Assignment 3

Rithwik Sivadasn and Mohammed Pahadwala

Business Overview

A prominent online store in the US, Bazaar.com, runs sponsored search ads on Google, Bing, Yahoo, and Ask. They also have a strong presence in display advertising and search engine advertising. Paid advertisements on Bazaar can be broadly divided into two groups based on keywords: branded and nonbranded. Brand keywords, such as "Bazaar," "Bazaar shoes," "Bazaar clothes," and so on, contain the "Bazaar" brand name, whereas nonbranded keywords, such as "shoes" or "dress," do not. Bob, a member of Bazaar's marketing analytics team, calculated a 320% ROI on their sponsored ad spending using traffic data from Google and Bing. His findings are dubious because those who searched for "Bazaar" already intended to go to that website, making the efficacy of branded keyword ads improbable. The following inquiries will be answered for a thorough examination overall in order to help us understand the relationship between the cause and effect of the search ads and their efficacy: 1. What's wrong with Bob's ROI analysis? 2. What is the treatment and control of the experiment? 3. Is the First Difference reliable to estimate the causal effect of the treatment? 4. How should we compare with Difference-in-Difference estimation or Pre-Post estimation? 5. Based on our new treatment effect, what should be the corrected ROI? How is it compared with Bob's ROI?

Analysis Performed

To estimate the correct causal impact of sponsored ads on Bazaar.com's traffic, we used Difference-in-Difference through the following steps: 1) Calculate the first difference for weekly average traffic (Ads + Organic) over the course of before and after the technical hitch. This gives us the raw effect of sponsored ads of Google search instead of the overall treatment effect 2) Compare the first level pre-post difference in Google and the first level pre-post difference in other search engines. This determines the true incremental effect as this step handles possible confounders like seasonal variations across weeks as well as market factors.

Findings and Takeways

We see that without the sponsored advertisements, we loose a lot of clicks and almost 94% of them are new customers that would not have come through organic channel. ### (a) Solution: There might be lot of people who intend to make purchase from bazaar.com and would search for it and because we are having a sponsored ad they would use it to reach bazaar.com, This means people who would have used organic search to reach bazaar.com could also be using. sponsored ads to reach bazaar.com. According to the data provided we have two search engines Google and Bing. For both we have organic and sponsored click while for google the sponsored lick stopped in the ninth week According to the Bing data there is no sudden jump in number of clicks for the 10th, 11th and 12th week ,which eliminates possibility of Seasonality Factor or any other abrupt changed , Whereas in google data we have a sudden jump in organic clicks when the sponsored ads were down, this means that most of the people who are using navigational keywords to search will convert even if they don't get sponsored ads. Also, Bob is assuming that probability of making a purchase of 12% applies for the branded keyword searches as well even though people who are using branded keywords have a higher conversion rate than the average conversion rate of a person on the website For eg. Suppose 100 customers visit through the sponsored ads. ROI calculated by bob is ROI = (($21 * 0.12 * 100) - (100 * $0.6)) / (100 * $0.6) = 3.2 ROI = 320%

Although only 50 were those who were actually attracted by the advertisement while others were the one who were looking specifically for bazaar.com and clicked on the advertisement because it was visible. Hence, corrected ROI would be: corr_ROI = ((21 * 0.12 * 50) - (100* 0.6)) / (100 * 0.6) = 1.1 corr_ROI = 110%

(b)

Solution: Unit of observation: The unit of observation in our analysis is a search engine, i.e. Google, Bing, Yahoo and Ask. We have multiple observations (for weeks 1-12) for each unit in our data set. Treatment: Switching off of sponsored ads for google search engine from 10th to 12th week Control : The clicks received on sponsored ads for Yahoo, Bing and Ask for 10th to 12th week ### (c) Solution: Loading the required libraries

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

Reading the data

```
sponsored_ads = read.csv('C:/Users/mdphd19/Documents/Carlsons/Fall/Causal/Homework/Assig
```

Making data transformations

```
sponsored_ads_goog = sponsored_ads %>% filter(platform=='goog')
sponsored_ads_goog$after = if_else(sponsored_ads_goog$week < 10,0,1)
sponsored_ads_goog$total_clicks = sponsored_ads_goog$avg_spons+sponsored_ads_goog$avg_or
```

