REGRESSION ANALYSIS

**Introduction**

The market historical data set of real estate valuation are collected from Sindian Dist., New Taipei City, Taiwan. This dataset gives us a chance to look into the data on what really influences the value of a house, which makes us realize about the tricky ways by which we can buy a house.The dataset consist of 6 independent variables(response) and 1 dependent variable(predictor).

**Attribute Information:**

The independent variables are as follows:   
X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)   
X2=the house age (unit: year)   
X3=the distance to the nearest MRT station (unit: meter)   
X4=the number of convenience stores in the living circle on foot (integer)   
X5=the geographic coordinate, latitude. (Unit: degree)   
X6=the geographic coordinate, longitude. (Unit: degree)   
  
The dependent variable is:   
Y= house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

Now we are going to do the regression analysis for thereal estate dataset for predicting the house price of unit area using the other independent variables. But before going to the analysis first we need to clean the data by removing the null values and outliers so that the data will be ready for analysis. Because existence of null values and outliers may reduce the Goodness of fit for the model. The significance of the model is given by the linear model and its summary, which shows the most important variables that affect the rate of a house in an area. So let’s do some basic analysis on the data.

Now we are importing the dataset and specifying the number of column’s, rows and showing its dimension.

data=read.csv("C:/Users/ANJU/Downloads/Real estate.csv")

**Dimension of the dataset**

dim(data)

## [1] 414 8

Inference: The data set shows 414 rows and 8 columns.

ncol(data)

## [1] 8

colnames(data)

## [1] "No"   
## [2] "X1.transaction.date"   
## [3] "X2.house.age"   
## [4] "X3.distance.to.the.nearest.MRT.station"  
## [5] "X4.number.of.convenience.stores"   
## [6] "X5.latitude"   
## [7] "X6.longitude"   
## [8] "Y.house.price.of.unit.area"

**Structure of the data**

str(data)

## 'data.frame': 414 obs. of 8 variables:  
## $ No : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ X1.transaction.date : num 2013 2013 2014 2014 2013 ...  
## $ X2.house.age : num 32 19.5 13.3 13.3 5 7.1 34.5 20.3 31.7 17.9 ...  
## $ X3.distance.to.the.nearest.MRT.station: num 84.9 306.6 562 562 390.6 ...  
## $ X4.number.of.convenience.stores : int 10 9 5 5 5 3 7 6 1 3 ...  
## $ X5.latitude : num 25 25 25 25 25 ...  
## $ X6.longitude : num 122 122 122 122 122 ...  
## $ Y.house.price.of.unit.area : num 37.9 42.2 47.3 54.8 43.1 32.1 40.3 46.7 18.8 22.1 ...

Inference: The structure of a dataset gives basic data structures present in the data. The basic data structures used in r include vectors, lists, matrices, data frames and factors. It is used to handle multiple values that means we can’t work with data having single value.

Null values of the data

colSums(is.na(data))

## No   
## 0   
## X1.transaction.date   
## 0   
## X2.house.age   
## 0   
## X3.distance.to.the.nearest.MRT.station   
## 0   
## X4.number.of.convenience.stores   
## 0   
## X5.latitude   
## 0   
## X6.longitude   
## 0   
## Y.house.price.of.unit.area   
## 0

**Inference: Null is a term used to represent a missing value .Here we are checking if there is any null values in our data because if there is any null values in the data it will affect our models and will reduce the significance. Hence, we want to omit the null values if there are any. The above code shows that there is no null values in the data.**

**Exploratory data analysis**

summary(data[3:8])

## X2.house.age X3.distance.to.the.nearest.MRT.station  
## Min. : 0.000 Min. : 23.38   
## 1st Qu.: 9.025 1st Qu.: 289.32   
## Median :16.100 Median : 492.23   
## Mean :17.713 Mean :1083.89   
## 3rd Qu.:28.150 3rd Qu.:1454.28   
## Max. :43.800 Max. :6488.02   
## X4.number.of.convenience.stores X5.latitude X6.longitude   
## Min. : 0.000 Min. :24.93 Min. :121.5   
## 1st Qu.: 1.000 1st Qu.:24.96 1st Qu.:121.5   
## Median : 4.000 Median :24.97 Median :121.5   
## Mean : 4.094 Mean :24.97 Mean :121.5   
## 3rd Qu.: 6.000 3rd Qu.:24.98 3rd Qu.:121.5   
## Max. :10.000 Max. :25.01 Max. :121.6   
## Y.house.price.of.unit.area  
## Min. : 7.60   
## 1st Qu.: 27.70   
## Median : 38.45   
## Mean : 37.98   
## 3rd Qu.: 46.60   
## Max. :117.50

