MY457: Solution Set 2 - Selection on Observables

WT 2025

1 Concepts

This question reviews some of the concepts covered in class. Mathematical notation can be a useful tool to explain concepts, but it's important that you understand and can explain the concepts clearly and concisely. If you want to support your explanations with mathematical notation, this page provides a tutorial on including mathematical notation in Rmarkdown.

Consider a simple study of the effect of a treatment $D_i \in 0, 1$ on Y_i for all $i \in 1, ..., N$ where believe that the conditional independence assumption holds, conditional on observable characteristics X.

1.1. Explain the notation $(Y_1, Y_0) \perp D|X$. Does this mean that $(Y_1, Y_0) \perp X$?

The first part represents the conditional ignorability assumption. It states that, given a certain value of $X = x$, there is no association between potential outcomes and treatment status.

This assumption does not mean that $(Y_1, Y_0) \perp X$. Conditional ignorability refers to potential outcomes being independent of treatment for a given value of X, however, potential outcomes may well be associated with X. In fact, if X is a candidate confounder, we would expect this!

1.2 Consider a regression-based estimator of the average treatment effect (ATE) through the following specification:

$$Y_i = \hat{\alpha} + \hat{\tau} D_i + \hat{\beta} X_i + \hat{\varepsilon}_i$$

Given conditional ignorability and common support, what are the underlying assumptions for the estimation of the ATE using regression?

1- Constant or homogeneous treatment effects. 2- Linearity in the relationship between outcomes Y, treatment D and covariates X.

1.3 When using a matching procedure, it is possible to use the same observation as a control unit for multiple treated units. What are the advantages of using matching with replacement compared to matching without replacement?

It is often the case that one control unit is the best match for multiple treated units. When this is the case, matching with replacement will ensure that each treated observations has its best match as the counterfactual.

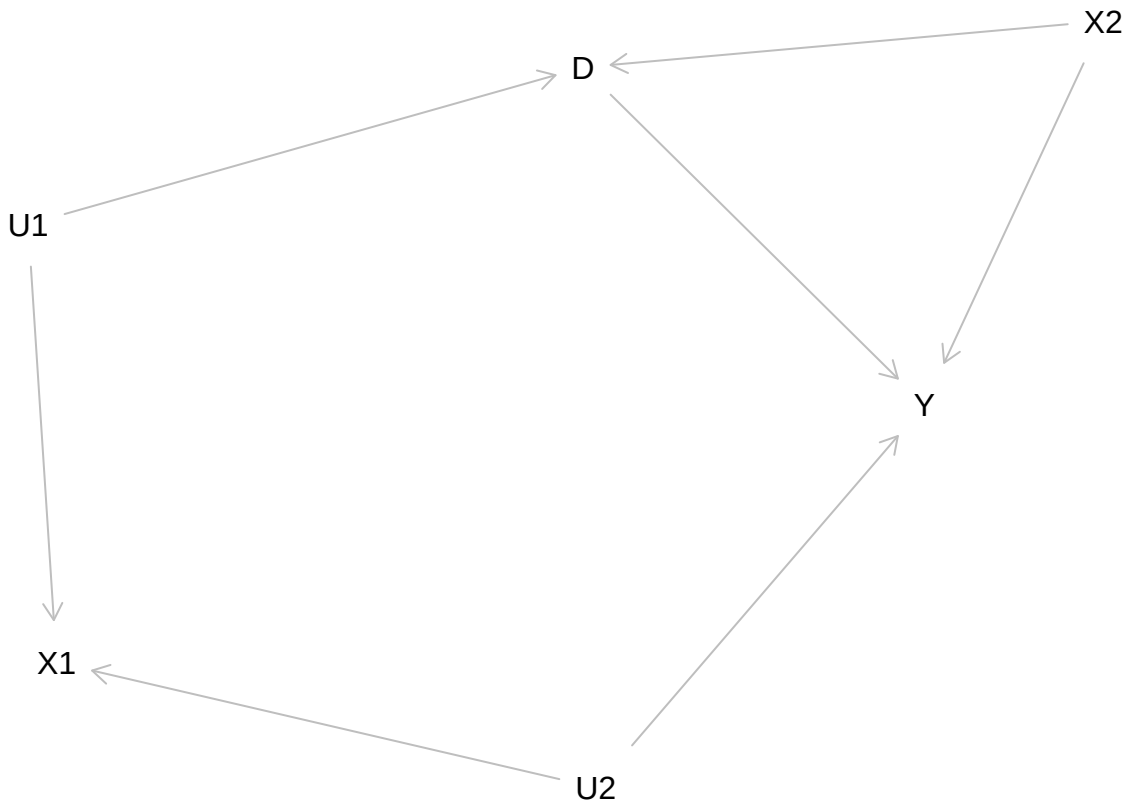
However, by using the same unit multiple times, we introduce some bias. The bias may arise from correlation between observations, or beacause there is not enough variance in the missing potential outcome.

