Econ 2120: Section 10

Instrumental Variables

Ashesh Rambachan

Fall 2018

Outline

Instrumental Variable Model

Intro to Weak IV

Potential Outcomes

Conditional Random Assignment

Estimating ATE under Strong Ignorability

Instrumental Variable Model and ATE

IV Bounds on ATE

Outline

Instrumental Variable Model

Intro to Weak IV

Potential Outcomes

Conditional Random Assignment

Estimating ATE under Strong Ignorability

Instrumental Variable Model and ATE

IV Bounds on ATE

Omitted Variables Bias

I'll follow the lecture notes and quickly cover the basic, linear IV model. We've seen this material before so it should be familiar.

Suppose that we are interested in the long regression

$$E[Y_i|FB_i, ED_i, A_i] = FB_i'\phi + ED_i\beta + A_i.$$

A_i is unobserved. Suppose there is an additional variable SUB_i that is observe – this is our **instrument**. It satisfied two exclusion restrictions:

(1): $E[Y_i|FB_i, SUB_i, ED_i, A_i] = FB_i'\phi + ED_i\beta + A_i$.

(2): $E^*[A_i|FB_i, SUB_i] = FB_i'\lambda$.

Linear IV estimator

Define the prediction errors

$$\begin{aligned}\epsilon_i &= A_i - E^*[A_i|FB_i, SUB_i] \\ U_i &= Y_i - E[Y_i|FB_i, SUB_i, ED_i, A_i]\end{aligned}$$

and write

$$\begin{aligned}A_i &= FB_i'\lambda + \epsilon_i, \\ Y_i = FB_i'\phi + ED_i\beta &+ A_i + U_i.\end{aligned}$$

Subbing in for A_i, we get that

$$\begin{aligned}Y_i &= FB_i'(\phi + \lambda) + ED_I\beta + (\epsilon_i + U_i) \\ &= FB_i'\delta + ED_i\beta + V_i\end{aligned}$$

where $\delta = \phi + \lambda,\ V_i = \epsilon_i + U_i$.

Linear IV estimator

FB_i, SUB_i are orthogonal to V_i. This will give us the moments needed to estimate β. Define

$$R_i = (FB_i' \ ED_i), \quad B_i = \begin{pmatrix} FB_i \\ SUB_i \end{pmatrix}, \quad \gamma = \begin{pmatrix} \delta \\ \beta \end{pmatrix}.$$

The exclusion restrictions give us that

$$Y_i = R_i\gamma + V_i, \ E[B_iV_i] = 0.$$

This fits into our Linear GMM framework from earlier and we're off to the races. See the lecture notes for details.

Outline

Instrumental Variable Model

Intro to Weak IV

Potential Outcomes

Conditional Random Assignment

Estimating ATE under Strong Ignorability

Instrumental Variable Model and ATE

IV Bounds on ATE

References

Presentation here follows the Vadim Marmer's (UBC) excellent lecture notes on weak instruments – see his website for the latest version.

Stock & Watson (2008) NBER SI Methods lectures are also a great resource.

Classic IV algebra

We observe $(y_{i1}, y_{i2}, Z_{i1}, Z_{i2})$, where $y_{i,1}$ is the dependent variable, y_{i2} is the single endogenous variable, Z_{i1} is L-dimensional vector of instruments and Z_{i2} is the M-dimensional vector of exogenous regressors.

The IV regression model is

$$
\begin{aligned}
y_{i1} &= \gamma y_{i2} + Z_{i2}'\beta + u_i \\
y_{i2} &= Z_{i1}'\pi_1 + Z_{i2}'\pi_2 + v_i,
\end{aligned}
$$

where $E[Z_{i1}u_i] = E[Z_{i1}v_i] = E[Z_{i2}u_i] = E[Z_{i2}v_i] = 0$. y_{2i} is endogenous if $E[u_i v_i] \neq 0$.

Denote the n-dimensional vectors y_1, y_2, Z_1, Z_2, u, v and, the model becomes

$$
\begin{aligned}
y_1 &= \gamma y_2 + \underset{n\times M}{Z_2}\beta + u \\
y_2 &= \underset{N\times L}{Z_1}\pi_1 + \underset{N\times M}{Z_2}\pi_2 + v.
\end{aligned}
$$

IV estimation

We'll residualize the first-stage and structural equations with respect to the included exogenous variables. Define

$$M_2 = I_n - Z_2(Z_2'Z_2)^{-1}Z_2'.$$

This is the **annihilator matrix** that projects vector onto the orthogonal complement of the space spanned by the columns of Z_2. It is symmetric and idempotent $M_2 = M_2'$, $M_2 M_2 = M_2$. So, the OLS estimator π_1 is

$$\hat{\pi}_1 = (Z_1' M_2 Z_1)^{-1} Z_1' M_2 y_2$$

from residualizing the first-stage.