**Inference: The exploratory data analysis gives the basic statistical calculations like mean, median etc... For each variable present in the data. From the above data we can see that X3.distance.to.the.nearest.MRT.station has the maximum mean with a value of 1083.89 and X4.number.of.convenience.stores has the minimum mean with a value of 4.**

**Correlation of the data**

cor(data[3:8])

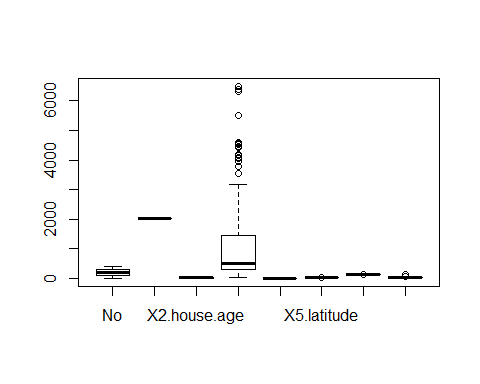
## X2.house.age  
## X2.house.age 1.00000000  
## X3.distance.to.the.nearest.MRT.station 0.02562205  
## X4.number.of.convenience.stores 0.04959251  
## X5.latitude 0.05441990  
## X6.longitude -0.04852005  
## Y.house.price.of.unit.area -0.21056705  
## X3.distance.to.the.nearest.MRT.station  
## X2.house.age 0.02562205  
## X3.distance.to.the.nearest.MRT.station 1.00000000  
## X4.number.of.convenience.stores -0.60251914  
## X5.latitude -0.59106657  
## X6.longitude -0.80631677  
## Y.house.price.of.unit.area -0.67361286  
## X4.number.of.convenience.stores  
## X2.house.age 0.04959251  
## X3.distance.to.the.nearest.MRT.station -0.60251914  
## X4.number.of.convenience.stores 1.00000000  
## X5.latitude 0.44414331  
## X6.longitude 0.44909901  
## Y.house.price.of.unit.area 0.57100491  
## X5.latitude X6.longitude  
## X2.house.age 0.0544199 -0.04852005  
## X3.distance.to.the.nearest.MRT.station -0.5910666 -0.80631677  
## X4.number.of.convenience.stores 0.4441433 0.44909901  
## X5.latitude 1.0000000 0.41292394  
## X6.longitude 0.4129239 1.00000000  
## Y.house.price.of.unit.area 0.5463067 0.52328651  
## Y.house.price.of.unit.area  
## X2.house.age -0.2105670  
## X3.distance.to.the.nearest.MRT.station -0.6736129  
## X4.number.of.convenience.stores 0.5710049  
## X5.latitude 0.5463067  
## X6.longitude 0.5232865  
## Y.house.price.of.unit.area 1.0000000

**Inference: Correlation is a method of statistical evaluation used to study the strength of a relationship between two numerically measured variables. Its value varies from -1 to 1. If the value of correlation is -1 which shows a strong negative correlation, if the value is 1 it means there exist a strong negative correlation and the value of correlation = 0 means there is a weak relation exist between the variables. And from the above data we can see there is a strong negative correlation exist between X3.distance.to.the.nearest.MRT.station and X6.longitude by a value of -0.8 and a positive correlation is shown by X4.number.of.convenience.stores and Y.house.price.of.unit.area by a value of 0.57.**

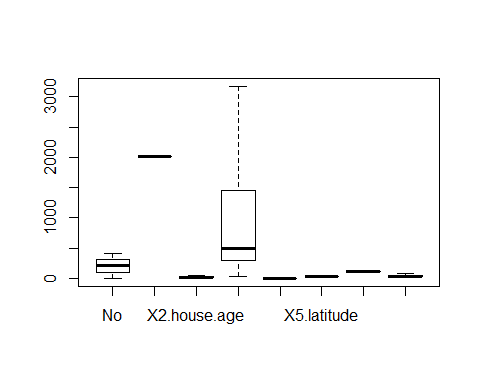
**Weak correlation is shown by the variablesX2.house.age and X3.distance.to.the.nearest.MRT.station by a value of 0.02 which is so close to zero.**