2 Simulations

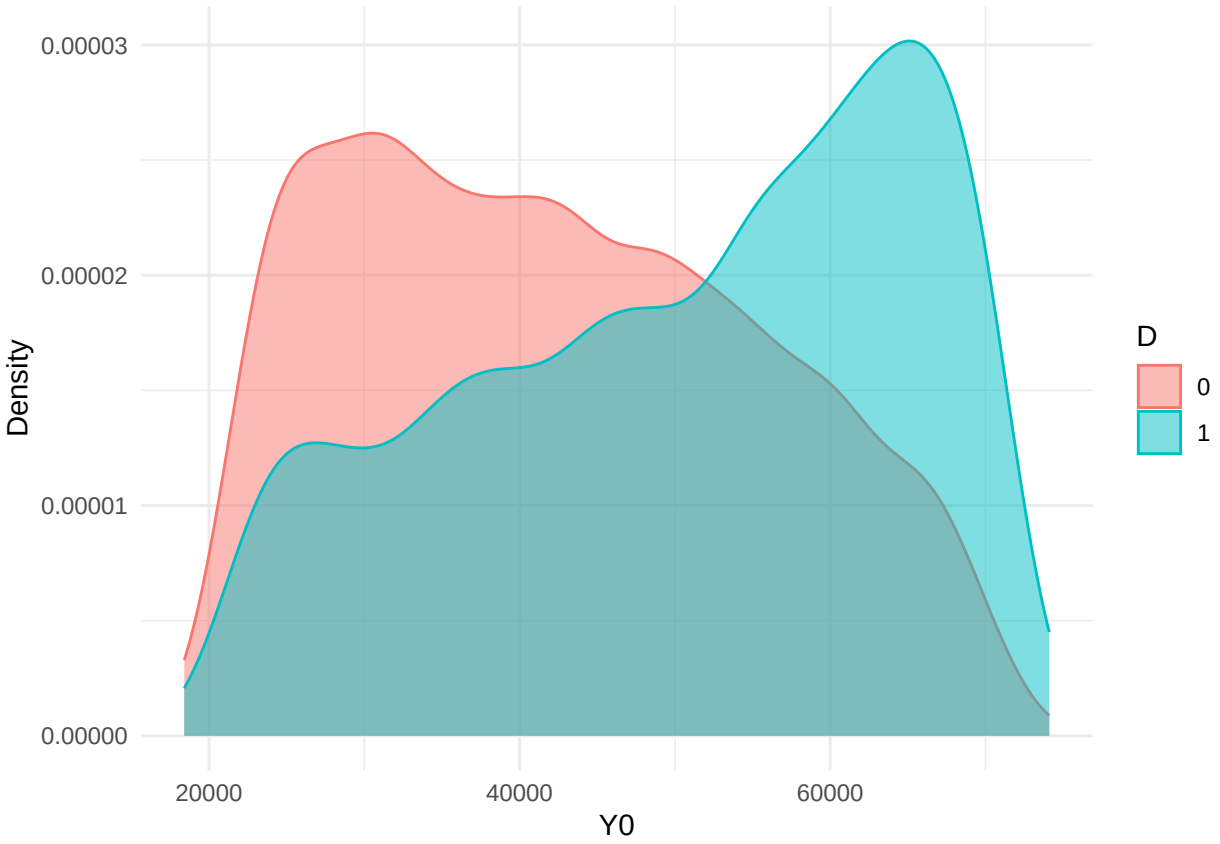
In this question we will use simulated data to test some of our intuitions about randomised experiments. The advantage of using a simulated dataset is that we have explicit control over the data generating process, and know the 'true' answer to any question we pose.

2.1. Suppose the government wants to implement an intensive course (D), designed for recently unemployed people to help them get back into the job market and improve their salaries (Y). As the resident data scientist you argued for D to be randomized, but unfortunately you were overruled by your seniors who have political machinations. They do, however, let you measure two covariates, $X1$ and $X2$. Below we simulate some data from this setting. Draw a DAG that represents the simulated data generating process (there are multiple packages available for drawing DAGs in RMarkdown). Using your DAG, explain in words what the code below does.

```r
set.seed(123)

n_obs <- 10000

tau <- 25000

# Create the dataset
U1 <- rbinom(n_obs, 1, 0.5)
U2 <- rbinom(n_obs, 1, 0.5)

X2 <- sample(1:50, n_obs, replace = TRUE)
X1 <- 10 + 1500*U1 + 1500*U2 + rnorm(n_obs, mean = 0, sd = 100)

Y0 <- 20000 + 1000*X2 + 1500*U2 + rnorm(n_obs, mean = 0, sd = 1000)
Y1 <- Y0 + tau

prob_d <- pnorm(X2, mean = 50, sd = 25) + 0.5*U1
prob_d <- pmin(prob_d, 1)
D <- rbinom(n_obs, 1, prob_d)

sim_data <- data.frame(U1, U2, X1, X2, Y0, Y1, D)

sim_data$Y <- ifelse(sim_data$D == 1, Y1, Y0)
```