Residualize the structural equation and then, regress y_1 on $Z_1 M_2 \hat{\pi}_1$, we get that

$$\hat{\gamma} = \frac{(M_2 Z_1 \hat{\pi}_1)' y_1}{(M_2 Z_1 \hat{\pi}_1)'(M_2 Z_1 \hat{\pi}_1)} = \frac{\hat{\pi}_1' Z_1' M_2 y_1}{\hat{\pi}_1' Z_1' M_2 Z_1 \hat{\pi}_1}$$

Asymptotics of IV estimator

Provided that the first-stage coefficients are fixed and different from zero, $\pi_1 \neq 0$, we can apply a WLLN and CLT.

We'll assume that

(1) The data $(y_{i1}, y_{i2}, Z_{i1}, Z_{i2})$ are i.i.d.

(2) $E[\begin{pmatrix} Z_{i1}Z_{i1}' & Z_{i1}Z_{i2}' \\ Z_{i2}Z_{i1}' & Z_{i2}Z_{i2}' \end{pmatrix}] = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}$ is finite and positive definite.

(3) $E[\begin{pmatrix} u_i \\ v_i \end{pmatrix} \begin{pmatrix} u_i \\ v_i \end{pmatrix}' |Z_{i1}, Z_{i2}] = \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix}$ is finite and positive definite.

Consistency of IV estimator

We have that $\hat{\pi}_1 = (Z_1' M_2 Z_1)^{-1} Z_1' M_2 y_2$ and so,

$$\begin{aligned} \hat{\pi}' Z_1' M_2 &= y_2' M_2 Z_1 (Z_1' M_2 Z_1)^{-1} Z_1' M_2 = y_2' P_{M_2 Z_1}, \\ \hat{\pi}' Z_1' M_2 Z_1 \hat{\pi} &= y_2' M_2 Z_1 (Z_1' M_2 Z_1)^{-1} Z_1' M_2 y_2 = y_2' P_{M_2 Z_1} y_2, \end{aligned}$$

where $P_{M_2 Z_1} = M_2 Z_1 (Z_1' M_2 Z_1)^{-1} Z_1' M_2$ is the projection matrix onto the space spanned by the columns of $M_2 Z_1$.

We can substitute this into the expression for $\hat{\gamma}$ and we get that

$$\hat{\gamma} = \frac{y_2' P_{M_2 Z_1} y_1}{y_2' P_{M_2 Z_1} y_2}.$$

Consistency of IV Estimator

We now substitute the structural equation into the numerator, $y_1 = \gamma y_2 + Z_2\beta + u$. We get that

$$
\begin{aligned}
\hat{\gamma} &= \gamma + \frac{y_2' P_{M_2 Z_1} u}{y_2' P_{M_2 Z_1} y_2} \\
&= \gamma + \frac{y_2' M_2 Z_1 (Z_1' M_2 Z_1)^{-1} Z_1' M_2 u}{y_2' M_2 Z_1 (Z_1' M_2 Z_1)^{-1} Z_1' M_2 y_2} \\
&= \gamma + \frac{(Z_1 \pi_1 + v)' M_2 Z_1 (Z_1' M_2 Z_1)^{-1} Z_1' M_2 u}{(Z_1 \pi_1 + v)' M_2 Z_1 (Z_1' M_2 Z_1)^{-1} Z_1' M_2 (Z_1 \pi_1 + v)} \\
&= \gamma + \frac{(Z_1' M_2 Z_1 \pi_1 + Z_1' M_2 v)' (Z_1' M_2 Z_1)^{-1} Z_1' M_2 u}{(Z_1' M_2 Z_1 \pi_1 + Z_1' M_2 v)' (Z_1' M_2 Z_1)^{-1} (Z_1' M_2 Z_1 \pi_1 + Z_1' M_2 v)}
\end{aligned}
$$

Using a LLN, we have that

$$
\frac{Z_1' Z_1}{n} \xrightarrow{p} E[Z_{i1} Z_{i1}'] = Q_{11}, \quad \frac{Z_1' Z_2}{n} \xrightarrow{p} Q_{12}, \quad \frac{Z_2' Z_2}{n} \xrightarrow{p} Q_{22}.
$$

Consistency of IV estimator (cont.)

