Causal_Homework_3

Ujjwal Khanna & Aurosikha Mohanty

2025-04-15

Importing Packages

```
suppressWarnings(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
```

```
suppressWarnings(library(ggplot2))
suppressWarnings(library(plm))
```

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

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

Overview

Bazaar.com is an e-commerce retailer that invests in sponsored search advertising to drive traffic to its website. These paid ads appear alongside organic results when users search for the brand name or relevant terms. Like many companies, Bazaar faces a critical question: Are these ads actually delivering a positive Return on Investment (ROI)?

A recent technical glitch unintentionally paused the company's Google ads for three weeks while ads on Yahoo, Ask and Bing continued running. This created a unique opportunity to study the true, causal effect of sponsored search ads on web traffic by comparing platforms where ads were paused vs. where they remained active.

Problem Statement

The goal of this analysis is to estimate the causal ROI of sponsored search ads using this natural experiment. We aim to determine whether turning off paid ads significantly impacted traffic from Google and how that compares to Yahoo, Ask and Bing, where ads remained live. Using a Difference-in-Differences (DiD) approach, along with panel data models, we seek to isolate the incremental value generated by the ads.

(a) What is Wrong with Bob's ROI Calculation?

Bob estimated ROI from sponsored ads as:

$$\text{ROI} = \frac{(0.12 \times \$21) - \$0.60}{\$0.60} = 320\%$$

While this calculation is mathematically correct, it is conceptually flawed. It assumes that all sponsored ad clicks are **incremental**, meaning the users would not have visited the site otherwise.

However, these ads are triggered by **branded keywords** like "Bazaar shoes," indicating that users are already familiar with the brand and are intentionally searching for the company. In the absence of a sponsored ad, it is highly likely that many of these users would click on Bazaar's **organic search result**, which appears for free.

As a result, sponsored ads may simply **cannibalize organic traffic**, not bring in new visitors. Bob's approach **overstates the ROI** by failing to account for this substitution effect.

To assess the true return on investment, we must measure the **incremental impact** of sponsored ads using a causal method—such as the **Difference-in-Differences** strategy we apply below.

(b) Define the Treatment and Control

This analysis leverages a **natural experiment** that occurred when Bazaar.com's sponsored search ads on Google were accidentally suspended during weeks 10–12. This unplanned outage created a clear separation between **pre-treatment** (weeks 1–9) and **post-treatment** (weeks 10–12) periods.

We define the treatment and control as follows:

– **Unit of Observation:** Platform-week (weekly data for each search engine) – **Treated Unit(s):** The `Google` platform, which experienced the ad suspension in the post period – **Control Unit(s):** The `Yahoo`, `Ask` and `Bing` platforms, which continued normal ad operations throughout

We introduce the following indicator variables to support our analysis:

- `treated = 1` for the Google platform, 0 for Yahoo, Ask and Bing
- `post = 1` for observations in weeks 10 through 12, 0 otherwise
- `did = treated × post`, the interaction term capturing the **causal effect** of suspending sponsored ads

This setup enables us to use a **Difference-in-Differences (DiD)** approach to isolate the effect of the ad suspension on overall traffic to Bazaar.com, controlling for temporal trends common across platforms.

