Causal Inference and Econometrics- Assignment 3

Pranvi Setia (5815963), Priyanka Saraswat (5815960)

2022-11-19

Business Goal

Bazaar wants to understand the impact of their sponsored advertisements and correctly measure its ROI, so they can allocate budgets more efficiently towards sponsored advertisements.

Experiment

Bazaar.com ran an experiment where they ran a paid campaign using branded keywords (containing the word "Bazaar") estimated the number of customers who landed on their website through sponsored and organic advertisement links. But there was a glitch in week 10 for the Google campaign so they were unable to capture the customer traffic through sponsored advertisements. So, they used the data obtained from weeks 1 through 9 to calculate the ROI from sponsored advertisements.

But one of the executives, Myra, believed that their ROI calculations were flawed because of overestimation. She believed that customers who were searching for their brand were already more likely to come to their website and it didn't matter whether or not they were shown a sponsored advertisement. She argued that these customers would still land on their website through organic links even if they weren't being shown any sponsored advertisements. Therefore, to understand the effect of sponsored advertisements on their web traffic for branded keywords campaign they decided to use the campaign data to find the difference in their total web traffic through sponsored and organic advertisements between in the period before and after the glitch, that is, till week 9 and after it, respectively.

We have employed the difference in difference approach to estimate the effect of the glitch (sponsored

advertisements) on total advertisement clicks across the platforms and use it to estimate the correct ROI of sponsored advertisements.

Data Description The data consists of name of the platform, the week number, number of clicks on sponsored advertisement links and the number of clicks on organic links on that platform for a given week. The campaign was run on platforms: Yahoo, Bing, Google and Ask, for 12 weeks.

```
#Importing required packages
suppressMessages(library(dplyr))
suppressMessages(library(ggplot2))
suppressWarnings(suppressMessages(library(plm)))
#Importing the data
did_ads <- read.csv("did_sponsored_ads.csv")
```

Data Analysis

1. What is Wrong with Bob's ROI Calculation?

Bob in his ROI calculation multiplies the probability of a customer making a purchase once they land on their website (12%) with the average margin per conversion ($21). Here, Bob is assuming that every customer who lands on their website is either clicking on the organic or sponsored advertisements and hence, incorrectly attributing the effect of advertisements on purchase. He is overestimating the traffic that lands on Bazaar.com by clicking on their advertisements and completely disregarding customers who land on their website through other means, such as, directly typing the website URL. Moreover, we are only interested in the ROI based on sponsored ads but we are capturing the traffic through all means, instead of just sponsored ads, consisting of organic clicks and other means.

Secondly, as per Myra's comment for the brand keyword campaigns the customer's are searching for the "Bazaar" keyword because they are already interested in Bazaar.com and they will come to their website even if they aren't shown and sponsored advertisements. So, the customers which could be captured through organic search results are also being misallocated to sponsored ads, thus boosting the organic ad clicks estimate, which is leading to overestimation of ROI.

2. Define the Treatment and Control.

Unit of Observation: The unit of observation is the average sponsored and organic ad clicks on a platform for a given week.

Treatment: Google platform is our treatment variable in this analysis

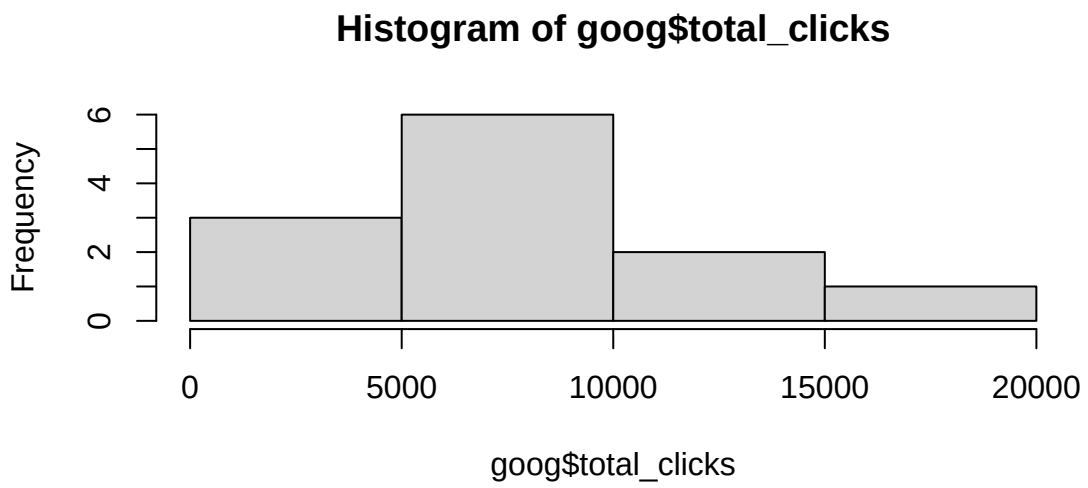
Control: Rest of the platforms, Yahoo, Bing and Ask, are part of the control group for this analysis.

The weeks 10-12 are being treated for the treatment group and the rest represent the control units.

