ECMT 676: Econometrics II

Homework III - Due date 03/04/2025

Sebastian Sanin-Restrepo*

Exercise 1

Consider a randomized control trial framework where we assume $(Y(1), Y(0)) \perp D$. Show that the ATE is identified and conditions under which your proposed estimator consistent and asymptotically normal. Under these assumptions what is the limiting distribution of your estimator? Be explicit about the formula for the asymptotic variance. You may assume that $\Pr(D = 1)$ is known and does not need to be estimated. (Hint: you will need to use the continuous mapping theorem. As a first step, it is useful to recall the central limit theorem for a vector of sample means.)

In a randomized control trial (RCT), the key assumption is that treatment assignment is independent of the potential outcomes, so $(Y(1), Y(0)) \perp D$. This independence is useful because we will have,

$$\mathbb{E}[Y|D = 1] = \mathbb{E}[Y(1)] \quad \text{and} \quad \mathbb{E}[Y|D = 0] = \mathbb{E}[Y(0)]$$

Thus, the average treatment effect ATE is identified as,

$$ATE = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y|D = 1] - \mathbb{E}[Y|D = 0]$$

A feasible estimator for the ATE is the difference of the means of the treated and non-treated groups:

$$\widehat{ATE} = \bar{Y}_1 - \bar{Y}_0 = \frac{1}{n_1} \sum_{i:D_i=1}^{n} {}_1Y_i - \frac{1}{n_0} \sum_{i:D_i=0}^{n} {}_0Y_i$$

where, n_1 and n_0 refer to the number of individuals in the treated and non-treated group respectively. In an RCT the probability of treatment is known and given by $p = \mathbb{P}(D = 1)$ and $1 - p = \mathbb{P}(D = 0)$, with $p \in (0, 1)$ ensuring overlap.

Under the assumption that the data $\{Y_i, D_i\}_{i=1}^n$ are *iid*, with finite moments $\mathbb{E}[Y_i^2] < \infty$, the Law of Large Numbers ensures that $\bar{Y}_1$ and $\bar{Y}_0$ converge in probability to $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$. Therefore, the estimator is consistent.

*Economics, Ph.D. Student. E-mail: sr.sebastian@tamu.edu

We can also apply the Central Limit Theorem for the two groups and obtain

$$\sqrt{n_1}\left(\bar{Y}_1 - \mathbb{E}[Y(1)]\right) \to_d N(0, \text{Var}(Y(1)))$$

$$\sqrt{n_0}\left(\bar{Y}_0 - \mathbb{E}[Y(0)]\right) \to_d N(0, \text{Var}(Y(0)))$$

Since, the overal sample is n, we have that approximately $n_1 \approx np$ and $n_2 \approx n(1-p)$. Then, the asymptotic variance of the sample mean for the treated group is given by,

$$\text{Var}(\bar{Y}_1) = \frac{\text{Var}(Y(1))}{np}$$

and similarly,

$$\text{Var}(\bar{Y}_0) = \frac{\text{Var}(Y(0))}{n(1-p)}$$

Since $\widehat{ATE} = \bar{Y}_1 - \bar{Y}_0$ and the two samples are independent, by the properties of variance we have that,

$$\text{Var}(\widehat{ATE}) = \frac{\text{Var}(Y(1))}{np} + \frac{\text{Var}(Y(0))}{n(1-p)}$$

Thus, by the Central Limit Theorem, and the Continuous Mapping Theorem (in which the continuous function is the sum of the two samples), the estimator is asymptotically normal as follows,

$$\sqrt{n}\left(\widehat{ATE} - ATE\right) \to_d N\left(0, \frac{\text{Var}(Y(1))}{np} + \frac{\text{Var}(Y(0))}{n(1-p)}\right)$$

Exercise 2

Consider the parameter $\Pr(Y(1) - Y(0) \geq 0)$. Is this parameter identified under our assumption $(Y(1), Y(0)) \perp D$? Why or why not? If it is identified propose an estimator. If it is not identified propose a research design under which you could identify the parameter.

Under the standard randomization assumption in which $(Y(1), Y(0)) \perp D$, we can identify the marginal distributions of $Y(1)$ and $Y(0)$ but not the joint distribution of $(Y(1), Y(0))$. Since the parameter $\Pr(Y(1) - Y(0) \geq 0)$ depends on the joint distribution, it is **not identified** without further assumptions. We cannot infer how $Y(1)$ and $Y(0)$ comove for the same individual because we never observe them jointly (as occurs by the fundamental problem of causal inference).