2.2 Generate a clear and compelling graph that shows the distribution of just Y_0, split by treatment status. Do you think we can assume that $(Y_1, Y_0) \perp D$?



We can clearly see that Y_0 is different for those who have treatment compared to those in control.

2.3 Check for balance in the covariates $X1$ and $X2$. What do you find?

Control	Treatment	P.value
1143.754	1913.499	0

Control	Treatment	P.value
21.63144	30.02665	0

The t-test suggest that the difference between the two groups is statistically significant, suggesting that there is no balance in covariates between the two groups.

2.4 As the resident data scientist you are faced with the task of estimating the effect of the course (D) on observed salaries (Y). First, naïvely estimate the ATE using a linear regression with only D as a covariate. Given what you know about the true ATE, do you find any evidence of selection bias from this approach?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42363.46	185.7472	228.0706	0
D	33405.94	277.9787	120.1745	0

Our true ATE is set to be 25,000. Our estimate is 33,405, almost 10,000 above the true one. We find evidence for selection bias since there is a big difference between the true and the observed ATE, though ideally we'd want to see this over repeated sampling.

2.5 Now include the variable $X2$. What happens to the estimate of the ATE? Why? Is $X2$ a good control or a bad control?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20717.97	25.8608900	801.1314	0
D	25005.27	26.2748734	951.6800	0
X2	1000.65	0.9088811	1100.9689	0

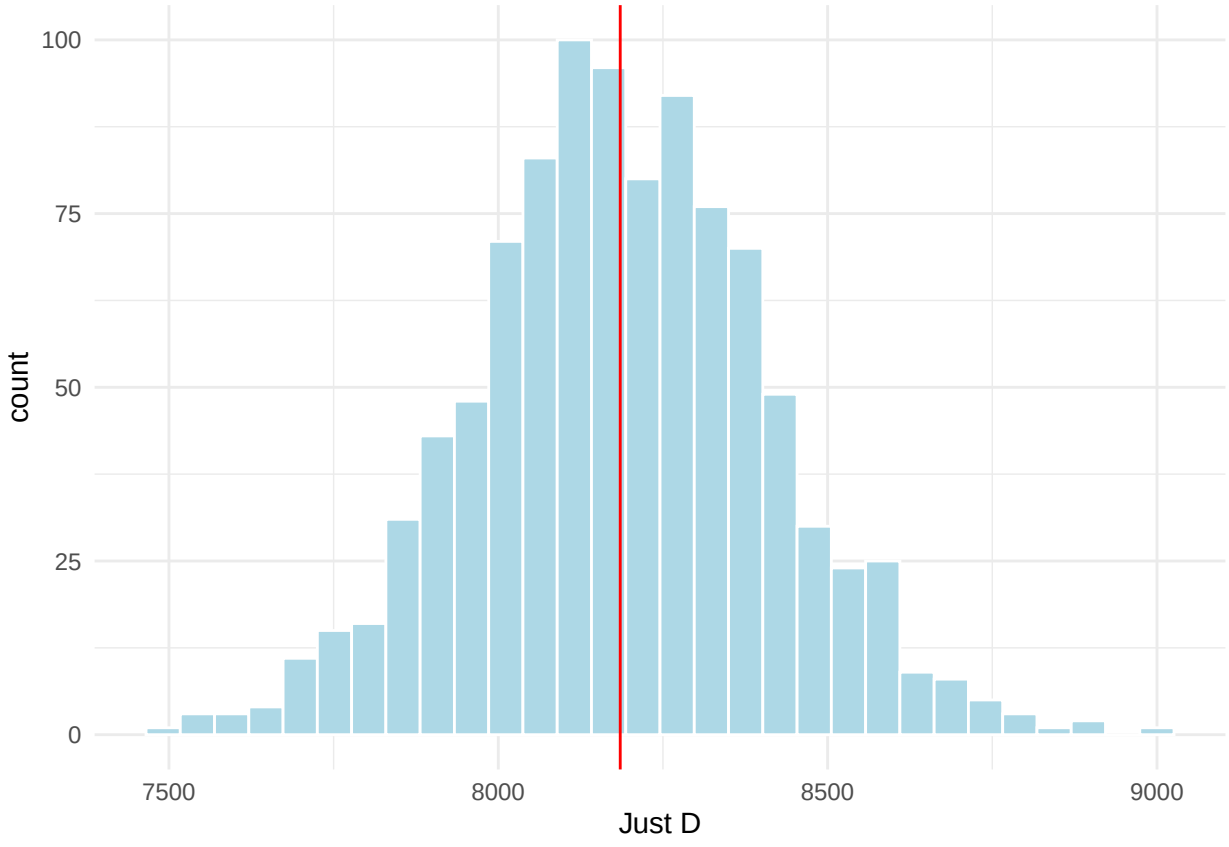
When we include X2 in the regression we get an estimate closer to the true ATE. X2 is a good control beacause it affects both the treatement status and potential outcomes. Controlling for X2 will account for the correlation between D and X2, effectively removing the selection bias part.

