Homework 4

Avnee Satija - 5992424 and Harshal Sable - 5949697

2025-05-09

Executive Summary

This study investigates the causal effects of search behavior, policy intervention, and ad ranking using three econometric methods: propensity score matching, synthetic control, and regression discontinuity.

In Question 1, we examine whether directed search behavior increases sales. After matching users on income, education, and behavioral history, we find that directed search sessionss have an incremental increase on overall sales and promoted product sales. Interestingly, non-promoted sales fall indicating that directed users are more focused and less likely to browse broadly. Matching improves comparability between groups, strengthening the causal interpretation.

In Question 2, we evaluate the impact of California's 1989 cigarette tax using synthetic control. By comparing actual cigarette sales in California to a synthetic version constructed from control states, we observe a clear post-treatment divergence. While trends were nearly identical pre-1989, California's actual sales dropped significantly afterward. This provides compelling evidence that the tax effectively reduced cigarette consumption.

In Question 3, we assess whether being ranked first in ad placement increases click-through rate (CTR) using regression discontinuity. By constructing a forcing variable based on bid differences between the top two ads, we isolate the causal impact of ranking. We find that being ranked #1 increases CTR by 15.8 percentage points (p = 0.002).

Together, these results demonstrate the power of causal inference techniques in uncovering real behavioral and policy-driven effects in digital and offline contexts.

QUESTION 1

Data Exploration

```r
library(readr)
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
```

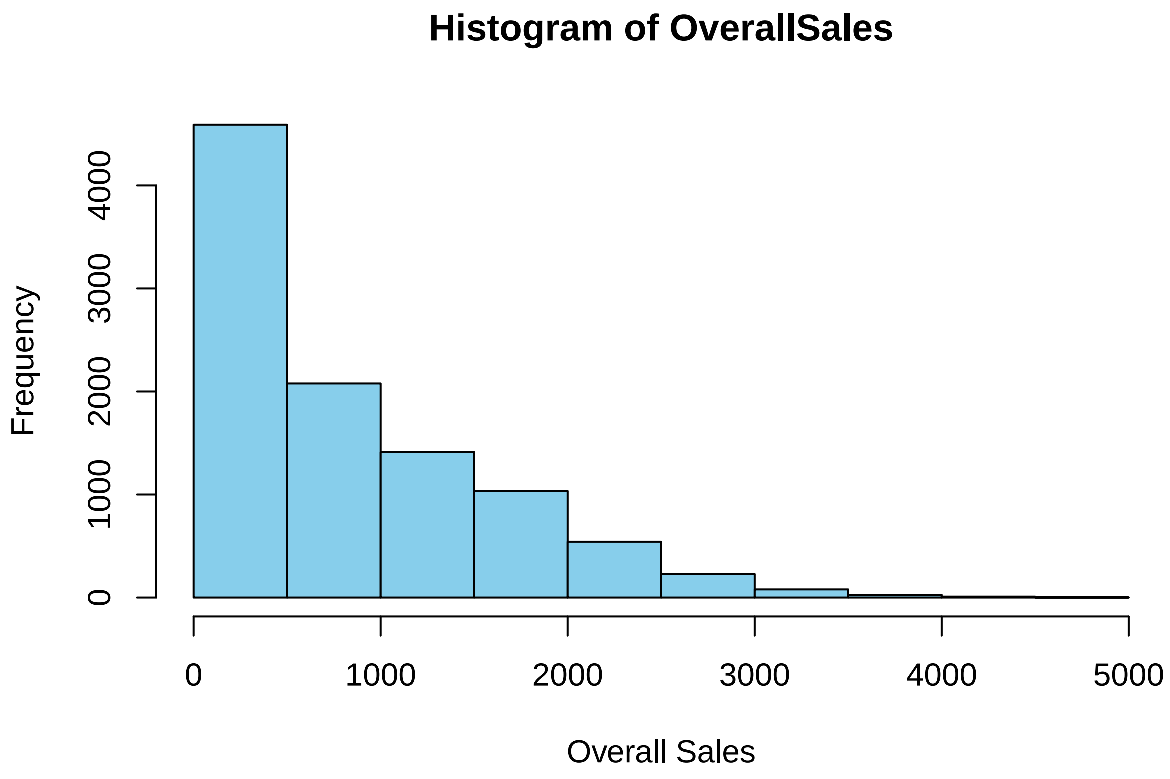
```r
matching <- read_csv("~/Desktop/Semester 2/Casual Inference/Homework/hw4/matching.csv")
```

```
## Rows: 10000 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (11): ConsumerID, Income, Education, NumSessions, DaysSinceLastPurchase,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Overall Sales Dustribution

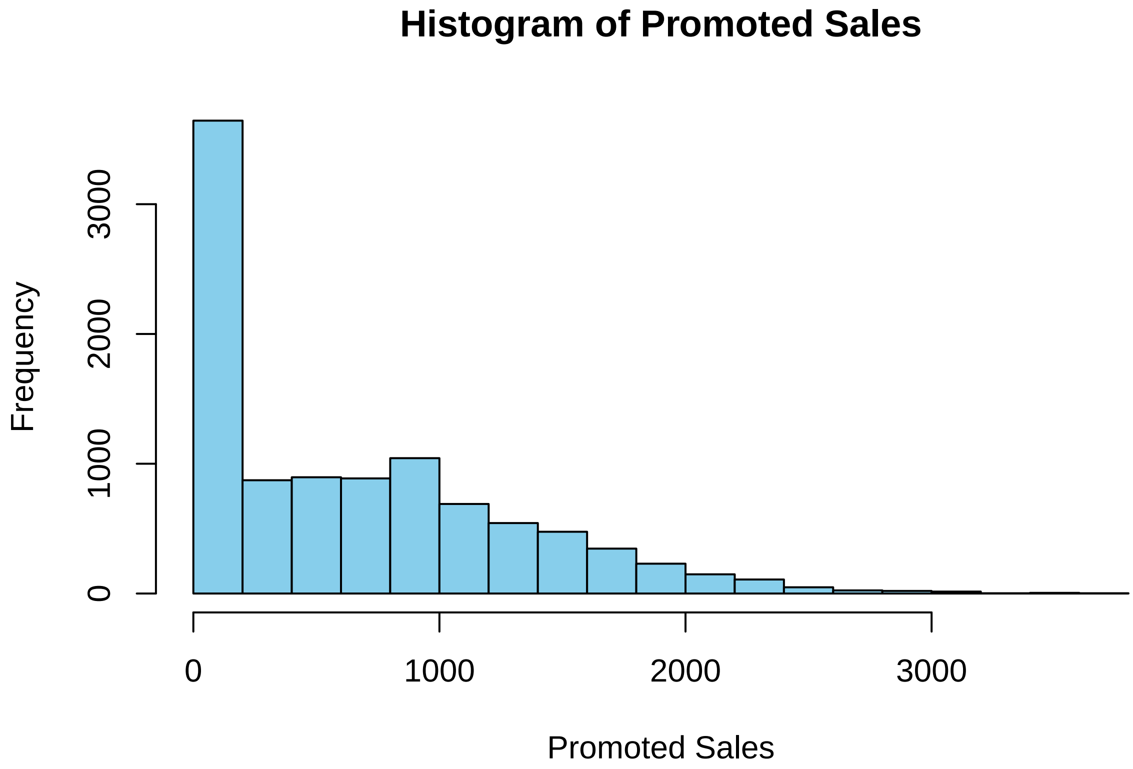
The histogram of Overall Sales reveals a highly right-skewed distribution, with the majority of observations concentrated in the lower range. A small number of sessions recorded very high sales, resulting in a long tail to the right. This skewness suggests that log-transformation may be appropriate before regression modeling to stabilize variance and reduce the influence of extreme values.

