SA2_Group_Assignment_Group-02

06/12/2021

Authors:

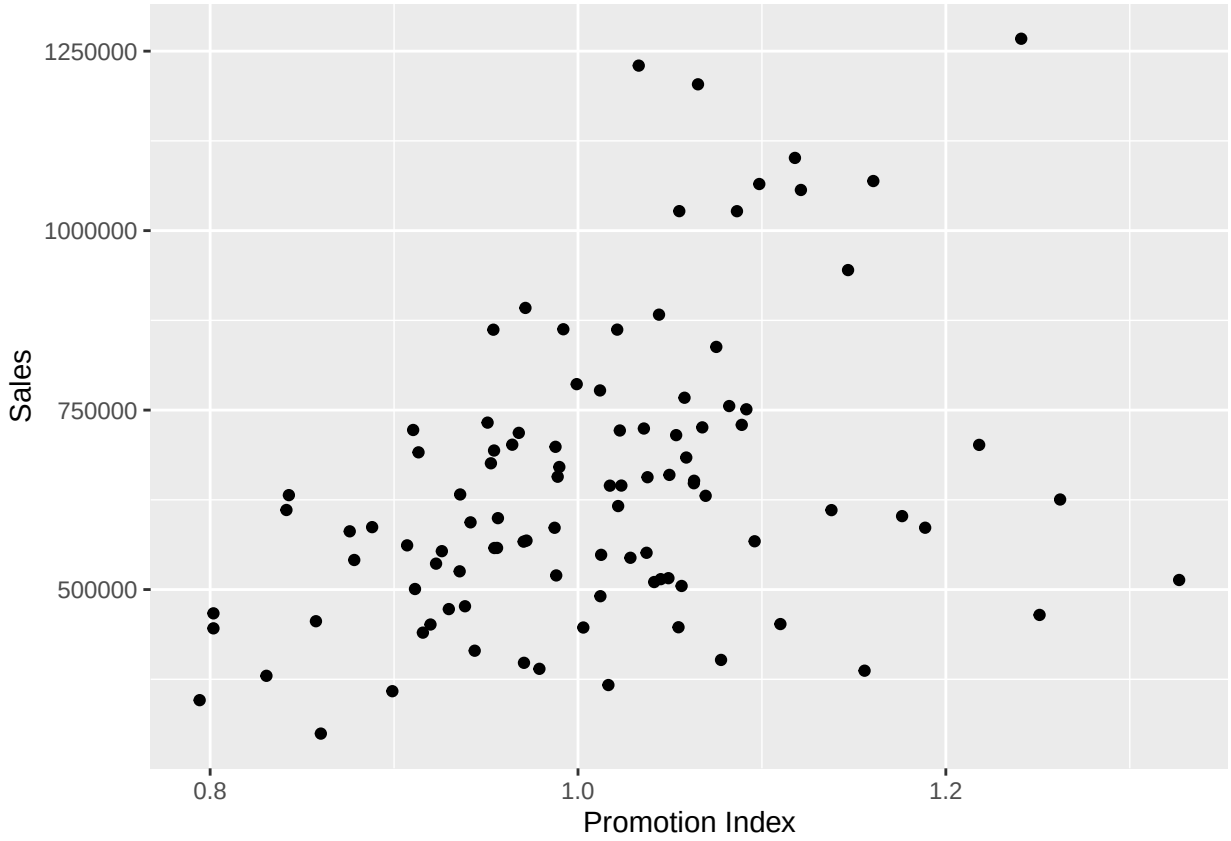
"Shephali Bhardwaj_12110010"
"Anil Jhamb_12110076"
"Ronil Bhan_12110079"
"Shreyas Namjoshi_12110103"

```
#installing required libraries, setting up working directory and reading data file
library("readxl")
library("ggplot2")
#setwd("D:\\ISBAMPBA\\Term2\\SA2\\Assignment")
walmart<-read_excel('walmart_data.xlsx')
```

Assumptions: alpha = 0.05(In All questions alpha value remains same unless specified in the question)

