SELECTION OF VARIABLES USING FORWARD SELECTION CRITERIA TO MODEL STORE SALES

**INTRODUCTION:**

***Objective:*** Here, our main objective is to build a model that can predict sales of store based on variables i.e. advertising costs, store size in square feet, % employee retention, customer satisfaction score, whether a promotion was run or not. We want to fit the best multiple linear regression model using the forward variable selection method.

***Data Description:*** We have collected data from 41 stores on advertising costs, store size in square feet, % employee retention, customer satisfaction score, whether a promotion was run or not and sales.

Variable of interests:

**Dependent variable**, is the variable for which we want to build a model i.e. the variable which we want to predict is **sales.**

**Independent variables**, are the variables that effects sales and they **are advertising costs, store size in square feet, % employee retention, customer satisfaction score, whether a promotion was run or not.**

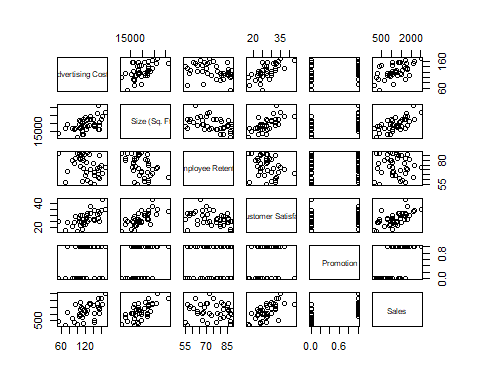
*#Reading Store dataset we are interested in.*  
**library**(readxl)  
Store\_Data <- **read\_excel**("Store Data.xlsx")  
  
*#Obtaining the first few records of the our dataset.*  
**head**(Store\_Data)

## # A tibble: 6 x 6  
## `Advertising Co~ `        Size (~ `          Empl~ `      Customer~  
## <dbl> <dbl> <dbl> <dbl>  
## 1 124. 22560 59 32  
## 2 154 31181 62 33  
## 3 124. 16314 78 28  
## 4 163 24205 66 43  
## 5 107 17574 82 22  
## 6 144. 19584 67 34  
## # ... with 2 more variables: ` Promotion` <dbl>, Sales <dbl>

*#Attaching the dataset to R so that we can access the variables present in the data framework without calling the data frame.*  
**attach**(Store\_Data)

**ANALYSIS:**

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

Interpretation: From the above scatter matrix i.e. figure 1 we observe that there exists a positive linear realtionship between advertising costs and sales,size(sq. ft) and sales,customer satisfaction and sales,promotion and sales but it is also observed that there exists negative linear relationship between Employee Retention and sales.

It is also observed from the scatter plot matrix that Advertising Costs and Size (Sq. Ft), Advertising Costs and Customer Satisfaction ,Advertising Costs and Promotion, Size (Sq. Ft) and Customer Satisfaction, Size (Sq. Ft) and Promotion, Employee Retention (%) and Promotion, Customer Satisfaction and Promotion possess positive linear relationship which means as the one variable increases there is corresponding increase in another variable. Also we observe that Advertising Costs and Employee Retention (%),Size (Sq. Ft) and Employee Retention (%),Customer Satisfaction and Employee Retention (%) have linearly negative relationship which means as the one variable increases there is corresponding decrease in another variable. Now since we observed that there also exist some kind of relationship between regressors therefore it indicates the presence of multicolinearity.

*#Loading the package 'Hmisc'.*  
**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

*#Now we are obtaining the matrix of coefficient of correlation with rounded to two decimal points.*  
**round**(**cor**(Store\_Data),2)

## Advertising Costs         Size (Sq. Ft)  
## Advertising Costs 1.00 0.51  
##         Size (Sq. Ft) 0.51 1.00  
##           Employee Retention (%) -0.32 -0.43  
##       Customer Satisfaction 0.66 0.67  
##       Promotion 0.18 0.46  
## Sales 0.58 0.67  
##           Employee Retention (%)  
## Advertising Costs -0.32  
##         Size (Sq. Ft) -0.43  
##           Employee Retention (%) 1.00  
##       Customer Satisfaction -0.35  
##       Promotion 0.09  
## Sales -0.08  
##       Customer Satisfaction       Promotion  
## Advertising Costs 0.66 0.18  
##         Size (Sq. Ft) 0.67 0.46  
##           Employee Retention (%) -0.35 0.09  
##       Customer Satisfaction 1.00 0.13  
##       Promotion 0.13 1.00  
## Sales 0.63 0.63  
## Sales  
## Advertising Costs 0.58  
##         Size (Sq. Ft) 0.67  
##           Employee Retention (%) -0.08  
##       Customer Satisfaction 0.63  
##       Promotion 0.63  
## Sales 1.00

Now we will check if there actually exists the multicolinearity between the variables.

From the above coefficient of correlation matrix we observe that there exists a slightly positive realtionship between advertising costs and sales,size(sq. ft) and sales,customer satisfaction and sales,promotion and sales but it is also observed that there exists slightly negative relationship between Employee Retention and sales.

