MSBX-5310: Assignment 3 - Solution

due: 2/11/19

Overview

Instructions

1. Due date: Feb 11th 2019, 8:00 AM.
2. Your task is to fill in all R code blocks that currently contain "#TBD" comments. Similarly, insert text responses wherever you see *TBD* in the markdown file.
3. PLEASE UNCOMMENT LINE 2 AND ADD YOUR NAME

Consulting project description

Your task is to evaluate the effectiveness of an online display ad campaign.

You have data from an experiment designed to measure the effectiveness of an online display advertising campaign. The experiment involves randomly assigning Internet users to a test or a control group based on cookies that uniquely identify each user visiting a site where the ad exchange (Rocket Fuel) can place an ad. Users in the test group see an ad for a newly released handbag by TaskaBella, Rocket Fuel's client. Users in the control group are shown a public service announcement that is unrelated to the advertised product. Based on the unique IDs, Rocket Fuel is able to track which users eventually purchased a handbag from TaskaBella, allowing the analyst to discern the effectiveness of the campaign.

Each row in the CSV file data set (*rocketfuel_data.csv*) represents a uniquely identified user in the ad campaign. For each user, the following six variables are provided:

`user_id`	Unique identifier of the user
`test`	1 if the user was exposed to the real ad
	0 if the user was in the control group and was shown a PSA
`converted`	1 if the user made a purchase, 0 otherwise
`tot_impr`	Total number of ad impressions the user encountered (treat=ad, control=PSA)
`mode_impr_day`	Day of the week on which the user encountered the most impressions (1=Mon,. . . ,7=Sun)
`mode_impr_hour`	Hour of the day (0-23) in which the user encountered the most impressions

For these data, `converted` is the outcome, and `test` is the treatment indicator. `user_id` uniquely identifies users (and rows). The remaining variables provide additional information observed during the experiment.

The client firm TaskaBella estimates that a conversion generates approximately $40 in incremental profit for the firm. The cost to serve ads in the experiment was $9 CPM ($9 per 1000 impressions).

Project (homework) workflow

1. Exploratory analysis
2. Randomization checks
3. ATE estimation
4. ROI calculation

1) Exploratory analysis

Read the data into R and perform some exploratory analysis. Show your work in the R chunks below, and provide text answers following the R chunk.

1.1) How many users are in the test and control conditions? (0.5 points)

```
# load data
DF = read.csv("rocketfuel_data.csv")

# users in control, treatment (test) conditions
sum(DF$test == 0)
```

```
[1] 23524
```

```
sum(DF$test == 1)
```

```
[1] 564577
```

There are 23524 users in the control condition and 564577 users in the test condition.

1.2) Conversion rates

The conversion rate for a group is the fraction of users that purchase (`converted`==1) in that group.

What is the conversion rate (in percentage) for the – a) test group and b) control group? (0.5 points)

```
sum(DF$converted[DF$test == 0])
```

```
[1] 420
```

```
sum(DF$converted[DF$test == 1])
```

```
[1] 14423
```

```
100*sum(DF$converted[DF$test == 0])/sum(DF$test == 0)
```

```
[1] 1.785411
```

```
100*sum(DF$converted[DF$test == 1])/sum(DF$test == 1)
```

```
[1] 2.554656
```

The number of people converted in the control and test conditions is 420 and 14423 respectively. Using the total number of users assigned to these two groups (from 1.1), we can calculate the conversions rates as – a) test group: 420/23524 = 1.79%, and b) control group: 14423/564577 = 2.55%

2) Randomization checks

Verify that Rocketfuel implemented the randomization correctly by examining whether the distributions of the variable `tot_impr` for the test and control groups are the same. If the average number of impressions that users see in each group is different, then the differences in their response rate can be (potentially) attributed to this instead of the ads that they see. We can examine the distribution of `tot_impr`for the two groups in three ways: (simple) mean comparison, distribution (histogram) comparison, and formal difference in means t-tests.