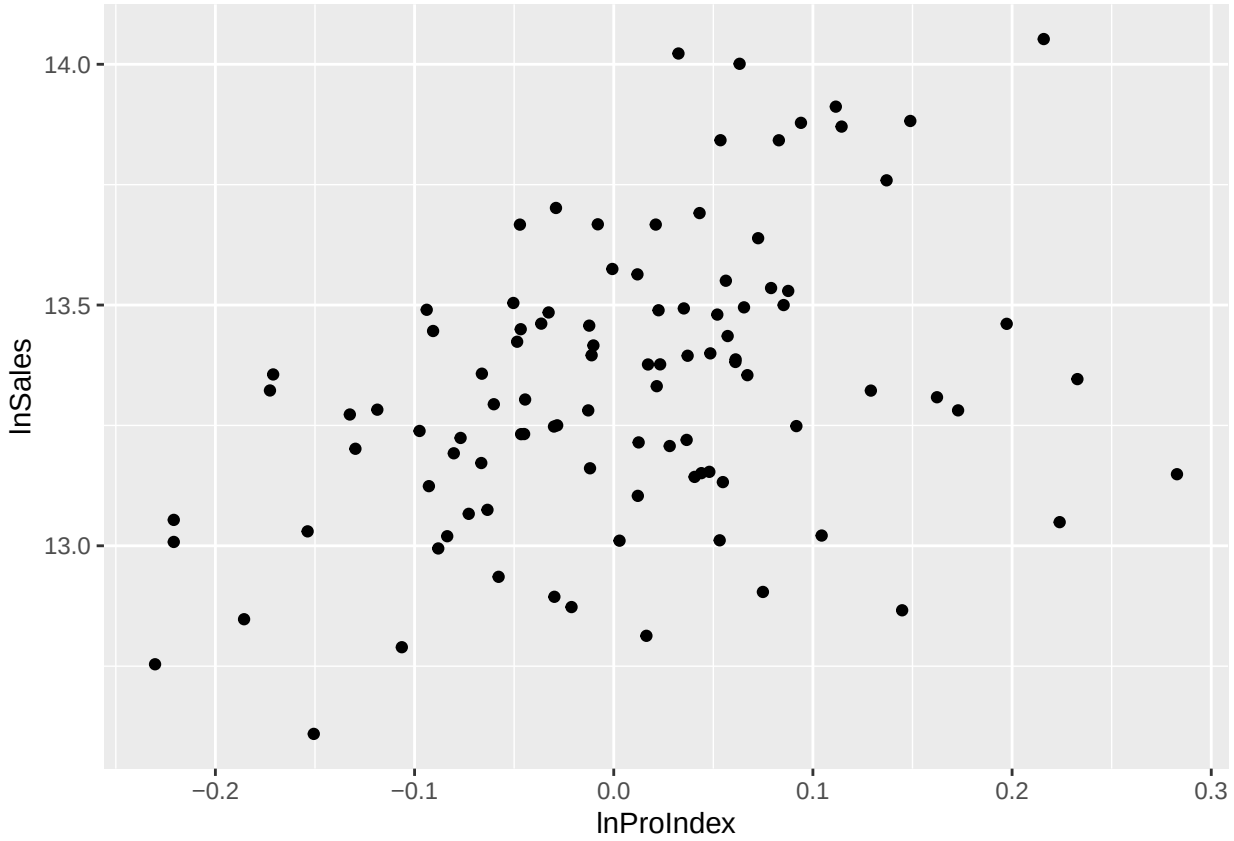
Question 1: a. create a scatterplot of sales and promotion index

```
attach(walmart)
qplot(`Promotion Index`,Sales)
```



Q1: b. create a scatterplot of log(sales) and log(promotion index)

```
walmart$lnSales<-log(Sales)
walmart$lnProIndex<-log(`Promotion Index`)
attach(walmart)
qplot(lnProIndex,lnSales)
```



Q1: c. comment the plots and decide which variable are a better fit for linear regression and why?

Explanation: When we look at the scatter plot of Sales vs Promotion Index, the relationship looks linear with few points spread out as the Sales and Promotion Index increases. Since the data range is high, few values can have huge impact than the other points. When we transform the Sales and Promotion index, the relationship becomes more linear and the values which were spread out becomes closer to the cloud of points. Thus it is better to have transformed variables, which improves the relation and helps in creating a linear model with better Adj R Square Values.