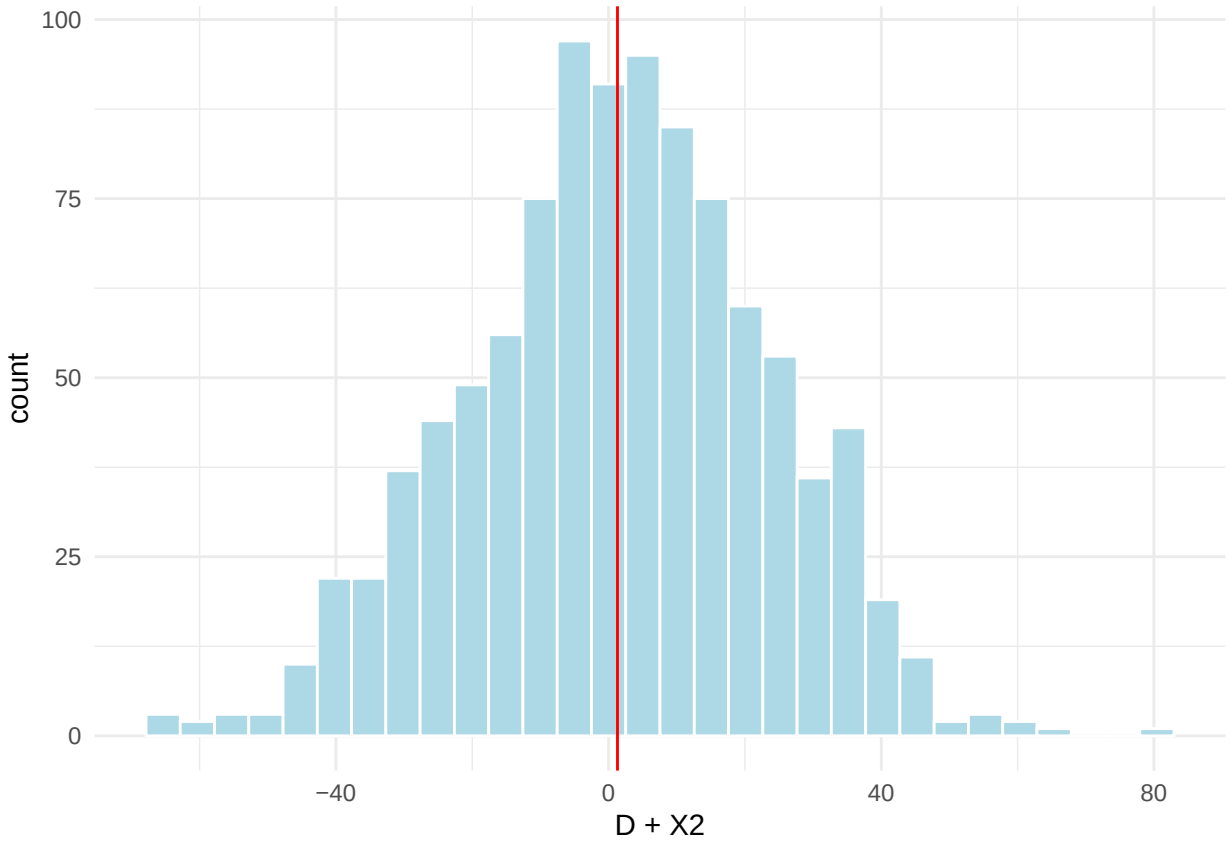
2.6 Now add to your linear regression the variable $X1$. What happens to the estimate of the ATE? Is $X1$ a good control or a bad control?

