Causal Assignment 2

Ishan Kotian[5991337] and Swapnil Sharma[5955550]

2025-03-07

Executive Summary

Business Context

Star Digital, a large multichannel video service provider, spends a large portion of their budget on advertising. Star Digital's advertising approach evolved along with the technical landscape, as they started to make greater investments in online advertising, including banner adverts. They actively assess each ad medium's return on investment (RoI) to get the most of their cash. Star Digital has created a controlled experiment to assess the efficacy of an internet advertising campaign. Participants were randomised to either the treatment group, which received advertisements from Star Digital, or the control group, which received advertisements for a charity on a variety of websites using ad-serving software.

Business Question

Star Digital intends to evaluate the causal relationship between display advertising and sales conversion using the experiment's findings. They specifically want to answer three key questions:

1. Is online advertising effective for Star Digital?
2. Is there a frequency effect of advertising on purchase?
3. Which sites should Star Digital advertise on?

Analysis Performed

Using the test group as the regressor and conversion as the answer, we conducted a logistic regression to investigate the impact of online advertising on Star Digital purchases in relation to the first question. Using a logistic regression model that added total ad impressions as an additional regressor and its interaction, we calculated the effect of an increase in ad impressions on purchase for the second question by summing up all of the impressions from the six sites. We utilised logistic regression to quantify the impact of ad impressions from particular sites on purchasing decisions, and we created a return on investment (RoI) metric to compare the efficacy of sites 1 through 5 and site 6. We add replace total impressions with impressions from sites 1 to 5 and 6 and their interaction as regressors. We calculate RoI using our estimated causal impact per impression on purchase to determine the most profitable site decision for Star Digital.

Main Take-aways

A. We lack sufficient data to draw the conclusion that Star Digital's online advertising is successful. Customers' chances of making a purchase at Star Digital are not substantially impacted by whether they are in the treatment or control group.

B. Customers' decision to buy from Star Digital is heavily influenced by the overall number of online ad impressions they receive. Both the treatment and control groups' chances of making a purchase at Star Digital rise with each ad impression, but the difference is much greater for those who viewed Star Digital advertisements. This suggests that people who use the internet more frequently are more likely to buy something overall, and that viewing advertisements from Star Digital also raises that likelihood.

C. Regardless of test group, the overall quantity of impressions from sites 1 through 5 significantly influences the decision to buy, whereas site 6 impressions had no discernible impact. However, the group exposed to Star Digital advertisements had a far higher chance of making a purchase than the control group for every additional impression on sites 1 through 5 or site 6.

D. We advise Star Digital to allocate its advertising money to site 6 instead of sites 1–5 based on the return on investment.

Pre-Analysis

```
library(dplyr)
library(ggplot2)
library(pwr)
star = read.csv("starDigital.csv")
```

First, we must load the necessary packages and read the dataset into a dataframe.

Data Description

Before running any analyses, we will first seek to understand the data a little bit better.

```
table(star$purchase)
```

```
##
##     0     1
## 12579 12724
```

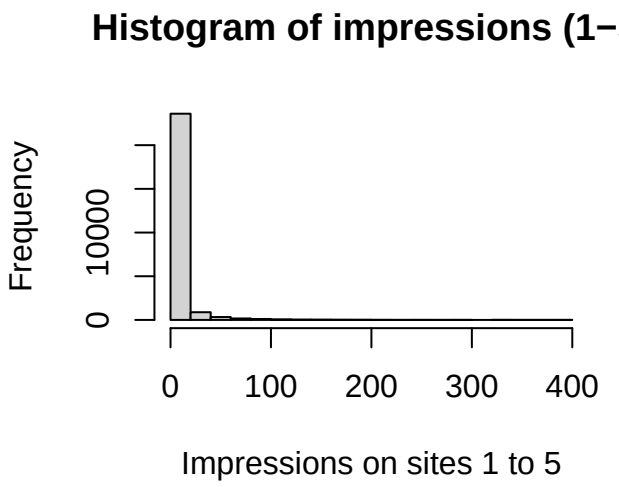
```
table(star[star$test == 0,]$purchase)
```

```
##
##    0    1
## 1366 1290
```

