LECTURE 6: IDENTIFICATION

Guido Imbens – Stanford University

Economics 272, GSB 507, Spring 2025

LEAMER

- Edward Leamer, “Let’s take the con out of Econometrics,” (1983):

  “He mentions the phrase "identification problem," which, though no one knows quite what he means, is said with such authority that it is totally convincing.”

  (Very important paper, highly recommended, more on this in Lecture 7)

THE MEANING OF IDENTIFICATION

- A complete definition of the concept of “identification” in econometrics is difficult to give. See Lewbel (2019) for discussion of many definitions.
  - Informally, we say a parameter is (point-)identified if given an infinitely large sample we can infer its value.
  - We say it is not identified if the infinitely large sample is uninformative about the parameter.
  - We say it is partially identified if the infinitely large sample narrows down the possible set of values.
- Part of the challenge is that you need to make precise what you mean about getting a larger sample. Easy in a random sampling case, not so easy in time series setting:
  - more time periods
  - more units within same time period,

  also challenging in spatial setting, or panel setting, or network setting.

FORMAL ARGUMENT

- Consider a sequence of populations, indexed by $k = 1, \ldots, \infty$.
- In population k we observe a matrix Y_k of dimension $N_k \times M_k$ (N_k may be sample size, and may go to infinity with k, M_k may be number of covariates, also increasing with k)
- Y_k has a distribution $f_k(y_k; \theta, \pi_k)$, indexed by a common parameter θ and nuisance parameters π_k (the latter indexed by k). Dimensions of these parameters may also be indexed by k, possibly infinite dimensional.
- Identification is about the ability to infer value or sets of possible values of θ from the sequence of Y_k.

EXAMPLE: IDENTIFICATION IN CROSS-SECTION SETTINGS

- Consider cross-section setting .
- Suppose our data consists of a random sample from some distribution, same for all populations:

$$Y_{k1}, \ldots, Y_{kN_k} \text{ i.i.d. } \sim f_Y(y)$$

- Then as $N_k \to \infty$, we can consistently estimate the cumulative distribution function $F_Y(\cdot)$:

$$\hat{F}_Y(y) = \frac{1}{N_k} \sum_{i=1}^{N} 1_{Y_{ki} \leq y} \xrightarrow{p} F_Y(y).$$

- So, let's pretend we know $F_Y(\cdot)$.

IDENTIFICATION IN CROSS-SECTION SETTINGS

- We also have a model that says that $F_Y(\cdot)$ has a particular form:

$$F_Y(y) = M(y; \theta, \lambda),$$

  $\theta \in \Theta, \lambda \in \Lambda$ (both θ and λ unknown), and $M(\cdot)$ known. θ and λ may be functions or finite dimensional parameters.

- We care about θ, and the other parameter λ is just a nuisance parameter.
- We say that θ is (point-)identified if we can infer the value of θ from knowledge of $F_Y(\cdot)$.

IDENTIFICATION IN TIME SERIES SETTINGS

- Sample is $Y_{k1}, \ldots, Y_{kT_k}$, where Y_{kt} is the value at time $\Delta_k t$.
- Suppose Y_{kt} is a stationary time series, and we are interested in the long run average.
- Now we can do consider the case where the sequence of populations has sample size $T_k = k$, and interval $\Delta_k = 1$. This is the standard asymptotic sequence where we get mnore observations further and further out into the future. The long run average is identified here for stationary time series.
- We can also think of the sequence $T_k = k$, $\Delta_k = 1/k$, so that we get more and more observations for the same fixed time period, say for the same year. This is infill asymptotics. We cannot learn the long-run average under this asymptotic sequence.

WEAK INSTRUMENT ASYMPTOTICS (STAIGER & STOCK, 1997)

- Model, for $i = 1, \ldots, N_k$:

$$Y_{ki} = \beta_0 + X_{ki}\beta_1 + \varepsilon_{ki} \qquad \text{(parameters not indexed by } k\text{)}$$

$$X_{ki} = \pi_{k0} + Z_{ki}\pi_{k1} + \eta_{ki} \qquad \text{(parameters indexed by } k\text{)}$$

$$Z_{ki} \perp\!\!\!\perp \left(\varepsilon_{ki}, \eta_{ki}\right)$$

- The sequence of populations satisfies a weak instrument condition

$$\pi_{k1} = C/\sqrt{N_k}$$

- What can we learn about common β_1?

