

# Assignment 3: Difference-in-Differences – Bazaar.com

Apurva Baru and Sameeksha Mohan

2025-04-10

## Measuring ROI on Sponsored Search Ads

### Introduction

Bazaar.com is an e-commerce platform that heavily relies on online traffic to drive sales. In an effort to assess the effectiveness of its sponsored ad campaigns, the company unintentionally ran a natural experiment: sponsored ads on Google were paused for a period of 3 months, while ads on other platforms continued uninterrupted. This unexpected pause presents an opportunity to evaluate the causal impact of sponsored ads on overall website traffic.

As an analyst, our task was to quantify how the suspension of ads affected traffic arriving from search platforms - specifically Google and determine whether the investment in sponsored advertising is justified. To do this, we applied causal inference techniques such as pre-post comparisons, parallel trends check, and Difference-in-Differences (DiD). These methods allow us to separate the true effect of the advertising pause from other underlying trends in traffic.

This document summarizes our analysis, describes the natural experiment, and presents evidence based recommendations for management based on empirical results.

### Understanding the current ROI calculation

Bob's current method for calculating return on investment (ROI) from sponsored search ads is conceptually flawed. He calculates ROI by simply dividing the revenue attributed to ad clicks by the total cost of those ads. While straightforward, this approach assumes all traffic and revenue from sponsored ads is incremental, meaning it would not have occurred without the ads.

However, in practice, many users who clicked on a sponsored link were already planning to visit Bazaar.com. Without the ad, they might have:

1. Clicked on the organic search result, which appears just below the sponsored link.
2. Typed the website URL directly into the browser.
3. Returned later via another unpaid channel.

Bob's ROI calculation ignores this counterfactual scenario. As a result:

1. He overstates the impact of the ads by attributing all observed sales to them.
2. He fails to isolate the true causal effect the portion of traffic and sales that would not have occurred without the ad.

To conclude, Bob's method captures both incremental and non-incremental revenue, leading to an inflated ROI. A more accurate assessment requires comparing the treated and untreated outcomes over time - which we do using causal inference techniques in the following analysis.

## Treatment and Control Definition

In the context of this case study, the goal is to evaluate the impact of sponsored search ads on web traffic to Bazaar.com. To understand the impact of turning off sponsored search ads, we compared traffic patterns across different platforms and time periods which is our unit of measure.

The analysis focuses on different search platforms:

- Google is our treatment group, because the sponsored ads were turned off unintentionally starting in week 10.
- Other platforms are in our control group, because sponsored ads continued as usual throughout the entire period.

We tracked weekly website traffic on all the platforms over 12 weeks. This gives us a natural comparison:

- Before week 10, all platforms had sponsored ads running.
- After week 9, Google had no sponsored ads, while other platforms remained unchanged.

By comparing changes in traffic over time between Google and other platforms, we can estimate the true impact of turning off the ads - separating it from other seasonal or industry-wide changes.

## Preparing the Dataset

The dataset contains weekly web traffic information for Bazaar.com across different platforms over a 12 week period. For each platform and week we have the average number of visitors coming through sponsored ads and organic (unpaid) search results. From these, we compute the total traffic to the site per week. A key event in this period is that sponsored ads on Google were turned off after Week 9 due to a glitch, creating a natural experiment to analyse how that change impacted total website traffic.

```
#importing required libraries and data columns  
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(ggplot2)
library(plm)
```

```
## Warning: package 'plm' was built under R version 4.4.3
```

```
##
## Attaching package: 'plm'
```

```
## The following objects are masked from 'package:dplyr':
##
##      between, lag, lead
```

```
data = read.csv("did_sponsored_ads.csv")
data$post = ifelse(data$week >= 10, 1, 0)
data$total_traffic <- data$avg_spons + data$avg_org
data$treated <- ifelse(data$platform == "goog", 1, 0)
data$log_total_traffic <- log(data$total_traffic + 1)
```

Website traffic data often has a right skewed distribution, meaning there are a few weeks or platforms with unusually high traffic compared to the rest. This kind of skew can distort statistical results and make it harder to detect the true impact of changes- like turning off sponsored ads.