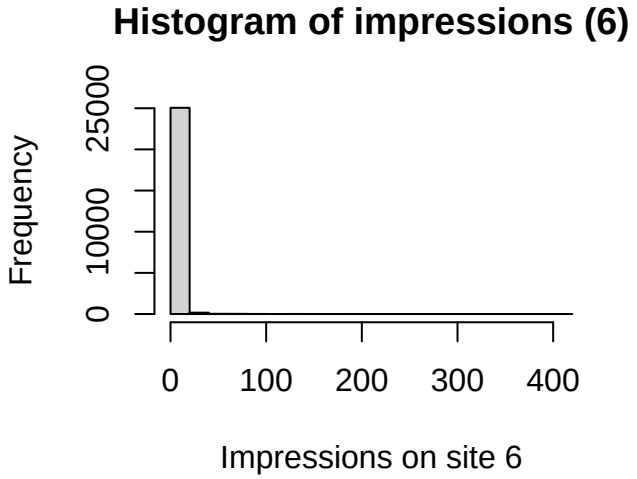
```
table(star[star$test == 1,]$purchase)
```

```
##
##     0     1
## 11213 11434
```

```
hist(star$sum1to5, xlab = 'Impressions on sites 1 to 5',
     main = 'Histogram of impressions (1-5)')
```




```
hist(star$imp_6, xlab = 'Impressions on site 6', main = 'Histogram of impressions (6)')
```



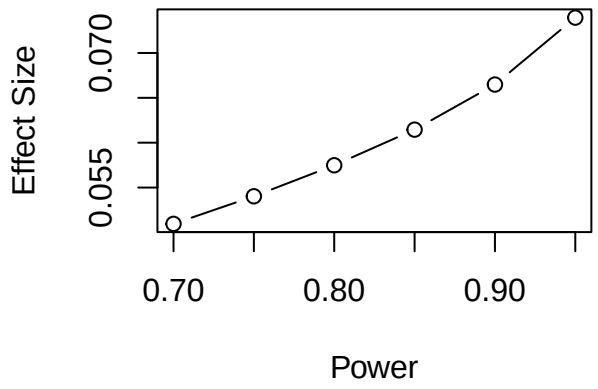
There appears to be a reasonably equal distribution of the outcome variable between purchases and non-purchases. With roughly 50% of each group making a purchase, it also seems that the treatment and control groups' purchasing decisions are equally divided. Furthermore, as the case explains, around 90% of the data received Star Digital advertisements, and approximately 10% received control charity advertising. According to the impressions histograms, the majority of users have comparatively few impressions on sites 1 through 5 and site 6, which are both highly skewed to the right.

Sample Size Analysis

We will now talk about how our sample size affects the inferences we can draw. We cannot perform the same power test that was covered in class because our sample does not meet the presumption that the treatment and control groups are of equal size because their sizes are not even. The pwr package contains a test for uneven groups that we will use instead. The probability of not detecting an effect that exists is 1 minus power. A higher power hence lessens the likelihood that a true effect would go unnoticed.

```r
ptab2 = cbind(NULL, NULL)
for (i in c(.7, .75, .8, .85, .9, .95)){
  pwrt = pwr.t2n.test(n1 = 2656, n2 = 22647, sig.level = .05, power = i,
                      alternative = "two.sided")
  ptab2 = rbind(ptab2, cbind(pwrt$power, pwrt$d))
}

plot(ptab2[,1], ptab2[,2], type = "b", xlab = "Power", ylab = "Effect Size")
```



The graph above illustrates how the amount of power and the smallest effect size that can be consistently observed increase with sample size and a standard significance threshold of 0.05. In other words, a greater effect is required to be reliably recognised at higher power levels.

Randomization Check

To ensure that the groups are approximately homogeneous outside of the therapy and that the randomisation between the treatment and control groups was adequate, we will lastly run t-tests. The results of our analysis might not be trustworthy if there are substantial differences between the groups.

```r
t.test(imp_1 ~ test, data = star)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp_1 by test
## t = -3.905, df = 3574.1, p-value = 9.596e-05
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.5961772 -0.1976257
## sample estimates:
## mean in group 0 mean in group 1
##       0.5756777       0.9725791
```

```r
# Output of tests for sites 2 to 5 are hidden to reduce redundancy as results are
# similar to that of site 1 (sites 3/5 significant, sites 2/4 insignificant)
invisible(t.test(imp_2 ~ test, data = star))
invisible(t.test(imp_3 ~ test, data = star))
invisible(t.test(imp_4 ~ test, data = star))
invisible(t.test(imp_5 ~ test, data = star))
t.test(imp_6 ~ test, data = star)
```

```
##
##  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
t.test(sum1to5 ~ test, data = star)
```

```
##
##  Welch Two Sample t-test
##
## data:  sum1to5 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
```

There appears to be a statistically significant difference between the treatment and control groups when we look at webpages 1 through 5 alone. This suggests that there may be differences in the demographics of the two groups. Due to their interchangeability and Star Digital's incapacity to display ads on a particular site, there is insufficient data to conclude that sites 1 through 5 differ considerably in terms of the overall number of impressions seen. Similarly, there is insufficient data to conclude that the treatment and control groups are not homogeneous with regard to their impressions on site 6.

