MY457: Problem Set 2

L. Cäcilia Präckel

Do/15/Feb

1 R Markdown

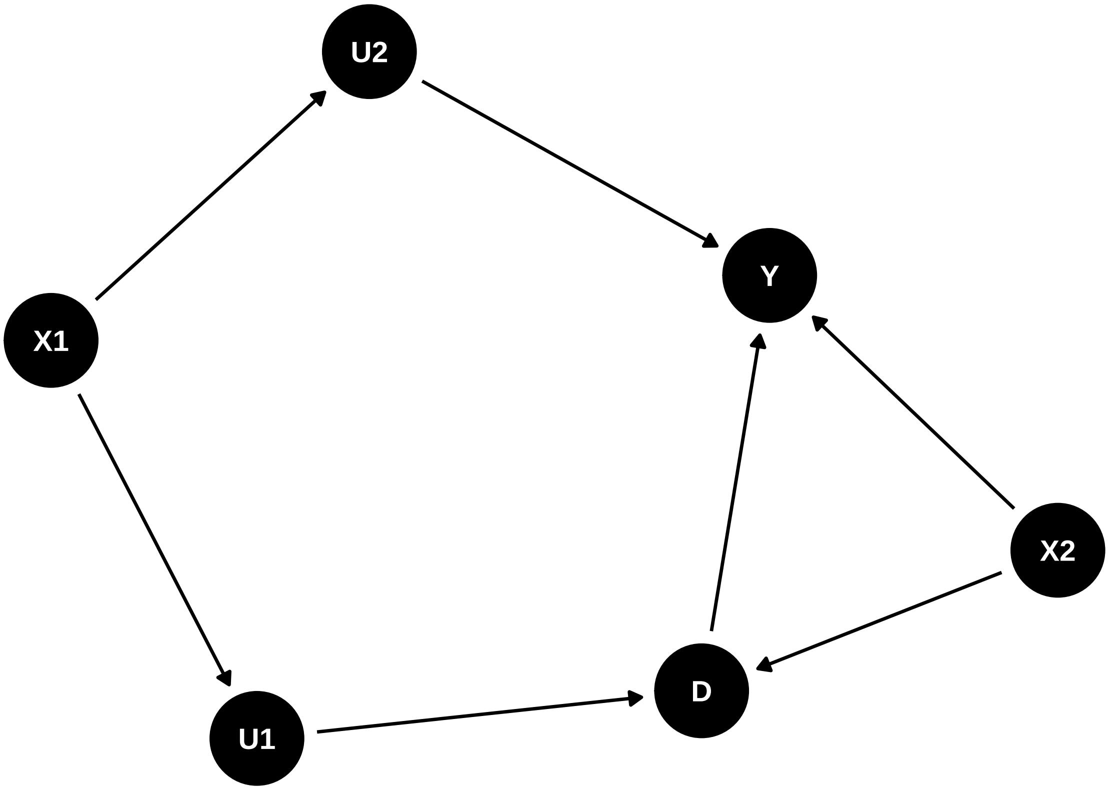
You must submit your problem set solutions as a .pdf. We strongly recommend that you use `R Markdown`, which allows you to integrate R code and your own writing/commentary into a single reproducible document. This specific template is set up as a `Bookdown` document written to .pdf. You should click the **Knit** button in `RStudio` to produce the .pdf document. The majority of your code should appear **only** in the code appendix at the end of the document. The main part of your document should be written text, plus tables, plots, and occasionally code chunks where relevant (e.g. if you write a particular function you want to discuss in detail).

1 Concepts 1.1 The notation $(Y_1, Y_0) \perp D|X$ means that the potential outcomes (Y_1, Y_0) are independent of the treatment D conditional on the covariates or pre-treatment variable X, this is also called the conditional independence assumption. It does not mean that (Y_1, Y_0) are independent from all X. In a sense, we make a little mini-experiment with the conditional independence assumption by saying that every observation that has the same X, the same pre-treatment, we assume that they also get the same D, may it be treatment or control. This is a strong assumption, because it means that we assume that e.g. people with the same characteristics (the same X) are put into the same treatment group (the same D), regardless of other factors that might influence their treatment.This could lead to a selection bias, and it is a form of confounding. For example, if we are interested in the effect of a job training program on earnings, we might expect that the potential earnings of a person who receives the training (Y_1) are different from the potential earnings of a person who does not receive the training (Y_0), even after controlling for the person's characteristics X.

