Data Exploration Project

Anh Nguyen

#### **Libraries**

library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.2 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.0  
✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
✔ purrr 1.0.1   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(fixest)  
library(ggplot2)  
library(rio)  
library(lubridate)  
library(Hmisc)

Attaching package: 'Hmisc'  
  
The following objects are masked from 'package:dplyr':  
  
 src, summarize  
  
The following objects are masked from 'package:base':  
  
 format.pval, units

library(dplyr)

#### **Data\_cleaning**

# Reading in the Google Trends data  
google\_trends <- list.files('Lab3\_Rawdata', pattern = 'trends\_up\_to\_', full.names = TRUE)  
  
my\_data <- import\_list(google\_trends, rbind = TRUE, fill = TRUE )

# Aggregating the Google Trends data  
my\_data <- my\_data %>%  
 mutate(week = str\_sub(monthorweek, start = 1, end = 10)) %>%  
 mutate(week = ymd(week)) %>%  
 mutate(month = floor\_date(week, unit = 'month'))

# Aggregating  
my\_data <- my\_data %>%  
 group\_by(schname, keyword) %>%  
 mutate(std\_index = (index - mean(index))/sd(index), na.rm = TRUE)

# Reading in the Scorecard data  
Scorecard\_data <- import('Most+Recent+Cohorts+(Scorecard+Elements).csv')  
id\_name <- import('id\_name\_link.csv')

# Merge in the Scorecard data  
id\_name <- id\_name %>%  
 group\_by(schname) %>%  
 mutate(n = n()) %>%  
 filter(n == 1)  
   
# Change column name in Scorecard\_data from UNITID to unitid  
colnames(Scorecard\_data)[colnames(Scorecard\_data) == 'UNITID'] <- 'unitid'  
  
# Join data together  
sch\_link <- inner\_join(my\_data, id\_name, by = 'schname')  
id\_link <- inner\_join(sch\_link, Scorecard\_data, by = 'unitid')  
  
# filter PREDDEG == 3 (Predominantly bachelor's-degree granting)  
id\_link <- id\_link %>%  
 filter(PREDDEG == 3)  
  
# export my usable data to a new csv files  
rio::export(id\_link, 'usable\_data.csv')

#### **Data Analysis**

# Reading the clean data  
data <- import('usable\_data.csv')

# Change column name  
colnames(data)[colnames(data) == 'md\_earn\_wne\_p10-REPORTED-EARNINGS'] <- 'Report Earnings'  
  
# Converting 'Report Earnings' to numeric  
data$`Report Earnings` <- as.numeric(data$`Report Earnings`)

Warning: NAs introduced by coercion

data <- na.omit(data)

# Calculate median, lower quartile, and upper quartile  
median\_earnings <- median(data$`Report Earnings`, na.rm = TRUE)  
lower\_quartile <- quantile(data$`Report Earnings`, probs = 0.25)  
upper\_quartile <- quantile(data$`Report Earnings`, probs = 0.75)  
  
# Categorize incomes into High and Low incomes  
data$`Income\_Category` <- ifelse(data$`Report Earnings` >= upper\_quartile, "High", "Low")  
  
# Select necessary variables  
data <- data [, c('unitid', 'schname', 'keyword', 'week', 'Report Earnings', 'Income\_Category', 'std\_index')]  
  
# Create 'after' variable based on the release date of the College Scorecard  
data <- data %>%  
 mutate(after = week >= as.Date('2015-09-06'),  
 treated = Income\_Category == 'High')  
  
# Calculate proportion index by grouping data by 'after' and 'treated' variables  
proportion\_data <- data %>%  
 group\_by(after, treated) %>%  
 dplyr::summarize(proportion\_std\_index = mean(std\_index, na.rm = TRUE), .groups = 'drop')  
  
print(proportion\_data)

# A tibble: 4 × 3  
 after treated proportion\_std\_index  
 <lgl> <lgl> <dbl>  
1 FALSE FALSE 0.0240  
2 FALSE TRUE 0.0619  
3 TRUE FALSE -0.103   
4 TRUE TRUE -0.267