```r
hist(matching$OverallSales, main = "Histogram of OverallSales", xlab = "Overall Sales", col = "skyblue"
```



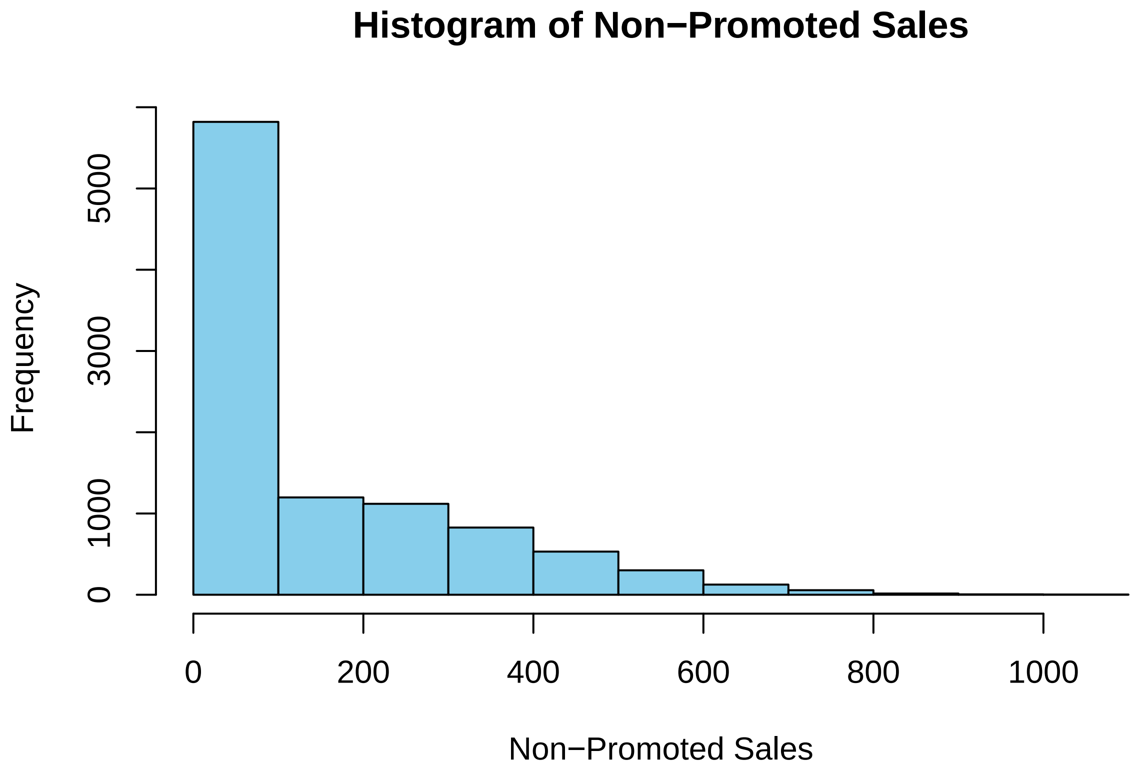
### Promoted Sales Dustribution The distribution of Promoted Sales is strongly right-skewed, with a large concentration of observations at low sales values. The frequency sharply drops as promoted sales increase, indicating the presence of a long right tail. This suggests that most users generate minimal promoted sales, while a small number drive significantly higher sales. Due to this skewness, a log-transformation may again be helpful in modeling to reduce heteroskedasticity and make regression estimates more reliable.

