Causal Inference - Assignment 3

Abhishikth Peri (5907987)

2023-11-11

Business Context

Bazaar.com is a leading US online retailer that leverages marketing analytics through both display and online search advertising to improve its revenue and reach.It runs paid search engine advertising on Google and Bing, where ads are curated in response to the search keywords of online customers. The advertisements in search engine domain are categorized as 'Branded' and 'Non-Branded' based on the query keywords - 'Bazaar', 'Bazaar shoes' etc. for Branded (which include the comapny name), and 'shoes','skirts' etc. for Non-Branded (which do not include company name).

Problem statement

Bob and Myra are senior members of the marketing analytics team who are interested in assessing the return on investment (ROI) of the sponsored advertising efforts. Bob claims that the sponsored ad-campaign yielded a massive 320% return on investment (ROI), while Maya opines that the number is too good to be true. We are tasked with evaluating the effectiveness of the ad-campaign from an ROI perspective, specifically addressing the following questions:

(a) Identify the mistake in Bob's RoI Calculation.
(b) Define the Treatment and Control of the experiment.
(c) Consider a First Difference Estimate to first understand how reliable is the pre-post difference in treatment cohort in estimating the causal effect of the treatment.
(d) Calculate the Difference-in-Differences.
(e) Correct the original ROI formulation to propose a new ROI calculation.

ROI discussion

Understanding the reasoning and computation provided by Bob for ROI, we highlight the following limitations:

1) Bob's ROI computation assumes that every customer who clicked on the sponsored ad is influenced and driven by the advertising. However, a closer look into the mechanism of these ads shows us that all those customers who were shown these sponsored ads had already queried for Bazaar products. This means that the customers would have still organically landed on the Bazaar website irrespective of the sponsored campaign since they already had an intent to visit the same. In other words, those customers were not logically driven by the sponsored campaign and would still land on the website despite the campaign.
2) If we assume x% of the sponsored traffic actually comes from genuinely clicking on sponsoored ads, then the (100-x)% represents the customer segment on which advertising expenditure is wasted. By mitigating this wasteful expenditure, Bazaar can realize savings.

Threats to Causal Inference

1) Influence of any confounding variables that might influence the advertising (and/or) click rate results that might not be taken into consideration can introduce omitted variable bias to the experiment. Example - competitor marketing campaigns, external promotions, changes in consumer behavior etc.
2) Taking into consideration the same strategies applied for all search engines for targeting ads, keyowrd mix, and the mix of potential customers visiting the company's website from both search engines is virtually identical; we can assume that there is no selection bias in this experiment.
3) Inaccurate measurement of click-through rates due to technical problems can intorduce measurement error into the experiment.

Solution ideation

To accurately estimate the causal impact of sponsored advertisements on website traffic of Bazaar.com, we will employ the Difference-in-Differences methodology. 1) Calculate the first difference for total weekly average traffic for pre and post stopping of sponsored ads in Google to understand the effect of sponsored ads on pre-post period. 2) Calculate the difference-in-difference estimate of treatment effect to understand pre-post difference with Google versus other search engines and determine true casual impact of sponsored ads on ROI.

```
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(plm)
```

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

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

```
library(ggplot2)
library(readxl)
```

Data exploration

We are provided with a dataset containing weekly average traffic stats for four search engines - Google, Bing, Yahoo and Ask for both sponsored and organic traffic buckets.

Load dataset

```r
bzdata = read.csv('did_sponsored_ads.csv')
```

1) What is wrong with Bob's ROI calculation?

Referring to the ROI discussion above, we have established that all clicks on sponsored ads cannot directly be attributed to the success of the sponsored campaign since the users might already have the intention to use the website regardless.

If we calculate ROI using Bob's formula for 100 users who clicked on sponsored ads considering 12% purchase probability and average profit margin per customer as $21 and average cost per sponsored click-ad as $0.6: Old ROI = [100((21*0.12)-0.6)]/(100*0.6) = 320%

```r
Old_ROI = ((100*21*0.12)-(100*0.6))/(100*0.6)
Old_ROI
```

```
## [1] 3.2
```

However, let's say only 30 of those customers were actually influenced by sponsored ads. New ROI = [(50*21*0.12)-(100*0.6)]/(100*0.6) = 110%

```r
New_ROI = ((50*21*0.12)-(100*0.6))/(100*0.6)
New_ROI
```

```
## [1] 1.1
```

Hence, we can see inflation of revenue. and ROI using Bob's ROI computation.