A research design to identify the parameter that could allow us to observe both potential outcomes for the same unit includes longitudinal designs in which we can observe the same individual with and without treatment so we can infer $Y(1)$ and $Y(0)$.

Exercise 3

A common intuitive way to estimate the treatment effect in this setting is to regress Y on D and X. The coefficient on D is then interpreted as the treatment effect. First, show that this approach can be justified by showing that $\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mathbb{E}[Y \mid D = 1, X] - \mathbb{E}[Y \mid D = 0, X]]$. In the case of a linear regression model, $Y = \gamma D + \beta' X + \epsilon$, what assumption are we implicitly adding to Assumption 2.1 in interpreting γ as the ATE?

Under Assumption 2.1 (Conditional Unconfoundedness) we have that,

$$\mathbb{E}[Y(1)|X] = \mathbb{E}[Y|D = 1, X] \quad \text{and} \quad \mathbb{E}[Y(0)|X] = \mathbb{E}[Y|D = 0, X]$$

Taking the difference and then expectations over X we obtain,

$$\begin{aligned}\mathbb{E}[Y(1) - Y(0)] =& \mathbb{E}[\mathbb{E}[Y(1) - Y(0)|X]] \\ =& \mathbb{E}[\mathbb{E}[Y(1)|X] - \mathbb{E}[Y(0|X]] \\ =& \mathbb{E}[\mathbb{E}[Y|D = 1, X] - \mathbb{E}[Y|D = 0, X]]\end{aligned}$$

Thus, the Average Treatment Effect (**ATE**) can be written as,

$$ATE = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mathbb{E}[Y|D = 1, X] - \mathbb{E}[Y|D = 0, X]]$$

Now, in the linear regression model $Y = \gamma D + \beta' X + \epsilon$, the coefficient γ is interpreted as the ATE if we add linearity as an assumption, that is, we assume that the conditional expectation of Y given D and X is linear in D and X. So,

$$\mathbb{E}[Y|D, X = x] = \gamma D + \beta' X$$

Notice that this implies that the potential outcomes satisfy,

$$\mathbb{E}[Y|D = 1, X = x] = \mathbb{E}[Y(1)|X = x] = \gamma + \beta' X$$

and,

$$\mathbb{E}[Y|D = 0, X = x] = \mathbb{E}[Y(0)|X = x] = \beta' X$$

Furthermore, we also need to add the assumption that the treatment effect is homogeneous across all values of X. Formally, it means that,

$$\mathbb{E}[Y(1) - Y(0)|X = x] = \gamma$$

This implies that the difference between the potential outcomes does not depend on x, so that any heterogeneity in the effect of treatment is ruled out.

Exercise 4

Assumption 2.1 allows us to identify more than just the ATE. In particular, we can identify the entire marginal CDF of $Y(1)$ (as well as the marginal CDF of $Y(0)$). By modifying the argument above, show that $\Pr(Y(1) \leq \tau)$ is identified for every $\tau \in \mathbb{R}$.

To identify the marginal CDF of $Y(1)$ at a point τ, we use Assumption 2.1. We have that,

$$\mathbb{P}(Y(1) \leq \tau | X = x) = \mathbb{P}(Y \leq \tau | D = 1, X = x)$$

By the Law of Iterated Expectations,

$$\mathbb{P}(Y(1) \leq \tau) = \mathbb{E}_X \left[\mathbb{P}(Y(1) \leq \tau | X)\right] = \mathbb{E}_X \left[\mathbb{P}(Y \leq \tau | D = 1, X)\right]$$

To express this in terms of observed variables using the inverse probability weightings we do

$$\mathbb{E}\left[\frac{1\{Y \leq \tau\} D}{\pi(X)}\right], \qquad \pi(X) = \mathbb{P}(D = 1 | X)$$

Conditioning on X yields,

$$\begin{aligned}
\mathbb{E}\left[\frac{1\{Y \leq \tau\} D}{\pi(X)} \middle| X\right] =& \frac{1}{\pi(X)} \mathbb{E}\left[1\{Y \leq \tau\} D \middle| X\right] \\
=& \frac{1}{\pi(X)} \mathbb{P}\left(Y \leq \tau, D = 1 \middle| X\right) \\
=& \frac{1}{\pi(X)} \mathbb{P}\left(Y \leq \tau | D = 1, X\right) \mathbb{P}\left(D = 1 | X\right) \\
=& \frac{1}{\pi(X)} \mathbb{P}\left(Y \leq \tau | D = 1, X\right) \pi(X) \\
=& \mathbb{P}\left(Y \leq \tau | D = 1, X\right)
\end{aligned}$$