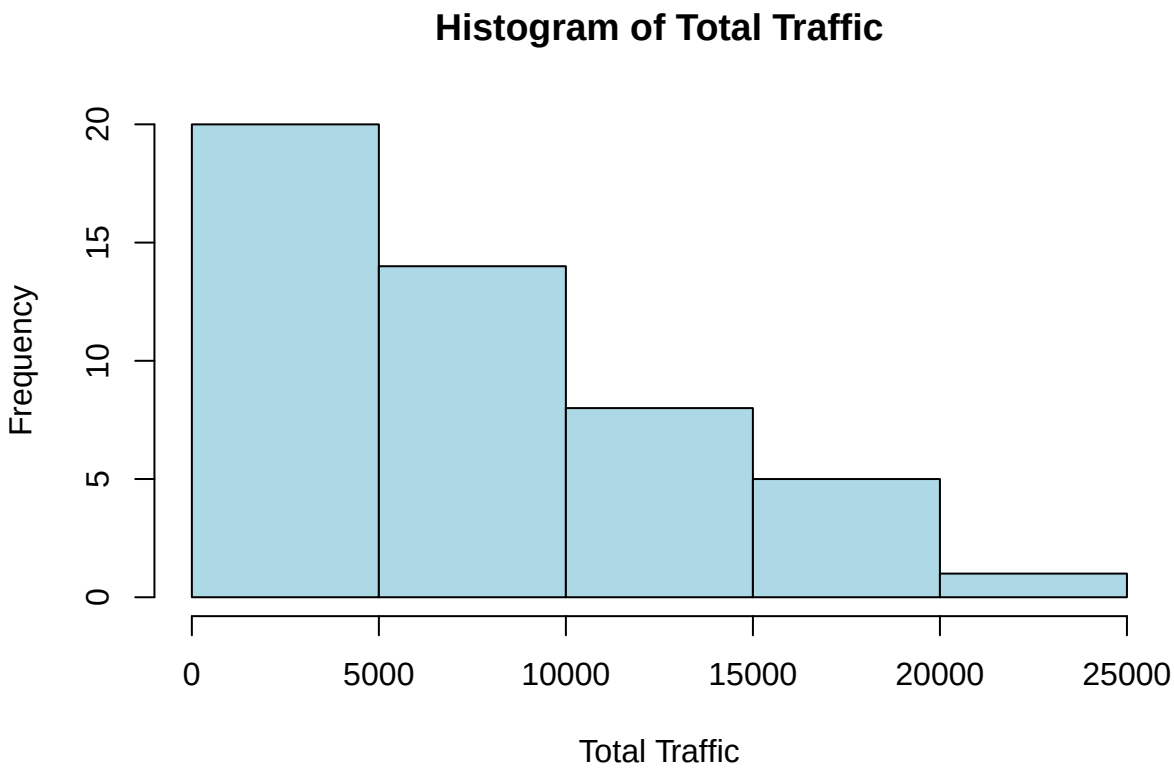
(c) First Difference Estimate (Pre-Post for Google)

We begin by estimating the treatment effect using a **first difference (pre-post)** approach, focusing solely on the treated unit—Google. The model specification is:

```
df <- read.csv("did_sponsored_ads.csv")
df <- df %>%
  mutate(total_traffic = avg_spons + avg_org,
    treated = ifelse(tolower(platform) == "goog", 1, 0),
    post = ifelse(week >= 10, 1, 0))
```

Plot histogram

```
hist(df$total_traffic,
     main = "Histogram of Total Traffic",
     xlab = "Total Traffic",
     col = "lightblue",
     border = "black")
```



- Since the distribution of the total traffic variable is right-skewed. We will need to use log-transform when implementing the pre-post regression for the experiment.

Run pre-post regression

```
## Filter only Google
df_google <- df %>%
  filter(tolower(platform) == "goog")
model_first_diff <- lm(log(total_traffic) ~ post, data = df_google)
summary(model_first_diff)
```

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

```
## lm(formula = log(total_traffic) ~ post, data = df_google)
##
## 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
```

Where: - `Post = 1` for weeks 10–12 (after ad suspension), and 0 otherwise

The estimated treatment effect is:

- **Coefficient on `Post`:** 0.001306, **p-value:** 0.998, **Interpretation** : Since the dependent variable, total traffic, is log transformed. The co-efficient is interpreted in terms of percentage change in total traffic. The coefficient of 0.001306 implies a ~0.13% increase in traffic after the ad suspension. Although this effect is statistically insignificant since the p-value is too large, 0.998.

Not a good idea to determine Causal effect of the treatment

While this estimate is simple to compute, it is not reliable as a causal measure. It assumes that no other external factors are affecting web traffic during this time. Without a control group to account for general trends (e.g., seasonality, marketing cycles), this estimate could be confounded by unrelated changes in user behavior or market conditions.

To isolate the true causal effect of removing sponsored ads, we need to compare Google to platforms that continued running ads, which leads us to the **Difference-in-Differences** approach.

(d) Difference-in-Differences Estimate

To estimate the causal impact of suspending sponsored search ads, we employ a **Difference-in-Differences (DiD)** regression using a two-way fixed effects model. This method controls for: - **Time-invariant platform characteristics** (e.g., some platforms driving more traffic than others), - **Week-specific effects** (e.g., seasonality or external shocks).

We restrict the analysis to **Google** (treated platform) and **Others** (control platforms), using panel data over 12 weeks.

