LAB 6 SALES OF A COMPANY

A.Mayuri(2348133)

Lab 6
Sales Of a Company

A.Mayuri(2348133)

2023-12-15

Case situation:

You are a VP of sales and have responsibility for 41 stores. You have collected data from the stores on advertising costs, store size in square feet, % employee retention, customer satisfaction score, whether a promotion was run or not, and sales. You want to build a model that can predict sales based on these five variables.

Fit a best multiple linear regression model to predict the sales using the forward selection backward elimination procedure. Prepare a report based on the above questions with introduction, analysis, and conclusions.

Step1: Import dataset

```
library(readxl)
vps <- read_excel("C:/Users/mayur/Desktop/Mstat/Semesters/Tri-sem2/Regression
/Dataset/vps.xlsx")
View(vps)
attach(vps)
```

Forward selection

Step2:

Fit a regression only with the constant term y=b0

```
fitstart_1=lm(sales~1,data=vps)
fitstart_1
```

```
##
## Call:
## lm(formula = sales ~ 1, data = vps)
##
## Coefficients:
```

```
## (Intercept)
##        1210
```

```
summary(fitstart_1)
```

```
##
## Call:
## lm(formula = sales ~ 1, data = vps)
##
## 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
```

Interpretation: 1210 is the intercept

Step3:forward procedure

```
fitall=lm(sales~.,data=vps)

fwd=step(fitstart_1,direction = "forward",scope=formula(fitall))
```

```
## Start:  AIC=520.8
## sales ~ 1
##
##               Df Sum of Sq      RSS    AIC
## + size         1   5737977  7091222 498.49
## + pro          1   5169224  7659975 501.66
## + cust_sat     1   5121575  7707624 501.91
## + Adv_cost     1   4385811  8443388 505.65
## <none>                     12829199 520.80
## + `%_emp_ret`  1     86968 12742231 522.52
## + store        1      1604 12827595 522.80
##
## Step:  AIC=498.49
## sales ~ size
##
##               Df Sum of Sq     RSS    AIC
## + pro          1   1729572 5361650 489.03
## + Adv_cost     1   1010282 6080940 494.19
## + cust_sat     1    786890 6304331 495.67
## + `%_emp_ret`  1    673165 6418057 496.40
## <none>                     7091222 498.49
## + store        1     24801 7066421 500.35
## 
## Step:  AIC=489.03
## sales ~ size + pro
## 
##                Df Sum of Sq     RSS    AIC
## + cust_sat     1   1678748 3682903 475.63
## + Adv_cost     1   1246351 4115299 480.18
## <none>                     5361650 489.03
## + `%_emp_ret`  1    138181 5223469 489.96
## + store        1     13876 5347775 490.92
## 
## Step:  AIC=475.63
## sales ~ size + pro + cust_sat
## 
##                Df Sum of Sq     RSS    AIC
## + Adv_cost     1    308276 3374626 474.05
## <none>                     3682903 475.63
## + `%_emp_ret`  1    136615 3546288 476.08
## + store        1     71008 3611894 476.83
## 
## Step:  AIC=474.05
## sales ~ size + pro + cust_sat + Adv_cost
## 
##                Df Sum of Sq     RSS    AIC
## + `%_emp_ret`  1    189767 3184859 473.67
## <none>                     3374626 474.05
## + store        1     27669 3346957 475.71
## 
## Step:  AIC=473.67
## sales ~ size + pro + cust_sat + Adv_cost + `%_emp_ret`
## 
##         Df Sum of Sq     RSS    AIC
## <none>               3184859 473.67
## + store  1     11055 3173804 475.53
```

```
fwd
```

```
## 
## Call:
```

```
## lm(formula = sales ~ size + pro + cust_sat + Adv_cost + `%_emp_ret`,
##     data = vps)
##
## Coefficients:
## (Intercept)         size          pro     cust_sat     Adv_cost  `%_emp_re
t`
##  -1.762e+03    2.122e-02    5.208e+02    4.000e+01    4.751e+00    8.087e+
00
```

note: smaller AIC explains better about the variability, hence the appropriate model suggested according to forward selection is a regression between dependent variable as sales and regressors as size, promotion, customer satisfaction, advertisement cost and employment retention.

