Difference-in-Differences – Bazaar.com

Raghuram Sirigiri

Business Overview

US retailer Bazaar uses both display and search advertising running paid search ads on Google and Yahoo. Bazaar releases its ads in response to key words used by customers and are classified broadly into Branded(such as 'Bazaar', 'Bazaar shoes', 'Bazaar clothes' and so on) and Non Branded keywords(such as 'shoes', 'dress' that do not contain 'Bazaar').

Using data from Google and Bing, Bazaar's marketing analytics team computed an ROI of 320% on their sponsored ad spending. But were skeptical because some the the user who searched bazaar might already have the intent visit Bazaar.com. This is an confusion they would like the analyst to clear. The goal is to understand the causal effect of the search ads.

Question to be answered:

1. What is wrong with Bob's ROI analysis?
2. What is treatment and control in the experiment?
3. Consider First Difference Estimate and is it a good estimate?
4. Calculate the Difference-in-Differences?
5. What if the fixed ROI Calculation?

Data Overview

Loading Packages

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

Loading and Transforming data

```
data_2 = read.csv("did_sponsored_ads.csv")
data_2 = data_2%>%
  mutate(treat = if_else(platform %in% c('goog'), 1, 0),
```

```
        post = if_else(week < 10, 0, 1),
        avg_total = avg_spons+avg_org)
```

Following data points were given for the data:

1. id - ID of the search engine
2. platform - Name of the search engine
3. week - Count of the week starting from 1
4. avg_spons - Count of clicks received form sponsored ads
5. avg_org - Count of clicks received from organic clicks

Transformed variables: 6. treatment - 1 for rows which are part of treatment else 0 7. post - 1 for weeks after 9 else 0 8. avg_total - Sum of click from sponsored and organic clicks

Sponsored clicks and organic clicks data is give for four different search engines Google, Yahoo, Bing, Ask and because of some glitch the ads on Google didn't work after week 9. Users did not see ads of Bazaar.com after week 9. So the data is zero for advertisement clicks after week 9 for Google.

Analysis Performed

1. Pre_post difference is calculated using the first difference approach.
2. Difference-in-Differences regression is performed to get the estimate.

Problem With Bob's ROI Calculation

The problem with Bob's investment ROI calculation is that the return value used in the calculations may not be the actual value realized with the Advertisements. User will see th is if they are searching for the specific key words related to Bazaar.com. So most the users may be interested in Bazaar.com and searching for it. So these users are not the users coming to the website just because they saw the advertisement. These might still visit the website even if the website advertisement is not there and the value generated by them is not the actual value.

Bob's ROI = $((21 * 0.12) - (0.6)) / (0.6) = 320\%$ $

1. This assumes that all 100 percentage of clicks are from user who do not have intent. But that is not the case as some of the branded words are used in publishing ads so users who search and have the intent to visit the site will also see these ads.
2. Some of the user might have seen the ad but did not react immediately and later searched for the website to visit. Since it a pay per click model that cost is not incurred by Bazaar.com.

Treatment and Control Definition

Unit of Observation

The search engine the users are s.earching keywords

Treatment

Treatment is the stop of sponsored ads in Google starting from week 10.

