Causal Inference Analysis

Hao Cheng

January 26, 2023

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


# Set working directory to source file location
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
```

Background Introduction

Bazaar.com is the leading online retailer in the United States using display and search engine advertising, running paid search ads on Google and Bing. It releases its ads in response to keywords from online customers and classifies them into branded and nonbranded. Brand keywords contain the brand name such as 'Bazaar shoes' and 'Bazaar guitar.' Nonbranded keywords include items without a brand name, such as 'shoes' and 'guitar.' Considering traffic data from Google and Bing, Bob, who is from Bazaar's marketing analytics team, computed that ROI is 320% associated with sponsored search ads. His result is problematic because people who search with the word 'Bazaar' already intended to visit Bazaar.com, so we doubt the effectiveness of branded keyword ads. Our goal is to understand the causal inference of the search ads and their point.

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

As case mentioned, the 12% conversion rate we observed is not purely based on sponsored traffic but on both the sponsored and organic links. Therefore, we must isolate the conversion rate for sponsored ads only to calculate the right ROI. Given the sponsored ads on branded keywords, people who would have used organic search could also use sponsored ads to reach the website. These people usually have a higher conversion

rate since they are already familiar with the brand. This fact could lead to a wrong conclusion about the conversion rate in sponsored ads.

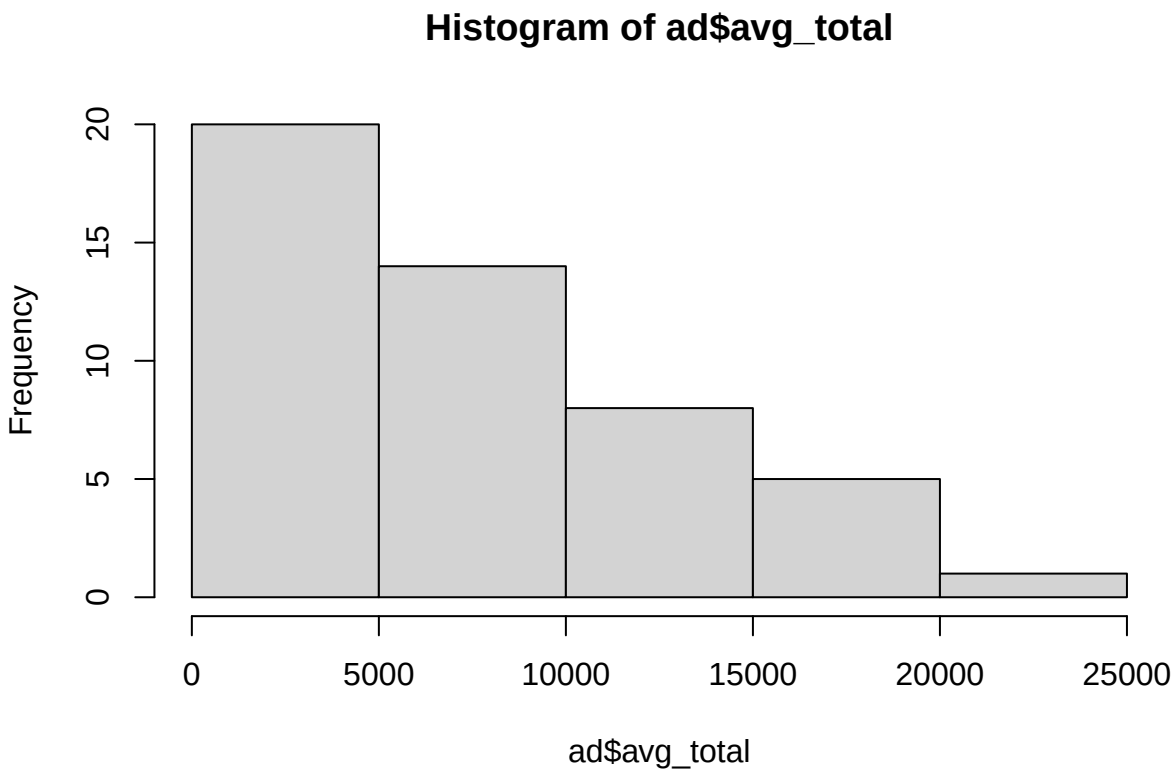
Besides, the margin per conversion is $21. This number is also biased since it is calculated by a combination of both sponsored and organic links. The actual margin could even be lower for those who click on the sponsored ads since they probably are still in the awareness phase.

(b) Define the Treatment and Control.

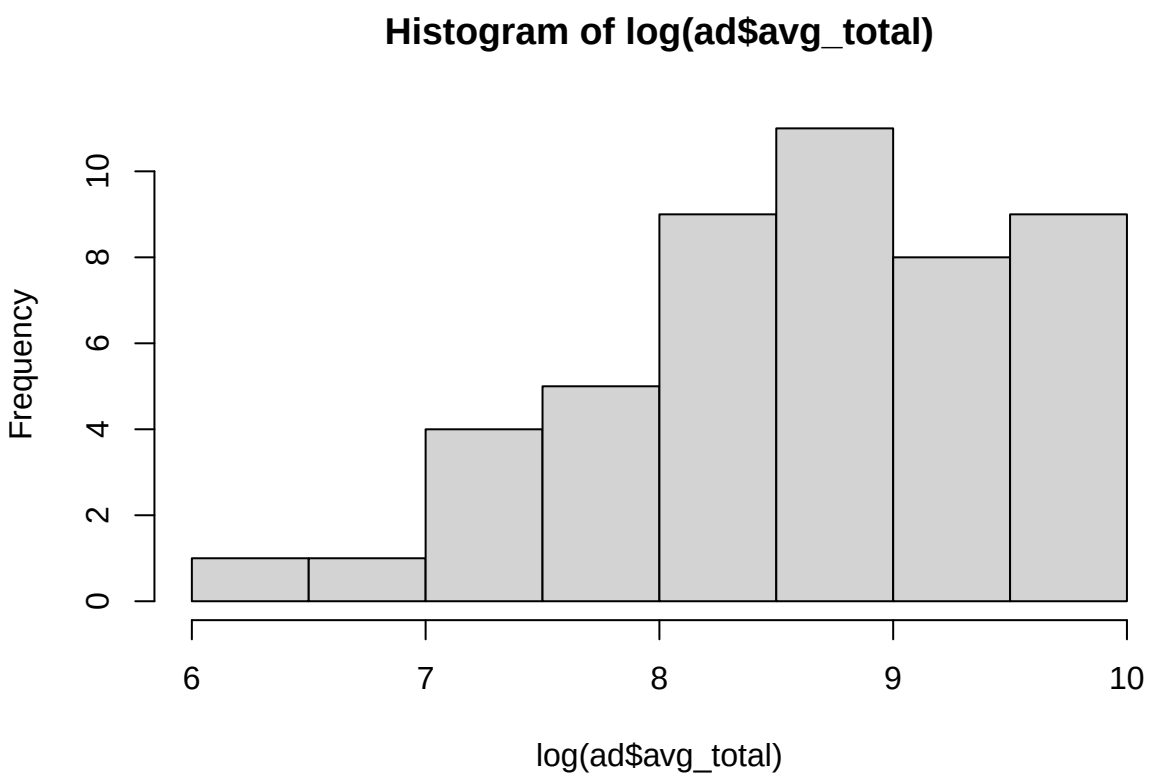
- Unit of observation: Weekly average clicks number on each platform.
- Treatment: Suspend sponsored ad campaign
- Treatment Group: Average clicks number of Google platform
- Control Group: Average clicks number of other platforms

(c) Consider a First Difference Estimate.

