Assignment 3_MSBA_6441_Causal Inference via Econometrics and Experimentation_AT_SM

Aditya Tomar, Shubham Midha

2022-11-21

(a) What is Wrong with Bob's RoI Calculation?

Keyword searches include the name of the company. These people most likely would already be aware of the website and would be actively looking for it.

For example: if a consumer used the search term "Bazaar shoes" and clicked on the sponsored ad that came up in the results. If the ad weren't there, the consumer would very likely click on organic link and company wouldn't even have to pay for the click cost for sponsored ads.

Also, it is mentioned in the case that the conversion probability and margin per conversion are the same for all consumers, irrespective of how they land at the website.

Hence for calculating the ROI of the sponsored campaign, we need to find out how many conversions can truly be attributed to sponsored ads.

(b) Define the Treatment and Control:

Platform is the unit of observation as the sponsored ads ran on Google (goog) for Weeks 1 - 9 and got stopped because of the technical glitch for Weeks 10 - 12.

Treatment is stopping of sponsored ads post week 9 for Google.

Treatment group: Google - Week 10, 11 & 12 Control groups: Yahoo, Bing, and Ask

Loading the data

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(plm)
```

```
##
## Attaching package: 'plm'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, lag, lead
```

```
library(readr)

# import data
data = read_csv("did_sponsored_ads.csv")
```

```
## Rows: 48 Columns: 5
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): platform
## dbl (4): id, week, avg_spons, avg_org
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Exploring and checking for number of weeks

```
# Data Exploration

# How many weeks does the data cover?
# What is the minimum and maximum week in the data?

min(data$week)
```

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

```
max(data$week)
```

```
## [1] 12
```

Data is for 12 weeks

Manipulating Data and adding additional treatment/control columns

```r
# Adding sponsored and organic traffic to compute total web traffic
data$total_traffic = data$avg_spons + data$avg_org

# Adding treatment variable - 1 for week 10, week 11, and week 12; 0 for rest
# Only for Google

data = data %>%
  mutate(treatment = ifelse((week>9 & platform=='goog'), 1,0))

# Adding a variable for indicating post period
data = data %>%
  mutate(post_flag = ifelse((week>9), 1,0))

# Adding a flag for Google
data = data %>%
  mutate(google_flag = ifelse((platform=='goog'), 1,0))
```

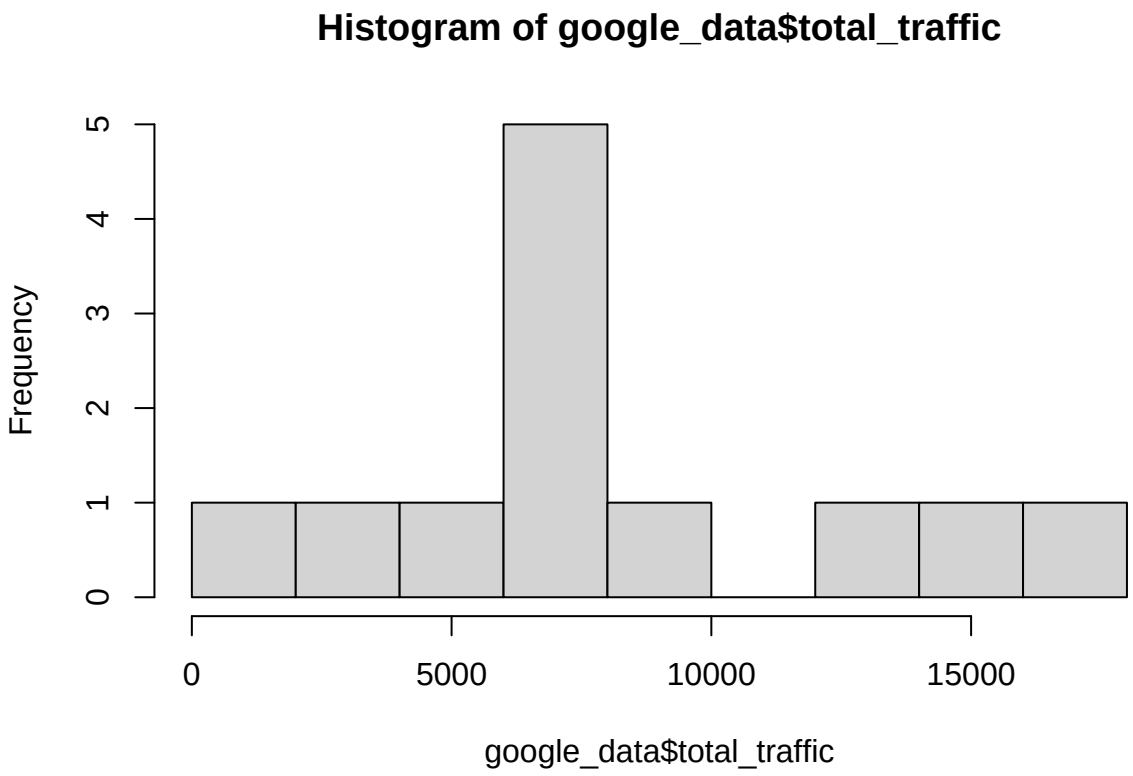
(c) Consider a First Difference Estimate:

