Econometrics II (Spring 2024)
Suggested Solutions to Problem Set 2

Shuhei Kainuma*

April 23, 2024

Question 1

a.

Take any observation $i \in \{1, \ldots, N\}$. The unit assignment probability is defined as (see Lecture 1)

$$p_i(\mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1})) = \sum_{\mathbf{D}: D_i = 1} \Pr(\mathbf{D} | \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1})),$$

where

$$\Pr(\mathbf{D} | \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1})) = \binom{N}{N_1}^{-1}.$$

This is because the assignment mechanism is the completely randomized experiment with $N_1 < N$. Hence, we can rewrite the unit assignment probability as

$$p_i(\mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1})) = \binom{N}{N_1}^{-1} \sum_{\mathcal{D}} \mathbb{1}\{\mathbf{D} : D_i = 1\}, \tag{1}$$

where $\sum_{\mathcal{D}} \mathbb{1}\{\mathbf{D} : D_i = 1\}$ is the total number of the assignments with $D_i = 1$. Now we compute this. The number of the assignments with $D_i = 1$ is equal to the number of all the possible assignments for the rest of the observations, which is choosing $N_1 - 1$ observations from $N - 1$ units. It is therefore given by $\sum_{\mathcal{D}} \mathbb{1}\{\mathbf{D} : D_i = 1\} = \binom{N-1}{N_1 - 1}$. Plugging this into Equation 1, we have

$$p_i(\mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1})) = \binom{N}{N_1}^{-1} \binom{N-1}{N_1 - 1} = \frac{\frac{(N-1)!}{(N_1-1)!(N-1-(N_1-1))!}}{\frac{N!}{N_1!(N-N_1)!}} = \frac{N_1}{N}.$$

*shuhei.kainuma@iies.su.se. Let me know if you spot any typos/errors!

b.

We want to show that using $\mathbb{V}\left(\hat{\beta}|\mathbf{D}\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega(\mathbf{D})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$,

$$\mathbb{V}\left(\hat{\beta}_D|\mathbf{D}\right) = \frac{\sum \sigma_i^2(1)D_i/N_1}{N_1} + \frac{\sum \sigma_i^2(0)(1-D_i)/(N-N_1)}{N-N_1}$$

Observe that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ D_1 & D_2 & \cdots & D_N \end{bmatrix} \begin{bmatrix} 1 & D_1 \\ 1 & D_2 \\ \vdots & \vdots \\ 1 & D_N \end{bmatrix} = \begin{bmatrix} \sum_{i\in\{1,\ldots,N\}} 1 & \sum_{i\in\{1,\ldots,N\}} D_i \\ \sum_{i\in\{1,\ldots,N\}} D_i & \sum_{i\in\{1,\ldots,N\}} D_i^2 \end{bmatrix} = \begin{bmatrix} N & N_1 \\ N_1 & N_1 \end{bmatrix},$$

so we have

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{NN_1 - N_1^2} \begin{bmatrix} N_1 & -N_1 \\ -N_1 & N \end{bmatrix}.$$

Since the conditional covariance matrix of the error is assumed to be

$$\Omega(\mathbf{D}) = \begin{bmatrix} \sigma_1^2(D_1) & 0 & \cdots & 0 \\ 0 & \sigma_2^2(D_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N^2(D_N) \end{bmatrix},$$

we have

$$\begin{aligned} \mathbf{X}'\Omega(\mathbf{D})\mathbf{X} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ D_1 & D_2 & \cdots & D_N \end{bmatrix} \begin{bmatrix} \sigma_1^2(D_1) & 0 & \cdots & 0 \\ 0 & \sigma_2^2(D_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N^2(D_N) \end{bmatrix} \begin{bmatrix} 1 & D_1 \\ 1 & D_2 \\ \vdots & \vdots \\ 1 & D_N \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1^2(D_1) & \sigma_2^2(D_2) & \cdots & \sigma_N^2(D_N) \\ \sigma_1^2(D_1)D_1 & \sigma_2^2(D_2)D_2 & \cdots & \sigma_N^2(D_N)D_N \end{bmatrix} \begin{bmatrix} 1 & D_1 \\ 1 & D_2 \\ \vdots & \vdots \\ 1 & D_N \end{bmatrix} \\ &= \begin{bmatrix} \sum_i \sigma_i^2(D_i) & \sum_i \sigma_i^2(D_i)D_i \\ \sum_i \sigma_i^2(D_i)D_i & \sum_i \sigma_i^2(D_i)D_i^2 \end{bmatrix}. \end{aligned}$$

