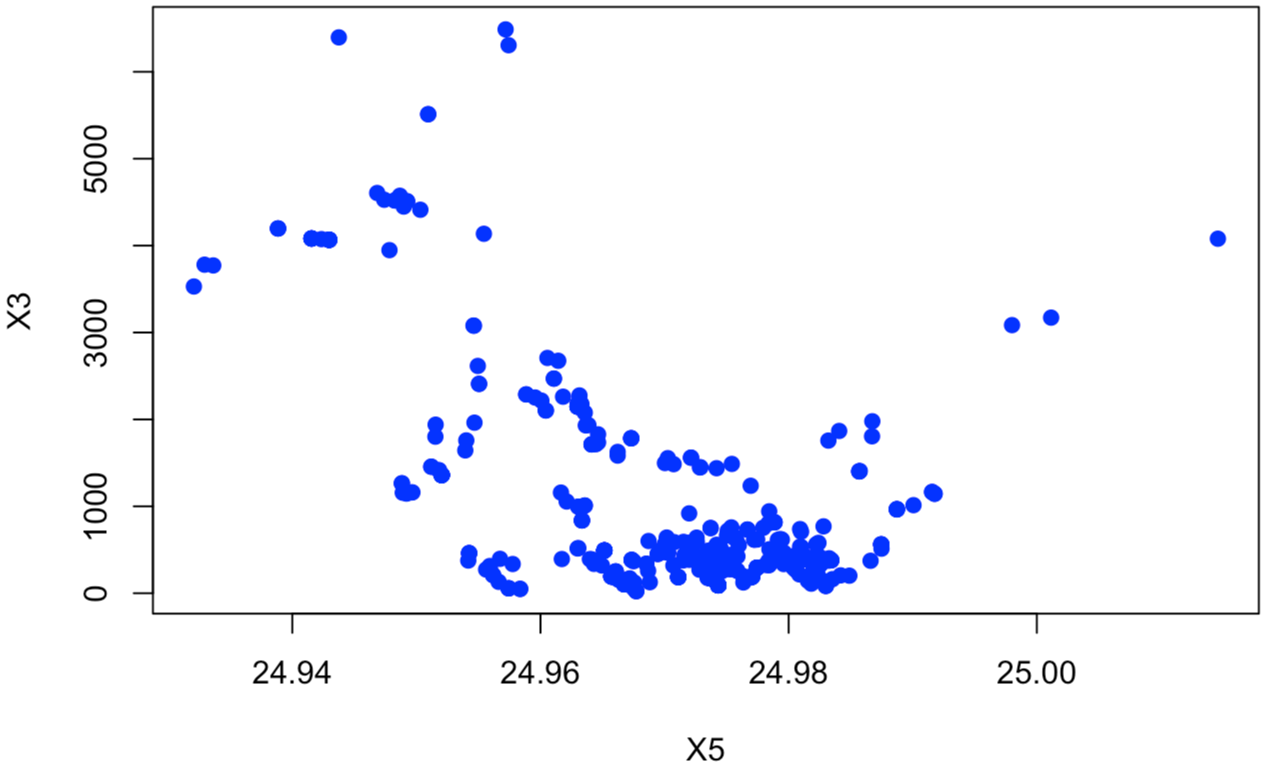
2.3 Variable Pair Plots



Most of these plots are not very noteworthy. However, there appears to be a peculiar relationship between X_3 and X_6, almost quadratic-like.



There is a similar albeit weaker association between X_3 and X_5:



The existence of these relationships makes sense: X_3 is the distance to the nearest MRT station. It makes sense that this distance would be affected by the latitude and longitude coordinates (X_5 and X_6, respectively) of the property.

There is also a relationship between X_4 and X_5, X_6 (i.e. between the number of department stores nearby and the latitude / longitude coordinates).

3. Methodology

We will first examine a naive full linear regression model, with all variables. Since we have a small number of predictors, we can use what is called the "Leaps and Bounds" method to select the best possible model. From there, we will perform a series of model diagnostics and adapt the model accordingly. Then, we will evaluate the predictive power of the model by calculating the testing data prediction error. Afterwards, we will consider a ridge regression model and a random forest model, and we will compare the results of these three.

3.1 Linear Regression (OLS) Model

In this section, we will analyze the obvious linear model. There is an inherent collinearity issue present in the data. This is an artifact resulting from our construction of X_7 from X_1, so we will remove X_1 from the model for the time being and proceed with our analysis.

Here are the results:

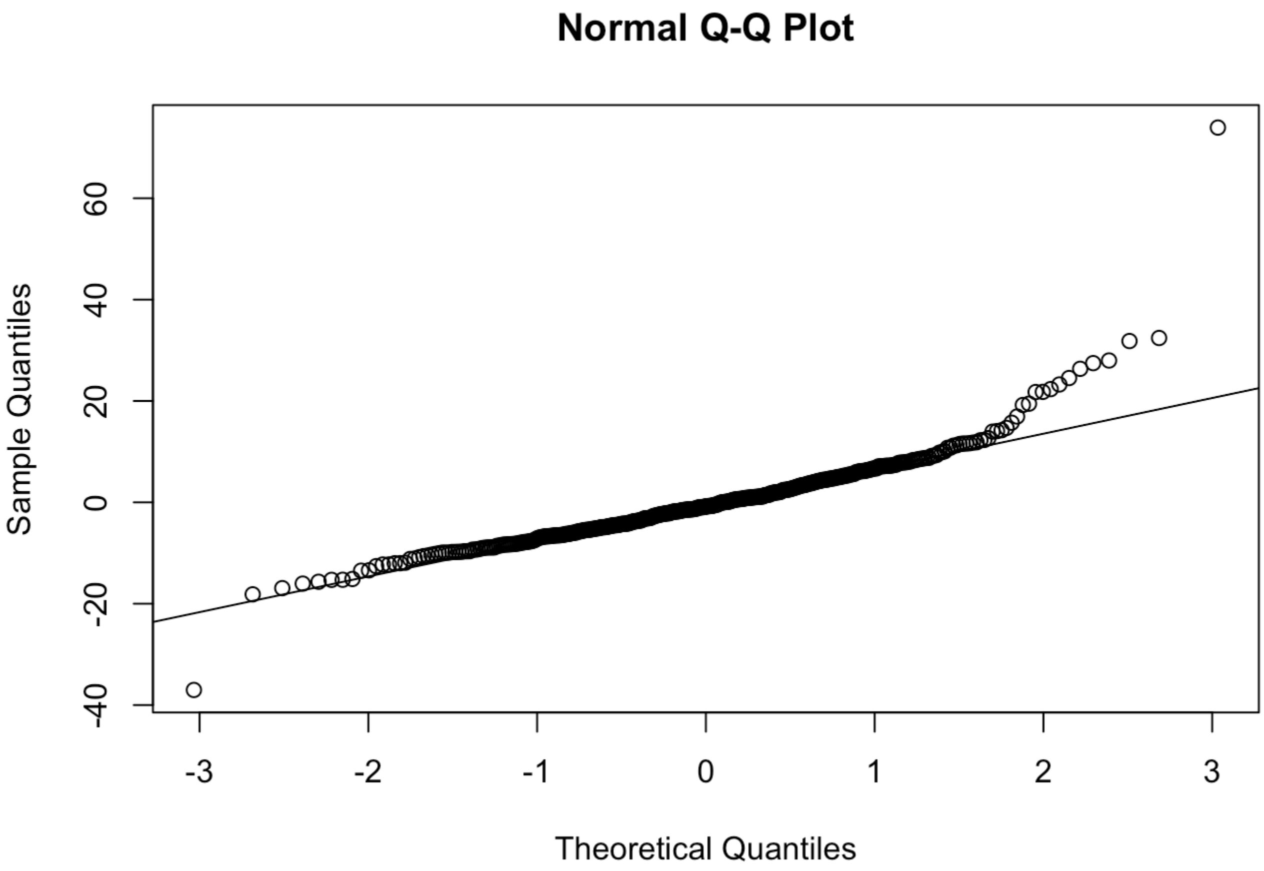
```
## (Intercept)          X2          X3          X4          X5          X6 
##        TRUE        TRUE        TRUE        TRUE        TRUE       FALSE 
##    X7August  X7December  X7February   X7January      X7July      X7June 
##       FALSE       FALSE       FALSE       FALSE        TRUE        TRUE 
##     X7March       X7May  X7November   X7October X7September 
##       FALSE       FALSE        TRUE       FALSE       FALSE 
```

According to these results, the best model is one in which only X_6 is removed. This gives us a linear model with an okay R^2 value of 0.5939969. Intuitively, this is the proportion of variation in the data explained by the model.

It should also be noted that only three months out of twelve were found to be significant. However, it is generally ill-advised to remove insignificant factor levels (in this case, months), so we retain the entirety of X_7.

We now proceeed with model diagnostics.

Normality Test:



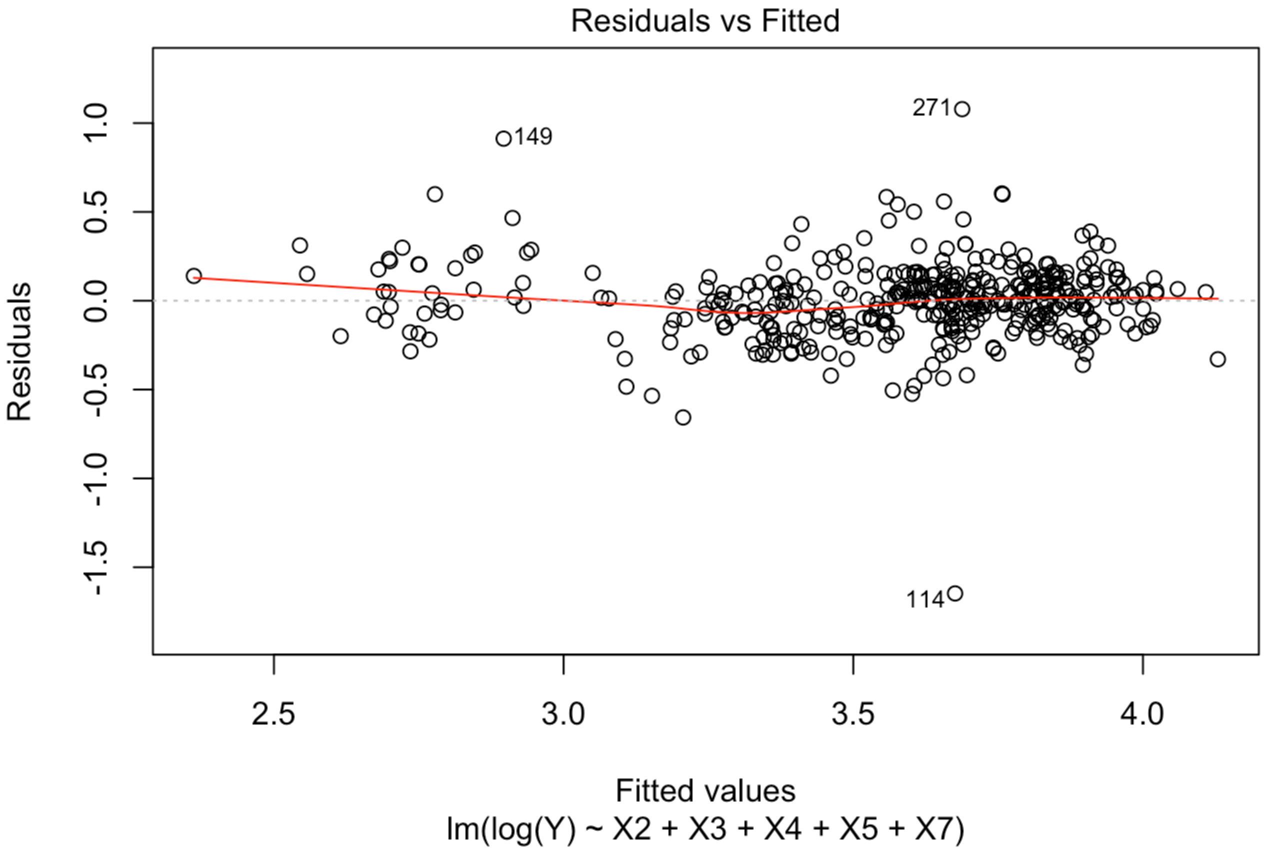
From these results, our normality assumption is mostly justified. There is some deviation from the theoretical quantiles towards the end, but this is not totally unusual.

Now we will test for homoscedasticity:

```
##
##  studentized Breusch-Pagan test
##
## data:  fit3
## BP = 25.246, df = 15, p-value = 0.04674
```

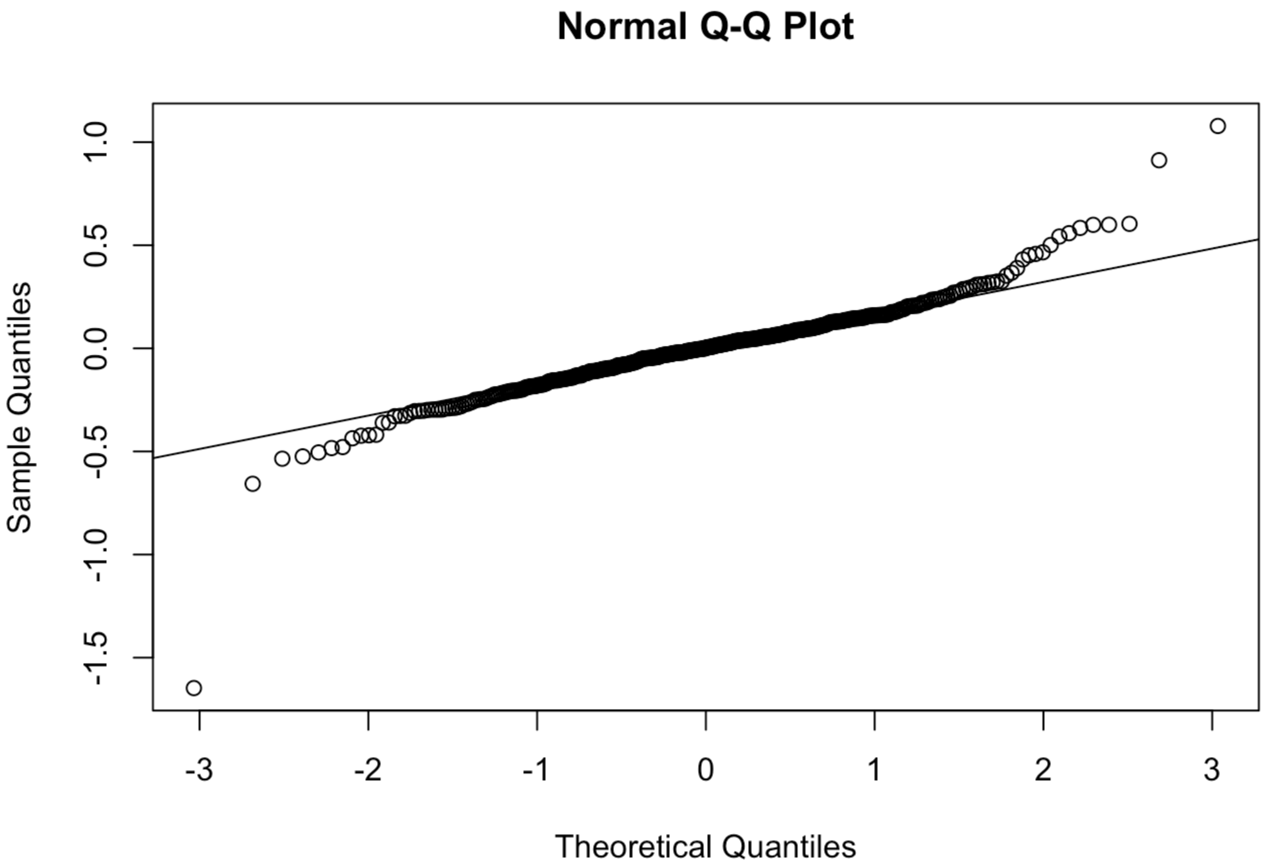
According to these results, our homoscedasticity assumption actually fails, So we will have to transform the response variable to see if we can remedy this. We will try a $\mathrm{Log}(\mathrm{Y})$ transformation:

```
##
##  studentized Breusch-Pagan test
##
## data:  fit4
## BP = 24.938, df = 15, p-value = 0.05078
```