Running the linear regression model

```
summary(lm(log(total_clicks)~after,data = sponsored_ads_goog))
```

```
##
```

```
## Call:
## lm(formula = log(total_clicks) ~ after, data = sponsored_ads_goog)
##
## 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
```

Looking at the regression results above, we can clearly see the difference in the pre and post period change in traffic from google. The average decrease in the total clicks after the shutdown of sponsored ads 0.13%. We cannot say that this decrease is because of shutdown of sponsored ads as there could be a natural decrease in number of users in that period on the internet. For this we need to compare it with other channels as well. As the p-value is high, we cannot conclude the differences with and without the sponsored ads ### (d) Solution:As there are multiple control groups, we need to make one control group that would be the best estimation of the multiple groups. Converting to wide format to get a synthetic control group

```
sponsored_ads$total_clicks = sponsored_ads$avg_spons + sponsored_ads$avg_org
search.wide <- sponsored_ads %>% pivot_wider(id_cols=c("week"),names_from=c("platform"),
search.wide.train = subset(search.wide,week<10)
```

Getting the synthetic control group based on yahoo,bing and ask

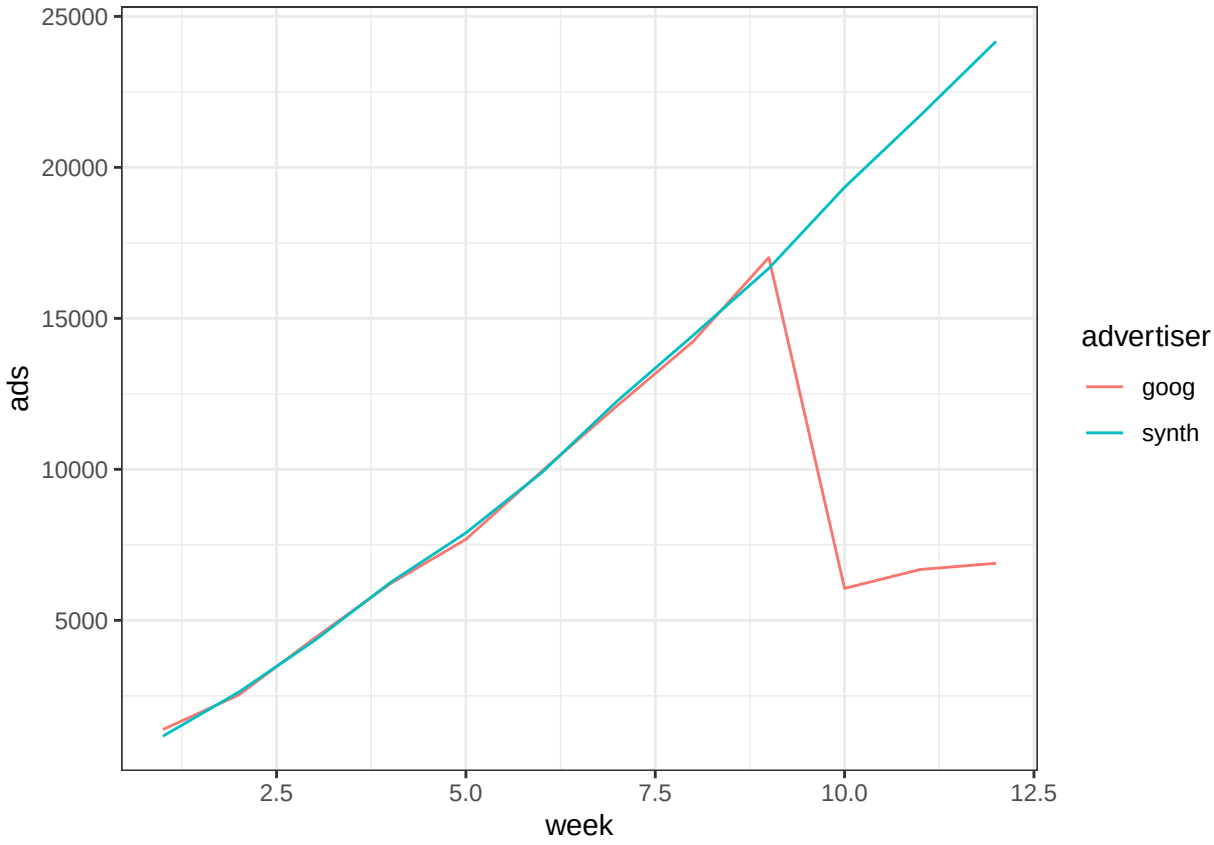
```
synth <- lm(formula=goog~yahoo+bing+ask,data=search.wide.train)
search.wide$synth <- predict(synth,newdata = search.wide)
```

