Name- Arnab Dey

21MDT0068

Problem definition

Considering a Meddicorp Company and its sales data in three regions of the United States: The South, the West, and the Midwest, which are further divided into territories overseen by regional sales managers, for the year 2003.

The problem is to detect **whether there is a linear relationship between its sales** in each of the territories **and the bonuses paid to the salespeople** working in those territories by the company.

P.T.O

<u>Methodology adopted</u>:

For determining the relationship, we take into account the effects of **advertising, bonus, market share** currently held by Meddicorp in each territory (MKTSHR), and **largest competitor's sales** in each territory (COMPET).

As such the variables to be used in the study include:

Y= Meddicorp's sales (in $000) in each territory for 2003

(SALES) = the amount Meddicorp spent on advertisement in each territory (in $00) in 2003

(ADV) = the total amount of bonuses paid in each territory (in $00) in 2003

(BONUS)

(MKTSHR) = Market share currently held by Meddicorp in each territory

(COMPET) = Largest competitor's sales in each territory.

P.T.O

The data is considered as follows:

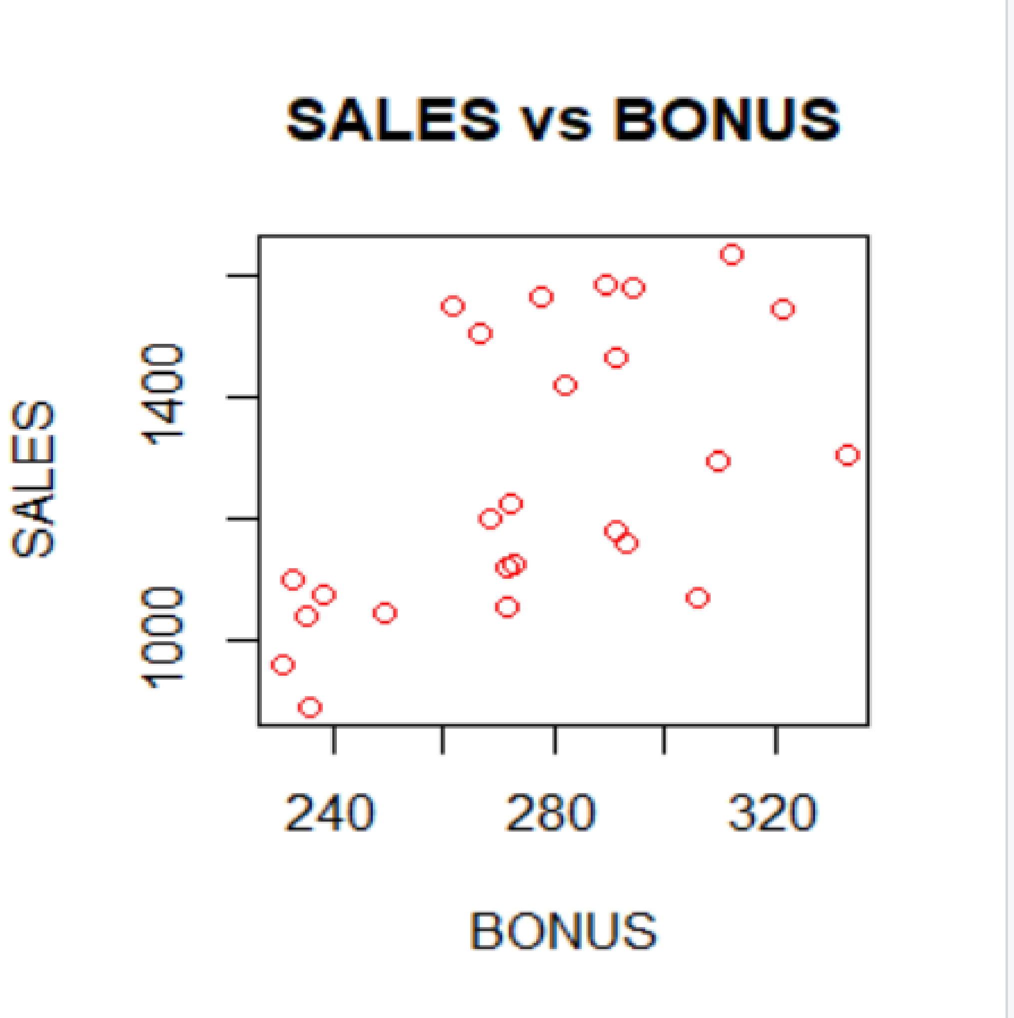
```
> df=read.csv(file='C:\\Users\\arnab\\Desktop\\Sales Data.csv')
> df
   Territory SALES.in..000. ADV.in..00. BONUS.in..00. MKTSHR.in.. COMPET.in.000.
1          1         963.50      374.27        230.98          33         202.22
2          2         893.00      408.50        236.28          29         252.77
3          3        1057.25      414.31        271.57          34         293.22
4          4        1183.25      448.42        291.20          24         202.22
5          5        1419.50      517.88        282.17          32         303.33
6          6        1547.75      637.60        321.16          29         353.88
7          7        1580.00      635.72        294.32          28         374.11
8          8        1071.50      446.86        305.69          31         404.44
9          9        1078.25      489.59        238.41          20         394.33
10        10        1122.50      500.56        271.38          30         303.33
11        11        1304.75      484.18        332.64          25         333.66
12        12        1552.25      618.07        261.80          34         353.88
13        13        1040.00      453.39        235.63          42         262.88
14        14        1045.25      440.86        249.68          28         333.66
15        15        1102.25      487.79        232.99          28         232.55
16        16        1225.25      537.67        272.20          30         273.00
17        17        1508.00      612.21        266.64          29         323.55
18        18        1564.25      601.46        277.44          32         404.44
19        19        1634.75      585.10        312.25          36         283.11
20        20        1159.25      524.56        292.87          34         222.44
21        21        1202.75      535.17        268.27          31         283.11
22        22        1294.25      486.03        309.85          32         242.66
23        23        1467.50      540.17        291.03          28         333.66
24        24        1583.75      583.85        289.29          27         313.44
25        25        1124.75      499.15        272.55          26         374.11
```