Main Analysis

We can now address Star Digital's three primary questions as we have a better understanding of the data we will be analysing.

Question 1:

Is online advertising effective for Star Digital?

```r
lm1 = glm(purchase ~ test, data = star, family = 'binomial')
summary(lm1)
```

```
##
## Call:
## glm(formula = purchase ~ test, family = "binomial", data = star)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.05724    0.03882  -1.474   0.1404
## test         0.07676    0.04104   1.871   0.0614 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35077  on 25302  degrees of freedom
## Residual deviance: 35073  on 25301  degrees of freedom
## AIC: 35077
##
## Number of Fisher Scoring iterations: 3
```

```r
exp(coef(lm1))
```

```
## (Intercept)        test
##   0.9443631   1.0797852
```

```r
# This transformation of fitted model coefficients is done here and in further analyses
# so that they can be interpreted as the odds ratio.
```

Interpretations

First, we will use logistic regression to determine whether being in the group that receives Star Digital ads has a significant increase in the odds of purchase. In the regression, the p-value for 'test' is 0.0614, which is greater than the acceptable significance level 0.05. This means there is not enough evidence to say whether the consumer is in the treatment or control group has an effect on whether he/she will eventually make a purchase at Star Digital. Hence, we do not have enough evidence to conclude that simply displaying online advertising is effective in stimulating purchases. Although we cannot confidently say that it is different from 0, we estimate that being part of the test group would increase the odds of purchasing Star Digital by 7.98%.

Question 2:

Is there a frequency effect of advertising on purchase? In particular, the question is whether increasing the frequency of advertising increases the probability of purchase?

```r
star$total_imp = rowSums(star[,4:9])

lm2 = glm(purchase ~ test + total_imp + test*total_imp, data = star, family = 'binomial')
summary(lm2)
```

```
##
## Call:
## glm(formula = purchase ~ test + total_imp + test * total_imp,
##     family = "binomial", data = star)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.169577   0.042895  -3.953 7.71e-05 ***
## test           -0.013903   0.045613  -0.305    0.761
## total_imp       0.015889   0.002876   5.524 3.32e-08 ***
## test:total_imp  0.015466   0.003207   4.823 1.42e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35077  on 25302  degrees of freedom
## Residual deviance: 34190  on 25299  degrees of freedom
## AIC: 34198
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(lm2))
```

```
##    (Intercept)           test      total_imp test:total_imp
##      0.8440214      0.9861932      1.0160156      1.0155865
```

Interpretations

We start by creating a new variable that adds up all of the website impressions in order to respond to this query. We must now ascertain how ad frequency affects purchases. We'll employ logistic regression once more. Our findings demonstrate that exposure to the Star Digital advertisements alone does not significantly boost sales. This bolsters the conclusion we came to in response to Question 1.

When examining how total impressions affect the likelihood of a purchase, we find a highly significant p-value that is significantly less than 0.05. This indicates that there is proof that a customer's overall ad impressions have an impact on whether or not they buy from Star Digital. The probability of making a purchase at Star Digital rise by approximately 1.6% for every additional ad impression in the control group, according to the coefficent of the total impression term, which is 0.0159. This suggests that whether or not people are viewing Star Digital advertisements, increased online engagement raises the likelihood of making a purchase from Star Digital.

Now that we are concentrating on the treatment group, we can observe that the p-value for the interaction between total impressions and being in the treatment group is significantly less than 0.05. This suggests that the treatment and control groups' responses to an extra ad impression differed significantly. Customers in the treatment group are anticipated to have an extra 1.5% rise in their chances of making a purchase from Star Digital for every ad impression, on top of the 1.6% increase in buy probabilities for the control group, according to the coefficient on the interaction term.

As we can see, it seems that the likelihood of a purchase is increased by more frequent promotion.

Question 3:

Which sites should Star Digital advertise on? In particular, should it put its advertising dollars in site 6 or in sites 1 through 5?

