Building a best multiple linear regression model for VP of sales data set.

Nithya S

2023-12-21

Objective

I am a VP of sales and have responsibility for 41 stores.I 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. I want to build a model that can predict sales based on these five variables.

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

URL for the dataset:

https://docs.google.com/spreadsheets/d/1X_s4VEOlm8dQT3LqTFQkotF3QZg-t45/edit?usp=sharing&ouid=110138716074493614410&rtpof=true&sd=true

First, we import the dataset in order to proceed further with the analysis. We import the dataset as follows:

```
library(readxl)
VP_sales <- read_excel("G:/My Drive/Linear Regression/Datasets/VP_sales-dataset.xlsx")
View(VP_sales)
attach(VP_sales)
```

We fit the best regression model by two methods: 1. Forward selection 2. Backward elimination

First we fit the best regression model by using forward selection

The algorithm for Forward Selection is:

1. Choose a significance level.

2. Fit a model with intercept term alone.

3. Construct the individual model for the independent variables with the above constant model. After fitting the model, we check for the least p-value [lesser than the significance level, here it is 15%], or least AIC value or t-value which is larger.

4. Two cases arise here, if there are any variables which are not significant, we ignore them and consider the significant variables.

5. After construction of the model with the significant variable, we fix this model and add other independent variables individually for this fixed model.

6. We see for the least p-value or lesser AIC value or the larger t-value.

7. We continue this process until all the significant variables have been fitted and insignificant variables are removed.

To get the forward selection model, we first fit a constant model. Then we fit a model with all the variables and then fit the best model as follows:

```
fit_start=lm(sales~1,data=VP_sales)
fit_start
```

```
##
## Call:
## lm(formula = sales ~ 1, data = VP_sales)
##
## Coefficients:
## (Intercept)
##        1210
```

```
summary(fit_start)
```

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

```
fit_all=lm(sales~.,data=VP_sales)
fit_all
```

```
##
## Call:
## lm(formula = sales ~ ., data = VP_sales)
##
## Coefficients:
## (Intercept)        store     Adv_cost         size  `%_emp_ret`
```

```
cust_sat
##  -1.769e+03    1.439e+00    4.580e+00    2.065e-02    7.805e+00
4.103e+01
##         pro
##   5.235e+02

forward=step(fit_start,direction="forward",scope=formula(fit_all))

## 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
```

forward

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

```
summary(forward)
```

```
##
## Call:
## lm(formula = sales ~ size + pro + cust_sat + Adv_cost + `%_emp_ret`,
##     data = VP_sales)
##
## 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
```

In the output, the value of AIC corresponds to the model value. As AIC is small, the model is a good fit. The most significant variable is the one with the least AIC value

We infer from the above output as: The final bet fitted model using forward selection at 15% significance level is

Y[sales] $= B_1X_1[size] + B_2X_2[promotion] + B_3X_3[customer\ satisfaction] + B_4X_4[advertising\ cost] + B_5X_5[employment\ retension] + E[error]$

The estimated model is

$Y^{hat} = -1.762e+03 + (2.122e-02)X_1 + (5.208e+02)X_2 + (4.000e+01)X_3 + (4.751e+00)X_4 + (8.087e+00)X_5$

We also see that at 5% significance level X_1[size], X_2[promotion]and X_3[customer satisfaction] are significance as their p-value is lesser than 0.05. At 15% level of significance, all the variables are significant. 71.63% of the data is explained by the independent variables for the response variable.

Now we fit the best regression model by using backward elimination

The algorithm for Backward elimination is:

1. Choose a significance level, here it is 15%.

2. Fit the regression model with all the independent variables. Here it is a full model.

3. Observe the p-value which is the highest

4. If the p-value is greater than 0.15, we remove the variable.

5. if there are no variables with p-value greater than 0.15, we stop the process and say it is the best fitted model. 6. If there are variables with p-value greater than 0.15, refit the model without the regressor.