MANY INSTRUMENT ASYMPTOTICS (BEKKER, 1994)

- Same model as Staiger & Stock (1997)
- In addition, the number of instruments, the dimension of $Z_{k,i}$ (and π_{k1}) is $D_k \propto N_k$, proportional to the sample size.
- There are many sequences of populations to consider!
- How do you choose between different asymptotic sequences?
  - Not always clear and not always agreement
  - Sometimes initially researchers focus on one sequence, and later decide other sequences are more appropriate.
  - Choice is about deciding that some aspects of the problem are not important relative to others.

ASYMPTOTIC SEQUENCES AND INFERENCE

- Binary data, Sample size $N_k = k$:

$$Y_{ki} \sim \text{Binomial}(1, p_k), , i = 1, \ldots, N_k$$

- What is the distribution of $X = \sum_{i=1}^{N_k} Y_{ki}$?
  - $\mathbb{E}[X_k] = p_k N_k, \mathbb{V}(X_k) = N_k p_k (1 - p_k)$
  - Exact distribution is difficult to approximate.
  - Approximation based on normal distribution, assuming $p_k \rightarrow p$, leads to

$$\sqrt{N}_k (X_k / N_k - p) \xrightarrow{d} \mathcal{N}(0, p(1 - p))$$

  - Approximation based on Poisson distribution, assuming $N_k p_k \rightarrow \lambda$, leads to

$$X_k \xrightarrow{d} \mathcal{P}(\lambda)$$

Better approximation for small p_k.

IDENTIFICATION: TAKEAWAYS

- What makes identification questions interesting and answers compelling?
  - It sheds new light on an existing problem.
  - It establishes identification in settings where it was unknown.
- Subject to fads:
  - Results that were viewed as interesting at some point may no longer be viewed as useful/credible.
  - Example: unobserved heterogeneity in duration models (Elbers & Ridder, 1984; Heckman and Singer, 1984)
- Delicate interplay between assumptions that can be credible (conditional independence relations) and those that are not convincing (functional form assumptions).

EXAMPLE I: ORDINARY LEAST SQUARES

- Now dropping explicit indexing by sequence of populations.
- We focus on cross-section setting. Here we assume the sequence consists of iid random variables, all with the same distribution, not changing with the population.
- Suppose

$$Y_i|X_i \sim \mathcal{N}(\alpha + \beta X_i, \sigma^2).$$

- Then slope coefficient β is identified, as long as $\mathbb{V}(X_i) > 0$.

$$\beta = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)}$$

EXAMPLE II: ORDINARY LEAST SQUARES IN A DIFFERENT SETTING

- Model:

$$Y_i | X_i \sim \mathcal{N}(\alpha + \beta X_i + \gamma(X_i/2), \sigma^2).$$

- Here β and γ are not identified.
- We can identify linear combination $\beta + \gamma/2$.
- So, β and γ are not even partially identified, we can learn nothing about their values, other than a linear combination.

EXAMPLE III: PARTIAL IDENTIFICATION IN SELECTION MODEL, MANSKI, 1990

- Suppose $(D, Y) \sim f(d, y)$, $D \in \{0, 1\}$, $Y_i \in \{0, 1\}$, and we observe

$$(D_i, D_i \cdot Y_i)$$

- So, we only observe Y_i if $D_i = 1$.
- Suppose we are interested in $\mu = E[Y_i] = \text{pr}(Y_i = 1)$.
- In large samples we know

$$\text{pr}(D_i = 1), \;\; \text{pr}(Y_i = 1 | D_i = 1).$$

EXAMPLE III: PARTIAL IDENTIFICATION IN SELECTION MODEL (CTD)

- We know nothing about $\text{pr}(Y_i = 1|D_i = 0)$ other than that it is between zero and one.
- We can decompose the marginal probability that $Y_i = 1$:

$$\text{pr}(Y_i = 1) = \text{pr}(D_i = 1)\text{pr}(Y_i = 1|D_i = 1) + (1 - \text{pr}(D_i = 1))\text{pr}(Y_i = 1|D_i = 0)$$

  with orange parts estimable, and blue parts completeley unknown (and than beteween zero and one).
- So,

$$\text{pr}(D_i = 1)\text{pr}(Y_i = 1|D_i = 1) \leq \text{pr}(Y = 1) \leq \text{pr}(D_i = 1)\text{pr}(Y_i = 1|D_i = 1) + (1 - \text{pr}(D_i = 1))$$

- Hence $\text{pr}(Y = 1)$ is partially identified, with the width of the identified set equal to $1 - \text{pr}(D_i = 1)$.