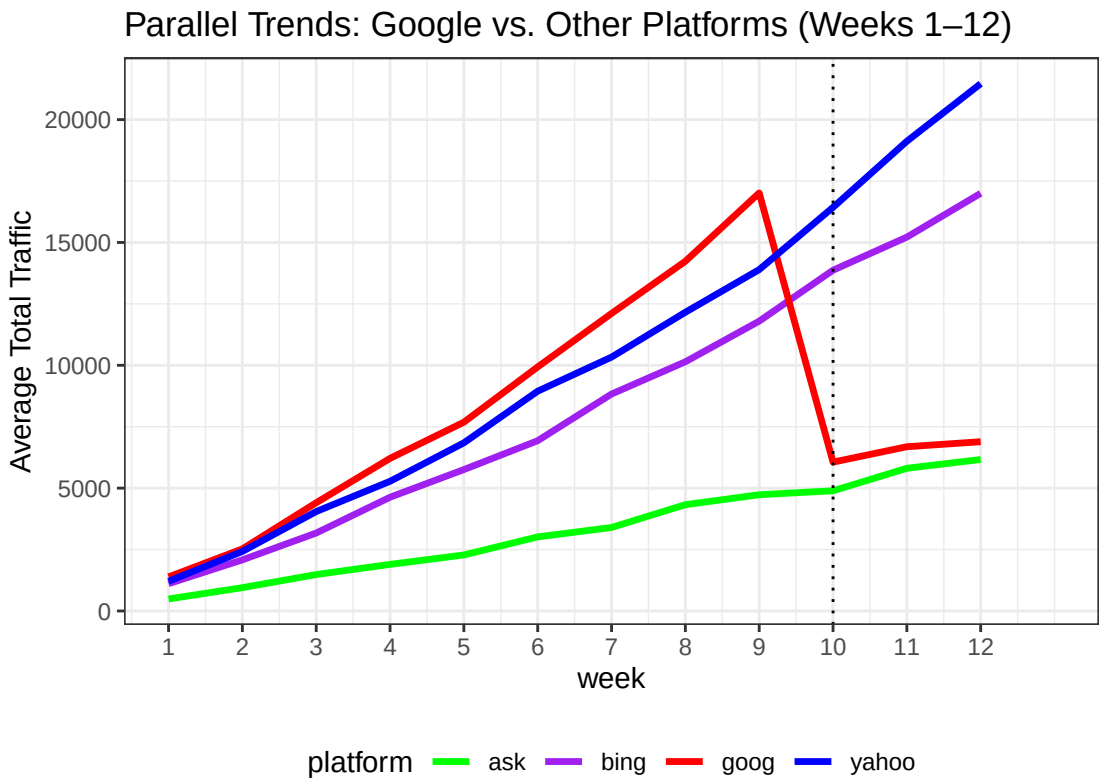
Parallel Trends Check

```
pre_data <- df %>%
  filter(platform %in% c("goog", "bing", "yahoo", "ask"))


week_ave <- pre_data %>%
  group_by(week, platform) %>%
  summarise(avg_traffic = mean(total_traffic), .groups = 'drop')
```

```r
ggplot(week_ave, aes(x = week, y = avg_traffic, color = platform)) +
  coord_cartesian(xlim = c(1, 13)) +
  geom_line(size = 1.2) +
  geom_vline(xintercept = 10, linetype = 'dotted') +
  scale_x_continuous(breaks = 1:12) +
  scale_color_manual(values = c("goog" = "red",
                                "yahoo" = "blue",
                                "ask" = "green",
                                "bing" = "purple")) +
  labs(title = "Parallel Trends: Google vs. Other Platforms (Weeks 1-12)",
       y = "Average Total Traffic") +
  theme_bw() +
  theme(plot.margin = unit(c(0.5, 2, 0.5, 0.1), "cm"),
        legend.position = "bottom",
        axis.text.x = element_text(margin = margin(t=2)))
```

```
## 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.
```



Interpretation

The plot shows weekly average total traffic by platform from weeks 1 to 12. Prior to the ad suspension in week 10 (dotted line), traffic trends across Google, Yahoo, Ask, and Bing appear to follow **parallel patterns**, supporting the key assumption of **parallel trends** in our Difference-in-Differences design.

After week 10, Google's traffic drops sharply while the others remain stable, providing visual evidence of the **causal effect** of suspending sponsored search ads on Google traffic.