Here we call for the full model and construct the moddel using backward elimination as follows:

```
fit_all

##
## Call:
## lm(formula = sales ~ ., data = VP_sales)
##
## Coefficients:
## (Intercept)        store     Adv_cost         size  `%_emp_ret`
cust_sat
##  -1.769e+03    1.439e+00    4.580e+00    2.065e-02    7.805e+00
4.103e+01
```

```
##         pro
##   5.235e+02
```

```
backward=step(fit_all,direction='backward')
```

```
## 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
```

```
backward
```

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

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

```
summary(backward)
```

```
##
## Call:
## lm(formula = sales ~ Adv_cost + cust_sat + pro, data = VP_sales)
##
## 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 **
## Adv_cost       4.456      2.370   1.880 0.068007 .
## cust_sat      45.130     11.991   3.764 0.000581 ***
## pro          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
```

We infer from the above output as, The final bet fitted model using backward elimination at 15% significance level is

$Y[sales] = B_0 + B_1X_1[advertising\ cost] + B_2X_2[customer\ satisfaction] + B_3X_3[promotion] + E[error]$

The estimated model is

$$Y^{hat} = -899.824 + 4.456X_1 + 45.130X_2 + 608.785X_3$$

We also see that at 5% significance level X_2[customer satisfaction] and X_3[promotion] are significance as their p-value is lesser than 0.05. At 15% level of significance, all the variables are significant. 71.38% of the data is explained by the independent variables for the response variable.

To select one of the best models obtained from forward selection and backward elimination, since proper selection cannot be made based on p-value, we see for the model which has larger adjusted R^2 value. Here we see that the adjusted R^2 value is greater for the model obtained using forward selection procedure.

So we say that the model obtained using forward selection is best model, and proceed for the further residual analysis. The model is

$$Y[sales] = B_0 + B_1X_1[size] + B_2X_2[promotion] + B_3X_3[customer\ satisfaction] + B_4X_4[advertising\ cost] + B_5X_5[employment\ retension] + E[error]$$

The estimated model is

$$Y^{hat} = -1.762e + 03 + (2.122e - 02)X_1 + (5.208e + 02)X_2 + (4.000e + 01)X_3 + (4.751e + 00)X_4 + (8.087e + 00)X_5$$

The analysis to be done for the model step by step is as follows:

1. Plot the data or find the correlation between the variables, in order to know the kind of relationship between the variables.

2. Fit the model using variable selection procedure [Forward selection, Backward elimination or Step wise selection]

3. Check for multi-collinearity.

4. Residual analysis

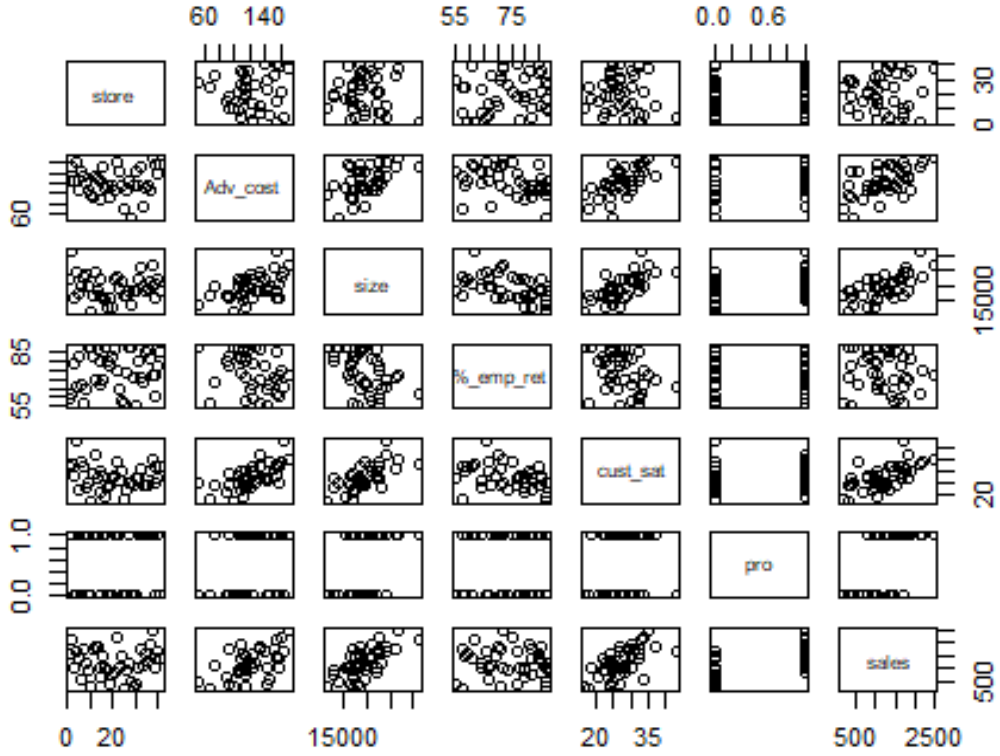
a) Residual v/s fitted values If there exists an influential point, remove that and refit the model and continue the same procedure.