2) Define treatment and control

1) We consider 'Google' as treatment group, and the rest three platforms as control groups (variable = treatment)
2) Within Google, we consider weeks 10, 11 and 12 as post treatment, and weeeks 1 - 9 s pre-treatment (variable = effect)
3) We sum the traffic from sponsored and organic channels into a singe variable - traffic

```r
tweek = c(10,11,12)
bzdata = bzdata %>% mutate(treatment = ifelse(platform == 'goog',1,0),
                           effect = ifelse(week %in% tweek,1,0),
                           traffic = avg_spons + avg_org)
```

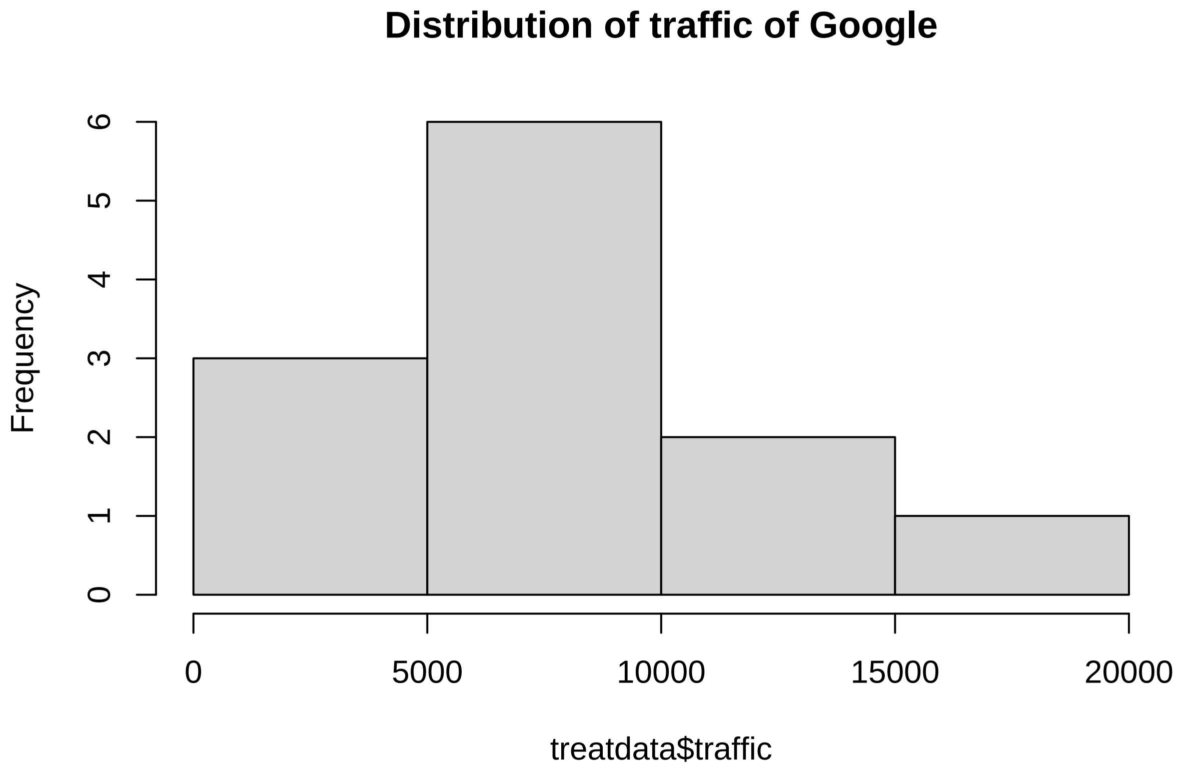
3) First difference estimate

Here, we calculate the percentage change in web traffic arriving from Google; (after – before) / before for the pre-post treatment cohort and estimate the value using a regression.

First, let us look at the distribution of data.

```r
treatdata = bzdata %>% filter(treatment== 1)

hist(treatdata$traffic, main = "Distribution of traffic of Google")
```



Here, since the data is skwewed, applying a log transform will help in effective model building.

```
gdata = bzdata %>% filter(platform == 'goog')

M1 = lm(log(traffic) ~ effect, data = gdata)
```

```
summary(M1)
```

```
##
## Call:
## lm(formula = log(traffic) ~ effect, data = gdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54933 -0.15495  0.03784  0.46975  0.95834
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.783506   0.248968  35.280 7.94e-12 ***
## effect      0.001306   0.497936   0.003    0.998
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7469 on 10 degrees of freedom
## Multiple R-squared:  6.88e-07,   Adjusted R-squared:   -0.1
## F-statistic: 6.88e-06 on 1 and 10 DF,  p-value: 0.998
```

```r
exp(coef(summary(M1))[2])-1
```

```
## [1] 0.001306972
```

From the above model, first difference = exp(0.001306)-1 = 13.06%. However, since p-value = 0.998 which is greater than 0.05 (level of significance), we cannot conclude the statistical significance of this test.