EXAMPLE IV: DURATION MODEL WITH UNOBSERVED HETEROGENEITY

- Suppose we observe a random sample of durations and covariates (T_i, X_i).
- Conditional on an unobserved scalar random variable $U_i = u$ and the covariate $X_i = x$, the hazard for the duration is

$$h(t|u, x) = \lambda(t) \exp(\beta x + u)$$

- U_i is independent of X_i (unobserved heterogeneity, frailty models).
- $\lambda(\cdot)$ and β and the distribution of U_i are unknown (some moments of U_i exist).
- Note: if the hazard is $h(t)$, the distribution function of the duration is

$$F(t) = \exp\left(-\int_0^t h(s)ds\right)$$

- Can we identify β? Answer not obvious.
- Elbers and Ridder (1983), and Heckman and Singer (1984): answer is yes.

EXAMPLE V: CHAMBERLAIN CONDITIONAL LOGIT MODEL

- Panel data set up with binary outcomes:

$$\text{pr}(Y_{it} = 1 | X_{i1}, \ldots, X_{iT}, \alpha_i) = \frac{\exp(X_{it}^\top \beta + \alpha_i)}{1 + \exp(X_{it}^\top \beta + \alpha_i)} \qquad , i = 1, \ldots, N, t = 1, \ldots, T$$

- α_i is fixed effect.
- $Y_{it}, Y_{it'}$ are independent conditional on α_i and $X_{i1}, \ldots, X_{iT}$.
- Fixed T (*e.g.* $T = 2$), so no consistent estimation of α_i is possible.
- Chamberlain:

$$\text{pr}(Y_{i1} = 1 | Y_{i1} + Y_{i2} = 1, X_{i1}, X_{i2}, \alpha_i) = \frac{\exp((X_{i1} - X_{i2})^\top \beta)}{1 + \exp((X_{i1} - X_{i2})^\top \beta)}$$

- Fixed effect can be removed by conditioning on $Y_{i1} + Y_{i2} = 1$, and β is identified.
- Does not work for probit or other binary response models.
  - What do we make of that?
  - Is this useful if it only works for logit and not probit?

EXAMPLE VI: HECKMAN SELECTION MODEL

- Interest in regression model

$$Y_i = \beta_0 + X_i^\top \beta_1 + \varepsilon_i \qquad \varepsilon \perp\!\!\!\perp X_i.$$

- Observe $(Y_i, X_i | W_i = 1)$, where

$$W_i = \begin{cases} 1 & \text{if } \pi_0 + Z_i^\top \pi + \eta_i \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

- ε_i and η_i may be correlated., but both are uncorrelated with (X_i, Z_i) In that case a regression of Y_i on X_i does not work (suffers from selection bias).

EXAMPLE VI: HECKMAN SELECTION MODEL

- If (ε_i, η_i) are jointly normal with correlation ρ, then

$$\mathbb{E}[Y_i | X_i, Z_i, W_i = 1] = \beta_0 + X_i^\top \beta_1 + \mathbb{E}[\varepsilon_i | \eta_i > -\pi_0 - Z^\top \pi]$$

$$= \beta_0 + X_i^\top \beta_1 + \frac{\rho \sigma_\varepsilon}{\sigma_\eta} \frac{\phi(-(\pi_0 + Z^\top \pi)/\sigma_\eta)}{\Phi((\pi_0 + Z^\top \pi)/\sigma_\eta)}$$

  (where $\phi(\cdot)$ and $\Phi(\cdot)$ are the Normal density and cdf respectively, and the ratio is the inverse Mills ratio.)

- β is identified.
- How satisfactory is this result:
  - What can we say without assuming Normality?
  - What can we say if Z_i and X_i are the same?
  - What if the conditional expectation in the population is not linear?

EXAMPLE VII: HONORE

- Censored Regression Model:

$$Y_{it}^* = \alpha_i + X_{it}^\top \beta + \varepsilon_{it} \qquad Y_{it} = \max(0, Y_{it}^*), \qquad t = 1, 2.$$

- α_i is fixed effect.
- Question: is β identified in fixed T setting (here $T = 2$)?
- Honore: yes. Also proposes consistent estimator.
- Assumptions include:
  - The distribution of $\varepsilon_{1i} - \varepsilon_{2i}$ conditional on $\varepsilon_{i1} + \varepsilon_{i2}$ and conditional on X_{i1}, X_{i2}, α_i is symmetric around zero.