Using these matrices, we want to compute $\mathbb{V}\left(\hat{\beta}_D|\mathbf{D}\right)$. Notice that $\mathbb{V}\left(\hat{\beta}_D|\mathbf{D}\right)$ is the bottom-right element of the conditional covariance matrix $\mathbb{V}\left(\hat{\beta}|\mathbf{D}\right)$.

To find this element, first we have

$$\begin{aligned} &(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega(\mathbf{D})\mathbf{X} \\ =&\frac{1}{NN_1 - N_1^2} \begin{bmatrix} N_1 & -N_1 \\ -N_1 & N \end{bmatrix} \begin{bmatrix} \sum_i \sigma_i^2(D_i) & \sum_i \sigma_i^2(D_i)D_i \\ \sum_i \sigma_i^2(D_i)D_i & \sum_i \sigma_i^2(D_i)D_i^2 \end{bmatrix} \\ =&\frac{1}{NN_1 - N_1^2} \begin{bmatrix} N_1\left(\sum_i \sigma_i^2(D_i) - \sum_i \sigma_i^2(D_i)D_i\right) & N_1\left(\sum_i \sigma_i^2(D_i)D_i - \sum_i \sigma_i^2(D_i)D_i^2\right) \\ -N_1\sum_i \sigma_i^2(D_i) + N\sum_i \sigma_i^2(D_i)D_i & -N_1\sum_i \sigma_i^2(D_i)D_i + N\sum_i \sigma_i^2(D_i)D_i^2 \end{bmatrix}. \end{aligned}$$

Thus, the bottom-right element of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega(\mathbf{D})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ is given by

$$
\begin{aligned}
&\mathbb{V}\left(\hat{\beta}_D|\mathbf{D}\right) \\
=&\frac{1}{NN_1-N_1^2}\begin{bmatrix}-N_1\sum_i\sigma_i^2(D_i)+N\sum_i\sigma_i^2(D_i)D_i & -N_1\sum_i\sigma_i^2(D_i)D_i+N\sum_i\sigma_i^2(D_i)D_i^2\end{bmatrix} \\
&\qquad\times\frac{1}{NN_1-N_1^2}\begin{bmatrix}-N_1\\N\end{bmatrix} \\
=&\frac{1}{(NN_1-N_1^2)^2}\left(\begin{array}{ll} N_1^2\underbrace{\sum_i\sigma_i^2(D_i)}_{=\sum\sigma_i^2(1)D_i+\sum\sigma_i^2(0)(1-D_i)} & -NN_1\underbrace{\sum_i\sigma_i^2(D_i)D_i}_{=\sum\sigma_i^2(1)D_i} \\ -NN_1\underbrace{\sum_i\sigma_i^2(D_i)D_i}_{=\sum\sigma_i^2(1)D_i} & +N^2\underbrace{\sum_i\sigma_i^2(D_i)D_i^2}_{=\sum\sigma_i^2(D_i)D_i=\sum\sigma_i^2(1)D_i}\end{array}\right) \\
=&\frac{1}{N_1^2\left(N-N_1\right)^2}\left(\underbrace{\left(N_1^2-2NN_1+N^2\right)}_{=(N-N_1)^2}\sum_i\sigma_i^2(1)D_i+N_1^2\sum_i\sigma_i^2(0)(1-D_i)\right) \\
=&\frac{\sum_i\sigma_i^2(1)D_i}{N_1^2}+\frac{\sum_i\sigma_i^2(0)(1-D_i)}{(N-N_1)^2} \\
=&\frac{\sum_i\sigma_i^2(1)D_i/N_1}{N_1}+\frac{\sum_i\sigma_i^2(0)(1-D_i)/(N-N_1)}{(N-N_1)}.
\end{aligned}
$$

c.