# Fit the DID regression model  
model <- feols(std\_index ~ after \* treated, data = data)  
etable(model, digits = 3)

model  
Dependent Var.: std\_index  
   
Constant 0.024\*\*\* (0.001)  
afterTRUE -0.127\*\*\* (0.003)  
treatedTRUE 0.038\*\*\* (0.003)  
afterTRUE x treatedTRUE -0.202\*\*\* (0.007)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID  
Observations 741,598  
R2 0.00602  
Adj. R2 0.00602  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Create a data frame for plotting  
plot\_data <- data.frame(  
 after = c(FALSE, TRUE, FALSE, TRUE),  
 treated = c(FALSE, FALSE, TRUE, TRUE),  
 mean\_std\_index = c(0.024, -0.102, 0.062, -0.267),  
 Treatment = factor(c("Untreated", "Untreated", "Treated", "Treated"), levels = c("Untreated", "Treated"))  
)  
  
# Plot the mean standardized index before and after the College Scorecard release  
ggplot(plot\_data, aes(x = after, y = mean\_std\_index, color = Treatment)) +  
 geom\_point(size = 3) +  
 geom\_line(aes(group = Treatment), linetype = "dashed") +  
 labs(title = "Mean Standardized Index Before and After College Scorecard Release",  
 x = "After College Scorecard Release",  
 y = "Mean Standardized Index",  
 color = "Treatment Group") +  
 scale\_color\_manual(values = c("Treated" = "deeppink", "Untreated" = "deepskyblue")) +  
 theme\_minimal()

A graph with blue and pink dots

Description automatically generated

# create treatment\_group variable  
data$treatment\_group <- ifelse(data$treated == "TRUE", 'Treated', 'Untreated')  
  
# model   
ggplot(data, aes(x = week, y = std\_index, color = treatment\_group)) +  
 stat\_summary(fun.data = "mean\_cl\_normal", geom = "line") +  
 geom\_vline(xintercept = as.numeric(as.Date('2015-09-06')), linetype = "dashed") +  
 annotate("text", x = as.Date('2015-09-06'), y = 2, label = "Release Date", vjust = 3, hjust = 0.5) +  
 labs(title = "Search Activity Before and After Release of College Scorecard",  
 x = "Year",  
 y = "Standardized Google Search Index",  
 color = "Treatment Group") +  
 scale\_color\_manual(values = c("Treated" = "deeppink", "Untreated" = "deepskyblue")) +  
 theme\_minimal()

A graph with blue and pink lines

Description automatically generated

#### **Write Up**

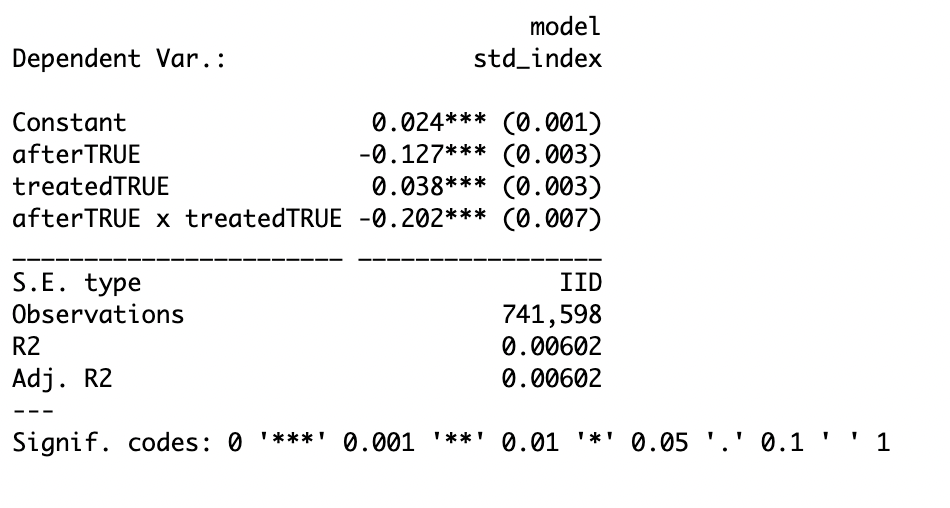
**Introduction:**

The College Scorecard, released in September 2015, provides valuable information about colleges, including graduate earnings. This analysis investigates whether the Scorecard release influenced search activity for colleges with high-earning graduates relative to low-earning ones. Using Google Trends data, I assess changes in search activity before and after the release.

