**Lab 3**

**Multiple Linear Regression with 2 Regressors**

Name: Jesto Peter

Title: Multiple Linear Regression

Date: 02/12/2022

Class: 2MSTAT

* **Introduction**

Multiple linear regression estimates the relationship between two or more independent variables and one dependent variable. We can use multiple linear regression when we want to know how strong the relationship is between two or more [independent variables](https://www.scribbr.com/methodology/independent-and-dependent-variables/#independent) and one dependent variable.

The [correlation](https://www.investopedia.com/terms/c/correlation.asp) coefficient is a statistical measure of the strength of a [linear relationship](https://www.investopedia.com/terms/l/linearrelationship.asp) between two variables. Its values can range from -1 to 1. A correlation coefficient of -1 describes a perfect [negative](https://www.investopedia.com/terms/n/negative-correlation.asp) or [inverse](https://www.investopedia.com/terms/i/inverse-correlation.asp) correlation and a correlation coefficient of +1 describes a perfect positive correlation.

* **Objective**

In this Lab Session we will be considering the dataset on Housing. We have the Dependent Variable as “Price” and 4 regressors as independent variables.

We have to:

1. Explain the need for MLR for the selected dataset.
2. Find the correlation matrix and identify the highly correlated variables. Interpret it.
3. Draw a matrix scatter diagram between the variables of interest using ggplots and the pair function.
4. Are the regressors independent of each other? Justify your answer with the help of Q3.
5. Fit a multiple linear regression model to the selected response (y) and regressor variables (highly correlated with 'y' only) and interpret the estimated coefficients.
6. Construct the 95% and 99% confidence intervals for the individual parameters in the model.

* **Procedure**

*#importing the dataset*

house=read.csv("D:/Regression Analysis/3 LAB (2ND DEC)/house1.csv")  
head(house)

## price bedrooms bathrooms sqft\_living sqft\_lot  
## 1 221900 3 1.00 1180 5650  
## 2 538000 3 2.25 2570 7242  
## 3 180000 2 1.00 770 10000  
## 4 604000 4 3.00 1960 5000  
## 5 510000 3 2.00 1680 8080  
## 6 1230000 4 4.50 5420 101930

**Interpretation:** Use of Multiple Linear Regression: Here Price is the Dependent Variable and the Other 4 columns are the regressors. We will be using MLR for the House Dataset to predict the price of the houses based on the no. of bedrooms, bathrooms, square feet living and square feet lot.

#2. Finding the correlation matrix and identifying the highly correlated variables. Interpret it.  
  
cor(house)

## price bedrooms bathrooms sqft\_living sqft\_lot  
## price 1.00000000 0.30833837 0.52513407 0.7020437 0.08965521  
## bedrooms 0.30833837 1.00000000 0.51588364 0.5766707 0.03170324  
## bathrooms 0.52513407 0.51588364 1.00000000 0.7546653 0.08773966  
## sqft\_living 0.70204372 0.57667069 0.75466528 1.0000000 0.17282566  
## sqft\_lot 0.08965521 0.03170324 0.08773966 0.1728257 1.00000000

**Interpretation:** We see that the most Highly Correlated Regressors are sqft\_living. Followed by the “No. of Bathrooms” and “Bedrooms”. We can see that the other variables have a positive weak correlation between them.

#3. Draw a matrix scatter diagram between the variables of interest using ggplots and the pair function.   
  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.3

library(GGally)

## Warning: package 'GGally' was built under R version 4.1.3

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(psych)

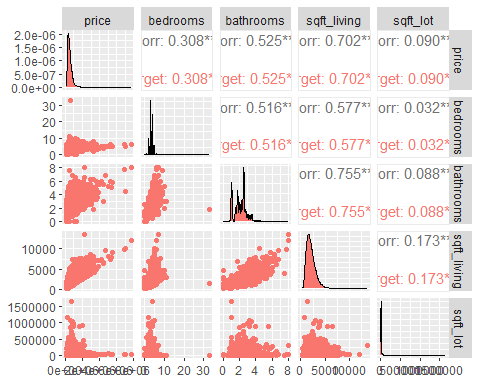
## Warning: package 'psych' was built under R version 4.1.3

