OM 386: Demand Analytics and Pricing
Assignment #1

Please paste your answers within this file and save it as "HW1_eid1_eid2_eid3" (Where eids refer to your group members' EIDs) on Canvas at appropriate place. If you used MS Excel or any other software to arrive at your answers, please submit the relevant files/annotated code as well.

Write the names of your team members here:

1. Meeth Yogesh Handa (EID: mh58668)
2. Rishabh Tiwari (EID: rt27739)
3. Saurabh Arora (EID: sa55445)
4. Rianna Patel (EID: rnp599)
5. Parthiv Borgohain (EID: pb25347)

Q1: Use Regression to Understand Prices, Promotions and Entry

You have been hired as a consultant for a major local grocery store. Store management is worried since Wal-Mart has entered the market by opening a "Wal-Mart Super-center" only 3 miles away from the local store. Management is interested in analyzing the impact on store sales of the Wal-Mart entry and whether or not a new strategy is required.

For analysis, management has given you access to one hundred weeks of sales data for the local store covering the period both pre- and post-entry of Wal-Mart.

Look at the data in *"HW1 Walmart.xls"* (The file is available on Canvas).

It has the following variables:

WEEK	Week number
Sales	Weekly sales
Promotion Index	Index of weekly promotion activity –higher promotion index indicates more products on promotion in the store
Walmart	Walmart dummy = 1 in the weeks after the Walmart opens, and 0 in the weeks before the Walmart opens

Feature Advertising Index	Index of feature advertising activity – higher feature advertising index indicates more feature advertising
Holiday	Holiday Dummy = 1 during major holiday weeks, and 0 for non-holiday weeks

A. Estimate the following regression model: Create the appropriate variables[1]. (5 points)

log(sales) = α + β_1 *log(promotion index)* + β_2 *WalMart*

Paste results here.

```
> # Defining first model
> LogModel1 = lm(logSales ~ logPromotionIndex + WalmartData$Walmart)
> LogModel1

Call:
lm(formula = logSales ~ logPromotionIndex + WalmartData$Walmart)

Coefficients:
        (Intercept)    logPromotionIndex  WalmartData$Walmart
            13.4768               0.9624              -0.3026
```

```
> summary(LogModel1)

Call:
lm(formula = logSales ~ logPromotionIndex + WalmartData$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 ***
logPromotionIndex    0.96244    0.23060   4.174 6.54e-05 ***
WalmartData$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
```

B. What is the interpretation of the coefficient on *log(promotion index)?* (5 points)

[1] I use functions Log and Ln interchangeably both referring to natural logarithm.

Since we are using a log-log regression model, the coefficient represents the elasticity of dependent variable (Sales) with respect to the independent variable (log(promotion index)). Thus, a coefficient of 0.96244 represents that a 1% increase in the promotion index results in a 0.96% increase in the overall Sales. Furthermore, we can observe that the coefficient is statistically quite significant and therefore, very relevant for our analysis.

C. What is the effect of Walmart entry? (5 points)

The coefficient obtained for binary variable Walmart = -0.30256 and is statistically significant with a very low p-value. Thus, it is seen that the opening of a Walmart store negatively affects the log of weekly sales. As log is an ever increasing function, the sales are directly proportional to the log of sales value. Therefore, we can hypothesize that the entry of Walmart leads to a decrease in weekly sales. This is subject to the fact that the independent variable is not associated with some other independent variable which is the real reason for the increase in sales.

D. Which independent variables are significant in explaining the variation in sales? (2 points)

In the above model (LogModel1), both the independent variables considered: log(promotion index) and Walmart are statistically significant with a confidence level of 0.1%.