Using the LLN results and that $M_2 = I_n - Z_2(Z_2'Z_2)^{-1}Z_2'$,

$$\begin{aligned}\frac{Z_1'M_2Z_1}{n} &= \frac{Z_1'Z_1}{n} - \frac{Z_1'Z_2}{n}(\frac{Z_2'Z_2}{n})^{-1}\frac{Z_2'Z_1}{n} \\ &\xrightarrow{p} Q_{11} - Q_{12}^{-1}Q_{22}^{-1}Q_{12}' = Q_{1\cdot 2},\end{aligned}$$

where $Q_{1\cdot 2}$ is positive definite. Moreover, we have that

$$\begin{aligned}\frac{Z_1'u}{n} &\xrightarrow{p} E[Z_{i1}u_i] = 0, \\ \frac{Z_2'u}{n} &\xrightarrow{p} E[Z_{i2}u_i] = 0.\end{aligned}$$

So,

$$\frac{Z_1'M_2u}{n} = \frac{Z_1'u}{n} - \frac{Z_1'Z_2}{n}(\frac{Z_2'Z_2}{n})^{-1}\frac{Z_2'u}{n} \xrightarrow{p} 0 - Q_{12}Q_{22}^{-1}0 = 0.$$

By a similar argument, $\frac{Z_1'M_2v}{n} \xrightarrow{p} 0$.

Consistency of IV estimator (cont.)

Using these results, it is immediate that

$$\hat{\gamma} \xrightarrow{p} \gamma + \frac{(Q_{1\cdot 2}\pi_1 + 0)' Q_{1\cdot 2}^{-1} 0}{(Q_{1\cdot 2}\pi_1 + 0)' Q_{1\cdot 2}^{-1} (Q_{1\cdot 2}\pi_1 + 0)'} = \gamma,$$

which holds because $\pi_1 \neq 0$ by assumption.

Limiting distribution of IV estimator

Begin with

$$\hat{\gamma} = \gamma + \frac{(Z_1' M_2 Z_1 \pi_1 + Z_1' M_2 v)'(Z_1' M_2 Z_1)^{-1} Z_1' M_2 u}{(Z_1' M_2 Z_1 \pi_1 + Z_1' M_2 v)'(Z_1' M_2 Z_1)^{-1}(Z_1' M_2 Z_1 \pi_1 + Z_1' M_2 v)}.$$

And re-write to get that

$$\sqrt{n}(\hat{\gamma} - \gamma) = \frac{n^{-1/2} \pi_1' Z_1' M_2 u + o_p(1)}{n^{-1} \pi_1' Z_1' M_2 Z_1 \pi_1 + o_p(1)}.$$

From the CLT, we have that

$$n^{-1/2} \begin{pmatrix} Z_1' u \\ Z_2' u \end{pmatrix} \xrightarrow{d} N(0, E[u_i^2 \begin{pmatrix} Z_{i1}^2 & Z_{i1} Z_{i2}' \\ Z_{i2} Z_{i1}' & Z_{i2}^2 \end{pmatrix}] = N(0, \sigma_u^2 Q)$$

under the assumption of homoskedasticity.

Limiting Distribution of IV estimator

Introduce notation

$$\begin{pmatrix}\Phi_1\\ \Phi_2\end{pmatrix} \sim N(0, \sigma_u^2 Q).$$

Using the CLT from earlier, we have that

$$\begin{aligned}\frac{Z_1' M_2 u}{\sqrt{n}} &= \frac{Z_1' u}{\sqrt{n}} - \frac{Z_1' Z_2}{n}(\frac{Z_2' Z_2}{n})^{-1}\frac{Z_2' u}{\sqrt{n}} \\ &\xrightarrow{d} \Phi_1 - Q_{12}' Q_{22}^{-1}\Phi_2 = \Phi_{1\cdot 2},\ \ V(\Phi_{1\cdot 2}) = \sigma_u^2 Q_{1\cdot 2}.\end{aligned}$$

So, $\frac{Z_1' M_2 u}{\sqrt{n}} \xrightarrow{d} \Phi_{1\cdot 2} \sim N(0, \sigma_u^2 Q_{1\cdot 2})$.

It then follows that

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} \frac{\pi' \Phi_{1\cdot 2}}{\pi_1' Q_{1\cdot 2}\pi_1} \sim \frac{N(0, \sigma_u^2 \pi_1' Q_{1\cdot 2}\pi_1)}{\pi_1' Q_{1\cdot 2}\pi_1} \sim N(0, \frac{\sigma_u^2}{\pi_1' Q_{1\cdot 2}\pi_1}).$$

Weak Instruments - Motivation

Note that the asymptotic variance of $\hat{\gamma}$ is inversely proportional to

$$\|\pi_1\|^2_{Q_{1\cdot 2}} = \pi_1' Q_{1\cdot 2}\pi_1.$$

So, when π_1 is close to zero, the asymptotic variance is larger. "Weaker" instruments correspond to instruments with smaller $\|\pi_1\|^2_{Q_{1\cdot 2}}$.