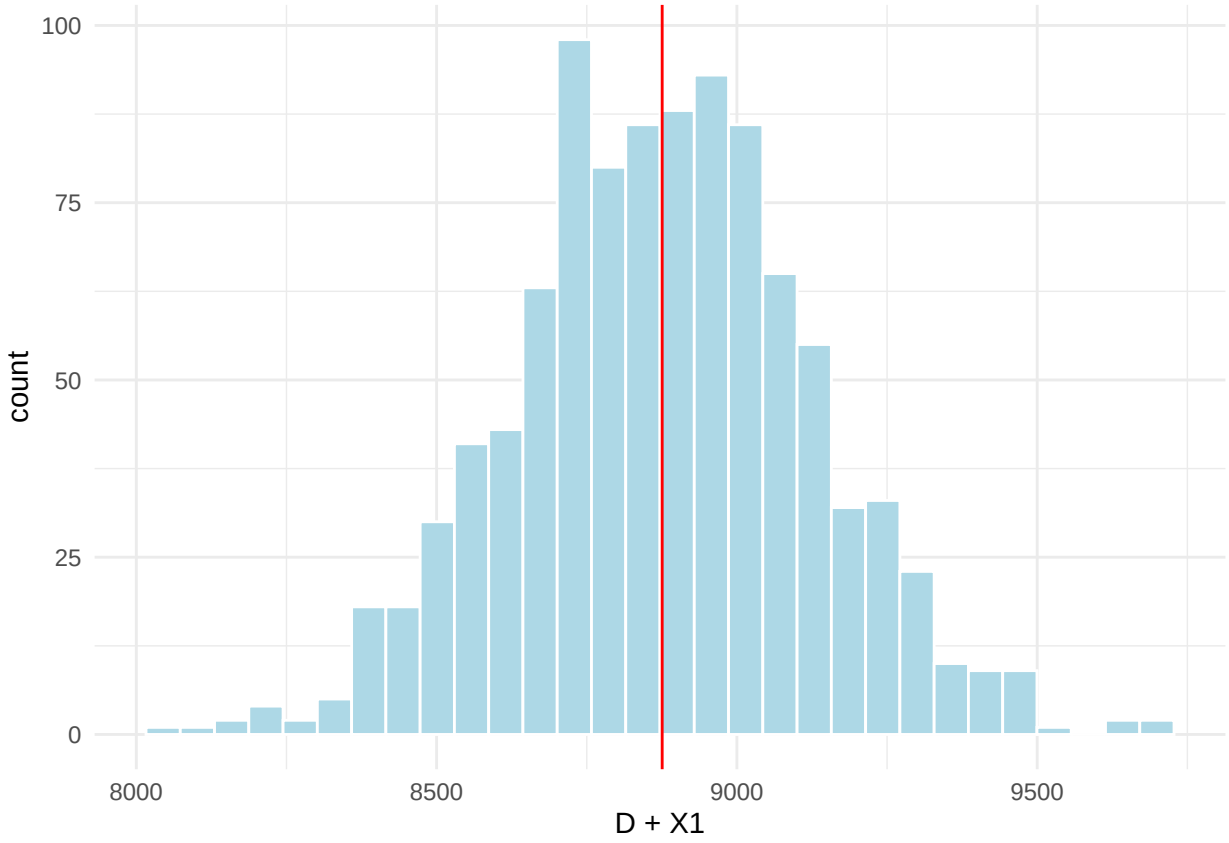
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	43494.5569201	244.452083	177.926718	0
D	34167.1658585	297.335067	114.911323	0
X1	-0.9889314	0.139409	-7.093739	0

We don't find the same when we control for X1. X1 is related to D, but it is not related to Y. Controlling for X1 opens a back-door path – we are conditioning on a collider. This is often called M-bias. X1 is a bad control.

2.7 (Extra credit): Show that your answers to Questions 2.4, 2.5, and 2.6 were not due to chance. Over at least 1,000 repeat simulations (note: you will want to remove the `set.seed` step), generate a fresh draw of the data and estimate the sample-specific ATE using the three different regression specifications. Calculate the difference between the estimated ATEs and what you know to be the true (fixed) value of the ATE, and store those differences. Finally, produce a histogram that shows the three distributions of the differences over your repeated samples, along with their means. What do you conclude?




100
75
50
25
0
count
−40
0
40
80
D + X2


100
75
50
25
0
count
8000
8500
9000
9500
D + X1

3 Replication

In this section, we will use real-world data to reinforce what we have learned. We will analyse the dataset employed in *Dollars on the Sidewalk: Should U.S. Presidential Candidates Advertise in Uncontested States?*.

