Assignment 3: Measuring ROI on Sponsored Search Ads

Aakriti Aneja, ID:5790726 ;Maria Moy, ID: 5479516

11/20/2022

Context and Background

Bazaar has data from both sponsored and organic clicks from four platforms over 12 weeks. The team wants to understand the impact of stopping sponsored search advertising for keywords and determine the ROI analysis for Google. Having the correct ROI allows Bazaar to know how to advertise better.

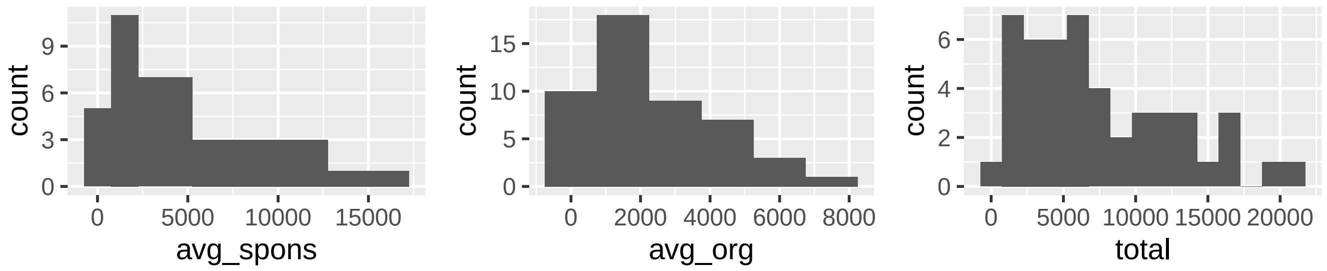
We will explain why Bob's calculation is wrong. Then, we will explain what the observation, treatment, and control are for the analysis. Next, we will calculate the first difference estimate using only the treated group (Google) and explain why this isn't the best method. Finally, we will discuss the assumptions of Difference-in-Differences (DiD) analysis, run a DiD regression, and calculate the ROI.

We have a Panel data here. We use this to create columns for treatment and post period.

```r
# after column is 1 for weeks 10-12
data_ad$after = ifelse(data_ad$week > 9, 1,0)
#google column is if the platform is google
data_ad$google =ifelse(data_ad$platform == 'goog', 1,0)
```

We also look at the distribution of sponsored ad clicks, organic clicks and total traffic to check for skewness

```r
hist1 <- ggplot(data_ad, aes(x=avg_spons)) + geom_histogram(binwidth=1500)
hist2 <- ggplot(data_ad, aes(x=avg_org)) + geom_histogram(binwidth=1500)
hist3 <- ggplot(data_ad, aes(x=total)) + geom_histogram(binwidth=1500)
grid.arrange(hist1, hist2, hist3, ncol = 3)
```



All there come out to be right skewed, so we may need to use log transform to estimate the effect.

(a) What is wrong with Bob's ROI Calculation? Bob's calculation is wrong because it assumes that people who visited using the sponsored ad would not have visited if not for the sponsored ad. As correctly pointed by Maya, there will be some customers that would scroll to Bazaar link without sponsored ad present, as they searched for the 'Bazaar' keyword in the first place, showing their interest in the website. The contribution of these customers is overestimating the ROI.

(b) Define the Treatment and Control

- B1. What is the unit of observation? Traffic on a Company and Week level
- B2. what is the treatment? Treatment is actually not getting sponsored advertisements
- B3. What are the units being treated? Google: week 10-12
- B4. Which are being controlled? Google: week 1-9, Yahoo: week 1-12, Bing: week 1-12, Ask: week 1-12

(c) Consider a first difference estimate We filter the data for Google only to do this analysis

```r
df1 <- data_ad %>% filter(data_ad$platform == 'goog')
summary(lm(log(total) ~ after, data = df1))
```

```
##
## Call:
## lm(formula = log(total) ~ after, data = df1)
##
## 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 ***
## after       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
```

- **C1. Estimate this value using a regression**: Treatment increases total traffic by ~0.13% points. However, with a p-valueof 0.998, it is statistically not different from zero.
- **C2. Explain why it would not be a good idea to solely rely on this number as our estimate of the causal effect of the treatment**: We are essentially ignoring any week over week trends in the data by relying on this number solely. In the post-period, We want to know how other websites are behaving. If there are shocks across all websites (e.g. if week 10 was Black Friday/cyber Monday), we want to be able to track the market shocks. We want to see other units are not treated in same time period.

