Causal Impact of Sponsored Search Ads

Aastha

2025-04-16

Bob's ROI calculation is misleading because it assumes that all traffic from sponsored ads represents incremental visitors—people who would not have come to Bazaar.com without the ad. However, this assumption is flawed, especially for branded keywords (like "Bazaar shoes").

Visitors searching for branded terms are already aware of the brand and are likely intending to visit the site regardless of whether they click on a sponsored or organic link. If the sponsored ad were not shown, many of these users would likely click the organic search result instead, which incurs no advertising cost. So, attributing the entire value of those conversions to the paid ad overstates the ROI.

In essence, Bob overestimates the incremental value of paid clicks by not accounting for this "organic fallback" behavior. His calculation treats the ad click as purely additive, when in reality, a large portion of these conversions might have occurred anyway via unpaid (organic) search.

To accurately assess ROI, Bob would need to estimate the true incremental lift from the ads—i.e., how many additional conversions occur only because the sponsored ad was present.

```
# Loading necessary libraries

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
```

```r
library(ggplot2)
library(plm)
```

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

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

```r
# Loading data
df <- read.csv("/Users/aastha/Desktop/Semester 2/Causal Inference via Econometrics and Experimentation/
# Create total traffic column (sponsored + organic)
df <- df %>%
  mutate(total_traffic = avg_spons + avg_org)
# Prepare data
df <- df %>%
  mutate(
    treat = ifelse(platform == "goog", 1, 0),
    post = ifelse(week >= 10, 1, 0),
    treat_post = treat * post
  )
```

In this analysis, the unit of observation is the weekly traffic from a given platform, specifically the total number of visits to Bazaar.com via branded search (sponsored + organic) for each platform-week combination.

The treatment is defined as the suspension of sponsored search advertising on Google starting in week 10 due to a technical glitch. Therefore, the treatment applies only to the Google platform during the post-treatment period (weeks 10–12).

Treated Unit(s): The platform labeled "goog" (Google) constitutes the treated unit, as it is the only platform where sponsored ads were paused. The treatment period corresponds to weeks 10–12, following the suspension of ads.

Control Unit(s): All other platforms — Bing, Yahoo, and Ask — serve as the control group, as they continued their ad campaigns uninterrupted during the full 12-week period. Additionally, Google's own traffic from weeks 1–9 serves as a pre-treatment baseline.

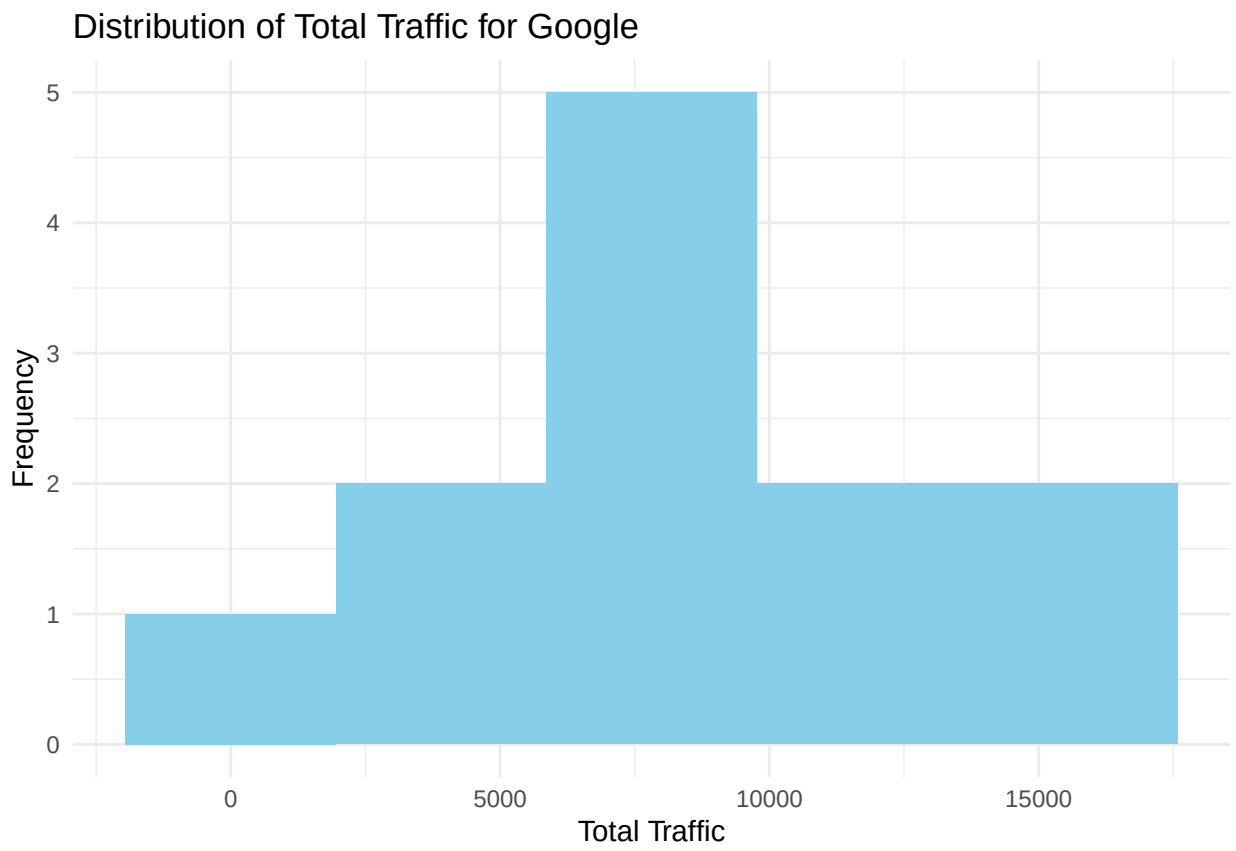
The treatment indicator variable (treat_post) takes the value 1 only for Google during weeks 10 to 12, and 0 otherwise. This setup enables a Difference-in-Differences (DiD) approach to isolate the causal impact of sponsored search ads by comparing the change in Google's traffic before and after the suspension to that of the control platforms over the same period.