Q2. Estimate the following regression model: Create the appropriate variables. log(sales) = a + b1*log(promotion index) +b2* Walmart

a. What is the interpretation of the coefficient on log(promotion index)?

```
model1<-lm(lnSales~lnProIndex+Walmart)
summary(model1)
```

```
##
## Call:
## lm(formula = lnSales ~ lnProIndex + Walmart)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.44747 -0.15693 -0.02471  0.16110  0.58820
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.47677    0.03279 410.959  < 2e-16 ***
## lnProIndex   0.96244    0.23060   4.174 6.54e-05 ***
## Walmart     -0.30256    0.04638  -6.524 3.11e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2298 on 97 degrees of freedom
## Multiple R-squared:  0.414,  Adjusted R-squared:  0.4019
## F-statistic: 34.26 on 2 and 97 DF,  p-value: 5.555e-12
```

Explanation:

Here b1=0.96244(coef. of log(Promotional Index).This implies that for every 1% increase in Promotional Index,the sales increase by 0.96244% assuming all other factors remain unchanged.

Q2 b. What is the effect of Wal-Mart entry?

For this lets have a model such as log(sales)=a+b1*log(Promotional Index)

```
model2<-lm(lnSales~lnProIndex)
summary(model2)
```

```
##
## Call:
## lm(formula = lnSales ~ lnProIndex)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62684 -0.18809  0.00513  0.17951  0.66020
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.32436    0.02746 485.152  < 2e-16 ***
## lnProIndex   1.16420    0.27270   4.269 4.54e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2742 on 98 degrees of freedom
## Multiple R-squared:  0.1568, Adjusted R-squared:  0.1482
## F-statistic: 18.23 on 1 and 98 DF,  p-value: 4.539e-05
```

Explanation:

We see that Adjusted R Square Value with only log(promotional index) as an independent variable is only 0.1482

Now Consider model as given in Q2, i.e log(sales) = a + b1*log(promotion index) +b2* Walmart

```
model3<-lm(lnSales~lnProIndex+Walmart)
summary(model3)
```

```
##
```

```
## Call:
## lm(formula = lnSales ~ lnProIndex + Walmart)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44747 -0.15693 -0.02471  0.16110  0.58820
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.47677    0.03279 410.959  < 2e-16 ***
## lnProIndex   0.96244    0.23060   4.174 6.54e-05 ***
## Walmart     -0.30256    0.04638  -6.524 3.11e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2298 on 97 degrees of freedom
## Multiple R-squared:  0.414,  Adjusted R-squared:  0.4019
## F-statistic: 34.26 on 2 and 97 DF,  p-value: 5.555e-12
```

We see below two observations:

1. Interpretation of b1- Sales for the weeks after which Walmart has Opened are "negatively impacted" by 30% as compared to sales for the weeks when walmart is not opened,assuming all other factors remain unchanged.
2. Addition of walmart variable increases the Adjusted R Square value of the model to 0.4019,Which is a significant increase.Thus we can say that Walmart further explains the variation in sales.

Q3. Which independent variables are significant in explaining the variation in sales?

Assuming the model eqn to be Sales= b0+b1*Promotion Index*+*b2*Walmart+b3*FeatureAdvertising Index*+*b4*Holiday

```
model4<-lm(Sales~`Promotion Index`+Walmart+`FeatureAdvertising Index`+Holiday)
summary(model4)
```

```
##
## Call:
## lm(formula = Sales ~ `Promotion Index` + Walmart + `FeatureAdvertising Index` +
##     Holiday)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -250601 -103492  -11463   88677  433546
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -339699     202091  -1.681 0.096063 .
## `Promotion Index`            591067     144805   4.082 9.31e-05 ***
## Walmart                     -198480      29566  -6.713 1.38e-09 ***
## `FeatureAdvertising Index`   466834     145326   3.212 0.001798 **
## Holiday                      192599      54236   3.551 0.000599 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 146700 on 95 degrees of freedom
```

```
## Multiple R-squared:  0.4914, Adjusted R-squared:   0.47
## F-statistic: 22.95 on 4 and 95 DF,  p-value: 2.745e-13
```

Explanation:

We see that all the variables `Promotion Index`, Walmart, `FeatureAdvertising Index`,Holiday have a significant impact on the Sales as explained by the p-values of each variable.Here p-value for each variable is less than assumed alpha=0.05

Also we consider the model as is from Q2 i.e log(sales) = a + b1*log(promotion index) +b2* Walmart

```
model3<-lm(lnSales~lnProIndex+Walmart)
summary(model3)
```

```
##
## Call:
## lm(formula = lnSales ~ lnProIndex + Walmart)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44747 -0.15693 -0.02471  0.16110  0.58820
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.47677    0.03279 410.959  < 2e-16 ***
## lnProIndex   0.96244    0.23060   4.174 6.54e-05 ***
## Walmart     -0.30256    0.04638  -6.524 3.11e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2298 on 97 degrees of freedom
## Multiple R-squared:  0.414,  Adjusted R-squared:  0.4019
## F-statistic: 34.26 on 2 and 97 DF,  p-value: 5.555e-12
```

Explanation:

We see that all the variables log(`Promotion Index`) and Walmart have a significant impact on the Sales as explained by the p-values of each variable.Here p-value for each variable is less than assumed alpha=0.05

Q4. The local store also engages in feature advertising by mailing ads to households. 'Feature Advertising Index' gives the feature advertising activity in a given week. You add the log of this variable to the regression. In addition to this, you also add a 'Holiday Dummy' equal to one if the corresponding week covers a major holiday. Add these two variables to the regression and re-estimate the model

log(sales) = b0 + b1 log(promotion index) + b2 WalMart + b3 log(feature index) + b4 Holiday Interpret the two newly estimated coefficients.

```
walmart$lnFindex<-log(`FeatureAdvertising Index`)
attach(walmart)
model5<-lm(lnSales~lnProIndex+Walmart+lnFindex+Holiday)
summary(model5)
```

```
##
## Call:
## lm(formula = lnSales ~ lnProIndex + Walmart + lnFindex + Holiday)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.45327 -0.15721 -0.00367  0.12272  0.46376
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.45697    0.03049 441.411  < 2e-16 ***
## lnProIndex   0.90240    0.21065   4.284 4.40e-05 ***
## Walmart     -0.30684    0.04224  -7.264 1.03e-10 ***
## lnFindex     0.71829    0.20623   3.483 0.000752 ***
## Holiday      0.26057    0.07728   3.372 0.001082 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2092 on 95 degrees of freedom
## Multiple R-squared:  0.5243, Adjusted R-squared:  0.5042
## F-statistic: 26.17 on 4 and 95 DF,  p-value: 1.224e-14
```

Explanation:

here b3=0.71829(coef. of log(FeatureAdvertising Index)). This means that increase in 1% of AdvertisingFeature Index will increases the sales by 0.71829% assuming all other factors remain unchanged.

here b4=0.26057(coef. of Holidays). This means that Sales are impacted positively by 26.05% during the weeks with major holidays as compared to when weeks do not have major holidays assuming all other factors remain unchanged.

Q5. Are the two new coefficients significant?

Assuming model equation to be same as in Q4 i.e log(sales) = b0 + b1 log(promotion index) + b2 WalMart + b3 log(feature index) + b4 Holiday

Explanation: Yes, Both the new coefficients are significant.The P Values are 0.000752 and 0.001082 for log(FeatureAdvertising Index) and Holiday respectively, which is less than assumed alpha=0.05.

Q6. You add a final variable to the regression: log(promotion Index) X WalMart, i.e., the Wal-Mart dummy multiplied by the log(promotion index) variable. Create this interaction variable. The full regression is now

log(sales) = b0 + b1 log(promotion index) + b2 WalMart + b3 log(feature index) + b4 Holiday + b5(log(promotion Index) x WalMart)

What is the interpretation of b5?

```
walmart$lnProIndex_Wal<-lnProIndex*Walmart
attach(walmart)
```

Explanation: b5(coef of log(Promotional Index)*walmart) is additional impact(%increase or %decrease) of Promotional Index on sales in presence of walmart assuming all other factors remain unchanged.