EXAMPLE VIII: THE CORRELATION OF POTENTIAL OUTCOMES

- Suppose we have a randomized experiment where we observe (W_i, Y_i) for a random sample from the population,
- $Y_i = Y_i(1)W_i + Y_i(0)(1 - W_i)$.
- What can we say about the correlation between $(Y_i(0), Y_i(1))$.
- Not compeletely identified, but some partial identification.

IDENTIFICATION OF TREATMENT EFFECTS WITH INSTRUMENTAL VARIABLES: RESULT I (HECKMAN (1990))

- Set up

$$Y_i(0) = \beta_0 + \varepsilon_{i0} \qquad Y_i(1) = \beta_1 + \varepsilon_{i1} \qquad W_i = 1_{\pi_0+\pi_1 Z_i+\eta_i>0}$$

$$\text{Key Assumption :} \qquad Z_i \perp\!\!\!\perp (\varepsilon_{i0}, \varepsilon_{i1}, \eta_i)$$

- We observe

$$(Z_i, W_i, Y_i), \quad Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$$

- Can the average treatment effect $\beta_1 - \beta_0$ be inferred from the joint distribution of (Z_i, W_i, Y_i)?
- Point identification of $\tau \equiv \beta_1 - \beta_0$ requires that the support of Z_i is unbounded (equal to the entire real line): identification at infinity.

LOCAL AVERAGE TREATMENT EFFECT VERSION

- Let's re-formulate the example. Suppose Z_i is binary: $Z_i \in \{0, 1\}$.
- Let $W_i(z)$ denote the treatment value if a unit is assigned instrument value z, corresponding to

$$W_i(z) = 1_{\pi_0 + \pi_1 z + \eta_i > 0},$$

  in Heckman's latent index version.
- Previous Assumption

$$Z_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1), W_i(0), W_i(1) \Big)$$

- Assume also:

$$W_i(1) \geq W_i(0)$$

LOCAL AVERAGE TREATMENT EFFECT VERSION (CTD)

- Define compliance types C_i:
  - a: always takers (units with $W_i(0) = W_i(1) = 1$),

    n: never takers (units with $W_i(0) = W_i(1) = 0$),

    c: compliers (units with $W_i(0) = 0$, $W_i(1) = 1$).
- Now consider

$$E[Y_i|Z_i = 1]$$

by iterated expectations this is equal to:

$$E[Y_i|C_i = a, Z_i = 1]\mathrm{pr}(C_i = a|Z_i = 1) + E[Y_i|C_i = n, Z_i = 1]\mathrm{pr}(C_i = n|Z_i = 1) + E[Y_i|C_i = c, Z_i = 1]\mathrm{pr}(C_i = c|Z_i = 1$$

$$= E[Y_i(1)|C_i = a, Z_i = 1]\mathrm{pr}(C_i = a) + E[Y_i(0)|C_i = n, Z_i = 1]\mathrm{pr}(C_i = n) + E[Y_i(1)|C_i = c, Z_i = 1]\mathrm{pr}(C_i = c)$$

$$= E[Y_i(1)|C_i = a] \cdot \mathrm{pr}(C_i = a) + E[Y_i(0)|C_i = n] \cdot \mathrm{pr}(C_i = n) + E[Y_i(1)|C_i = c] \cdot \mathrm{pr}(C_i = c)$$

LOCAL AVERAGE TREATMENT EFFECT VERSION (CTD)

- Similarly,

$$E[Y_i|Z_i = 0] = E[Y_i(1)|C_i = a]\cdot \mathrm{pr}(C_i = a) + E[Y_i(0)|C_i = n]\cdot \mathrm{pr}(C_i = n) + E[Y_i(0)|C_i = c]\cdot \mathrm{pr}(C_i = c)$$

- So:

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = \mathrm{pr}(C_i = c) \times \Bigg(E[Y_i(1)|C_i = c] - E[Y_i(0)|C_i = c] \Bigg)$$

LOCAL AVERAGE TREATMENT EFFECT VERSION (CTD)

- Next,

$$E[W_i|Z_i = 1] = E[W_i(1)|Z_i = 1] = E[W_i(1)] = \mathrm{pr}(C_i = a|Z_i = 1) + \mathrm{pr}(C_i = c|Z_i = 1)$$

- Similarly

$$E[W_i|Z_i = 0] = E[W_i(0)|Z_i = 1] = E[W_i(0)] = \mathrm{pr}(C_i = a|Z_i = 1)$$

- So

$$E[W_i|Z_i = 1] - E[W_i|Z_i = 0] = \mathrm{pr}(C_i = c|Z_i = 1)$$

Combining this all:

$$\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[W_i|Z_i = 1] - E[W_i|Z_i = 0]} = E[Y_i(1) - Y_i(0)|C_i = c]$$

the local average treatment effect (Imbens & Angrist, 1994).