**Boxplot**

boxplot(data)



boxplot(data,outline =FALSE)

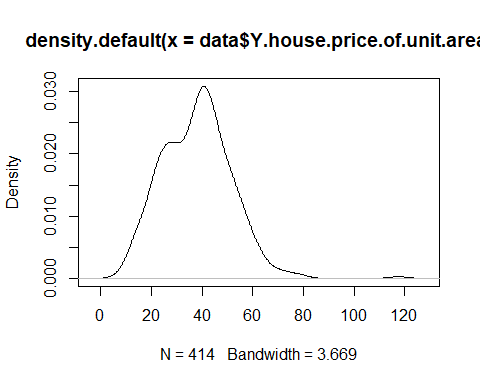


**Inference: Boxplot is usually plotted to find out if there are outliers in our data. An outlier is a data point on a graph or in a set of results that is very much bigger or smaller than the next nearest data point.**

**We have found some amount of outliersin the variable X3.distance.to.the.nearest.MRT.station. Outliers in our data can affect the model fitting, but when we remove a large number of outliers it affects the fitting of model. Therefore we will hide the outliers present in the data using the code outline= false, which won’t show the outliers present.**

**Density plot**

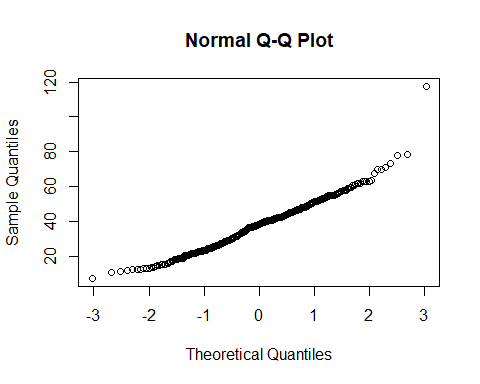
density\_plot=density(data$Y.house.price.of.unit.area)  
plot(density\_plot)



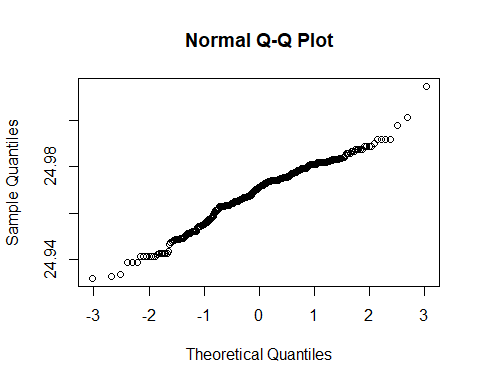
**Inference: A density plot is a representation of distribution of a numeric variable. It uses a kernel density estimate to show the probability density function of the variable. Density plots can be considered as plots of smoothed histograms. The visual representation of density plot will show whether the variable is normally distributed or not. The above graph shows that the predictor variable Y.house.price.of.unit.area has a normal distribution since the shape of the graph is almost a bell shaped curve which is the representation of normal distribution.**

Finding the type of distribution

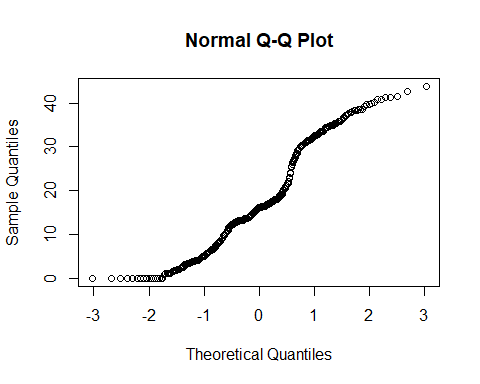
qqnorm(data$Y.house.price.of.unit.area)



qqnorm(data$X5.latitude)



qqnorm(data$X2.house.age)



**Inference: q-q norm is a generic function for producing normal Q-Q plot which is a probability plot for comparing the two probability distribution by plotting their quantiles against each other. X-axis represents theoretical quantiles and y-axis represents sample quantiles. If we get a rough straight line after plotting both sets of quantiles which indicates that it is normally distributed.The above graphs shows a straight line which indicates that the type of distribution is normal distribution.**

linearmodel1=lm(Y.house.price.of.unit.area~.,data=data)  
linearmodel2=lm(data$Y.house.price.of.unit.area~data$X3.distance.to.the.nearest.MRT.station,data=data)

AIC(linearmodel1,linearmodel2)