b) Normality test: Shapiro test or QQ plot

c) Autocorrelation: ACF (Autocorrelation function) plot, Durbin Watson test If all these conditions are satisfied, we say that the model is the best multiple linear regression model.

Firstly, we plot the data as follows:

```
pairs(VP_sales[1:7])
```



Since, we find difficulty in finding the relationship of the independent variables with the dependent variable, we go for the correlation function as below:

```
library(Hmisc)

##
## Attaching package: 'Hmisc'

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

rcorr(as.matrix(VP_sales))

##           store Adv_cost  size %_emp_ret cust_sat  pro sales
## store      1.00     0.06 -0.05      0.13    -0.12 0.00  0.01
## Adv_cost   0.06     1.00  0.51     -0.32     0.66 0.18  0.58
## size      -0.05     0.51  1.00     -0.43     0.67 0.46  0.67
## %_emp_ret  0.13    -0.32 -0.43      1.00    -0.35 0.09 -0.08
## cust_sat  -0.12     0.66  0.67     -0.35     1.00 0.13  0.63
## pro        0.00     0.18  0.46      0.09     0.13 1.00  0.63
## sales      0.01     0.58  0.67     -0.08     0.63 0.63  1.00
##
## n= 41
```

```
##
##
## P
##           store  Adv_cost size   %_emp_ret cust_sat pro    sales
## store            0.6962   0.7612 0.4191    0.4557   0.9795 0.9447
## Adv_cost  0.6962          0.0006 0.0442    0.0000   0.2719 0.0000
## size      0.7612 0.0006          0.0048    0.0000   0.0023 0.0000
## %_emp_ret 0.4191 0.0442   0.0048           0.0231   0.5830 0.6088
## cust_sat  0.4557 0.0000   0.0000 0.0231             0.4187 0.0000
## pro       0.9795 0.2719   0.0023 0.5830    0.4187          0.0000
## sales     0.9447 0.0000   0.0000 0.6088    0.0000   0.0000
```

We get the correlation matrix as above.

Secondly, we construct the model using variable selection procedure. Here, we used both Forward selection and Backward elimination procedure and selected model obtained using Forward selection procedure.

Now, we check for multicollinearity as follows:

```
library(car)

## Loading required package: carData

vif(forward)

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

Since all the values are less than 5, we say that there is no multicollinearity in the model. So we proceed further.

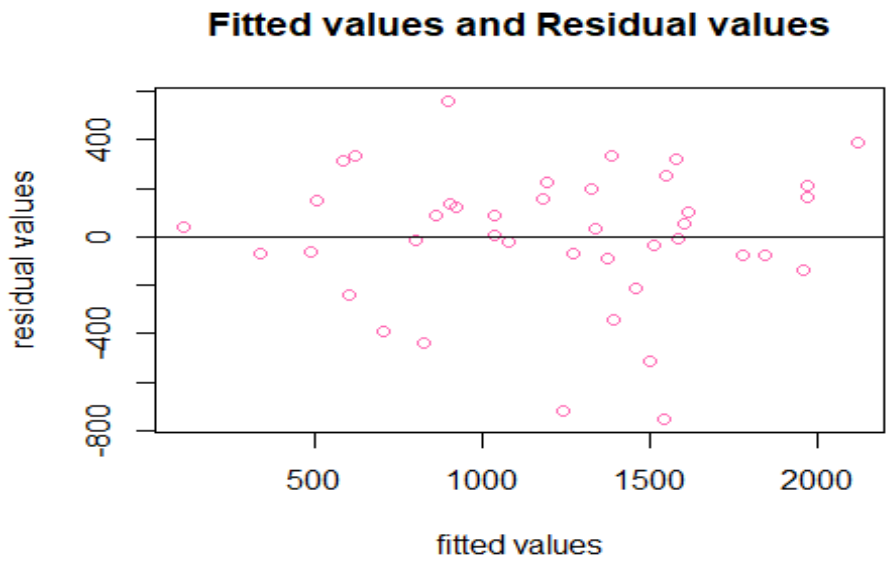
Now, we go for residual analysis. First, we get the residuals of the model and plot the onatained residuals with fitted values in order to understand the linearity, constant variance and outliers of the model.

