Assignment4

Amy Chiu (chiu0109), Tiffany Chen (chen8541)

2023-11-11

```
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)
```

```
## Warning: package 'plm' was built under R version 4.3.1

##
## Attaching package: 'plm'

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

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.1
```

1. What is Wrong with Bob's RoI Calculation?

Overestimate of Revenue: The ROI calculation overestimate the sales growth from sponsored search ads. In fact, part of people is not influenced by the sponsored ads. These conversions occur organically and are not influenced by the ad. Instead, they are interested in Bazaar.com and would visit the website through an organic link, irrespective of the existence of sponsored ads. Hence, This porportion of revenue should not be excluded from ROI calculation.

2. Define the Treatment and Control.

Google was assigned as the treatment group, and the stoppage of sponsored search ads on Google occurred during the last three weeks of the experiment, while Bing, Yahoo, and Ask served as the control groups.

3. Consider a First Difference Estimate.

To implement the linear regression model, we introduced a new binary column named 'treatment' to signify whether the data points belonged to the treatment group. Additionally, we created a 'ttl' column, combining the average organic traffic and sponsored ads traffic, and an 'after' column to indicate whether the data point was recorded after the removal of sponsored ads.

The linear regression analysis was conducted exclusively using Google data, with the dependent variable being 'ttl' and the independent variable 'after'. However, the p-value associated with the coefficient of the 'after' variable was notably high. This suggests insufficient confidence in asserting its effectiveness on total traffic. Potential reasons for this could include the oversight of other influential factors on organic traffic, such as overall market fluctuations.

While the obtained results were statistically insignificant, the coefficient of 'after' was -1846. Despite the lack of significance, this coefficient implies that, in the absence of sponsored ads, the average total traffic would decrease by 1846 units. To gain a more comprehensive understanding of the true causal effect of sponsored ads, it is imperative to incorporate additional information that considers other potential influencing factors.

```
df = read.csv('did_sponsored_ads.csv', header = TRUE)
df$ttl = df$avg_spons + df$avg_org
df <- df %>% mutate(treatment = ifelse(platform == 'goog', 1, 0))
df <- df %>% mutate(after = ifelse(week >= 10, 1, 0))

df_goog = df %>% filter(platform == 'goog')
model1 = lm(ttl ~ factor(after), data = df_goog)
summary(model1)
```

```
##
## Call:
## lm(formula = ttl ~ factor(after), data = df_goog)
##
## 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 ***
## factor(after)1    -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
```

4. Calculate the Difference-in-Differences.

Check parallel assumption

The line plot suggested that the treatment group and control group were not parallel. Consequently, we chose not to proceed with dynamic Difference-in-Differences (DiD) analysis to assess pre-treatment parallelism. Despite potential limitations in the reliability of linear regression results, we conducted a DiD regression comparing the treatment group and control group to assess the genuine impact of sponsored ads. The dependent variable was 'ttl', with independent variables including 'treatment' and 'after'.

Interpretation:

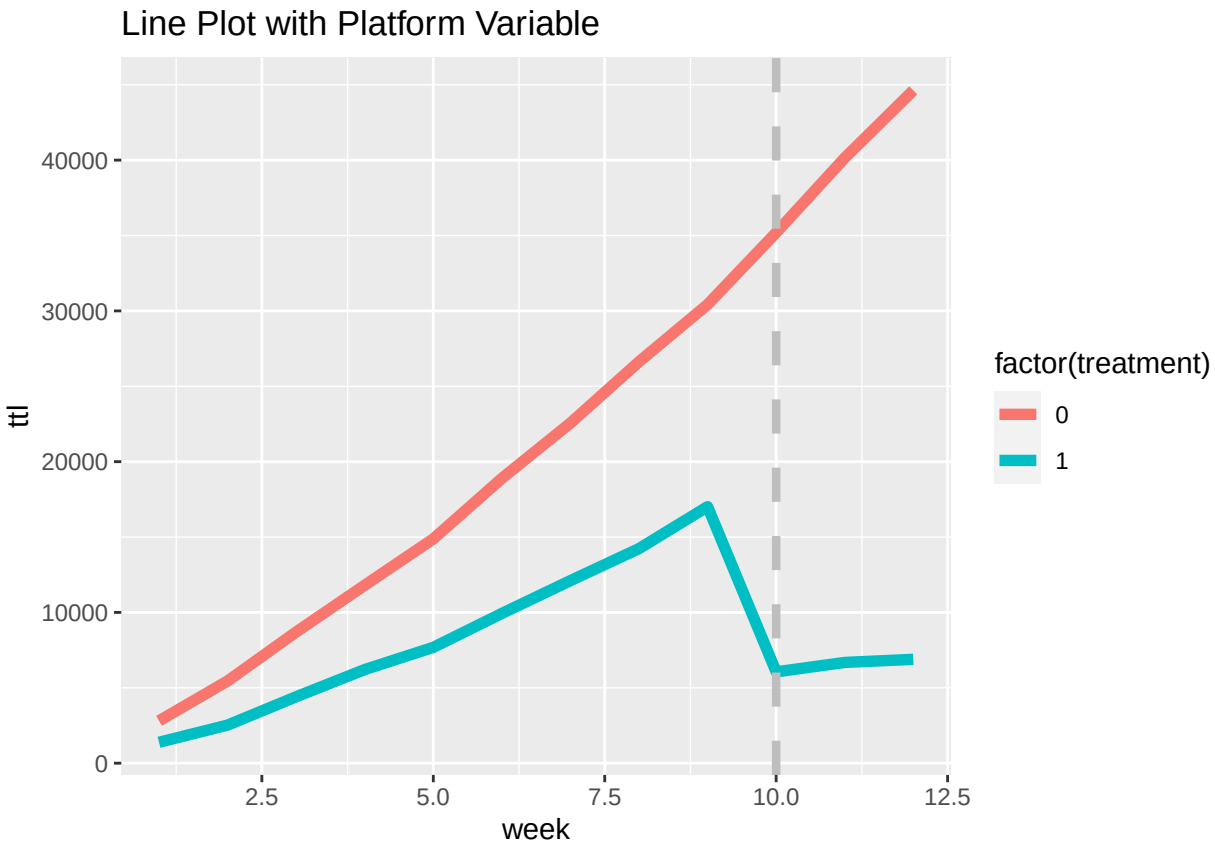
The average weekly traffic decreased by 9910.6 units in the absence of sponsored ads. This reduction in total traffic surpasses the previous estimate of the 'after' coefficient, which was -1846. When compared, the impact on overall traffic was more pronounced when not considering other platform information.

```
## check parallel
df_group <- df %>% group_by(treatment, week) %>% summarise(ttl_sum = sum(ttl))
```

```
## `summarise()` has grouped output by 'treatment'. You can override using the
## `.groups` argument.
```

```
ggplot(df_group, aes(x = week, y = ttl_sum, color = factor(treatment))) +
  geom_line(size = 2) +
  labs(x = "week", y = "ttl", title = "Line Plot with Platform Variable") +
  geom_vline(xintercept = 10, linetype = "dashed", color = "grey", size = 1.5)
```

```
## 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.
```




```
# Run difference in difference regression model
model2 = lm(ttl ~ factor(treatment) * factor(after), data = df)
summary(model2)
```

```
##
## Call:
## lm(formula = ttl ~ factor(treatment) * factor(after), data = df)
##
## 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 ***
## factor(treatment)1                 3124.9     1765.0   1.770  0.08357 .
## factor(after)1                     8064.7     1765.0   4.569 3.94e-05 ***
## factor(treatment)1:factor(after)1 -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
```

5

Bob should exclude clicks from individuals already familiar with the website and actively seeking it. Including them in the calculation might result in an overestimation of the Return on Investment (ROI). So we try to calculate traffic driven by organic search results using difference in difference first, and then use the value to get true traffic driven by sponsored ads.

```
# calculate the traffic driven by organic search results
did_org <- lm(avg_org ~ treatment + after + treatment * after, data=df)
summary(did_org)
```

```
##
## Call:
## lm(formula = avg_org ~ treatment + after + treatment * after,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1928.78  -847.92   -52.67   825.00  2067.33
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1489.7      215.4   6.917 1.51e-08 ***
## treatment          777.0      430.7   1.804   0.0781 .
## after             1984.1      430.7   4.607 3.49e-05 ***
## treatment:after   2293.2      861.4   2.662   0.0108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1119 on 44 degrees of freedom
## Multiple R-squared:  0.6043, Adjusted R-squared:  0.5773
## F-statistic:  22.4 on 3 and 44 DF,  p-value: 5.881e-09
```

A = 9910 (traffic driven by the ads, which we got from (d)) B = 2293 (traffic driven by organic search results)

true traffic by ads = A / (A+B) = 9910 / (9910+2293) = 0.812

```
ROI_new = (21 * 0.12 * 0.812- 0.6)/0.6
ROI_new
```

```
## [1] 2.4104
```