Real Estate Price Prediction

STAT 425 Final Project Report

Date: 10 December 2021

Team Members:

Mayank Agarwal
Yash Bajaj
Yash Kalyani

Work Distribution

Team Member	Contribution in Code	Contribution in Report
Mayank Agarwal	Generating numerical summary, creating pairwise plot, MLR model fitting	Objective, Data description, numerical summary, pairwise plot analysis, Discussion, Conclusion
Yash Bajaj	Generating histograms and boxplots, MLR diagnostics, Smoothing splines	Histogram analysis, MLR diagnostic, Price vs Month box plot analysis, Smoothing Splines, Conclusion
Yash Kalyani	Data mutation, correlation matrix, generating price vs variable plot, MLR Diagnostics, Random Forest	Correlation matrix, Scatterplot analysis, MLR Diagnostic, Random Forest, LATEX Report writing

Contents

Chapter 1

Introduction

1.1 Objective

In this project, we are analyzing the real estate valuation data.

Using this dataset, we are trying to predict the house price based on a number of factors. We will compare different Regression models and test their performance. In the first part of this project,we perform Exploratory Data Analysis to understand the predictors and how they affect the house price. We look at their numerical and graphical summaries.

Next we will use different Regression models to create two prediction models. The models will be based on Multiple Linear Regression, Smoothing Splines and Random Forests. To train our models, we split the dataset into a training and testing set. The first prediction model we use is Multiple Linear Regression, while the second model is a Smoothing splines model. We compare these models on a testing set and compare the prediction errors. Another non-parametric prediction model, Random Forest, is also used to predict, and the results are compared with the previous models.

While implementing the prediction models we perform several methods for variable selection and model diagnostics.

1.2 Data Description

The dataset can be found on the UCI machine learning repository[1] and was posted on the website on 18 August, 2018. It consists of market historical data of real estate valuation collected from Scindian District, New Taipei City, Taiwan.

It was previously used in a 2018 publication, Yeh, I.C., and Hsu, T.K.(2018). "*Building real estate valuation models with comparative approach through case-based reasoning.*"

The original dataset consists of 414 rows and 7 columns. We also add another column to represent the month. Hence, the columns in the dataset are:

- **Date** (X1, numeric) = the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)

The House price not only depends on what kind of house it is but also on the market at the time of transaction.

- **Age** (X2, numeric) = the house age (unit: year)
  The House Age at the transaction date affects the depreciation of the house and also how the quality of living of the house will be.

- **Distance** (X3, numeric) = the distance to the nearest MRT station (unit: meter)
  The distance to each MRT station was calculated by the coordinate of the house and the MRT station and then applying the minimization operation.

- **Convenience** (X4, numeric) = the number of convenience stores in the living circle on foot (integer)
  The number of convenience stores in the vicinity of a house is an essential factor of the house price. This represents the number of convenience stores within a distance of 500m of the house.

- **Latitude and Longitude** (X5 and X6, numeric) = the geographic coordinates, latitude and longitude. (unit: degree)
  This determines the distance to the city's downtown, which is an important factor in determining the time and money spent for tasks like shopping and going to work.

- **Month** (X7, categorical) = the name of the month when the transaction took place.

- **Price** (Y, numeric) = house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

Chapter 2

Exploratory Data Analysis

The dataset has 7 numeric variables and 1 categorical variable. The numeric variables are Date, Age, Distance, Convenience, Latitude, Longitude, and Price. The categorical variable is Month.

The first step for our data analysis was to check the dataset for any missing data. We found no missing observation in the dataset.

2.1 Numerical Summary

```
      date            age             distance       convenience       latitude       longitude         price             month      
 Min.   :2013   Min.   : 0.000   Min.   :  23.38   Min.   : 0.000   Min.   :24.93   Min.   :121.5   Min.   :  7.60   May     : 58  
 1st Qu.:2013   1st Qu.: 9.025   1st Qu.: 289.32   1st Qu.: 1.000   1st Qu.:24.96   1st Qu.:121.5   1st Qu.: 27.70   June    : 47  
 Median :2013   Median :16.100   Median : 492.23   Median : 4.000   Median :24.97   Median :121.5   Median : 38.45   January : 46  
 Mean   :2013   Mean   :17.713   Mean   :1083.89   Mean   : 4.094   Mean   :24.97   Mean   :121.5   Mean   : 37.98   November: 38  
 3rd Qu.:2013   3rd Qu.:28.150   3rd Qu.:1454.28   3rd Qu.: 6.000   3rd Qu.:24.98   3rd Qu.:121.5   3rd Qu.: 46.60   March   : 32  
 Max.   :2014   Max.   :43.800   Max.   :6488.02   Max.   :10.000   Max.   :25.01   Max.   :121.6   Max.   :117.50   October : 31  
                                                                                                                     (Other) :162  
```