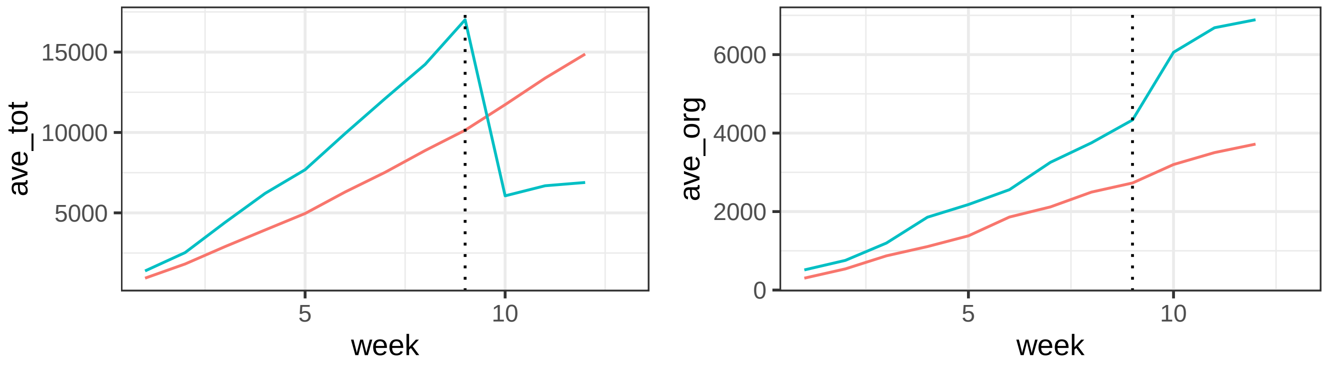
(d) Calculate the Difference-in-Differences. For this analysis we have to make two assumptions
- **SUTVA Assumption** - There are no unmodeled spillovers, that treatment does not affect the control. The treatment (not having sponsored ads on Google), is not changing the behavior of the control group, both sponsored and organic ads, on other websites. People using a search engine on one platform usually only use one platform, so there is no unmodelled spillover, hence no SUTVA Violation.
- **Parallel trends assumption**: Without the treatment, treated subjects would have continued in parallel with the control.

```
#regressing total on after and treatment(google)
summary(lm(total ~ google * factor(week), data = data_ad))
```

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

We see here that none of the p-values of the interaction term are significant - so we can safely proceed with the parallel trends assumption

```
tot_plot <- ggplot(week_ave, aes(x = week, y = ave_tot, color = factor(google))) + geom_line(s
org_plot <- ggplot(week_ave, aes(x = week, y = ave_org, color = factor(google)), show.legend =
grid.arrange(tot_plot, org_plot, ncol = 2)
```



The red line shows trends for Google and blue shows for the control group. This provide some visual evidence for parallel trend in control and treatment group.

We move to running the difference in difference analysis

```
summary(lm(log(total)~ after*google, data=data_ad))
```

```
##
## Call:
## lm(formula = log(total) ~ after * google, data = data_ad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0608 -0.5442  0.1414  0.5811  1.2861
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.2532     0.1543  53.505  < 2e-16 ***
## after          1.1176     0.3085   3.623 0.000751 ***
## google         0.5303     0.3085   1.719 0.092629 .
## after:google  -1.1163     0.6170  -1.809 0.077241 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8015 on 44 degrees of freedom
## Multiple R-squared:  0.2415, Adjusted R-squared:  0.1898
## F-statistic:  4.67 on 3 and 44 DF,  p-value: 0.006435
```

```
#Transforming the interaction coefficient to be able to interpret.
(exp(-1.1163)- 1)*100
```

```
## [1] -67.25107
```

- **D1. What is the new treatment effect estimate?**: The interaction (after:google) is the effect of the treatment.There is a total click decrease of 67.25%. While this is not statistically significant at the 0.05 level, it does have a p-value of 0.077 indicating some relationship.

