REGRESSION MODEL TO PREDICT SALES PRICE OF HOUSE BASED ON THE BUILDING CHARACTERSTICS OF THE HOUSE

**INTRODUCTION**

***Objective:*** Here our main objective is to perform a complete regression analysis for the given dataset of sales price and building characterstics of 24 houses i.e. we want to build regression model that gives the relationship between sales price of houses and the building characterstics of houses that helps in predicting the sales price of the house based on it’s building characterstics and check the presence of multicolinearity. Further we also want to perform residual analysis of the fitted model and validate all the assumptions associated with the model.

***Data Description:*** The dataset that is to be analyzed consists 24 observations about the sales price and various building characterstics of a house. Our dataset consists of 10 variables which are described below,

***y,*** denotes the ***sales price of house*** in thousand of dollars.

***x1***, denotes ***taxes(local, school, county)***in thousand of dollars.

***x2***, denotes ***number of bathrooms.***

***x3*,** denotes ***lot size*** (in thousand of square fit).

***x4***, denotes ***living space*** (in thousands of square feet).

***x5***, denotes ***number of garage stalls.***

***x6***, denotes ***number of rooms.***

***x7***, denotes ***number of bedrooms.***

***x8***, denotes ***age of the home*** (in years).

***x9***, denotes ***number of fireplaces****.*

Here, ***y is the response variable*** and ***x1,x2,…,x10 are regressors***.

*#Reading HouseSale dataset we are interested in.*  
**library**(readxl)  
HouseSale <- **read\_excel**("HouseSale.xlsx")  
  
*#Obtaining the first few records of the our 'HouseSale' dataset.*  
**head**(HouseSale)

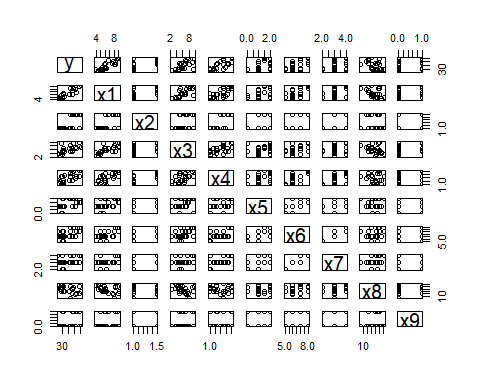
## # A tibble: 6 x 10  
## y x1 x2 x3 x4 x5 x6 x7 x8 x9  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 25.9 4.92 1 3.47 0.998 1 7 4 42 0  
## 2 29.5 5.02 1 3.53 1.5 2 7 4 62 0  
## 3 27.9 4.54 1 2.28 1.18 1 6 3 40 0  
## 4 25.9 4.56 1 4.05 1.23 1 6 3 54 0  
## 5 29.9 5.06 1 4.46 1.12 1 6 3 42 0  
## 6 29.9 3.89 1 4.46 0.988 1 6 3 56 0

**ANALYSIS**

*#We use attach function to access the variables present in the dataframe without calling the dataframe.*  
**attach**(HouseSale)

Step 1 - Firstly, we want to obtain a matrix of scatter plot between all the variables and check if they are linearly related.

*#Here we are plotting a matrix of scatter diagram between our variables of interest check the presence of linear relationship between variables.*  
**pairs**(HouseSale)

From figure 1 we observe that the dependent variable y i.e. sales price of house is linearly positively related with all the independent variables xi (i=1,2,…,10) i.e. building characterstics of house except the one i.e. age of the house (in years), the sales price of the house is linearly negatively related.

But still we conclude that the dependent variable that is sales price of house and all the regressors that is building characterstics of house are linearly related.

Step 2 - Next, we want to obtain the matrix of coefficient of correlation and check the presence of multicolinearity.