Step4: Fitting and significance of variable.

```
summary(fwd)
```

```
##
## Call:
## lm(formula = sales ~ size + pro + cust_sat + Adv_cost + `%_emp_ret`,
##     data = vps)
##
## 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         2.122e-02  2.052e-02   1.034 0.308304
## pro          5.208e+02  1.187e+02   4.386 0.000101 ***
## cust_sat     4.000e+01  1.446e+01   2.766 0.009005 **
## Adv_cost     4.751e+00  2.384e+00   1.993 0.054107 .
## `%_emp_ret`  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
```

```
library(car)
```

```
## Loading required package: carData
```

```
vif(fwd)

##        size         pro    cust_sat    Adv_cost `%_emp_ret`
##    2.966729    1.579684    2.579197    1.819381    1.431279
```

here we observe that the significant variables at a 0.1 level of significance(for convenience) are promotion,customer satisifaction,advertisement cost. also all the variables do not have multicollilinearity since vif<5.

Step5: Choosing the model

```
fit1=lm(sales~pro+cust_sat+Adv_cost,data=vps)
fit1

## 
## Call:
## lm(formula = sales ~ pro + cust_sat + Adv_cost, data = vps)
## 
## Coefficients:
## (Intercept)          pro     cust_sat     Adv_cost  
##    -899.824      608.785       45.130        4.456

summary(fit1)

## 
## Call:
## lm(formula = sales ~ pro + cust_sat + Adv_cost, data = vps)
## 
## 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 ** 
## pro          608.785     96.399   6.315 2.35e-07 ***
## cust_sat      45.130     11.991   3.764 0.000581 ***
## Adv_cost       4.456      2.370   1.880 0.068007 .  
## ---
## 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
```

here we observe that all the variables are significant.the adjusted R^2 0.7138 > 0.5 thus the model is significant. hence the model is y=-899.82 + 608 * x1 + 45.13 *x2 + 4.456* x3 + e where y = sales x1 = promotion x2 = customer satisifaction x3 = advertisement cost

Backward Selection:

```
fitall=lm(sales~.,data=vps)

bwd2=step(fitall,direction = "backward",scope=formula(fitall))

## Start:  AIC=475.53
## sales ~ store + Adv_cost + size + `%_emp_ret` + cust_sat + pro
## 
##               Df Sum of Sq     RSS    AIC
## - store        1     11055 3184859 473.67
## - size         1     91525 3265330 474.70
## <none>                     3173804 475.53
## - `%_emp_ret`  1    173153 3346957 475.71
## - Adv_cost     1    322186 3495991 477.50
## - cust_sat     1    703122 3876927 481.74
## - pro          1   1761184 4934988 491.63
## 
## Step:  AIC=473.67
## sales ~ Adv_cost + size + `%_emp_ret` + cust_sat + pro
## 
##               Df Sum of Sq     RSS    AIC
## - size         1     97258 3282117 472.91
## <none>                     3184859 473.67
## - `%_emp_ret`  1    189767 3374626 474.05
## - Adv_cost     1    361429 3546288 476.08
## - cust_sat     1    696073 3880932 479.78
## - pro          1   1750531 4935390 489.63
## 
## Step:  AIC=472.91
## sales ~ Adv_cost + `%_emp_ret` + cust_sat + pro
## 
##               Df Sum of Sq     RSS    AIC
## - `%_emp_ret`  1    113860 3395978 472.31
## <none>                     3282117 472.91
## - Adv_cost     1    372946 3655063 475.32
## - cust_sat     1   1405913 4688030 485.52
## - pro          1   3362431 6644548 499.83
## 
```

```
## Step:  AIC=472.31
## sales ~ Adv_cost + cust_sat + pro
##
##              Df Sum of Sq     RSS    AIC
## <none>                    3395978 472.31
## - Adv_cost  1    324381 3720358 474.05
## - cust_sat  1   1300060 4696037 483.59
## - pro       1   3660523 7056501 500.29
```

```
bwd2
```

```
##
## Call:
## lm(formula = sales ~ Adv_cost + cust_sat + pro, data = vps)
##
## Coefficients:
## (Intercept)     Adv_cost     cust_sat          pro
##    -899.824        4.456       45.130      608.785
```