E. 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

$$log(sales) = \alpha + \beta_1\, log(promotion\ index) + \beta_2\, WalMart + \beta_3\, log(feature\ index) + \beta_4\, Holiday$$

Paste results here. (5 points)

```
> # Defining second model
> LogModel2 = lm(logSales ~ logPromotionIndex + WalmartData$Walmart + logFeatureAdvertising + WalmartData$Holiday)
> LogModel2

Call:
lm(formula = logSales ~ logPromotionIndex + WalmartData$Walmart +
    logFeatureAdvertising + WalmartData$Holiday)

Coefficients:
          (Intercept)      logPromotionIndex    WalmartData$Walmart  logFeatureAdvertising    WalmartData$Holiday
              13.4570                 0.9024                -0.3068                 0.7183                 0.2606
```

```
> summary(LogModel2)

Call:
lm(formula = logSales ~ logPromotionIndex + WalmartData$Walmart +
    logFeatureAdvertising + WalmartData$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 ***
logPromotionIndex      0.90240    0.21065   4.284 4.40e-05 ***
WalmartData$Walmart   -0.30684    0.04224  -7.264 1.03e-10 ***
logFeatureAdvertising  0.71829    0.20623   3.483 0.000752 ***
WalmartData$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
```

F. Interpret the two newly estimated coefficients. (4 points)

The log of the feature-advertising index and holiday-based indicator variable have been identified as two new variables that have a positive correlation with the log of sales value. The T statistic of the coefficient suggests that the likelihood of these two coefficients' magnitudes being zero is less than 0.1%. So it is extremely unlikely statistically. Therefore, we can infer that the sales (approximately the log of sales) are directly proportional to the level of advertising index. 1% increase in feature advertising leads to a 0.7% increase in sales. Additionally, it has been observed that weekly sales increase when there is a major holiday in the week.

G. Are the two new coefficients significant? (2 points)

Yes, the two new coefficients are statistically significant as their t-statistic has very small p values (less than 0.1%).

You add a final variable to the regression: *log(promotion Index)* $\times$ *WalMart*, i.e., the Wal-Mart Dummy multiplied by the log(promotion index) variable. Create this interaction variable. The full regression model is now:

$$log(sales) = \alpha + \beta_1\, log(promotion\ index) + \beta_2\, WalMart + \beta_3\, log(feature\ index)$$

$$+ \beta_4\, Holiday + \beta_5\, (\log(\text{promotion Index}) \times WalMart_t)$$

H. What is the interpretation of β_5? (5 points)

β_5 examines the relationship between the changes in the log of weekly sales value and the combined effect of the promotional index and Walmart variables. Since the Walmart variable is binary or an indicator, the combination of the continuous (log of promotional index) and binary (Walmart) variables expresses a hidden behavior of the promotional index. In other words, it creates a variable related to the value of the promotional index that only becomes active when the Walmart variable is present. Therefore, β_5 illustrates how the value of the log of sales changes in relation to the value of the promotional index in the weeks after Walmart store is opened.

I. Estimate the regression. Paste results here. (5 points)

Result:

```
> # Defining third model
> LogModel3 = lm(logSales ~ logPromotionIndex + WalmartData$Walmart + logFeatureAdvertising + WalmartData$Holiday + logPro
dex:WalmartData$Walmart)
> LogModel3

Call:
lm(formula = logSales ~ logPromotionIndex + WalmartData$Walmart +
    logFeatureAdvertising + WalmartData$Holiday + logPromotionIndex:WalmartData$Walmart)

Coefficients:
                          (Intercept)                    logPromotionIndex                   WalmartData$Walmart
                              13.4488                               1.4620                              -0.2986
                logFeatureAdvertising                  WalmartData$Holiday  logPromotionIndex:WalmartData$Walmart
                               0.7369                               0.2292                              -0.8642
```

```
> summary(LogModel3)

Call:
lm(formula = logSales ~ logPromotionIndex + WalmartData$Walmart +
    logFeatureAdvertising + WalmartData$Holiday + logPromotionIndex:WalmartData$Walmart)

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 ***
logPromotionIndex                    1.46201    0.35470   4.122 8.10e-05 ***
WalmartData$Walmart                 -0.29863    0.04185  -7.136 1.98e-10 ***
logFeatureAdvertising                0.73694    0.20350   3.621 0.000475 ***
WalmartData$Holiday                  0.22915    0.07787   2.943 0.004096 **
logPromotionIndex:WalmartData$Walmart -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
```