## df AIC  
## linearmodel1 9 2990.939  
## linearmodel2 3 3091.072

**Inference: linear model describe a continuous response variable as a function of one or more predictor variable. Here we have built a multiple linear regression model1 using 6 independent variables and 1 dependent variable. Model 2 is built using one dependent and independent variable.**

**After building two models we have determined the AIC value of two models which stands for “Akaike’s Information Criteria” showing which model is better. If we are given a set of models the AIC estimates the quality of each model relative to each of the other model. Smaller the AIC value better the model. And here we obtained a smaller AIC value for model 1. Therefore the first model is better than the second.**

summary(linearmodel1)

##   
## Call:  
## lm(formula = Y.house.price.of.unit.area ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -36.003 -5.196 -0.990 4.181 75.384   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) -1.404e+04 6.788e+03 -2.068  
## No -3.593e-03 3.653e-03 -0.984  
## X1.transaction.date 5.079e+00 1.559e+00 3.259  
## X2.house.age -2.708e-01 3.855e-02 -7.026  
## X3.distance.to.the.nearest.MRT.station -4.521e-03 7.189e-04 -6.289  
## X4.number.of.convenience.stores 1.129e+00 1.882e-01 6.000  
## X5.latitude 2.247e+02 4.458e+01 5.040  
## X6.longitude -1.442e+01 4.863e+01 -0.297  
## Pr(>|t|)   
## (Intercept) 0.03927 \*   
## No 0.32590   
## X1.transaction.date 0.00121 \*\*   
## X2.house.age 9.04e-12 \*\*\*  
## X3.distance.to.the.nearest.MRT.station 8.28e-10 \*\*\*  
## X4.number.of.convenience.stores 4.37e-09 \*\*\*  
## X5.latitude 7.02e-07 \*\*\*  
## X6.longitude 0.76691   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.858 on 406 degrees of freedom  
## Multiple R-squared: 0.5834, Adjusted R-squared: 0.5762   
## F-statistic: 81.21 on 7 and 406 DF, p-value: < 2.2e-16

**Inference: The main parameter in the summary of linear model is the P value.**

**When we are looking for p value we should consider the hypothesis. And our null hypothesis is there is no relationship between the variables and the alternative hypothesis is there is relationship between the variables. When the p value is less than 0.05 we will reject the null hypothesis. In order to check the statistical significance, the value of P is p value<0.05. And it can be visually interpreted by looking the number of stars at the end of the row. More the number of stars beside the p value more significant the variable and here 3 stars are beside the p value and it means it is closer to 0 and more significant. Here p value= 2.2 e-16 which is less than 0.05 and it is closer to zero. Therefore we reject the null hypothesis and accept the alternative hypothesis. So we can conclude that there exist a statistical significance between the variables x2,x3,x4 and x5.**

**Next important parameter is t- value which is the ratio of departure of the estimated value of the parameter from its hypothesized value to its standard error**

**t- Value = estimate /std.error**

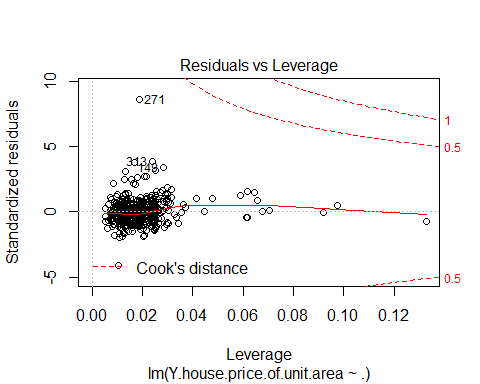
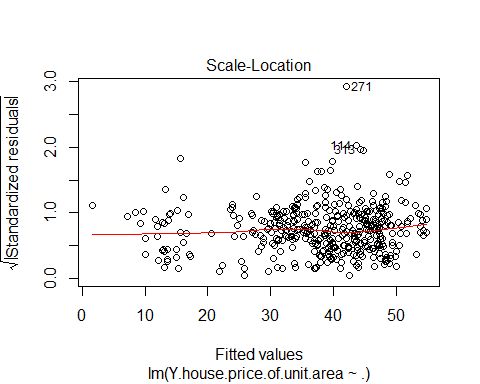
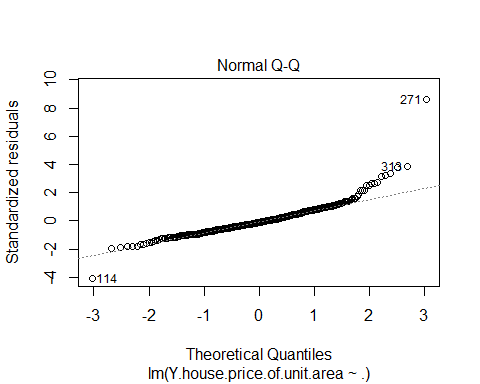
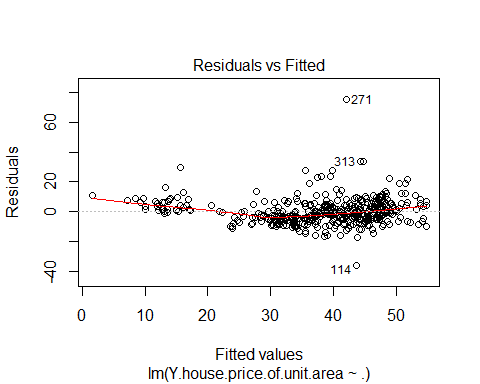
**It measures the size of the difference, relative to the variation in the sample data. The greater the magnitude of T, greater the evidence against the null hypothesis. Even the negative t- value indicates an evidence against null hypothesis. So, here the variables x4 and X5 shows a large t-value which indicates their significance.**