The asymptotics above assumed that π_1 is fixed as $n \to \infty$. As a result, the estimation error for γ is always 'small" relative to π for large enough n.

> The estimation error is $O_p(\frac{1}{\sqrt{n}})$.
>
> However, when we are dealing with finite samples, it is possible that the estimation error and π_1 have similar magnitude.
>
> $\Rightarrow$ These asymptotic results will provide a poor approximation to the behavior of $\hat{\gamma}$ in finite samples.

Weak Instruments – Intro

Suppose there is a single endogenous regressor, no exogenous regressors and a single instrument:

$$\begin{aligned} y_1 &= \gamma y_2 + u \\ y_2 &= \pi_1 Z_1 + v. \end{aligned}$$

The IV estimator is then

$$\begin{aligned} \hat{\gamma} &= \frac{Z_1' y_1}{Z_1' y_2} \\ &= \gamma + \frac{Z_1' u}{\pi_1 Z_1' Z_1 + Z_1' v} \\ &= \gamma + \frac{n^{-1/2} Z_1' u}{\pi_1 n^{-1/2} Z_1' Z_1 + n^{-1/2} Z_1' v} \end{aligned}$$

Weak Instruments – Intro

We can write this as

$$\hat{\gamma} = \gamma + \frac{O_p(1)}{\pi_1 n^{-1/2} Z_1' Z_1 + O_p(1)},$$

where $n^{-1/2} Z_1' u = O_p(1), n^{-1/2} Z_1' v = O_p(1)$ can be justified by a CLT. These terms are the **noise** due to estimation – they're a function of the errors.

Note that the term

$$n^{-1/2} Z_1' Z_1 = n^{1/2}(n^{-1} Z_1' Z_1) \xrightarrow{\infty}$$

by the WLLN. The term, $\pi_1 n^{-1/2} Z_1' Z_1$ is the **signal** in the data about γ. Note that provided $\pi_1 \neq 0$, the signal component blows up to infinity and the signal will dominate the noise in large samples.

Weak Instruments – Intro

If $\pi_1 = 0$, then Z_1 is irrelevant and the data contains no information about γ. Then,

$$\hat{\gamma} - \gamma = \frac{n^{-1/2} Z_1' u}{n^{-1/2} Z_1' v} \xrightarrow{d} \frac{Z_u}{Z_v}, \quad \begin{pmatrix} Z_u \\ Z_v \end{pmatrix} \sim N(0, \Sigma).$$

Weak instrument case: The data contains only "some" information about γ. Model this by assuming that the signal and noise are of the same order of magnitude. That is,

$$\pi_1 n^{-1/2} Z_1' Z_1 \xrightarrow{p} C,$$

for some constant C. We accomplish this by assuming $\pi_1 = \frac{C}{n^{1/2}}$ and so, the signal component then converges to $C \cdot E[Z_{i1}^2]$. The signal will no longer dominate the noise and $\hat{\gamma}$ is inconsistent.

> Known as "local-to-zero asymptotics" and first proposed by Stock & Staiger (1997) for formalizing the problem of weak instruments.

Outline

Instrumental Variable Model

Intro to Weak IV

Potential Outcomes

Conditional Random Assignment

Estimating ATE under Strong Ignorability

Instrumental Variable Model and ATE

IV Bounds on ATE

Potential Outcomes

Dominant framework for defining causal effects in statistics and econometrics.

Begin with a brief review following the notation set up in lecture. We focus attention on the case of a binary treatment $T_i \in \{0, 1\}$.

Potential Outcomes

For each unit i, pair of potential outcomes $(Y_i(0), Y_i(1))$.

> Notation implicitly imposes the stable unit treatment value assumption (SUTVA).
>
> Potential outcomes for any unit do not vary with the treatments assigned to other units and there are no hidden versions of the treatment.

Potential outcomes give the value of the outcome Y_i that would be observed if unit i receives either control $(T_i = 0)$ or treatment $(T_i = 1)$. Observed outcome is

$$Y_i \equiv Y_i(T_i) = \begin{cases} Y_i(1) & \text{if} \quad T_i = 1 \\ Y_i(0) & \text{if} \quad T_i = 0 \end{cases}$$

or

$$Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i).$$

Causal Effects

Causal effects are defined for each unit i using i's potential outcomes.

The **causal effect for unit** i is

$$\tau_i = Y_i(1) - Y_i(0).$$

Causal effect of treatment is allowed to vary across units i.e. the treatment effects may be *heterogeneous*.