Q7. Estimate the regression. Paste results here. Is the effect of promotions on store sales higher or lower after Wal-Mart enters?

```
model6<-lm(lnSales~lnProIndex+Walmart+lnFindex+Holiday+lnProIndex_Wal)
summary(model6)
```

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

```
## lm(formula = lnSales ~ lnProIndex + Walmart + lnFindex + Holiday +
##     lnProIndex_Wal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41021 -0.15980 -0.00894  0.11572  0.50489
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    13.44880    0.03034 443.264  < 2e-16 ***
## lnProIndex      1.46201    0.35470   4.122 8.10e-05 ***
## Walmart        -0.29863    0.04185  -7.136 1.98e-10 ***
## lnFindex        0.73694    0.20350   3.621 0.000475 ***
## Holiday         0.22915    0.07787   2.943 0.004096 **
## lnProIndex_Wal -0.86417    0.44409  -1.946 0.054651 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2062 on 94 degrees of freedom
## Multiple R-squared:  0.5427, Adjusted R-squared:  0.5184
## F-statistic: 22.31 on 5 and 94 DF,  p-value: 1.107e-14
```

Explanation:

The P Value for b5 is 0.054651 which is greater than assumed alpha=0.05 and hence the b5 has non significant contribution to sales. But we will still keep b5 in the model and calculate the sales basic the model given in Q6 i.e log(sales) = b0 + b1 log(promotion index) + b2 WalMart + b3 log(feature index) + b4 Holiday + b5(log(promotion Index) x WalMart)

Lets Calculate Predicted Value of Sales with Walmart=0 and Walmart=1 Case 1: walmart=0, promotionalIndex=0.8880742, FeatureAdvertising Index=0.8705762 and holiday=0 The equation becomes

```
y_pred<-13.44880 + (1.46201*log(0.8880742)) + 0 + (0.73694*log(0.8705762)) + 0 + 0
y_pred
```

```
## [1] 13.17312
```

```
sales_y_pred<-exp(y_pred)
sales_y_pred
```

```
## [1] 526033.2
```

Predicted Sales comes out to be 526033.2 when walmart =0

Case 2: walmart=1, promotionalIndex=0.8880742, FeatureAdvertising Index=0.8705762 and holiday=0 The equation becomes

```
y_pred_wal<-13.44880 + (1.46201*log(0.8880742)) -(0.29863*1) + (0.73694*log(0.8705762)) + 0 - (0.86417*
y_pred_wal
```

```
## [1] 12.97707
```

```
sales_y_pred_wal<-exp(y_pred_wal)
sales_y_pred_wal
```

```
## [1] 432382.8
```

Predicted Sales comes out to be 432382.8 when walmart =1

We clearly see that Sales in presence walmart is lower than sales in absence of walmart assuming all other

factors remaining unchanged, This is also evident from the coeff of walmart variable.Since the coef. is negative, we can say that addition of walmart is negatively impacting sales.

Q8. What does the estimate for b5 imply about the possibility of the local store using promotional activity to fight Wal-Mart? What strategy would you recommend to the local store?

The P Value for b5 is 0.054651 which is greater than assumed alpha=0.05 and hence the b5 has non significant contribution to sales. we will remove the interaction variable and recalculate the regression without interaction variable(same as q4) Model Equation becomes: log(sales) = b0 + b1 log(promotion index) + b2 WalMart + b3 log(feature index) + b4 Holiday

```
summary(model5)
```

```
##
## Call:
## lm(formula = lnSales ~ lnProIndex + Walmart + lnFindex + Holiday)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45327 -0.15721 -0.00367  0.12272  0.46376
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.45697    0.03049 441.411  < 2e-16 ***
## lnProIndex   0.90240    0.21065   4.284 4.40e-05 ***
## Walmart     -0.30684    0.04224  -7.264 1.03e-10 ***
## lnFindex     0.71829    0.20623   3.483 0.000752 ***
## Holiday      0.26057    0.07728   3.372 0.001082 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2092 on 95 degrees of freedom
## Multiple R-squared:  0.5243, Adjusted R-squared:  0.5042
## F-statistic: 26.17 on 4 and 95 DF,  p-value: 1.224e-14
```

As can be seen in the model, log('Promotional Index') and log('Advertisingfeature index') have positive impact on sales,hence we suggest investing further in Promotion and featureadvertising which would lead to higher sales.