Figure 2.1: Numerical Summary

In Figure 2.1, a numerical summary of the data is shown. The Average Price per unit of a house is 37.98 units, with prices in the range between 7.6 and 117.5 units (10000 Taiwan dollars per Ping). The average age of a house is 17.7 years old from the transaction date. The oldest house present in the dataset is 43.8 years old. There are on average 4 convenience stores near a house, and a maximum of 10 convenience stores around some houses. The average transaction date was in 2013, but the last transaction took place in 2014. The average distance of a house from an MRT station is 1083 metres away. The minimum distance is 23.38 metres and the maximum distance is 6488.02 metres. The average longitude is 121.5 and the average latitude is 24.97 degrees. We also note that just the mean and quartile values of the columns are not enough to get a good description of the data, so we create histograms, correlation matrix, and scatter plots of these variables.

The values and sampling variance of regression coefficients can be highly dependent on the particular predictors chosen for the model. So we check the correlation between all pairs of variables. From the correlation matrix in Figure 2.2, we observe that there is a high correlation between longitude - distance and price - distance. This means that there

	date	convenience	price	age	distance	latitude	longitude
date	1.000000000	0.009544199	0.08752927	0.01754234	0.06088009	0.03501631	-0.04106508
convenience	0.009544199	1.000000000	0.57100491	0.04959251	-0.60251914	0.44414331	0.44909901
price	0.087529272	0.571004911	1.00000000	-0.21056705	-0.67361286	0.54630665	0.52328651
age	0.017542341	0.049592513	-0.21056705	1.00000000	0.02562205	0.05441990	-0.04852005
distance	0.060880095	-0.602519145	-0.67361286	0.02562205	1.00000000	-0.59106657	-0.80631677
latitude	0.035016305	0.444143306	0.54630665	0.05441990	-0.59106657	1.00000000	0.41292394
longitude	-0.041065078	0.449099007	0.52328651	-0.04852005	-0.80631677	0.41292394	1.00000000

Figure 2.2: Correlation Matrix

is a high chance that distance might be a good predictor for price. It also shows that we may need to remove longitude from our model to avoid collinearity.

2.2 Graphical Summary

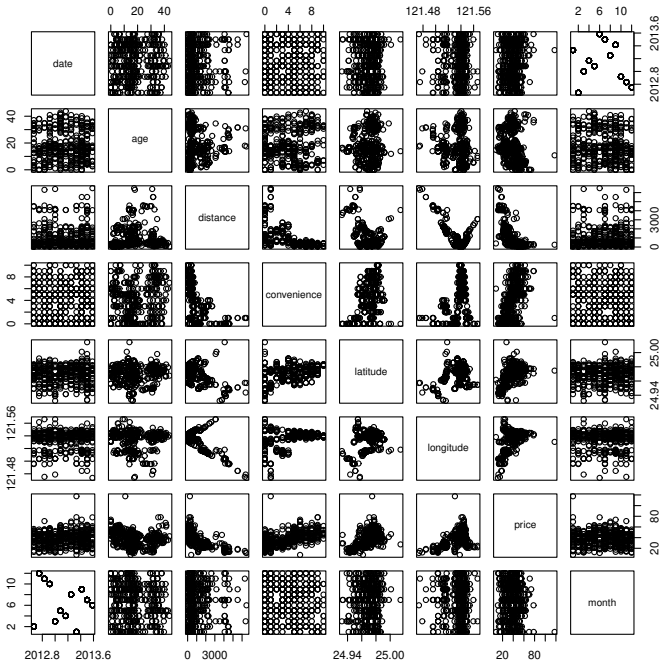
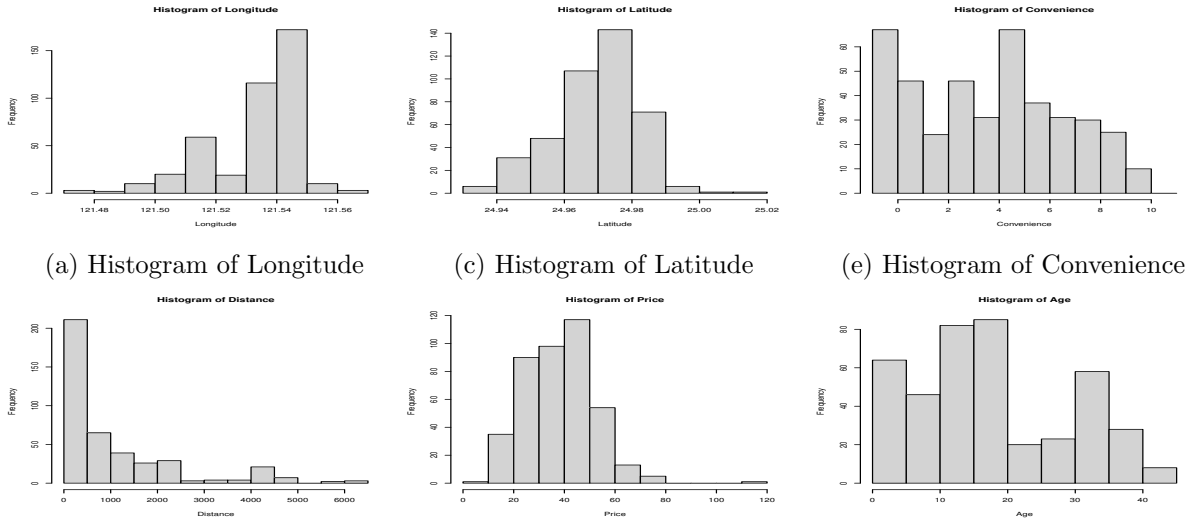