**Methodology:**

To address the research question, I categorized colleges into treated (high-earning) and untreated (low-earning) groups based on their median earnings. I then conducted a difference-in-differences (DID) regression analysis, considering the impact of the College Scorecard release on search activity. Additionally, I visualized the trends in search activity using two graphs.

**Regression Analysis:**



I performed a DID regression with the following model:

*std\_index = β0​ + β1​×after + β2​×treated + β3​×(after×treated) + ϵ*

**Results:**

The regression model provides valuable insights into the impact of the College Scorecard release on search activity, particularly for colleges categorized as high-earning (treated) and low-earning (untreated).

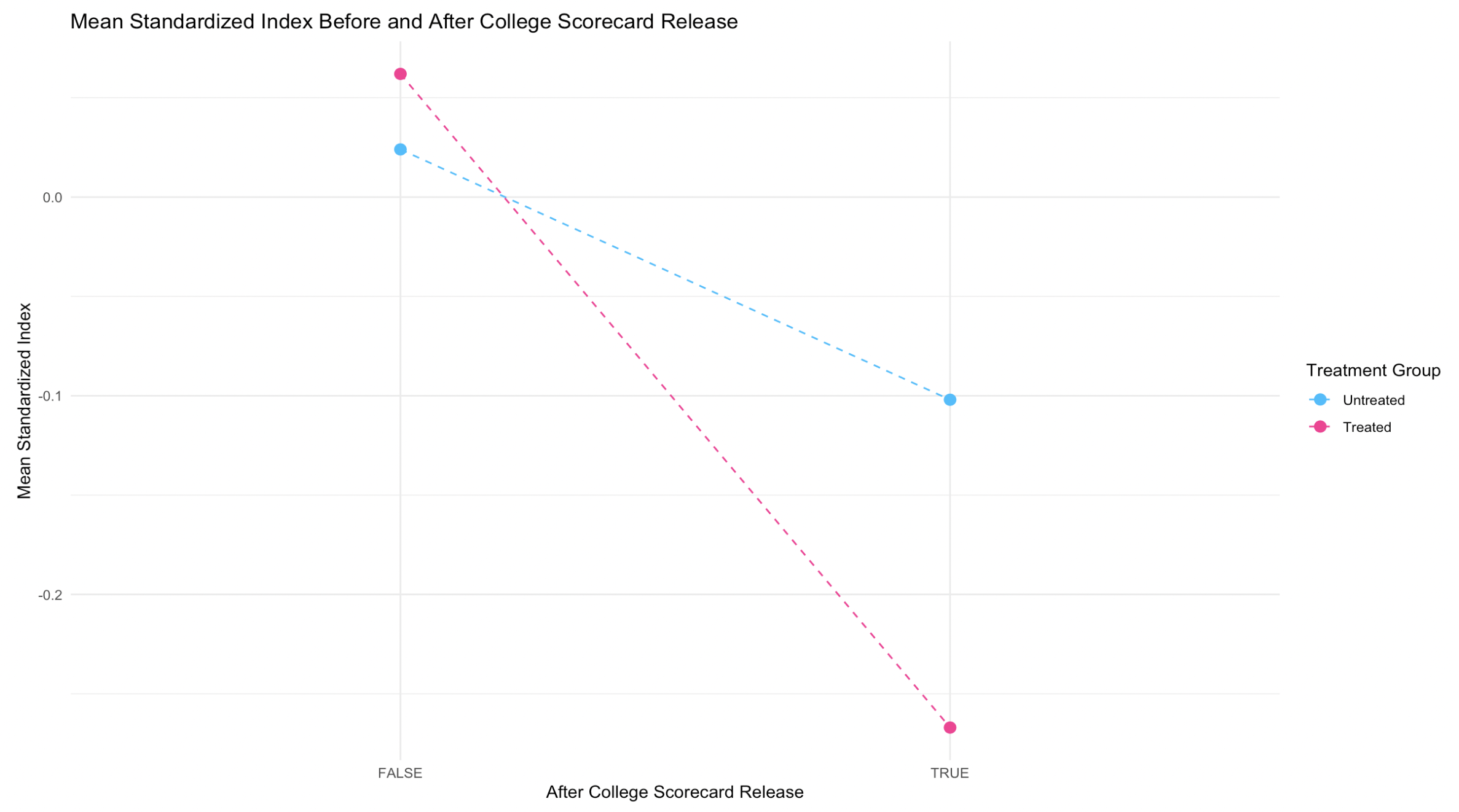
The constant term (*β*0​=0.024, *SE*=0.001) provides a baseline representing the mean level of the standardized search index for untreated colleges before the release. This constant allows us to understand the initial search activity level for low-earning colleges prior to any intervention.