In the USA, political parties spend over 750 million dollars on television adverts during campaigns. However, because of the nature of the US political system, many parties only advertise in states they deem competitive. However, beyond votes, advertising might also increase the amount of financial contributions people make to political campaigns. The authors study whether ads lead to greater levels of contributions to campaigns. To do this, they consider areas in non-competitive that are exposed to 'spillover' ads from bordering competitive states as treated units, and areas in non-competitive states that are not exposed to these ads as control units. They use a matching procedure to estimate the ATT of being exposed to at least a thousand ads.

3.1 In your own words, explain how the identification strategy that the authors use to estimate the effect of ad exposure on campaign contributions works. In your view, is this a case of selection-on-observables? (Hint: Consult pages 324 and 328 in the paper.)

The authors study zip-codes in non-competetitive states, where they shouldn't get much exposure to ads. They then focus on zip-codes that get spillover ads (treated) and compare them to zip-codes that do not have spillover ads (control). They match these two types of zip-codes and estimate an ATT. This does seem like selection-on-observables, but it is a little unclear what the observables are that drive the design.

3.2 Read into `R` the replication data file `dollars_on_the_sidewalk.dta`. These data are at the zip code level, and include data from both competitive and non-competitive states, indicated by the dummy variable `NonComp` which takes a value of 1 if non-competitive. For only those zip codes in non-competitive states, define a binary treatment variable (`Treated`) based on whether the zip code received more than a 1000 ads (`TotAds`). How many units are in the treated group, and how many are in the control group?

Var1	Freq
0	11165
1	6237

We have 11,165 zip codes in control and 6,237 in treatment.

3.3 What is the mean campaign contribution (`Cont`) for each level of `Treated`? Estimate the naïve ATE using these two quantities.

D	Cont
0	19.91722
1	20.16180

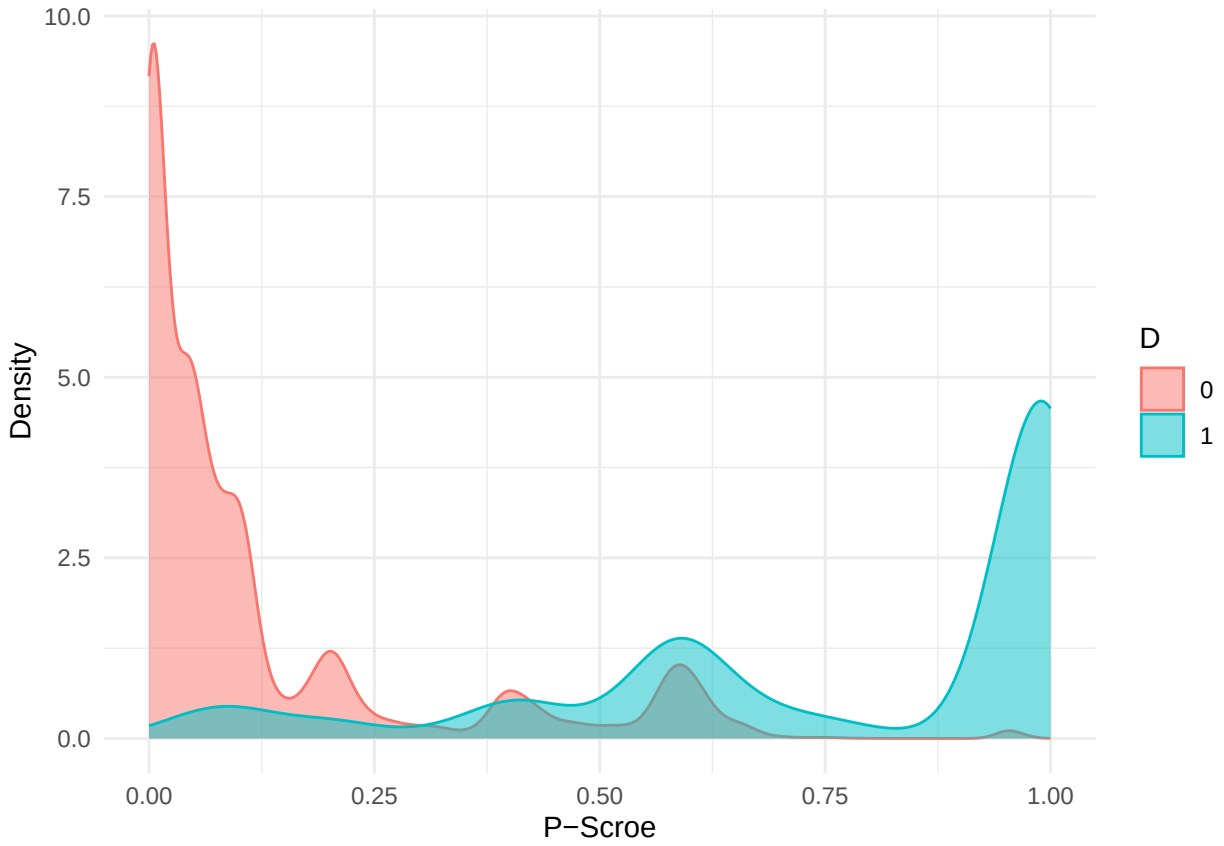
Diff.
0.244578

When we estimate it naively, we find an ATE of 0.24.

