Star Digital Causal Analysis

Yun-Chien, Yen

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, they conducted an experiment to understand the incremental impact of advertising on sales. They 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.

```
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)
```

4. Check outliers

We choose 0.99 percentile outlier.

```
# imp1 to imp5
quantile(data$imp1to5,c(0.9,0.95,0.97,0.98,0.99,0.995,0.999))
```

```
##     90%     95%     97%     98%     99%   99.5%   99.9%
##  13.000  28.000  43.000  59.000  90.000 134.490 261.698
```

```
outlier1 = quantile(data$imp1to5, 0.99)[[1]]
data$imp1to5<-ifelse(data$imp1to5 > outlier1, outlier1, data$imp1to5)

# imp6
quantile(data$imp_6,c(0.9,0.95,0.97,0.98,0.99,0.995,0.999))
```

```
##    90%    95%    97%    98%    99%  99.5%  99.9%
##  4.000  6.000  9.000 12.000 20.000 30.000 83.698
```

```
outlier2 = quantile(data$imp_6,0.99)[[1]]
data$im_6<-ifelse(data$imp_6 > outlier2, outlier2, data$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 0.01, 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.

```
# 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 = -0.64533, df = 3309.8, p-value = 0.5188
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.7211013  0.3639661
## sample estimates:
## mean in group 0 mean in group 1
##        5.267319        5.445887
```

```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.

```r
# 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 = 'two.sided')
```

```
##
##      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
```

we cannot conclude that the mean purchase proportion of treatment groups is higher than that of the control group. That is, we can't tell whether online advertising is significantly effective for the company or not.

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.

```
treatment = data %>% filter(test == 1)
summary(lm(purchase ~ imp_all, treatment))
```

```
##
## Call:
## lm(formula = purchase ~ imp_all, data = treatment)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.89562 -0.48357  0.05179  0.51280  0.52006
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.4763150  0.0034981  136.16   <2e-16 ***
## imp_all     0.0036299  0.0001537   23.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.494 on 22645 degrees of freedom
## Multiple R-squared:  0.02404,    Adjusted R-squared:  0.02399
## F-statistic: 557.7 on 1 and 22645 DF,  p-value: < 2.2e-16
```

The p-value of imp_all is smaller than 0.05, which means if the number of ad impressions increase 1 unit, customers will have 0.36299 more probability to 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.

```
summary(lm(purchase ~ imp1to5 , treatment))
```

```
##
## Call:
## lm(formula = purchase ~ imp1to5, data = treatment)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1055 -0.4733 -0.1055  0.5196  0.5338
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.4661979  0.0035162   132.6   <2e-16 ***
## imp1to5     0.0071029  0.0002416    29.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4907 on 22645 degrees of freedom
## Multiple R-squared:  0.03678,    Adjusted R-squared:  0.03673
## F-statistic: 864.6 on 1 and 22645 DF,  p-value: < 2.2e-16
```

```
summary(lm(purchase ~ imp_6, treatment))
```

```
##
## Call:
## lm(formula = purchase ~ imp_6, data = treatment)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0637 -0.5020  0.4019  0.5017  0.5017
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.4983241  0.0034406  144.84  < 2e-16 ***
## imp_6       0.0036950  0.0005118    7.22 5.38e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4994 on 22645 degrees of freedom
## Multiple R-squared:  0.002296,   Adjusted R-squared:  0.002252
## F-statistic: 52.12 on 1 and 22645 DF,  p-value: 5.377e-13
```

Each additional increase in the company's ad impression of websites 1 through 5 will increase purchase by 0.71029%. Each additional point increase in the company's ad impression of website 6 will increase purchase by 0.3695%.

```
# cost
cost_imp1to5 <- 25/1000
cost_imp_6 <- 20/1000

#calculate the cost
cost_1to5 <- cost_imp1to5/0.0071029
cost_1to5
```

```
## [1] 3.519689
```

```
cost_6 <- cost_imp_6/0.0036950
cost_6
```

```
## [1] 5.41272
```

For the cost of advertising on different websites for one thousand impressions, website 1 through 5's cost is $0.025 per impression, and website 6's cost is $0.02 per impressio. Thus, the cost for per increase in purchase is $3.52 for website 1 to 5, and $5.41 for website 6. The cost for websites 1 to 5 is cheaper than website 6. Hence, we will recommend Star Digital to invest more money on websites 1 to 5.

Executive Summary

We can conclude three points in the following:

1. We use t-test and find that we cannot tell significant difference from the purchase rate of the treatment group and that of the control group. Therefore, we cannot determine whether the advertisement is effective for the company.
2. There is frequency effect of advertising on purchase. User's exposure to the company's advertisement can significantly increase the chances of purchase. To be more specific, each additional ad impression will lead to a 0.36299% chance increase in purchase.
3. Star Digital should invest more on site 1 through 5 since it costs less than site 6 on budget. In order to get 1 unit increase in purchase, the company should spend $5.41 on sites 6. However, it only needs to spend $3.52 on site 1 through 5 to get the same results.