The coefficient for the afterTRUE variable (*β*1​=−0.127, *SE*=0.003) captures the change in search activity for untreated colleges after the Scorecard release. The negative coefficient suggests a significant decrease in search activity following the release, indicating a general decline in interest in college-related searches post-release.

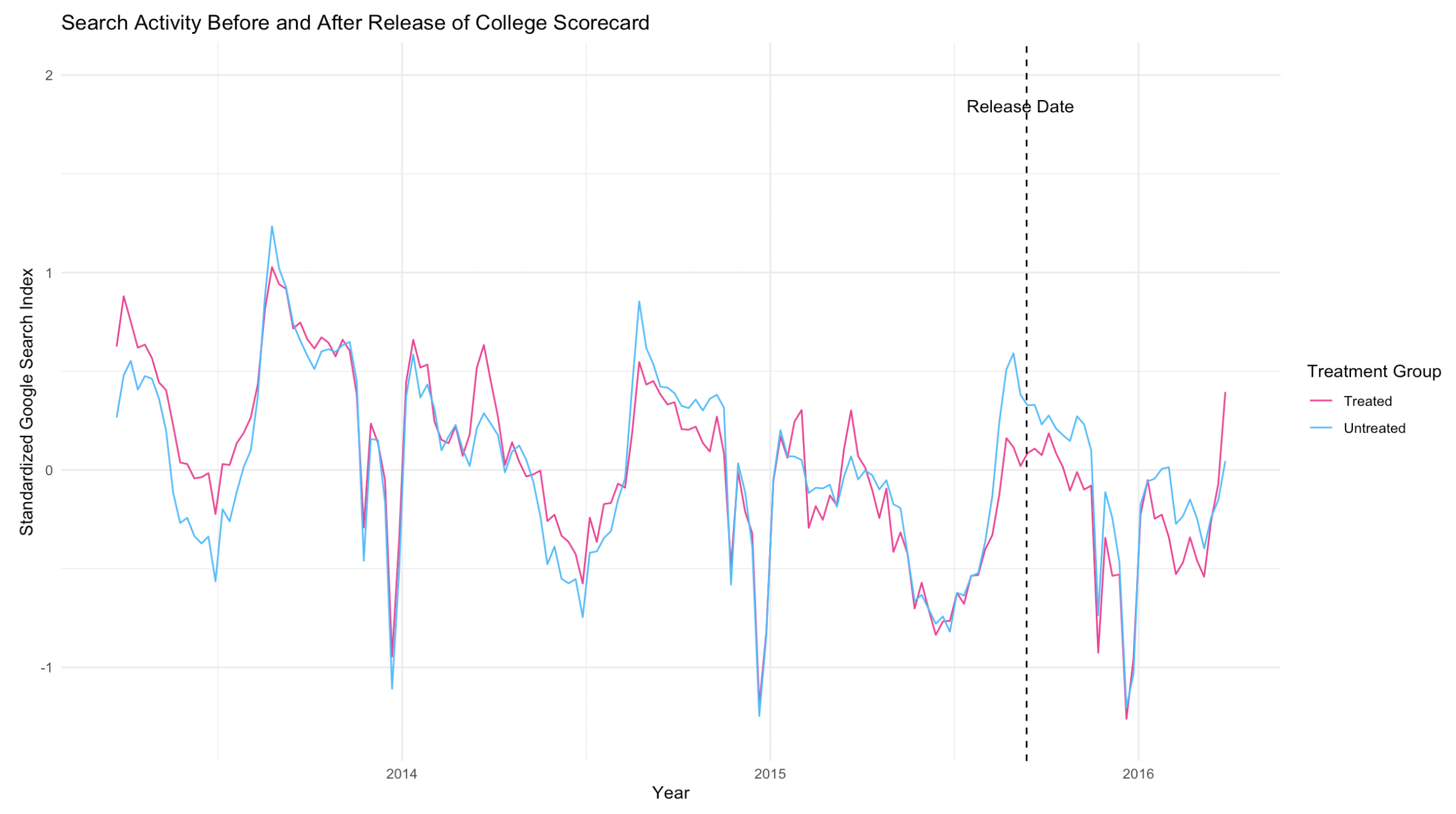
The treatment effect is represented by the coefficient for the treatedTRUE variable (*β*2​=0.038, *SE*=0.003). This coefficient signifies the average difference in search activity between treated (high-earning) and untreated (low-earning) colleges before the Scorecard release. The positive coefficient indicates that, before the release, high-earning colleges had higher search activity compared to low-earning colleges.

