Star Digital Case

Group 16

Andrew Gillson, Hengzhen Chang, Hui-Lun Kuo, Siva Kallur, Yuqing Zhang

February 12, 2019

I. Executive Summary

1) Situation/Background:

Star Digital is a large online video service provider. They have steadily increased their online advertising budget and want to know how effective their online advertisements are to increasing sales. Star Digital is conducting an experiment to determine if online advertising causes sales to increase.

2) Complication:

Star digital does not know if advertisements are effective or what the best websites are to place the advertisements. There are two key questions. The first key question is whether online advertising is effective and if so, how much? The second question is which websites are the best place to place ads.

3) Key Takeaways:

Our group has determined three conclusions from our study. The first is that online advertisements increases purchases by 2%. The second conclusion is that increasing exposure to the advertisement increases likelihood of purchase. The final conclusion is that Start Digital should post its advertisements on sites 1 through 5.

II. Data Exploration

1) Variable Exploration and Construction

In this data set, we have the following attributes: the unique id of each consumer, whether or not each consumer made purchase eventually, which group the consumers were assigned, the number of ad impressions each consumer saw in each of the 6 channels and the total number of ad impressions each consumer saw in 1 through 5 channels. For the further analysis, we created another variable, the total number of ad impressions each consumer saw in all 6 channels (details are shown in Appendix – Step 2: Data manipulation).

2) Log Transformation

After looking at the distribution of the total impression number, we found that the total number of impression is heavily centered on the left side of the histogram (as shown in the graph 1 in Appendix: Number of impressions). The distribution seems not to be linear. When we are examing the correlation between the number of impressions and the probability of purchase, for statistical reasons, we need to make the variable linear in order to apply the linear regression. Therefore, we followed the industrial convention to take the log of the total impressions. After the log transformation, the distribution tends to become more linear, as shown in graph 2 (Log of the number of impressions) in the Appendix.

3) Missing Value Check

There is no missing values in the data. We have 25303 unique consumers in this data. Among all the consumers, nearly 12 percent were placed in the control group, who saw the charity ads, and the other 88 percent were placed in the test group, who saw company's campaign ads (as shown in graph 3 in the Appendix).

4) Randomization Check

Before interpreting the experiment result, we want to make sure that there is no systematic error related to the experimental design, we conducted a randomization check on the sample. We believe that the number of impression is indicative of user watching habit. A user who generally pays more attention to advertisements will receive a greater number of impressions compared to a user who ignores advertisement. As a result, the effect of a same advertisement will be different on these two subjects. By running three t-test of total number of impressions from website 1-6, 1-5 and only 6 on test and control groups, we were checking whether unobserved confounds such as user watching habit might be a concern when we interpret the results of the experiment. According to our test result, user watching habits are not different across test and control groups in the three tests. Therefore, we can conclude that the selection of test and control groups is randomized (as shown in Appendix – Step 3: Randomization Check).

III. Resolution

1) Online advertising is effective for Star Digital

The experiment results indicated that the online advertising contributes to Star Digital's sales by boosting customers' likelihood of purchase.

To determine whether online advertising is effective or not, we ran a statistical t-test of purchase on the test and control group (details of the t-test are shown in Appendix – Step 4: Discover the advertising effect). The t-test told us whether there is a significant difference in probabilities of purchase between two groups.

According to the t-test result, the purchase probability in test group is 50.5%, whereas the purchase probability in control group is 48.6%. The difference between those two probabilities is 2% approximately. A p-value of 0.06 is shown in the test result. It told us that the t-test result is statistically significant and there is only around 6.1% probability of making the wrong conclusion.

In conclusion, the t-test result told us that if shown the advertising, users will be 2% more likely to make purchases.

2) Increasing the frequency improves probability of purchase

Increasing the frequency of advertising helps to increase customer's probability of purchase. Frequency of advertising can be reflected by the number of impressions. According to the experiment result, an 1% increase in impressions is going to increase the possibility of purchase by 0.13%.

To discover whether increasing the frequency of advertising increases the probability of purchase, we fitted a generalized linear regression of purchase probabilities on the log of the total number of impressions and whether the user belongs to test group or control group (details of the generalized linear regression are shown in the Appendix – Step 6: Discover effects of increasing advertising frequency). As mentioned in Log Transformation part of the Data Exploration section, we took the log of the total impressions in order to make the distribution of the variables applicabe for the generalized linear regression.