```
fitted_values=fitted.values(forward)
fitted_values

##         1         2         3         4         5         6         7
8
## 1585.8171 1973.6242  921.8645 1779.9439 1183.3415 1239.2223 1616.4176
1269.7594
##         9        10        11        12        13        14        15
16
##  860.6755  335.8247 1384.8975 1956.9149 1551.6708 1034.2237  503.8085
1514.5878
##        17        18        19        20        21        22        23
24
## 1079.5429  621.4554  602.9353  797.8462 1544.5760 1972.1672 1393.4163
895.6600
##        25        26        27        28        29        30        31
```

```
32
##  103.7892  584.0683 1457.7629  486.8447  705.5299  822.2437 1503.9172
1844.5400
##        33        34        35        36        37        38        39
40
## 1036.4486 1326.1643 1374.2406 1581.2304 2121.7839  903.0605 1194.3280
1606.1790
##        41
## 1336.6758
```

```
plot(fitted_values,resid(forward),col="hotpink",main="Fitted values and
Residual values",xlab="fitted values",ylab="residual values")
abline(0,0)
```



From the plot, we say that, the residuals are in a horizontal band fashion and they fluctuate more or less in a random manner inside, this indicates that there are no visible model defects. So no conclusion can be drawn based on the linearity of the variables and constant variance. However we have shown above that the errors have no constant variance.

Now, to see if there are any outliers or influential points in the model:

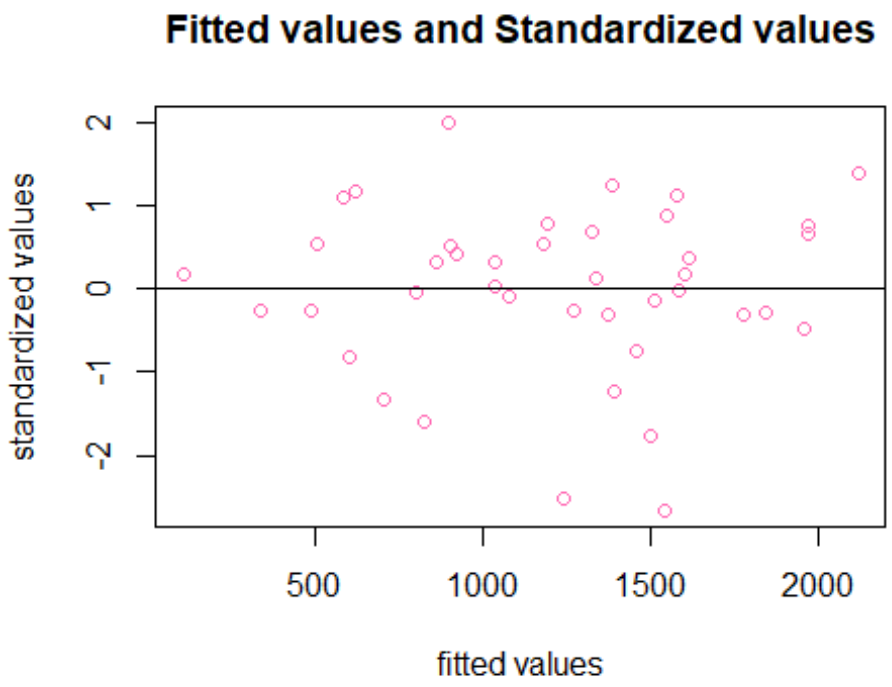
We check the same by using two methods

1. Standardized residuals

2. Studentized residuals

Let us use standardized residuals. The procedure is as follows:

```
plot(fitted_values,rstandard(forward),col="hotpink",main="Fitted values and
Standardized values",xlab="fitted values",ylab="standardized values")
abline(0,0)
```



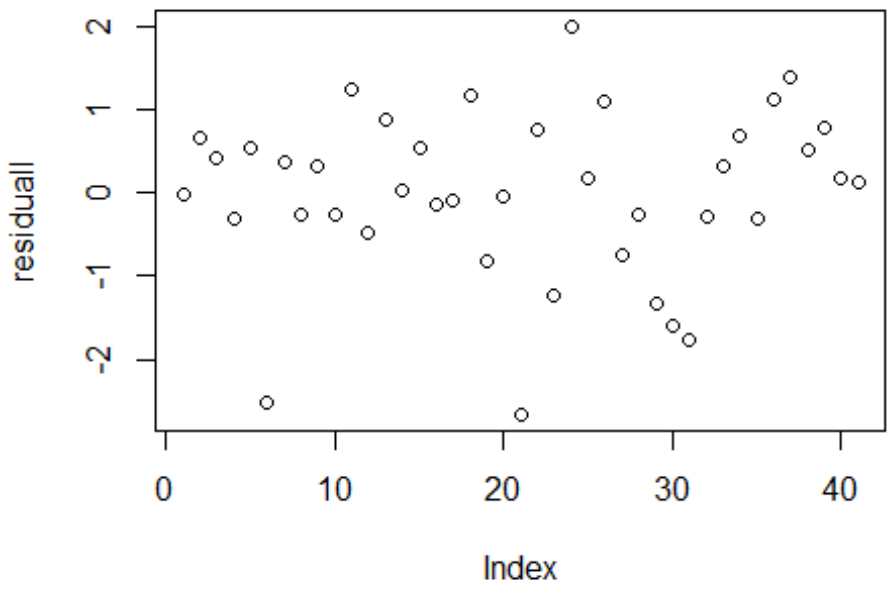
From the above plot we say that no values lie outside the range of modulus 3. So there are no outliers in the dataset.

To be very precise we can even check in the residual values for outliers as follows:

```
residuall=rstandard(forward)
residuall
```

```
##           1           2           3           4           5           6
## -0.01730985  0.65386508  0.41653948 -0.30953821  0.53846604 -2.52661893
##           7           8           9          10          11          12
##  0.37268355 -0.26486342  0.32974147 -0.25390314  1.23629500 -0.49294696
##          13          14          15          16          17          18
##  0.87945665  0.02718511  0.53326631 -0.13221113 -0.08109210  1.16031483
##          19          20          21          22          23          24
## -0.82464008 -0.05296134 -2.66067697  0.74485734 -1.23634746  1.99453947
##          25          26          27          28          29          30
##  0.16860946  1.09077703 -0.74876991 -0.26092847 -1.33689974 -1.60815232
##          31          32          33          34          35          36
## -1.76183857 -0.27732803  0.32099889  0.67821503 -0.32265090  1.10816154
##          37          38          39          40          41
##  1.38615651  0.50963688  0.78377992  0.18272214  0.12234812
```