*#To obtain the coefficient of correlation matrix of regressors to check if there exist multicolinearity.*  
**round**(**cor**(HouseSale),2)

## y x1 x2 x3 x4 x5 x6 x7 x8 x9  
## y 1.00 0.87 0.71 0.65 0.71 0.46 0.53 0.28 -0.40 0.27  
## x1 0.87 1.00 0.65 0.69 0.73 0.46 0.64 0.37 -0.44 0.15  
## x2 0.71 0.65 1.00 0.41 0.73 0.22 0.51 0.43 -0.10 0.20  
## x3 0.65 0.69 0.41 1.00 0.57 0.20 0.39 0.15 -0.35 0.31  
## x4 0.71 0.73 0.73 0.57 1.00 0.36 0.68 0.57 -0.14 0.11  
## x5 0.46 0.46 0.22 0.20 0.36 1.00 0.59 0.54 -0.02 0.10  
## x6 0.53 0.64 0.51 0.39 0.68 0.59 1.00 0.87 0.12 0.22  
## x7 0.28 0.37 0.43 0.15 0.57 0.54 0.87 1.00 0.31 0.00  
## x8 -0.40 -0.44 -0.10 -0.35 -0.14 -0.02 0.12 0.31 1.00 0.23  
## x9 0.27 0.15 0.20 0.31 0.11 0.10 0.22 0.00 0.23 1.00

Interpretation: From the above matrix of coefficient of correlation we observe that sales price of house(y) and taxes(in thousands of dollars)(x1), sales price of house(y) and number of bathrooms(x2), sales price of house(y) and lot size(in thousands of square feet)(x3), sales price of house(y) and living space(in thousands of square feet)(x4) are highly positively related whereas sales price of house(y) and number of garage stalls(x5),sales price of house(y) and number of rooms(x6),sales price of house(y) and number of bedrooms(x7),sales price of house(y) and number of fireplaces(x9) are slightly positively related.It is also observed that sales price of house(y) and age of the house(in years)(x8) is slightly negatively related.

Now on observing the relationship among regressors we observe that taxes(in thousands of dollars)(x1) and number of bathroom(x2),taxes(in thousands of dollars)(x1) and lot size(in thousand square feet)(x3), taxes(in thousands of dollars)(x1) and living space(in thousand square feet)(x4),taxes(in thousands of dollars)(x1) and number of garage stalls(x5),taxes(in thousands of dollars)(x1) and number of rooms (x6),taxes(in thousands of dollars)(x1) and number of bedrooms(x7),taxes(in thousands of dollars)(x1) and number of fireplaces(x9), number of bathroom(x2) and lot size(in thousand square feet)(x3),number of bathroom(x2) and living space(in thousand square feet)(x4) ,number of bathroom(x2) and number of garage stalls(x5),number of bathroom(x2) and number of rooms (x6),number of bathroom(x2) and number of bedrooms(x7),number of bathroom(x2) and number of fireplaces(x9), lot size(in thousand square feet)(x3) and living space(in thousand square feet)(x4) ,lot size(in thousand square feet)(x3) and number of garage stalls(x5) ,lot size(in thousand square feet)(x3) and number of rooms (x6) ,lot size(in thousand square feet)(x3) and number of bedrooms(x7) ,lot size(in thousand square feet)(x3) and number of fireplaces(x9) ,living space(in thousand square feet)(x4) and number of garage stalls(x5),living space(in thousand square feet)(x4) and number of rooms (x6) ,living space(in thousand square feet)(x4) and number of bedrooms(x7),living space(in thousand square feet)(x4) and number of fireplaces(x9), number of garage stalls(x5) and number of rooms (x6) ,number of garage stalls(x5) and number of bedrooms(x7) ,number of garage stalls(x5) and number of fireplaces(x9) ,number of rooms (x6) and number of bedrooms(x7),number of rooms (x6) and age of home(in years)(x8),number of rooms (x6) and number of fireplaces(x9), number of bedrooms(x7) and age of home(in years)(x8), age of home(in years)(x8) and number of fireplaces(x9) are positively related to each other

we also observe that the variables taxes(in thousands of dollars)(x1) and age of home(in years)(x8),number of bathroom(x2) and age of home(in years)(x8),lot size(in thousand square feet)(x3) and age of home(in years)(x8), living space(in thousand square feet)(x4) and age of home(in years)(x8),number of garage stalls(x5) and age of home(in years)(x8) are negatively related to each other.

Since from the coefficient of correlation matrix we observe that the regressors are slightly positively or negatively related i.e. the coefficient of correlation is not very high for most of the pairs hence we cannot claim the presence of multicolinearity.

Step 3 - Now since we want to fit the best multiple linear regression model to predict the sales price we need to select the most significant variables among all the regressors i.e. the variables having non zero regression coefficient and we do it with the help of stepwise variable selection method.