The fitted generalized linear regression result illustrated that for users who are shown the charity advertising, a 1% increase in impression is going to increase the possibility of purchase by 0.11%. For users who are shown Star Digital advertising, a 1% increase in impression is going to increase the possibility of purchase by 0.13%.

There is a correlation between the number of impressions and purchase regardless if the consumer saw the actual advertisement or the control adversistment. This suggests that there is a positive relationship between internet use and purchase. This is intuitive; the more some is online the more online service they are likely to purchase. We can see from the difference in the slope of the test and control group that consumers that viewed the real advertisement have a higher likelihood of purchase (shown in the graph 4 of the Appendix: Test group purchase slope vs Control Group purchase slope).

In the Impressions and Average Purchase plot we can see that impact of the number of times a customer sees an ad and purchases the service stabilizes around 18 views. 95% of consumers saw either the actual advertisement or the control advertisement less than 30 times (shown in the graph 5 of the Appendix: Impressions and Average Purchase).

We also saw that most users only see the impression 5 times or less. We recommend increasing average frequency up to 18 impressions per viewer to maximize purchases. A cost/benefit analysis will have to be done to determine the optimal number of impressions a user is exposed to.

3) Sites 1 through 5 are more attractive options for Star Digital

From the above results, we understand that advertising is a better option for Star Digital and with increasing frequency of number of impressions, the likelihood of purchase is increasing. Now the question arises which site is more profitable for Star Digital? Should Star Digital put its advertising dollars in site 1 through 5 or in site 6?

To answer this we fitted a generalized linear regression of purchase probabilities on the log of the total number of impressions from sites 1 through 5 and site 6 along with whether the user belongs to test group or control group.

The fitted generalized linear regression result (shown in Appendix – Step 7: Identifying the right option) illustrated that for users who are shown the Star Digital advertising, a 1% increase in impressions in sites 1 through 5 is going to increase the possibility of purchase by 0.117% (refer Appendix 1 for interpretation reference) and a 1% increase in impressions in site 6 will increase the possibility of purchase by 0.029%.

A purchase results in a contribution of $1,200 for star digital. So a 1% increase in impressions costs $0.25 and $0.20 in sites 1 through 5 and site 6 which is likely to generate $1.41 (0.117% of $1,200) and $0.35 (0.029% of $1,200) respectively which makes advertising in sites 1 through 5 more attractive option for Start Digital.

IV. Limitations & Concerns

There are two concerns we have about the experience. The first concern is the effect on purchases with low impressions. The second is that the study is underpowered. Based on our experiment data, there're 8700 subjects that have zero impressions from website 1 to website 5; and over 80 percent of the samples from overall websites (website 1 to website 6) have less than 5 impressions. This might be a potential problem that we're making conclusions based on very low impressions, which might not, in reality, indicative to the impact on purchases. Also, the other problem is that from website 1 to website 5, about 8700 subjects have never viewed any of the ads, whilst taking into consideration of website 6 there's no such cases. This implied potential bias between website 1~5 compared to website 6 - website 6 has more impressions generated at the first place.

We are also concerned with the test's power. With the differences between the average purchases between the two groups, 0.505 and 0.486 in this case, we conducted a power test to see if the sample size is valid in making our conclusion.

The statistical power in an experiment is how likely the experiment is to distinguish an actual effect from one by chance. It's the likelihood that the test is correctly rejecting the null hypothesis when there is an effect there to be detected. We used conventional power value 0.8 and significance level (alpha) 0.1 in the analysis (details of the statistical power analysis are shown in Appendix – Step 5: Statistical Power Analysis).

The power analysis told us that if we want to discover a 0.019 difference in the average purchase probability given a significant level of 0.1 and a power value of 0.8, we need a sample size of 34253 in the experiment. However, we had only 2656 observations in the control group and 22647 observations in the test group in the experiment. The sample size in the actual experiment is smaller than the required sample size according to the power analysis result. Therefore, the experiment was underpowered, which might miss a real effect on purchase by not taking enough data, or fail to notice an important side-effect.