Figure 2.3: Pairwise Scatter Plot Of Variables

Next, we wanted to check the behaviour of the variables with each other. So, a pairwise scatter plot was made (shown in Figure 2.3).

The pairwise plots of the predictors and the response variable are plotted. The plot between convenience and price shows that as the number of convenience stores increases, so does the unit price of the houses increase. The plot between latitude, longitude and convenience stores suggest that most of the convenience stores are concentrated in a particular geographic location.

Next we plot the histograms of the numerical variables. We don't plot the histogram for the variable "Date" since it is the variable month encoded in numeric and it is more



(a) Histogram of Longitude (c) Histogram of Latitude (e) Histogram of Convenience

(b) Histogram of Distance (d) Histogram of Price (f) Histogram of Age

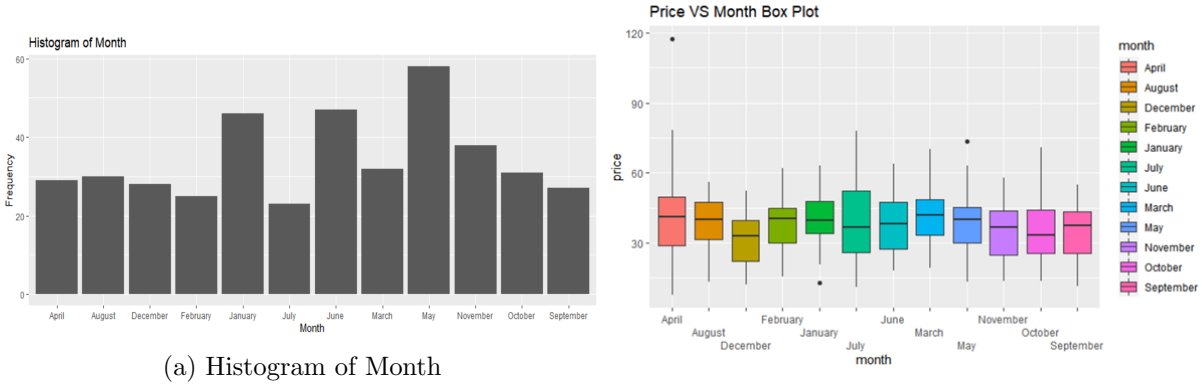
Figure 2.4: Histogram of Numerical Variables

meaningful to look at the descriptive statistics for "Month". The histograms of the variables are shown in Figure 2.4 We can see that histograms of both Latitude (Figure 2.4c) and Longitude (Figure 2.4a) are left-skewed. Most of the houses have Longitude between 121.53- 121.55 degrees, and Latitude between 24.96-24.98 degrees. This is expected since the data has been taken from Scindian District, New Taipei City, Taiwan, which means only houses from a specific location are present in the dataset. The histogram of Convenience (Figure 2.4e) shows that nearly 67 houses have no convenience store near to them. The most common number of convenience stores are 0 and 5. The histogram of Distance(metres) (Figure 2.4b) shows that the histogram is highly right-skewed. There are over 200 houses that have MRT stations within 500 metres and around 60 that have MRT stations between 500-1000m distance. This can mean that most of the houses have better value as the people living there would not have to spend a high amount on transportation. The histogram of Price (our target variable) (Figure 2.4d) shows that most of the houses are priced at 20-50 unit price (10000 New Taiwan dollars per Ping). There are around 50 houses that are priced at 50-60 unit price range. There are very few houses that are priced at the 110-120 unit price range, suggesting that they must be very high premium houses. The histogram of Age(in years) (Figure 2.4f) shows that most of the houses have the age of 10-20 years from the transaction date. Around 60 houses are less than 5 years old and 60 houses are between 30-35 years. This means that most of the houses available are less than 20 years old.