Simply observing the treated unit (Google) by calculating first difference that is, the % change in web traffic arriving from Google; (after – before) / before This estimate is the pre-post difference in the treated cohort;

Problem: However, this approach assumes that the first nine weeks were systematically similar to the last three weeks

```r
google_data = data %>% filter(platform == 'goog')

hist(google_data$total_traffic,breaks=10)
```




```
summary(lm(total_traffic ~ treatment, data=google_data))
```

```
##
## Call:
## lm(formula = total_traffic ~ treatment, data = google_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7003.9 -2630.1  -172.5  2088.4  8625.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8390       1598   5.252 0.000373 ***
## treatment      -1846       3195  -0.578 0.576238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4793 on 10 degrees of freedom
## Multiple R-squared:  0.0323, Adjusted R-squared:  -0.06447
## F-statistic: 0.3337 on 1 and 10 DF,  p-value: 0.5762
```

Problem: However, this approach assumes that the first nine weeks were systematically similar to the last three weeks This is not the case as we can see a clear uptrend and hence we should control for the variation in weeks or try difference in differences method.

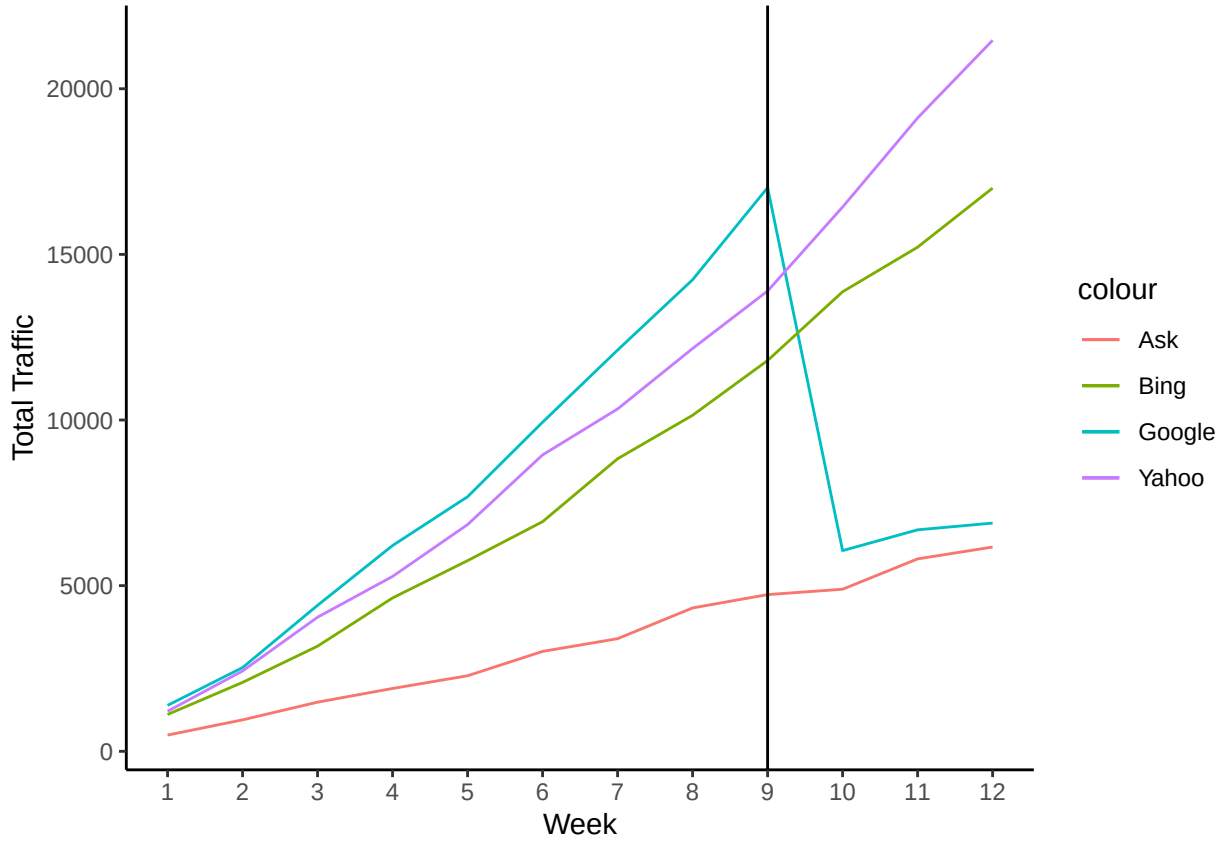
Also the p-value for the coefficients is 0.576 which is above our acceptable threshold of 0.05, hence we cannot interpret the coefficient meaningfully