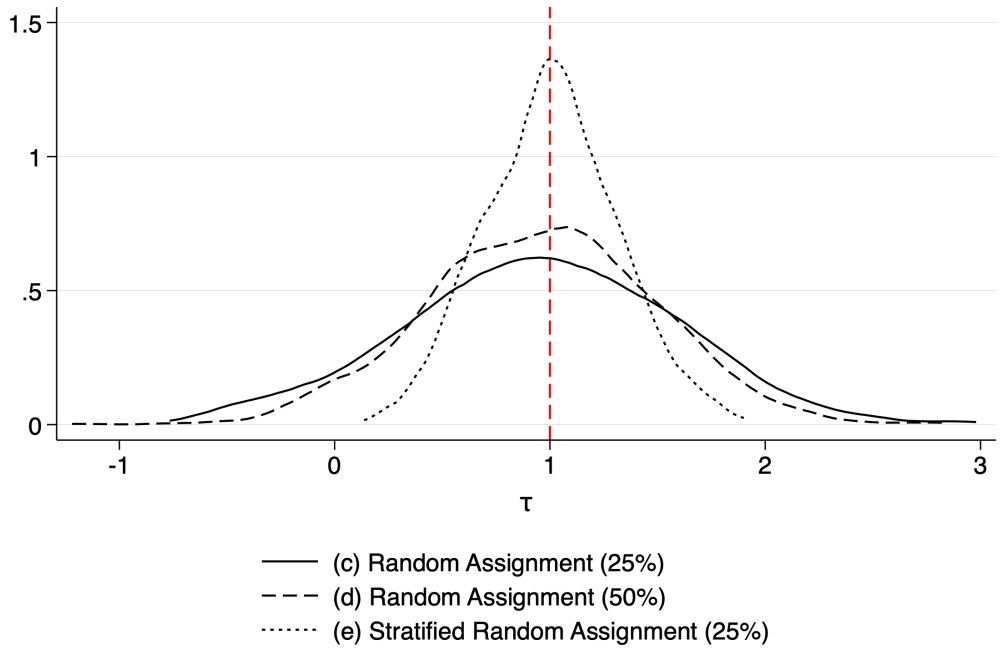
Suppose that for all i, $\sigma_i^2(D_i)=\sigma^2(D_i)$. We want to show that $\mathbb{E}_{\mathbf{D}\in\mathcal{D}}\left[\mathbb{V}\left(\hat{\beta}_D|\mathbf{D}\right)\right]=\frac{\sigma^2(1)}{N_1}+\frac{\sigma^2(0)}{N-N_1}$. From the expression we have obtained in Question 1.b. above, we have

$$
\begin{aligned}
\mathbb{V}\left(\hat{\beta}_D|\mathbf{D}\right)&=\frac{\sum_i\sigma_i^2(1)D_i/N_1}{N_1}+\frac{\sum_i\sigma_i^2(0)(1-D_i)/(N-N_1)}{(N-N_1)} \\
&=\frac{\frac{\sigma^2(1)}{N_1}\sum_i D_i}{N_1}+\frac{\frac{\sigma^2(0)}{N-N_1}\sum_i(1-D_i)}{(N-N_1)}.
\end{aligned}
$$

Since $\mathbb{E}_{\mathbf{D}\in\mathcal{D}}[D_i]=\Pr(D_i=1)$ is the unit assignment probability and we have derived this $=N_1/N$ in Question 1.a., it follows that