```
hist(matching$PromotedSales, main = "Histogram of Promoted Sales", xlab = "Promoted Sales", col = "skybl
```



### Non-promoted Sales Dustribution The Non-Promoted Sales distribution is heavily right-skewed, with most values clustered at the low end of the sales spectrum (e.g., under 100 units). A steep drop-off occurs as sales increase, and only a small number of sessions show high non-promoted sales. This skewed pattern suggests the presence of extreme values and again supports the use of log transformation in downstream regression analysis to stabilize variance and reduce the influence of outliers.

```
hist(matching$NonpromotedSales, main = "Histogram of Non-Promoted Sales", xlab = "Non-Promoted Sales", c
```



### T-Test Analysis: Directed vs. Undirected Search Sessions We conducted independent two-sample t-tests to compare sales outcomes between sessions classified as Directed Search (group 1) and Undirected Search (group 0) across three dimensions:

```
t.test( OverallSales ~ DirectedSearchUsage, data = matching)
```

```
##
##  Welch Two Sample t-test
##
## data:  OverallSales by DirectedSearchUsage
## t = -32.01, df = 3530.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -619.7359 -548.1990
## sample estimates:
## mean in group 0 mean in group 1
##        662.3696       1246.3370
```

Directed search sessions are associated with a significantly higher average in overall sales. The extremely small p-value (below 0.05) and tight confidence interval indicate that this difference is not due to random variation.

```
t.test( PromotedSales ~ DirectedSearchUsage, data = matching)
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  PromotedSales by DirectedSearchUsage
## t = -43.483, df = 3283.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -679.5183 -620.8815
## sample estimates:
## mean in group 0 mean in group 1
##        510.1724       1160.3723
```

Directed sessions lead to markedly higher promoted sales. This again reflects a strong treatment effect, supported by robust statistical significance (a p-value < 0.05)

```
t.test( NonpromotedSales ~ DirectedSearchUsage, data = matching)
```

```
##
##  Welch Two Sample t-test
##
## data:  NonpromotedSales by DirectedSearchUsage
## t = 18.394, df = 4678.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  59.17314 73.29179
## sample estimates:
## mean in group 0 mean in group 1
##       152.19720        85.96473
```

In contrast, undirected search sessions tend to yield more non-promoted sales, suggesting they may be driven more by discovery behavior rather than targeted purchase intent. P-value here is significant as well as it is below 0.05.

While the differences in outcomes are statistically significant, these t-tests do not control for confounding variables. Users who engage in directed searches may differ systematically from undirected users in terms of income, experience, or purchasing behavior. As such, the treatment (directed search) and control (undirected search) groups are not comparable at baseline.Therefore, to estimate the causal effect of directed search more accurately, we proceed with propensity score matching to balance the two groups on observed covariates.

Part a) We estimate a log-linear model to assess the impact of directed search behavior on all sales variables. The variable log is used to handle skewness and zero values in the original distribution.

```
model1 <- lm(log(OverallSales + 1) ~ DirectedSearchUsage, data = matching)
summary(model1)
```

```
##
## Call:
## lm(formula = log(OverallSales + 1) ~ DirectedSearchUsage, data = matching)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7724 -0.7368  0.3776  1.6080  3.2296
##
```

```
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.17956    0.02463  210.32   <2e-16 ***
## DirectedSearchUsage  1.59287    0.05202   30.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.169 on 9998 degrees of freedom
## Multiple R-squared:  0.08573,    Adjusted R-squared:  0.08564
## F-statistic: 937.5 on 1 and 9998 DF,  p-value: < 2.2e-16
```

Holding other factors constant, users in directed search sessions are associated with a log increase of 1.593 in OverallSales + 1. Back-transforming this effect:

```
100 * (exp(1.59287) - 1)
```

```
## [1] 391.7843
```

So, on average, directed search increases overall sales by approximately 392%, a statistically significant and substantial effect ($p < 0.05$).

```
model2 <- lm(log(PromotedSales + 1) ~ DirectedSearchUsage, data = matching)
summary(model2)
```

```
##
## Call:
## lm(formula = log(PromotedSales + 1) ~ DirectedSearchUsage, data = matching)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7309 -0.6833  0.4071  1.5436  3.1575
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.97592    0.02392  207.99   <2e-16 ***
## DirectedSearchUsage  1.75498    0.05054   34.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.107 on 9998 degrees of freedom
## Multiple R-squared:  0.1076, Adjusted R-squared:  0.1075
## F-statistic:  1206 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```
100 * (exp(1.75498) - 1)
```

```
## [1] 478.3332
```

Switching from undirected to directed search increases the log of promoted sales (+1) by 1.755, a highly significant effect ($p < 0.05$).

Exponentiating this, directed search leads to an estimated 478% increase in promoted sales, on average, relative to undirected sessions.

```r
model3 <- lm(log(NonpromotedSales + 1) ~ DirectedSearchUsage, data = matching)
summary(model3)
```

```
##
## Call:
## lm(formula = log(NonpromotedSales + 1) ~ DirectedSearchUsage,
##     data = matching)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5938 -2.1150  0.1486  2.0775  4.5842
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.59385    0.02601  138.17   <2e-16 ***
## DirectedSearchUsage -1.47886    0.05494  -26.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.291 on 9998 degrees of freedom
## Multiple R-squared:  0.06757,    Adjusted R-squared:  0.06747
## F-statistic: 724.5 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```r
100 * (exp(-1.47886) - 1)
```

```
## [1] -77.21027
```

Directed search is associated with a decrease of 1.479 units in the log of non-promoted sales (+1), and the result is highly statistically significant ($p < 0.05$).

Part b)

We applied nearest neighbor matching (1:1 without replacement) using a logistic model to estimate the probability of receiving Directed Search treatment. Matching was based on log-transformed income, education, prior session count, days since last purchase, and historical total purchases. A tight caliper of 0.001 was used to ensure close matches.