Additionally, the assumption of stable and constant market conditions without factoring in any external variables like seasonal variations, nature/intent of search etc. on website traffic significantly undermines the effectiveness of this method.

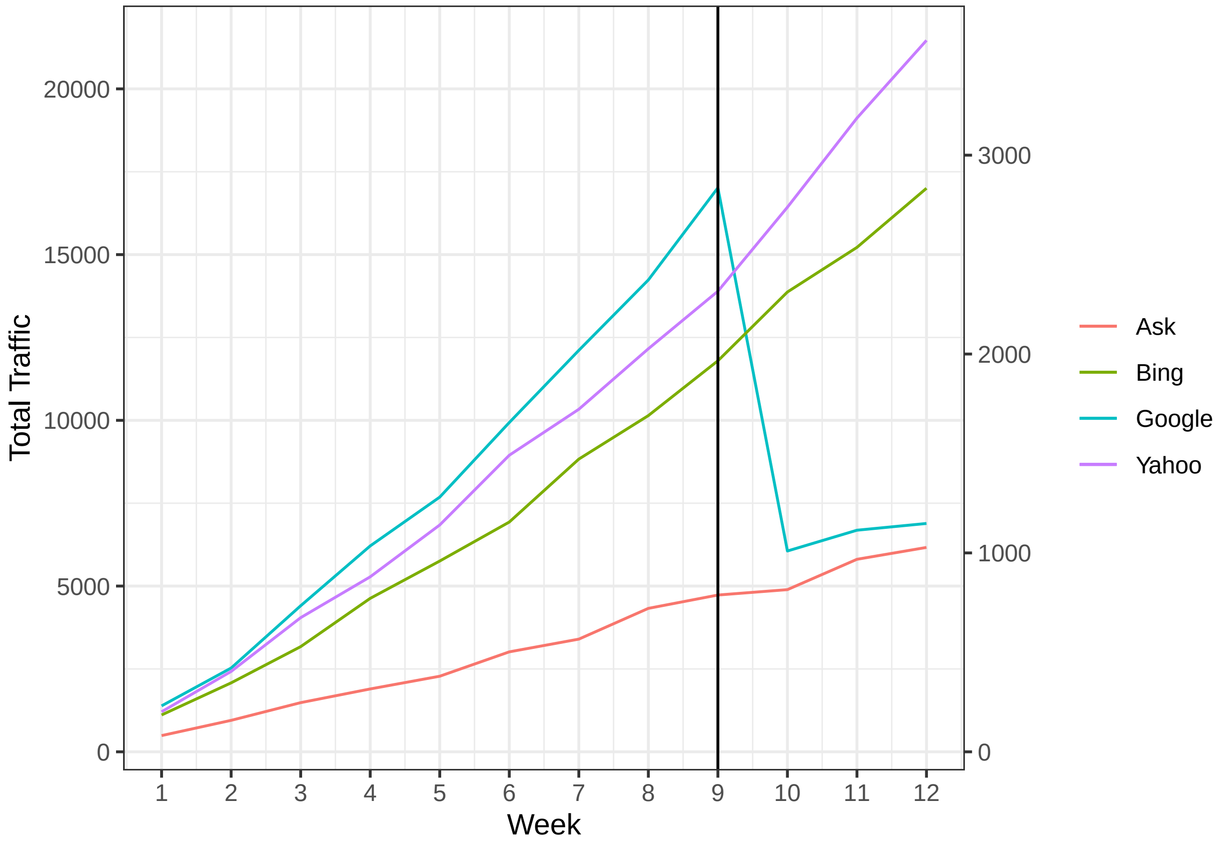
Hence, we cannot rely on first difference estimate alone, and need to use difference in differences methodology to determine causality of the experiment. DiD is a more robust approach for this usecase since we would be taking into consideration both treatment (Google) and control groups (Ask, Yahoo, Bing) which can lessen the aforementioned concerns to a certain extent.

4) Difference-in-differences

As a pre-requisite, let us first check for any parallel trends to validate that the differences that we obtain are comparable in the first place. Key intuition - 1) We should not see a negative trend amogst differences since we then cannot derive any insights. 2) Treatment and control group should exhibit similar behaviour for pre-treatment.

```r
tb = bzdata %>%  filter(platform %in%  c('bing')) %>% select(week, traffic)
ty = bzdata %>%  filter(platform %in%  c('yahoo')) %>% select(week, traffic)
ta = bzdata %>%  filter(platform %in%  c('ask')) %>% select(week, traffic)

ggplot(bzdata %>% filter(platform == 'goog'), aes(x=week, y= traffic, color = 'Google')) +
  geom_line() +
  geom_line(aes(x=week, y= traffic, color = 'Bing'), data = tb) +
  geom_line(aes(x=week, y= traffic, color = 'Yahoo'), data = ty) +
  geom_line(aes(x=week, y= traffic, color = 'Ask'), data = ta) +
  geom_vline(xintercept = 9,color='black') +
  scale_y_continuous(sec.axis = sec_axis(~./6)) +
  scale_x_continuous(breaks = seq(1, 12, by = 1)) +
  labs(y = "Total Traffic", x = "Week") +
  theme_bw() +
  theme(legend.title = element_blank())
```