It is also observed from the coefficient of correlation matrix that **Advertising Costs and Size (Sq. Ft)** has a slightly positive relationship with coefficient of correlation value **0.51**, similarly the other pairs of regressors i.e. **Advertising Costs and Customer Satisfaction** with correlation coefficient **0.66**, **Advertising Costs and Promotion** with correlation coefficient **0.18**, **Size (Sq. Ft) and Customer Satisfaction** with correlation coefficient **0.67**, **Size (Sq. Ft) and Promotion** with correlation coefficient **0.46**, **Employee Retention (%) and Promotion** with correlation coefficient **0.09**, **Customer Satisfaction and Promotion** with correlation coefficient **0.13**, possess slightly positive relationship .Also we observe that **Advertising Costs and Employee Retention (%)** with correlation coefficient **-0.32** ,**Size (Sq. Ft) and Employee Retention (%)** with correlation coefficient **-0.43**,Customer **Satisfaction and Employee Retention (%)** with correlation coefficient **-0.35** , have slightly negative relationship .

Hence from the above analysis we observe that we don’t have multicolinearity in this case between regressors because for multicolinearity to exist the correlation coefficient should be greater than 0.7 between the regressors which is not the case here.

Here our main aim is to select the variables that have significant effect on the model using forward selection method and fit a best multiple linear regression model and we proceed for the same by following these steps.

*#Firstly we are fitting the initial regression model with only the intercept term.*  
reg\_fit1=**lm**(Sales**~**1,data=Store\_Data)  
reg\_fit1

##   
## Call:  
## lm(formula = Sales ~ 1, data = Store\_Data)  
##   
## Coefficients:  
## (Intercept)   
## 1210

*#Obtaining the summary of regression model fitted initially with only the intercept term for the dataset.*  
**summary**(reg\_fit1)

##   
## Call:  
## lm(formula = Sales ~ 1, data = Store\_Data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1063.98 -310.98 -12.98 449.02 1298.02   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1209.98 88.45 13.68 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 566.3 on 40 degrees of freedom

*#Fitting a regression model to predict sales with all the 5 regressors.*  
fit\_all=**lm**(Sales**~**.,data=Store\_Data)

*#Now since we want to fit the best multiple linear regression model to predict the sales we need to select the significant variables i.e. the variables having non zero regression coefficient and we do it with the help of forward variable selection method.*  
  
*#Now we are choosing a best multiple linear regression model by AIC(Akaike information criterion) in a stepwise procedure*  
frwd\_fit=**step**(reg\_fit1,direction = "forward",scope=**formula**(fit\_all))

## Start: AIC=520.8  
## Sales ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + `        Size (Sq. Ft)` 1 5737977 7091222 498.49  
## + `      Promotion` 1 5169224 7659975 501.66  
## + `      Customer Satisfaction` 1 5121575 7707624 501.91  
## + `Advertising Costs` 1 4385811 8443388 505.65  
## <none> 12829199 520.80  
## + `          Employee Retention (%)` 1 86968 12742231 522.52  
##   
## Step: AIC=498.49  
## Sales ~ `        Size (Sq. Ft)`  
##   
## Df Sum of Sq RSS AIC  
## + `      Promotion` 1 1729572 5361650 489.03  
## + `Advertising Costs` 1 1010282 6080940 494.19  
## + `      Customer Satisfaction` 1 786890 6304331 495.67  
## + `          Employee Retention (%)` 1 673165 6418057 496.40  
## <none> 7091222 498.49  
##   
## Step: AIC=489.03  
## Sales ~ `        Size (Sq. Ft)` + `      Promotion`  
##   
## Df Sum of Sq RSS AIC  
## + `      Customer Satisfaction` 1 1678748 3682903 475.63  
## + `Advertising Costs` 1 1246351 4115299 480.18  
## <none> 5361650 489.03  
## + `          Employee Retention (%)` 1 138181 5223469 489.96  
##   
## Step: AIC=475.63  
## Sales ~ `        Size (Sq. Ft)` + `      Promotion` + `      Customer Satisfaction`  
##   
## Df Sum of Sq RSS AIC  
## + `Advertising Costs` 1 308276 3374626 474.05  
## <none> 3682903 475.63  
## + `          Employee Retention (%)` 1 136615 3546288 476.08  
##   
## Step: AIC=474.05  
## Sales ~ `        Size (Sq. Ft)` + `      Promotion` + `      Customer Satisfaction` +   
## `Advertising Costs`  
##   
## Df Sum of Sq RSS AIC  
## + `          Employee Retention (%)` 1 189767 3184859 473.67  
## <none> 3374626 474.05  
##   
## Step: AIC=473.67  
## Sales ~ `        Size (Sq. Ft)` + `      Promotion` + `      Customer Satisfaction` +   
## `Advertising Costs` + `          Employee Retention (%)`