```r
library(MatchIt)
match_output = matchit(DirectedSearchUsage ~ log(Income) + Education + NumSessions +
                       DaysSinceLastPurchase + HistoricalTotalPurchases,
                       data = matching, method = "nearest",
                       distance = "glm",
                       link = "logit",
                       caliper =  0.001,
                       ratio = 1,
                       replace = FALSE)
summary(match_output)
```

```
##
## Call:
## matchit(formula = DirectedSearchUsage ~ log(Income) + Education +
##     NumSessions + DaysSinceLastPurchase + HistoricalTotalPurchases,
##     data = matching, method = "nearest", distance = "glm", link = "logit",
##     replace = FALSE, caliper = 0.001, ratio = 1)
##
## Summary of Balance for All Data:
##                         Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance                       0.5906        0.1182          1.7730     2.3001
## log(Income)                    3.8034        1.3472          2.0759     0.5355
## Education                     15.4641       14.8536          0.3022     1.0197
## NumSessions                    5.0156        5.0447         -0.0132     0.9675
## DaysSinceLastPurchase          4.5695        4.4481          0.0206     0.9895
## HistoricalTotalPurchases       4.4289        4.4666         -0.0065     1.0011
##                         eCDF Mean eCDF Max
## distance                   0.4126   0.6764
## log(Income)                0.4229   0.6622
## Education                  0.0360   0.1273
## NumSessions                0.0038   0.0110
## DaysSinceLastPurchase      0.0095   0.0239
## HistoricalTotalPurchases   0.0058   0.0183
##
## Summary of Balance for Matched Data:
##                         Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance                       0.3590        0.3590          0.0001     1.0002
## log(Income)                    2.9481        2.9581         -0.0085     1.4760
## Education                     15.3363       15.2817          0.0270     1.1645
## NumSessions                    5.0724        4.9811          0.0414     1.1004
## DaysSinceLastPurchase          4.6886        4.3108          0.0640     1.1082
## HistoricalTotalPurchases       4.5828        4.4735          0.0189     1.5203
##                         eCDF Mean eCDF Max Std. Pair Dist.
## distance                   0.0001   0.0033          0.0003
## log(Income)                0.0350   0.1091          0.4209
## Education                  0.0120   0.0457          1.1713
## NumSessions                0.0109   0.0412          1.1170
## DaysSinceLastPurchase      0.0183   0.0401          0.8151
## HistoricalTotalPurchases   0.0132   0.0356          0.8443
##
## Sample Sizes:
##           Control Treated
## All          7759    2241
## Matched       898     898
## Unmatched    6861    1343
## Discarded       0       0
```

```r
matched_data = match.data(match_output)
```

Only a subset of observations met the strict caliper criteria for matching. This enhances covariate balance but reduces the sample size, which is an acceptable trade-off for more reliable causal inference. The resulting matched sample (n = 1,796) is now suitable for outcome comparison via regression.

```r
matched_model1 <- lm(log(OverallSales + 1) ~ DirectedSearchUsage, data = matched_data)
summary(matched_model1)
```

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

```
## lm(formula = log(OverallSales + 1) ~ DirectedSearchUsage, data = matched_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5073 -0.2663  0.3748  1.1493  2.8568
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.45845    0.06664   81.91   <2e-16 ***
## DirectedSearchUsage  1.04883    0.09424   11.13   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.997 on 1794 degrees of freedom
## Multiple R-squared:  0.06458,    Adjusted R-squared:  0.06406
## F-statistic: 123.9 on 1 and 1794 DF,  p-value: < 2.2e-16
```

After matching on user characteristics, directed search is associated with a 185% higher overall sales, and the effect is statistically significant ($p < 0.05$). This estimate reflects a more credible causal effect, free from confounding bias in observable covariates.

```r
100 * (exp(1.04883) - 1)
```

```
## [1] 185.431
```

After propensity score matching, sessions classified as Directed Search are associated with an estimated 185.4% higher overall sales on average compared to Undirected Search sessions, holding other covariates balanced.

```r
matched_model2 <- lm(log(PromotedSales + 1) ~ DirectedSearchUsage, data = matched_data)
summary(matched_model2)
```

```
##
## Call:
## lm(formula = log(PromotedSales + 1) ~ DirectedSearchUsage, data = matched_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4624 -0.2117  0.4251  1.0703  2.7322
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.30571    0.06481   81.86   <2e-16 ***
## DirectedSearchUsage  1.15673    0.09166   12.62   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.942 on 1794 degrees of freedom
## Multiple R-squared:  0.08154,    Adjusted R-squared:  0.08103
## F-statistic: 159.3 on 1 and 1794 DF,  p-value: < 2.2e-16
```

```
100 * (exp(1.15673) - 1)
```

```
## [1] 217.9519
```

After propensity score matching, directed search sessions are associated with an average 218% increase in promoted sales, relative to undirected sessions. This effect is statistically significant ($p < 0.05$) and accounts for observable user-level confounders.

```
matched_model3 <- lm(log(NonpromotedSales + 1) ~ DirectedSearchUsage, data = matched_data)
summary(matched_model3)
```

```
##
## Call:
## lm(formula = log(NonpromotedSales + 1) ~ DirectedSearchUsage,
##     data = matched_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5090 -2.1333 -0.3656  2.2709  4.4941
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.50898    0.08023   43.74   <2e-16 ***
## DirectedSearchUsage -1.37563    0.11346  -12.12   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.404 on 1794 degrees of freedom
## Multiple R-squared:  0.07574,    Adjusted R-squared:  0.07522
## F-statistic:   147 on 1 and 1794 DF,  p-value: < 2.2e-16
```

```
100 * (exp(-1.37563) - 1)
```

```
## [1] -74.73196
```

After matching, directed search sessions are associated with a 74% lower non-promoted sales compared to undirected sessions. This effect is statistically significant ($p < 0.05$), indicating that directed users are much less likely to purchase unpromoted items, even after controlling for observed covariates.

Q2 (Synthetic Control):

```
library(tidyr)
library(dplyr)
library(ggplot2)
library(glmnet)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-8
```

```r
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```r
library(Synth)
```

```
## ##
## ## Synth Package: Implements Synthetic Control Methods.

