eBay Assignment Marketing Analytics

Debarati Mazumdar, Maura Oray

10/21/2018

knitr::opts\_chunk$set(echo = TRUE)

**Introduction**

Is online advertising effective? This is a natural question that would interest any business using an online platform. With more and more businesses joining the e-commerce sector, especially with Amazon taking 49.1% of e-commerce sales [[1]](#footnote-21), how does a company like eBay get ahead?

library(dplyr)  
library(ggplot2)  
library(tidyverse)  
library(randomizeR)  
  
#The required libraries for the case analysis

setwd("~/Desktop/BUS-256/MarketingAnalytics") #Set the working directory for easier collaboration  
ebay <- read.csv("eBayData.csv")  
summary(ebay)

## date dma isTreatmentPeriod isTreatmentGroup  
## 4/1/12 : 210 Min. :500.0 Min. :0.0000 Min. :0.0000   
## 4/10/12: 210 1st Qu.:552.0 1st Qu.:0.0000 1st Qu.:0.0000   
## 4/11/12: 210 Median :627.5 Median :1.0000 Median :1.0000   
## 4/12/12: 210 Mean :641.1 Mean :0.5487 Mean :0.6762   
## 4/13/12: 210 3rd Qu.:724.0 3rd Qu.:1.0000 3rd Qu.:1.0000   
## 4/14/12: 210 Max. :881.0 Max. :1.0000 Max. :1.0000   
## (Other):22470   
## revenue   
## Min. : 367   
## 1st Qu.: 26386   
## Median : 53919   
## Mean : 120850   
## 3rd Qu.: 116900   
## Max. :2601390   
##

str(ebay)

## 'data.frame': 23730 obs. of 5 variables:  
## $ date : Factor w/ 113 levels "4/1/12","4/10/12",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ dma : int 500 501 502 503 504 505 506 507 508 509 ...  
## $ isTreatmentPeriod: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ isTreatmentGroup : int 0 1 1 1 1 0 1 1 1 1 ...  
## $ revenue : num 76719 2096177 34994 34199 641014 ...

dim(ebay)

## [1] 23730 5

Here we see that the data set has 5 variables with 23,730 observations. The minimum revenue is $367, maximum is $2,601,390 with a median of $53,919.

In terms of the data types, date is a factor, DMA, isTreatmentPeriod and isTreatmentGroup are integers, and revenue is numerical.

**a. Convert the Date Column and Transform Variables**

ebay$rDate <- as.Date(ebay$date, format="%m/%d/%y")   
head(ebay)

## date dma isTreatmentPeriod isTreatmentGroup revenue rDate  
## 1 4/1/12 500 0 0 76718.74 2012-04-01  
## 2 4/1/12 501 0 1 2096176.54 2012-04-01  
## 3 4/1/12 502 0 1 34993.85 2012-04-01  
## 4 4/1/12 503 0 1 34198.75 2012-04-01  
## 5 4/1/12 504 0 1 641014.21 2012-04-01  
## 6 4/1/12 505 0 0 327989.31 2012-04-01

#Convert to factors  
ebay[3:4] <- lapply(ebay[3:4], factor)

In the above code, we added a column called ‘rDate’ and transformed the date from a factor to a date type.

We also decided to convert the Treatment Period and Treatment Group variables from integers to factors, because the variables are essentially categorical. Converting to factors will make for better statistical analysis and regression.

**b. Determine Treatment Period start date**

ebay\_treatment = subset(ebay[order(ebay$rDate),]) #order date  
head(ebay\_treatment)

## date dma isTreatmentPeriod isTreatmentGroup revenue rDate  
## 1 4/1/12 500 0 0 76718.74 2012-04-01  
## 2 4/1/12 501 0 1 2096176.54 2012-04-01  
## 3 4/1/12 502 0 1 34993.85 2012-04-01  
## 4 4/1/12 503 0 1 34198.75 2012-04-01  
## 5 4/1/12 504 0 1 641014.21 2012-04-01  
## 6 4/1/12 505 0 0 327989.31 2012-04-01

ebay\_treatment1 = subset(ebay[order(ebay$rDate), ],isTreatmentPeriod == "1") #13,020 observations   
head(ebay\_treatment1)