V. Appendix

Preparation

1) Install packages

```r
suppressPackageStartupMessages({
  library(plyr) # For descriptive visualization
  library(MESS) # For statistical power test
})
```

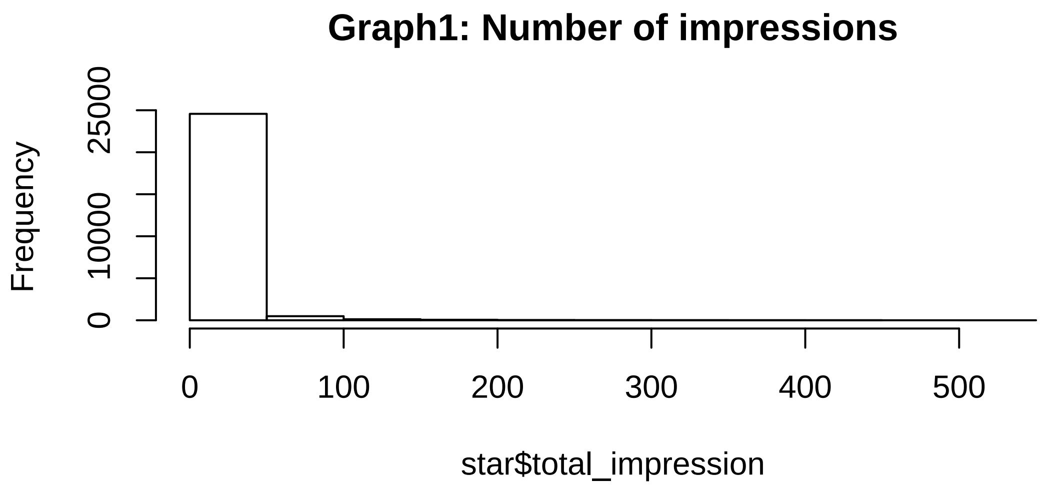
2) Load the data

```r
star <- read.csv("starDigital.csv")
```

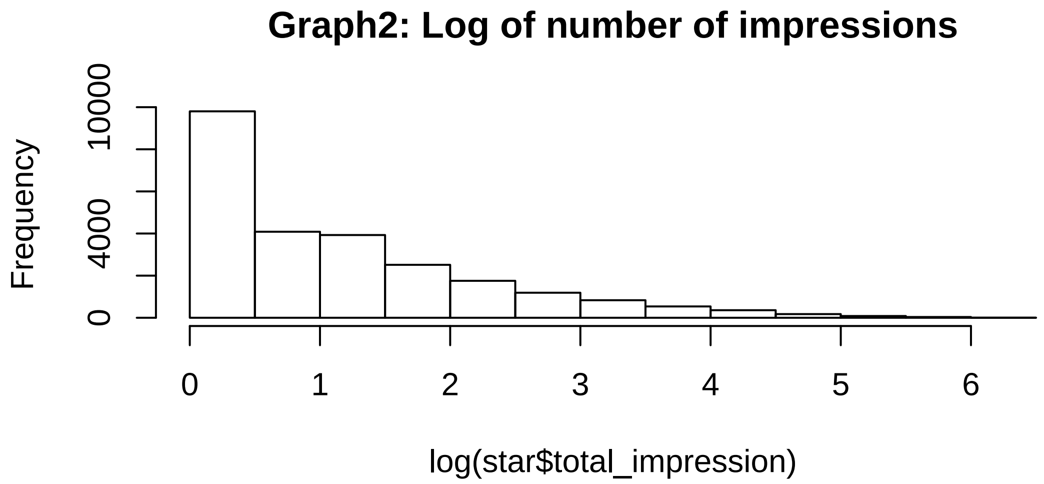
Step 1: Basic visualization and summary statistic

```r
# Construct a new column: Total number of ad Impressions
star$total_impression <- star$sum1to5 + star$imp_6
```