J. Is the effect of promotions on store sales higher or lower after Wal-Mart enters? (2 points)

The coefficients of the variables associated with promotional_index values can assist us in understanding the influence of the promotional index on sales after taking into account the Walmart indicator variable. The adjusted coefficient becomes (β_1 + β_5 * walmart), and the estimate for β_5 has a negative value in this sample dataset. As a result, we can conclude that the impact of the promotional index on the log of sales will decrease when Walmart is included/opened. However, a point to note is that the negative coefficient of the interaction term (promotion index:Walmart) is significant only at the 10% level which is less than the 5% usually considered as the threshold.

K. What does the estimate for β_5 imply about the possibility of the local store using promotional activity to fight Wal-Mart? What strategy would you recommend to the local store? (5 points)

Upon analyzing the coefficients related to the two variables and interaction terms, the following equation was developed: log_sales= 13.44 + 1.462**log_promotional_index -0.29***Walmart + 0.736*log_feature_advertising + 0.229*Holiday -0.86**Walmart**log_promotional_index.

By keeping other variables constant and focusing on the margins where Walmart values are 1, the equation becomes –

log_sales= 1.462*log_promotional_index -0.29**1 -0.86**1**log_promotional_index

= 13.44 - 0.29 + (0.6*log_promotional_index).

Therefore, it is suggested that the store owner should increase promotions temporarily and try to keep the other variables constant.

Q2: Use Two-Stage Regression to Test/Overcome Endogeneity

The Omitted_Variable_Price-Demand_Dataset excel file posted on Canvas provided the historical demand and price data points of a product. A data analyst suspected that besides the price, there are other independent factors that influence the demand and the price data is related to these factors. Unfortunately, such data are unavailable. The data analyst claimed that s/he uncovered two so-called instrument variables that s/he believes are related to the price, and can be related to the demand of the product only through the price (and thus not related to the unobservable factors).

Your team was approached to conduct the analysis following the data analyst's claim. Report your procedure (step 1, 2, …) and results concisely:

a) check whether both of the claimed instrumental variables are related to the demand and price;

```
> # Displaying a correlation matrix to find out if the IVs are related to Demand, Price
> cor_matrix = cor(OVData)
> cor_matrix
             D          P         IV1         IV2
D    1.0000000 -0.7524704 -0.7233842 -0.6257961
P   -0.7524704  1.0000000  0.8390225  0.7381610
IV1 -0.7233842  0.8390225  1.0000000  0.8797868
IV2 -0.6257961  0.7381610  0.8797868  1.0000000
```

Based on the results obtained from the above correlation matrix, it can be observed that:

1) Both IV1, IV2 have a strong negative correlation with Demand (D)
2) Both IV1, IV2 have a strong positive correlation with Price (P)

Thus, both IV1, IV2 are related to demand and price and further analysis is required to determine their effect on the two variables and assess the statistical significance.

b) conduct two stages IV regressions;

Stage 1 Regression (Price ~ IV1 + IV2):

```
> summary(Regr11)

Call:
lm(formula = Price ~ IV1 + IV2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.92325 -0.23708 -0.00906  0.24603  1.17580

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.363e+00  1.613e-02  208.44   <2e-16 ***
IV1          1.000e+00  6.220e-02   16.08   <2e-16 ***
IV2         -6.510e-11  4.020e-02    0.00        1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3549 on 481 degrees of freedom
Multiple R-squared:  0.704,     Adjusted R-squared:  0.7027
F-statistic: 571.9 on 2 and 481 DF,  p-value: < 2.2e-16
```