## ## See https://web.stanford.edu/~jhain/synthpage.html for additional information.
```

```r
library(ggthemes)
library(patchwork)
```

```r
library(readr)
smoking <- read_csv("~/Desktop/Semester 2/Casual Inference/Homework/hw4/smoking.csv")
```

```
## New names:
## Rows: 1209 Columns: 8
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): state dbl (7): ...1, year, cigsale, lnincome, beer, age15to24, retprice
## i Use ‘spec()‘ to retrieve the full column specification for this data. i
## Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
## * ‘‘ -> ‘...1‘
```

To prepare the dataset for synthetic control analysis, we first created lagged outcome variables by reshaping the cigarette sales data from long to wide format. Specifically, we extracted sales (cigsale) for the years 1975, 1980, and 1988, and pivoted them into separate columns (cigsale_1975, cigsale_1980, cigsale_1988) for each state. These lagged values serve as key predictors of cigarette consumption trends leading up to the treatment period.

Next, we ensured that the state_id variable used to uniquely identify each state was stored as a numeric type. This step avoids errors during the synthetic control setup and guarantees proper indexing during matching and model estimation.

```r
smoking$state_id <- as.numeric(factor(smoking$state))
## long to wide dataset
lag_data <- smoking %>%
  filter(year %in% c(1975, 1980, 1988)) %>%
  select(state, year, cigsale) %>%
  pivot_wider(names_from = year, values_from = cigsale, names_prefix = "cigsale_")

smoking <- left_join(smoking, lag_data, by = "state")
```

```r
# Coerce again and confirm
smoking$state_id <- as.numeric(smoking$state_id)
str(smoking$state_id)  # Should return num
```

```
##  num [1:1209] 29 32 10 21 14 27 22 25 2 36 ...
```

Synthetic Control Summary We used the Synth package to estimate the effect of California's cigarette tax by comparing its sales to a weighted combination of control states.

- Predictors: lnincome, retprice, age15to24, beer, and lagged sales (1975, 1980, 1988)
- Treatment unit: California
- Pre-treatment period: 1970–1988
- Plot window: 1970–2000
- The plot compares actual vs. synthetic cigarette sales, with a red line at 1989 marking the policy change.
- A post-1989 gap indicates the causal effect of the tax.

```r
smoking <- as.data.frame(smoking)
treatment_state <- "California"
treatment_id <- unique(smoking$state_id[smoking$state == treatment_state])

dataprep_out <- dataprep(
  foo = smoking,
  predictors = c("lnincome", "retprice", "age15to24", "beer"),
  predictors.op = "mean",
  special.predictors = list(
    list("cigsale", 1975, "mean"),
    list("cigsale", 1980, "mean"),
    list("cigsale", 1988, "mean")
  ),
  dependent = "cigsale",
  unit.variable = "state_id",
  time.variable = "year",
  treatment.identifier = treatment_id,
  controls.identifier = setdiff(unique(smoking$state_id), treatment_id),
  time.predictors.prior = 1970:1988,
  time.optimize.ssr = 1970:1988,
  unit.names.variable = "state",
  time.plot = 1970:2000
)
```

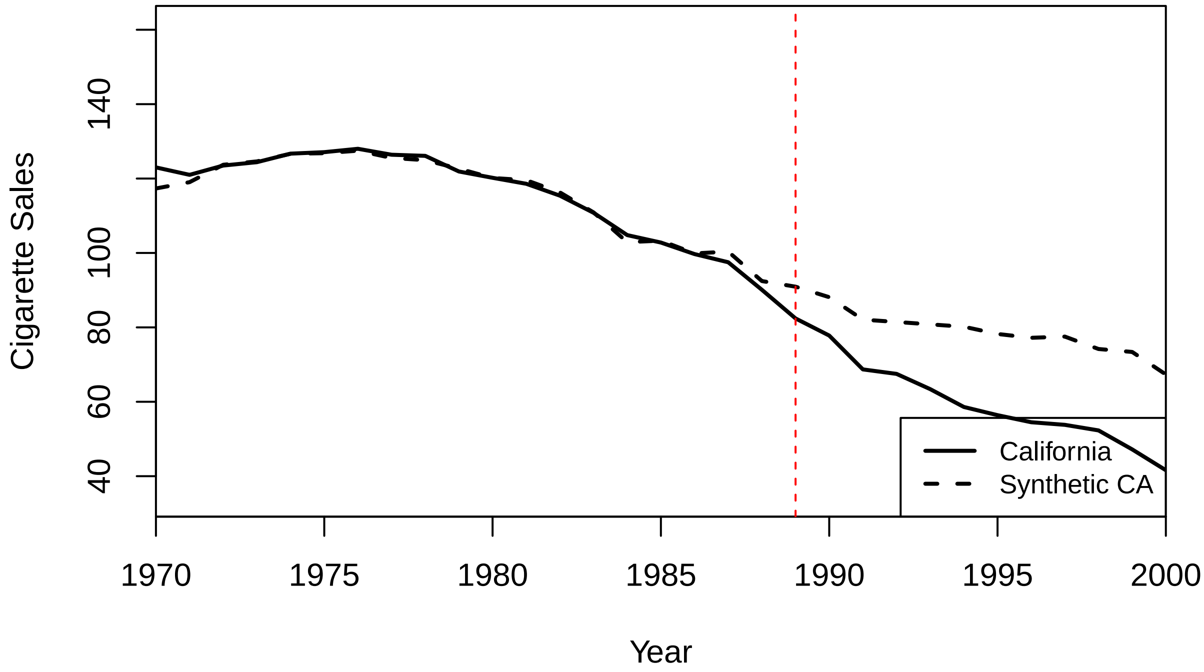
```
##
##  Missing data- treated unit; predictor: lnincome ; for period: 1970
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data- treated unit; predictor: lnincome ; for period: 1971
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data- treated unit; predictor: beer ; for period: 1970
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data- treated unit; predictor: beer ; for period: 1971
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data- treated unit; predictor: beer ; for period: 1972
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data- treated unit; predictor: beer ; for period: 1973
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data- treated unit; predictor: beer ; for period: 1974
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data- treated unit; predictor: beer ; for period: 1975
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data- treated unit; predictor: beer ; for period: 1976
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data- treated unit; predictor: beer ; for period: 1977
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data- treated unit; predictor: beer ; for period: 1978
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data- treated unit; predictor: beer ; for period: 1979
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data- treated unit; predictor: beer ; for period: 1980
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data- treated unit; predictor: beer ; for period: 1981
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data- treated unit; predictor: beer ; for period: 1982
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data- treated unit; predictor: beer ; for period: 1983
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data - control unit: 1 ; predictor: lnincome ; for period: 1970
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data - control unit: 1 ; predictor: lnincome ; for period: 1971
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data - control unit: 1 ; predictor: beer ; for period: 1970
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data - control unit: 1 ; predictor: beer ; for period: 1971
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data - control unit: 1 ; predictor: beer ; for period: 1972
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
##
##  Missing data - control unit: 1 ; predictor: beer ; for period: 1973
##  We ignore (na.rm = TRUE) all missing values for predictors.op.
```