To address this, we:

- Checked the skewness of the total traffic variable to assess whether the data was heavily skewed.
- Applied a log transformation to compress extreme values and make the distribution more symmetrical.

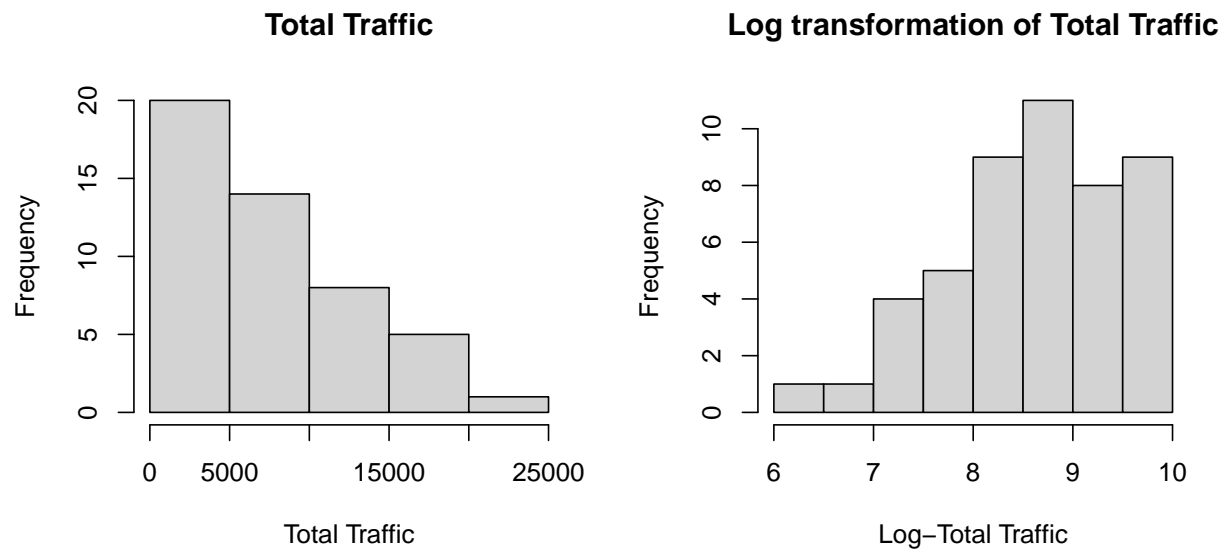
This transformation helps:

- Stabilize variance across time and platforms
- Make patterns in the data clearer
- Improve the reliability of our regression estimates

By using the log of traffic rather than raw traffic numbers, we ensure that our results are more robust and easier to interpret when comparing percentage changes in traffic.

```
#checking the skewness of the data
par(mfrow = c(1, 2))

hist(data$total_traffic, main = "Total Traffic", xlab = "Total Traffic")
hist(data$log_total_traffic, main = "Log transformation of Total Traffic", xlab = "Log-Total Traffic")
```



```
par(mfrow = c(1, 1))
```

### First Difference Estimate - A simple pre-post analysis on Google

To begin understanding the potential impact of pausing sponsored ads on Google, we performed a simple pre-post analysis using only data from the treated platform. We calculated a first difference estimate, which captures the average change in web traffic before and after the sponsored ads were turned off. This was done by running a regression of total traffic on a post indicator (1=after the ad pause, 0=before) using only Google data.

```
#filtering data for google and performing first difference estimate
data_google<- data %>% filter(platform=="goog")

reg<- lm(log(total_traffic+1) ~ post, data = data_google)
summary(reg)
```

```
##
## Call:
## lm(formula = log(total_traffic + 1) ~ post, data = data_google)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54882 -0.15494  0.03783  0.46963  0.95819
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.783723   0.248907  35.289 7.92e-12 ***
## post         0.001243   0.497814   0.002  0.998
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7467 on 10 degrees of freedom
```

```
## Multiple R-squared:  6.235e-07, Adjusted R-squared:  -0.1
## F-statistic: 6.235e-06 on 1 and 10 DF,  p-value: 0.9981
```

**Regression Results:** The regression output shows that the coefficient on post is 0.0012, which is statistically insignificant ( $p=0.998$ ). This suggests almost no detectable change in total traffic before versus after the add was turned off on Google.