## date dma isTreatmentPeriod isTreatmentGroup revenue rDate  
## 10711 5/22/12 500 1 0 85870.94 2012-05-22  
## 10712 5/22/12 501 1 1 2456220.82 2012-05-22  
## 10713 5/22/12 502 1 1 43084.81 2012-05-22  
## 10714 5/22/12 503 1 1 44031.41 2012-05-22  
## 10715 5/22/12 504 1 1 706243.12 2012-05-22  
## 10716 5/22/12 505 1 0 386673.17 2012-05-22

We also found the treatment and pre-treatment periods. The pre-treatment period (0), where both groups were shown ads, starte on 04/01/2012, and the treatment period (1), where control group continued to see ads but the treatment group did not, started on 05/22/2012. Therefore, the treatment group viewed ads for approximately 7 weeks.

**Frequency Table**

a = table(ebay$isTreatmentGroup, ebay$isTreatmentPeriod)  
t = as.data.frame(a)  
names(t)[1] = 'Treatment Group' #header for first column  
names(t)[2] = 'Treatment Period' #header for second column  
t

## Treatment Group Treatment Period Freq  
## 1 0 0 3468  
## 2 1 0 7242  
## 3 0 1 4216  
## 4 1 1 8804

prop = table(ebay$isTreatmentGroup, ebay$isTreatmentPeriod)  
prop.table(prop)

##   
## 0 1  
## 0 0.1461441 0.1776654  
## 1 0.3051833 0.3710072

prop

##   
## 0 1  
## 0 3468 4216  
## 1 7242 8804

We decided to view the categorical data (treatment period and groups) as a frequency table for clearer understanding. As shown above, we can see that 14.6% of the population is in the control group/pre-treatment period (0/0), 17.8% is in the control group/treatment period (0/1), 30.5% is in the treatment group/pre-treatment period (1/0), and finally 37.1% is in the treatment group/treatment period (1/1). We can see that there is a data disparity in the division of groups, as the treatment group is larger.

**c. Regression of Treatment Group and Revenue for pre and post treatment periods**

# Understanding the Notations:  
# Control Group (0)  
# Treatment Group (1)  
# Pre-Treatment Period (0), Ads shown to both groups  
# Treatment Period (1), Control Group shown ads, Treatment Group not shown ads  
  
ebay1<- ebay %>%  
filter(isTreatmentGroup == 1)   
ebay\_log <- lm(log(revenue) ~ isTreatmentPeriod, data = ebay1)  
summary(ebay\_log)

##   
## Call:  
## lm(formula = log(revenue) ~ isTreatmentPeriod, data = ebay1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.0038 -0.7490 -0.0274 0.6929 3.8268   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.94865 0.01472 743.988 <2e-16 \*\*\*  
## isTreatmentPeriod1 -0.03940 0.01987 -1.983 0.0474 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.252 on 16044 degrees of freedom  
## Multiple R-squared: 0.0002451, Adjusted R-squared: 0.0001828   
## F-statistic: 3.933 on 1 and 16044 DF, p-value: 0.04737

We hypothesize that not showing the ads to the Treatment Group will lead to a decrease in revenue during the treatment period.

After performing the regression, we see that the intercept for the group during the period when ads are shown is 10.94 (in dollars) and for the period when ads are not shown is (10.94 -.03) = 10.91, which is an insignificant difference.

Thus, the interpretation is that for every one ad shown in the pre-treatment period there is $10.94 increase in revenue keeping all other factors constant. The intercept of 10.94 is true at the 99% confidence level, and 10.91 at the 90% confidence level. The p-value is signifcant (p < 0.05), thus we can say that we should reject the null hypothesis and conclude that there is a decrease in revenue once the ads are not shown to the group.

**d. Randomization Test**

# Pre-Treatment Regression  
  
ebay2<- ebay %>%  
filter(isTreatmentPeriod == 0)  
ebay\_pre = lm(log(revenue) ~ isTreatmentGroup , data = ebay2)  
summary(ebay\_pre)

##   
## Call:  
## lm(formula = log(revenue) ~ isTreatmentGroup, data = ebay2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.9962 -0.7502 -0.0285 0.7331 3.8229   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.96273 0.02037 538.128 <2e-16 \*\*\*  
## isTreatmentGroup1 -0.01408 0.02477 -0.568 0.57   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.2 on 10708 degrees of freedom  
## Multiple R-squared: 3.017e-05, Adjusted R-squared: -6.322e-05   
## F-statistic: 0.323 on 1 and 10708 DF, p-value: 0.5698