2.1) Mean comparison

Using the describe command, summarize `tot_impr` for two the groups of users (in test and control conditions). What is the mean of this variable each of these groups? Are the means similar? (0.5 points)

```
library(psych)
describe(DF$tot_impr[DF$test==0])
```

```
   vars     n  mean    sd median trimmed   mad min max range skew kurtosis
X1    1 23524 24.76 42.86     12   15.82 14.83   1 907   906 5.65    53.35
     se
X1 0.28
```

```
describe(DF$tot_impr[DF$test==1])
```

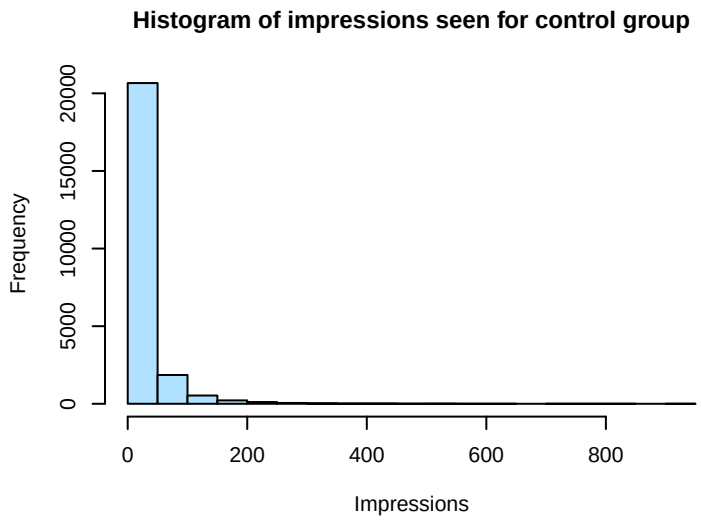
```
   vars      n  mean    sd median trimmed   mad min  max range skew
X1    1 564577 24.82 43.75     13   16.29 14.83   1 2065  2064  7.5
   kurtosis   se
X1   112.08 0.06
```

The mean of tot_impr in the control condition is 24.76 and 24.82 in the test condition. These numbers look pretty similar.

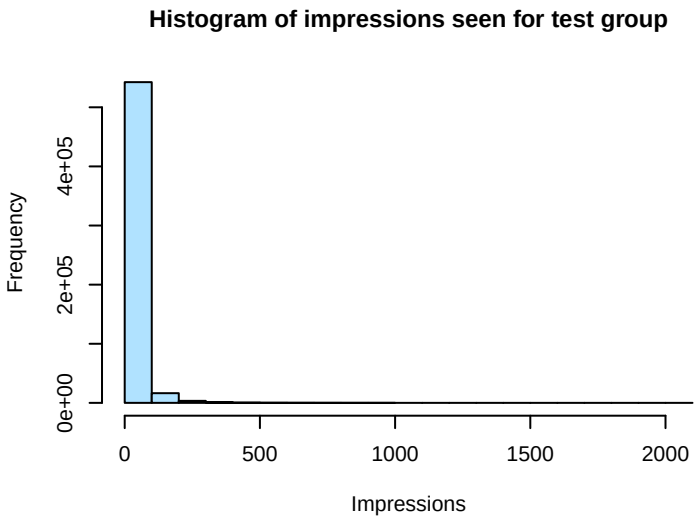
2.2) Distribution (histogram) comparison

To further understand how the distribution of `tot_impr` looks for the two groups, plot the histograms of `tot_impr` for each of the two groups (test and control). Do the two histograms look similar? (0.5 points)

```
par(cex = 0.65)
hist(DF$tot_impr[DF$test==0],
     col  = "lightskyblue1",
     main = "Histogram of impressions seen for control group",
     xlab = "Impressions")
```




```
hist(DF$tot_impr[DF$test==1],
     col  = "lightskyblue1",
     main = "Histogram of impressions seen for test group",
     xlab = "Impressions")
```



Yes, the histograms look quite similar.

2.3 Formal difference in means t-test

Finally, conduct a t-test to examine whether the differences (if any) in `tot_impr` across the two groups is statistically significant? (0.5 points)

```
t.test(tot_impr ~ test, data = DF)
```

```
    Welch Two Sample t-test

data:  tot_impr by test
t = -0.218, df = 25608, p-value = 0.8274
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6217286  0.4972735
sample estimates:
mean in group 0 mean in group 1
       24.76114        24.82337
```

The t-test tells us that the difference in impressions is not significant across groups.

Based on the above analyses, can you conclude that the randomization was done correctly? (0.5 points)

For the tot_impr variable, it appears the randomization was done properly.