Define

$$m(t) = E[Y_i(t)] \quad \text{for} \quad t \in \{0, 1\}.$$

This is **average potential outcome function**. The **average treatment effect** (ATE) is

$$ATE = E[Y_i(1) - Y_i(0)].$$

For now, $E[\cdot]$ is an expectation over some super-population.

Will return to this.

Fundamental Problem of Causal Inference

The **fundamental problem of causal inference** is that given a treatment assignment, we only observe one potential outcome for each unit.

> Phrase coined by Rosenbaum (1986).

Causal inference is a **missing data problem**. How do we solve this?

> Note that it is not enough to simply observe (Y_i, T_i) for many units due to **selection bias**.

Selection Bias

There is selection bias if comparing the average outcome of units that received the treatment to the average outcome of units that received the control does not equal ATE.

Define

$$r(t) = E[Y_i | T_i = t]$$

to be the **regression function** for $t \in \{0, 1\}$. There is selection bias if

$$r(t) \neq m(t).$$

Example: "Roy Model" - individuals select into college based on expectations of earnings.

Worst-case bounds on ATE

Can we learn anything about ATE without additional assumptions? *Yes*.

Suppose that $Y(0), Y(1) \in [0, 1]$. And consider, $E[Y(1)]$. We have that

$$E[Y(1)] = E[Y(1)|T = 1]P(T = 1) + E[Y(1)|T = 0]P(T = 0).$$

We can identify $E[Y(1)|T = 1], P(T = 1), P(T = 0)$ in the data and $E[Y(1)|T = 0]$ is not identified. But, by assumption, we have that

$$0 \leq E[Y(1)|T = 0] \leq 1.$$

So, we immediately have that

$$\begin{aligned} E[Y(1)|T = 1]P(T = 1) \leq E[Y(1)] & \\ & E[Y(1)] \leq E[Y(1)|T = 1]P(T = 1) + P(T = 0). \end{aligned}$$

These are **worst case bounds** on $E[Y(1)]$.

Worst-case bounds on ATE

We can similarly derive worst-case bounds on $E[Y(0)]$. We have that

$$E[Y(0)] = E[Y(0)|T = 1]P(T = 1) + E[Y(0)|T = 0]P(T = 0).$$

We can identify $P(T = 1), E[Y(0)|T = 0], P(T = 0)$ from the data and $E[Y(0)|T = 1]$ is not identified. Using the assumption $Y(0) \in [0, 1]$, we have

$$\begin{aligned} E[Y(0)|T = 0]P(T = 0) &\leq E[Y(0)] \\ E[Y(0)] &\leq P(T = 1) + E[Y(0)|T = 0]P(T = 0) \end{aligned}$$

Worst-case bounds on ATE

$ATE = E[Y(1)] - E[Y(0)]$. So, an upper-bound on ATE is

$$E[Y(1)|T = 1]P(T = 1) + P(T = 0) - E[Y(0)|T = 0]P(T = 0)$$

and a lower-bound on ATE is

$$E[Y(1)|T = 1]P(T = 1) - P(T = 1) - E[Y(0)|T = 0]P(T = 0).$$

The length of this interval is 1, so only looking at the data halved the bounds on ATE.

Random Assignment

A sufficient assumption to identify ATE is **random assignment**. That is,

$$(Y_i(1), Y_i(0)) \perp T_i$$

or the treatment is independent of the set of potential outcomes.

Under this assumption,

$$\begin{aligned} r(1) &= E[Y_i | T_i = 1] \\ &= E[Y_i(1) | T_i = 1] \\ &= E[Y_i(1)] = m(1) \end{aligned}$$

Similarly, $r(0) = m(0)$. As a result, ATE is identified by the regression function:

$$ATE = E[Y_i(1) - Y_i(0)] = E[Y_i | T_i = 1] - E[Y_i | T_i = 0].$$

Sources of Uncertainty

Two sources of uncertainty in potential outcomes model:

- Sampling-based uncertainty
- Design-based uncertainty

Sampling-based Uncertainty

Sample of units $i = 1, \ldots, N$ drawn randomly from a super-population of interest.

> If we resampled from the population, we would observe a different set of units.

View the pair of potential outcomes $(Y_i(0), Y_i(1))$ as a random vector with $(Y_i(0), Y_i(1))$ i.i.d. from some distribution F.

Design-based Uncertainty

Design-based uncertainty arises due to the random assignment of units in our sample to treatments.

The units $i = 1, \ldots, N$ are now fixed but the treatment T_i is random.

Each time we randomize the treatment we would observed different outcomes.

What does it matter?