NULL Hypothesis: Revenue for the Control Group is the same as for the Treatment Group during the Pre-Treatment period

NULL Hypothesis: Revenue for the Control Group is the same as for the Treatment Group during the Treatment period

The intercept slope for the Control Group for pre-treatment period is 10.96 and the intercept for the Treatment Group is nearly the same: (10.96 - 0.01) = 10.95 , thus for every ad shown to the Control Group and Treatment Group there is a $10.96 earn in sales. As we see that the p-value is insignificant (p > 0.05), we fail to reject the null hypothesis.

We also see that the R-Square value is extremely small, indicating that there is poor goodness of fit, and that the values are scattered and do not fall on the regression slope line.

**e. Post-treatment data to check the effectiveness of ads with and without DMA**

# Part I   
ebay$dma <- as.numeric(ebay$dma)  
ebay5<- ebay %>%  
filter(isTreatmentPeriod == 1)  
ebay\_dma = lm(log(revenue) ~ dma + isTreatmentGroup , data = ebay5)  
summary(ebay\_dma)

##   
## Call:  
## lm(formula = log(revenue) ~ dma + isTreatmentGroup, data = ebay5)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.6610 -0.7768 -0.0160 0.7094 3.9473   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.317517 0.069154 178.117 <2e-16 \*\*\*  
## dma -0.002185 0.000104 -21.005 <2e-16 \*\*\*  
## isTreatmentGroup1 -0.007630 0.022259 -0.343 0.732   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.188 on 13017 degrees of freedom  
## Multiple R-squared: 0.03279, Adjusted R-squared: 0.03264   
## F-statistic: 220.7 on 2 and 13017 DF, p-value: < 2.2e-16

#Part II  
ebay4<- ebay %>%  
filter(isTreatmentPeriod == 1)  
ebay\_dma1 = lm(log(revenue) ~ isTreatmentGroup , data = ebay4)  
summary(ebay\_dma1)

##   
## Call:  
## lm(formula = log(revenue) ~ isTreatmentGroup, data = ebay4)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.0038 -0.7546 -0.0288 0.7419 3.8268   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.916740 0.018610 586.595 <2e-16 \*\*\*  
## isTreatmentGroup1 -0.007494 0.022632 -0.331 0.741   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.208 on 13018 degrees of freedom  
## Multiple R-squared: 8.422e-06, Adjusted R-squared: -6.839e-05   
## F-statistic: 0.1096 on 1 and 13018 DF, p-value: 0.7406

In Part I, when we include DMA as an independent variable, we observe that the intercept for the Control Group is 12.308 (12.31 - 0.002) and for the Treatment Group the intercept is 12.301 (12.31 - 0.002 - 0.007), which indicates that there is a negligible difference when DMA is added to the regression model.

We also observe that the p-value is < 0.05, thus we fail to reject the null hypothesis. While the difference when adding DMA to the groups is insignifcant, we can extrapolate that the difference could be much more pronounced as the population increases (i.e. as customers increase).

**f. Log of revenue**

Log is taken for revenue as it reduced the skewness in the variables and the results. The minimum revenue in the data set is $367 and the maximum $2,601,390, which shows a huge range. Using a logarithmic scale is also useful when dealing with economies of scale. Finally, if we take the regression with simple revenue (without a log) the intercept value is very large and the p-value is large as well, which skews the analysis.

**g. Conclusion: Effectiveness of Advertising**

For this dataset we can conclude that advertisements can be effective in increasing sales for a company but the sample chosen for the ads should be the correct representation of the total population. From the tests in the analysis we can say that the ads were effective and customer actually got influenced in buying the product/ service. The Control group also tells us the exact sample of data and as a company we can study the consumer behaviour and profile such customers as “people with similar buying patterns”

1. Lunden, Ingrid. “Amazon’s Share of the US e-Commerce Market Is Now 49%, or 5% of All Retail Spend.” TechCrunch, TechCrunch, 13 July 2018, techcrunch.com/2018/07/13/amazons-share-of-the-us-e-commerce-market-is-now-49-or-5-of-all-retail-spend/. [↑](#footnote-ref-21)