1.2 We can only use the linear regression model to estimate the average treatment effect (ATE) if 1) we believe that there is a linear relation between the treatment and the outcome, i.e. if we believe that the potential outcomes are produced by a linear function and 2) if the treatment effect is constant, which is a direct consequence from the assumption number one.

1.3 Matching with replacement can lead to better matches than matching without replacement because for each unit in the treatment group we can find the best match in the control group drawing from the whole sample. This is especially useful if the treatment group is big and the control group is small. However, matching with replacement allows to match the same control unit to multiple treatment units, which can lead to a bias in the estimation of the treatment effect.

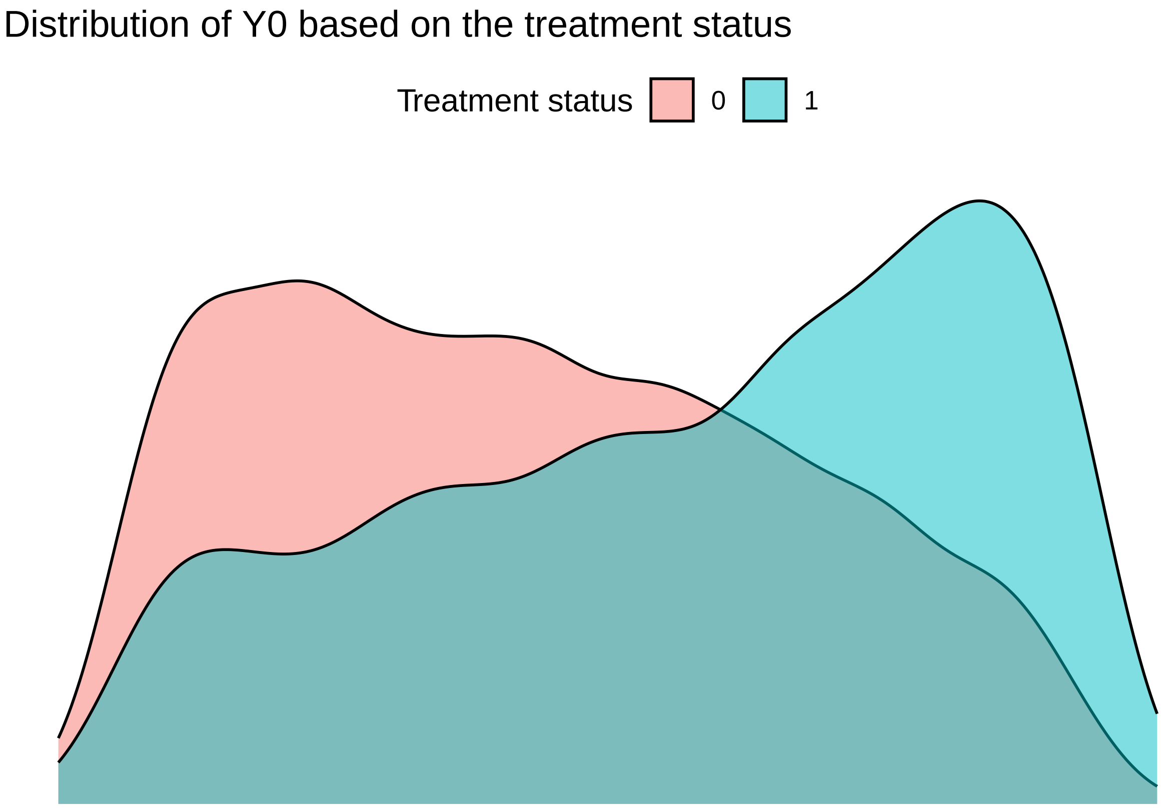
2 Simulations 2.1 The following DAG explains the data generation process for the simulated dataset. The potential outcomes of Y (Y1 and Y0) are influenced by the treatment variable D, the pre-treatment variable X2 and the unobserved variable U2. The treatment variable D is influenced by the pre-treatment variable X2 and the unobserved variable U1. Because X2 influences both D and Y, it is a confounder. The pre-treatment variable X1 influences both unobserved variables U1 and U2.