```
ad = read.csv("did_sponsored_ads.csv")
ad$avg_total = ad$avg_org + ad$avg_spons
hist(ad$avg_total)
```




```
hist(log(ad$avg_total))
```




```
ad = ad%>%mutate(after = ifelse(week<10, 0, 1))
ad = ad%>%mutate(treatment = ifelse(id==3, 1, 0))

# Create treatment subset
google = ad%>%filter(id==3)

# Calculate the mean avg_total in the two time periods(after)
google %>%
    group_by(after)%>%
    summarise(avg_week_total = mean(avg_total),
          avg_week_spons = mean(avg_spons),
          avg_week_org = mean(avg_org))
```

```
## # A tibble: 2 x 4
##   after avg_week_total avg_week_spons avg_week_org
##   <dbl>          <dbl>          <dbl>        <dbl>
```

```
## 1     0          8390.          6123.        2267.
## 2     1          6544              0         6544
```

```
# First difference
summary(lm(avg_total ~ after, data=google))
```

```
##
## Call:
## lm(formula = avg_total ~ after, data = google)
##
## 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 ***
## after          -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
```

```
# % Loss of clicks due to absence of sponsored ads
(6544-8390) / 8390
```

```
## [1] -0.2200238
```

With the first difference method, we can see that the treatment effect (no sponsored ad) causes around -1846 decrease in total traffic for the after period of the Google platform. But the p-value is greater than 0.05, indicating that there is no evidence that this treatment has effect on average total click.

The reason why this number is not solely reliable is that we ignore the natural variant of the website traffic. That said, perhaps in the post-period, the website traffic shows a significantly different trend compared to the pre-period. The estimation with this model could not capture this element and hence might lead to a wrong conclusion.

(d) Calculate the Difference-in-Differences.

```
summary(plm(log(1+avg_total) ~ treatment*after,
            data=ad,
            model='within',
            effect='twoways',
            index=c('id','week')))
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = log(1 + avg_total) ~ treatment * after, data = ad,
##     effect = "twoways", model = "within", index = c("id", "week"))
##
## Balanced Panel: n = 4, T = 12, N = 48
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.1054856 -0.0273133  0.0054203  0.0230813  0.1153218
##
## Coefficients:
##                  Estimate Std. Error t-value  Pr(>|t|)
## treatment:after -1.11611    0.04465 -24.997 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    2.2098
## Residual Sum of Squares: 0.10766
```

```
## R-Squared:      0.95128
## Adj. R-Squared: 0.92844
## F-statistic: 624.836 on 1 and 32 DF, p-value: < 2.22e-16
```

```
 # Calculate the mean avg_total in the two time periods(after)
ad %>%
    group_by(treatment, after)%>%
    summarise(avg_week_total = mean(avg_total),
              avg_week_spons = mean(avg_spons),
              avg_week_org = mean(avg_org))
```