$$
\begin{aligned}
\mathbb{E}_{\mathbf{D}\in\mathcal{D}}\left[\mathbb{V}\left(\hat{\beta}_D|\mathbf{D}\right)\right]&=\mathbb{E}_{\mathbf{D}\in\mathcal{D}}\left[\frac{\frac{\sigma^2(1)}{N_1}\sum_i D_i}{N_1}+\frac{\frac{\sigma^2(0)}{N-N_1}\sum_i(1-D_i)}{(N-N_1)}\right] \\
&=\frac{\frac{\sigma^2(1)}{N_1}\sum_i\overbrace{\mathbb{E}_{\mathbf{D}\in\mathcal{D}}[D_i]}^{=N_1/N}}{N_1}+\frac{\frac{\sigma^2(0)}{N-N_1}\sum_i\overbrace{\mathbb{E}_{\mathbf{D}\in\mathcal{D}}[(1-D_i)]}^{=(N-N_1)/N}}{(N-N_1)} \\
&=\frac{\sigma^2(1)}{N_1}+\frac{\sigma^2(0)}{N-N_1}.
\end{aligned}
$$

Figure 1: Q2.f. Distributions of the Coefficient Estimates in (c), (d), and (e)



Question 2

a

Since $\tau_i \sim N(1, 1)$,

$$\mathbb{E}\left[Y_i(1, X_i) - Y_i(0, X_i)\right] = \mathbb{E}\left[(\tau_i + 5 \times X_i + \varepsilon_i) - (5 \times X_i + \varepsilon_i)\right] = \mathbb{E}\left[\tau_i\right] = 1.$$

b

See the do-file for the Stata implementation. In my sample, the average treatment effect is 0.984.

c–e

See the do-file for the Stata implementation.

f

See the do-file for the Stata implementation. Figure 1 presents the distribution of the coefficient estimates each from the simulation in (c), (d), (e). The estimates exhibit the smallest variability under the stratification in (e), followed by (d) in which 50% of observations are treated.

g

See the do-file for the Stata implementation. Table 1 presents the key summary statistics of the simulation results: the standard deviation of the coefficient estimates, the average standard error, and the fraction of estimates that have a p-value < 0.05, each in (c), (d), and (e).

Table 1: Q2.g. Summary Statistics of the Simulation Results

	Question		
	(c)	(d)	(e)
Standard deviation			
Coefficient Estimates	0.631	0.531	0.307
Mean			
SE	0.651	0.569	0.663
% p-value < 0.05	0.052	0.033	0.000

Comparing the simple random assignments in (c) and in (d), the standard deviation of the estimates is smaller in (d) than in (c). This is related to the optimal fraction of treated observations given the variance of the potential outcomes (See Lecture 4 slides), i.e., the variance of the difference-in-means estimator is minimised if we have more observations for either of treatment or control group whose corresponding potential outcome is more noisy (has larger variance). Formally, you can derive the minimiser of $Var\left(\hat{\beta}\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}$, where $n_0 = n - n_1$, as

$$n_1^* = \frac{n}{1 + \sigma_0/\sigma_1}, \quad n_0^* = \frac{n}{1 + \sigma_1/\sigma_0},$$

with $n_1^* = n_0^* = n/2$ when $\sigma_1^2 = \sigma_0^2$. In our case, $Y_i(1, X_i)$ has larger variance than $Y_i(0, X_i)$, thus allocating 50% of the observations gives the estimates with smaller variability than under 25% allocation. Note that the average standard errors in (c) and (d) are more or less similar to the standard deviation of the estimates, and the reduced variability of the estimates can be observed here as well.

Then, when comparing these two assignments to the stratified random assignment in (e), it becomes evident that the estimates in (e) exhibit substantially smaller variability. The intuition is similar to the example in Lecture 4: by stratifying based on X_i, we effectively exclude potential assignments where the distribution of X_i is unbalanced, which would lead to estimates largely deviating from the true value. However, the average standard error is as large in (e) as in (c), which is due to the fact that the regression of Y_i on D_i uses the same information in both (c) and (e), with $n_1/n = 0.25$, and does not take this stratification into account when computing the standard error.

The discrepancy between the standard deviation of the estimates and the mean of the standard errors in (e) indicates that the standard errors are excessively large. This is reflected in the fractions of the estimates with p-values smaller than 0.05. The fraction should be around 5%, as seen in (c) and (d), but none of the estimates in (e) come with a p-value smaller than 0.05.