From the DAG we can see that the treatment variable D is influenced by the confounder X2 and the unobserved variable U1. The potential outcomes Y1 and Y0 are influenced by the treatment variable D, by X2 and the unobserved variable U2. The pre-treatment variable X1 influences the unobserved variables U1 and U2.

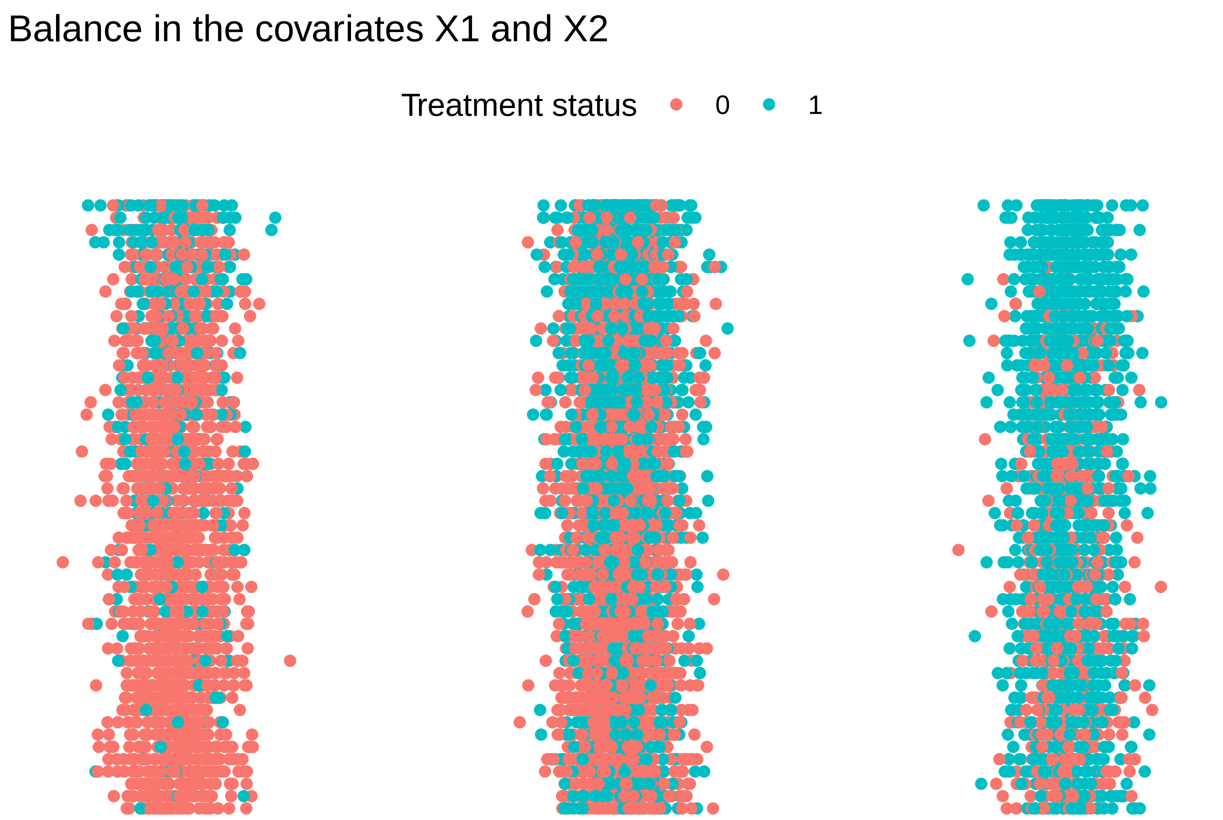
Explanations for the code used to generate the data are written as comments in the code (see appendix).