The boxplots of price and different months of transactions (Figure 2.5b) show that no particular month is more significant than the other. There are a few outliers seen in this plot.

We then plot all variables with respect to our target variable (Price) to get any graphical insight and determine their relationship. Their Scatter plots are visible in Figure 2.6

The Graph of Age vs Price shows that the price remains almost constant with a small curve with respect to age with very few outliers to this norm. The relation between



(a) Histogram of Month

(b) Price vs Month Boxplot

Figure 2.5: Descriptive Statistics of the categorical variable "Month"

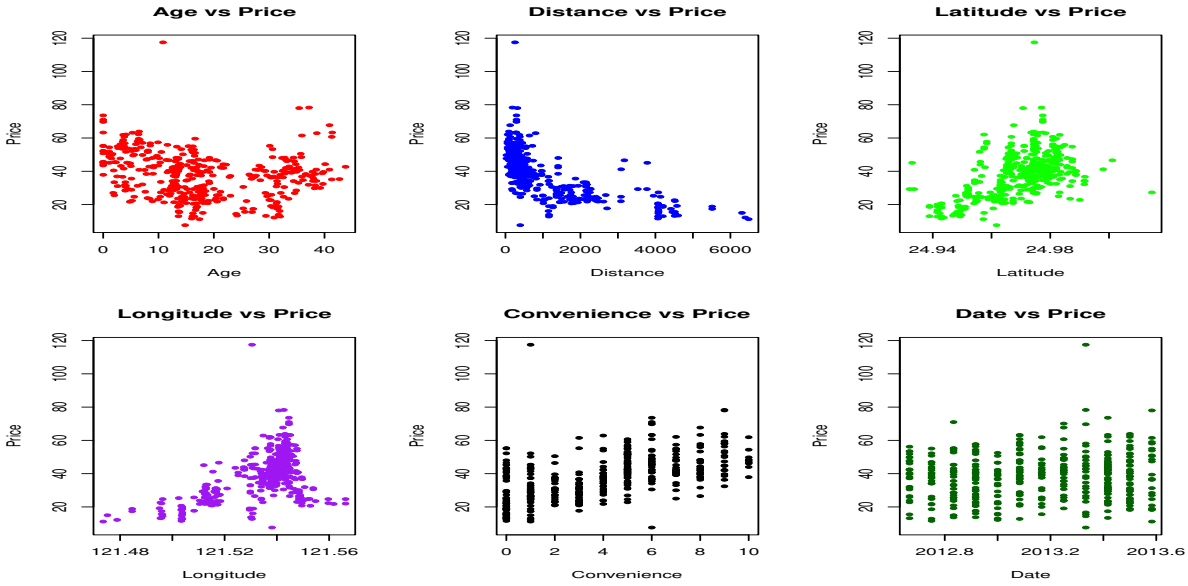


Figure 2.6: Scatterplots of predictors vs Price

Distance and Price seems to be polynomial. It is also visible that the houses with a small Distance to the MRT stations have higher prices. This could be due to the accessibility these houses provide. This is also visible in the Convenience vs Price plot, where the price increases with the number of convenience stores. We also see that price increases when the variable Latitude and Longitude increases, although the nature of their relationship is not obvious.

Chapter 3

Methods

3.1 Multiple Linear Regression

The first method that we will be implementing is Multiple Linear Regression. It is represented as -

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + e_i$,where y is our response variable, and β are the coefficients of x.

We start by shuffling the data and then dividing our data into training (2/3) and testing (1/3) sets. We train our model using the training data and measure its predictive performance using the testing data.

```
Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.458e+04  1.148e+04  -3.012  0.00285 **
date          1.378e+01  5.617e+00   2.452  0.01485 *
age          -2.842e-01  5.004e-02  -5.678 3.62e-08 ***
distance     -4.346e-03  6.166e-04  -7.048 1.61e-11 ***
convenience   1.192e+00  2.390e-01   4.986 1.13e-06 ***
latitude      2.761e+02  5.732e+01   4.818 2.46e-06 ***
monthAugust   2.535e+00  3.472e+00   0.730  0.46609
monthDecember -2.761e+00  2.653e+00  -1.041  0.29888
monthFebruary -3.461e+00  2.914e+00  -1.188  0.23598
monthJanuary  -4.173e-01  2.267e+00  -0.184  0.85412
monthJuly     -6.965e+00  4.187e+00  -1.663  0.09742 .
monthJune     -7.526e+00  3.537e+00  -2.128  0.03428 *
monthMarch    -5.320e+00  2.793e+00  -1.905  0.05792 .
monthMay      -8.718e+00  3.075e+00  -2.836  0.00493 **
monthNovember -3.106e+00  2.523e+00  -1.231  0.21947
monthOctober  -4.478e-01  2.770e+00  -0.162  0.87171
monthSeptember         NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.218 on 261 degrees of freedom
Multiple R-squared:  0.5876,    Adjusted R-squared:  0.5639
F-statistic: 24.79 on 15 and 261 DF,  p-value: < 2.2e-16
```