##   
## Attaching package: 'psych'

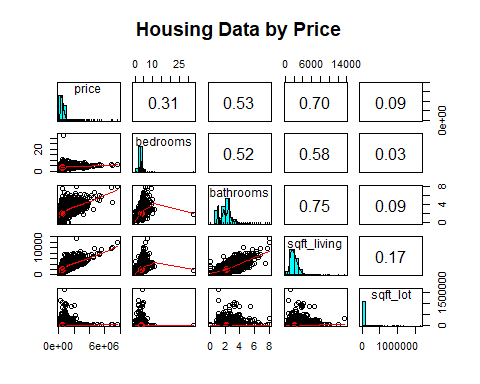
## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

ggpairs(house,aes(colour='target'),pch=21,main="Housing Data by Price")

## Warning in warn\_if\_args\_exist(list(...)): Extra arguments: "pch", "main" are  
## being ignored. If these are meant to be aesthetics, submit them using the  
## 'mapping' variable within ggpairs with ggplot2::aes or ggplot2::aes\_string.



library(ggplot2)  
library(GGally)  
library(psych)  
  
pairs.panels(house,pch=21,main="Housing Data by Price")

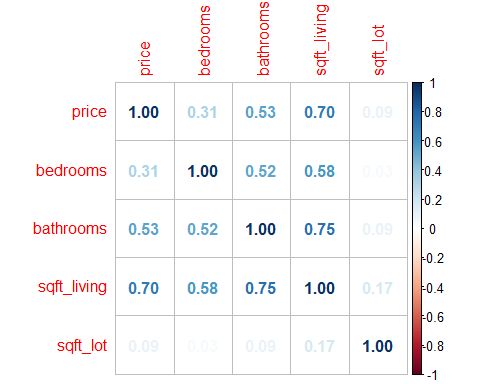


#4. Are the regressors independent of each other? Justify your answer with the help of Q3.  
library(corrplot )

## Warning: package 'corrplot' was built under R version 4.1.3

## corrplot 0.92 loaded

cor1=round(cor(house),2)  
corrplot(cor1,method = "number")



**Answer:** All the regressors have some positive correlation between them. The correlation between “Bedrooms” and “sqft\_lot” is very close to 0 and therefore they have a very weak positive correlation. But none of them are independent of each other.

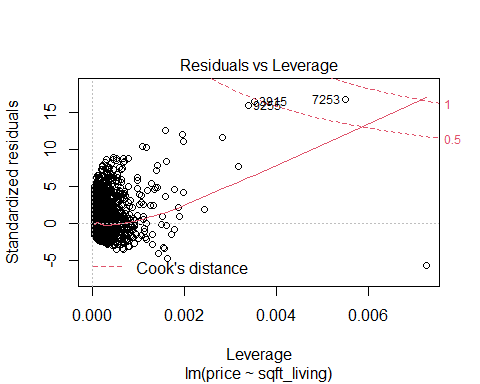
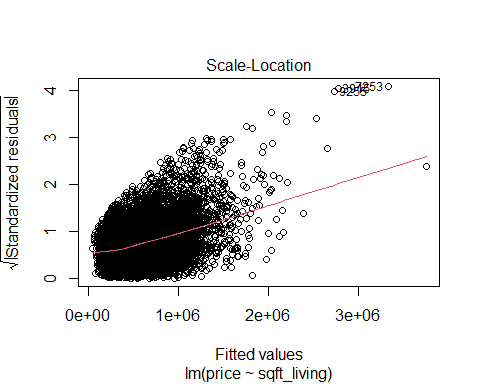
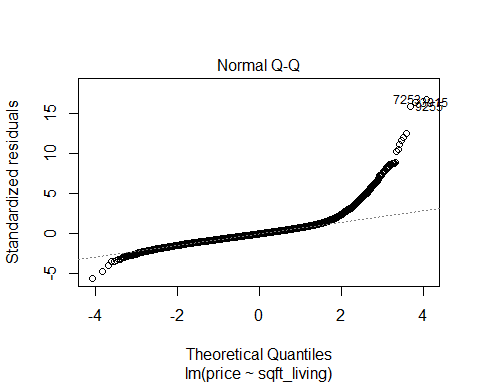
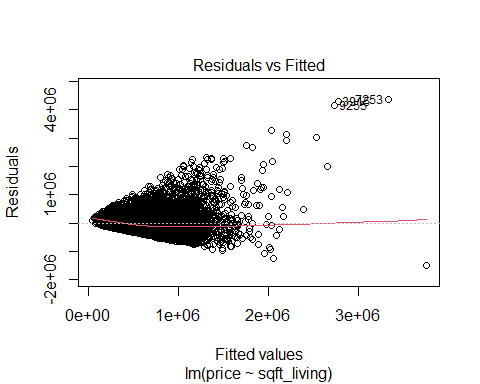
#5. Fit a multiple linear regression model to the selected response (y) and regressor variables (highly correlated with 'y' only) and interpret the estimated coefficients.  
  