note: smaller AIC explains better about the variability, hence the appropriate model suggested according to backward selection is a regression between dependent variable as sales and regressors as size,promotion,customer satisfaction, advertisement cost.

```
fit1=lm(sales~pro+cust_sat+Adv_cost,data=vps)
fit1
```

```
##
## Call:
## lm(formula = sales ~ pro + cust_sat + Adv_cost, data = vps)
##
## Coefficients:
## (Intercept)          pro     cust_sat     Adv_cost
##    -899.824      608.785       45.130        4.456
```

```
summary(fit1)
```

```
##
## Call:
## lm(formula = sales ~ pro + cust_sat + Adv_cost, data = vps)
##
## 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 **
## pro          608.785     96.399   6.315 2.35e-07 ***
## cust_sat      45.130     11.991   3.764 0.000581 ***
## Adv_cost       4.456      2.370   1.880 0.068007 .
## ---
## 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
```

Here we observe that all the variables are significant.the adjusted R^2 0.7138 > 0.5 thus the model is significant.

Hence the model is,

y=-899.82 + 608 * x1 + 45.13 *x2 + 4.456* x3 + e

where ,

y = sales x1 = promotion x2 = customer satisifaction x3 = advertisement cost

Validating the Built Model

Multicollinearity

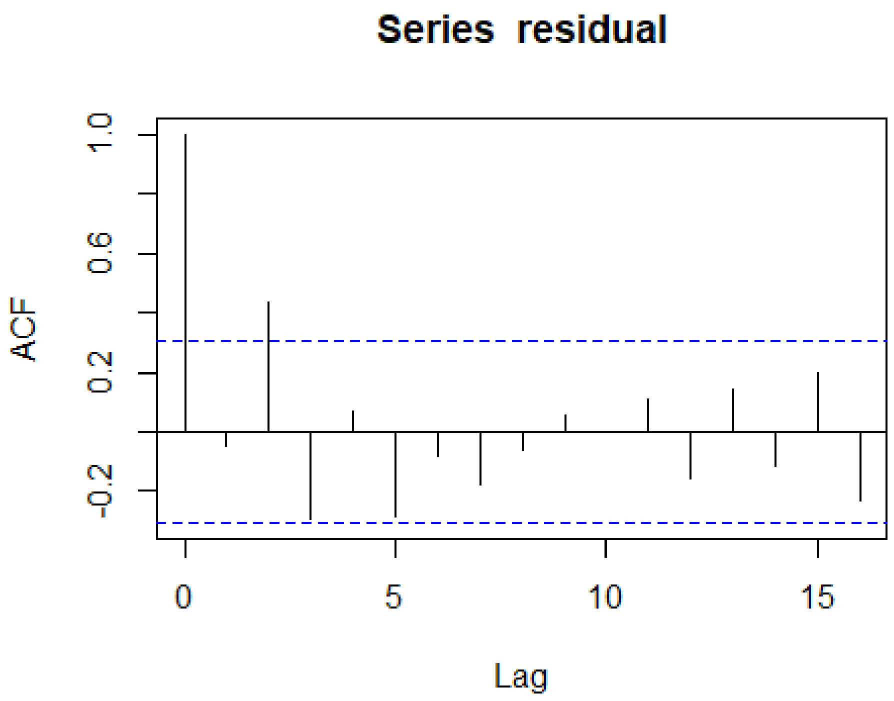
```
library(car)
vif(fit1)

##      pro cust_sat Adv_cost
## 1.032233 1.757675 1.783104
```

Interpretation :

Here we observe that all the vif values are less than 5. thus there exsist no multi-collinearity in the model.

Autocorrelation

```
residual=resid(fit1)
acf(residual)
```



Interpretation and Recommendation:

we observe there exist an auto-correlation in the model. thus, we need to include correction measures such as,

1)removing a few variables and refitting the model.

2)relax the assumptions by using transformation.

3)adopt nonlinear regression modeling