Difference-in-Differences Regression Analysis on Sponsored Search Ads

Yi Hsiang (Royce) Yen

2025-04-08

Overview

This analysis leverages a natural experiment described in the Columbia Business School case "Measuring ROI on Sponsored Search Ads" by Kinshuk Jerath. The objective is to estimate the causal impact of sponsored search advertising on Bazaar.com's web traffic and conversion, using a Difference-in-Differences (DiD) regression framework. In particular, the analysis aims to provide a more accurate assessment of the Return on Investment (ROI) of branded keyword advertising campaigns.

Case Introduction

Bazaar.com had been running sponsored ad campaigns on both Google and Bing for 12 weeks. However, due to a technical glitch, the Google campaign was suspended starting in week 10, while the Bing campaign continued uninterrupted. The dataset reflects weekly traffic to Bazaar.com generated by consumers clicking on either sponsored or organic links after searching for branded keywords (e.g., "Bazaar shoes") on these platforms. In such searches, both types of links typically appear on the results page.

Bob, a member of the marketing analytics team, conducted a simple ROI analysis using the following assumptions:

- Average cost per sponsored click: 0.60 USD
- Probability of conversion per click: 12%
- Average margin per conversion: 21 USD

He concluded that each sponsored click generated an average revenue of 2.52 USD, resulting in an ROI of 320%, calculated as: (2.52 - 0.60) / (0.60) = 320%

While this number seems impressive, it led the team to question the validity of the assumptions and methodology. Two main concerns emerged:

1. Is Bob's ROI calculation valid, or does it overstate the true impact of advertising?
2. Do sponsored ads actually generate incremental traffic (sponsored + organic), or are they simply redirecting traffic that would have arrived via organic clicks anyway?

Key Questions

In this report, I applied a DiD regression methodology to address these core business questions in the context of Bazaar.com's sponsored search advertising. Specifically, the analysis is structured to address the following sub-questions:

(a) What is wrong with Bob's ROI calculation?

(b) Define the treatment and control groups. What is the unit of observation, and which units are treated or not?

(c) Estimate and interpret the First Difference (pre-post) effect using only the treated group. Why is it problematic to rely solely on this estimate?

(d) Estimate and interpret the treatment effect using a Difference-in-Differences regression.

(e) Propose a revised ROI calculation based on the causal treatment effect estimated through DiD analysis.

Question (a): What is wrong with Bob's ROI calculation?

The main issue with Bob's calculation is that it assumes all conversions from sponsored clicks are incrementally caused by the ads, treating correlation as causation. However, this overlooks the counterfactual: for branded keywords like 'Bazaar shoes,' many users might have clicked the organic link if the sponsored ad were absent, as they are already searching for Bazaar.com. This substitution effect means Bazaar could acquire some customers at no cost, suggesting Bob's ROI of 320% overstates the true causal impact by failing to account for cannibalization of organic traffic. A more rigorous method, such as Difference-in-Differences, is needed to isolate the ads' incremental contribution, which may differ significantly from Bob's estimate.

Question (b): Define the treatment and control groups. What is the unit of observation, what is the treatment, and which units are treated or not?

I'll define the following components for our DiD analysis:

- Unit of Observation: Each observation corresponds to a platform-week — that is, one row per platform (e.g., Google, Bing, Yahoo, Ask) per week (1–12). For each unit, we observe the sum of the number of sponsored clicks and organic clicks, which we define as total clicks.
- Treatment: The treatment is defined as the absence of sponsored ads on a platform during a given week. i.e. Whether we stopped running sponsored ads on that platform.
- Treatment units: Google in weeks 10–12, because this is where the sponsored ads were turned off due to the glitch.
- Control units = Bing, Yahoo, Ask in all weeks, because they never stopped their sponsored ads
- Outcome Variable of Interest: Total traffic (sponsored + organic clicks)

```
# Load Library & Data Set
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
```

```
data = read.csv('did_sponsored_ads.csv')
data <- data %>% mutate(total_clicks = avg_spons + avg_org)
```

Question (c): Estimate and interpret the First Difference (pre-post) effect using only the treated group. Why is it problematic to rely solely on this estimate?

If we could only observe the treated unit, an approach would be to calculate the first difference(i.e. the % change in web traffic arriving from Google; (after-before)/before). This estimate is the pre-post difference in the treated cohort.

We can estimate this using a regression:

```
# 1. Filter data for only Google
google_data <- data %>% filter(platform == "goog")
# 2. Create a binary column "post" to indicate
# whether this week is pre or post the treatment
google_data <- google_data %>% mutate(post = ifelse(week > 9, 1, 0))
# 3. Regress total clicks on post, and interpret the coefficient on post
summary(lm(total_clicks ~ post, data = google_data))
```

```
##
## Call:
## lm(formula = total_clicks ~ post, data = google_data)
##
## 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)     8390       1598   5.252 0.000373 ***
## post           -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
```

```r
first_difference = (-1846)/8390
print(first_difference)
```

```
## [1] -0.2200238
```

According to this simple pre-post estimator, pausing sponsored ads on Google is associated with a 1,846-click decrease in weekly total clicks, or a 22% reduction (-1,846 / 8,390).

However, this estimate is problematic because it does not account for market-wide trends or seasonality that might affect traffic independently of the treatment. For instance, control platforms (Bing, Yahoo, Ask) show increased traffic in weeks 10-12, suggesting a general upward trend. Relying solely on this estimate conflates the treatment effect with these external factors, making it an unreliable measure of causality.

This estimate reflects only the pre-post difference within the treated group and does not control for market-wide trends or seasonality. As such, it cannot be interpreted as a causal effect of the treatment, but rather serves as a preliminary correlation. A more rigorous approach is needed to isolate the treatment effect — which we address through a DiD analysis in the next section.