Stage 2 Regression (Demand ~ Predicted Price):

```
> Regr12 = lm(Demand ~ Pred_Price1)
> Regr12

Call:
lm(formula = Demand ~ Pred_Price1)

Coefficients:
(Intercept)  Pred_Price1
     101.45       -10.43

> summary(Regr12)

Call:
lm(formula = Demand ~ Pred_Price1)

Residuals:
    Min      1Q  Median      3Q     Max
-17.588  -3.191   0.016   3.428  16.285

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.4520     1.5450   65.66   <2e-16 ***
Pred_Price1 -10.4318     0.4535  -23.00   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.444 on 482 degrees of freedom
Multiple R-squared:  0.5233,    Adjusted R-squared:  0.5223
F-statistic: 529.1 on 1 and 482 DF,  p-value: < 2.2e-16
```

c) check whether both of the claimed instrumental variables are useful;

After running stage 1 regression of Price over IV1, IV2 we get the following results:

```
> summary(Regr11)

Call:
lm(formula = Price ~ IV1 + IV2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.92325 -0.23708 -0.00906  0.24603  1.17580

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.363e+00  1.613e-02  208.44   <2e-16 ***
IV1          1.000e+00  6.220e-02   16.08   <2e-16 ***
IV2         -6.510e-11  4.020e-02    0.00        1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3549 on 481 degrees of freedom
Multiple R-squared:  0.704,     Adjusted R-squared:  0.7027
F-statistic: 571.9 on 2 and 481 DF,  p-value: < 2.2e-16
```

From the above results, we can conclusively prove that IV2 is not useful in predicting the true price since the coefficient obtained is not statistically significant. IV1 on the other hand has an extremely small p-value and hence, is quite useful and relevant.

d) check whether the price alone has endogeneity issues as claimed;

```
# Checking for potential endogeneity issue
EndoRegr = lm(Demand ~ Price + Residuals)
EndoRegr
summary(EndoRegr)
```

```
> summary(EndoRegr)

Call:
lm(formula = Demand ~ Price + Residuals)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9700  -3.4892  -0.0275   3.1921  13.5551

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.4520     1.4259  71.151  < 2e-16 ***
Price       -10.4318     0.4185 -24.924  < 2e-16 ***
Residuals     4.4837     0.7692   5.829 1.02e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.024 on 481 degrees of freedom
Multiple R-squared:  0.5948,    Adjusted R-squared:  0.5931
F-statistic: 353.1 on 2 and 481 DF,  p-value: < 2.2e-16
```

The coefficient for the residual term obtained after Stage 1 regression is statistically extremely significant with a confidence interval of 0.1%

Thus, it can be determined that the price alone does indeed have endogeneity issues as claimed.

e) derive the demand and price model.

Demand model:

Demand = 101.45 – 10.43(True Price) + Error term

```
> summary(Regr12)

Call:
lm(formula = Demand ~ Pred_Price1)

Residuals:
    Min      1Q  Median      3Q     Max
-17.588  -3.191   0.016   3.428  16.285

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.4520     1.5450   65.66   <2e-16 ***
Pred_Price1 -10.4318     0.4535  -23.00   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.444 on 482 degrees of freedom
Multiple R-squared:  0.5233,    Adjusted R-squared:  0.5223
F-statistic: 529.1 on 1 and 482 DF,  p-value: < 2.2e-16
```

Price model received for given data:

True Price = 3.36 + IV1 + Error term (Note: The coefficient of IV2 is extremely close to zero and insignificant)

```
> summary(Regr11)

Call:
lm(formula = Price ~ IV1 + IV2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.92325 -0.23708 -0.00906  0.24603  1.17580

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.363e+00  1.613e-02  208.44   <2e-16 ***
IV1          1.000e+00  6.220e-02   16.08   <2e-16 ***
IV2         -6.510e-11  4.020e-02    0.00        1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3549 on 481 degrees of freedom
Multiple R-squared:  0.704,     Adjusted R-squared:  0.7027
F-statistic: 571.9 on 2 and 481 DF,  p-value: < 2.2e-16
```