Run DiD regression with fixed effects

The model specification is:

```r
model_did <- plm(log(total_traffic) ~ treated * post,
                 data = df,
                 index = c("platform", "week"),
                 effect = "twoways",
                 model = "within")

# Show summary
summary(model_did)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = log(total_traffic) ~ treated * post, data = df,
##     effect = "twoways", model = "within", index = c("platform",
##         "week"))
##
## Balanced Panel: n = 4, T = 12, N = 48
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -0.105366 -0.027308  0.005391  0.023194  0.115109
##
## Coefficients:
##                Estimate Std. Error t-value  Pr(>|t|)
## treated: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
```

Where: - `Post = 1` for weeks 10–12 (after ad suspension), and 0 otherwise - `Treated = 1` for Google, and `0` for Yahoo/Ask/Bing - `Post × Treated` captures the **causal impact** of turning off ads

Key Results

- **Estimated treatment effect (`post:treated`)**: `-1.116`
- **p-value**: `< 0.001`

Interpretation

Since the p-value is quite small, the estimated coefficient of `-1.116` is statistically significant. The interaction term implies a significant drop in log-transformed total traffic due to the suspension of Google's branded ads. Translating this into percentage terms:

```r
# Log coefficient from DiD regression
log_coef <- -1.116
impact_pct <- (exp(log_coef) - 1) * 100
# Print result
cat("Impact (%):", round(impact_pct, 1), "%\n")
```

```
## Impact (%): -67.2 %
```

This means that removing sponsored search ads caused a **statistically significant ~67% decrease** in traffic to Bazaar.com from Google, relative to the control platforms.

Comparison to the Pre-Post Estimate

The earlier **first-difference model (Part c)**, which looked only at pre-post changes for Google, suggested no real impact. That estimate ignored broader temporal trends, which may have masked the effect.

By contrast, the DiD approach provides a **more credible causal estimate** by accounting for shared changes across all platforms and isolating the effect of the ad suspension specifically on the treated group.

Why This Matters

This analysis demonstrates that removing ads caused a **substantial and statistically significant drop in traffic**, underscoring the importance of running controlled comparisons. Had we relied on pre-post differences alone, we would have **underestimated the true business impact** of ad removal.

(e) Fixing Bob's ROI Calculation

Bob originally calculated an ROI of 320% based on the assumption that all sponsored clicks are incremental. However, our Difference-in-Differences analysis found that removing branded ads led to a **66.6% drop in traffic** from Google—indicating true incrementality.

To calculate the revised ROI, we use actual observed data:

- **Week 9 sponsored clicks (before the outage):** 12,681

- **Estimated traffic drop due to ad suspension:** -67.2%
- **Incremental traffic from ads:** -8,451 clicks
- **Conversion rate:** 12%
- **Margin per conversion:** $21
- **Cost per click:** $0.60

Revised ROI Calculation

```
# Assumptions and Inputs
sponsored_clicks <- 12681        # Week 9 sponsored clicks
traffic_lift_pct <- 0.672
conversion_rate <- 0.12
margin_per_conversion <- 21
cost_per_click <- 0.60

# Step 1: Estimate incremental traffic from ads
incremental_traffic <- sponsored_clicks * traffic_lift_pct
incremental_traffic
```

```
## [1] 8521.632
```

```
# 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
```

```r
# Step 4: ROI
roi <- (incremental_revenue - ad_spend) / ad_spend


# Display results
cat("Incremental Revenue: $", round(incremental_revenue, 2), "\n")
```

```
## Incremental Revenue: $ 21474.51
```

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

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

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

```
## ROI: 182.24 %
```

Interpretation

Based on the estimated 67.2% lift in traffic attributed to branded keyword ads, we estimate an ROI of **182.24%**. This means that for every $1 Bazaar.com spent on branded ads, they earned **$2.82 in revenue**, resulting in **$1.82 in profit**.

This suggests that the sponsored search ads were **highly effective** in driving incremental traffic that likely would **not have arrived organically**, validating the value of continuing investment in branded keyword advertising.

Conclusion

Using a causal estimate from a Difference-in-Differences model, we find that Bazaar.com's investment in sponsored ads for branded keywords delivers a **strong positive return**. With an ROI of over **180%**, the ads are contributing significant incremental traffic and revenue.