3. First Difference Estimate.

First, we will create new variables to capture the total clicks and filter the dataset for only Google platform campaign data. Google's total ad clicks in weeks 10-12 is a part of our treatment group and Google's total ad clicks in weeks 1-9 forms our control group for this part of our experiment.

```
did_ads <- did_ads %>% mutate(total_clicks = avg_spons + avg_org)
did_ads <- did_ads %>% mutate(after = ifelse(week<=9,0,1))
goog = did_ads %>% filter (platform=="goog")
#Plotting the histogram for total clicks on Google
hist(goog$total_clicks)
```



Since the histogram for total ad clicks on Google is not skewed we can run a simple linear regression model to estimate the effect of treatment on Google in the after period on total ad clicks.

Our regression equation is as follows: total_ads = Beta0 + Beta1*after + error

For this test, our hypothesis is: H0: Beta1 = 0 OR having sponsored ads does not have any effect on the total ad clicks Ha: Beta1 not equal to 0 OR having sponsored ads does have an effect on the total ad clicks

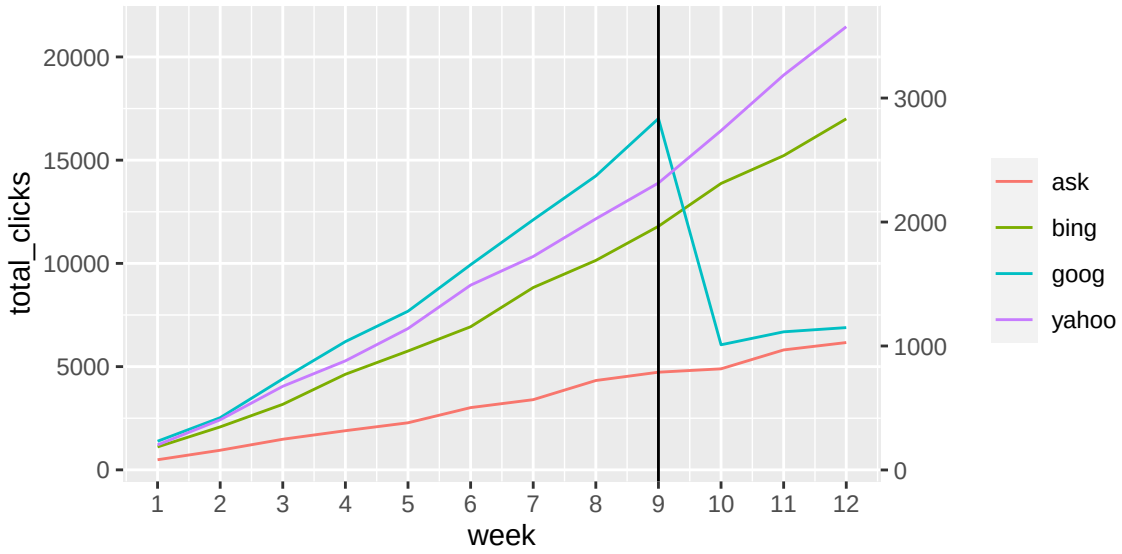
```
summary(lm(1+total_clicks ~ after, data = goog))
```

```
##
## Call:
## lm(formula = 1 + total_clicks ~ after, data = goog)
##
## 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)     8391       1598   5.252 0.000372 ***
## 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
```

The p-value for test is 0.57, which is greater than 0.05, we cannot the null hypothesis. Therefore, there is no significant effect of sponsored ads on total ad clicks for Google. Not having any sponsored effects decreases the total average ad clicks by 1846 for a given week on Google, but this value is statistically insignificant.

But the results obtained might not be a good reliable estimate for the captured effect, because our data still may be influenced by some omitted variable bias, which could be effecting the total ad clicks for all platforms. We can check if there is some trend or seasonality in data by plotting the total clicks for platforms across weeks.

```
ggplot(did_ads, aes(x = week, y = total_clicks, color = factor(platform))) +
   geom_line() + geom_vline(xintercept = 9,color='black') +
  scale_y_continuous(sec.axis = sec_axis(~./6)) +
  scale_x_continuous(breaks = seq(1, 12, by = 1)) +
  theme(legend.title = element_blank())
```



We can observe from the plot that the total ad clicks for all platforms are increasing which might be due to an observed variable that is affecting all the platforms. To capture this effect of the omitted variable bias we need to compare the trend observed on Google with other platforms (Yahoo, Bing and Ask) using the difference-in-difference model taking into account the parallel trends across these platforms.