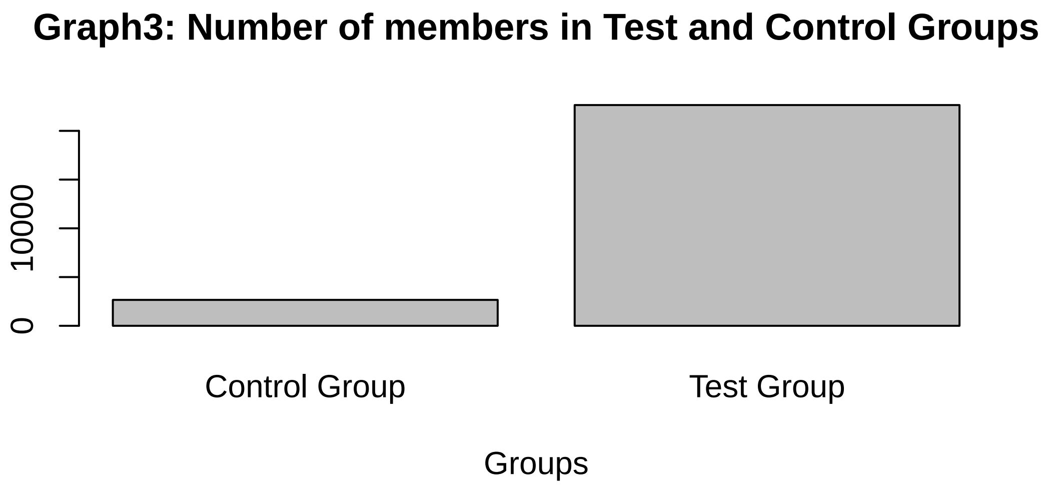
```r
hist(star$total_impression,main = "Graph1: Number of impressions")
```



```r
hist(log(star$total_impression),main = "Graph2: Log of number of impressions")
```



```
barplot(table(star$test),
        main = "Graph3: Number of members in Test and Control Groups",
        xlab="Groups",
        names.arg=c("Control Group", "Test Group"))
```



Step 2: Data manipulation

```
# Construct a new column: Total Impression
# star$total_impression <- star$sum1to5 + star$imp_6
```

Step 3: Randomization check

```
t.test(data=star,total_impression~test)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  total_impression by test
## t = 0.12734, df = 3204.4, p-value = 0.8987
## alternative hypothesis: true difference in means 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
```

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

```
##
##  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 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
```

```
t.test(data=star,imp_6~test)
```

```
##
##  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 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
```

Step 4: Discover the advertising effect

```
t.test(data=star,purchase~test)
```

```
##
##  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 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
```

Step 5. Statistical power analysis

```
power_t_test(n=NULL,type=c("two.sample"),
             alternative="two.sided",
             power=0.8,
             sig.level=0.1,
             delta=0.505-0.486)
```

```
##
##      Two-sample t test power calculation
##
##               n = 34253.07
##           delta = 0.019
##              sd = 1
##       sig.level = 0.1
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Step 6: Discover effects of increasing advertising frequency

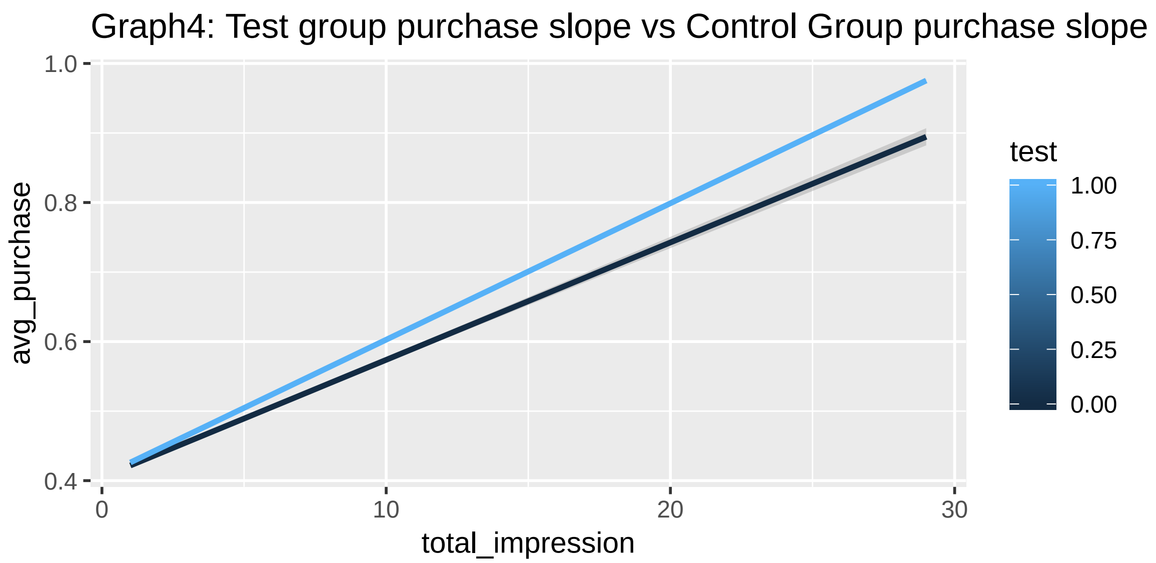
```
summary(glm(data=star,purchase~test*log(total_impression + 1)))
```