Now, taking the expectation over X gives us,

$$\mathbb{E}\left[\frac{1\{Y \leq \tau\} D}{\pi(X)}\right] = \mathbb{E}_X[\mathbb{P}\left(Y \leq \tau | D = 1, X\right)] = \mathbb{P}\left(Y \leq \tau | D = 1\right) = \mathbb{P}\left(Y(1) \leq \tau\right)$$

Thus, for every $\tau \in \mathbb{R}$, the marginal CDF of $Y(1)$ is identified by the observable expression,

$$\mathbb{P}(Y(1) \leq \tau) = \mathbb{E}\left[\frac{1\{Y \leq \tau\} D}{\pi(X)}\right]$$

To estimate this we can use the sample analogs,

$$\hat{F}_{Y(1)}(\tau) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1\{Y_i \leq \tau\} D_i}{\hat{\pi}(X_i)}\right]$$

where $\hat{\pi}(X_i)$ is an estimator of the propensity score that could be addressed by using

logistic, probit or linear probability model.

Exercise 5

Where was Assumption 2.1(ii) used in the identification argument above?

Assumption 2.1(ii), the **strong overlap** condition, ensures that for every x in the support of X, the propensity score satisfies $\eta \leq \pi(X) \leq 1 - \eta$ for some $\eta > 0$. This is used in several steps of the argument. First, it is necessary to define the inverse probability weights. Notice that the estimator is computed using $\frac{1}{\pi(X)}$ and $\frac{1}{1-\pi(X)}$. For these to be well-defined, we need the denominator to be nonzero. It is zero if $\pi(X)$ approaches to zero or one. So $\pi(X)$ must be bounded away from zero and one (Assumption 2.1).

Later in the argument in the step,

$$\mathbb{E}\left[\frac{YD}{\pi(X)} - \frac{Y(1-D)}{1-\pi(X)}\right] = \mathbb{E}\left[\frac{1}{\pi(X)}\mathbb{E}[Y(1) \mid X]\pi(X) - \frac{1}{1-\pi(X)}\mathbb{E}[Y(0) \mid X](1-\pi(X))\right]$$

the multiplication by $\pi(X)$ or $1-\pi(X)$ cancels the inverses, but this cancellation is valid only if those estimators are nonzero. Again, Assumption 2.1 is relevant for this step.

Exercise 6

Describe, step-by-step, how you would use the above results to test $H_0 : \text{ATE} = 0$ against $H_1 : \text{ATE} \neq 0$.

To test the hypothesis $H_0 : \text{ATE} = 0$ against $H_1 : \text{ATE} \neq 0$ we can proceed in the following steps:

1. Estimate the propensity score: Using a proper model (i.e., logistic regression, probit, linear probability or nonparametric methods) estimate $\pi(X) = \mathbb{P}(D = 1|X)$. Obtain $\hat{\pi}(X_i)$.

2. Use the inverse probability weighting (IPW) estimator for the **ATE** as we did before. Formally,
$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i D_i}{\hat{\pi}(X_i)} - \frac{Y_i(1-D_i)}{1-\hat{\pi}(X_i)}\right]$$

3. Estimate the Asymptotic variance by
$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i D_i}{\hat{\pi}(X_i)} - \frac{Y_i(1-D_i)}{1-\hat{\pi}(X_i)} - \widehat{ATE}_i\right)^2$$
and compute standard errors as $\widehat{SE} = \frac{\hat{V}}{n}$

4. Build the t-statistic under the null hypothesis $H_0 : ATE = 0$, as

$$t_{ATE} = \frac{\widehat{ATE}}{\widehat{SE}}$$

From the results above, we know that:

$$\sqrt{n}(\widehat{ATE} - ATE) \to_d N(0, V)$$

Therefore, under H_0 the t-statistci follows a $N(0,1)$ distribution. For a two sided-test, the rejection region is given by $|t_{ATE}| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. So, we can compare the computed $|t_{ATE}|$ statistic, to 1.96 (for $\alpha = 0.05$). If $|t_{ATE}| > 1.96$ then we reject H_0.

Exercise 7

Show that $\mathbb{E}\left[g\left(Y_i(1)\right) \mid \text{Complier }\right]$ and $\mathbb{E}\left[g\left(Y_i(0)\right) \mid \text{Complier }\right]$ are identified for any function $g : \mathbb{R} \to \mathbb{R}$. For what kind of functions can we additionally identify $\mathbb{E}\left[g\left(Y_i(1) - Y_i(0)\right) \mid \text{Complier }\right]$?