2.2 The below plot shows us the distribution of Y0. Y in our case is the income of the people who were offered the job training. Y0 is the potential outcome of the income when not getting treatment. In a perfect dataset, the income of people who were offered the job training would be distributed in the same way among the people in question as the income of people who were not offered the job training.



The plot shows us that the distribution of Y0 is different for the actual treatment and the actual control group (indicated by the treatment status). This means we cannot assume that the potential outcomes (Y_1, Y_0) are independent of the treatment D. Instead, they might be conditional on the covariates or pre-treatment variable $X1, X2, U1, U2$. Consequently, we shouldn't estimate the average treatment effect (ATE) using the linear regression model, because the treatment effect is not constant.

2.3 Checking for the balance of the observations in the covariates X1 and X2 helps us to identify whether the conditional independence assumption holds or not. Additionally, we can further investigate if we can use the linear model to adequately model estimate the ATE.



The plot shows us that the covariates X1 and X2 are not balanced between the treatment and control group. If we imagine X1 and X2 to be two characteristics of people, e.g. X1 is the age and X2 is the education level, then the plot shows us that the people who were offered the job training (treatment group) are different from the people who were not offered the job training (control group) in terms of age and education level.

2.4 When we estimate the average treatment effect (ATE) using a simple linear regression model, the ATE is estimated to be 33405.94, which is not very close to the true ATE of 25000. This is because the linear regression model is not able to estimate the ATE correctly, as shown above. This can be seen as a selection bias.

```
## [1] 33405.94
```

2.5 When we estimate the average treatment effect (ATE) using the linear regression model including X2, the ATE is estimated to be 25005.27. This is very close to the true ATE. Apparently, X2 gives an important piece of explanation for the treatment effect. However, as we can see from th dag, X2 is a common cause and a confounder of D and Y. Hence, we should control for X2 in the linear regression model to see the real effect of the treatment.

```
## [1] 25005.27
```

2.6 When we estimate the average treatment effect (ATE) using the linear regression model including X1, the ATE is estimated to be 34167.17. This is not close to the true ATE. X1 is a common cause of D and Y, a confounder, which makes it reasonable to control for X1. Also, X1 influences the unobservable U1 and by controlling for X1, we might take away the noise of Y that might be induced only by U1. However, X1 also influences U2, which affects D and hence controlling for X1 could amplify the bias in D. In this case, I would see more benefit in controlling for X1 than not controlling for it.

```
## [1] 34167.17
```

3 Replication 3.1 Identification strategy In the study of Urban and Niebler, the authors try to find out about the effect of advertisment on the voting behaviour. They compare non-competitive states with no advertisment and non-compeitive states with spill-over advertisment from neighbouring states aiming to find out about the treatment effect advertisement. They match areas on their zip code using propensity scores. The aim is to find out about the counterfactual of how households would have voted if they had not been exposed to advertisement. This is a typical missing data question. To find the most explaining variables, they included the observed variables density, median household income, percentage of Hispanics and African American, and college graduates.