3.4 The authors use propensity score matching to more credibly estimate the average effect of `Treated` on `Cont` for the treated. Using whatever pre-treatment covariates you deem appropriate, fit a logistic regression to estimate the propensity score for each unit in the data. (Hint: Use `family = binomial(link = "logit")` in the `glm` function. To predict probabilities, use `predict()` but remember to set `type = "response"`.)

We estimate a logit model using *Income*, *Percent of Hispanics*, *Percent of Blacks*, *Population Density* and *State* as covariates.

3.5 In a single well-formatted and compelling graph, show the propensity score distributions for the two groups. What do you conclude?



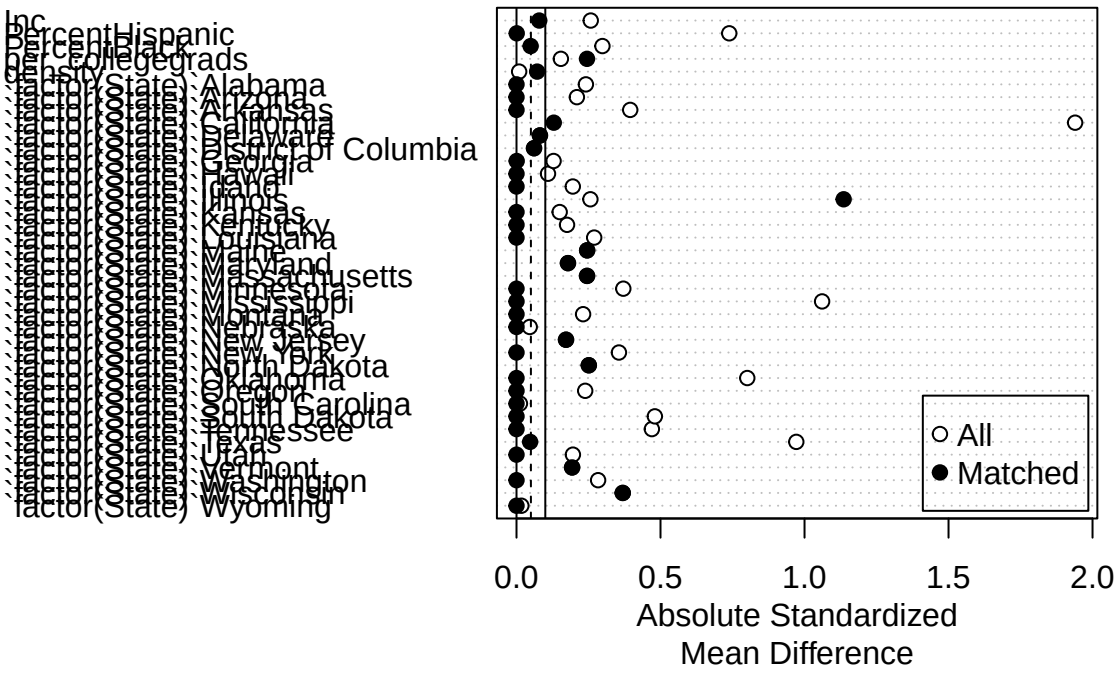
The density of both groups shows overlapping p-scores. This means that we are able to find a similar unit for treatment and control.

3.6 Implement 1:1 matching with replacement using the propensity scores. How many observations are left in your matched dataset? Explain what has happened. (Hint: You can use a canned package like `MatchIt` to estimate new p-scores and find the nearest matches, or the `Matching` package using your previously estimated p-scores, or for extra credit you can hand-roll your own matching function using your previously estimated p-scores.)

```
## [1] 7835
```

We end up with 7,835 observations. Out of which 6,237 are in treatment and 1,598 in control.

3.7 Show balance in your matched data for a range of different pre-treatment covariates. Again, `MatchIt` or `Matching` or similar packages may help. What do you find?



We find balance for almost all variables. In the demographics, the only covariate for which we do not find balance is the percentage of college graduates.