In order to assess the efficacy of sites 1 through 5 and site 6, we must first create a return on investment (ROI) statistic. To do this, we evaluated the profitability of several locations using the following formula as our criterion:

ROI = ((Value of Purchase * Increase in Odds of Purchase) - Cost of Impression) / Cost of Impression

```
lm3 = glm(purchase ~ test + sum1to5 + imp_6 + test*sum1to5 + test*imp_6,
          data = star, family = 'binomial')
summary(lm3)
```

```
##
## Call:
## glm(formula = purchase ~ test + sum1to5 + imp_6 + test * sum1to5 +
##     test * imp_6, family = "binomial", data = star)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.166556   0.042533  -3.916 9.01e-05 ***
## test         -0.006087   0.045314  -0.134 0.893139
## sum1to5       0.019452   0.003443   5.650 1.61e-08 ***
## imp_6         0.003978   0.004294   0.927 0.354179
## test:sum1to5  0.014617   0.003794   3.852 0.000117 ***
## test:imp_6    0.013483   0.005405   2.494 0.012616 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35077  on 25302  degrees of freedom
## Residual deviance: 34166  on 25297  degrees of freedom
## AIC: 34178
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(lm3))
```

```
##  (Intercept)         test      sum1to5        imp_6 test:sum1to5   test:imp_6
##    0.8465755    0.9939313    1.0196422    1.0039860    1.0147240    1.0135741
```

Interpretations

Our findings from Question 1 that simply existing in the treatment or control group has no discernible impact on the likelihood of making a purchase are supported by our final regression summary, which uses logistic regression and evaluates the impact of advertisements on certain websites.

The odds of a purchase are significantly impacted by the total impressions from sites 1 through 5, but there is insufficient data to conclude that the amount of impressions from site 6 has a significant impact at the 95% confidence level. More precisely, we calculate that the likelihood of making a purchase from Star Digital rises by 1.96% for every extra impression on sites 1 through 5. However, the projected improvement in the probability of purchase is only 0.40% for every additional impression on site 6.

Furthermore, we have evidence that the treatment group's purchase behaviour changes differently with more impressions than the treatment group, as indicated by the extremely low p-value for the interaction terms of being in the treatment group and the total number of impressions on each group of sites. For every

extra impression from sites 1 through 5, the treatment group's purchase probabilities are predicted to rise by 1.47%, which is higher than the control group's anticipated increase of 1.96%. For every extra site 6 ad, the treatment group's probabilities of making a purchase are predicted to rise by 1.36% more than those of the control group.

Advertisements on sites 1 through 5 cost $25 per 1000 impressions, whereas advertisements on site 6 cost $20 per 1000 impressions. Thus, the following formula is used to determine our ROI:

```
ROI_site1to5 = ((1200 * .0147240) - (25 / 1000)) / (25 / 1000)
ROI_site1to5
```

```
## [1] 705.752
```

```
ROI_site6 = ((1200 * .0135741) - (20 / 1000)) / (20/1000)
ROI_site6
```

```
## [1] 813.446
```

As we can see, there is a higher ROI in investing in ads on site 6 and we conclude that Star Digital should put its advertising dollars in site 6.

Concerns and Limitations

Here we discuss key threats and limitations that could affect the results and conclusions.

Threats to Causal Inference

Selection Bias: The sample may not represent the broader customer base. Since the analysis uses a choice-based sample (with about half making purchases), interpretations like purchase odds and RoI could be skewed. A transformation is needed to generalize findings to the population, as the true conversion rate was much lower than the 50% sample rate.

Neglecting Variable Bias: The lack of demographic data raises concerns. If factors like age or gender are linked to both subscription likelihood and website engagement, the experiment could be biased. Randomization should control for this, but differences between treatment and control groups suggest some unaccounted factors.

Simultaneity Bias: Since Star Digital assigned treatment and control, and purchase decisions shouldn't affect ad frequency, simultaneity isn't a concern.

Measurement Error: The impressions measurement might be unreliable, as we don't know how long ads were shown or if users saw them. Additionally, ads blocked by software might still be counted, leading to errors.

Limitations

The experimental design has several limitations that affect the conclusions. It's unclear whether participants knew they were part of the experiment, which could alter their behavior. There's also a risk of interference if participants share content online, making causal estimates less precise. Since the study evaluates a single ad campaign, participants may have been exposed to other Star Digital ads through different channels, potentially influencing results. Additionally, viewing Star Digital's ads might drive purchases from competing companies, suggesting market share could be a valuable metric. Finally, there's no information on the timing of conversions, and purchases occurring long after ad exposure may be considered view-through conversions.

Business Recommendations

- Refocus Advertising Strategy: The study suggests that simple exposure to ads does not significantly drive purchases, making personalized targeting and interactive formats a viable alternative.
- Increase Ad Frequency: There is evidence that ad impressions positively impact purchase likelihood, so optimizing ad delivery is a logical step.
- Strategic Site Selection: Since Site 6 yields a better ROI with a lower cost per impression ($20 per thousand vs. $25 for other sites), reallocating the budget there is recommended.
- Enhanced Data Collection: Since demographics and behavioral factors might influence purchases, integrating advanced profiling can improve targeting accuracy.