```r
synth_out <- synth(dataprep_out)
```

```
##
## X1, X0, Z1, Z0 all come directly from dataprep object.
##
##
## ****************
##  searching for synthetic control unit
##
##
## ****************
## ****************
## ****************
##
## MSPE (LOSS V): 3.069261
##
## solution.v:
##  0.0010901 0.009765174 0.001041528 0.01119902 0.5811205 0.3235615 0.07222212
##
## solution.w:
##  3.7609e-06 3.0555e-06 0.09383158 0.1106215 4.3302e-06 3.8177e-06 1.3774e-06 1.53224e-05 3.0171e-06 8
```

```r
path.plot(synth.res = synth_out, dataprep.res = dataprep_out,
          Ylab = "Cigarette Sales", Xlab = "Year",
          Legend = c("California", "Synthetic CA"), Legend.position = "bottomright")

abline(v=1989,lty=2,col="red")
```



Actual vs. Synthetic Sales Visualization

We plot California's actual cigarette sales against its synthetic control from 1970 to 2000. The vertical dashed line at 1989 marks the policy intervention (tax increase). A persistent divergence between the two lines after 1989 visually reflects the estimated causal impact of the tax on reducing cigarette consumption in California.

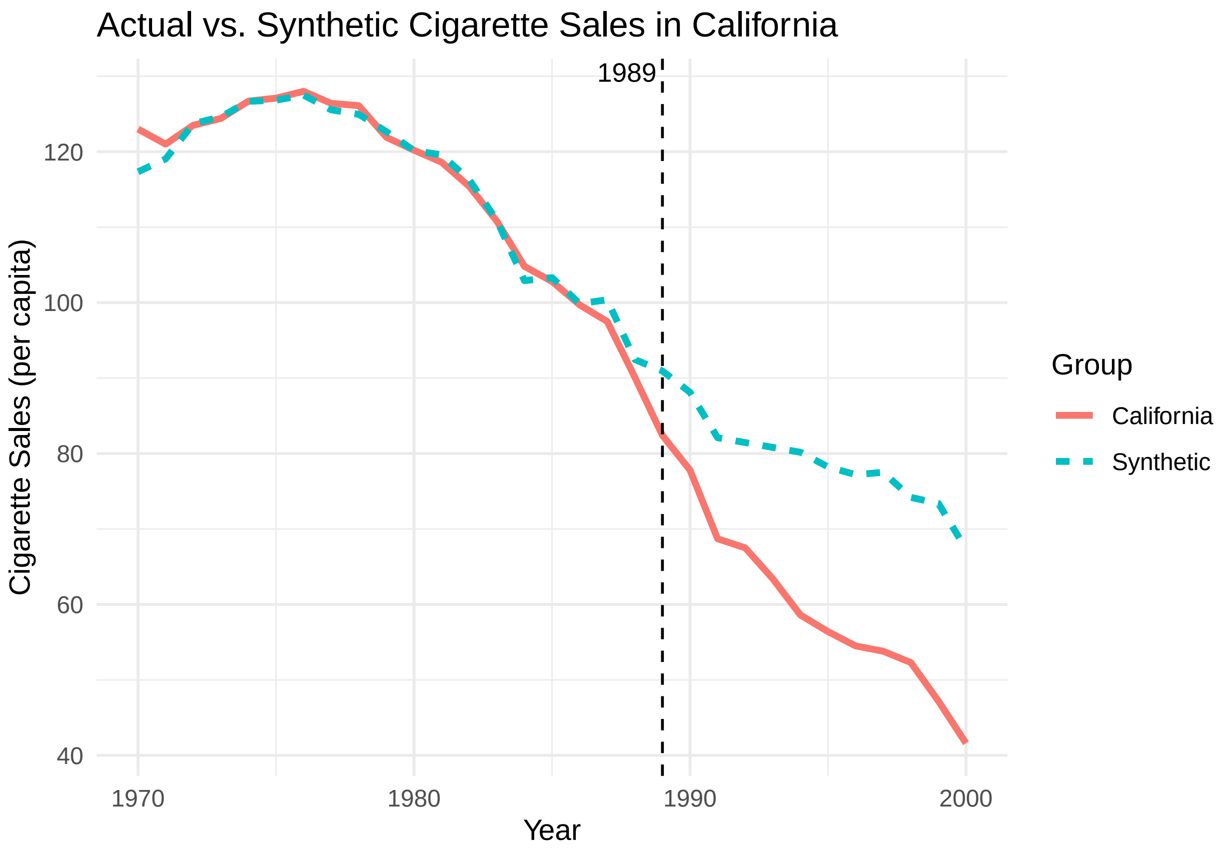
```r
synth_sales <- dataprep_out$Y0plot %*% synth_out$solution.w
actual_sales <- dataprep_out$Y1plot