Converting the data back into long format to perform regression

```
search_synth_long = search.wide %>% select(week,goog,synth) %>%
  pivot_longer(cols = c("goog","synth"),names_to = "advertiser",
               values_to = "ads")
search_synth_long$treatement = ifelse(search_synth_long$advertiser =='goog',1,0 )
search_synth_long$after = ifelse(search_synth_long$week>9,1,0)
```

Running a DID model with fixed effects within the advertiser and the week

```
ggplot(search_synth_long, aes(x = week, y = ads, color = advertiser)) +
  geom_line() +
  theme_bw()
```




```
options(repr.plot.width = 1, repr.plot.height =0.5)
summary(plm( ads ~ after*treatement, data = search_synth_long,
             model = "within",
             effect = "twoway",
             index = c("advertiser","week")))
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = ads ~ after * treatement, data = search_synth_long,
##     effect = "twoway", model = "within", index = c("advertiser",
##         "week"))
##
```

```
## Balanced Panel: n = 2, T = 12, N = 24
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -1.0422e+03 -8.7358e+01  4.5475e-13  8.7358e+01  1.0422e+03
##
## Coefficients:
##                    Estimate Std. Error t-value  Pr(>|t|)
## after:treatement -15200.70     609.58 -24.936 2.461e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    264120000
## Residual Sum of Squares: 4180400
## R-Squared:      0.98417
## Adj. R-Squared: 0.9636
## F-statistic: 621.812 on 1 and 10 DF, p-value: 2.4606e-10
```

DID assumes that:- The two groups are growing at similar rates before one is ”treated” Hence going with that assumption and looking at the value above, we can say that bazaar.com is loosing customers at an average rate of 15,200 clicks from week 10. It differs from the earlier pre-post estimate by taking in account the effect on clicks if the sponsored ads were not closed. It tells that the average clicks would increase with the previous trend and hence the difference is large. ### (e) Solution: From above, we will be using the same synthetic control to get average increase in organic clicks