*#Fitting a regression model initially with only intercept term.*  
fit\_start=**lm**(y**~**1,data=HouseSale)  
  
*#Now we fit a regression model to predict sales with all the 9 regressors.*  
fit\_all=**lm**(y**~**.,data=HouseSale)  
  
*#Below, we fit best multiple linear regression model with the help of stepwise elimination method.*   
stepwise\_fit=**step**(fit\_start,direction="both", scope=**formula**(fit\_all))

## Start: AIC=87.08  
## y ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + x1 1 635.04 196.47 54.459  
## + x2 1 418.90 412.61 72.267  
## + x4 1 416.53 414.98 72.404  
## + x3 1 348.76 482.75 76.035  
## + x6 1 232.20 599.31 81.225  
## + x5 1 177.07 654.44 83.337  
## + x8 1 131.32 700.19 84.959  
## <none> 831.51 87.085  
## + x7 1 65.90 765.61 87.103  
## + x9 1 59.22 772.29 87.311  
##   
## Step: AIC=54.46  
## y ~ x1  
##   
## Df Sum of Sq RSS AIC  
## + x2 1 28.56 167.91 52.689  
## + x9 1 16.35 180.12 54.374  
## <none> 196.47 54.459  
## + x4 1 7.88 188.59 55.477  
## + x5 1 3.88 192.58 55.980  
## + x3 1 3.25 193.21 56.058  
## + x7 1 1.48 194.98 56.277  
## + x6 1 1.39 195.08 56.288  
## + x8 1 0.24 196.22 56.429  
## - x1 1 635.04 831.51 87.085  
##   
## Step: AIC=52.69  
## y ~ x1 + x2  
##   
## Df Sum of Sq RSS AIC  
## <none> 167.91 52.689  
## + x9 1 10.921 156.99 53.075  
## + x7 1 7.478 160.43 53.596  
## + x5 1 6.645 161.27 53.720  
## + x3 1 4.655 163.26 54.015  
## + x6 1 4.241 163.67 54.075  
## + x8 1 4.034 163.88 54.106  
## - x2 1 28.556 196.47 54.459  
## + x4 1 0.057 167.85 54.681  
## - x1 1 244.697 412.61 72.267

In order obtain the best regression model(model with the most significant regressors) we use the stepwise variable selection procedure. Initially we observe that the AIC value is 87.08. We proceed to select the variables in the following steps:

step 1 - Initially, a model with only the intercept term is built and we observe that the regressor taxes(in thousands of dollars) has least AIC value that is 54.459 which is also less than previous AIC i.e.87.08 therefore we add the variable taxes(in thousands of dollars) into our model. Now the AIC value for the model with taxes(in thousands of dollars) as the regressor is 54.459.

step 2 - Now in the next step we observe that the variable number of bathrooms is having least AIC value among all the other variables i.e. 52.689 which is also less than previous AIC i.e. 54.459. Therefore now we select the variable number of bathrooms as regressor and include it into our model.Finally we stop at this stage because now there are no more variable whose AIC value is less than the previous AIC value that is 52.689.

Hence by the stepwise selection method the variables that are selected as regressor into our model are taxes(in thousand of dollars) and bathroom. Hence all these variable will be having a significant effect on sales price of house. Finally the model with least AIC value is selected which is our best multiple linear regression model and is given by,

***y=B0+(B1\*x1)+(B2\*x2)***

i.e. ***Sales=B0+(B1\*taxes)+(B2\*no. of bathrooms)***

Hence the best multiple linear regression model for the given dataset is obtained.

*#Next, we are obtaining the summary of the best fitted model.*   
**summary**(stepwise\_fit)

##   
## Call:  
## lm(formula = y ~ x1 + x2, data = HouseSale)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.7644 -1.9510 -0.2057 1.8040 5.4172   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.1120 2.9961 3.375 0.00286 \*\*   
## x1 2.7170 0.4911 5.532 1.73e-05 \*\*\*  
## x2 6.0985 3.2271 1.890 0.07267 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.828 on 21 degrees of freedom  
## Multiple R-squared: 0.7981, Adjusted R-squared: 0.7788   
## F-statistic: 41.5 on 2 and 21 DF, p-value: 5.067e-08

From the above summary we observe that

Our best fitted regression model is,

***y=10.1120+(2.7170x1)+( 6.0985x2)***

***i.e. Sales=10.1120+(2.7170taxes)+(2.7170no. of bathrooms)***

Since the adjusted coefficient of determination is 0.7788 > 0.5 hence our fitted regression model is of good quality.