3.2 There are 14124 units in treated group and 16447 units in the control group. The treated group is bigger than the control group but before we can say anything about the balance of the covariates, we would need to check the distribution of the covariates in the treatment and control group, such as the mean contribution below.

```
## # A tibble: 30,571 x 89
##    zip   TotalPop MedianHHInc PerCapitaHHInc MaleOver65 FemaleOver65
##    <chr>    <dbl>       <dbl>          <dbl>      <dbl>        <dbl>
##  1 01001    16475       45735          22490       1234         2133
##  2 01002    36776       42567          18212        999         1451
##  3 01005     5079       50395          20518        273          382
##  4 01007    12997       52425          21923        500          636
##  5 01008     1234       52663          23680         56           60
##  6 01010     3350       50181          23697        158          208
##  7 01011     2085       45511          19963         97          120
##  8 01012      521       50938          19137         14           20
##  9 01013    22963       32412          16608       1410         2423
## 10 01020    29325       37282          20205       2204         3200
## # i 30,561 more rows
## # i 83 more variables: PercentOver65 <dbl>, Rural <dbl>, Urban <dbl>,
## #   PercentWhite <dbl>, PercentBlack <dbl>, PercentHispanic <dbl>,
## #   amount <dbl>, AmountRep <dbl>, AmountDem <dbl>, AmountRCand <dbl>,
## #   AmountDCand <dbl>, AmountRComm <dbl>, AmountDComm <dbl>, rep <dbl>,
## #   dem <dbl>, meanrep <dbl>, meandem <dbl>, ‘_merge1‘ <dbl+lbl>,
## #   StCtyFIPS <chr>, StDMACode <dbl>, ‘_merge2‘ <dbl+lbl>, DMACode <dbl>, ...
```

```
## # A tibble: 2 x 2
##   Treated     n
##     <dbl> <int>
## 1       0 14124
## 2       1 16447
```

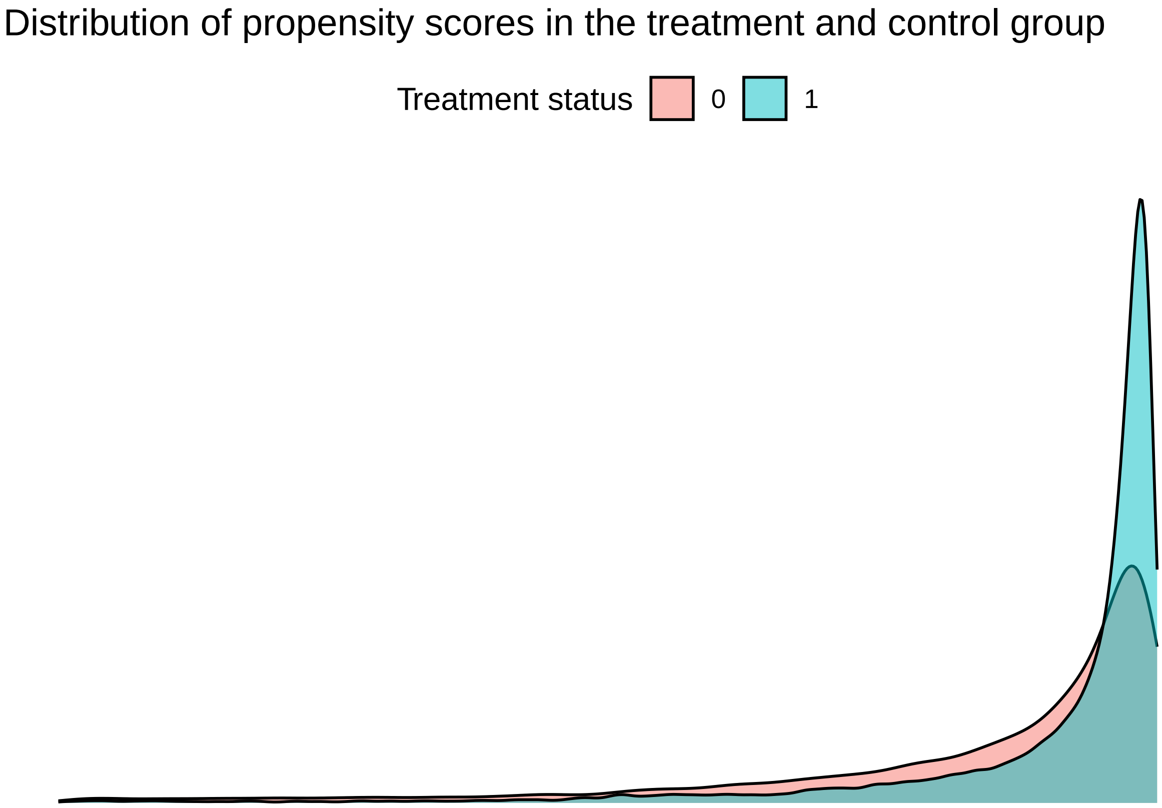
3.3 The mean campaign contribution for the control group is 22,365.72 and 16,946.62 for the treatment group. The naive ATE, which is the differences between those means, is 5,419.1. This naive ATE does not solve our missing data problem and we need further calculation.

```
## # A tibble: 2 x 2
##   Treated mean_cont
##     <dbl>     <dbl>
## 1       0    22366.
## 2       1    16947.
## # A tibble: 1 x 1
##     ate
##   <dbl>
## 1 5419.
```