(a) Summary of model

```
Start:  AIC=1246.07
price ~ date + age + distance + convenience + latitude + month

Step:  AIC=1246.07
price ~ age + distance + convenience + latitude + month

             Df Sum of Sq   RSS    AIC
- month      11    1310.7 23490 1240.0
<none>                    22180 1246.1
- latitude    1    1972.5 24152 1267.7
- convenience 1    2112.2 24292 1269.3
- age         1    2740.0 24920 1276.3
- distance    1    4221.7 26401 1292.3

Step:  AIC=1239.97
price ~ age + distance + convenience + latitude

             Df Sum of Sq   RSS    AIC
<none>                    23490 1240.0
- latitude    1    2012.5 25503 1260.7
- convenience 1    2263.4 25754 1263.5
- age         1    2595.6 26086 1267.0
- distance    1    4457.0 27947 1286.1

Call:
lm(formula = price ~ age + distance + convenience + latitude,
    data = train)

Coefficients:
(Intercept)          age     distance  convenience     latitude
 -6.723e+03   -2.701e-01   -4.299e-03    1.206e+00    2.709e+02
```

(b) Backward elimination using AIC

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.723e+03  1.401e+03  -4.797 2.66e-06 ***
age         -2.701e-01  4.927e-02  -5.482 9.59e-08 ***
distance    -4.299e-03  5.984e-04  -7.184 6.51e-12 ***
convenience  1.206e+00  2.356e-01   5.119 5.81e-07 ***
latitude     2.709e+02  5.613e+01   4.827 2.31e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.293 on 272 degrees of freedom
Multiple R-squared:  0.5632,    Adjusted R-squared:  0.5568
F-statistic: 87.68 on 4 and 272 DF,  p-value: < 2.2e-16
```

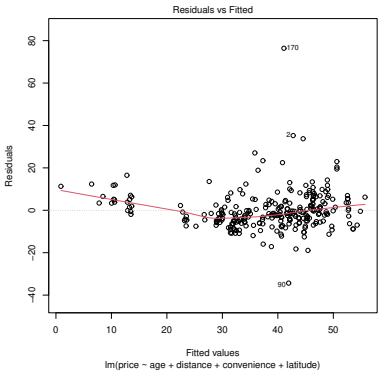
(c) Model Summary after removing Date and Month

Figure 3.1: Multiple Linear Regression Model

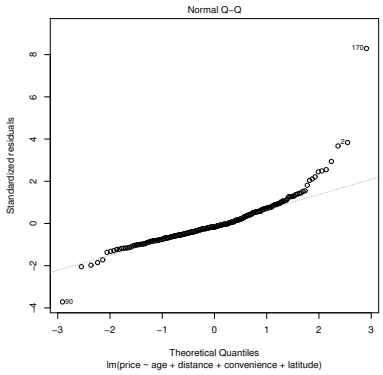
Then, we move to fitting the model. Since we know that there exists a high correlation between Distance and Longitude, we remove the predictor Longitude and fit our linear model on the remaining predictors. We get an AIC (Akaike Information criterion) of 1246.07 and R-squared 0.5876 (shown in Figure 3.1a).

We now want to remove the unnecessary predictors from our model. This can be done using Backward elimination with AIC as the criterion. We observe that the lowest AIC value is achieved when we drop the Month and Date predictors. Thus, we now fit a linear model consisting of Distance, Age, Convenience and Latitude as the predictors. This gives an AIC value 1239.97 (shown in Figure 3.1b).

Now, we want to run the diagnostic tests for our model. We plot the Residuals vs Fitted values (Figure 3.2a) to check whether our model follows the constant variance assumption but we notice that we have a fan shaped graph, which means we don't have constant variance. We also notice that the QQ plot (Figure 3.2b) is not a straight line, which could mean a non-normality of errors. We run the Shapiro-Wilks Test and get a p-value < 0.05 (Figure 3.2c). So we reject the null hypothesis of normality of errors.



(a) Residual vs Fitted



(b) QQ Plot