(d) Calculate the Difference-in-Differences

```
bing_data = data %>% filter(platform == 'bing')
yahoo_data = data %>% filter(platform == 'yahoo')
ask_data = data %>% filter(platform == 'ask')

# Before running DiD, we need to check for parallel trend assumption in
# pre-period

ggplot(google_data, aes(x=week, y= total_traffic, color = 'Google')) +
  geom_line() +
  geom_line(aes(x=week, y= total_traffic, color = 'Bing'), data = bing_data) +
  geom_line(aes(x=week, y= total_traffic, color = 'Yahoo'), data = yahoo_data) +
  geom_line(aes(x=week, y= total_traffic, color = 'Ask'), data = ask_data) +
  geom_vline(xintercept = 9,color='black') +
  scale_x_continuous(breaks = seq(1, 12, by = 1)) +
  labs(y = "Total Traffic", x = "Week") +
  theme_classic()
```



```r
# Evaluating parallel trends assumption
summary(lm(total_traffic ~ google_flag*factor(week), data=data))
```

```
##
## Call:
## lm(formula = total_traffic ~ google_flag * factor(week), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8710.7  -111.8    87.3  1422.3  6586.3
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     936.3     2465.2   0.380 0.707414
## google_flag                     449.7     4930.3   0.091 0.928087
## factor(week)2                   881.3     3486.3   0.253 0.802574
## factor(week)3                  1964.7     3486.3   0.564 0.578291
## factor(week)4                  2998.3     3486.3   0.860 0.398274
## factor(week)5                  4023.3     3486.3   1.154 0.259840
## factor(week)6                  5361.0     3486.3   1.538 0.137190
## factor(week)7                  6584.7     3486.3   1.889 0.071069 .
## factor(week)8                  7940.0     3486.3   2.278 0.031955 *
## factor(week)9                  9204.3     3486.3   2.640 0.014337 *
## factor(week)10                10794.3     3486.3   3.096 0.004932 **
## factor(week)11                12445.3     3486.3   3.570 0.001550 **
## factor(week)12                13940.3     3486.3   3.999 0.000529 ***
## google_flag:factor(week)2       259.7     6972.5   0.037 0.970600
## google_flag:factor(week)3      1055.3     6972.5   0.151 0.880960
## google_flag:factor(week)4      1826.7     6972.5   0.262 0.795571
## google_flag:factor(week)5      2274.7     6972.5   0.326 0.747075
## google_flag:factor(week)6      3187.0     6972.5   0.457 0.651723
## google_flag:factor(week)7      4140.3     6972.5   0.594 0.558196
## google_flag:factor(week)8      4909.0     6972.5   0.704 0.488177
## google_flag:factor(week)9      6424.7     6972.5   0.921 0.365997
## google_flag:factor(week)10    -6122.3     6972.5  -0.878 0.388613
## google_flag:factor(week)11    -7146.3     6972.5  -1.025 0.315616
## google_flag:factor(week)12    -8437.3     6972.5  -1.210 0.238030
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4270 on 24 degrees of freedom
## Multiple R-squared:  0.6819, Adjusted R-squared:  0.3771
## F-statistic: 2.237 on 23 and 24 DF,  p-value: 0.0278
```