```
##
## Call:
## glm(formula = purchase ~ test * log(total_impression + 1), data = star)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.09442  -0.45599   0.02381   0.50632   0.59761
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.325299   0.017149  18.969   <2e-16 ***
## test                           -0.013256   0.018148  -0.730   0.4651
## log(total_impression + 1)       0.111224   0.009944  11.185   <2e-16 ***
## test:log(total_impression + 1)  0.019800   0.010506   1.885   0.0595 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.234956)
##
##     Null deviance: 6325.5  on 25302  degrees of freedom
## Residual deviance: 5944.2  on 25299  degrees of freedom
## AIC: 35165
##
## Number of Fisher Scoring iterations: 2
```

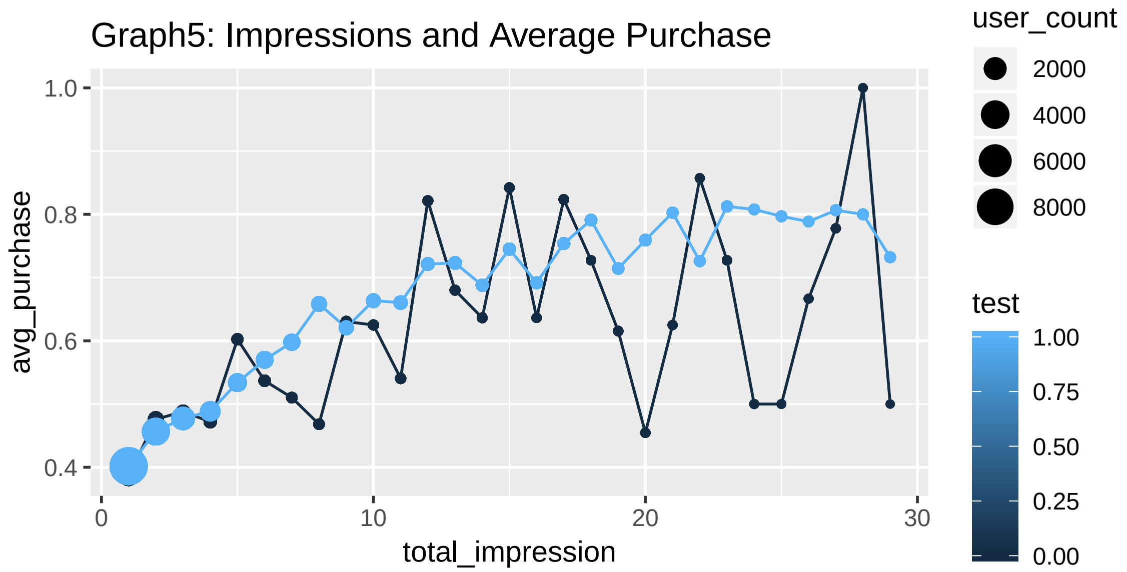
```
# install packages
suppressPackageStartupMessages({
  library(dplyr) # For descriptive visualization
  library(MESS) # For statistical power test
  library(ggplot2)
})

star_grouped <- star %>% filter(total_impression < 30) %>%
  group_by(test, total_impression) %>%
  mutate(avg_purchase = mean(purchase)) %>% mutate(user_count = n())