```
plot(residuall)
```



Now we have numerically seen that there are no outliers in the dataset. Since, there are no outliers in the model, we proceed further.

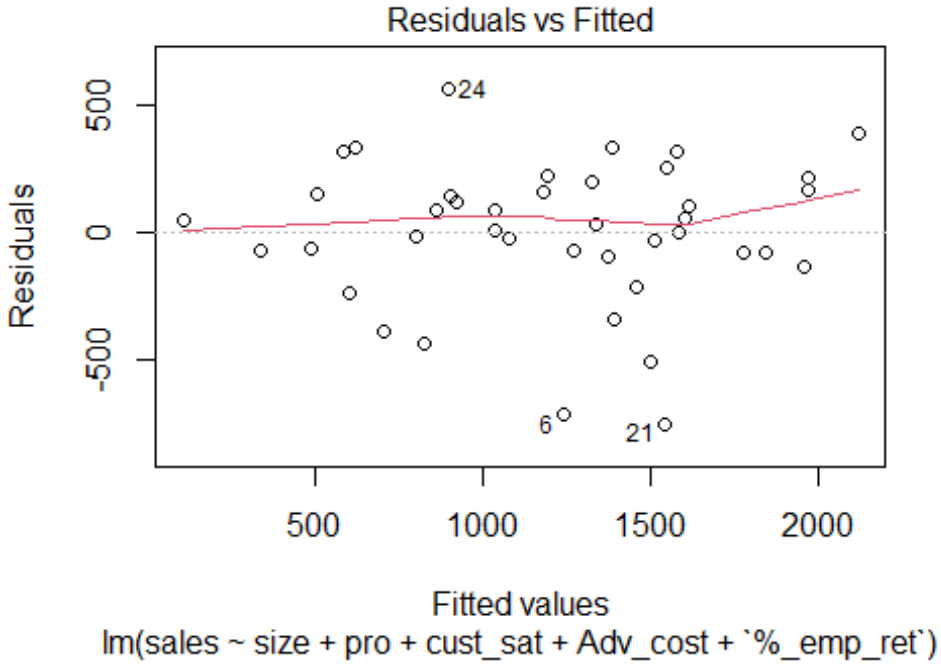
We check if the normality assumption is satisfied by using

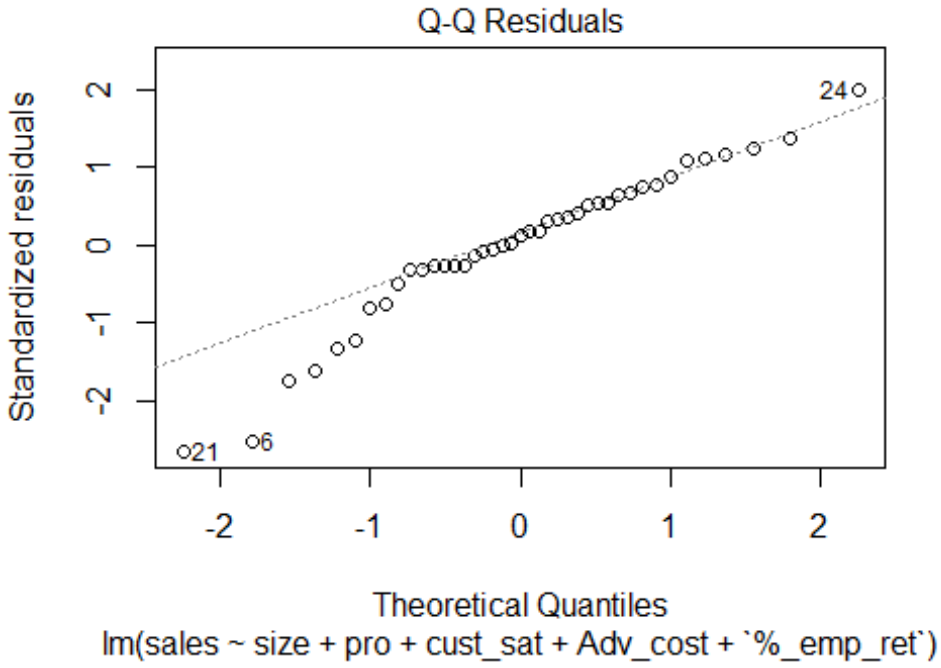
1. QQ plot

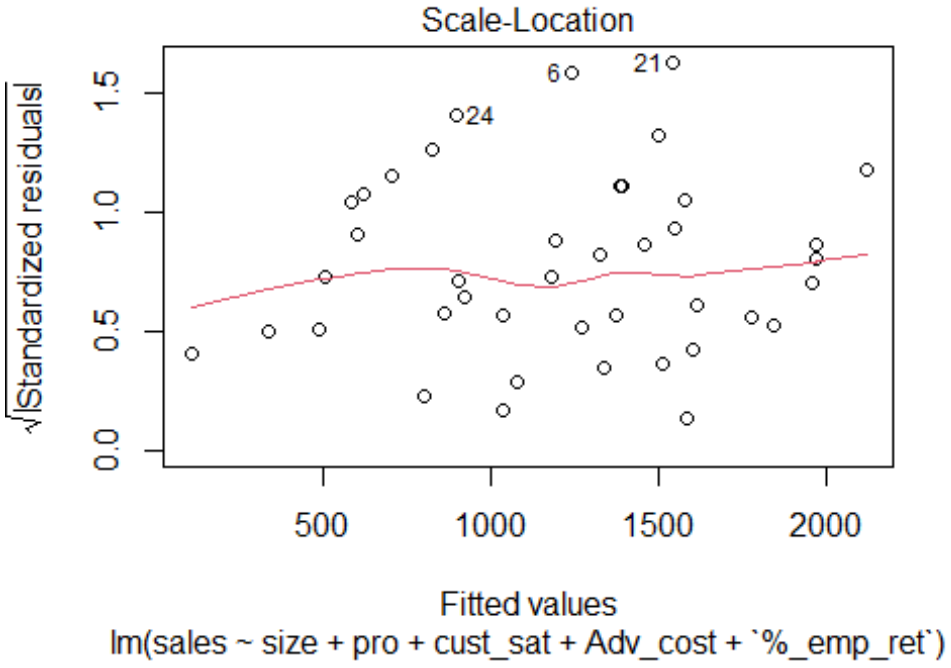
2. Shapiro test

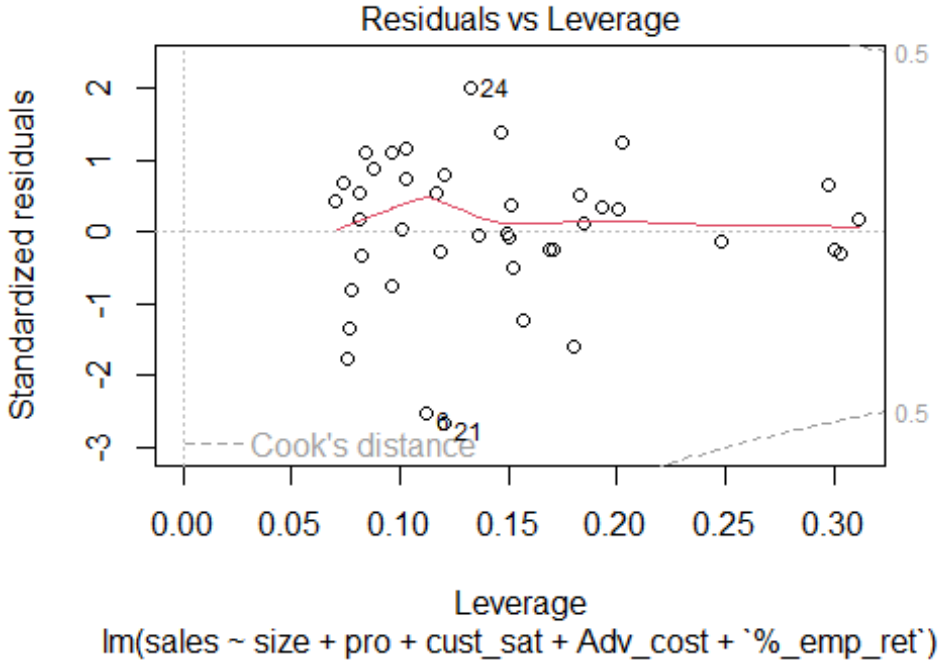
The QQ plot is as follows:

```
plot(forward)
```


Residuals vs Fitted
Residuals
Fitted values
lm(sales ~ size + pro + cust_sat + Adv_cost + `%_emp_ret`)
24
6
21
500
0
-500
500
1000
1500
2000


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(sales ~ size + pro + cust_sat + Adv_cost + `%_emp_ret`)
24
21
6
2
1
0
-1
-2

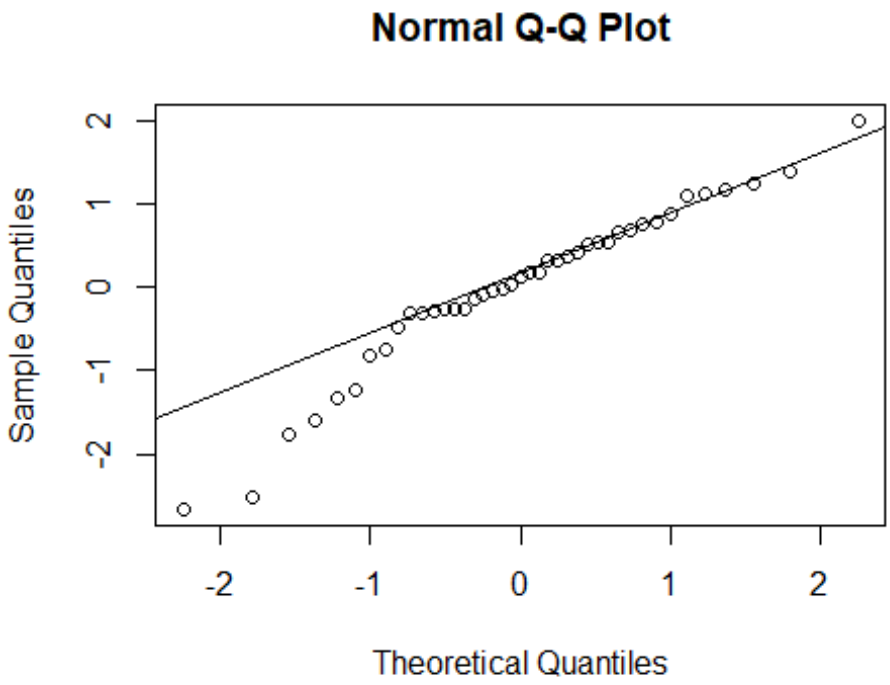