years <- dataprep_out$tag$time.plot
plot_data <- data.frame(
  Year = years,
  California = as.numeric(actual_sales),
  Synthetic = as.numeric(synth_sales)
)

plot_data_long <- pivot_longer(plot_data, cols = c("California", "Synthetic"),
                                names_to = "Group", values_to = "CigaretteSales")


ggplot(plot_data_long, aes(x = Year, y = CigaretteSales, color = Group, linetype = Group)) +
  geom_line(size = 1.2) +
  geom_vline(xintercept = 1989, linetype = "dashed", color = "black") +
  annotate("text", x = 1989, y = max(plot_data_long$CigaretteSales, na.rm = TRUE),
           label = "1989", vjust = -0.5, hjust = 1.1, size = 3.5) +
  labs(title = "Actual vs. Synthetic Cigarette Sales in California",
       x = "Year", y = "Cigarette Sales (per capita)",
       color = "Group", linetype = "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.
```



QUESTION 3

```
rd <- read_csv("~/Desktop/Semester 2/Casual Inference/Homework/hw4/rd.csv")
```

```
## Rows: 50000 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (3): auction_id, bid, ctr
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
library(rddtools)
```

```
## Loading required package: AER

## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival

## Loading required package: np

## Nonparametric Kernel Methods for Mixed Datatypes (version 0.60-18)
## [vignette("np_faq",package="np") provides answers to frequently asked questions]
## [vignette("np",package="np") an overview]
## [vignette("entropy_np",package="np") an overview of entropy-based methods]

##
## Please consider citing R and rddtools,
## citation()
## citation("rddtools")
```

```r
library(rdrobust)
library(rdd)
```

```
## Loading required package: Formula
```

```r
library(dplyr)
library(ggplot2)
```

RDD Forcing Variable Construction

- Ranking: Within each auction_id, bids are ranked from highest to lowest.
- Filtering: Only ads ranked 1 or 2 are retained.
- Forcing Variable (z):
- Calculated as the difference between the top two bids.
- Assigned positive for rank 1 and negative for rank 2.
- This symmetric forcing variable z is centered at 0, enabling causal estimation of the effect of being ranked 1 (vs. 2) on click-through rates using RDD.

```r
rd <- rd %>%
  group_by(auction_id) %>%
  mutate(rank = rank(-bid, ties.method = "first")) %>%
  ungroup()

rd_filtered <- rd %>%
  filter(rank %in% c(1, 2))

rd_filtered <- rd_filtered %>%
  group_by(auction_id) %>%
  mutate(
    bid_diff = max(bid) - min(bid),
    z = ifelse(rank == 1, bid_diff, -bid_diff)
  ) %>%
  ungroup()
```

```r
robust_rdd = rdrobust(rd_filtered$ctr,rd_filtered$z, c=0)
```

```
## Warning in rdrobust(rd_filtered$ctr, rd_filtered$z, c = 0): Mass points
## detected in the running variable.
```

```r
summary(robust_rdd)
```

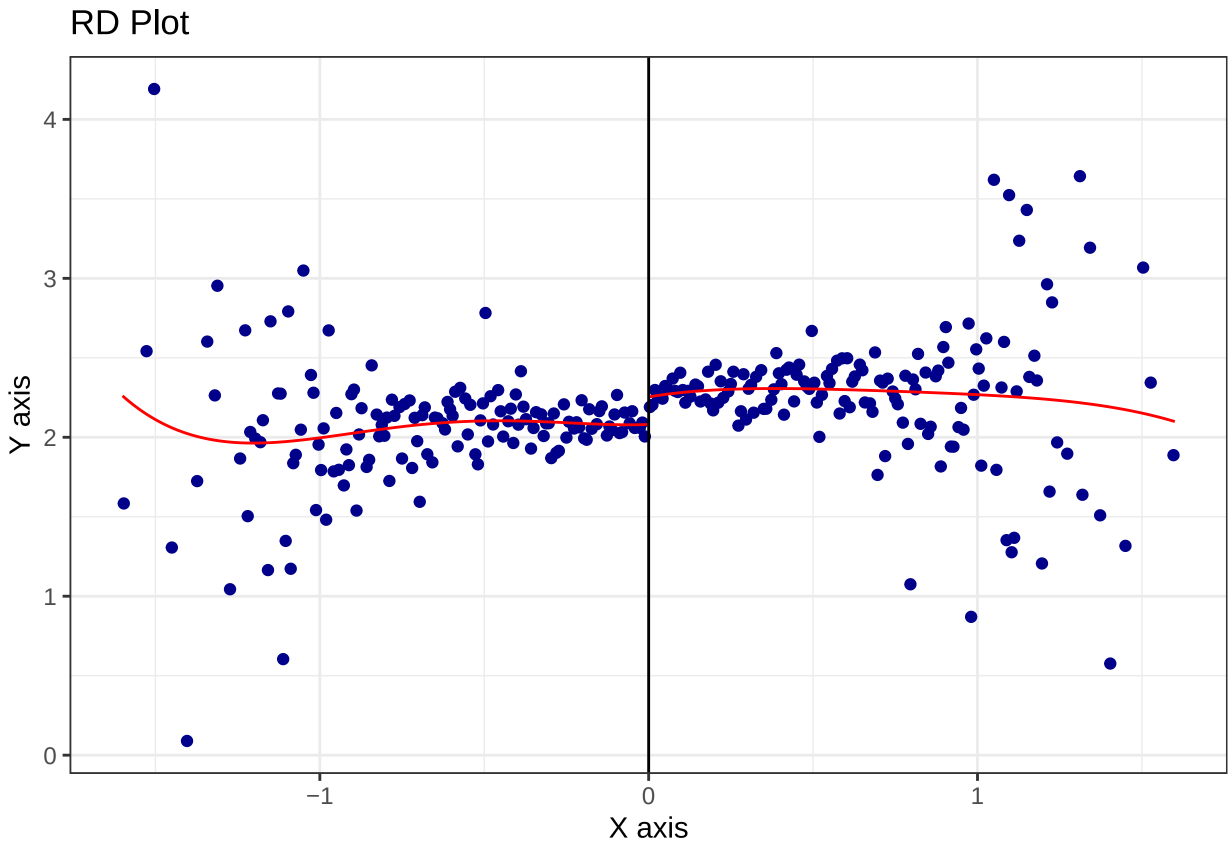
```
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                 20000
## BW type                        mserd
## Kernel                    Triangular
## VCE method                        NN
##
## Number of Obs.                  9837        10163
## Eff. Number of Obs.             7118         7444
## Order est. (p)                     1            1
## Order bias  (q)                    2            2
## BW est. (h)                    0.339        0.339
## BW bias (b)                    0.602        0.602
## rho (h/b)                      0.564        0.564
## Unique Obs.                      457          458
```