This just pushes us over the 0.05 threshold and thus, though not decisively, we can technically say that we have homoscedasticity in this model. This is not a very statistically sound conclusion, so perhaps more convincingly, we turn to the residual-fitted plot: the red line is mostly straight; in practice, it will typically never be exactly straight due to the randomness of the data, but this is good enough.

Retesting for normality in this new model:



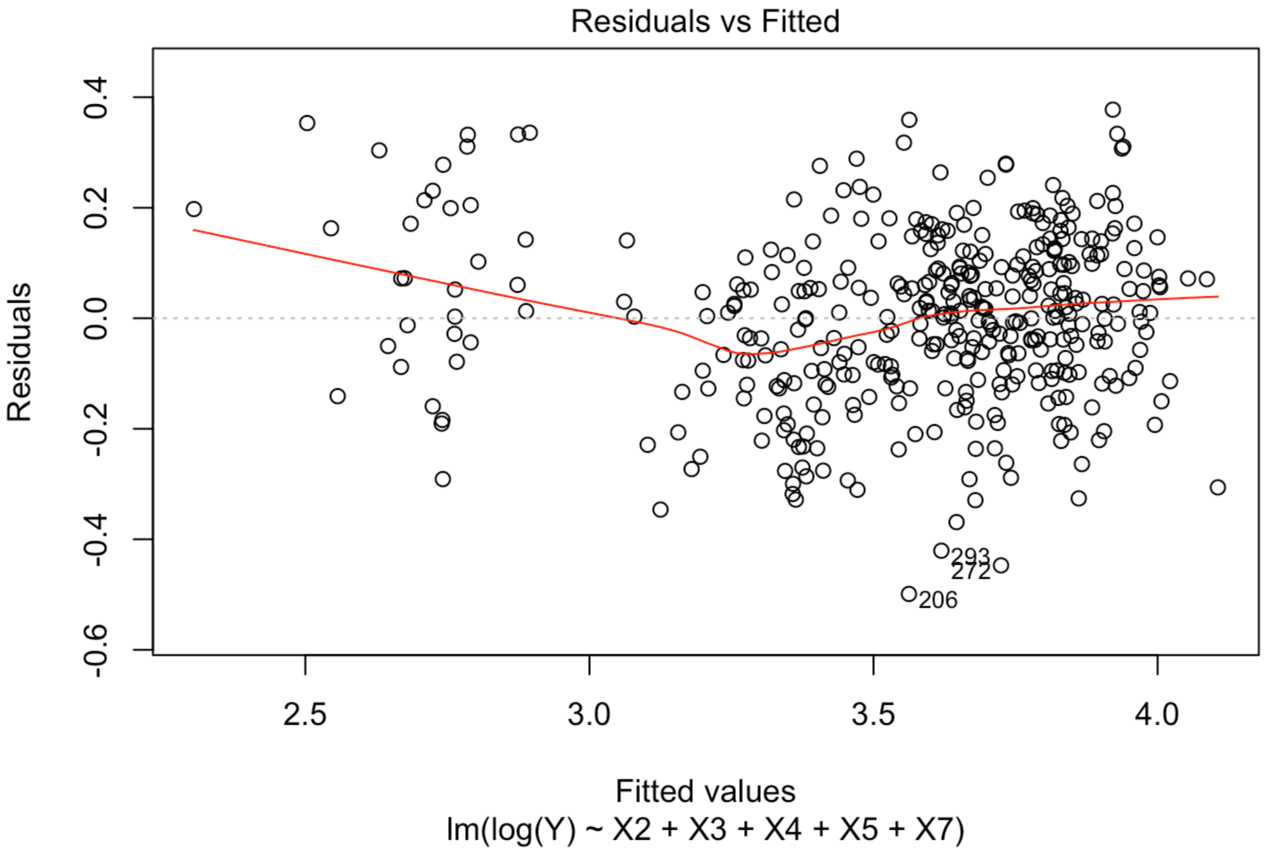
The quantiles follow the QQ line closely, with some deviation towards the end. We can say that the data is approximately normally distributed.