**Interpretation:** At first glance, this might seem like the ad pause had no effect. However, this simple estimate does not account for any external trends or seasonal changes in traffic that may have affected both platforms over time. For example, overall search volume might have changed during the period due to unrelated factors like promotions, holidays or broader internet trends.

**Conclusion:** While the pre-post estimate suggests no major traffic drop, we cannot rely on this result alone to judge the impact of removing sponsored ads. A more robust approach, like Difference-in-Differences, allows us to compare changes on Google with a similar platform that did not undergo the ad pause, helping us better isolate the true effect of the treatment.

## Parallel Trends check

To ensure that our Difference-in-Differences(DiD) analysis is valid, we need to verify that traffic trends for the treatment and control platforms were similar before the intervention (i.e. the parallel trends assumption).

A regression analysis on pre-treatment data reveals that Google had a significantly faster-growing traffic trend compared to the control platforms, even before the change in ad policy. This violates the parallel trends assumption and suggests that a simple DiD estimate might overstate the causal effect of sponsored search ads. We advise interpreting the treatment effect with caution and recommend exploring additional methods or robustness checks to account for these diverging pre-treatment trends.

```
#parallel trends

pre_data <- subset(data, week <= 9)

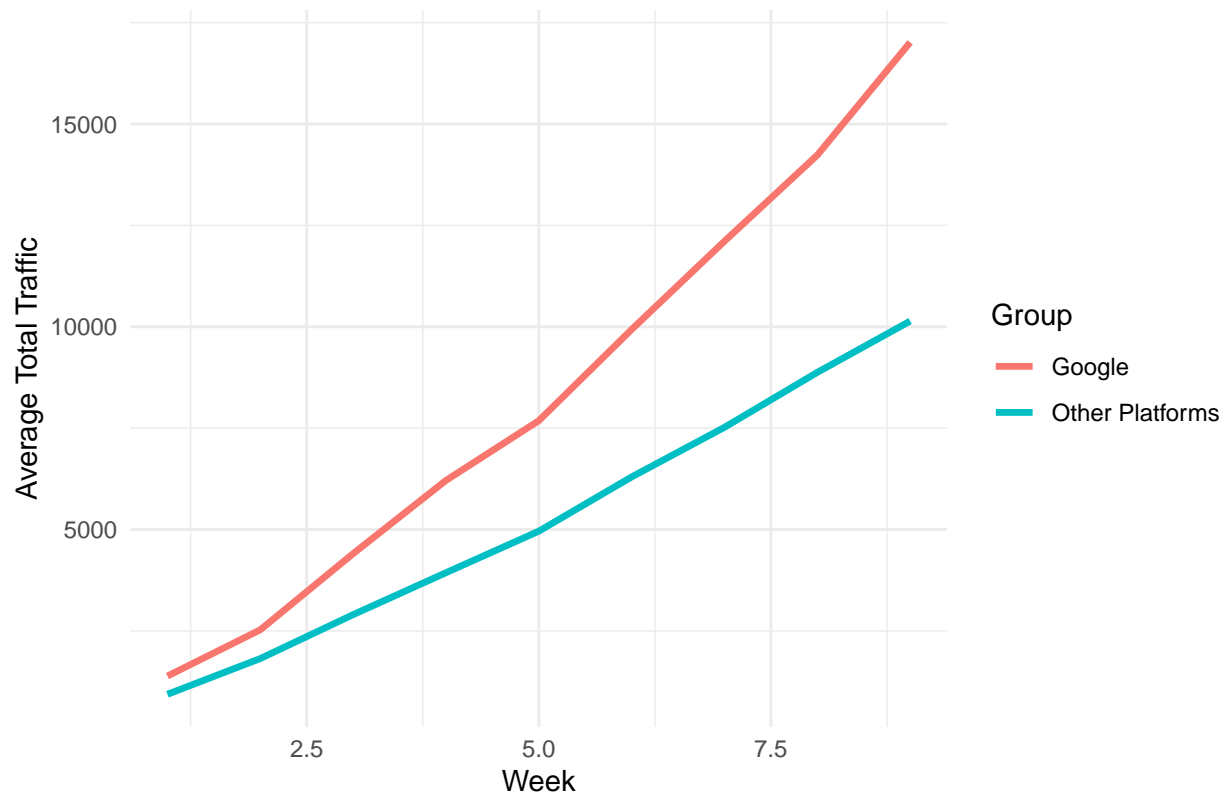
# Create a new group column
pre_data <- pre_data %>%
  mutate(group = ifelse(platform == "goog", "Google", "Other Platforms"))

# Aggregate total traffic by group and week
pre_trend_summary <- pre_data %>%
  group_by(week, group) %>%
  summarise(avg_traffic = mean(total_traffic, na.rm = TRUE), .groups = "drop")

# Plot
ggplot(pre_trend_summary, aes(x = week, y = avg_traffic, color = group)) +
  geom_line(size = 1.2) +
  labs(title = "Pre-Treatment Parallel Trends Check",
       y = "Average Total Traffic", x = "Week",
       color = "Group") +
  theme_minimal()
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

## Pre-Treatment Parallel Trends Check