4. Calculate the Difference-in-Differences

We will first check the assumptions to back our DiD analysis. Hence we consider the following assumptions:

1. There aren't any SUTVA(Stable unit treatment value assumptions) violations. Given that all the

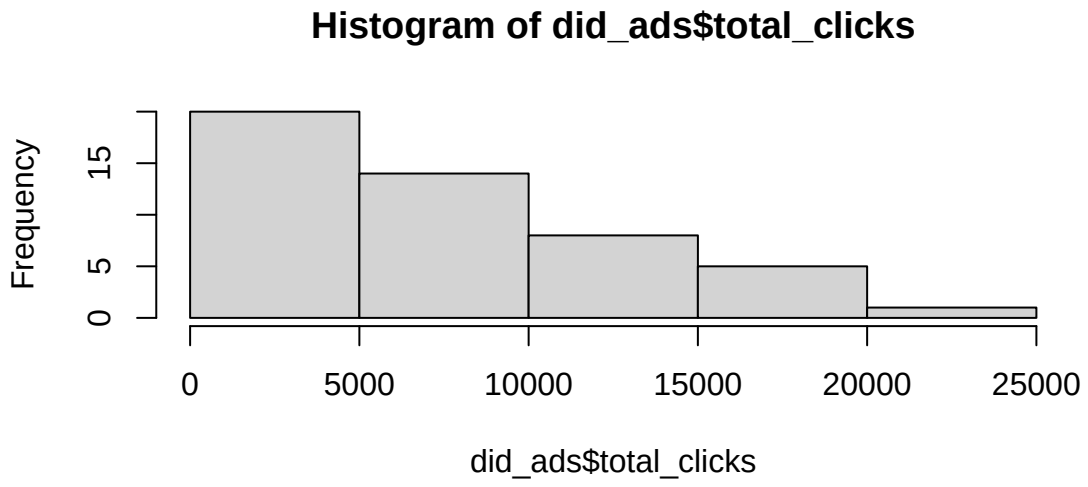
browsers are independent of each other so, there is not spillage of the data in terms of clicks on the ads.

2. There isn't any Anticipation effect i.e., the platform, google did not expect the unavailability of sponsored ads in week 10-12, given that it was due to a technical glitch.

3. The total clicks from different platforms are following a parallel trend, i.e., the total clicks for all the platforms were moving in parallel in the pre treatment period. To check the validity of this assumption we can perform a test to check the beta coefficients for pre and post period with treatment. for pre period, for Google the beta coefficients should be 0, that means Google is not affecting total clicks in pre period in any way.

Parallel Trends:

```
#Adding a treatment variable for our DiD analysis, where google being the treatment so, it'
did_ads <- did_ads %>% mutate(treatment = ifelse(platform == "goog", 1, 0))
#Check the distribution of total clicks
hist(did_ads$total_clicks)
```




```
summary(lm(total_clicks ~ treatment * factor(week),data=did_ads))
```

```
##
```

```
## Call:
## lm(formula = total_clicks ~ treatment * factor(week), data = did_ads)
##
## 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
```

As we can see the p-values for all interaction terms between treatment and weeks are greater than 0.05, so none of them are significant and the parallel trends assumption holds true.

Now, to perform the DiD analyses with the treatment effect, we are running regression on *log of total_clicks* as our dependent variable because the above histogram is right skewed, and on the right hand side we will have *after and treatment interaction*, to measure the treatment with pre and post period in consideration.

```
summary(lm(log(total_clicks)~ after*treatment, data = did_ads))
```

```
##
## Call:
## lm(formula = log(total_clicks) ~ after * treatment, data = did_ads)
##
## 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 ***
```

```
## treatment         0.5303     0.3085   1.719 0.092629 .
## after:treatment  -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
```

From the above results, we can say that in the post period, for Google the total clicks decreased by *(exp(-1.1163)-1)X100 = 67.25%* without the sponsored adds as compared to the total clicks for other platforms in the post period. This is our post treatment effect.

To measure the statistical significance of this effect, we look at the p-value, which is $0.07 > 0.05$ (the threshold p-value we consider to be in 95% confidence interval). So, we can say the effect discussed above is also statistically significant only if our confidence interval is 90%.

In the pre-post estimate, we saw the post effect for Google was -1846, i.e., the total ad clicks reduced by *1846/8391 = 22%* when there were no sponsored ads on Google in week 10-12. This effect value is different from the effect we estimated using DiD treatment analysis, in this the total clicks reduced for Google by 67.25% for weeks 10-12. The difference in the effect values is because the pre-post estimate does not consider the systematic effect on the total clicks including other platforms and other omitted variables into consideration. We use Did analysis to compare similar platforms(browsers) that may be impacted by some external factors in a similar fashion, so in did analysis the effect of those external factors are removed.

5. Improving the ROI Calculation

The increase in traffic due to organic ads is 67.25%, which we want to see is equivalent how many actual clicks:

```r
coefficients(lm(total_clicks~ after*treatment, data = did_ads))["after:treatment"]
```

```
## after:treatment
##       -9910.593
```

So, the sponsored ads decrease the total clicks by 9910. But this effect also consists of the effect of an increase in organic clicks due to removal of sponsored ads. To estimate this effect, we find the effect of treatment and after on averag organic ad clicks.

```r
coefficients(lm(avg_org~ after*treatment, data = did_ads))["after:treatment"]
```

```
## after:treatment
##        2293.222
```

So, the organic ads increase by 2293 in the absence of sponsored ads. Therefore, to estimate the correct ROI, we will calculate the correct traffic through sponsored ads by subtracting the number from the sponsored ads, which is equal to:

```r
correct_traffic = (1-(2293/(9910+2293)))
#ROI =
((correct_traffic*21*0.12)-0.6)/0.6
```

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

That is actual total ROI is 241% as compared to the 320% estimated earlier.