We observe that the SALES variable and BONUS are highly correlated with a value of **0.5680652 (close to 1)**

```
> cor(df$SALES.in..000.,df$BONUS.in..00.)
[1] 0.5680652
> plot(df$BONUS.in..00.,df$SALES.in..000., main="SALES vs BONUS", xlab="BONUS",
+       ylab="SALES",col="red")
> 
```

P.T.O

We do scatterplot of SALES vs BONUS as follows:



We observe that there appears to be a strong possible linear relationship between **SALES** and **BONUS**.

P.T.O

We have acquired data for **25 such territories** for each of the three regions in US, where we consider variables as **ADV**, **COMPET and MKTSHR** in addition to **BONUS** to explain possible variations in the output variable **SALES,** as they seem to show **positive linear relationship** with **SALES. (Evident** from their **correlation values as follows)**

Between SALES and Advertisement cost

```
> cor(df$SALES.in..000.,df$ADV.in..00.)
[1] 0.9003288
```

Between SALES and largest competitors sale

```
> cor(df$SALES.in..000.,df$COMPET.in.000.)
[1] 0.3770662
```

Between SALES and Market share of the company in that particular territory:

```
> cor(df$SALES.in..000.,df$MKTSHR.in..)
[1] 0.02311166
```

P.T.O

On the **basis of these results** and the **assumptions** as:

1. **linearity and additivity of the relationship** between dependent and independent variables,

2. **statistical independence of the errors**,

3. **homoscedasticity (constant variance)** of the errors (a) versus time (in the case of time series data) (b) versus the predictions, (c) versus any independent variable,
4. **normality of the error distribution**

5. **Mean of the errors as zero**.

we develop a **linear regression model** as:

$SALES_i = b0 + b1 * BONUS_i + b2 * ADV_i + b3 * COMPET_i + b4 * MKTSHR_i + e_i$

for ith observation from the data.

Where,

b0, b1 , b2 , b3 ,b4 are **coefficients of the model**.

And **e_i= error term** associated with ith observation.

P.T.O

```
> fit=lm(df$SALES.in..000.~df$BONUS.in..00.+df$ADV.in..00.+df$MKTSHR.in..+df$COMPET.in.000.)
> fit

Call:
lm(formula = df$SALES.in..000. ~ df$BONUS.in..00. + df$ADV.in..00. +
    df$MKTSHR.in.. + df$COMPET.in.000.)

Coefficients:
      (Intercept)   df$BONUS.in..00.     df$ADV.in..00.     df$MKTSHR.in..  df$COMPET.in.000.
        -593.5790             1.9056             2.5133             2.6531            -0.1208
```

Using **R programming** , we obtain the values of the model coefficients as **1.9056, 2.5133, 2.6531, -0.1208** for **BONUS, ADV, MKTSHR and COMPET**, respectively.