Also since the p value is negligible it can be concluded that the performance of the fitted regression model is good.

Step 4 - Here, in this step we want to check the presence of multicolinearity between regressors. We do it by obtaining the variance inflation factors corresponding to each regressor of the fitted model.

*#Loading the package 'car' required for obtaining the variance inflation factor.*  
**library**(car)

## Warning: package 'car' was built under R version 4.0.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.0.3

*#Obtaining the variance inflation factor for all the regressors of the fitted regression model.*  
**vif**(stepwise\_fit)

## x1 x2   
## 1.736559 1.736559

From the above table we observe that the all the regressors have variance inflation factor between 1 to 5 which indicates the presence of moderate correlation between regressors.Hence, the multicolinearity is not present between the regrssors.

Step 5 - Now, since we have the best regression model, also there’s no multicolinearity between regressors now we check the goodness of fit by validating the assumptions associated with the fitted model.

The following are the assumptions regarding the fitted model,

1. The relationship between y and x1,x2,…x10 is linear.
2. Errors have zero mean.
3. Assumption of homoscedasticity, i.e.the errors have constant variance.
4. Errors are uncorrelated.
5. Errors are normally distributed random variables.

Note: Here, in this problem we’ll be validating the assumptions for the residuals since we have been provided with the sample of house sales price and their building characterstics.

1. To check the relationship between y and x1,x2,…x10 is linear.

* From the scatter plot matrix obtained above we can conclude that the relationship between y and x1,x2,…x10 is linear.

1. To check if the errors have mean zero.

*#Obtaining the residual terms of the fitted model.*  
residuals=**resid**(stepwise\_fit)  
  
*#Obtaining the mean of the residuals.*  
**mean**(residuals)

## [1] -1.249001e-16

From the above calculation we observe that mean of residuals is 0 hence it can be concluded that the errors has mean zero.

1. To check assumption of homscedasticity i.e. variance of residuals is constant.

*#Loading the package 'Hmisc' required to obtain the plot.*  
**library**(Hmisc)

## Warning: package 'Hmisc' was built under R version 4.0.3

## Loading required package: lattice

## Warning: package 'lattice' was built under R version 4.0.3

## Loading required package: survival

## Warning: package 'survival' was built under R version 4.0.3

## Loading required package: Formula

## Warning: package 'Formula' was built under R version 4.0.3

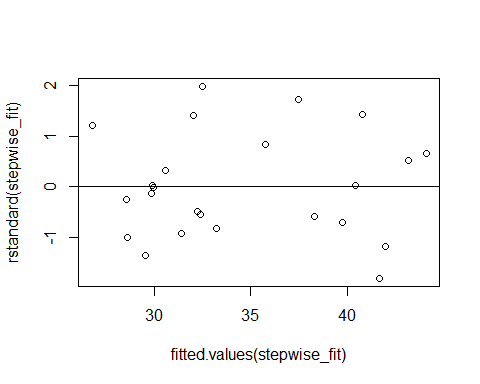
## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.0.3

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

*#Plot of fitted values against residuals.*  
fitted\_values<-**fitted.values**(stepwise\_fit)  
**plot**(**fitted.values**(stepwise\_fit),**rstandard**(stepwise\_fit))  
**abline**(0,0)

From the figure 2 i.e. plot of fitted values against residuals we observe that all the points are evenly spread around the horizontal line which means the errors have constant variance.

*#Loading the package 'lmtest' required to perform studentized Breusch-Pagan test.*  
**library**(lmtest)

## Warning: package 'lmtest' was built under R version 4.0.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.0.3

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

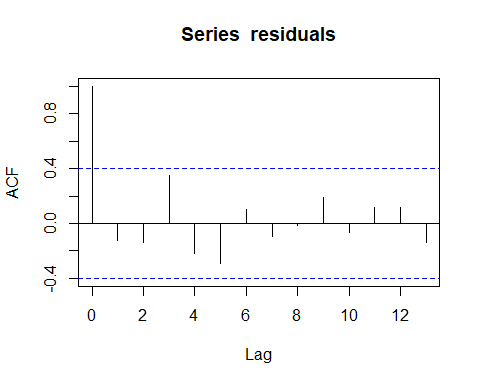
*#Performing studentized Breusch-Pagan test to check the assumption of homoscedasticy.*  
**bptest**(stepwise\_fit)

##   
## studentized Breusch-Pagan test  
##   
## data: stepwise\_fit  
## BP = 0.23315, df = 2, p-value = 0.89

Here, we observe that the p value obtained is 0.89 which is greater than 0.05 hence we fail to reject null hypothesis and conclude that errors have constant variance.