```
## `summarise()` has grouped output by 'treatment'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 4 x 5
## # Groups:   treatment [2]
##   treatment after avg_week_total avg_week_spons avg_week_org
##       <dbl> <dbl>          <dbl>          <dbl>        <dbl>
## 1         0     0          5265.          3775.        1490.
## 2         0     1         13330.          9856.        3474.
## 3         1     0          8390.          6123.        2267.
## 4         1     1          6544              0         6544
```

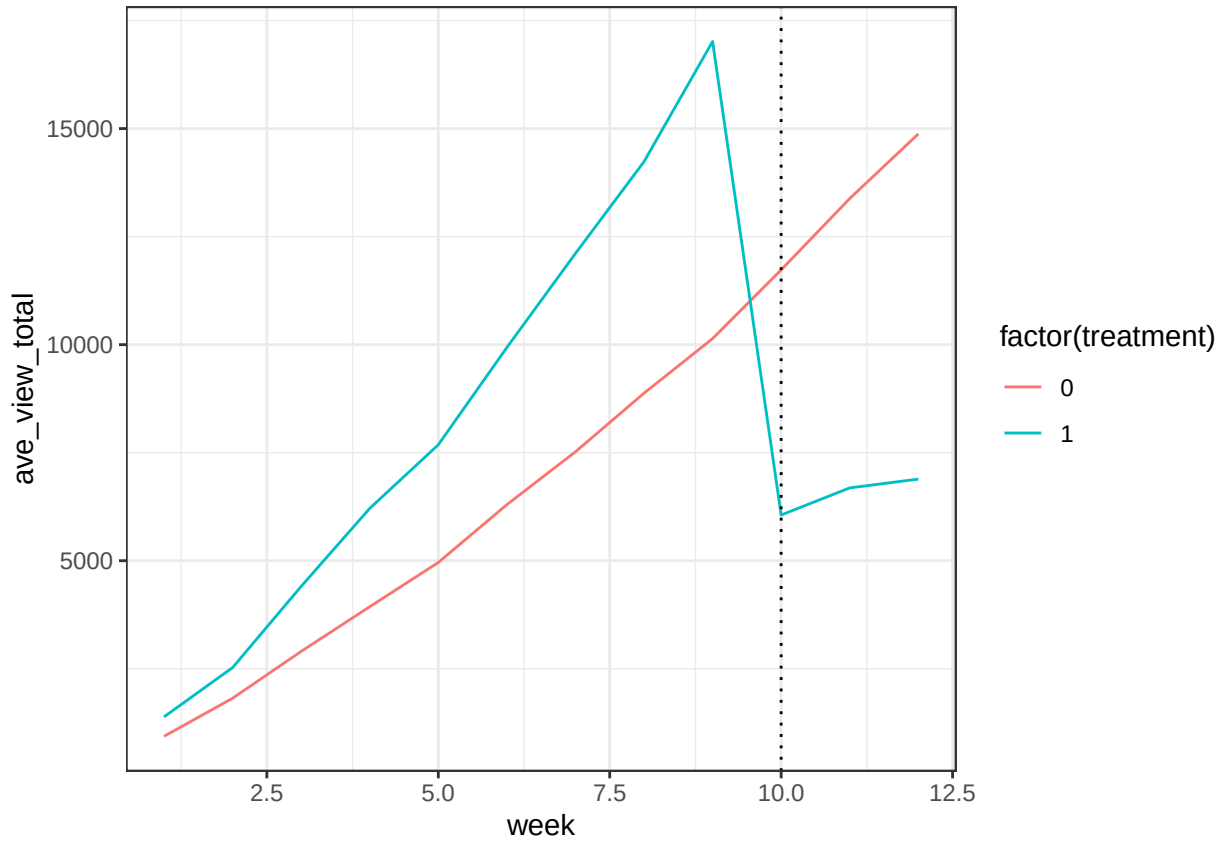
```
# The real % Loss of clicks due to absence of sponsored ads with DiD
(6544-8390) / 8390 - (13330-5265) / 5265 # 175% decrease
```

```
## [1] -1.751838
```

```
# Group data by week and treatment and calculate average values for plotting
week_ave = ad %>% group_by(week, treatment) %>% summarise(ave_view_total = mean(avg_total),
                                                          ave_view_org = mean(avg_org),
                                                          ave_view_spons = mean(avg_spons))
```

```
## `summarise()` has grouped output by 'week'. You can override using the
## `.groups` argument.
```

```
ggplot(week_ave, aes(x = week, y = ave_view_total, color = factor(treatment))) +
  geom_line() +
  geom_vline(xintercept = 10, linetype='dotted') +
  theme_bw()
```



With the DiD model, we can discover that the difference in difference effect of the treatment is -9910.6, which is way lower than the coefficient we estimate by the First Difference method. This shows the real impact of suspending sponsored ads on branded keywords. More specifically, this DiD model captures the difference in total traffic for the post-period with and without the treatment effect, which the first difference model could not capture.

(e) Fix Bob's RoI Calculation.

This ROI calculation is still based on the information provided by Bob (e.g the conversion rate and the margin), which might be not very accurate as we discussed in (a).

- Incremental weekly traffic attribute to sponsored ad: 9911

- Incremental gain from these clicks: 9911 * 0.12 * 21 = 24975.72
- Average weekly clicks from sponsored search: 6123
- Weekly cost of sponsored search: 6123 * 0.60 = 3673.8

ROI = (24975.72 - 3673.8) / 3673.8 = 580%