3.4 The propensity score is the probability of being treated given the observed covariates. It is the conditional probability of receiving the treatment given the observed covariates. Below we see the first ten observations with their propensity scores.

```
##    [,1]    [,2]
## 1  "01001" "0.586396851072955"
## 2  "01002" "0.54523871951436"
## 3  "01005" "0.591512428665537"
## 4  "01007" "0.582441594029386"
## 5  "01008" "0.590586682266538"
## 6  "01010" "0.588438547559168"
## 7  "01011" "0.594192342253107"
## 8  "01012" "0.59777302885376"
## 9  "01013" "0.487604961721251"
## 10 "01020" "0.566009722103774"
```

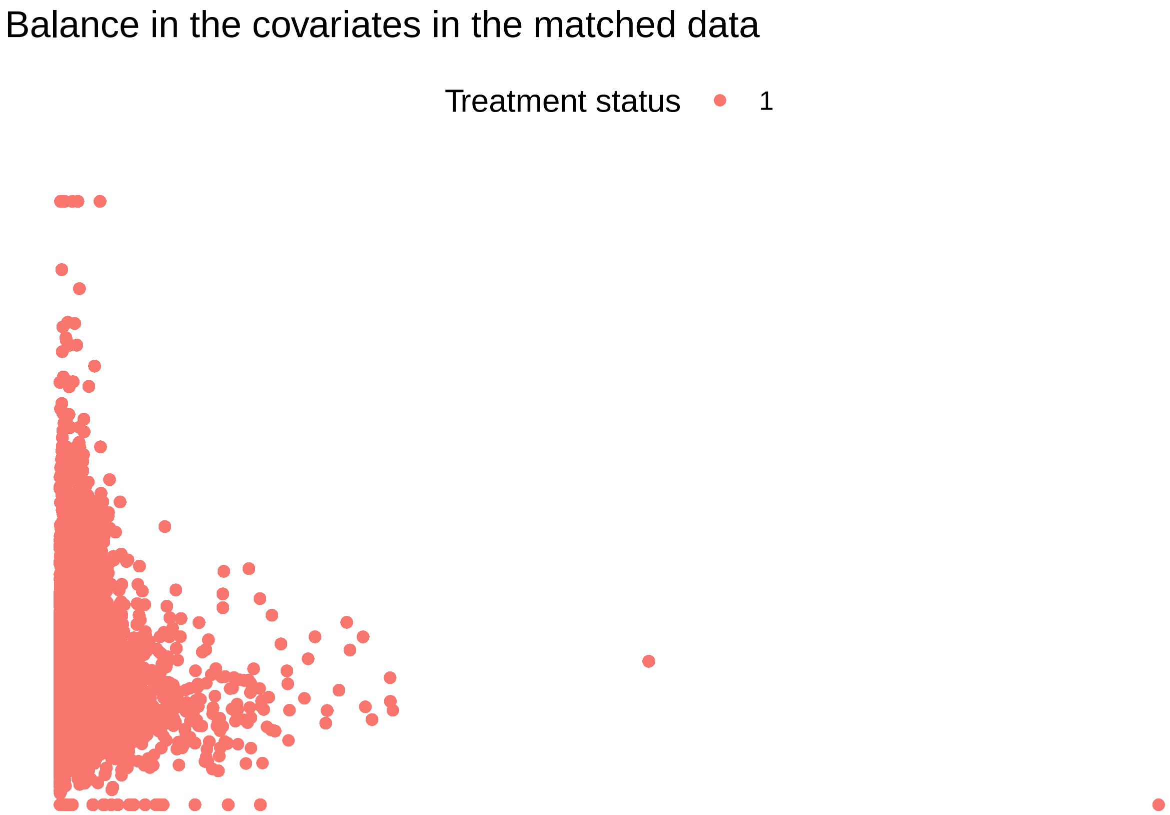
3.5 The plot below shows us the distribution of propensity scores in the treatment and control group. The propensity score for the treatment group is much higher than for the control group, which means, the probability to get treated was higher for some households and was dependent on one or more of the covariates $density, MedianHHInc, PercentHispanic, PercentBlack$. This means we cannot assume that the potential outcomes (Y_1, Y_0) are independent of the treatment D.



