Project Title:

Real Estate Price Prediction using Multiple Linear Regression

Objective

1. Explain the need for MLR for the selected dataset.
2. Identifying the highly correlated variables using the Correlation matrix.
3. Showing a matrix scatter diagram between the variables of interest.

4. Detecting Multicollinearity

5. Fitting a multiple linear regression model to the selected response (y) and regression variables (highly correlated with 'y' only) and interpreting the estimated coefficients.

6.Constructing 95% and 99% confidence intervals for the individual parameters in the model.

Data description

Data gives the information about the price (per acre) of the real estate based on dependent variables such as house age, distance from metro station, longitude, latitude, number of convenience factors.

Regression models *are used to describe relationships between variables by fitting a line to the observed data. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.*

We are using multiple linear regression on this dataset as house price will depend on these given factors. Considering house price as the dependent variable and others as independent variable

The formula for multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

- y = *the predicted value of the dependent variable*
- β_0 = *the y-intercept (value of y when all other parameters are set to 0)*
- $\beta_1 X_1$ = *the regression coefficient* (β_1) *of the first independent variable* (X_1) *(a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)*
- … = *Do the same for however many independent variables you are testing*
- $\beta_n X_n$ = *the regression coefficient of the last independent variable*
- ϵ = *model error (a.k.a. how much variation there is in our estimate* y*)*

#Reading the data

```
df=read.csv ("C:/Users/ahuja/Downloads/Real estate.csv")

head(df)

##   No X1.transaction.date X2.house.age X3.distance.to.the.nearest.MRT.stati
on
## 1  1            2012.917         32.0                               84.878
82
## 2  2            2012.917         19.5                              306.594
70
## 3  3            2013.583         13.3                              561.984
50
## 4  4            2013.500         13.3                              561.984
50
## 5  5            2012.833          5.0                              390.568
40
## 6  6            2012.667          7.1                             2175.030
00
##   X4.number.of.convenience.stores X5.latitude X6.longitude
## 1                              10    24.98298     121.5402
## 2                               9    24.98034     121.5395
## 3                               5    24.98746     121.5439
## 4                               5    24.98746     121.5439
## 5                               5    24.97937     121.5425
## 6                               3    24.96305     121.5125
##   Y.house.price.of.unit.area
## 1                       37.9
## 2                       42.2
## 3                       47.3
## 4                       54.8
## 5                       43.1
## 6                       32.1
```