Now, the regression model is :

(a) SALES= **-593.5790 + 1.9056** * BONUS **+ 2.5133** * ADV **+ (-0.1208)** * COMPET **+ 2.6531*** MKTSHR

(b) For an observation as 500,250,200 and 40 for the variables

BONUS ,ADV ,COMPET , MKTSHR**, the conditional mean of sales** is:

E(SALES | BONUS ,ADV ,COMPET , MKTSHR)

= **-593.5790 + 1.9056** * 500 **+ 2.5133** * 250 **+ (-0.1208)** * 200 **+ 2.6531*** 40

= **<u>1069.51</u>**

(c) Interpretation

If we find the summary of the overall fit of the model using R programming, we have the following observations:

(ANOVA **TABLE** for the fit of the linear regression model)

```
> summary(fit)

Call:
lm(formula = df$SALES.in..000. ~ df$BONUS.in..00. + df$ADV.in..00. +
    df$MKTSHR.in.. + df$COMPET.in.000.)

Residuals:
    Min      1Q  Median      3Q     Max
-187.00  -73.95   16.93   55.64  125.51

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -593.5790   259.2640  -2.289   0.0331 *
df$BONUS.in..00.     1.9056     0.7426   2.566   0.0184 *
df$ADV.in..00.       2.5133     0.3143   7.997 1.17e-07 ***
df$MKTSHR.in..       2.6531     4.6362   0.572   0.5735
df$COMPET.in.000.   -0.1208     0.3719  -0.325   0.7487
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 93.78 on 20 degrees of freedom
Multiple R-squared:  0.8592,    Adjusted R-squared:  0.831
F-statistic:  30.5 on 4 and 20 DF,  p-value: 2.944e-08
```

We observe that the R^2 value for the model is **0.8592** which signifies that **85.92%** of the variation in **SALES** can be explained by the variables **BONUS, ADV, MKTSHR and COMPET,** while the remaining variations are due to error terms associated with the model.

To test the statistical significance of the input variables, let us consider the **level of significance** at we will conduct the test be **5% i.e 0.05**.

Considering the **null hypothesis that the regression coefficients of the model are 0**

And the **alternate hypothesis that the regression coefficients are not 0,**

we observe that the p-values of **BONUS and ADV** are **lesser than the Level of Significance= 0.05**. So we reject the Null hypothesis for these two variables and conclude that they are **statistically significant for the prediction purpose.**

On the other hand, variables **MKTSHR and COMPET** have **p values more than the level of significance= 0.05** as 0.5735 and 0.7487 respectively. So we don't have enough evidence in our data acquired to reject the Null hypothesis for these two variables. So we conclude that they are **statistically insignificant** and are therefore **not important for explaining the variation in SALES.**

(d)

The regression coefficients calculated for the variables corresponding to **bonus, advertisement, market share and competitors** were b1=**1.9056** , b2=**2.5133 ,** b3**=2.6531**

and b4**=-0.1208.**

The confidence intervals for bonus are given as:

[b1- (t statistic value at 0.05 and degree of freedom = n-5) * standard error of b1, b1+ (t statistic value at 0.05) * standard error of b1]

Where, n=25.

When Level of significance= 0.05

i.e [1.9056- 2.08* 0.7426, 1.9056+ 2.08* 0.7426]

i.e [**0.360992, 3.450208**].

C.L=3.04

When Level of significance= 0.01

[1.9056- 2.84* 0.7426, 1.9056+ 2.84* 0.7426]

i.e **[-0.203384, 4.014584].**

C.L = 4.21

The confidence intervals for advertisement are given as:

[b2- (t statistic value at 0.05 and degree of freedom = n-5) * standard error of b2, b2+ (t statistic value at 0.05) * standard error of b2]

Where, n=25.

When Level of significance= 0.05

i.e [2.5133 – 2.08 * 0.3143, 2.5133 + 2.08 * 0.3143]

i.e **[1.859556, 3.167044**].

C.L=1.31

When Level of significance= 0.01:

[2.5133 - 2.84* 0.3143, 2.5133 + 2.84* 0.3143]

i.e [**1.620688, 3.405912**].

C.L= 1.8

The confidence intervals for market share are given as:

[b3- (t statistic value at 0.05 and degree of freedom = n-5) * standard error of b3, b3+ (t statistic value at 0.05) * standard error of b3]

Where, n=25.

When Level of significance= 0.05

i.e [2.6531 – 2.08 * 4.6362, 2.6531 + 2.08 * 4.6362]

i.e [**-6.990196, 12.2964**].

C.L= 19.28

When Level of significance= 0.01 :

[2.6531 - 2.84* 4.6362, 2.6531 + 2.84* 4.6362]

i.e [**-10.51371, 15.81991**].

C.L= 26.32

The confidence intervals for largest competitor's sale are given as:

[b4- (t statistic value at 0.05 and degree of freedom = n-5) * standard error of b4, b4+ (t statistic value at 0.05) * standard error of b4]

Where, n=25.

When Level of significance= 0.05

[-0.1208 – 2.08 * 0.3719, -0.1208 + 2.08 * 0.3719]

i.e **[<u>-0.894352, 0.652752</u>]**.