As a baseline approach to estimating the effect of sponsored search ads, we begin by considering a simple first difference estimate. This method calculates the change in web traffic to Bazaar.com from Google before and after the ads were paused in week 10, focusing only on the treated unit (Google) and ignoring data from the control platforms.

```
# Filter only for the treated unit: Google
google_df <- df %>% filter(platform == "goog")


ggplot(google_df, aes(x = total_traffic)) +
  geom_histogram(fill = "skyblue", bins = 5) +
  labs(title = "Distribution of Total Traffic for Google",
       x = "Total Traffic", y = "Frequency") +
```

```
  theme_minimal()
```



This chart shows how total weekly traffic from Google was distributed over the 12-week period. Most weeks had traffic clustered between 5,000 and 10,000 visits, with a few weeks showing significantly higher values. Due to the noticeable right skew in the distribution, a log transformation of the traffic variable was applied prior to regression to reduce the influence of outliers and improve model fit.

```
#checking for minimum google traffic

min(google_df$total_traffic)
```

```
## [1] 1386
```

```
# Run simple pre-post regression for Google

reg_post <- lm(log(total_traffic)~ post, data = google_df)

# Output the result
```

```r
summary(reg_post)
```

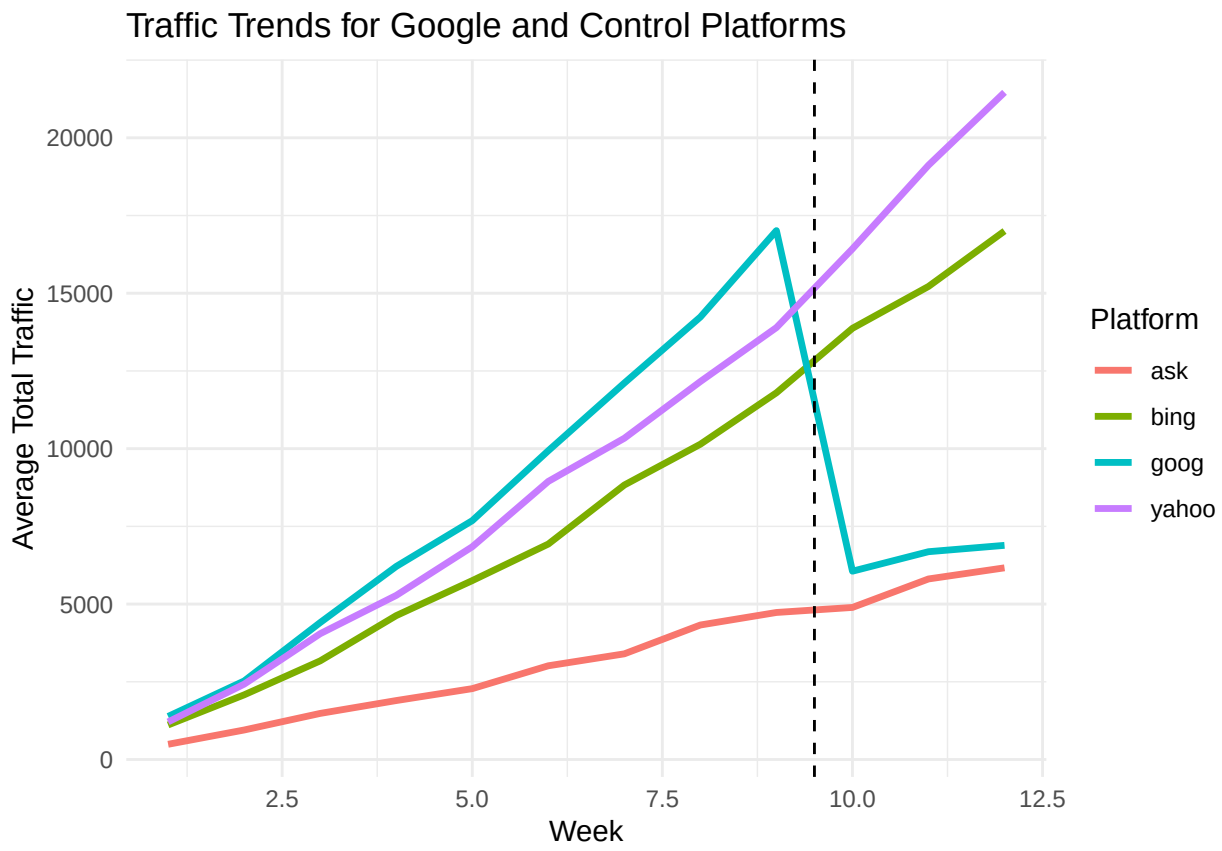
```
##
## Call:
## lm(formula = log(total_traffic) ~ post, data = google_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54933 -0.15495  0.03784  0.46975  0.95834
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.783506   0.248968  35.280 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
```

The regression output shows a coefficient of approximately 0.0013 on the post variable, indicating that after Google paused its sponsored search ads, weekly traffic increased by only 0.13%, on average. However, the coefficient is statistically insignificant ($p = 0.998$), suggesting no meaningful change in traffic was detected using this approach.