```
        Shapiro-Wilk normality test

data:  residuals(simple.model.linear)
W = 0.8272, p-value < 2.2e-16
```

(c) Shapiro-Wilks Test

Figure 3.2: Diagnostics after removal of Date and Month

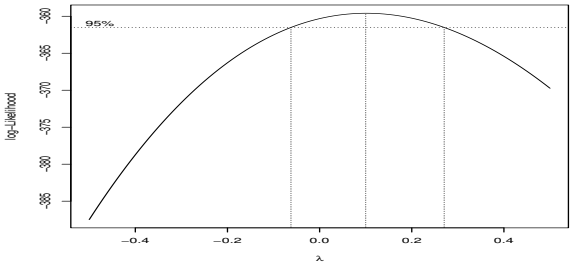
A possible remedy of non-normality of errors is to apply BoxCox Transformation to the response variable. From the BoxCox plot (Figure 3.3a), we get a lambda value of 0. We also note that 1 doesn't fall in the 95% CI, so we reject the null hypothesis that no transformation is required. Thus, we apply log transformation to the response.

After fitting a new model on the log-transformed response (Figure 3.3c), we run Backward elimination again to see if we can drop any further variables, but dropping any other predictor causes the AIC to increase, so we don't remove any other predictor. We then plot the Residuals vs Fitted values plot (Figure 3.3b) and QQ plot (Figure 3.3d) again, and notice that further transformation may be required. So we plot the Residuals with each predictor (Figure 3.4). We observe that the Residuals vs Distance (Figure 3.4b) plot shows a polynomial relation, so we try including higher order polynomials for the Distance variable.

We use the Forward Elimination and start by adding higher degree terms to the model until the last added term is not significant. We stop at d=3 since the fourth degree is not significant. Thus, our model now includes the predictors : Age, Distance, Distance^2, Distance^3, Convenience, and Latitude (Figure 3.5a).

We again test for constant variance of errors using Breusch-Pagan Test, which fails to reject the null hypothesis (Figure 3.5b). We now look for any unusual points in the dataset. We have 13 high leverage points. Using a t-test to conduct the outlier test, we notice there are 2 outliers but no influential points. Since the removal of 2 data points shouldn't change the true fundamental relation between the predictors and response, we remove the outliers.

We train our model again on this clean dataset and notice all our predictors are significant. We also plot the Residuals vs Fitted value and observe that there is no pattern. We also

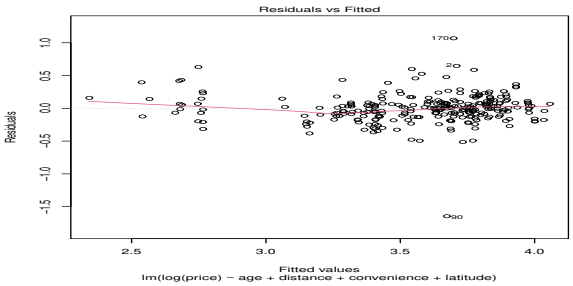


(a) Box-Cox Plot

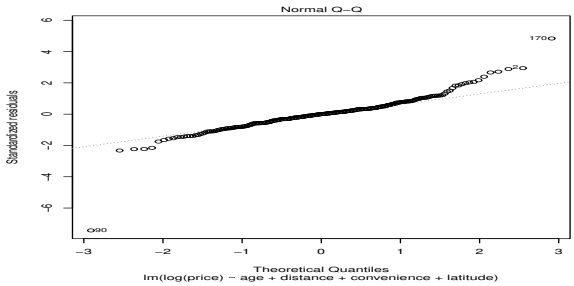
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.205e+02  3.358e+01  -6.568 2.58e-10 ***
age         -6.788e-03  1.180e-03  -5.750 2.39e-08 ***
distance    -1.534e-04  1.434e-05 -10.703  < 2e-16 ***
convenience  2.789e-02  5.646e-03   4.939 1.37e-06 ***
latitude     8.982e+00  1.345e+00   6.679 1.35e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2227 on 272 degrees of freedom
Multiple R-squared:  0.6874,    Adjusted R-squared:  0.6828
F-statistic: 149.6 on 4 and 272 DF,  p-value: < 2.2e-16
```

