Lab 6: Model Selection and Variable Selection

Jeevan Koshy - 1740256

August 28, 2019

Aim

To establish a best multiple regression model to predict the sales using forward, backward and stepwise procedure.

H0: The difference in advertising costs, size (in sq. ft), employee retention percentage, customer satisfaction and promotion have no significant effect on the sales of different stores.

H1: The difference in advertising costs, size (in sq. ft), employee retention percentage, customer satisfaction and promotion have a significant effect on the sales of different stores.

Procedure

library(readxl)

## Warning: package 'readxl' was built under R version 3.5.3

fdata<- read\_excel("d:/Downloads/data\_6.xlsx")  
View(fdata)

fullmodel=lm(Sales~., data=fdata)  
formula(fullmodel)

## Sales ~ Store + `Advertising Costs` + `Size (Sq. Ft)` + `% Employee Retention` +   
## `Customer Satisfaction` + Promotion

summary(fullmodel)

##   
## Call:  
## lm(formula = Sales ~ ., data = fdata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -752.50 -91.56 16.19 181.44 549.65   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.769e+03 6.395e+02 -2.766 0.00911 \*\*   
## Store 1.439e+00 4.180e+00 0.344 0.73286   
## `Advertising Costs` 4.580e+00 2.465e+00 1.858 0.07187 .   
## `Size (Sq. Ft)` 2.065e-02 2.085e-02 0.990 0.32907   
## `% Employee Retention` 7.805e+00 5.731e+00 1.362 0.18217   
## `Customer Satisfaction` 4.103e+01 1.495e+01 2.745 0.00961 \*\*   
## Promotion 5.235e+02 1.205e+02 4.344 0.00012 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 305.5 on 34 degrees of freedom  
## Multiple R-squared: 0.7526, Adjusted R-squared: 0.709   
## F-statistic: 17.24 on 6 and 34 DF, p-value: 4.892e-09

Forward Elimination Method

f1=lm(Sales~1, data=fdata)  
summary(f1)

##   
## Call:  
## lm(formula = Sales ~ 1, data = fdata)  
##   
## 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

step(f1, direction="forward", scope=formula(fullmodel))

## 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  
## + Store 1 1604 12827595 522.80  
##   
## 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  
## + Store 1 24801 7066421 500.35  
##   
## 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  
## + Store 1 13876 5347775 490.92  
##   
## 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  
## + Store 1 71008 3611894 476.83  
##   
## 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  
## + Store 1 27669 3346957 475.71  
##   
## Step: AIC=473.67  
## Sales ~ `Size (Sq. Ft)` + Promotion + `Customer Satisfaction` +   
## `Advertising Costs` + `% Employee Retention`  
##   
## Df Sum of Sq RSS AIC  
## <none> 3184859 473.67  
## + Store 1 11055 3173804 475.53

##   
## Call:  
## lm(formula = Sales ~ `Size (Sq. Ft)` + Promotion + `Customer Satisfaction` +   
## `Advertising Costs` + `% Employee Retention`, data = fdata)  
##   
## Coefficients:  
## (Intercept) `Size (Sq. Ft)` Promotion   
## -1.762e+03 2.122e-02 5.208e+02   
## `Customer Satisfaction` `Advertising Costs` `% Employee Retention`   
## 4.000e+01 4.751e+00 8.087e+00

cm=lm(Sales~fdata$`Advertising Costs`+fdata$`Size (Sq. Ft)`+fdata$`Customer Satisfaction`+fdata$Promotion+fdata$`% Employee Retention`, data=fdata)  
summary(cm)

##   
## Call:  
## lm(formula = Sales ~ fdata$`Advertising Costs` + fdata$`Size (Sq. Ft)` +   
## fdata$`Customer Satisfaction` + fdata$Promotion + fdata$`% Employee Retention`,   
## data = fdata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -752.58 -78.54 33.32 165.38 560.34   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.762e+03 6.311e+02 -2.792 0.008432 \*\*   
## fdata$`Advertising Costs` 4.751e+00 2.384e+00 1.993 0.054107 .   
## fdata$`Size (Sq. Ft)` 2.122e-02 2.052e-02 1.034 0.308304   
## fdata$`Customer Satisfaction` 4.000e+01 1.446e+01 2.766 0.009005 \*\*   
## fdata$Promotion 5.208e+02 1.187e+02 4.386 0.000101 \*\*\*  
## fdata$`% 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

From the above obtained results, it is observed that the p-value is less than 0.05. Hence, we reject the null hypothesis. Therefore, it can be concluded that the difference in advertising costs, size (in sq. ft), employee retention percentage, customer satisfaction and promotion have a significant effect on the sales of different stores.