1. To validate the assumption that the errors are uncorrelated or not.

*#Obtaining the acf plot to check if the residuals uncorrelated i.e. to check if there is no autocorrelation in our residual series.*  
**acf**(residuals)





From figure 3 we observe that all the lags are inside the threshhold line hence we conclude that the residuals are not autocorrelated.

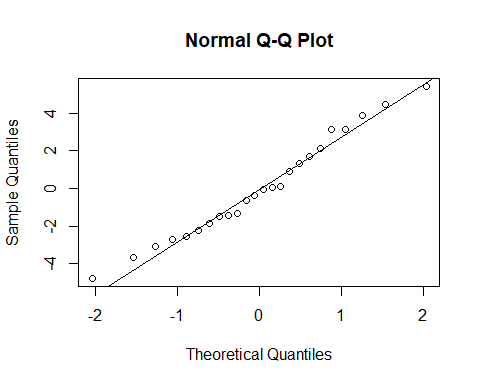
*#Performing Durbin-Watson test to check if the residuals are uncorrelated.*  
**dwtest**(stepwise\_fit)

##   
## Durbin-Watson test  
##   
## data: stepwise\_fit  
## DW = 2.1484, p-value = 0.577  
## alternative hypothesis: true autocorrelation is greater than 0

From Durbin-Watson test we observe that p value=0.577 which is greater than 0.05 therefore we fail to reject the null hypothesis and conclude that the residuals not autocorrelated which means errors are uncorrelated.

1. To check normality assumption i.e to check if the residuals are normally distributed.

*#Obtaining the qqplot to check if the residuals are normally dstributed or not.*  
**qqnorm**(residuals)  
**qqline**(residuals)

From the figure 4 i.e. qqplot we observe that the points on the graph almost form a straight line hence we conclude that errors are normally distributed.

*#Now, performing Shapiro-Wilk normality test to check if the residuals are normaally distributed or not.*  
**shapiro.test**(residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: residuals  
## W = 0.97571, p-value = 0.8058

On performing Shapiro-Wilk normality test we observe that the p value obtained is 0.8058 which is greater than 0.05 hence we fail to reject the null hypothesis and conclude that the errors are normally distributed random variables.

Since all the assumptions made about model are validated to be true we conclude that our model is a good fit.

**CONCLUSION:**

From the above analysis we come to the following conclusion,

On using the stepwise variable selection method we obtained the best multiple linear regression model (the model that consists of most significant variables among all the independent variables) of our data set, which is given by,

**y=10.1120+(2.7170*x1)+( 6.0985*x2)**

**i.e. Sales=10.1120+(2.7170*taxes)+(6.0985*no. of bathrooms)**

Hence, from the above regression model we conclude that the taxes(in thousands of dollars) and number of bathrooms are having a significant effect on the sales price of the house.

From the fitted regression model we observe that the intercept B0 = 10.1120 which indicates that the sales price of house is 10.1120(in thousands of dollars) when there is no additional tax and there are no bathrooms in the house.

We also observe that the coefficient of taxes i.e. B1 in the fitted regression model is 2.7170 which indicates that when the number of bathrooms in the house is kept constant then for one unit of change in tax the sales price of house increases by 2.7170(in thousands of dollars) amount.

Also the coefficient of number of bathroom i.e. B2 in the fitted regression model is 6.0985 which indicates that when the tax is kept constant then for increase in one bathroom the sales price of house increases by 6.0985(in thousands of dollars) amount.

Since the adjusted coefficient of determination is 0.7788 , hence we conclude that 77.88% of total variation of sales price of house is explained by the taxes(in thousands of dollar) and number of bathrooms in the house. Also since adjusted coefficient of determination is 0.7788 which is greater than 0.5 which means that our fitted regression model is of good quality.

We can also say that since the p value is negligible for this dataset we can conclude that the performance of regression model is good.

From the above analysis we also observed that the variance inflation factor for each of the regressors in our fitted regression model is between 1 and 5 which indicates that absence of multicolinearity between the regressors.

Further on doing the residual analysis we observe that all our assumptions about the fitted regression model are validated and are satisfied and hence we conclude that the fitted multiple regression model obtained is a good fit.