Star Digital Causal Analysis

Hao Cheng

Background Introduction

Star Digital, a multi-channel video service provider would like to know whether it should invest more on online advertising, especially on banner advertising. Therefore, We conducted an experiment to understand the incremental impact of advertising on sales. We randomly assigned consumers into test and control groups based on exposure of ads from a charity organization and Start Digital. The goal is to analyze the effectiveness of experiment, increase purchase frequency, and find the target sites for budget management.

Experiment Design

(a) Treatment and control group

Treatment variable: whether the software places campaign ads to customers or not

Treatment group: 90% of customers who were shown Star Digital Ads

Control group: 10% of customers who were shown charity organization ads

(b) The unit of analysis

Customers viewing online advertisements

(c) Testing method

A/B testing

Threat of causal inference

1. Omitted variable bias:

The customer personal information such as gender and age might be omitted. It is likely that these factors are correlated to the final purchasing. For example, younger generation is more likely to subscribe because they addict more to social media and networks.

2. Simultaneity bias:

In some cases, not only impressions influence on purchase decision, dependent variable(purchase) can affect independent variable(impressions). For instance, consumers may be impressed more on specific sites after subscription.

3. Measurement error:

We cannot accurately count and check if users really view the ads, since some extension tools might block the ads.

4. Selection bias:

There is no evidence about which sample of customers are selected in the experiment. It is possible that consumers in the experiment are mostly low financial level and cannot afford the subscription.

Exploratory Data Analysis

This dataset includes 1 id column, 6 numerical independent variables (imp_1 ~ imp_6), 1 binary treatment variable (test), and 1 binary dependent variable (purchase).
We conduct data processing to view the statistics and check the assumption.

1. Descriptive summary

```
summary(data[3:8])
```

```
##  test          imp_1              imp_2             imp_3
##  0: 2656   Min.   :  0.0000   Min.   :  0.000   Min.   :  0.00000
##  1:22647   1st Qu.:  0.0000   1st Qu.:  0.000   1st Qu.:  0.00000
##            Median :  0.0000   Median :  0.000   Median :  0.00000
##            Mean   :  0.9309   Mean   :  3.428   Mean   :  0.09477
##            3rd Qu.:  0.0000   3rd Qu.:  2.000   3rd Qu.:  0.00000
##            Max.   :296.0000   Max.   :373.000   Max.   :148.00000
##      imp_4            imp_5
##  Min.   :  0.00   Min.   : 0.00000
##  1st Qu.:  0.00   1st Qu.: 0.00000
##  Median :  0.00   Median : 0.00000
##  Mean   :  1.59   Mean   : 0.04897
##  3rd Qu.:  0.00   3rd Qu.: 0.00000
##  Max.   :225.00   Max.   :51.00000
```

2. Check missing values

```
sum(is.na(data))
```

```
## [1] 0
```

3. Data Transformation

We combine the numbers of impressions that the consumer saw at website1 through 5, and all websites.

```r
data=data %>% mutate(imp1to5=imp_1+imp_2+imp_3+imp_4+imp_5)
data=data %>% mutate(imp_all=imp_1+imp_2+imp_3+imp_4+imp_5+imp_6)
```

Before experiments

1. Randomization Check

We conducted t.test to see whether the control and treatment groups have the similar average number of imp_1to5 and imp_6. It shows that p-values of both imp_1to5 and imp6 are larger than 0.05, which means the numbers of impression 1 to 5 and impression 6 are not different between the control and treatment groups. That is, the experiment is successfully randomized.

```r
# p-value = 0.5188 > alpha(0.05), do not reject H0.
t.test(imp1to5 ~ test,data=data)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp1to5 by test
## t = -0.071371, df = 3268.6, p-value = 0.9431
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.8402427  0.7812196
## sample estimates:
## mean in group 0 mean in group 1
##        6.065512        6.095024
```

```r
# p-value = 0.6661 > alpha(0.05), do not reject H0.
t.test(imp_6 ~ test,data=data)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp_6 by test
## t = 0.43156, df = 2898.4, p-value = 0.6661
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.3176712  0.4969729
## sample estimates:
## mean in group 0 mean in group 1
##        1.863705        1.774054
```

```r
# p-value = 0.8987 > alpha(0.05), do not reject H0.
t.test(imp_all ~ test,data=data)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp_all by test
```

```
## t = 0.12734, df = 3204.4, p-value = 0.8987
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.8658621  0.9861407
## sample estimates:
## mean in group 0 mean in group 1
##        7.929217        7.869078
```

2. Power Test

We check whether the sample size is less than or larger than the minimum required, we use alpha=0.05 and beta=0.2. If we would like to detect 0.1% change in purchase rate, we need at least 174 samples in each group. For this case, we have more than 20000 samples in treatment and more than 2000 samples in control group. Therefore, it is an overpowered study.

```
# treatment
treat<-filter(data,test==1)
p1<-mean(treat$purchase)
n1<-nrow(treat)
s1<-sqrt(p1*(1-p1)/n1)

# control
control<-filter(data,test==0)
p2<-mean(control$purchase)
n2<-nrow(control)
s2<-sqrt(p2*(1-p2)/n2)

power.t.test(delta = 0.001,sd=s1, sig.level = 0.05, type = 'two.sample', power = 0.8, alternative = 'tw
```

```
##
##      Two-sample t test power calculation
##
##               n = 174.2365
##           delta = 0.001
##              sd = 0.003322339
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Three experiments

1. The Effectiveness of Online Advertising for Star Digital

We performed t-test to check if the campaign ads (treatment) affects the purchase (dependent variables).