```
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional     0.182     0.031     5.807     0.000     [0.121 , 0.244]
##         Robust         -         -     4.926     0.000     [0.108 , 0.250]
## =============================================================================
```

Using a regression discontinuity design centered at a bid difference cutoff of 0, we estimate that being ranked 1 (vs. 2) increases click-through rate by approximately 18.2 percentage points. This effect is statistically significant with a p-value 0f 0.00 which < 0.05 and reflects the local causal impact of just securing the top ad slot.

```
rdplot(rd_filtered$ctr, rd_filtered$z, c=0)
```

```
## [1] "Mass points detected in the running variable."
```



RDD Visualization Interpretation This scatterplot shows the relationship between the forcing variable (z, bid difference) on the X-axis and the outcome (likely CTR) on the Y-axis. The vertical black line at $z = 0$ marks the cutoff between rank 2 and rank 1.

The red curve indicates a fitted regression line on either side of the cutoff. A visible upward jump at the cutoff suggests that ads ranked 1 experience higher click-through rates compared to those ranked 2 — consistent with a positive local treatment effect.

Let me know if you want this paired with your rdrobust() result.

```r
# Okay let's run our RDD regression now. Our estimate falls to about 5% with this tighter bandwidth.

rdd_model = lm(ctr ~ z + ctr + z*ctr, data =rd_filtered)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on
## the right-hand side and was dropped

## Warning in model.matrix.default(mt, mf, contrasts): problem with term 2 in
## model.matrix: no columns are assigned
```

```r
summary(rdd_model)
```

```
##
## Call:
## lm(formula = ctr ~ z + ctr + z * ctr, data = rd_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1065 -0.7598 -0.0029  0.7676  3.3277
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.187155   0.007068 309.424  < 2e-16 ***
## z           0.157649   0.051034   3.089  0.00201 **
## ctr:z       0.036544   0.021269   1.718  0.08579 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9965 on 19997 degrees of freedom
## Multiple R-squared:  0.006362,   Adjusted R-squared:  0.006263
## F-statistic: 64.02 on 2 and 19997 DF,  p-value: < 2.2e-16
```

RDD Linear Model Interpretation We estimate the local treatment effect of being ranked 1 (vs. 2) on click-through rate (CTR) using a regression around the bid difference cutoff ($z = 0$).

Being ranked 1 increases CTR by 15.8 percentage points, and the effect is statistically significant ($p < 0.05$). The interaction term (ctr:z) is not significant ($p = 0.086$), suggesting that slope differences across the cutoff are modest and not driving the observed jump.

Assumptions for the Study

Unconfoundedness (PSM): After matching on observed covariates (e.g., income, education, historical purchases), treatment assignment (directed vs. undirected search) is assumed to be as-good-as-random. This ensures that differences in outcomes can be attributed to treatment.

Overlap (PSM): There is sufficient overlap in covariate distributions between treated and control units, so that every treated unit has comparable controls — ensured via caliper-based matching.

Log-Linearity (Log Models): The log-transformed outcome variables (e.g., $\log(\text{OverallSales} + 1)$) are linearly related to the binary treatment indicator. This transformation addresses skewness and satisfies homoscedasticity assumptions.

Synthetic Control Validity: Exogeneity of Treatment (California): The cigarette tax policy was implemented independently of unobserved shocks to cigarette consumption.

Pre-Treatment Fit: Synthetic control closely matches California's sales trends before 1989, validating the assumption that the synthetic counterfactual is credible.

Continuity (RDD): All other factors affecting CTR are assumed to change smoothly around the bid difference cutoff. Only treatment assignment (rank 1 vs. 2) changes discontinuously at z = 0.

Local Randomization (RDD): Near the threshold, ad rank assignment is effectively random. This justifies causal interpretation of the observed jump in CTR.

Confidence Level: A 95% confidence level is used across all models to assess statistical significance.