Control

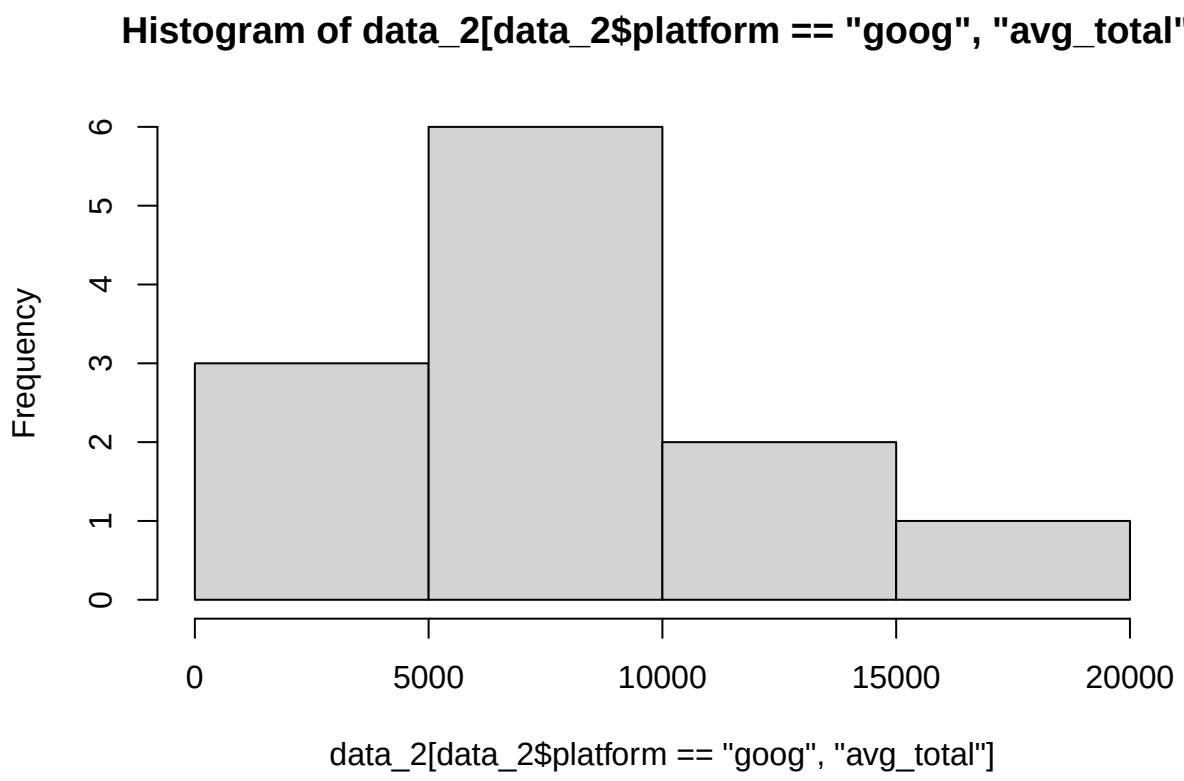
Clicks from Yahoo, Bing and Ask are used as control

First Difference Estimate

Using the data of only Google first difference estimate is calculated.

Data Distribution

```
hist(data_2[data_2$platform=='goog',"avg_total"])
```



Data is skewed so we can use log transform while performing analysis.

Analysis

```
model = lm(log(avg_total)~post, data = data_2[data_2$platform=='goog',])

summary(model)
```

```
##
## Call:
## lm(formula = log(avg_total) ~ post, data = data_2[data_2$platform ==
##     "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 ***
## post        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
```

```
print(paste("Percentage Change: ",(exp(model$coefficients[2])-1)*100))
```

```
## [1] "Percentage Change:  0.130697194056317"
```

Interpretation

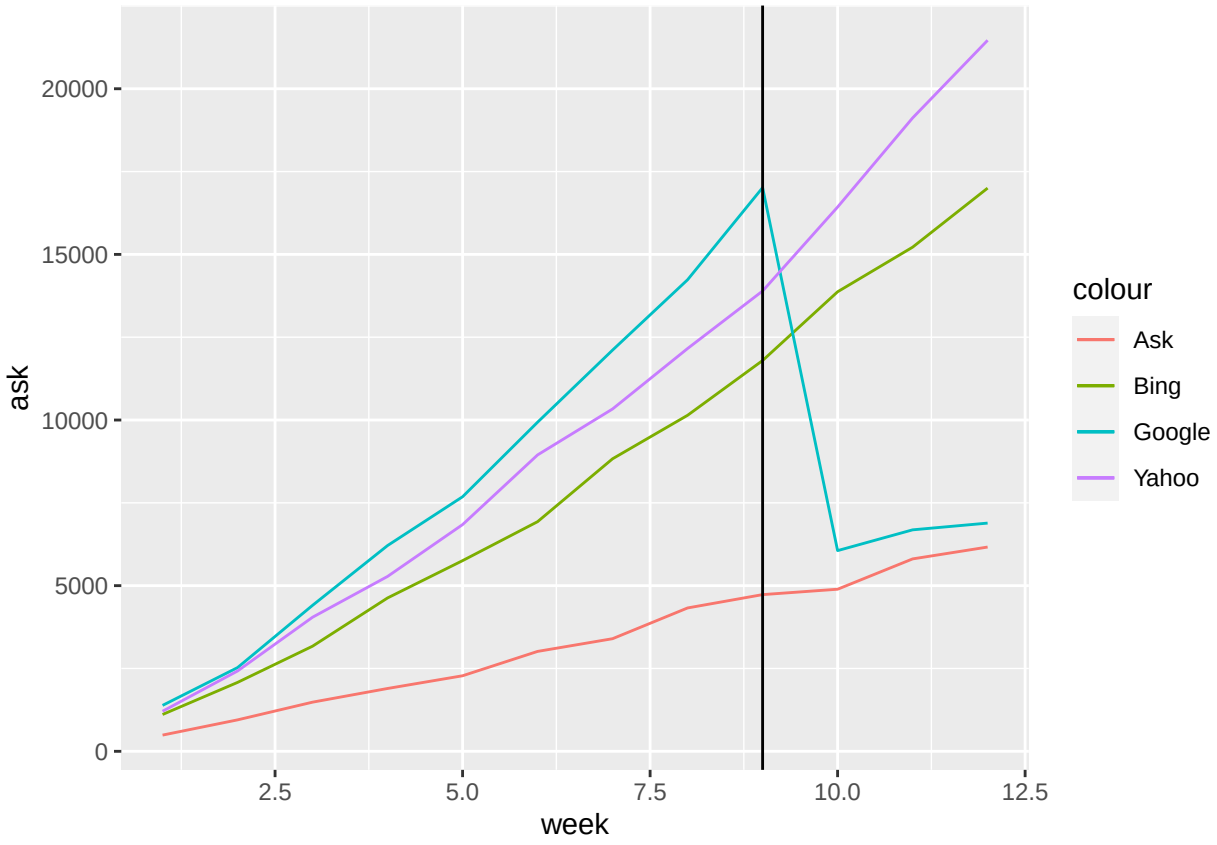
From the above we get that the percentage change in the total clicks is 0.13% and the p-value is 0.998(>0.05). As the p-value is not significant we cannot conclude anything from this analysis.