Question 3

a

See the do-file for the Stata implementation. In my simulated sample,

- the true average treatment effect of D_i, $\frac{1}{N}\sum_i [Y_i(1, X_i(1)) - Y_i(0, X_i(0))] = -2.47$,
- the sample average of τ_i, $\frac{1}{N}\sum_i \tau_i = 0.99$.

The difference between these two numbers stems from the fact that now X_i depends on D_i. Hence, there are two channels that D_i would affect the outcome, directly and indirectly through X_i.

Given the DGP, the population average treatment effect in the population would be

$$\begin{aligned}\mathbb{E}\left[Y_i(1, X_i(1)) - Y_i(0, X_i(0))\right] &= \underbrace{\mathbb{E}\left[\tau_i\right]}_{=1} + 5 \times (\underbrace{\mathbb{E}[\mathbb{1}(\epsilon_i > 1)]}_{1-\Phi(1)=\Phi(-1)} - \underbrace{\mathbb{E}[\mathbb{1}(\epsilon_i > -1)]}_{1-\Phi(-1)=\Phi(1)}) \\ &\approx -2.413,\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the stadandard normal. This calculation is done in the do-file.

Note that the individual treatment effect can be decomposed in the following way:

$$\begin{aligned}Y_i(1, X_i(1)) - Y_i(0, X_i(0)) &= [Y_i(1, X_i(d)) - Y_i(0, X_i(d))] + [Y_i(1-d, X_i(1)) - Y_i(1-d, X_i(0))] \\ &\equiv \underbrace{\xi(d)}_{=\tau_i} + \underbrace{\delta(1-d)}_{=5\times(X_i(1)-X_i(0))}.\end{aligned}$$

where $\xi(d)$ is called the (natural) direct effect while $\delta(1-d)$ is the (natural) indirect (or causal mediation) effect, for each $d \in \{0, 1\}$, in the literature of the mediation analysis. Intuitively, $\xi(d)$ is the treatment effect of D_i while fixing other variables that are also dependent on D_i at some treatment status $d \in \{0, 1\}$. Similarly, $\delta(1-d)$ is the treatment effect of D_i through the other variables that are affected by D_i, while shutting down the *direct* effect channel.

b

See the do-file for the Stata implementation. The average coefficent estimate on D_i across all 1000 assignments is reported in the first column of Table 2.

Table 2: Q3.b.-c. Average Coefficent Estimates on D_i

	Average estimate of coefficient on D_i
(b) Full sample	-2.466
(c) Subgroup $\{i : X_i = 1\}$	2.152
(c) Subgroup $\{i : X_i = 0\}$	2.206

c

See the do-file for the Stata implementation. The average coefficent estimate on D_i across all 1000 assignments for each $X_i \in \{0, 1\}$ is reported in the second and third column of Table 2.

d

With the full sample in (b), the average estimate is close to the true average treatment effect shown in part (a) above. This demonstrates the unbiasedness of the difference-in-means estimator for the sample average treatment effect computed in part (a).

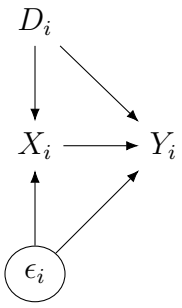
On the other hand, the average estimates from the two subgroups, defined by $X_i \in \{0, 1\}$, substantially differ from the sample average treatment effect or the average τ_i. In a sense, these are highly biased estimates. Why is it the case? Intuitively, this is because conditioning X_i induces the imbalance in the error ϵ_i between the treatment and control groups. Given that the value of X_i is determined by ϵ_i under the different rule in each group of D_i, conditioning X_i means we compare the treated and control observations that are selected differently. Approximating each subgroup average difference in Y_i, we have