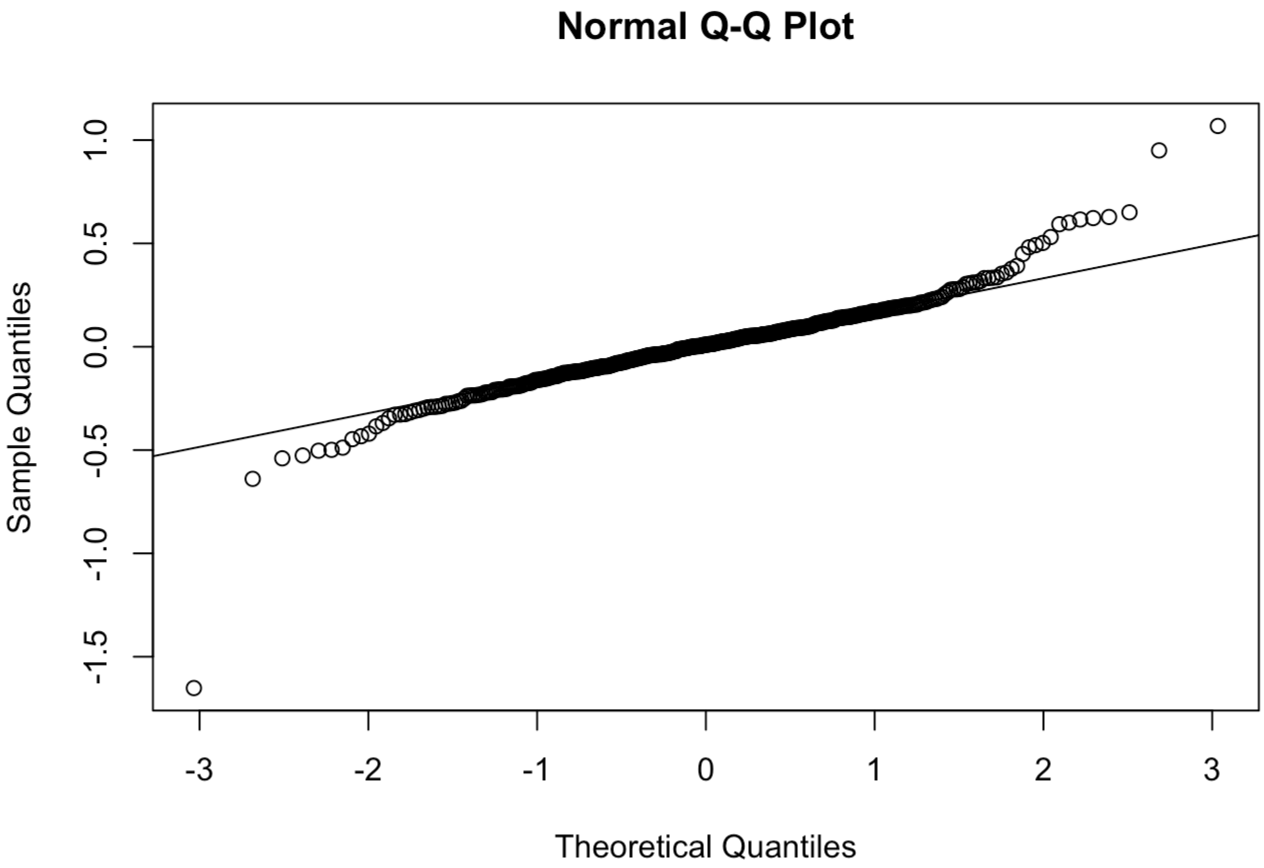
We will now handle highly influential and outlier points. We will use Cook's Distance to see if any highly influential points need to be removed from the model, with the standard cutoff of $\frac{4}{n}$, where n is our sample size. We will also remove points with absolute studentized residuals greater than 3.

We find that 23 points in our data fit the above criteria and must be removed.

After removing, we test our assumptions one more time:

```
##
##  studentized Breusch-Pagan test
##
## data:  fit5
## BP = 24.938, df = 15, p-value = 0.05078
```


Residuals vs Fitted
Residuals
0.4
0.2
0.0
-0.2
-0.4
-0.6
2.5
3.0
3.5
4.0
293
272
206
Fitted values
lm(log(Y) ~ X2 + X3 + X4 + X5 + X7)



We have a heavy-tailed but still Gaussian distribution, as indicated by the QQ plot.

Note that the residual vs fitted plot looks worse than before, but this partially due to the huge scale reduction on the y-axis; the red line is not the be-all-end-all; it will never be perfect. For our purposes, the residuals do not contain any significant pattern, therefore we may proceed.