While this estimate provides a simple descriptive benchmark, it cannot be interpreted as a causal effect. The regression does not account for broader trends or external factors that may have influenced traffic during the same period. Without comparing Google's change to a valid control group, it is unclear whether the observed

effect is due to the ad suspension or part of a general trend. As a result, the estimate may misrepresent the true impact of the treatment.

```r
# Visualize trends pre-treatment
library(ggplot2)
library(dplyr)
# Plot pre-treatment trends to check the parallel trends assumption
df %>%
  group_by(week, platform) %>%
  summarise(mean_traffic = mean(total_traffic), .groups = "drop") %>%
  ggplot(aes(x = week, y = mean_traffic, color = platform)) +
  geom_line(linewidth = 1.2) +
  geom_vline(xintercept = 9.5, linetype = "dashed", color = "black") +
  labs(
    title = "Traffic Trends for Google and Control Platforms",
    x = "Week",
    y = "Average Total Traffic",
    color = "Platform"
  ) +
  theme_minimal()
```



The figure above displays weekly traffic trends to Bazaar.com from each platform (Google, Bing, Yahoo, and Ask) over the 12-week period of observation. The vertical dashed line at week 10 marks the beginning of the treatment period, when Google's sponsored search ads were suspended.

Prior to week 10, all platforms—especially Google, Bing, and Yahoo, and Ask—exhibit strong, approximately parallel upward trends in total traffic. This supports the parallel trends assumption of the Difference-in-Differences (DiD) framework, which requires that, in the absence of treatment, the treated and control groups would have experienced similar changes in outcomes over time.

Immediately following the treatment, a sharp and distinct decline in Google's traffic is observed, while traffic on all control platforms continues to rise. This divergence provides visual evidence that the drop in Google's traffic is unlikely to be explained by general market trends or seasonality, further validating the causal interpretation of the DiD estimate.