**summary**(frwd\_fit)

##   
## Call:  
## lm(formula = Sales ~ `        Size (Sq. Ft)` + `      Promotion` +   
## `      Customer Satisfaction` + `Advertising Costs` + `          Employee Retention (%)`,   
## data = Store\_Data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -752.58 -78.54 33.32 165.38 560.34   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.762e+03 6.311e+02 -2.792 0.008432 \*\*   
## `        Size (Sq. Ft)` 2.122e-02 2.052e-02 1.034 0.308304   
## `      Promotion` 5.208e+02 1.187e+02 4.386 0.000101 \*\*\*  
## `      Customer Satisfaction` 4.000e+01 1.446e+01 2.766 0.009005 \*\*   
## `Advertising Costs` 4.751e+00 2.384e+00 1.993 0.054107 .   
## `          Employee Retention (%)` 8.087e+00 5.600e+00 1.444 0.157602   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 301.7 on 35 degrees of freedom  
## Multiple R-squared: 0.7517, Adjusted R-squared: 0.7163   
## F-statistic: 21.2 on 5 and 35 DF, p-value: 1.052e-09

**INTERPRETATION:**

Step 1 - From the above analysis, we observe in the forward selection method that initially the AIC value is 520.8. Then a model with only the intercept term is built and we observe that the regressor Size( Sq. Ft) has least AIC value that is 498.49 which is also less than previous AIC i.e. 520.8 therefore we add the variable Size (Sq. Ft) into our model. Now the AIC value for the model with Size (Sq. Ft) as the regressor is 498.49.

Step 2 - Now in the next step we observe that Promotion is having least AIC value among all the other variables i.e. 489.03 which is also less than previous AIC i.e. 498.49. Therefore now we select the variable Promotion as regressor and include it into our model.

Step 3 - In the next step we proceeded in the similar manner and we observed that the variable customer satisfaction is having least AIC value i.e. 475.63 which is less than previous AIC i.e. 489.03. Therefore now we select the variable customer satisfaction as regressor and include it into our model.

Step 4 - In the next step we observe that the variable Advertising Costs is having least AIC value i.e. 474.05 which is less than previous AIC i.e. 475.63. Therefore now we select the variable Advertising Costs as regressor and include it into our model.

Step 5 - In the next step we observe that the variable Employee Retention (%) is having least AIC value i.e. 473.67which is less than previous AIC i.e. 474.05. Therefore now we select the variable Employee Retention (%) as regressor and include it into our model.

Hence by the forward selection method the variables that are selected as regressor into our model are Size( Sq. Ft), Promotion, customer satisfaction, Advertising Costs, Employee Retention (%). Hence all these variable will be having a significant effect on sales. Finally the model with least AIC value is selected which is our best multiple linear regression model and is given by,

Sales=B0+(B1\*Size*)+(B2\**Promotion)+(B3\*customer satisfaction*)+(B4\**Advertising Costs)+(B5\*Employee Retention)

i.e. Sales=-1.762e+03+(2.122e-02\*Size*)+(5.208e+02\**Promotion)+(4.000e+\*customer satisfaction*)+( 4.751e+00\**Advertising Costs)+(8.087e+00\*Employee Retention)

From the fitted model, we observe that the intercept Beta0=-1.762e+03, which basically gives the average value of response variables when the regressor variables are zero but here size of the store is one of the regressor which cannot be 0 hence here the intercept does not have any physical meaning.

We also observe that the coefficient of Size of the store (Sq. Ft) i.e. Beta1 is 2.122e-02 which means that when the promotion, customer satisfaction, advertising costs, employee retention (%) is kept constant then for one unit of change in store size (Sq. Ft), sales changes 2.122e-02 times of store size (Sq. Ft).

We also observe that the coefficient of promotion of the store i.e. Beta2 is 5.208e+02 which means that when the store size, customer satisfaction, advertising costs, employee retention (%) is kept constant then for one unit of change in promotion, sales changes 5.208e+02 times of promotion.

We also observe that the coefficient of customer satisfaction i.e. Beta3 is 4.000e+01 which means that when the promotion, store size, advertising costs, employee retention (%) is kept constant then for one unit of change in customer satisfaction, sales changes 4.000e+01 times of customer satisfaction.

We also observe that the coefficient of Advertising Costs i.e. Beta4 is 4.751e+00 which means that when the promotion, customer satisfaction, store size, employee retention (%) is kept constant then for one unit of change in Advertising Costs, sales changes 4.751e+00 times of Advertising Costs.

We also observe that the coefficient of Employee Retention (%) i.e. Beta5 is 8.087e+00 which means that when the promotion, customer satisfaction, advertising costs, store size is kept constant then for one unit of change in Employee Retention (%), sales changes 8.087e+00 times of Employee Retention (%).

We also observe that since the p value for the model is negligible therefore we conclude that the performance of the regression model is good.

Hence the best multiple linear regressiom model is obtained.