# MY457: Problem Set 2 - Selection on Observables

Wed/12/Feb

This problem set is formative and will not contribute to your final grade. However, it is designed to build and deepen your conceptual understanding of the material and to practice applying the material in R. Using tools and resources such as ChatGPT and Stack Overflow is allowed and encouraged. Bear in mind that you are ultimately responsible for the work you submit, and that these tools often provide incorrect solutions. Make sure that however you use these tools it aligns with your best interests, and enhances your learning in this class.

This problem set must be submitted on Moodle by 11am on Wed/19/Feb. You must also use the provided `.Rmd` template to produce a .pdf with your answers. If your submission is late, is not a .pdf, or is not appropriately formatted, you will not receive feedback on your work.

## 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 $(Y_1, Y_0) \perp X$?

**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?

**1.3** 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?

## 2 Simulations

In this question 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 $(Y_1, Y_0) \perp D$?

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

**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?

**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?

**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?

**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?

# 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.)

**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?

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

**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"`.)

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

**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.)

**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?

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

**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?