Our final model is

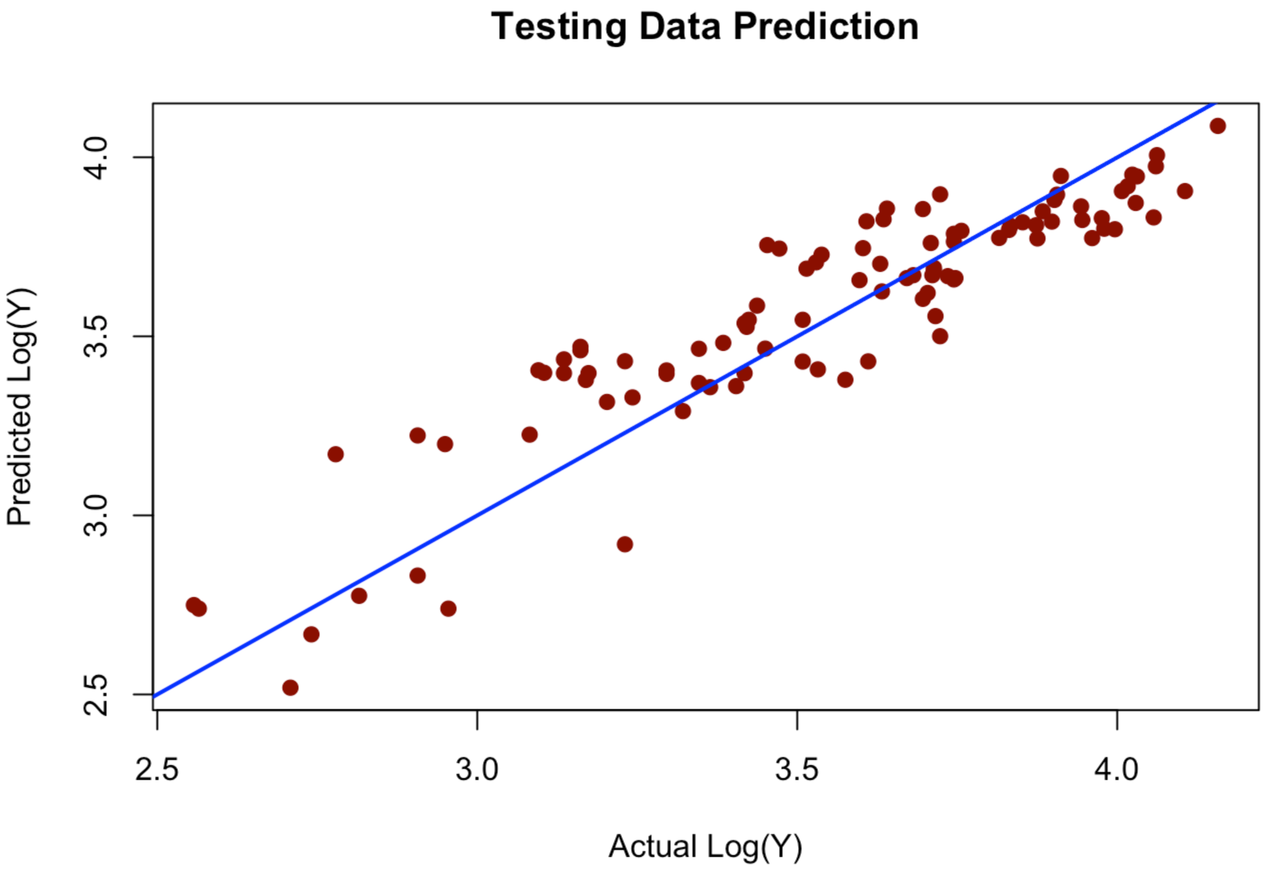
$$\mathrm{LogY} \sim \mathrm{X}_2 + \mathrm{X}_3 + \mathrm{X}_4 + \mathrm{X}_5 + \mathrm{X}_7$$

with an R^2 value of 0.8249352, which is quite good and is a marked improvement from the previous 0.594 value from before.

3.2: Prediction

We will now evaluate the testing prediction error of our model.

The following is an Actual vs. Predicted plot for the response in the testing data set:

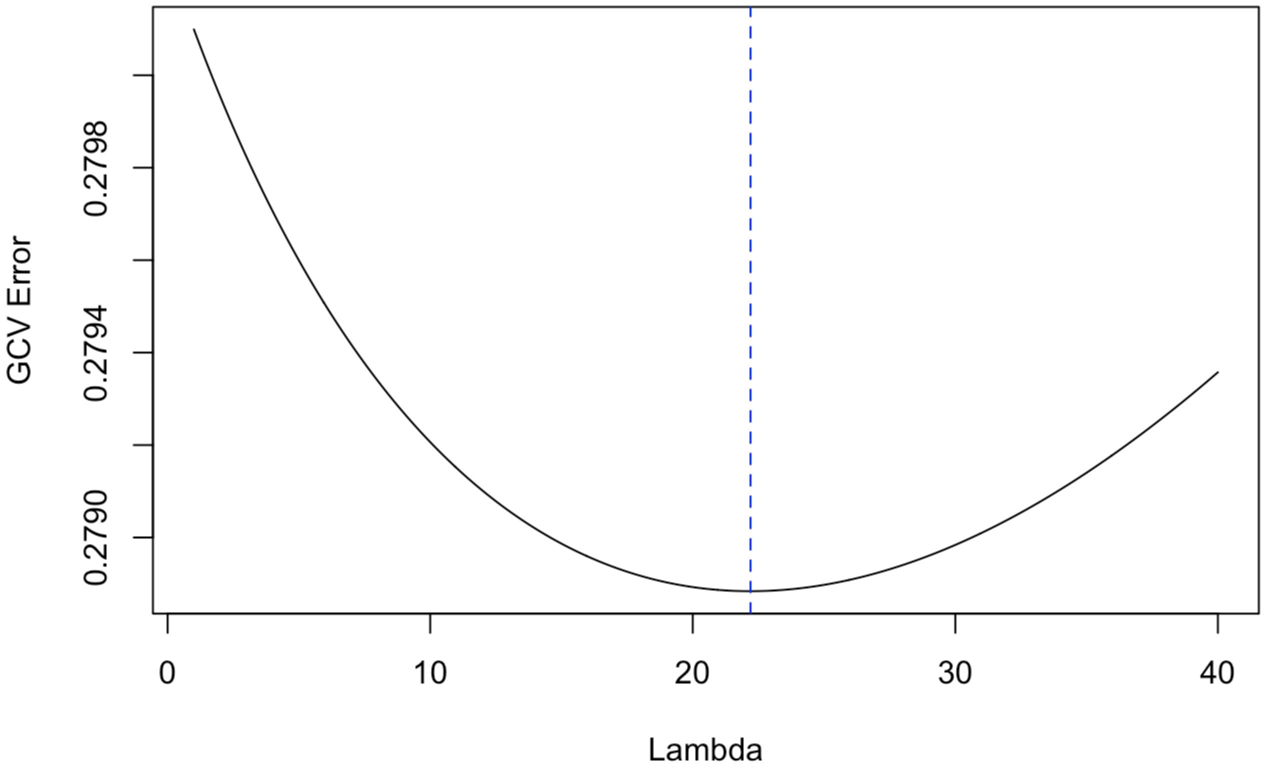


This plot shows a reasonable scattering pattern, which is what we would expect.

After randomly splitting the data into a training (75%) and testing (25%) set, the mean squared prediction error (calculated by exponentiating the prediction) of the model is 27.5665631. We will use this for comparison later.

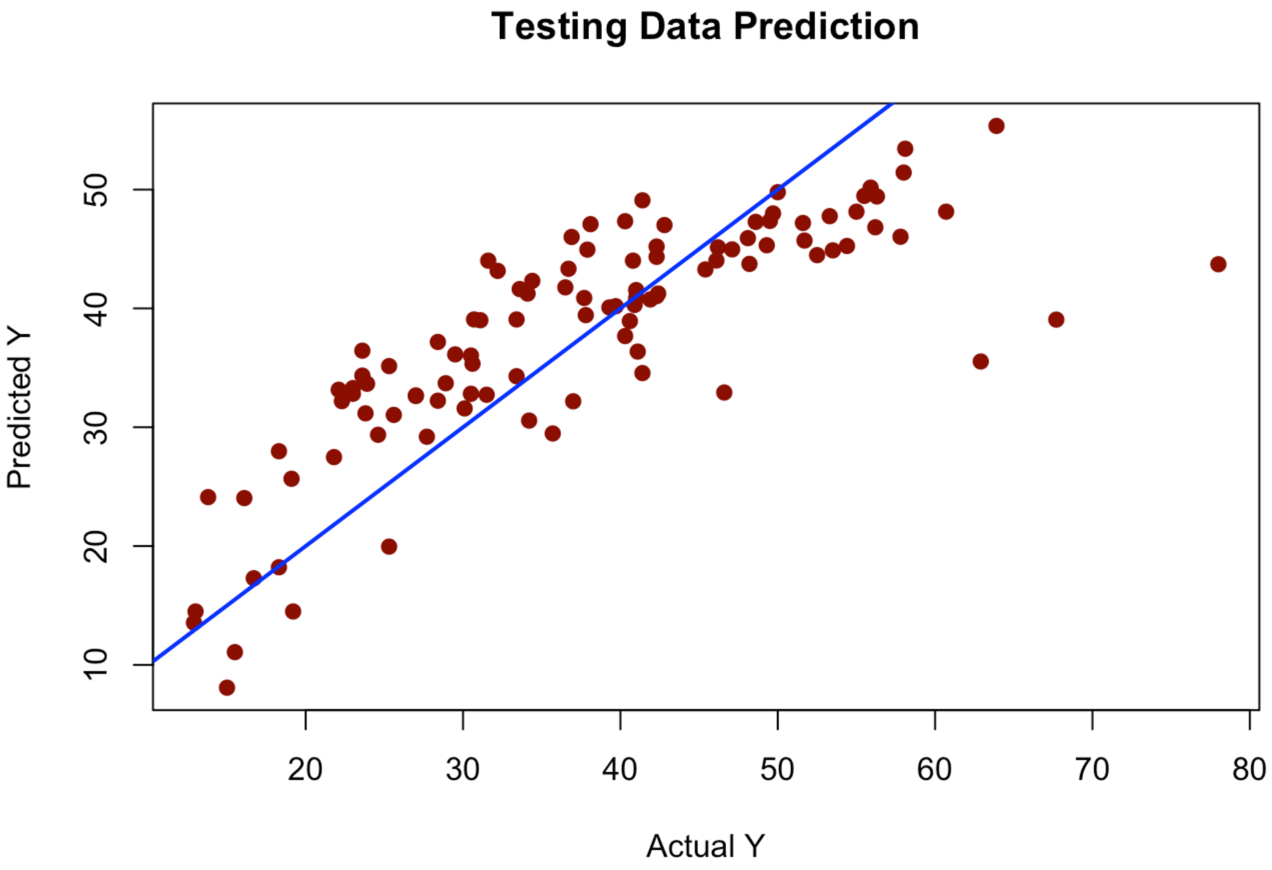
3.3 Penalized (Ridge) Regression Model

We will now construct another regression model, this time using ridge regression to add a penalty. We will construct a grid of λ values, perform GCV, and select the λ that minimizes the GCV error. We will train this model using the same training data as in section 3.1. The response will not be transformed, but we will still keep X_6 out of the model.



This is a plot of λ vs. the error. According to the plot, our best λ value is 22.2.

Actual vs. fitted plots for both the training and testing data:



The scattering pattern observed here is more erratic than that from the OLS model; there are more extreme values towards the right end.