The regression results indicate a significant interaction effect between the treatment group and the post-treatment period (*after*×*treated*). The coefficient for this interaction term suggests a decrease in search activity for high-earning colleges relative to low-earning colleges after the College Scorecard release (*β*3​=−0.202, *SE*=0.007). This implies that the release of the College Scorecard had a greater negative impact on search activity for high-earning colleges compared to low-earning ones.

**Graphical Analysis:**



Line Plot of Mean Standardized Index: This graph illustrates the mean standardized search index before and after the College Scorecard release, differentiated by treatment group. The dashed lines represent trends for treated (high-earning) and untreated (low-earning) colleges, with pink indicating treated and blue indicating untreated. I observe a noticeable decline in search activity for treated (high-earning) colleges after the release date, indicating a negative impact of the College Scorecard release on search interest.



Time Series Plot of Standardized Search Index: This graph displays the trends in the standardized search index over time, with a dashed vertical line indicating the release date of the College Scorecard. The lines represent search activity for treated and untreated colleges, with pink indicating treated and blue indicating untreated. The graph shows fluctuations in search activity before and after the release date. The decline in search activity for treated colleges after the release date aligns with the regression results, confirming the negative impact of the Scorecard release on search interest for high-earning colleges.

**Conclusion:**

The comprehensive analysis of the impact of the College Scorecard release on search activity for high-earning colleges yields several key insights. Firstly, the regression results reveal a significant decrease in search activity for high-earning colleges following the release, indicating a negative response to the availability of earnings information provided by the Scorecard. This finding suggests that students may have shifted their attention away from high-earning colleges towards other options, possibly due to heightened scrutiny or reevaluation of their college choices in light of the disclosed earnings data.

Moreover, the interaction effect between treatment status (high-earning vs. low-earning) and the post-release period underscores the nuanced nature of the intervention’s impact. Specifically, treated (high-earning) colleges experienced a more pronounced decline in search activity compared to untreated (low-earning) colleges, indicating a disproportionate negative effect on colleges with higher median earnings. This differential response highlights the importance of considering contextual factors, such as institutional characteristics and student preferences, when evaluating the effectiveness of policy interventions in the higher education landscape.

In a broader context, these finding underscores the importance of transparent information in shaping student preferences and decision-making regarding college choices. Policymakers and educators should consider the implications of such initiatives on student behavior and college selection processes.