Evaluating this visualization, we can see that there are no parallel trends for pre-treatment weeks. We see positive divergence for all search engines until week 9, after which only Google shows convergence (decrease in traffic). Hence, we conclude that Difference-in-differences analysis can be applied for this usecase.

```r
DiD_model = lm(traffic ~ treatment * factor(week), data=bzdata)
summary(DiD_model)
```

```
##
## Call:
## lm(formula = traffic ~ treatment * factor(week), data = bzdata)
##
## 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
## treatment                   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 ***
## treatment:factor(week)2     259.7     6972.5   0.037 0.970600    
## treatment:factor(week)3    1055.3     6972.5   0.151 0.880960    
## treatment:factor(week)4    1826.7     6972.5   0.262 0.795571    
## treatment:factor(week)5    2274.7     6972.5   0.326 0.747075    
## treatment:factor(week)6    3187.0     6972.5   0.457 0.651723    
## treatment:factor(week)7    4140.3     6972.5   0.594 0.558196    
## treatment:factor(week)8    4909.0     6972.5   0.704 0.488177    
## treatment:factor(week)9    6424.7     6972.5   0.921 0.365997    
## treatment:factor(week)10  -6122.3     6972.5  -0.878 0.388613    
## treatment:factor(week)11  -7146.3     6972.5  -1.025 0.315616    
## treatment: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
```

Using the above test, we see positive interaction coefficients for reatment and week from week 1 until 9 and negative coefficients from 10 - 12. This change in sign indicates stoppage of sposored advertisements in Google starting week 10. Moreover, p-value less than 0.05 shows us that the results are statistically significant.

Now, we run a DiD model using treatment, effect and interaction between treatment and effect as independent variables to get the average traffic and establish true causal relationship of sponsored advertisements on observed click rates.

```
DiD_independent = lm(traffic ~ treatment + effect + treatment * effect, data=bzdata)
summary(DiD_independent)
```

```
## 
## Call:
## lm(formula = traffic ~ treatment + effect + treatment * effect, 
##     data = bzdata)
## 
## 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 ***
## treatment          3124.9     1765.0   1.770  0.08357 .  
## effect             8064.7     1765.0   4.569 3.94e-05 ***
## treatment:effect  -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
```

Discontinuing sponsored ads on Google results in a weekly loss of 9910 clicks for Bazaar on average. The newly estimated treatment effect, derived by comparing the control and treatment groups, highlights the causal impact of sponsored ads. This approach is superior and more accurate than the pre-post estimate because it allows us to analyze the behavior of both control and treatment groups and their temporal variations within a single model.

5) Fixing Bob's original ROI computation using new treatment effect

Since the stoppage of sponsored ads on Google staring week 10, users would organically land on Bazaar website as they already intended. Hence, these users need not be considered for sponsored campaign and revenue generation.

```r
DiD_org = lm(avg_org ~ treatment * effect, data=bzdata)
summary(DiD_org)
```

```
##
## Call:
## lm(formula = avg_org ~ treatment * effect, data = bzdata)
##
## 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 ***
## treatment           777.0      430.7   1.804   0.0781 .
## effect             1984.1      430.7   4.607 3.49e-05 ***
## treatment:effect   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
```

Using above model, we obtain average traffic for aforementioned users. Here, we see interaction term of ‘effect’. and ‘treatment’ is 2293, which means that post stoppage of sponsored ads, click rate increased by 2293 clicks on average per week.

Referring to DiD_independent model, we see that stoppage of sponsored ads caused loss in 9910 clicks on average per week. Since 2293 clicks attribute to no increase in revenue since they are generated from organic search queries, we can sum 9910 and 2293 to get a total of 12,203 clicks (new treatment effect in did_independent + new treatment effect in did_organic)

Proportion of true traffic = 9910/12203 = 0.8120954 = 81.2%

ROI_final = ((Margin*proportion*probability)-(cost per click))/(cost per click)

```r
ROI_final = (21 * 0.12 * 0.8120954 - 0.6)/0.6
ROI_final
```

```
## [1] 2.410801
```

Based on new treatment effect, corrected ROI = 241% which is smaller compared to original 320% but is still very impressive for a sponsored advertising campaign. Hence, sponsored advertising campaign is a positive aspect for the revenue of Bazaar.com