Notice that when $Z_i = 1$, both *always takers* (AT) and *compliers*(C) receive treatment. Thus, the observed moment for treated individuals is:

$$\mathbb{E}[g(Y_i(1))D_i|Z_i = 1] = \mathbb{E}[g(Y_i(1))|\text{AT}]\mathbb{P}(\text{AT}) + \mathbb{E}[g(Y_i(1))|\text{C}]\mathbb{P}(\text{C})$$

On the other hand, for $Z_i = 0$ ATs still receive the treatment, while Cs do not:

$$\mathbb{E}[g(Y_i(1))D_i|Z_i = 0] = \mathbb{E}[g(Y_i(1))|\text{AT}]\mathbb{P}(\text{AT})$$

Subtract the two equations to cancel out the AT's contribution,

$$\mathbb{E}[g(Y_i(1))D_i|Z_i = 1] - \mathbb{E}[g(Y_i(1))D_i|Z_i = 0] = \mathbb{E}[g(Y_i(1))|\text{C}]\mathbb{P}(\text{C})$$

For the case of the probability of the treatment given Z we identify the share of compliers as:

$$\mathbb{P}(D_i = 1|Z_i = 1) - \mathbb{P}(D_i = 1|Z_i = 0) = \mathbb{P}(C)$$

Hence,

$$\mathbb{E}[g(Y_i(1))|C] = \frac{\mathbb{E}[g(Y_i(1))D_i|Z_i = 1] - \mathbb{E}[g(Y_i(1))D_i|Z_i = 0]}{\mathbb{P}(D_i = 1|Z_i = 1) - \mathbb{P}(D_i = 1|Z_i = 0)}$$

Notice that $\mathbb{E}[g(Y_i(1))|C]$ is identified since both the numerator and denominator are built from observable quantities. Now, for $\mathbb{E}[g(Y_i(0))|C]$ both never takers (NT) and

compliers do not receive the treatment, so under $Z_i = 0$,

$$\mathbb{E}[g(Y_i(0))(1 - D_i)|Z_i = 0] = \mathbb{E}[g(Y_i(0))|\text{NT}]\mathbb{P}(\text{NT}) + \mathbb{E}[g(Y_i(0))|\text{C}]\mathbb{P}(\text{C})$$

On the other hand, for $Z_i = 1$ NTs still do not receive the treatment, while Cs do:

$$\mathbb{E}[g(Y_i(0))(1 - D_i)|Z_i = 1] = \mathbb{E}[g(Y_i(0))|\text{NT}]\mathbb{P}(\text{NT})$$

Subtract the two equations to cancel out the NT's contribution,

$$\mathbb{E}[g(Y_i(0))(1 - D_i)|Z_i = 0] - \mathbb{E}[g(Y_i(0))(1 - D_i)|Z_i = 1] = \mathbb{E}[g(Y_i(0))|\text{C}]\mathbb{P}(\text{C})$$

and dividing by $\mathbb{P}(D_i = 1|Z_i = 1) - \mathbb{P}(D_i = 1|Z_i = 0)$, we obtain,

$$\mathbb{E}[g(Y_i(0))|C] = \frac{\mathbb{E}[g(Y_i(0))(1 - D_i)|Z_i = 0] - \mathbb{E}[g(Y_i(0))(1 - D_i)|Z_i = 1]}{\mathbb{P}(D_i = 1|Z_i = 1) - \mathbb{P}(D_i = 1|Z_i = 0)}$$

again, everything is a function of observables. Thus, for any function $g : \mathbb{R} \to \mathbb{R}$, we can identify $\mathbb{E}[g(Y_i(0))|C]$ and $\mathbb{E}[g(Y_i(1))|C]$.

Now, identifying the expectation of a function of the treatment effect $\mathbb{E}[g(Y_i(1) - Y_i(0))]$ is more challenging since the joint distribution of $(Y_i(1), Y_i(0))$ among compliers is generally not identified. Knowing about the marginal distributions does not pin down the moments of the difference unless we assume further structure in the function.

- Linear (or affine transformations): If g is linear, $g(x) = a + bx$, then,

  $$\mathbb{E}[a + b(Y_i(1) - Y_i(0))|\text{C}] = a + b\left(\mathbb{E}[Y_i(1)|\text{C}] - \mathbb{E}[Y_i(0)|\text{C}]\right)$$

  since both $\mathbb{E}[Y_i(1)|\text{C}]$ and $\mathbb{E}[Y_i(0)|\text{C}]$ are identified, the expectation of the linear function is identified.