- **D2.How does it compare with the pre-post estimate, and what does this say about problems with relying on the post estimator?**:
  - We found the difference in part (c) was not statistically significant. However, in this DiD calcuation, we find that there is a difference, there is a decrease in clicks by 67.25%.
  - An issue with relying on the post-estimator is that it shows ads visited in the post-period using only Google's information and we aren't seeing the other company's information.
  - Relying on post estimator means we are overlooking the natural trend in data and incorrectly concluding that there is no difference in pre and post period as we did in part (c).

(e) Given Your Treatment Effect Estimate, Fix Bob's RoI Calculation. We see that the decrease in the total clicks in post period for Google from the DiD analysis is 67.25%. We hypothesize that some part of this decrease is going into the organic clicks (as Maya's hunch) and the rest is going into external factors eg., to competitors.

In this natural experiment we can estimate the part of the decrease that went into the organic ad clicks - inherently estimating the proportion of customers who would organically click on Bazaar links in absence of sponsored ads.

```
summary(lm(log(avg_org) ~ after*google, data=data_ad))
```

```
##
## Call:
## lm(formula = log(avg_org) ~ after * google, data = data_ad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1859 -0.4715  0.1132  0.5002  1.1203
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.0380     0.1402  50.188  < 2e-16 ***
## after           1.0333     0.2805   3.684 0.000625 ***
## google          0.4851     0.2805   1.730 0.090680 .
## after:google    0.2284     0.5609   0.407 0.685826
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7287 on 44 degrees of freedom
## Multiple R-squared:  0.3651, Adjusted R-squared:  0.3218
## F-statistic: 8.434 on 3 and 44 DF,  p-value: 0.0001539
```

```
(exp(0.2284)-1)*100
```

```
## [1] 25.65879
```

Lets look at this in terms of actual click numbers, as the percentage we shall not be able to subtract the increase in organic ad clicks from sponsored ad clicks

```
summary(lm(total~after*google, data=data_ad))
```

```
##
## Call:
## lm(formula = total ~ after * google, data = data_ad)
##
## 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 ***
## after          8064.7     1765.0   4.569 3.94e-05 ***
## google          3124.9     1765.0   1.770  0.08357 .
## after:google  -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
```

```
summary(lm(avg_org~after*google, data=data_ad))
```

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

Here, We observe the following two things:

- In absence of sponsored ads, the total clicks go down by 9910. These should ideally constitute the decrease in sponsored ad clicks along with decrease in organic clicks
- However, When there are no sponsored ads, organic clicks increase by 2293.These are the clicks that Bob is overestimating and attributing to the sponsored ads - when these clicks would have anyway come ‘organically’

So, the actual decrease in sponsored ad click is (9910+2293), and the overestimated click count in sponsored ads is 2293. If we remove this overestimated component from the revenue obtained from sponsored ads, we shall be able to fix the ROI calculation

```
# $21 is return when a customer makes a purchase
# 12% probability of customer purchasing once they land on the website
# 2293 - increase in organic clicks
# 9910 - total decrease in click - constitutes sponsored ads decrease as well as organic cl
# So, revenue and cost per click come out to be
revenue = 21 * 0.12 *(1-(2293/(9910+2293)))
cost = 0.60
# and the ROI comes out to be
(revenue - cost)*100/cost
```

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

We observe from the above calculation that the updated ROI is 241%, way lower than Bob’s estimate of 320%