Since we observe statistically significant coefficients for a few weeks prior to week 9, our parallel trends assumptions is violated

```r
# Performing the DiD
did <- lm(total_traffic ~ google_flag + post_flag + google_flag * post_flag, data=data)
summary(did)
```

```
##
```

```
## Call:
## lm(formula = total_traffic ~ google_flag + post_flag + google_flag *
##     post_flag, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8437.7 -3231.0  -510.5  3591.6  8630.0
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5265.0      882.5   5.966 3.79e-07 ***
## google_flag             3124.9     1765.0   1.770  0.08357 .
## post_flag               8064.7     1765.0   4.569 3.94e-05 ***
## google_flag:post_flag  -9910.6     3530.0  -2.808  0.00741 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4586 on 44 degrees of freedom
## Multiple R-squared:  0.3274, Adjusted R-squared:  0.2816
## F-statistic: 7.141 on 3 and 44 DF,  p-value: 0.0005211
```

From the results obtained, we can say that due to the technical glitch causing the absence of Google sponsored ads, there is a decrease of 9910 clicks for the Google platform. In terms of statistical significance, the p-value(0.007) is lower than the the threshold of 0.05, implying that it is statistically significant.

This confirms our hypothesis that there might be few users who would visit the website via organic links in the absence of sponsored ads, however we would lose few customers too.

(e) Given Your Treatment Effect Estimate, Fix Bob's RoI Calculation

The ROI calculation that was calculated by Bob had the following metrics: An average revenue per click of 0.12 x 21, or $2.52, which implies an ROI of (2.52– 0.60)/(0.60), or 320.0%.

The discussion revolves Bob and Myra where the latter suggests that they can save money by showing the ads only to relevant people. The keywords searches already include names of bazaar.com, which means that users were aware of what they were searching - which makes them discuss if they even need to show ads to the users.

```
# regressing the average organic clicks for the treatment, pre_post
organic <- lm(avg_org ~ google_flag + post_flag + google_flag * post_flag, data=data)
summary(organic)
```

```
##
## Call:
## lm(formula = avg_org ~ google_flag + post_flag + google_flag *
##     post_flag, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1928.78  -847.92   -52.67   825.00  2067.33
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)             1489.7      215.4   6.917 1.51e-08 ***
## google_flag              777.0      430.7   1.804   0.0781 .
## post_flag               1984.1      430.7   4.607 3.49e-05 ***
## google_flag:post_flag   2293.2      861.4   2.662   0.0108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1119 on 44 degrees of freedom
## Multiple R-squared:  0.6043, Adjusted R-squared:  0.5773
## F-statistic:  22.4 on 3 and 44 DF,  p-value: 5.881e-09
```

First, out of all the traffic, we first calculate the proportion of true traffic: By this we mean that we want to find the proportion of clicks truly motivated by sponsored ad between it and clicks by customers who would still visit Bazaar.com irrespective of sponsored ads.

Looking at our co-efficient we obtained: X = 9,910 (clicks truly motivated by sponsored ad) Y = 2,293.2 (clicks by customers who would still visit Bazaar.com irrespective of sponsored ads)

Proportion of true traffic = X/(X+Y) = 9,910/(9,910+2,293) = 0.8120954 (81%)

We know calculate the new return of investment:

```
New_ROI = ((21 * 0.12 * 0.8120954- 0.6)/0.6)*100
New_ROI
```

```
## [1] 241.0801
```

With respect to the new estimated treatment effect, the new ROI is 241.08%.

Without the wrong expansion of supported promotion income, the recently determined return on initial capital investment of 320% is diminished to 241%, which is as yet a decent profit.