Calculate the Difference-in-Differences

Parellel Trends Assumption Check

To conduct Difference-in-Differences analysis we need to verify parallel trends assumption.

```
wide_data=spread(data_2%>%select(week, platform, avg_total), platform, avg_total)
ggplot(wide_data, aes(week)) +
  geom_line(aes(y = ask, colour = "Ask")) +
  geom_line(aes(y = bing, colour = "Bing")) +
  geom_line(aes(y = goog, colour = "Google")) +
  geom_line(aes(y = yahoo, colour = "Yahoo")) +
  geom_vline(xintercept = 9)
```



As we can see that the data is not parallel before week 9 so the parallel trends assumption failed in this case.

As there are multiple platforms in control group we can use synthetic control instead of DiD to calculate

Creating Synthetic Data

```
synth_model <- lm(formula=goog~yahoo+bing+ask,data=wide_data[wide_data$week<10,])
wide_data$synth <- predict(synth_model,newdata = wide_data)
summary(synth_model)
```

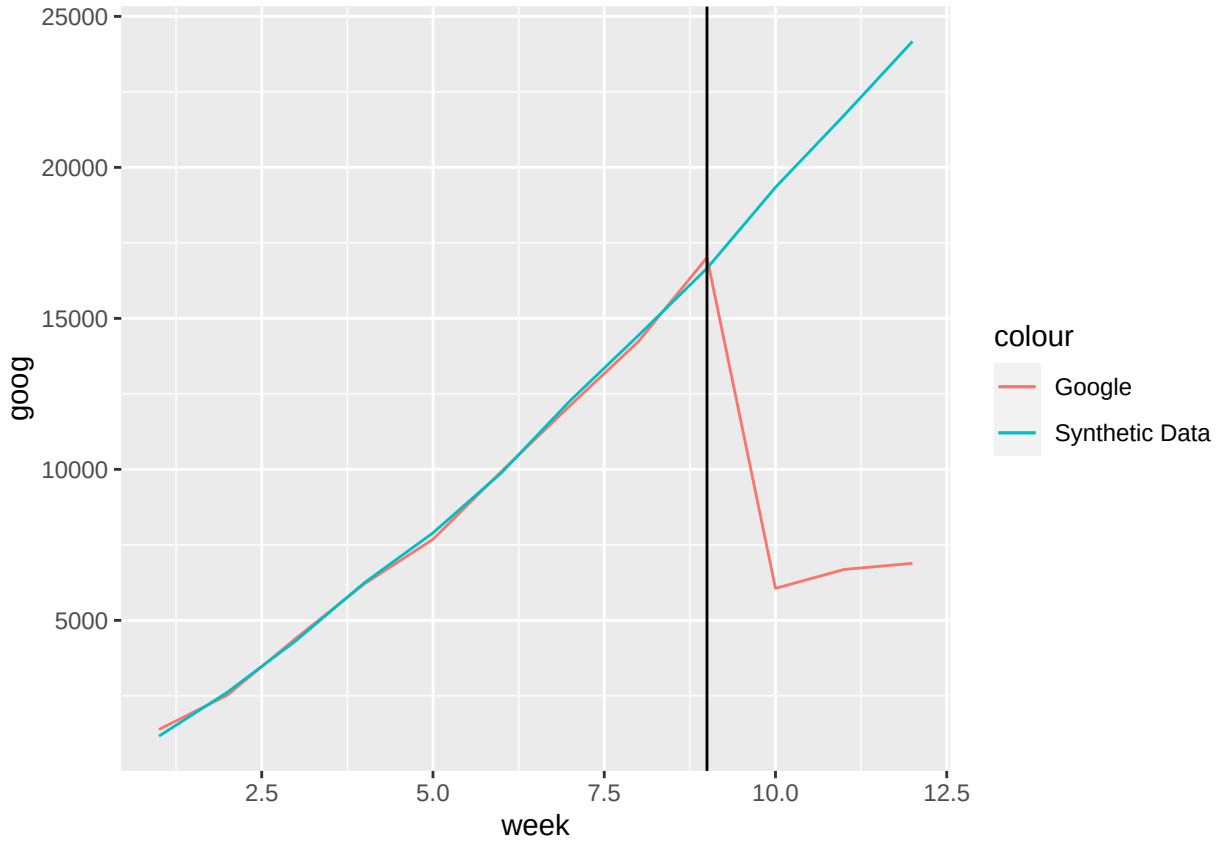
```
##
## Call:
```

```
## lm(formula = goog ~ yahoo + bing + ask, data = wide_data[wide_data$week <
##     10, ])
##
## Residuals:
##       1       2       3       4       5       6       7       8       9
##  221.60  -94.70   79.09  -36.73 -212.43   44.95 -157.42 -202.40  358.04
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -447.6371   187.8053  -2.384   0.0629 .
## yahoo          0.2125     0.4302   0.494   0.6423
## bing           0.9939     0.3597   2.763   0.0397 *
## ask            0.5132     0.9337   0.550   0.6062
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 247.7 on 5 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9979
## F-statistic:  1244 on 3 and 5 DF,  p-value: 1.332e-07
```

```
long_data = gather(wide_data, key='platform', value='avg_total', goog, synth)%>%
  select(week, platform,avg_total) %>%
  mutate(treat = if_else(platform %in% c('goog'), 1, 0),
         post = if_else(week < 10, 0, 1))
```