(c) Model Summary

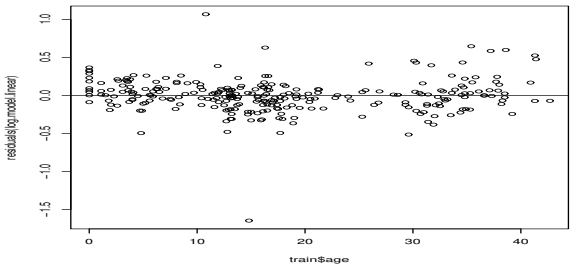


(b) Residuals vs Fitted Values

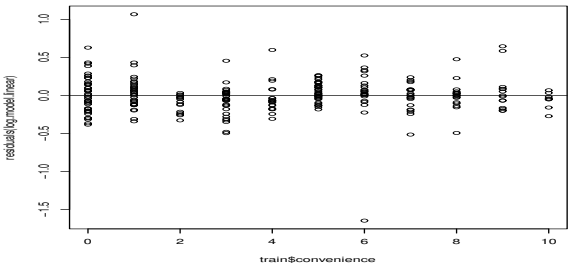


(d) QQ Plot

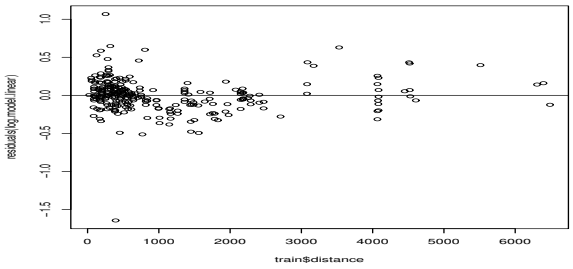
Figure 3.3: Log Transformed Model



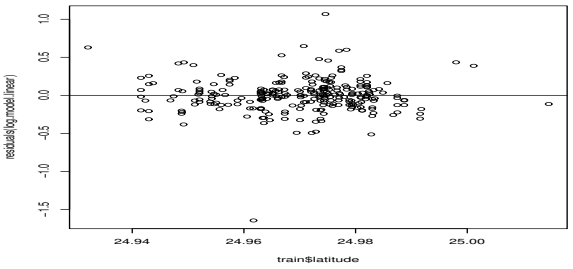
(a) Residuals vs Age



(c) Residuals vs Convenience



(b) Residuals vs Distance



(d) Residuals vs Latitude

Figure 3.4: Residuals vs Predictors

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.381e+02  3.202e+01  -7.437 1.37e-12 ***
age           -6.599e-03  1.118e-03  -5.903 1.07e-08 ***
distance      -5.236e-04  6.579e-05  -7.959 4.82e-14 ***
I(distance^2)  1.440e-07  2.857e-08   5.039 8.58e-07 ***
I(distance^3) -1.398e-11  3.308e-12  -4.225 3.27e-05 ***
convenience    1.535e-02  5.756e-03   2.667  0.00811 **
latitude       9.695e+00  1.283e+00   7.559 6.34e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2101 on 270 degrees of freedom
Multiple R-squared:  0.7238,    Adjusted R-squared:  0.7177
F-statistic: 117.9 on 6 and 270 DF,  p-value: < 2.2e-16
```

(a) Model Summary

```
        studentized Breusch-Pagan test

data:  cubic.loglinear.model
BP = 1.9351, df = 6, p-value = 0.9256
```

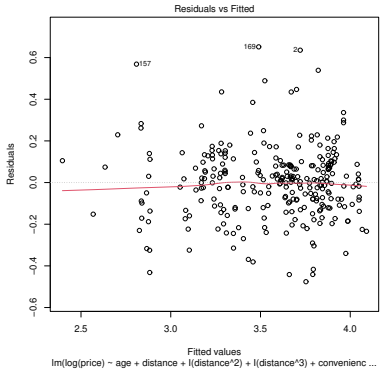
(b) Breusch-Pagan Test

Figure 3.5: Model After adding cubic terms

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.147e+02  2.683e+01  -8.002 3.72e-14 ***
age           -6.589e-03  9.339e-04  -7.056 1.47e-11 ***
distance      -4.959e-04  5.516e-05  -8.989  < 2e-16 ***
I(distance^2)  1.328e-07  2.391e-08   5.556 6.65e-08 ***
I(distance^3) -1.285e-11  2.766e-12  -4.645 5.33e-06 ***
convenience    2.048e-02  4.849e-03   4.223 3.30e-05 ***
latitude       8.754e+00  1.075e+00   8.147 1.43e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1754 on 268 degrees of freedom
Multiple R-squared:  0.7903,    Adjusted R-squared:  0.7857
F-statistic: 168.4 on 6 and 268 DF,  p-value: < 2.2e-16
```

(a) Model Summary



(b) Residual vs Fitted Values

```
        Durbin-Watson test

data:  clean.cubic.loglinear.model
DW = 1.9394, p-value = 0.3104
alternative hypothesis: true autocorrelation is greater than 0
```

(c) Durbin-Watson Test

Figure 3.6: Model After removing outliers

run the Durbin-Watson test to check for correlation of errors but fail to reject the null hypothesis. Thus, our model diagnostics are now complete.