Furthermore,

1. The intercept, i.e., the average value of sales = -1.762e+03
2. Each unit of advertising cost adds 4.751e+00 when adjusted for all the other regressors.
3. Each unit of store size adds 2.122e-02 when adjusted for all the other regressors.
4. Each unit of customer satisfaction adds 4.000e+01 when adjusted for all the other regressors.
5. Each unit of promotion adds 5.208e+02 when adjusted for all the other regressors.
6. Each unit of employee retention adds 8.087e+00 when adjusted for all the other regressors.

Backward Elimination Method

step(fullmodel, direction="backward")

## Start: AIC=475.53  
## Sales ~ Store + `Advertising Costs` + `Size (Sq. Ft)` + `% Employee Retention` +   
## `Customer Satisfaction` + Promotion  
##   
## Df Sum of Sq RSS AIC  
## - Store 1 11055 3184859 473.67  
## - `Size (Sq. Ft)` 1 91525 3265330 474.70  
## <none> 3173804 475.53  
## - `% Employee Retention` 1 173153 3346957 475.71  
## - `Advertising Costs` 1 322186 3495991 477.50  
## - `Customer Satisfaction` 1 703122 3876927 481.74  
## - Promotion 1 1761184 4934988 491.63  
##   
## Step: AIC=473.67  
## Sales ~ `Advertising Costs` + `Size (Sq. Ft)` + `% Employee Retention` +   
## `Customer Satisfaction` + Promotion  
##   
## Df Sum of Sq RSS AIC  
## - `Size (Sq. Ft)` 1 97258 3282117 472.91  
## <none> 3184859 473.67  
## - `% Employee Retention` 1 189767 3374626 474.05  
## - `Advertising Costs` 1 361429 3546288 476.08  
## - `Customer Satisfaction` 1 696073 3880932 479.78  
## - Promotion 1 1750531 4935390 489.63  
##   
## Step: AIC=472.91  
## Sales ~ `Advertising Costs` + `% Employee Retention` + `Customer Satisfaction` +   
## Promotion  
##   
## Df Sum of Sq RSS AIC  
## - `% Employee Retention` 1 113860 3395978 472.31  
## <none> 3282117 472.91  
## - `Advertising Costs` 1 372946 3655063 475.32  
## - `Customer Satisfaction` 1 1405913 4688030 485.52  
## - Promotion 1 3362431 6644548 499.83  
##   
## Step: AIC=472.31  
## Sales ~ `Advertising Costs` + `Customer Satisfaction` + Promotion  
##   
## Df Sum of Sq RSS AIC  
## <none> 3395978 472.31  
## - `Advertising Costs` 1 324381 3720358 474.05  
## - `Customer Satisfaction` 1 1300060 4696037 483.59  
## - Promotion 1 3660523 7056501 500.29

##   
## Call:  
## lm(formula = Sales ~ `Advertising Costs` + `Customer Satisfaction` +   
## Promotion, data = fdata)  
##   
## Coefficients:  
## (Intercept) `Advertising Costs` `Customer Satisfaction`   
## -899.824 4.456 45.130   
## Promotion   
## 608.785

be=lm(Sales~fdata$`Advertising Costs`+fdata$`Customer Satisfaction`+fdata$Promotion, data=fdata)  
summary(be)

##   
## Call:  
## lm(formula = Sales ~ fdata$`Advertising Costs` + fdata$`Customer Satisfaction` +   
## fdata$Promotion, data = fdata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -754.81 -126.44 -8.95 222.87 495.28   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -899.824 262.839 -3.423 0.001525 \*\*   
## fdata$`Advertising Costs` 4.456 2.370 1.880 0.068007 .   
## fdata$`Customer Satisfaction` 45.130 11.991 3.764 0.000581 \*\*\*  
## fdata$Promotion 608.785 96.399 6.315 2.35e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 303 on 37 degrees of freedom  
## Multiple R-squared: 0.7353, Adjusted R-squared: 0.7138   
## F-statistic: 34.26 on 3 and 37 DF, p-value: 8.975e-11