**Now standard error shows the standard deviation of the samples.**

**Linear Diagnostic**

plot(linearmodel1)

**Inference: Linear diagnostic plot shows the residuals in four different way. Residual VS fitted, Normal Q-Q plot, Scale location and residual VS leverage.**



**Inference: Linear diagnostic plot shows the residuals in four different way. Residual VS fitted, Normal Q-Q plot, Scale location and residual VS leverage.**

**Residual VS fitted: This is the most frequently fitted plot, which is the scatter plot of residuals on the Y axis and fitted values in the x axis. The plot helps us to detect non- linearity, unequal-error variance and outliers. The above graph of residual VS fitted is an example of a well behaved plot which means the residuals bounce randomly around the 0 line. Therefore the assumption that the relationship is linear is reasonable.**

**Normal Q-Q plot: normal Q-Q plot is a probability plot for comparing the two probability distribution by plotting their quantiles against each other. X-axis represents theoretical quantiles and y-axis represents sample quantiles. If we get a rough straight line after plotting both sets of quantiles which indicates that it is normally distributed. So the type of distribution is normal distribution.**

**Scale Location: This is also called spread location plot. This plot shows if the residuals are spread equally along the ranges of predictors. The plot is good if we get a horizontal line with equally spread points. In the above graph red line is approximately horizontal but the spread of magnitude seems to be lowest in the fitted values close to zero which suggests a heteroskedasticity.**

**Residual VS leverage: The plot helps us to identify the influential data points in the model. Outliers can be influential and some points within a normal range in the model could be very influential. One main character of this plot is cooks distance, which is represented by the dotted lines. Any point lying outside the cooks distance is considered as the outliers.**

**And the above plot of residual VS leverage shows that there is no such outliers present as all values are inside the cooks distance. Therefore there is no influential points.**

confint(linearmodel1,level=0.99)

## 0.5 % 99.5 %  
## (Intercept) -3.160557e+04 3.529408e+03  
## No -1.304852e-02 5.861522e-03  
## X1.transaction.date 1.045362e+00 9.112812e+00  
## X2.house.age -3.706086e-01 -1.710753e-01  
## X3.distance.to.the.nearest.MRT.station -6.381182e-03 -2.660398e-03  
## X4.number.of.convenience.stores 6.421774e-01 1.616378e+00  
## X5.latitude 1.093108e+02 3.400350e+02  
## X6.longitude -1.402685e+02 1.114214e+02

**Inference: The confidence interval is a type of estimate computed from statistics of the observed data. The 99 % confidence interval defines a range of values that can be 99 % certain contains the population mean and since the data is small the confidence interval is more precise.**

**Conclusion**

After the analysis of real estate data set we conclude that for predicting the house rate the more significant variables that help us to determine on what basis we decide the house rate are distance to the MRT station, number of convenient stores and latitude. We have reached on such conclusion based on the linear model which shows the significance between the variables. Also the data follows a normal distribution, which shows the goodness of the model. The diagnostic plots also plays a major role since that shows that residuals are equally spreaded and the outliers has no influence on the model. So we can conclude that our model is better.