3.6 I am using the Matching package to match the observations on the previously calculated propensity scores. 2.685 observations are left after matching. This means that these observations in the control group were not matched to observations in the treatment group. This means that the 16,436 treated observation were not matched using all 14.104 control observations, but because of the method of replacement, only with 11.309 control group observations. To me, this seems to be a good balance in the covariates between the treatment and control group, although not a very good one.

```
## [1] 2685
```

3.7



Balance in the covariates in the matched data

3.8

```
## [1] NA
```

2 Code appendix

```
# this chunk contains code that sets global options for the entire .Rmd.
# we use include=FALSE to suppress it from the top of the document, but it will still appear in the app
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE, linewidth=60)

# you can include your libraries here:
library(haven)
library(tidyverse)
library(ggdag)
library(ggplot2)
theme_set(theme_dag())

# and any other options in R:
options(scipen=999)

dagify(U1 ~ X1,
       U2 ~ X1,
       Y ~ U2,
       Y ~ X2,
       D ~ X2,
       D ~ U1,
       Y ~ D
       ) %>%
  ggdag()
set.seed(123)
# Create a simulated dataset of 10000 observations with an ATE of 25000
n_obs <- 10000
tau <- 25000
# Generate two binary variables which have either the value 0 or 1 for each observation
U1 <- rbinom(n_obs, 1, 0.5)
U2 <- rbinom(n_obs, 1, 0.5)
# Generate a continuous variable X2 which has values between 1 and 50 for each observation
X2 <- sample(1:50, n_obs, replace = TRUE)
# Generate a continuous variable X1 based on a linear model which is influenced by U1 and U2 and includ
X1 <- 10 + 1500*U1 + 1500*U2 + rnorm(n_obs, mean = 0, sd = 100)
# Generate the potential outcome Y0 (control group) based on a linear model which is influenced by X2 a
Y0 <- 20000 + 1000*X2 + 1500*U2 + rnorm(n_obs, mean = 0, sd = 1000)
# Define the potential outcome Y1 (treatment group) as Y0 + the treatment effect
Y1 <- Y0 + tau
# Generate the probability of the treatment variable D based on a probit model which uses X2 with a mea
prob_d <- pnorm(X2, mean = 50, sd = 25) + 0.5*U1
# pmin is the min-function for vectors, which in this case makes sure that we only choose values for th
prob_d <- pmin(prob_d, 1)
# Assign the treatment variable D based on the previously calculated probabilities of D being 1
D <- rbinom(n_obs, 1, prob_d)
# Write all the calculated and generated data into a single dataframe
sim_data <- data.frame(U1, U2, X1, X2, Y0, Y1, D)
# Create a new variable Y, which is the actual observed outcome for Y. If an observation has been assig
sim_data$Y <- ifelse(sim_data$D == 1, Y1, Y0)

sim_data %>%
  ggplot(aes(x = Y0, fill = factor(D))) +
```

```r
  geom_density(alpha = 0.5) +
  labs(title = "Distribution of Y0 based on the treatment status",
       x = "Y0",
       y = "Density",
       fill = "Treatment status") +
  theme(legend.position = "top")

# Check for balance in the covariates X1 and X2
sim_data %>%
  ggplot(aes(x = X1, y = X2, color = factor(D))) +
  geom_point() +
  labs(title = "Balance in the covariates X1 and X2",
       x = "X1",
       y = "X2",
       color = "Treatment status") +
  theme(legend.position = "top")

# Estimate the ATE using the linear regression model
lm_ate <- lm(Y ~ D, data = sim_data) %>%
  # Extract the ATE
  broom::tidy() %>%
  filter(term == "D") %>%
  pull(estimate)
lm_ate
# Estimate the ATE using the linear regression model including X2
lm_ate_x2 <- lm(Y ~ D + X2, data = sim_data) %>%