```
# p-value = 0.06139 > 0.05(alpha), do not reject H0
t.test(purchase~test, data = data)
```

```
##
##  Welch Two Sample t-test
##
## data:  purchase by test
## t = -1.8713, df = 3309.2, p-value = 0.06139
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.039289257  0.000916332
## sample estimates:
## mean in group 0 mean in group 1
##       0.4856928       0.5048792
```

The p-value is a small number (although slightly greater >5%), but we conclude that it is marginally significant. This implies that the test group have positive effect on purchase. Therefore, the ads are effective.

2. Relationship between Impressions and Purchase

We use simple linear regression models on the treatment group to find out whether the change in number of impressions would result in changes of purchase.

```
summary(lm(purchase ~ test*imp_all, data))
```

```
##
## Call:
## lm(formula = purchase ~ test * imp_all, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89562 -0.47994 -0.05711  0.51280  0.53228
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4651265  0.0101335  45.900  < 2e-16 ***
## test1         0.0111885  0.0107209   1.044   0.2967
## imp_all       0.0025937  0.0004131   6.278 3.49e-10 ***
## test1:imp_all 0.0010362  0.0004408   2.351   0.0188 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4942 on 25299 degrees of freedom
## Multiple R-squared:  0.02317,    Adjusted R-squared:  0.02306
## F-statistic:   200 on 3 and 25299 DF,  p-value: < 2.2e-16
```

Looking at the effect of total impressions on the odds of purchase, we see a very significant p-value(3.49e-10), much lower than 0.05. This means that there is evidence that the total number of ad impressions for each consumer effects whether they make a purchase at Star Digital or not. The coefficient of the total impression term is 0.0025937, which means that there is around 0.25% increase in purchasing at Star Digital if incrase 1 impression in the control group. This indicates that more online activity increases the frequency of purchasing at Star Digital, regardless of whether they are seeing Star Digital ads.

As for treatment group, the p-value for the interaction between being in the treatment group and total impressions is under 0.05 (0.0188). This means that there is evidence that there is difference in the effect of an additional ad impression between the treatment and control group. And the coefficient on the interaction

term is around 0.1%, indicating increasing in purchase odds for the control group, consumers in the treatment group are expected to have an additional 0.1% increase of purchasing at Star Digital.

In conclusion, it appears that a higher frequency of advertising does increase the probability of purchase.

3. Choosing between Website 6 or Websites 1 through 5

We use simple linear regression models on the treatment group to compare the average impact on site1 to site 5 and that on site 6 purchase. Then, We used ROI to make business decision. ROI= ((Value of Purchase * Increase of Purchase) - Cost of Impression) / Cost of Impression

```
summary(lm(purchase ~ test*imp1to5 , data ))
```

```
##
## Call:
## lm(formula = purchase ~ test * imp1to5, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.00529 -0.48502 -0.05804  0.51498  0.53311
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4668879  0.0100139  46.624  < 2e-16 ***
## test1         0.0142322  0.0105882   1.344    0.179
## imp1to5       0.0031003  0.0004744   6.536 6.46e-11 ***
## test1:imp1to5 0.0007978  0.0005031   1.586    0.113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4943 on 25299 degrees of freedom
## Multiple R-squared:  0.02272,    Adjusted R-squared:  0.0226
## F-statistic: 196.1 on 3 and 25299 DF,  p-value: < 2.2e-16
```

```
summary(lm(purchase ~ test*imp_6, data))
```

```
##
## Call:
## lm(formula = purchase ~ test * imp_6, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0637 -0.5020  0.3724  0.5017  0.5165
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.4834860  0.0098438  49.116   <2e-16 ***
## test1       0.0148381  0.0104279   1.423   0.1548
## imp_6       0.0011841  0.0009256   1.279   0.2008
## test1:imp_6 0.0025109  0.0010577   2.374   0.0176 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.4995 on 25299 degrees of freedom
## Multiple R-squared:  0.002258,   Adjusted R-squared:  0.00214
## F-statistic: 19.09 on 3 and 25299 DF,  p-value: 2.334e-12
```

The p-value for both [test:imp1to5] and [test:imp_6] are all smaller than 0.05, indicating that there is evidence that there is difference in the effect of an additional ad impression between the treatment and control group for "site1 to site5" and "site 6".

```
ROI_site1to5 = ((1200 * 0.0007301) - (25 / 1000)) / (25 / 1000)
ROI_site1to5
```

```
## [1] 34.0448
```

```
ROI_site6 = ((1200 * 0.0014738) - (20 / 1000)) / (20/1000)
ROI_site6
```

```
## [1] 87.428
```

In conclusion, Star Digital should put its advertising dollars in Site 6, because sit 6 has higher ROI.