```r
#or

qqnorm(residuall) # QQ plot-of residual
```

```
qqline(residuall) # plots the points
```



If not all, majority of the points fall on the line thus the quartile of normal and residual are almost same, hence it indicates that the residuals follow a normal distribution. However the assumption of normality has to be verified by using a statistical test.(but to know if the deviation of the points lying away from the line, we use the test to further confirm the normality.)

The Shapiro test is as follows:

Hypothesis to testing for normality:

H_0: Errors follow normal distribution.

v/s

H_1: Errors do not follow normal distribution.

```
shapiro.test(residuall)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuall
## W = 0.94973, p-value = 0.06849
```

At 0.15 level of significance the p value is 0.06849 which is lesser than 0.15, thus we reject null and and say that the residuals do not follow normal distribution. Hence one of the assumption of errors is not satisfied.

The assumption of normality is not satisfied. We can fix this problem by performing a transformation. In our model, we can take the logarithm transformation. Then we can refit our model and proceed for further analysis. By doing this transformation, the issue of normality can be fixed.

Conclusion

Model building is done using both forward selection and backward selection.

Model built using forward selection was considered, as it have a greater adjusted R^2 value.

The model is

$$Y[sales] = B_0 + B_1X_1[size] + B_2X_2[promotion] + B_3X_3[customer\ satisfaction] + B_4X_4[advertising\ cost] + B_5X_5[employment\ retension] + E[error]$$

The estimated model is

$$Y^{hat} = -1.762e + 03 + (2.122e - 02)X_1 + (5.208e + 02)X_2 + (4.000e + 01)X_3 + (4.751e + 00)X_4 + (8.087e + 00)X_5$$

There was no multicollinearity in the model.

Residual analysis was performed. Residual analysis includes, plot analysis of fitted values and residual values, which inferred the residuals are in a horizontal band fashion and they fluctuate more or less in a random manner inside, this indicates that there are no visible model defects. So no conclusion can be drawn based on the linearity of the variables and constant variance. However we have shown above that the errors have no constant variance. Outliers presence was checked, and there were no outliers in the model. Later, normality check was done which turned out that, the errors do not follow a normal distribution.

Since the assumptions of the model was violated, we stop the procedure here and say that, the fitted model is a poor fit.

The model can be remodeled by performing certain transformations and proceeding further.

If this does not fix the problem, then non-linear transformation model can be constructed.