We defined the average treatment effect in terms of some target super-population

$$ATE = E[Y_i(1) - Y_i(0)]$$

where the expectation is taken over some population distribution F for the pair $(Y_i(0), Y_i(1))$.

> Imbens & Wooldridge (2007) refers to this as the **population average treatment effect** (PATE). Sampling uncertainty in addition to design uncertainty.

Alternative definition conditions on the given sample and defines

$$ATE = N^{-1} \sum_{i=1}^{N} Y_i(1) - Y_i(0)$$

.

> Imbens & Wooldridge (2007) refers to this as the *sample average treatment effect* (SATE). Only design uncertainty.

These are different objects!

Outline

Instrumental Variable Model

Intro to Weak IV

Potential Outcomes

Conditional Random Assignment

Estimating ATE under Strong Ignorability

Instrumental Variable Model and ATE

IV Bounds on ATE

Conditional Random Assignment

Treatment is not randomly assigned but is "as if" randomly assigned among similar units.

> additionally observe some covariates W_i that take values in the set $\mathcal{W}$ and assume that the treatment is **randomly assignment conditional on** W_i or
>
> $$(Y_i(1), Y_i(0)) \perp T_i | W_i.$$

This is referred to as **conditional random assignment**, **unconfoundedness** or **selection on observables**.

Conditional Average Treatment Effects

Let $m(t|w)$ be the average potential outcome function conditional on $W_i = w$ and let $r(t, w)$ denote the regression function.

$$m(t|w) = E[Y_i(t)|W_i = w] \quad \text{and} \quad r(t, w) = E[Y_i|T_i = t, W_i = w].$$

With conditional random assignment, the regression function identifies the average potential outcome function conditional on $W_i = w$.

$$\begin{aligned} r(t, w) &= E[Y_i|T_i = t, W_i = w] \\ &= E[Y_i(t)|T_i = t, W_i = w] \\ &= E[Y_i(t)|W_i = w] = m(t|w). \end{aligned}$$

Conditional Average Treatment Effects

Can use the regression function to identify the conditional average treatment effect (CATE)

$$CATE(w) = E[Y_i(1) - Y_i(0)|W_i = w]$$

Under conditional random assignment,

$$\begin{aligned} r(1,w) - r(0,w) &= E[Y_i|T_i = 1, W_i = w] - E[Y_i|T_i = 0, W_i = w] \\ &= E[Y_i(1)|T_i = 1, W_i = w] - E[Y_i(0)|T_i = 0, W_i = w] \\ &= E[Y_i(1)|W_i = w] - E[Y_i(0)|W_i = w] = CATE(w). \end{aligned}$$

Identifying ATE?

Not sufficient to identify the average potential outcome function nor the average treatment effect.

> May be parts of the covariate distribution $w \in \mathcal{W}$ in which there are no individuals in both treatment and control.
>
> $\Rightarrow$ we cannot estimate the regression $r(1, w)$ or $r(0, w)$ at this value w.

Need additional assumption

$$0 < P(T_i = 1 | W_i = w) < 1$$

for all $w \in \mathcal{W}$. Known as **overlap**.

If the treatment T_i satisfies conditional random assignment and overlap, referred to as a **strongly ignorable treatment assignment**.

Identifying ATE?

With both assumptions, can identify the average potential outcome functions and the average treatment effect by averaging over the distribution of W_i.

$$\begin{aligned} m(t) &= E[Y_i(t)] \\ &= E[E[Y_i(t)|W_i]] \\ &= E[r(t, W_i)] \end{aligned}$$

and

$$\begin{aligned} ATE &= E[Y_i(1) - Y_i(0)] \\ &= E[E[Y_i(1) - Y_i(0)|W_i]] \\ &= E[r(1, W_i) - r(0, W_i)]. \end{aligned}$$

Exercise 1

We have data on a random sample of individuals ($i = 1, \ldots, n$), with observations on an outcome Y_i, an indicator $T_i = 0, 1$ for which of two treatments was received, and a vector of individual characteristics W_i. There is a pair of random variables, $Y_i(0)$; $Y_i(1)$ and

$$Y_i = 1(T_i = 1)Y_i(1) + 1(T_i = 0)Y_i(0).$$

There are two regression functions corresponding to the average potential outcomes

$$m(0|w) = E[Y_i(0)|W_i = w], \quad m(1|w) = E[Y_i(1)|W_i = w].$$

Let $\gamma = E[Y_i(1) - Y_i(0)] = ATE$.

Exercise 1 (continued)

Suppose that the treatment is assigned by

$$T_i = 1 \quad \text{if} \quad m(1|W_i) - m(0|W_i) \geq C_i$$

where C_i is unobserved.