```
trend_test <- lm(total_traffic ~ week * treated, data = pre_data)
summary(trend_test)
```

```
##
## Call:
## lm(formula = total_traffic ~ week * treated, data = pre_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5174.7  -478.1   390.3  1119.9  3990.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -534.7      907.2  -0.589   0.5598
## week           1159.9      161.2   7.195 3.6e-08 ***
## treated        -806.5     1814.5  -0.444   0.6597
## week:treated    786.3      322.4   2.439   0.0205 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2163 on 32 degrees of freedom
## Multiple R-squared:  0.7815, Adjusted R-squared:  0.761
## F-statistic: 38.14 on 3 and 32 DF, p-value: 1.114e-10
```

## DiD Analysis

To measure the true causal impact of stopping sponsored search advertisements, we employed a Difference-in-Differences (DiD) regression approach using panel data covering four major platforms (Google, Yahoo, Bing, Ask) across a 12-week period.

We constructed the following variables for our analysis:

- `treat = 1` for Google (the treated group that stopped sponsored ads), and 0 for all other platforms (the control platforms).
- `post = 1` for weeks 10 to 12 (the post-treatment period), and 0 for weeks 1 to 9 (the pre-treatment period). `total_traffic` = the sum of `avg_spons` and `avg_org`, representing the combined weekly traffic from both sponsored and organic sources.

To account for skewness in the traffic distribution, we applied a log transformation on the outcome variable. Additionally, we included two-way fixed effects in our regression — one for platform and one for week — to control for:

- Time-invariant platform characteristics (e.g., Google may naturally have higher baseline traffic),
- Time-varying factors affecting all platforms simultaneously (e.g., seasonal trends or external events).