$$\begin{aligned}
\mathbb{E}[Y_i|D_i = 1, X_i = 1] - \mathbb{E}[Y_i|D_i = 0, X_i = 1] &= \mathbb{E}[\tau_i + 5 + \epsilon_i|\epsilon_i > 1] - \mathbb{E}[5 + \epsilon_i|\epsilon_i > -1] \\
&= \mathbb{E}[\tau_i] + \mathbb{E}[\epsilon_i|\epsilon_i > 1] - \mathbb{E}[\epsilon_i|\epsilon_i > -1] \\
&= 1 + \underbrace{\frac{\phi(1)}{1 - \Phi(1)} - \frac{\phi(-1)}{1 - \Phi(-1)}}_{\approx 1.238} \\
&\approx 2.238,
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[Y_i|D_i = 1, X_i = 0] - \mathbb{E}[Y_i|D_i = 0, X_i = 0] &= \mathbb{E}[\tau_i + 5 + \epsilon_i|\epsilon_i \leq 1] - \mathbb{E}[5 + \epsilon_i|\epsilon_i \leq -1] \\
&= \mathbb{E}[\tau_i] + \mathbb{E}[\epsilon_i|\epsilon_i \leq 1] - \mathbb{E}[\epsilon_i|\epsilon_i \leq -1] \\
&= 1 \underbrace{- \frac{\phi(1)}{\Phi(1)} + \frac{\phi(-1)}{\Phi(-1)}}_{\approx 1.238} \\
&\approx 2.238,
\end{aligned}$$

where $\phi(\cdot)$ represents the probability density function of the standard normal. The terms after 1 in each average difference describe the average difference in ϵ_i between the treatment and control groups. For instance, in the subgroup with $X_i = 1$, the observations with $D_i = 1$ has on average larger ϵ_i than those with $D_i = 0$, by the definition of how X_i is generated. Hence, by conditioning on $X_i = 1$, we introduce a positive correlation between D_i and ϵ_i, both of which affect Y_i, inflating the average treatment effect estimate. In other words, conditioning on the variable $X_i(D_i)$, we suffer from the same problem as what we typically refer to the omitted variable bias or selection bias.

Figure 2: Q3. DAG



Practical Implications This result has at least three important implications. First, we should not control for variables affected by D_i. These are part of what we call *Bad Controls*.[1] Even when the treatment assignment is randomised so that it is independent of the unobserved factors, controlling for such variables could introduce bias in the estimation, because X_i can be endogenous and so correlated with the unobservables. This is illustrated in Figure 2. Additionally, by controlling for the channel through which D_i affects Y_i, the estimated effect may not capture all we are interested in.

Second, while the decomposition of the individual average treatment effect in part (a) above may imply that we could somehow obtain the direct effect estimate by fixing other (mediator) variables, this is generally not the case. The reason is, as we have seen here, fixing such variables may also change the distributions of the error term ϵ_i in the groups we compare. In a sense, conditioning on such variables changes the composition of the observations in each group. Thus, the conditional comparison generally does not capture "part of the treatment effect not explained by the variables we condition on," and it may not represent any meaningful quantity. To decompose the treatment effects into direct and indirect components, we need additional assumptions.

Lastly, the result in (c) implies that X_i does not necessarily have to be an independent variable in the regression explicitly. As demonstrated in Table 2, X_i can be a variable or factor that defines the sample we analyze. Therefore, even if we do not include X_i in the regression, the same issue arises if the sample at hand is generated similarly (thus effectively looking at the sub-sample defined by X_i).

One example can be such that D_i is a new medicine, Y_i mortality, and X_i is hospitalisation. Even if the new medicine reduces the mortality (just like in our simulation), you might get an estimate with the opposite sign if we condition on whether one is hospitalised or not. A group of people who get the new (effective) medicine but still hospitalised could be systematically different from those who didn't get treatment and have become hospitalised, in a way that the former might experience a higher mortality rate due to some latent characteristics. Then, our "conditional" estimate could be positive—the new medicine seemingly increasing the mortality rate when it does the opposite in fact.

[1] See Chapter 3 of Angrist and Pischke (2008) *Mostly Harmless Econometrics*. The bias we observe here is sometimes referred to as *post-treatment bias*. Note that other categories of variables also fall under *bad controls*, i.e., you should not control for them in the regression. In some cases, even variables determined before treatment assignment can be *bad controls*.