library(mctest)  
  
#Multiple Linear Model for Highly Correlated Variables (Only sqft\_living is highly correlated with Price)  
mlr=lm(price~sqft\_living,house)  
mlr

##   
## Call:  
## lm(formula = price ~ sqft\_living, data = house)  
##   
## Coefficients:  
## (Intercept) sqft\_living   
## -43867.6 280.8

summary(mlr)

##   
## Call:  
## lm(formula = price ~ sqft\_living, data = house)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1478255 -147564 -24169 106283 4360147   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -43867.602 4405.455 -9.958 <2e-16 \*\*\*  
## sqft\_living 280.807 1.938 144.924 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 261600 on 21611 degrees of freedom  
## Multiple R-squared: 0.4929, Adjusted R-squared: 0.4928   
## F-statistic: 2.1e+04 on 1 and 21611 DF, p-value: < 2.2e-16

plot(mlr)



**Interpretation:** Using the Coefficients we obtain the equation:

**Y= -43867.6 +280.8\*(sqft\_living)**

#Multiple Linear Model for All Variables  
mlr1=lm(price~.,house)  
mlr1

##   
## Call:  
## lm(formula = price ~ ., data = house)  
##   
## Coefficients:  
## (Intercept) bedrooms bathrooms sqft\_living sqft\_lot   
## 78910.978 -59453.896 6263.206 314.508 -0.378

#Individual Multi Collinearity Diagnostics  
#To Check VIF values since the regressors are highly intercorrelated in some cases.  
#if VIF <5 we can accept the model  
imcdiag(mlr1)

##   
## Call:  
## imcdiag(mod = mlr1)  
##   
## All Individual Multicollinearity Diagnostics Result  
##   
## VIF TOL Wi Fi Leamer CVIF Klein IND1  
## bedrooms 1.5418 0.6486 3902.3876 5853.8523 0.8054 5.9061 0 1e-04  
## bathrooms 2.3838 0.4195 9967.7745 14952.3537 0.6477 9.1318 1 1e-04  
## sqft\_living 2.6995 0.3704 12241.5809 18363.2211 0.6086 10.3410 1 1e-04  
## sqft\_lot 1.0412 0.9604 297.0982 445.6679 0.9800 3.9887 0 1e-04  
## IND2  
## bedrooms 0.8779  
## bathrooms 1.4503  
## sqft\_living 1.5728  
## sqft\_lot 0.0990  
##   
## 1 --> COLLINEARITY is detected by the test   
## 0 --> COLLINEARITY is not detected by the test  
##   
## bathrooms , coefficient(s) are non-significant may be due to multicollinearity  
##   
## R-square of y on all x: 0.5087   
##   
## \* use method argument to check which regressors may be the reason of collinearity  
## ===================================

**Interpretation:** This diagnostic tell us that bathrooms and sqft\_living are correlated. Also VIF values are less than 5 for all independent variables so we can consider all the terms for the model, even if the correlation coefficient is high for the independent variables.

*#6. Construct the 95% and 99% confidence intervals for the individual parameters in the model.*

confint(mlr,'sqft\_living',level=0.95)

## 2.5 % 97.5 %  
## sqft\_living 277.0088 284.6046

**Interpretation:** The 95% confidence interval for the regression coefficient (sqft\_living) is **[277.0088, 284.6046].**

confint(mlr,'sqft\_living',level=0.99)

## 0.5 % 99.5 %  
## sqft\_living 275.8153 285.7981

**Interpretation:** The 99% confidence interval for the regression coefficient ( sqft\_living) is **[275.8153, 285.7981]**

* **CONCLUSION**

Hence in this lab session we found the multiple regression model for the Housing Dataset. We found the Correlation matrix and the Matrix Scatter Plot to understand the correlation between all the variables in the model. We found that none of the variables had a correlation of 0 and hence all had multicorrelation between them.

We found the linear model for the dependent variable and the highly correlated variables i.e sqft\_living.

We found 2 independent variables had a correlation of 0.75 which is a high correlation, and hence we found the VIF(Variance Inflation Factor) , in which all values were less than 5 and hence acceptable for the model.

Finally, we found the 95% and 99% confidence intervals for the individual parameters.