Chcecking Parellel Trends for Synthetic Data

```
ggplot(wide_data, aes(week)) +
  geom_line(aes(y = goog, colour = "Google")) +
  geom_line(aes(y = synth, colour = "Synthetic Data")) +
  geom_vline(xintercept = 9)
```



From the above plot parallel trends can be seen between synthetic data and Google data.

Difference-in-Differences using synthetic data

```
model = plm( log(avg_total) ~ post*treat, data = long_data,
             model = "within",
             effect = "twoway",
             index = c("platform","week"))

summary(model)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = log(avg_total) ~ post * treat, data = long_data,
##     effect = "twoway", model = "within", index = c("platform",
##         "week"))
##
## Balanced Panel: n = 2, T = 12, N = 24
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
```

```
## -0.080376 -0.013338  0.000000  0.013338  0.080376
##
## Coefficients:
##              Estimate Std. Error t-value  Pr(>|t|)
## post:treat -1.211633   0.040595 -29.847 4.167e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1.6701
## Residual Sum of Squares: 0.018539
## R-Squared:      0.9889
## Adj. R-Squared: 0.97447
## F-statistic: 890.856 on 1 and 10 DF, p-value: 4.1672e-11
```

```
print((exp(model$coefficients[1])-1)*100)
```

```
## post:treat
##  -70.22892
```

Interpretation

Here the value of the coefficient of the interaction term is -1.21 with $p<0.05$, hence we can conclude that this is significant. Using the coefficient we determine that the percentage of total users lost if ads are stopped is ~70%.

Fixed RoI Calculation.

Percentage users from Sponsored Ads

```
data_2 %>% filter(platform=='goog') %>% summarize(mean(avg_spons)/mean(avg_total)*100)
```

```
##   mean(avg_spons)/mean(avg_total) * 100
## 1                              57.9235
```

From the above 57 percent users on average come from sponsored links.

Updated Calculation

In this case we can see that the we are losing 70% of the user if we stop posting ads on Google. But on average we are paying for 58% user come from the ad clicks for which bazaar.com is paying.

So of the 100 users that visit the website:

- 70 are coming because of search ads
- 58 users are clicking on links and amount is paid for those users.

So return on Investment is:

```
((21 * 0.12 * 70) - (0.6 * 58)) / (0.6 * 58)*100
```

```
## [1] 406.8966
```

The return on investment is 406% for the paid ads on Google.