Question 4

Let D be the variable indicating the treatment status, with $\Pr(D = 1) \in (0, 1)$, $\mathbf{X}$ the vector of covariates, and $e(\mathbf{x}) = \Pr(D = 1|\mathbf{X} = \mathbf{x})$. We will show that the distribution of covariates is balanced across treatment and control groups if and only if propensity score is constant.

($\Rightarrow$) Suppose that the distribution of covariates is balanced across the two groups, that is, $D \perp\!\!\!\perp \mathbf{X}$. Then, by this independence, we have $e(\mathbf{x}) = \Pr(D = 1|\mathbf{X} = \mathbf{x}) = \Pr(D = 1)\ \forall x \in \text{Supp}(\mathbf{X})$. This shows that the propensity score is constant.

($\Leftarrow$) Suppose that the propensity score is constant, i.e., there exists $e \in (0, 1)$ such that $\forall \mathbf{x} \in \text{Supp}(\mathbf{X}),\ e(\mathbf{x}) = e$. Then, observe that

$$\Pr(D = 1|\mathbf{X}) = e(\mathbf{X}) = e = \mathbb{E}[e] = \mathbb{E}[e(\mathbf{X})] = \mathbb{E}[\Pr(D = 1|\mathbf{X})] = \mathbb{E}[\mathbb{E}[D|\mathbf{X}]] = \mathbb{E}[D] = \Pr(D = 1),$$
$$\Pr(D = 0|\mathbf{X}) = 1 - \Pr(D = 1|\mathbf{X}) = 1 - \Pr(D = 1) = \Pr(D = 0),$$

implying $D \perp\!\!\!\perp \mathbf{X}$. Thus, the distribution of covariates is balanced across the two groups.

When X is continuous Let $e \equiv \Pr(D = 1)$. Using the notation from Lecture 5, the balanced covariate distribution is equivalent to $f_{\mathbf{X}|D=1}(\mathbf{x}) = f_{\mathbf{X}|D=0}(\mathbf{x})\ \forall \mathbf{x} \in \text{Supp}(\mathbf{X})$. For any $d \in \{0, 1\}$ and $\mathbf{x} \in \text{Supp}(\mathbf{X})$, we have

$$f_{\mathbf{X}=\mathbf{x}|D=d}(\mathbf{x}) = \frac{f_{\mathbf{X},D}(\mathbf{X} = \mathbf{x}, D = d)}{\Pr(D = d)} = \frac{\Pr(D = d|\mathbf{X} = \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{\Pr(D = d)}.$$

Then, for any $\mathbf{x} \in \text{Supp}(\mathbf{X})$,

$$\begin{aligned}
& f_{\mathbf{X}|D=1}(\mathbf{x}) = f_{\mathbf{X}|D=0}(\mathbf{x}) \\
\Leftrightarrow\ & \frac{\Pr(D = 1|\mathbf{X} = \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{\Pr(D = 1)} = \frac{\Pr(D = 0|\mathbf{X} = \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{\Pr(D = 0)} \\
\Leftrightarrow\ & \frac{e(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{\Pr(D = 1)} = \frac{1 - e(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{\Pr(D = 0)} \\
\Leftrightarrow\ & \frac{e(\mathbf{x})}{e} = \frac{1 - e(\mathbf{x})}{1 - e} \\
\Leftrightarrow\ & e(\mathbf{x}) = e.
\end{aligned}$$

Table 3: Q5.b.-d. Regression Results

	(b) $X_i = 0$	(b) $X_i = 1$	(c)	(d)
D_i	1.0006	1.0267	1.0087	2.4975
	(0.0316)	(0.0502)	(0.0270)	(0.0545)
X_i			5.0160	
			(0.0269)	
Constant	-0.0080	4.9899	-0.0120	1.0921
	(0.0221)	(0.0444)	(0.0209)	(0.0424)
Observations	6000	4000	10000	10000

Standard errors in parentheses.