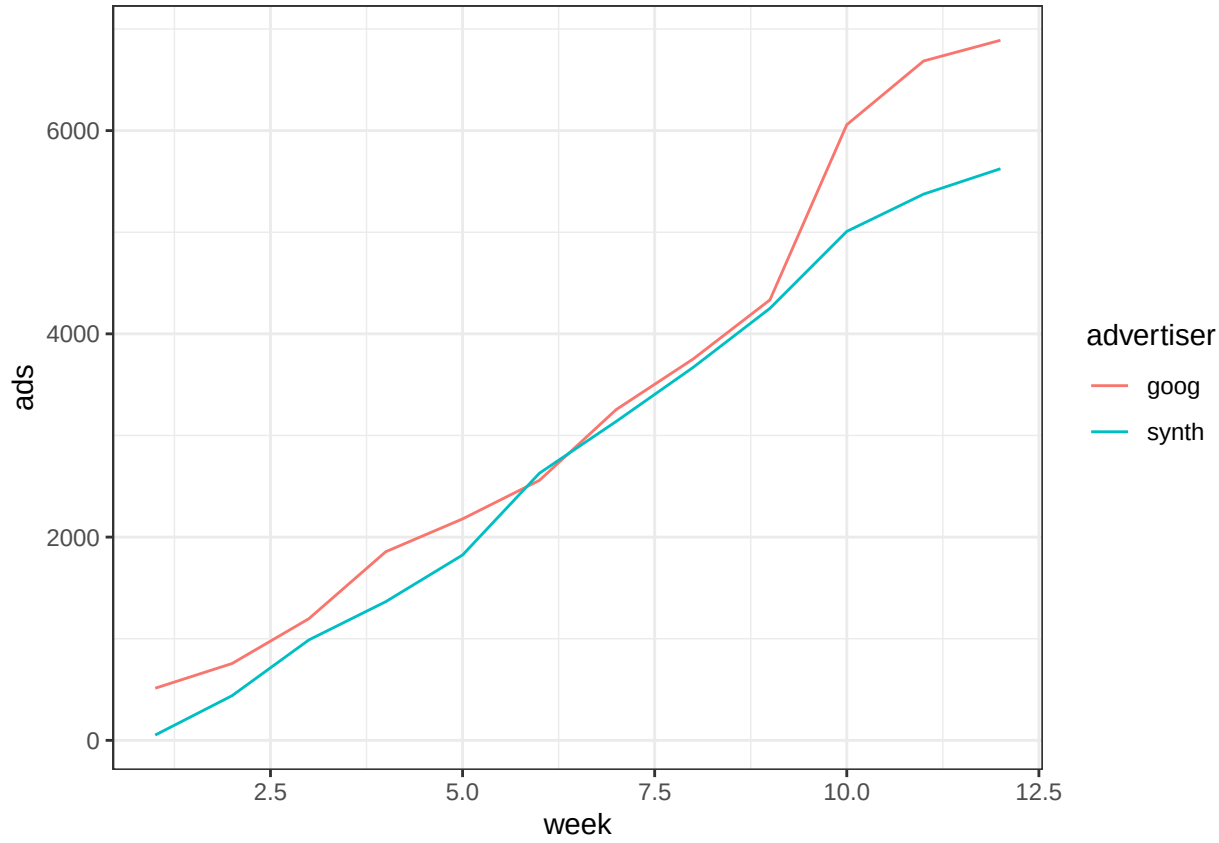
```
search.wide.organic <- sponsored_ads %>% pivot_wider(id_cols=c("week"),names_from=c("pla
search.wide.organic$synth <- predict(synth,newdata = search.wide.organic)
```

Converting the data back into long format to perform regression

```
search_synth_org_long = search.wide.organic %>% select(week,goog,synth) %>%
  pivot_longer(cols = c("goog","synth"),names_to = "advertiser",
               values_to = "ads")
search_synth_org_long$treatement = ifelse(search_synth_org_long$advertiser =='goog',1,0
search_synth_org_long$after = ifelse(search_synth_org_long$week>9,1,0)
```

Running a DID model with fixed effects within the advertiser and the week to get the increase in organic advertisements

```
ggplot(search_synth_org_long, aes(x = week, y = ads, color = advertiser)) +
  geom_line() +
  theme_bw()
```



```
summary(plm( ads ~ after*treatement, data = search_synth_org_long,
             model = "within",
             effect = "twoway",
             index = c("advertiser","week")))
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = ads ~ after * treatement, data = search_synth_org_long,
##     effect = "twoway", model = "within", index = c("advertiser",
##         "week"))
##
## Balanced Panel: n = 2, T = 12, N = 24
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -1.4890e+02 -6.6073e+01  1.1369e-13  6.6073e+01  1.4890e+02
##
## Coefficients:
##                  Estimate Std. Error t-value  Pr(>|t|)
## after:treatement   980.95     121.23  8.0917 1.065e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1247900
## Residual Sum of Squares: 165340
## R-Squared:      0.86751
```

```
## Adj. R-Squared: 0.69526
## F-statistic: 65.4753 on 1 and 10 DF, p-value: 1.0655e-05
```

From above, we see that 980 clicks increased because of treatment effect, i.e closing of sponsored ads. These would be the clicks that would have come from sponsored ads but then got shifted to organic channel. Earlier in question 4 we got 15200 clicks that would be reduced because of shutting down of sponsored ads. From this we can say that the increase in organic clicks are the ones that would have been part of sponsored ads but got shifted to the organic channel. Getting the ratio of clicks in the sponsored ads that would have been organic ads comes out to be : corrected ROI is: corr_ROI = ((21 * 0.12 * (1- 980/15200) - (1* 0.6)) / (1 * 0.6)

```
((21 * 0.12 * (1- 980/15200)) - (1* 0.6)) / (1 * 0.6)
```

```
## [1] 2.929211
```

Corrected ROI is 292%