  # Extract the ATE
  broom::tidy() %>%
  filter(term == "D") %>%
  pull(estimate)
lm_ate_x2
# Estimate the ATE using the linear regression model including X1
lm_ate_x1 <- lm(Y ~ D + X1, data = sim_data) %>%
  # Extract the ATE
  broom::tidy() %>%
  filter(term == "D") %>%
  pull(estimate)
lm_ate_x1
# Load the data
ad_data <- read_dta("dollars_on_the_sidewalk.dta")

# To the data add a new variable called "treated" which is 1 if the TotAds > 1000  and 0 if not
ad_data <- ad_data %>%
  mutate(Treated = ifelse(TotAds > 1000, 1, 0))
ad_data

# Display number of units in treatment and control group
ad_data %>%
  count(Treated)

# Display the mean campaign contribution Cont for the treatment and control group
ad_data %>%
  group_by(Treated) %>%
  summarise(mean_cont = round(mean(Cont*1000), 2))

# Estimate naive ATE
naive_ate <- ad_data %>%
  group_by(Treated) %>%
  summarise(mean_cont = round(mean(Cont*1000), 2)) %>%
  summarise(ate = abs(diff(mean_cont)))
naive_ate
# Estimate the propensity score using a logistic regression model and the covariates density, median ho

# Estimate the propensity score using a logistic regression model
ps_model <- glm(Treated ~ density + MedianHHInc + PercentHispanic + PercentBlack, data = ad_data, famil

# Predict the propensity score
ps = predict(ps_model, type = "response")

# Add propensity score to data
ad_data <- ad_data %>%
  mutate(ps = ps)

# Print the first 10 observations of the zip code and propensity score
head(cbind(ad_data$zip, ad_data$ps), 10)

# Plot the distribution of propensity scores in the treatment and control group
ad_data %>%
  ggplot(aes(x = ps, fill = factor(Treated))) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribution of propensity scores in the treatment and control group",
       x = "Propensity score",
       y = "Density",
       fill = "Treatment status") +
  theme(legend.position = "top")

# Load library
library(Matching)

# Match the observations on the propensity scores
# First, drop NAs in ad_data$Cont and ad_data$ps
ad_data <- ad_data %>%
  drop_na(ps)
ad_data <- ad_data %>%
  drop_na(Cont)
matched_data <- Match(Y = ad_data$Cont,
                      # Tr to indicate which observation is treated
                      Tr = ad_data$Treated,
                      # X for which variable to match on, here: p-score
                      X = ad_data$ps,
                      # M for number of matches, here 1:1
                      M = 1,
                      replace = TRUE,
                      # Accelarate the matching process for large dataset
                      version = "fast")
```

```
# Number of observations left after matching
n_distinct(matched_data$index.treated) - n_distinct(matched_data$index.control)
# Check for balance in the covariates in the matched data
matched_data_analysis <- ad_data[matched_data$index.treated, ]
matched_data_analysis %>%
  ggplot(aes(x = density, y = MedianHHInc, color = factor(Treated))) +
  geom_point() +
  labs(title = "Balance in the covariates in the matched data",
       x = "Density",
       y = "Median household income",
       color = "Treatment status") +
  theme(legend.position = "top")

# Estimate the ATE using the matched data and linear regression
lm_ate_matched <- lm(Cont ~ Treated, data = matched_data_analysis) %>%
  # Extract the ATE
  broom::tidy() %>%
  filter(term == "Treated") %>%
  pull(estimate)
lm_ate_matched
# this chunk generates the complete code appendix.
# eval=FALSE tells R not to re-run (``evaluate'') the code here.
```