Question 5

a

See the do-file for the Stata implementation.

b

See the do-file for the Stata implementation. The first two columns in Table 3 corresponds to the coefficent estimates on D_i for the sub-samples with $X_i = 0$ and $X_i = 1$.

c

See the do-file for the Stata implementation. The third column in Table 3 reports the estimated coefficents on D_i and X_i using the full sample.

d

See the do-file for the Stata implementation. The last column in Table 3 presents the coefficent estimate on D_i for the full sample.

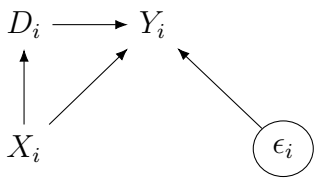
e

In this question, the treatment status D_i depends on X_i, which is different from Question 2 where D_i is generated independently and X_i depends on D_i. Because X_i does not depend on D_i here, from Question 2, the (population) average treatment effect is $\mathbb{E}[\tau_i] = 1$, and the sample average treatment effect should be similar.[2]

Since X_i also affect Y_i, X_i is now a confounder, meaning that we need to somehow adjust for X_i in order to identify the effect of D_i on Y_i. Given each value of X_i, the assignment mechanism is a completely randomised experiment, indicating that the $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i \mid X_i$, the conditional independence. Question 5 demonstrates this point in a simplified setting.

[2]In my sample, it is about 0.99.

Figure 3: Q5. DAG



In the first two columns of Table 3, X_i is *controlled for* by restricting the sample to the sub-group defined by each value of X_i. This way, we create two separate sub-samples, within which we have a completely randomized experiment, allowing the average treatment effect of D_i on Y_i to be identified. Hence, the simple difference-in-means estimator is unbiased. Indeed, the results in Table 3 show that we successfully obtain estimates in this simulation that are close to the true average treatment effect.

In the third column of Table 3, X_i is included in the regression as an additional independent variable. Since X_i is binary, this regression model is fully saturated in X_i. Then, as in Lecture 5, the regression coefficient becomes a weighted average of the conditional average treatment effect for each value of X_i, where the weights depend on the variance of D_i given a value of X_i. Because in our case the treatment effect itself does not depend on X_i, the coefficient estimate makes little difference.

Lastly, the fourth column of Table 3 shows the coefficient estimate substantially different from the true average treatment effect. This is due to the fact that X_i becomes the source of omitted variable bias (OVB). To see this, consider the population regression model:

$$
\begin{aligned}
\tau^{Q4.d.} &= \frac{Cov(Y_i, D_i)}{Var(D_i)} \\
&= \frac{Cov(\tau_i \times D_i + 5 \times X_i + \epsilon_i, D_i)}{Var(D_i)} \\
&= \underbrace{\frac{Cov(\tau_i \times D_i, D_i)}{Var(D_i)}}_{=\mathbb{E}[\tau_i|D_i=1]\times\frac{Var(D_i)}{Var(D_i)}} + \frac{5 \quad \overbrace{Cov(X_i, D_i)}^{(\mathbb{E}[D_i|X_i=1]-\mathbb{E}[D_i|X_i=0])\,Var(X_i)}}{\underbrace{Var(D_i)}_{=\Pr(D_i=1)\Pr(D_i=0)}} \\
&= \mathbb{E}[\tau_i] + 5 \times (\Pr[D_i = 1|X_i = 1] - \Pr[D_i = 1|X_i = 0]) \times \frac{\Pr(X_i = 1)\Pr(X_i = 0)}{\Pr(D_i = 1)\Pr(D_i = 0)} \\
&= 1 + 5 \cdot 0.3 \cdot \frac{0.4 \cdot 0.6}{\underbrace{(0.8 \cdot 0.4 + 0.5 \cdot 0.6)}_{=0.62}\underbrace{(1 - (0.8 \cdot 0.4 + 0.5 \cdot 0.6))}_{=1-0.62=0.38}} \\
&\approx 2.528,
\end{aligned}
$$

which is close to the estimate in the last column of Table 3. The second term corresponds to the bias, which is a typical OVB formula: the structural parameter on X_i multiplied by the coefficient from the regression of X_i on D_i. There is a positive association between X_i and D_i, and X_i affects Y_i positively, which are translated into upward bias of the OLS.