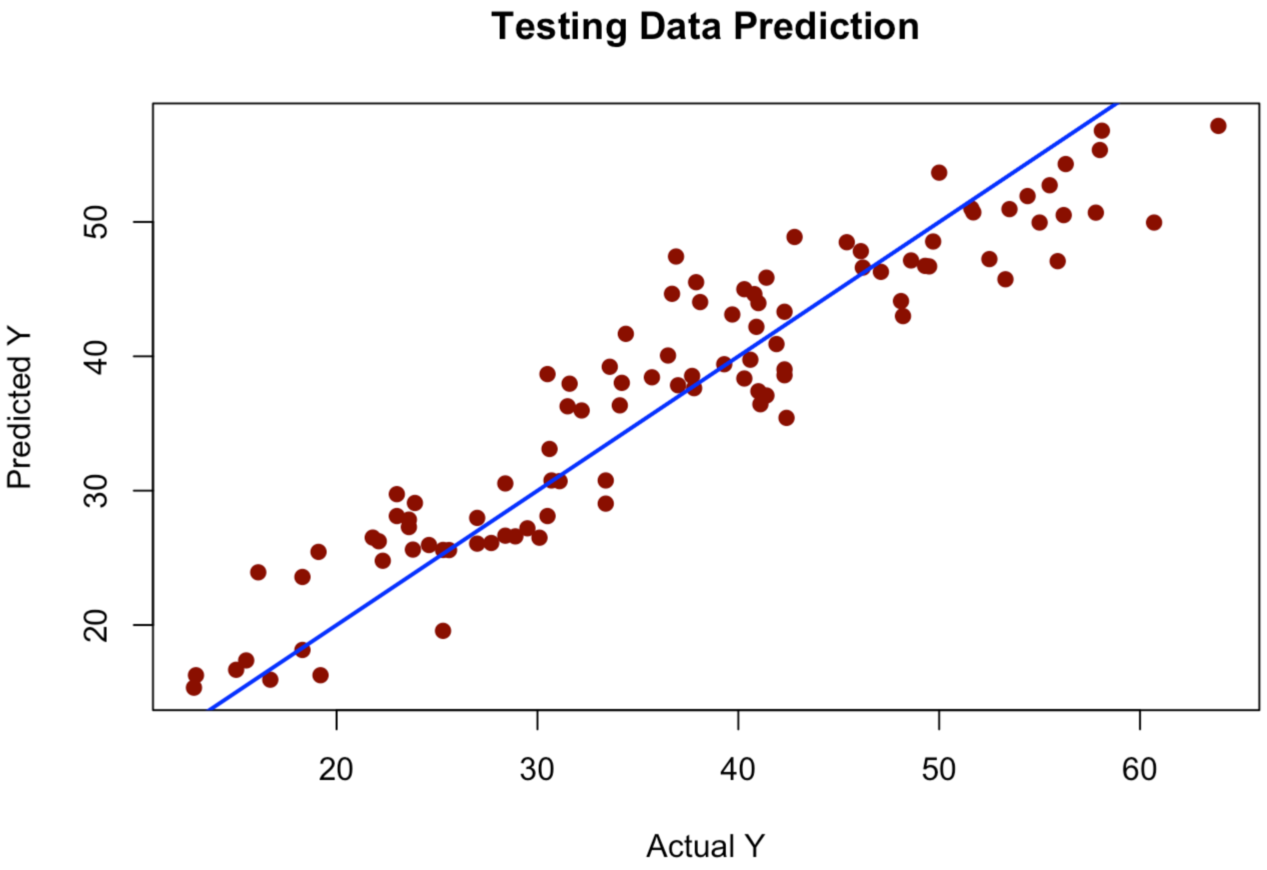
After predicting on the same testing set in section 3.1, the MSE is 66.9575255. This is worse than before, indicating that the ridge model may not be as good a fit as the OLS for the data.

However, though perhaps not totally valid, if we refit the ridge with the same extreme points removed as in the OLS model, we obtain an MSE of 37.2117705, which is significantly better, but still slightly worse than the OLS model.

3.4. Random Forest Model

We will now fit a random forest model. We will consider all the predictors (except X_1 again due to collinearity), and no transformation of the response. Also, no data points will be removed.

Here is the actual vs. fitted plot:



The plot exhibits a very good scattering pattern, much better than the prior two models.

The MSE for the random forest model is 18.0712615. This is also a huge improvement on the previous two models.

4. Conclusion

Three regression models were fit in this report: a standard OLS model, a ridge regression model, and a random forest model. Some predictors were removed from the analysis as well as extreme data points, and the response was log-transformed for the OLS model. What we found:

Model	**Prediction MSE**
OLS	27.567
Ridge	37.212
Random Forest	18.071

- The random forest model performed the best out of the three models with respect to the testing MSE
- Parameter tuning was necessary for the ridge model, optional for the random forest model, and not required for the OLS model
- Between OLS and ridge, OLS seemed to performs better. However, when removing the same outlier points as in the OLS model, we get a major performance boost in the ridge model
- In the OLS model, the log-transformed response may make interpretation more difficult

Therefore, in building a model to predict real estate valuation in Taiwan, our final recommendations are to use the random forest. It performed much better than the other two models, requires no transformations, and is relatively intuitive. Furthermore, though not explored in this report, the random forest model can be further tuned to possibly improve the performance even more, and reduce its complexity.