(1) Suppose that C_i is independent of $(Y_i(0), Y_i(1))$ conditional on W_i. Show that γ is identified.

The propensity score

The probability of treatment given a value of W_i is incredibly useful function in causal inference. Called the **propensity score**, denoted

$$e(w) = E[T_i | W_i = w] = P(T_i = 1, W_i = w).$$

The propensity score

Theorem

Suppose that the treatment is conditionally independent of the potential outcomes given W_i. Then,

$$(Y_i(1), Y_i(0)) \perp T_i | e(W_i).$$

The propensity score

Proof of theorem:

Show that $P(T_i = 1|Y_i(0), Y_i(1), e(W_i)) = P(T_i = 1|e(W_i)) = e(W_i)$. We have that

$$\begin{aligned} & P(T_i = 1|Y_i(0), Y_i(1), e(W_i)) = E[T_i|Y_i(0), Y_i(1), e(W_i)] \\ & = E[E[T_i|Y_i(0), Y_i(1), e(W_i), W_i]|Y_i(0), Y_i(1), e(W_i)] \\ & = E[E[T_i|Y_i(0), Y_i(1), W_i]|Y_i(0), Y_i(1), e(W_i)] \\ & = E[E[T_i|W_i]|Y_i(0), Y_i(1), e(W_i)] \\ & = E[e(W_i)|Y_i(0), Y_i(1), e(W_i)] = e(W_i) \end{aligned}$$

Outline

Instrumental Variable Model

Intro to Weak IV

Potential Outcomes

Conditional Random Assignment

Estimating ATE under Strong Ignorability

Instrumental Variable Model and ATE

IV Bounds on ATE

Estimating ATE under strong ignorability

Assume that conditional random assignment and overlap both hold. We observe a data set consisting of N observations. For each observation, we observe the triple (Y_i, T_i, W_i).

enormous literature that studies the properties and relative benefits of these methods

Provide a very (very) brief introduction to some of these techniques.

Regression

Suppose $\hat{r}(t, w)$ is a consistent estimators of the conditional expectation function $r(t, w) = E[Y_i | T_i = t, W_i = w]$.

The simplest estimator of the average treatment effect simply averages over the empirical distribution of W_i:

$$A\hat{T}E = \frac{1}{N} \sum_{i=1}^{N} \hat{r}(1, W_i) - \hat{r}(0, W_i).$$

Regression

Consider two cases in lecture:

- T_i, W_i are both binary and take on values $T_i \in \{0, 1\}$ and $W_i \in \{w_0, w_1\}$.
    - Population linear predictor that includes all dummy variables is the conditional expectation function
- T_i is binary and W_i is a scalar.
    - $r(0, w)$ and $r(1, w)$ are continuous functions
    - Proposed using polynomials or splines to approximate the conditional expectation functions. Examples of *non-parametric regression*

Intuition of this approach:

- Use estimated regression function to impute the missing potential outcomes. For instance, if $T_i = 1$, then we observe $Y_i(1)$ and impute $Y_i(0)$ using $\hat{r}(0, W_i)$.

Matching

Idea: impute the missing potential outcomes by looking at the observed outcomes of the "nearest neighbors" in the opposite treatment group.

Let $l_m(i)$ be the index l that satisfies $T_l \neq T_i$ and

$$\sum_{j: T_j \neq T_i} 1(\|W_j - W_i\| \leq \|W_l - W_i\|) = m.$$

> In English, $l_m(i)$ is index of the unit in the opposite treatment group that is the m-th closest unit to i in terms of covariate distance based on the norm $\| \cdot \|$.

Matching

Let $\mathcal{L}_M(i)$ be the set of indices for the first M matches of unit i. The imputed potential outcomes are given by

$$\hat{Y}_i(0) = \begin{cases} Y_i \quad \text{if} \quad T_i = 0, \\ \frac{1}{M} \sum_{j \in \mathcal{L}_M(i)} Y_j \quad \text{if} \quad T_i = 1 \end{cases}$$

$$\hat{Y}_i(1) = \begin{cases} \frac{1}{M} \sum_{j \in \mathcal{L}_M(i)} Y_j \quad \text{if} \quad T_i = 0, \\ Y_i \quad \text{if} \quad T_i = 1. \end{cases}$$

Matching

Number of matches M must be selected as well as the distance measure $\| \cdot \|$.

- Euclidean distance, $d_e(w_i, w_j) = (w_i - w_j)'(w_i - w_j)$.
- Common to standardize the covariates by using **Mahalanobis distance** with

$$d_M(w_i, w_j) = (w_i - w_j)'\Sigma_W^{-1}(w_i - w_j),$$

where Σ_W is the covariance matrix of the covariates that must be estimated.