#Considering only the numerical independent variable variables

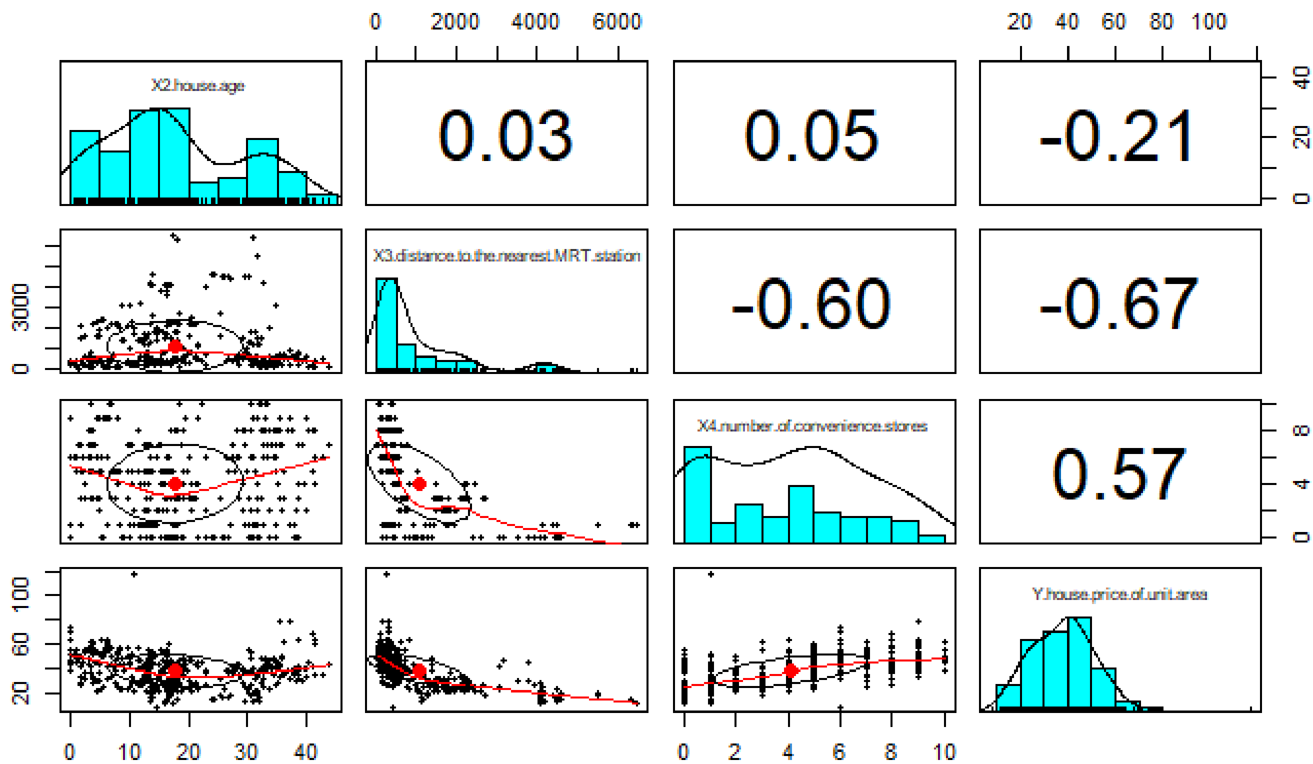
```
df1=df[,c(-1,-2,-6,-7)]

#Seperating independent and dependent variables
ind=df1$Y.house.price.of.unit.area
dep=df1[,-1]

cor(df1)

##                                         X2.house.age
## X2.house.age                              1.00000000
## X3.distance.to.the.nearest.MRT.station    0.02562205
## X4.number.of.convenience.stores           0.04959251
## Y.house.price.of.unit.area               -0.21056705
##                                         X3.distance.to.the.nearest.MRT.stat
ion
## X2.house.age                                                          0.02562
205
## X3.distance.to.the.nearest.MRT.station                                1.00000
000
## X4.number.of.convenience.stores                                      -0.60251
914
## Y.house.price.of.unit.area                                           -0.67361
286
##                                         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
## Y.house.price.of.unit.area                                   0.57100491
##                                         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
## Y.house.price.of.unit.area                               1.0000000
```

From the correlation matrix ,we can see that

#1. Correlation between house age and house price is -0.21 ie they are negatively correlated.
#2. Correlation between Area to nearest metro station and house price is -0.67 ie they are negatively correlated.

#3 Correlation between no of convenience and house price is 0.57 ie they are positively correlated.

```
pairs(df1,col="blue")
```



Through the scatter plots, we can observe that regressor variables are not independent There is a high correlation between regressor variables

#Fitting the multiple linear regression model

```
mlr=lm(ind~df1$X2.house.age+df1$X3.distance.to.the.nearest.MRT.station+df1$X4
.number.of.convenience.stores)
mlr

##
## Call:
## lm(formula = ind ~ df1$X2.house.age + df1$X3.distance.to.the.nearest.MRT.s
tation +
##     df1$X4.number.of.convenience.stores)
##
## Coefficients:
##                                (Intercept)
##                                  42.977286
##                           df1$X2.house.age
##                                  -0.252856
## df1$X3.distance.to.the.nearest.MRT.station
##                                  -0.005379
##        df1$X4.number.of.convenience.stores
##        1.297442
```

Hence the linear regression model is given by:

$$Y = 42.977286 - 0.252856X_1 - 0.005379X_2 + 1.297442X_3$$

$$where\ Y = House\ price$$

$$X_1 = House\ age$$

$$X_3 = Distance\ to\ nearest\ metro\ station$$

$$X_4 = Number\ of\ convenience\ available$$

```
summary(mlr)

##
## Call:
## lm(formula = ind ~ df1$X2.house.age + df1$X3.distance.to.the.nearest.MRT.s
tation +
##     df1$X4.number.of.convenience.stores)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.304  -5.430  -1.738   4.325  77.315
##
## Coefficients:
```

```
##                                               Estimate Std. Error t value
## (Intercept)                                  42.977286   1.384542  31.041
## df1$X2.house.age                             -0.252856   0.040105  -6.305
## df1$X3.distance.to.the.nearest.MRT.station -0.005379   0.000453 -11.874
## df1$X4.number.of.convenience.stores          1.297443   0.194290   6.678
##                                              Pr(>|t|)
## (Intercept)                                   < 2e-16 ***
## df1$X2.house.age                             7.47e-10 ***
## df1$X3.distance.to.the.nearest.MRT.station  < 2e-16 ***
## df1$X4.number.of.convenience.stores          7.91e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.251 on 410 degrees of freedom
## Multiple R-squared:  0.5411, Adjusted R-squared:  0.5377
## F-statistic: 161.1 on 3 and 410 DF,  p-value: < 2.2e-16
```

P value is less than level of significance , hence we reject the null hypothesis ie there exist a linear relationship between dependent and independent variables.