Question (d): Estimate and interpret the treatment effect using a Difference-in-Differences regression.

Next, I'll follow the steps to calculate the difference in differences estimate of the treatment effect.

```r
# 1. Create a binary column "post" to indicate
# whether this week is pre or post the treatment
data <- data %>% mutate(post = ifelse(week > 9, 1, 0))
# 2. Create a binary column "treatment" to indicate
# whether this platform is the treatment group "Google" or not
data <- data %>% mutate(treatment = ifelse(platform == "goog", 1, 0))
# 3. Regress total clicks on post + treatment
did_model <- lm(total_clicks ~ post*treatment, data = data)
summary(did_model)
```

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

The DiD analysis is based on the following regression model: Avg_total_weekly_clicks = 5265 + 8064.7 * post + 3124.9 * treatment -9910.6 * post * treatment

We can derive the following insights by interpreting the coefficients:

- *Regardless of the effect of the treatment, Google has a higher baseline of weekly clicks then the control group (Bing, Yahoo, Ask combined).*
  - Google has a weekly click of 5265 + 3124.9 = 8,389.9, while the control group (Bing, Yahoo, Ask) has a weekly click of only 5265 clicks. The difference (3124.9 clicks) reflects Google as a platform gets more traffic.
- *The control group (Bing, Yahoo, Ask) experienced an increase of 8,064.7 total weekly clicks in the post-treatment period.*
  - For Bing, Yahoo, Ask, the total weekly clicks pre-treatment is 5265 clicks; the total weekly clicks post-treatment is 5265 + 8064.7 = 13,329.7. Therefore, the increase is 13329.7 - 5265 = 8064.7, reflecting a general upward trend in traffic across these platforms during weeks 10-12.
- *The estimated treatment effect is 9910.6 total weekly clicks.*
  - This indicates that, relative to the control platforms (Bing, Yahoo, Ask), pausing sponsored ads on Google is associated with a reduction of 9910.6 total clicks per week. In contrast, the simple pre-post estimate (using only Google data) showed a decline of only 1846 clicks per week. The Difference-in-Differences estimate is thus substantially larger in magnitude.

The pre-post estimator, which only compares Google's pre- and post-treatment traffic, assumes that Google's traffic would have remained constant had sponsored ads not been paused. However, this assumption is flawed, as it ignores the observed changes on other platforms during the same period.

The control group's total clicks increased by 8,064.7 in the post-treatment period (from 5,265 to 13,329.7), reflecting a market-wide upward trend. In contrast, the DiD approach accounts for this trend by subtracting it from Google's observed change, providing a more credible counterfactual for what Google's traffic would have been without the treatment.

This comparison allows us to isolate the causal effect of pausing sponsored ads, rather than conflating it with platform-invariant time effects that also impacted the rest of the market. Thus, the larger magnitude of the DiD estimate highlights the bias in the pre-post method, which underestimates the true causal impact by failing to control for external time effects.

Consequently, the DiD estimate offers a more accurate and causally valid measure of the impact of ad suspension, leveraging real behavioral data from similar, unaffected platforms rather than assuming a general "market trend."

Question (e): Propose a revised ROI calculation based on the causal treatment effect estimated through DiD analysis

To revise the ROI calculation, we first need to clarify a few things.

1. This ROI estimation is only applicable on the ROI of running sponsor ads on Google.

- Since Google was the only platform that received the treatment, this Difference-in-Differences causal analysis is only applicable for Google. For other platforms, we don't know the exact estimate of the incremental difference of sponsor ads will have on these platforms.

2. We have to update the correct amount of clicks used during our ROI calculation.

- We can still follow the formula ROI = (Revenue - Cost) / Cost. However, revenue and cost should be updated accordingly.
  - Revenue: The revenue should be the increased revenue if we continue to run sponsor ads on Google. In other words, this is the incremental increase of total weekly clicks, which according to our DiD analysis, is 9910.6 clicks. Therefore the revenue is 9910.6 * 21 * 0.12 = 24,974.712 USD
  - Cost: The cost should be the cost we need to cover if we continue to run sponsor ads on Google. To get this number, we shouldn't use the incremental increase of total weekly clicks. Instead, we should think about "How much will it cost if we run sponsor ads on Google." This can be estimated by calculating the average sponsored clicks during week 1~9. By calculating this, we get a better sense of how much it will actually cost for us to run sponsor ads per week.

```
# Calculate the average sponsored clicks during week 1~9
google_spons_avg_weekly_cost = data %>% filter(treatment == 1, post == 0) %>% summarize(mean(avg_spons)
print(google_spons_avg_weekly_cost)
```

```
##   mean(avg_spons)
## 1        6123.222
```

Therefore the cost is 6123.222 * 0.6 = 3,673.9333 USD

```
- ROI: The ROI can therefore be updated as (24,974.712 - 3,673.9333) / (3,673.9333) = 5.798 = 579.8%
```

Unlike Bob's approach, which attributes all sponsored clicks to the ads (yielding 320%), this revised ROI uses the DiD-estimated incremental clicks (9,910.6) as the true contribution of sponsored ads, avoiding overestimation from cannibalized organic traffic. The cost reflects Google's pre-treatment average sponsored clicks (6,123.222), as this represents the actual ad spend if the campaign continued. Thus, ROI = (24,974.712 - 3,673.9333) / 3,673.9333 = 5.798 = 579.8%, should be a more accurate measure of the ads' causal impact.

Conclusion

This analysis demonstrates that sponsored ads on Google generate a substantial incremental traffic increase (9,910.6 clicks/week), yielding an ROI of 579.8%, far exceeding Bob's overstated 320%. The DiD method provides a causal estimate by controlling for market trends, highlighting the value of analytics in advertising decisions.