Inverse Propensity Score Weighting

Conditional random assignment conditional on W_i implies conditional random assignment conditional on $e(W_i)$.

Suggests another route: estimate the propensity score $e(W_i)$, then throw away the covariates W_i and proceed using only the estimated propensity score in our analysis.

Inverse Propensity Score Weighting

Relies on following result:

$$E[\frac{T_i Y_i}{e(W_i)}] = E[Y_i(1)] \quad \text{and} \quad E[\frac{(1-T_i)Y_i}{1-e(W_i)}] = Y_i(0).$$

We'll show first equality:

$$\begin{aligned} E[\frac{T_i Y_i}{e(W_i)}] &= E[\frac{T_i Y_i(1)}{e(W_i)}] \\ &= E[E[\frac{T_i Y_i(1)}{e(W_i)}|W_i]] \\ = E[E[\frac{e(W_i) Y_i(1)}{e(W_i)}]] &= E[Y_i(1)]. \end{aligned}$$

Inverse Propensity Score Weighting

Follows that ATE can be written as

$$ATE = E[\frac{T_i Y_i}{e(W_i)} - \frac{(1-T_i)Y_i}{1-e(W_i)}].$$

If propensity score is known exactly, a valid estimator is

$$A\hat{T}E = \frac{1}{N}\sum_{i=1}^{N} \frac{T_i Y_i}{e(W_i)} - \frac{(1-T_i)Y_i}{1-e(W_i)}.$$

Of course it must be estimated...

Outline

Instrumental Variable Model

Intro to Weak IV

Potential Outcomes

Conditional Random Assignment

Estimating ATE under Strong Ignorability

Instrumental Variable Model and ATE

IV Bounds on ATE

IV Model – Treatment Effect Heterogeneity

We'll now present the IV model with heterogeneous treatment effects. Then, we will place sufficient restrictions on the treatment effect heterogeneity so that the IV estimator will deliver an ATE.

Assume that treatment T_i is not randomly assigned but there is an instrument S_i that is randomly assigned and correlated with T_i.

For each individual i, there is a potential treatment function $T_i(\cdot)$ that can be evaluated at any $s \in \mathcal{S}$. $T_i(s)$ is the treatment realized for individual i at instrument level s. We observe

$$T_i = T_i(S_i).$$

For each individual i, there is a potential outcome function $Y_i(\cdot, \cdot)$ that can be evaluated at any level of the treatment and subsidy. We observe

$$Y_i(T_i, S_i)$$

IV Model – Treatment Effect Heterogeneity

We make two key assumptions:

(1) **Exclusion restriction**:

$$Y_i(t, s_1) = Y_i(t, s_2) \quad \forall s_1, s_2 \in \mathcal{S}.$$

(2) **Random assignment of instrument**:

$$\{\{Y_i(t), t \in \mathcal{T}\}, \{T_i(s), s \in \mathcal{S}\}\} \perp\!\!\!\perp S_i.$$

IV Model – Constant Treatment Effects

Suppose that $T_i \in \{0,1\}$, $S_i \in \{0,1\}$ and that the treatment effects are constant:

$$Y_i(1) = Y_i(0) + K$$

for all i. Then, the IV estimator identifies $ATE = K$. See the notes for the derivation.

Next time – we'll consider what the IV estimator identifies when we do not restrict the heterogeneity of the treatment effects.

Outline

Instrumental Variable Model

Intro to Weak IV

Potential Outcomes

Conditional Random Assignment

Estimating ATE under Strong Ignorability

Instrumental Variable Model and ATE

IV Bounds on ATE

IV Bounds on ATE

Under the IV assumptions, we can derive bounds on the ATE without treatment effect homogeneity assumptions. Suppose again that $Y \in [0,1]$.

We have that

$$\begin{aligned} E[Y(1)] &= E[Y(1)|Z=z] \\ &= E[Y(1)|Z=z, T=1]P(T=1|Z=z) \\ &+ E[Y(0)|Z=z, T=0]P(T=0|Z=z) \\ &\leq E[Y(1)|Z=z, T=1]P(T=1|Z=z) + P(T=0|Z=z) \end{aligned}$$

This holds for all $z \in \mathcal{Z}$. So, an upper-bound is

$$E[Y(1)] \leq \min_z E[Y(1)|Z=z, T=1]P(T=1|Z=z) + P(T=0|Z=z).$$

Similarly, a lower bound is

$$\max_z E[Y(1)|Z=z, T=1]P(T=1|Z=z) \leq E[Y(1)].$$

Great Practice Q: Work out the IV bounds on ATE for a binary instrument, $Z = \{0,1\}$ – this is a very simple calculation.