3.8 Based on your matched sample, use linear regression to estimate the ATT as the difference-in-means between the treatment categories.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.807596	2.725609	4.332094	0.0000150
D	8.354206	3.054887	2.734702	0.0062577

We find an effect of 8.35 when estimating the ATT. We know we are estimating the ATT instead of the ATE because we match the units in treatment with some control units, thus focusing on the effect on the treated.

3.9 (Extra Credit): What do you think of the research design used in this paper? Are there any weaknesses or concerns? Can you come up with any improvements to the design?

One major concern is that people living close to the border are different from those living closer to the interior. This may have issues for external validity. There may also be a compliance issue of a different type, whereby individuals who live in border zip codes may work or study in competitive states. If this is the case, individuals would be influenced by events happening in competitive states while living in non-competitive states, again suggesting meaningful differences. An improvement might be to only study people who live on borders, but take as treated those who live on competitive borders and those who live on non-competitive borders as control.

4 Appendix

```
# 2-----
## 2.1
dag <- dagitty('
  U1 -> X1
  U2 -> X1
  U2 -> Y
```

```
  X2 -> Y
  D -> Y
  U1 -> D
  X2 -> D
  ')

plot(dag)

## 2.2
ggplot(sim_data) + aes(x = Y0, fill = factor(D), color = factor(D)) +
  geom_density(alpha = 0.5) +
  xlab("Y0") +
  ylab("Density") +
  labs(fill = "D", color = "D") +
  theme_minimal()

## 2.3
t.test(X1~D, data = sim_data)

t.test(X2~D, data = sim_data)

## 2.4
summary(lm(Y~D, data = sim_data))

## 2.5
summary(lm(Y~D+X2, data = sim_data))

## 2.6
summary(lm(Y~D+X1, data = sim_data))

## 2.7
ATE1 <- NULL
ATE2 <- NULL
ATE3 <- NULL

set.seed(321)
for (i in 1:1000){
  D <- rbinom(n_obs, 1, prob_d)
  sim_data <- data.frame(U1, U2, X1, X2, Y0, Y1, D)
  sim_data$Y <- ifelse(sim_data$D == 1, Y1, Y0)

  reg1 <- lm(Y~D, data = sim_data)
  reg2 <- lm(Y~D+X2, data = sim_data)
  reg3 <- lm(Y~D+X1, data = sim_data)


  ATE1[i] <- coef(reg1)[2]
  ATE2[i] <- coef(reg2)[2]
  ATE3[i] <- coef(reg3)[2]
}

ggplot(data.frame(ATE1)) + aes(x = ATE1 - tau) +
  geom_histogram(color = "white", fill = "lightblue") +
```

```r
  geom_vline(xintercept = mean(ATE1-tau), color = "red") +
  xlab("Just D") +
  theme_minimal()

ggplot(data.frame(ATE2)) + aes(x = ATE2 - tau) +
  geom_histogram(color = "white", fill = "lightblue") +
  geom_vline(xintercept = mean(ATE2-tau), color = "red") +
  xlab("D + X2") +
  theme_minimal()

ggplot(data.frame(ATE3)) + aes(x = ATE3 - tau) +
  geom_histogram(color = "white", fill = "lightblue") +
  geom_vline(xintercept = mean(ATE3-tau), color = "red") +
  xlab("D + X1") +
  theme_minimal()


# 3 -------
## 3.2
data <- read_dta("./dollars_on_the_sidewalk.dta")

data <- data %>% filter(NonComp == 1) %>%
  mutate(D = ifelse(TotAds > 1000, 1, 0)) %>%
  na.omit()

table(data$D)

## 3.3
data %>% group_by(D) %>%
  summarise(
    Cont = mean(Cont, na.rm = T)
  )

mean(data$Cont[data$D == 1]) - mean(data$Cont[data$D == 0])

## 3.4
p_score <- glm(D~Inc+PercentHispanic+PercentBlack+per_collegegrads+density+factor(State),
  data = data,
  family = binomial(link = "logit"))
data$p_score <- predict(p_score, type = "response")

## 3.5
ggplot(data) + aes(x = p_score, fill = factor(D), color = factor(D)) +
  geom_density(alpha = 0.5) +
  xlab("P-Scroe") +
  ylab("Density") +
  labs(fill = "D", color = "D") +
  theme_minimal()

## 3.6
m_out <- MatchIt::matchit(D~Inc+PercentHispanic+PercentBlack+per_collegegrads+density+factor(State),
  data = data,
  method = "nearest",
```

```r
  link = "logit",
  ratio = 1,
  replace = T,
  distance = "mahalanobis")

m_data <- MatchIt::match.data(m_out)
dim(m_data)[1]

## 3.7
plot(summary(m_out))

## 3.8
summary(lm(Cont~D, data = m_data))
```