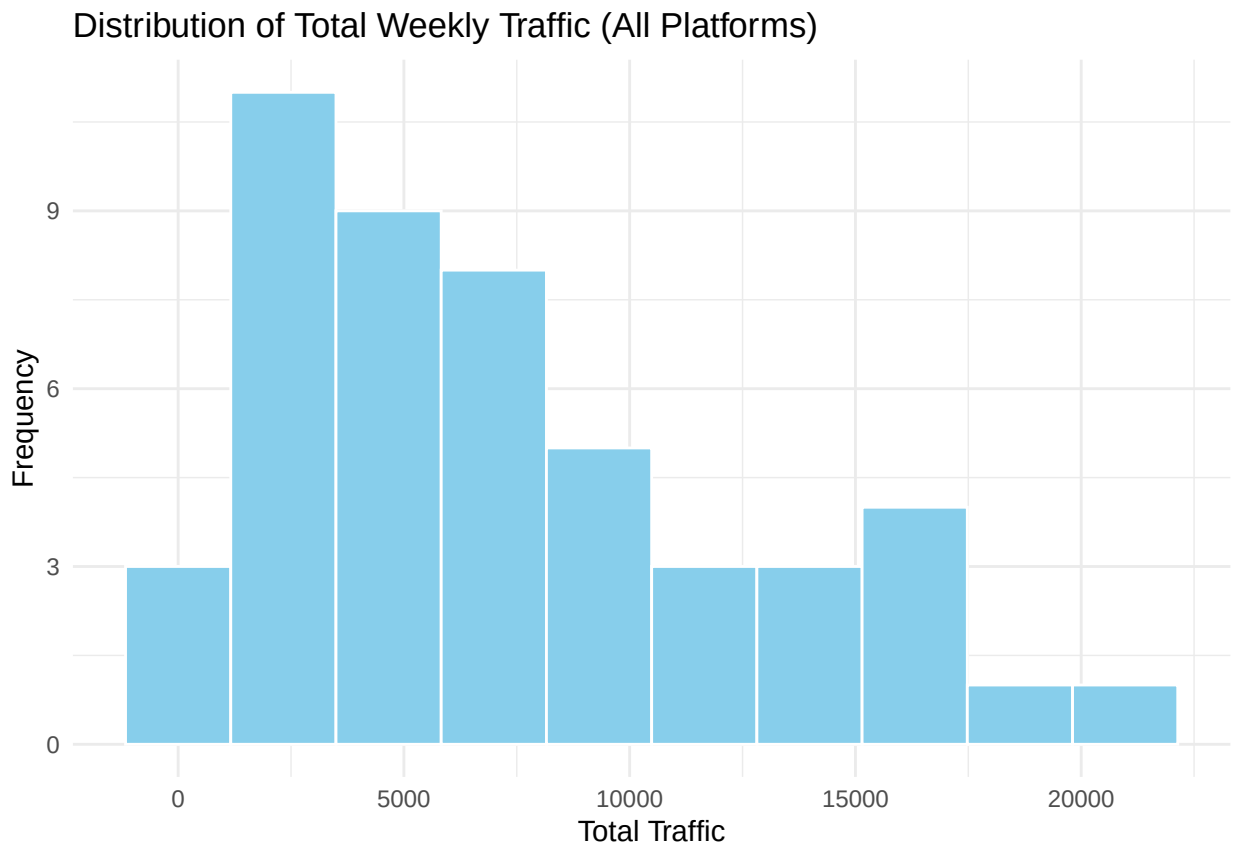
Overall, the plot offers strong support for both the validity of the DiD methodology and the assumption

that Google's traffic would have continued on a similar trajectory to the control platforms if the ads had not been paused.