## Regression plot of impressions impact of purchase
ggplot(data = star_grouped, aes(x = total_impression, y = avg_purchase, color = test)) +
  geom_smooth(aes(group = test),method="lm") +
  ggtitle('Graph4: Test group purchase slope vs Control Group purchase slope')
```



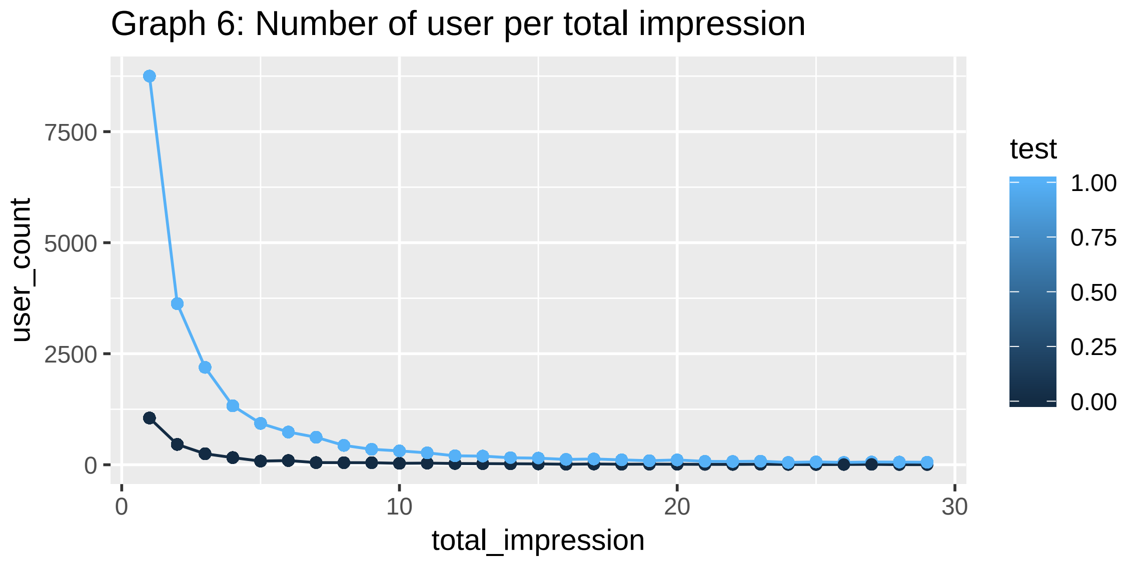
```
## plot of total impression and average purchase
ggplot(data = star_grouped, aes(x = total_impression, y = avg_purchase, color = test)) +
  geom_line(aes(group = test)) + geom_point(aes(size = user_count)) +
  ggtitle('Graph5: Impressions and Average Purchase')
```



```
### plot of total impression count of people

ggplot(data = star_grouped, aes(x = total_impression, y = user_count, color = test)) +
  geom_line(aes(group = test)) + geom_point() +
```

```r
  ggtitle('Graph 6: Number of user per total impression')
```



Step 7: Identifying the right option (channel options sites 1to5 or site 6)

```r
summary(lm(data=star, purchase ~ test*(sum1to5+imp_6)))
```

```
##
## Call:
## lm(formula = purchase ~ test * (sum1to5 + imp_6), data = star)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96643 -0.48127 -0.06395  0.51493  0.53375
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4653462  0.0101323  45.927  < 2e-16 ***
## test         0.0121117  0.0107286   1.129    0.259
## sum1to5      0.0030780  0.0004747   6.484 9.11e-11 ***
## imp_6        0.0008997  0.0009167   0.982    0.326
## test:sum1to5 0.0007301  0.0005037   1.449    0.147
## test:imp_6   0.0014738  0.0010489   1.405    0.160
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4941 on 25297 degrees of freedom
## Multiple R-squared:  0.02359,    Adjusted R-squared:  0.0234
## F-statistic: 122.3 on 5 and 25297 DF,  p-value: < 2.2e-16
```

```r
summary(lm(data=star, purchase ~ test*(log(sum1to5+1)+log(imp_6+1))))
```

```
##
## Call:
## lm(formula = purchase ~ test * (log(sum1to5 + 1) + log(imp_6 +
##     1)), data = star)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10809 -0.44782 -0.00085  0.50422  0.61802
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.372196   0.015671  23.751   <2e-16 ***
## test                  -0.006372   0.016579  -0.384    0.701
## log(sum1to5 + 1)       0.103791   0.008695  11.936   <2e-16 ***
## log(imp_6 + 1)         0.014120   0.013252   1.065    0.287
## test:log(sum1to5 + 1)  0.014502   0.009175   1.581    0.114
## test:log(imp_6 + 1)    0.015267   0.013995   1.091    0.275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4833 on 25297 degrees of freedom
## Multiple R-squared:  0.06576,    Adjusted R-squared:  0.06558
## F-statistic: 356.1 on 5 and 25297 DF,  p-value: < 2.2e-16
```