C.L = 1.542

When Level of significance= 0.01 :

[-0.1208 - 2.84* 0.3719, -0.1208 + 2.84* 0.3719]

i.e **<u>[-1.176996, 0.935396]</u>.**

C.L= 2.1

Therefore, we observe that the confidence length obtained at Level of significance = 0.05 is shorter than that of when obtained at Level of significance = 0.01.

This signifies that **we are more confident with our prediction when we consider the confidence length at Level of significance 0.05 .**

(e)

```
> summary(fit)

Call:
lm(formula = df$SALES.in..000. ~ df$BONUS.in..00. + df$ADV.in..00. +
    df$MKTSHR.in.. + df$COMPET.in.000.)

Residuals:
    Min      1Q  Median      3Q     Max
-187.00  -73.95   16.93   55.64  125.51

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -593.5790   259.2640  -2.289   0.0331 *
df$BONUS.in..00.     1.9056     0.7426   2.566   0.0184 *
df$ADV.in..00.       2.5133     0.3143   7.997 1.17e-07 ***
df$MKTSHR.in..       2.6531     4.6362   0.572   0.5735
df$COMPET.in.000.   -0.1208     0.3719  -0.325   0.7487
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 93.78 on 20 degrees of freedom
Multiple R-squared:  0.8592,	Adjusted R-squared:  0.831
F-statistic:  30.5 on 4 and 20 DF,  p-value: 2.944e-08
```

F- Test for overall fit of the regression model :

Defining the Null and Alternate Hypothesis

H0: The overall regression is not significant.

H1: The overall regression is significant.

Level of significance: 0.05 and 0.01

F statistic obtained for the regression model is 30.5 on 4 and 20 DF. Therefore, **Fcal = 30.5.**

Test Criteria

p-value: 2.944e-08 is much lesser than **Fcal=30.5.**

Therefore, **we have enough evidence in the data to reject H0.**

Decision and Conclusion

We reject the H0 and we conclude that:

The overall regression model is significant at Level of significance 0.05 and 0.01.

Recommendation:

The variables MKTSHR and COMPET corresponding to market share of the company and the Largest competitor’s sales does not explain much variation in the output variable SALES owing to their statistical insignificance. In particular, **MKTSHR has comparatively high standard error** (4.6362), so considering it in the model **may affect the overall forecast accuracy.**

So its recommended to drop these two variables from the regression model.

Before dropping the two variables we got the ANOVA table as :

```
> summary(fit)

Call:
lm(formula = df$SALES.in..000. ~ df$BONUS.in..00. + df$ADV.in..00. +
    df$MKTSHR.in.. + df$COMPET.in.000.)

Residuals:
    Min      1Q  Median      3Q     Max
-187.00  -73.95   16.93   55.64  125.51

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -593.5790   259.2640  -2.289   0.0331 *
df$BONUS.in..00.     1.9056     0.7426   2.566   0.0184 *
df$ADV.in..00.       2.5133     0.3143   7.997 1.17e-07 ***
df$MKTSHR.in..       2.6531     4.6362   0.572   0.5735
df$COMPET.in.000.   -0.1208     0.3719  -0.325   0.7487
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 93.78 on 20 degrees of freedom
Multiple R-squared:  0.8592,    Adjusted R-squared:  0.831
F-statistic:  30.5 on 4 and 20 DF,  p-value: 2.944e-08
```

Now, after dropping the two variables **MKTSHR and COMPET,** we see an **increase in adjusted R^2 value from 83.1% to 84.17 %.**

(ANOVA table after dropping variables **MKTSHR and COMPET)**

```
> summary(fit)

Call:
lm(formula = df$SALES.in..000. ~ df$BONUS.in..00. + df$ADV.in..00.)

Residuals:
     Min       1Q   Median       3Q      Max 
-165.256  -84.616    6.317   54.117  131.372 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -516.3932   189.9277  -2.719   0.0125 *  
df$BONUS.in..00.    1.8557     0.7159   2.592   0.0166 *  
df$ADV.in..00.      2.4734     0.2753   8.983 8.18e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 90.76 on 22 degrees of freedom
Multiple R-squared:  0.8549,	Adjusted R-squared:  0.8417 
F-statistic: 64.81 on 2 and 22 DF,  p-value: 6.001e-10
```

This shows an improvement in the model over the previous one.

Some ideas:

For better model accuracy for prediction of the SALES we can consider accepting other factors like **the consumers' perceived need for medicines at the household level**, **the cost of medicines, the purchasing habits of consumers, the literacy level and consumers' idea about efficacy and power of medicine, together with polypharmacy and polytherapy.**