Interpretation

From the above obtained results, it is observed that the p-value is less than 0.05. Hence, we reject the null hypothesis. Therefore, it can be concluded that the difference in advertising costs, size (in sq. ft), employee retention percentage, customer satisfaction and promotion have a significant effect on the sales of different stores.

Furthermore,

1. The intercept, i.e., the average value of sales = -899.824
2. Each unit of advertising cost adds 4.456 when adjusted for all the other regressors.
3. Each unit of customer satisfaction adds 45.130 when adjusted for all the other regressors.
4. Each unit of promotion adds 608.785 when adjusted for all the other regressors.

Stepwise Elimination Method

step(fullmodel, direction="both")

## Start: AIC=475.53  
## Sales ~ Store + `Advertising Costs` + `Size (Sq. Ft)` + `% Employee Retention` +   
## `Customer Satisfaction` + Promotion  
##   
## Df Sum of Sq RSS AIC  
## - Store 1 11055 3184859 473.67  
## - `Size (Sq. Ft)` 1 91525 3265330 474.70  
## <none> 3173804 475.53  
## - `% Employee Retention` 1 173153 3346957 475.71  
## - `Advertising Costs` 1 322186 3495991 477.50  
## - `Customer Satisfaction` 1 703122 3876927 481.74  
## - Promotion 1 1761184 4934988 491.63  
##   
## Step: AIC=473.67  
## Sales ~ `Advertising Costs` + `Size (Sq. Ft)` + `% Employee Retention` +   
## `Customer Satisfaction` + Promotion  
##   
## Df Sum of Sq RSS AIC  
## - `Size (Sq. Ft)` 1 97258 3282117 472.91  
## <none> 3184859 473.67  
## - `% Employee Retention` 1 189767 3374626 474.05  
## + Store 1 11055 3173804 475.53  
## - `Advertising Costs` 1 361429 3546288 476.08  
## - `Customer Satisfaction` 1 696073 3880932 479.78  
## - Promotion 1 1750531 4935390 489.63  
##   
## Step: AIC=472.91  
## Sales ~ `Advertising Costs` + `% Employee Retention` + `Customer Satisfaction` +   
## Promotion  
##   
## Df Sum of Sq RSS AIC  
## - `% Employee Retention` 1 113860 3395978 472.31  
## <none> 3282117 472.91  
## + `Size (Sq. Ft)` 1 97258 3184859 473.67  
## + Store 1 16788 3265330 474.70  
## - `Advertising Costs` 1 372946 3655063 475.32  
## - `Customer Satisfaction` 1 1405913 4688030 485.52  
## - Promotion 1 3362431 6644548 499.83  
##   
## Step: AIC=472.31  
## Sales ~ `Advertising Costs` + `Customer Satisfaction` + Promotion  
##   
## Df Sum of Sq RSS AIC  
## <none> 3395978 472.31  
## + `% Employee Retention` 1 113860 3282117 472.91  
## + Store 1 28743 3367235 473.96  
## - `Advertising Costs` 1 324381 3720358 474.05  
## + `Size (Sq. Ft)` 1 21351 3374626 474.05  
## - `Customer Satisfaction` 1 1300060 4696037 483.59  
## - Promotion 1 3660523 7056501 500.29

##   
## Call:  
## lm(formula = Sales ~ `Advertising Costs` + `Customer Satisfaction` +   
## Promotion, data = fdata)  
##   
## Coefficients:  
## (Intercept) `Advertising Costs` `Customer Satisfaction`   
## -899.824 4.456 45.130   
## Promotion   
## 608.785

Conclusion

A multiple regression model has been established to predict the sales using forward, backward and stepwise procedure. And it can be observed that the difference in advertising costs, size (in sq. ft), employee retention percentage, customer satisfaction and promotion have a significant effect on the sales of different stores.