```
# did analysis

did_plm <- plm(log(1+total_traffic) ~ treated * post,
               data = data, index= c("platform","week"),
               model = "within",
               effect = "twoways")
summary(did_plm)

## Twoways effects Within Model
##
## Call:
## plm(formula = log(1 + total_traffic) ~ treated * post, data = data,
##      effect = "twoways", model = "within", index = c("platform",
##      "week"))
##
## Balanced Panel: n = 4, T = 12, N = 48
##
## Residuals:
##      Min.      1st Qu.      Median      3rd Qu.      Max.
## -0.1054856 -0.0273133  0.0054203  0.0230813  0.1153218
##
## Coefficients:
##              Estimate Std. Error t-value Pr(>|t|)
## treated:post -1.11611    0.04465 -24.997 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    2.2098
```

```
## Residual Sum of Squares: 0.10766
## R-Squared:      0.95128
## Adj. R-Squared: 0.92844
## F-statistic: 624.836 on 1 and 32 DF, p-value: < 2.22e-16
```

**Results and Interpretation:** The key coefficient of interest - the interaction term `treat * post` - was estimated at -1.1161 and found to be highly statistically significant ( $p < 2.2e-16$ ). This means that after Google stopped running sponsored ads, its traffic dropped by approximately 67.3% relative to other platforms that continued their advertising as usual.

**Comparison with Pre-Post Estimate:** In previous section, we computed a simple pre-post estimate using only Google's data. However, that method failed to account for:

1. Market-wide trends affecting all platforms,
2. Normal week-to-week growth or variation,
3. The counterfactual scenario (i.e., what would have happened to Google if it hadn't stopped ads).

In contrast, our DiD model compares Google's change to those of similar platforms over the exact same period, making the estimate much more robust and causally valid.

The DiD approach effectively isolates the treatment effect by leveraging the parallel trends assumption, which - as supported by our visual check - appears to hold in the pre-treatment period. This means that in the absence of treatment, traffic trends on Google would have continued similarly to those on the control platforms.

By comparing changes over time between the treated and control groups, DiD removes the influence of confounding variables that are constant within units or common across time.

**Conclusion:** The DiD regression reveals that stopping sponsored search ads led to a significant 67.3% decline in total traffic to Google. This estimate is substantially larger than the pre-post estimate and offers a more accurate representation of the causal impact. It highlights the limitations of relying solely on within group comparisons and reinforces the importance of using control groups and fixed effects to obtain unbiased treatment effect estimates.

## ROI Calculation

Based on our DiD estimate, we find that turning off sponsored search ads led to a 67.3% drop in traffic. Assuming that only this incremental portion of traffic was truly generated by the ads, we recalculate the ROI as follows: Using 1261 sponsored clicks in Week 9 and an estimated 67.3% incremental lift, we get approximately 845 incremental clicks. With \$2.52 revenue per click and \$0.60 cost per click, the adjusted ROI is:

```
treatment_effect <- exp(-1.116) - 1
S <- data$avg_spons[data$platform == "goog" & data$week == 9]
Incremental_Clicks = S * abs(treatment_effect)
Revenue= Incremental_Clicks*2.52 # revenue per sponsored click is $2.52
Cost= S* 0.60 # cost per sponsored ads is $0.60
adj_roi = (Revenue - Cost) / Cost

cat("The correct ROI is: ", adj_roi*100,"%")
```

```
## The correct ROI is: 182.4132 %
```



This is significantly lower than Bob's original estimate of 320%, but it more accurately reflects the true incremental return of sponsored search advertising.

## Concluding Thoughts

Through a Difference in Differences analysis using panel data across four platforms, we find that stopping sponsored search ads on Google led to a statistically significant **67.3%** drop in total traffic, confirming that these ads play a substantial role in driving visits.

While the initial prepost estimate underestimated this effect, our DiD model correctly isolated the causal impact by accounting for platform specific and time specific factors.

Using this treatment effect, we recalculated the return on investment (ROI) and found that, when accounting for only incremental traffic, the corrected ROI is **182.41%** - significantly lower than Bob's overestimated 320%, but still highly profitable.

The adjusted ROI of 182.41% confirms that sponsored ads positively impact traffic, though not as dramatically as initially estimated. This suggests the need for more efficient ad budget allocation, continuing sponsored ads, but with a sharper focus on optimizing spend. To better understand long-term value, Bazaar.com should also explore customer retention and lifetime value, offering a more complete view of the ads' overall impact.