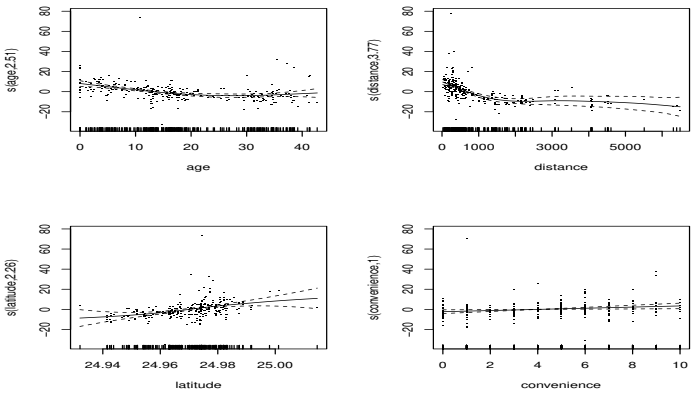
We now proceed to test the model performance on the testing data. One of the key metrics is Root Mean Squared Error to check how good our model can predict the price of a house while reducing the error as far as possible. For our model we get RMSE of 0.2.

3.2 Smoothing Splines

Smoothing Splines are a nonparametric approach to regression problems. The model is represented as $y_i = f(x_i) + e_i$ We select $\hat{f}$ to minimize the mean squared error subject to a function roughness penalty:

$$\frac{1}{n}\sum(Y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx.$$

The first term is a measure of how close the fitted values are to the observed values and the second term is a measure of function roughness. The solution of this equation is a natural cubic spline in each sub interval. The function that we try to estimate here is continuous in the first and second-order derivatives. Since we are using multiple predictors to build our model we use the library mgcv in R.



(a) Transformation functions for the model fit by mgcv

```
Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.0404     0.4926   77.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                edf Ref.df      F  p-value
s(age)        2.511  3.126 16.161  < 2e-16 ***
s(distance)   3.770  4.591 19.062  < 2e-16 ***
s(latitude)   2.260  2.949 12.297 7.54e-07 ***
s(convenience) 1.000  1.000  6.133   0.0139 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.655   Deviance explained = 66.7%
GCV = 69.876  Scale est. = 67.217     n = 277
```

(b) Model Summary

Figure 3.7: Smoothing Splines Model

We fit the model on the predictors that we found were significant in our previous approach, i.e., Age, Distance, Convenience and Latitude. We now proceed to test the model performance on the testing data. For our model we get RMSE of 7.0.

3.3 Random Forests

Regression Trees are another nonparametric approach to regression problems. Since simple decision trees can be prone to overfitting and instability due to their non-robustness, ensemble methods like bagging and boosting are commonly applied to get a better predictive model.

Random Forest was first proposed by Tim Kam Ho in 1995 and later extended by Breiman in 2005. Random Forest uses Bootstrap Aggregation (Bagging) and feature bagging to effectively build a forest made of many trees. They are commonly used for both regression and classification problems. The algorithm builds an ensemble of smaller decision trees by Bootstrap Aggregation. Each decision tree in the random forest provides a classification or prediction. By aggregating over the ensemble, final predictions are then made.

We now proceed to test the model performance on the testing data. For our model we get RMSE of 6.3.

Method	R^2	RMSE on Testing Set
Multiple Linear Regression	0.790	0.20
Smoothing Splines	0.667	7.01
Random Forest	0.692	6.35

Table 3.1: Performance of Models

Chapter 4

Discussion and Conclusion

In this project, we have analyzed several numerical and categorical variables for building prediction models to predict house prices like Multiple Linear Regression, Smoothing Splines and Random Forest. Their predictive performance can be summarised in the table below.

From Table 3.1, we can observe that the lowest (best) Root mean squared error (RMSE) is observed to be 0.2 for the Multiple Linear Regression model. The next best model Random Forest gives us an RMSE of 6.3 and with Smoothing splines we get an RMSE of 7.0.

Apart from their predictive performance, we also need to compare these methods to provide a suitable recommendation for this dataset.

The Linear regression model is the simplest way to express the relationship between the multiple predictor variables and predicted variables. The speed of Linear regression is very fast as it doesn't include complicated calculations and predicts fast when the data amount is large and hence is computationally efficient. The model clearly tells us the impact of each predictor variable which are independent of each other on the response variable. It also has a few disadvantages. The outliers can have a huge impact on the output. Our predictors may be correlated and hence diagnostics are required to remove them, thus, reducing the dimensions as well.

Nonparametric models like Smoothing splines can be needed in some applications where we have very little idea of an appropriate form for the model. They can be useful in interpolating values in noisy data. Since extrapolation requires some assumptions about how the function will behave outside the range of the data, parametric methods do better here as they are more transparent about how they will behave. Nonparametric regression fits are also hard to interpret in higher dimensions where visualization is difficult.

Random Forests have an easy application and can be parallelized. They are robust to outliers and noise and provide good accuracy. Although bagging gives a higher predictive performance, it loses the interpretability of predictors. This puts it at a slight disadvantage when compared to methods like Multiple Linear Regression

Thus, for this dataset, we suggest the use of Multiple Linear Regression models, since they have a better predictive performance and allow interpretability of their results.