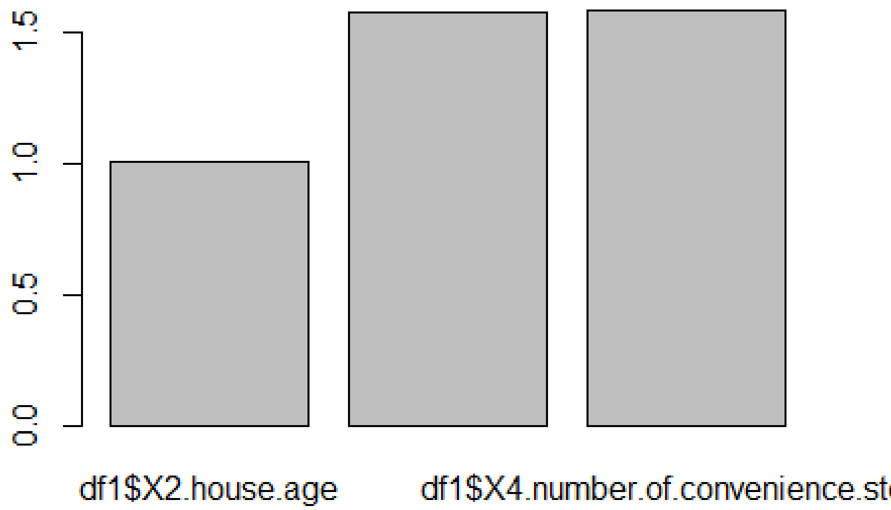
Also Adjusted R^2value is 0.5377, hence model is good fit for the data.

Since the independent variables are correlated, multicollinearity exists, Hence we need to find VIF for the linear model.

```
v=vif(mlr)
v
```

```
##                           df1$X2.house.age
##                                   1.007349
## df1$X3.distance.to.the.nearest.MRT.station
##                                   1.577579
##        df1$X4.number.of.convenience.stores
##        1.580431
```

```
barplot(v)
```



Since VIF is less than 5 for all the variables, we can continue with multi linear regression model.

#Calculating 95% confidence interval

```
confint(mlr,level=0.95)

##                                                2.5 %       97.5 %
## (Intercept)                               40.255598691 45.698973721
## df1$X2.house.age                          -0.331693553 -0.174018100
## df1$X3.distance.to.the.nearest.MRT.station -0.006269685 -0.004488574
## df1$X4.number.of.convenience.stores         0.915513969  1.679370983
```

The 95% confidence interval is given by (40.255598691,45.698973721)

#Calculating 99% confidence interval

```
confint(mlr,level=0.99)

##                                                0.5 %      99.5 %
## (Intercept)                               39.39426549 46.56030692
## df1$X2.house.age                          -0.35664335 -0.14906830
## df1$X3.distance.to.the.nearest.MRT.station -0.00655152 -0.00420674
## df1$X4.number.of.convenience.stores         0.79464495  1.80024000
```

The 99% confidence interval is given by (39.39426549,46.56030692)

Conclusion

1. *From the correlation matrix ,we can see that*

a. *Correlation between house age and house price is -0.21 ie they are negatively correlated.*

b. *Correlation between Area to nearest metro station and house price is -0.67 ie they are negatively correlated.*

c. *Correlation between no of convenience and house price is 0.57 ie they are positively correlated*

2. *Hence the linear regression model is given by:*

$$Y = 42.977286 - 0.252856X_1 - 0.005379X_2 + 1.297442X_3$$

$$where\ Y = House\ price$$

$$X_1 = House\ age$$

$$X_3 = Distance\ to\ nearest\ metro\ station$$

$$X_4 = Number\ of\ convenience\ available$$

3. *VIF is less than 5 for all the variables, we can continue with multi linear regression model*

4. *The 95% confidence interval is given by (40.255598691,45.698973721)*

5. *The 99% confidence interval is given by (39.39426549,46.56030692)*