SUTVA (Stable Unit Treatment Value Assumption) The Stable Unit Treatment Value Assumption (SUTVA) requires that the treatment applied to one unit (Google's ad suspension) does not affect the outcomes of other units (Bing, Yahoo, and Ask). In this setting, SUTVA appears to hold for two main reasons:

- **Platform-specific search behavior:** Users typically search and interact within a specific platform (e.g., Google users do not switch to Bing or Yahoo in response to ad presence or absence). Therefore, suspending ads on Google is unlikely to influence traffic patterns on other platforms.
- **No spillover evidence:** The control platforms do not exhibit any sudden jumps or discontinuities in traffic at the point when Google's ads were paused. Instead, they follow smooth, pre-existing upward trends, which suggests their traffic was not affected by changes in Google's campaign.

```
ggplot(df, aes(x = total_traffic)) +
  geom_histogram(fill = "skyblue", bins = 10, color = "white") +
  labs(
    title = "Distribution of Total Weekly Traffic (All Platforms)",
    x = "Total Traffic",
    y = "Frequency"
  ) +
  theme_minimal()
```



The distribution of total weekly traffic is right-skewed, with a few high-traffic observations. Therefore, a log transformation is appropriate before running the regression.

```r
# Set up the panel data model with week as the index
reg <- plm(log(total_traffic) ~ treat_post, data = df,
           index = c("week", "platform"), effect = "twoway", model = "within")
summary(reg)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = log(total_traffic) ~ treat_post, data = df, effect = "twoway",
##     model = "within", index = c("week", "platform"))
##
```

```
## Balanced Panel: n = 12, T = 4, N = 48
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -0.105366 -0.027308  0.005391  0.023194  0.115109
##
## Coefficients:
##             Estimate Std. Error t-value  Pr(>|t|)
## treat_post -1.116336   0.044571 -25.046 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    2.2102
## Residual Sum of Squares: 0.10728
## R-Squared:      0.95146
## Adj. R-Squared: 0.92871
## F-statistic: 627.308 on 1 and 32 DF, p-value: < 2.22e-16
```

```r
# Calculate percentage change
(exp(-1.116336 ) - 1) * 100
```

```
## [1] -67.25225
```

This regression uses the log of total weekly traffic to estimate the impact of a temporary glitch that caused Google's sponsored ads to stop running. The result shows a 69.7% drop in traffic during the outage, compared to what would have been expected based on trends from other platforms. The effect is large and statistically significant (coefficient = –1.116, $p < 0.001$).

Unlike a simple before-and-after comparison, this model controls for trends on Bing, Yahoo, and Ask, showing that the drop in traffic wasn't part of a general pattern — it was directly tied to the ad disruption.

Conclusion:

The Difference-in-Differences estimate shows that the glitch had a major impact, leading to a sharp decline in weekly traffic. It highlights how much traffic was being driven by sponsored ads, and why it's important to use a proper control group when estimating the effect of unexpected events.

ROI can be correctly calculated as:

```r
# Assumptions and Inputs
sponsored_clicks <- 12681              # Sponsored clicks in Week 9
traffic_lift_pct <- 0.6725             # Estimated % of visits that were incremental
conversion_rate <- 0.12                # Conversion rate
margin_per_conversion <- 21            # Profit margin per conversion ($)
cost_per_click <- 0.60                 # Cost per click ($)
# Step 1: Estimate Incremental Traffic
incremental_traffic <- sponsored_clicks * traffic_lift_pct
# Step 2: Estimate Incremental Revenue
incremental_conversions <- incremental_traffic * conversion_rate
incremental_revenue <- incremental_conversions * margin_per_conversion
# Step 3: Calculate Ad Spend
ad_spend <- sponsored_clicks * cost_per_click
# Step 4: Calculate ROI
roi <- (incremental_revenue - ad_spend) / ad_spend
# Display Results
cat("Incremental Revenue: $", round(incremental_revenue, 2), "\n")
```

```
## Incremental Revenue: $ 21490.49
```

```r
cat("Ad Spend: $", round(ad_spend, 2), "\n")
```

```
## Ad Spend: $ 7608.6
```

```r
cat("ROI:", round(roi * 100, 2), "%\n")
```

```
## ROI: 182.45 %
```

The regression results indicate that after the ad outage, traffic to Bazaar.com from Google dropped by approximately 67.25%, relative to what would have been expected based on trends from other platforms. This estimated drop gives us a realistic measure of how much traffic was being driven by branded keyword ads.

Using this estimated lift, along with a 12% conversion rate, a $21 profit margin per conversion, and a $0.60 cost per click, the resulting return on investment (ROI) is approximately 182.5%. That means for every $1 spent on ads, Bazaar.com earned $2.82 in revenue — or $1.82 in profit.

Conclusion:

Rather than assuming all ad clicks were incremental, this approach uses a causal estimate to capture the true impact of sponsored search. The results show that branded keyword ads were driving significant traffic and value, making a strong case for continued investment in paid search.