3) Average treatment effect (ATE) estimation & application

3.2) Compute the treatment effect "by hand"

Calculate the ATE as the difference in mean outcomes across the treatment and control conditions. Report your ATE estimate as a percentage. Was the campaign effective?

```
100*(mean(DF$converted[DF$test == 1])-mean(DF$converted[DF$test == 0]))
```

```
[1] 0.7692453
```

The treatment effect is 0.00769, or 0.769%. This implies the treatment increases conversions (sales) by 0.769%. The campaign is effective to the extent that the treatment effect is positive. We formally assess the return on investment for the campaign in Section 4 below.

3.3) Compute the treatment effect by regression

Use a regression to estimate the treatment effect (ATE). Does your estimate match the "by hand" calculation? What is the standard error of the ATE?

```
lm1 = lm(converted ~ test, data = DF)
summary(lm1)
```

```
Call:
lm(formula = converted ~ test, data = DF)

Residuals:
     Min       1Q   Median       3Q      Max
-0.02555 -0.02555 -0.02555 -0.02555  0.98215

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.017854   0.001023   17.46  < 2e-16 ***
test        0.007692   0.001044    7.37  1.7e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1568 on 588099 degrees of freedom
Multiple R-squared:  9.236e-05, Adjusted R-squared:  9.066e-05
F-statistic: 54.32 on 1 and 588099 DF,  p-value: 1.703e-13
```

Same estimate. The standard error is 0.001044, or 0.1044%.

4) Return on investment (ROI)

4.1) Campaign incremental conversions

For the users in the test group, how many extra conversions can be attributed to the ad campaign? In other words, what is the incremental number of conversions from the ad campaign? (1 point)

Hint: the ATE is incremental (causal) effect of the campaign on conversion for each user. The total effect of the campaign on conversion is the number of users in the treatment condition times the ATE.

```
N_treat = sum(DF$test == 1)
ate = lm1$coefficients["test"]
inc_conv = N_treat*ate
inc_conv
```

```
    test
4342.982
```

The number of users in the test group is 564577. With the ad campaign, the conversion for this group increases by 0.769%. So the conversions that can be attributed to the ad campaign is 564577 x 0.00769 = 4342.982.

4.2) Campaign incremental profit

Recall from the overview above that TaskaBella gets on average $40 for each conversion.

How much more money did TaskaBella make by running the campaign (excluding advertising costs)? In other words, what is the incremental profit from the ad campaign? (0.5 point)

```
inc_pft = 40*inc_conv
inc_pft
```

```
    test
173719.3
```

The number of extra users that TaskaBella got through the ad campaign is 4342.982 and each additional user comes with a margin of $40. Thus the extra profit from the ad campaign is 4342.982 x $40 = $173,719.3.

4.2) Campaign cost

What was the cost of the campaign? (1 point)

Hint: the relevant number of impressions is contained in the `tot_impr` variable.

```
sum(DF$tot_impr)
```

```
[1] 14597182
```

```
cost = sum(DF$tot_impr)/1000*9
cost
```

```
[1] 131374.6
```

The total cost of the ad campaign is the sum of the impressions in the campaign (the sum of `tot_impr`) multiplied by the cost per impression (which is $9 per 1000 impressions.)

So the cost of the campaign is (14597182/1000)x$9 = 131,374.6

4.3) ROI calculation

Calculate the ROI of the campaign. Percentage ROI is defined as: 100*(incremental_profit - campaign_cost)/campaign_cost. (1 point)

```
roi = 100*(inc_pft-cost)/cost
roi
```

```
    test
32.23198
```

ROI is given by 100x(173,719.3 - 131,374.6)/131,374.6 = 32.2%

4.4) Control group opportunity cost

If the ad campaign had been shown to the control group as well, how much additional profit would have been generated? (1 point)

```
40*sum(DF$test == 0)*ate
```

```
    test
7238.291
```

The key thing to recognize here is that some users in the control group would have purchased anyway. The opportunity cost is the counterfactual gain we would have gotten from advertising to the control group. So the opportunity cost is:

(value of converted user) x (number of users in control group) x (incremental effect of campaign on conversion for exposed users)

where the last term is is the ATE

This gives $40 x 23524 x 0.007692 = $7238.29