- Additively decompose functions: If g could be written in the form:

  $$g(Y_i(1) - Y_i(0)) = h(Y_i(1)) - k(Y_i(0))$$

  for some functions $h(\cdot)$ and $k(\cdot)$ then,

  $$\mathbb{E}[g(Y_i(1) - Y_i(0))|\text{C}] = (\mathbb{E}[h(Y_i(1))|\text{C}] - \mathbb{E}[k(Y_i(0))|\text{C}])$$

Exercise 8

Show that 2SLS with a binary instrument and a binary treatment identifies the LATE. That is, show that $\beta = \frac{\text{Cov}(Y,Z)}{\text{Cov}(D,Z)}$ is equivalent to the identifying formula for the LATE in (3.1).

To show that the 2SLS estimator with binary instrument and binary treatment identifies the LATE, we start with the estimator,

$$\beta = \frac{Cov(Y,Z)}{Cov(D,Z)}$$

because Z is binary, we can express the covariates as conditional expectations. Denoting $p = \mathbb{P}(Z=1)$. Then, by definition of covariance, we have:

$$\text{Cov}(Y,Z) = \mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z]$$

Since $\mathbb{E}[YZ] = \mathbb{P}(Z=1)\mathbb{E}[Y|Z=1] = p\mathbb{E}[Y|Z=1]$, $\mathbb{E}[Y] = p\mathbb{E}[Y|Z=1] + (1-p)\mathbb{E}[Y|Z=0]$, and $\mathbb{E}[Z] = p$. Thus,

$$\text{Cov}(Y,Z) = p\mathbb{E}[Y|Z=1] - p\left[p\mathbb{E}[Y|Z=1] + (1-p)\mathbb{E}[Y|Z=0]\right]$$

Simplify this expression,

$$\begin{aligned}\text{Cov}(Y,Z) =& p\mathbb{E}[Y|Z=1] - p^2\mathbb{E}[Y|Z=1] - p(1-p)\mathbb{E}[Y|Z=0] \\ =& p(1-p)[\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]]\end{aligned}$$

Similarly, we can write $Cov(D,Z)$ as,

$$\begin{aligned}\text{Cov}(D,Z) =& \mathbb{E}[DZ] - \mathbb{E}[D]\mathbb{E}[Z] \\ =& p(1-p)\left[\mathbb{P}(D=1|Z=1) - \mathbb{P}(D=1|Z=0)\right]\end{aligned}$$

Therefore, we can write the β as:

$$\begin{aligned}\beta =& \frac{\text{Cov}(Y,Z)}{\text{Cov}(D,Z)} = \frac{p(1-p)[\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]]}{p(1-p)\left[\mathbb{P}(D=1|Z=1) - \mathbb{P}(D=1|Z=0)\right]} \\ =& \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{P}(D=1|Z=1) - \mathbb{P}(D=1|Z=0)} = \text{LATE}\end{aligned}$$

Exercise 9

In the context of the linear IV model, we typically assumed that our instrument was "strong" in the sense that $\text{Cov}(Z,D) \geq \delta > 0$. How can we interpret strong identification in the context of the LATE model? In other words what must we assume about the

response types / potential treatments to ensure that $\text{Cov}(Z, D)$ is bounded away from zero?

Recall that in the binary instrument setting under monotonicity we have:

$$\mathbb{P}(D = 1|Z = 1) = \mathbb{P}(AT) + \mathbb{P}(C)$$

and

$$\mathbb{P}(D = 1|Z = 0) = \mathbb{P}(AT)$$

Thus, the difference in treatment probabilities is,

$$\mathbb{P}(D = 1|Z = 1) - \mathbb{P}(D = 1|Z = 0) = \mathbb{P}(C)$$

Since,

$$\begin{aligned}\text{Cov}(Z, D) =& \mathbb{P}(Z = 1)(1 - \mathbb{P}(Z = 1)) \left[\mathbb{P}(D = 1|Z = 1) - \mathbb{P}(D = 1|Z = 0)\right] \\ =& \mathbb{P}(Z = 1)(1 - \mathbb{P}(Z = 1))\mathbb{P}(C)\end{aligned}$$

So, in the LATE context saying that $\text{Cov}(Z, D) \geq \delta > 0$ traduces to, saying that $\mathbb{P}(Z = 1)(1 - \mathbb{P}(Z = 1))\mathbb{P}(C) \geq \delta > 0$, that could be rewritten,

$$\mathbb{P}(C) \geq \frac{\delta}{\mathbb{P}(Z = 1)(1 - \mathbb{P}(Z = 1))} > 0$$

indicating that for Z to be a strong instrument, the proportion of compliers should be bounded away from zero. This means there must be a significant share of individuals for whom the instrument affects the treatment decision, ensuring that the first-stage relationship is strong.