BOUNDS ON THE AVERAGE TREATMENT EFFECT

- Suppose in this setting we want to get the overall average effect

$$\tau = E[Y_i(1) - Y_i(0)]$$

- We can get the average effect for compliers.
- We cannot get the average effect for always takers or never takers.
- We can bound their average effects if the outcome is bounded (say binary).
- For always takers we can get

$$E[Y_i(1)|C_i = z]$$

but the data are not informative about

$$E[Y_i(0)|C_i = z]$$

beyond the fact that this expectation is between 0 and 1 for binary outcomes.

BOUNDS ON THE AVERAGE TREATMENT EFFECT

- Writing the average effect as

$$E[Y_i(1)-Y_i(0) = E[Y_i(1)-Y_i(0)|C_i = a]\text{pr}(C_i = a)+E[Y_i(1)-Y_i(0)|C_i = n]\text{pr}(C_i = n)+E[Y_i(1)-Y_i(0)|C_i = c]\text{pr}(C_i = c$$

- we can estimate all components other than

$$E[Y_i(0)|C_i = a] \qquad \text{and} \qquad E[Y_i(1)|C_i = n]$$

- Hence the lower bound on $E[Y_i(1) - Y_i(0)]$ is

$$E[Y_i(1) - 1|C_i = a]\text{pr}(C_i = a) + E[Y_i(1) - Y_i(0)|C_i = n]\text{pr}(C_i = n) + E[0 - Y_i(0)|C_i = c]\text{pr}(C_i = c)$$

and the upper bound on $E[Y_i(1) - Y_i(0)]$ is

$$E[Y_i(1) - 0|C_i = a]\text{pr}(C_i = a) + E[Y_i(1) - Y_i(0)|C_i = n]\text{pr}(C_i = n) + E[1 - Y_i(0)|C_i = c]\text{pr}(C_i = c)$$

CHANGES-IN-CHANGES MODEL (ATHEY & IMBENS, 2006)

- Suppose we have a difference-in-differences set up. We have two groups, treatment and control, and observations from two time periods, pre and post. (repeated cross-section data, not proper panel.)
- Standard DID estimator:

$$\tau = \left(\overline{Y}_{T,\mathrm{post}} - \overline{Y}_{C,\mathrm{post}}\right) - \left(\overline{Y}_{T,\mathrm{pre}} - \overline{Y}_{C,\mathrm{pre}}\right)$$

- This is functional form dependent. It assumes there is an additive fixed effect in levels,

$$Y_i(0) = \alpha_{G_i} + \delta T_i + \eta_i,$$

but not in logarithms, or some other transformation of the outcome.

CHANGES-IN-CHANGES MODEL (CTD)

- Alternative, functional form free approach: assume that in the absence of the intervention, the outcomes satisfy

$$Y_i(0) = h(U_i, T_i),$$

with $h(u, t)$ strictly increasing in u.
The distribution of U_i is allowed to vary **across** groups, but not over time **within** groups, so that a key assumption is that

$$U_i \perp T_i \Big| G_i.$$

CHANGES-IN-CHANGES MODEL (CTD)

- The standard DID model embodies three additional assumptions, namely

$$U_i - \mathbb{E}[U_i|G_i] \perp G_i \qquad \text{(additivity, only mean varies by group)}$$

$$h(u,t) = \phi(u + \delta t), \qquad \text{(single index model)}$$

for a strictly increasing function $\phi(\cdot)$, and

$$\phi(\cdot) \text{ is the identity function.} \qquad \text{(identity transformation)}$$

These three assumptions combined lead to

$$Y_{it}(0) = \alpha_{G_i} + \delta t + \eta_i,$$

(where $\alpha_g = \mathbb{E}[U_i|G_i = g]$ and $\eta_i = U_i - \mathbb{E}[U_i|G_i]$.) This gets us back to the standard DID model.

CHANGES-IN-CHANGES MODEL (CTD)

- Identification requires that we can infer the distribution of $Y(0)|G_i = 1, T_i = 1$ from the distributions of $Y_i(0)|G_i = g, T_i = t$ for $(g, t) \neq (1, 1)$. The distribution of $Y(0)|G = 1, T = 1$ is identified under these assumptions:

$$F_{Y(0),11}(y) = F_{Y(0),10}\left(F^{-1}_{Y(0),00}\left(F_{Y(0),01}(y)\right)\right).$$

where $F_{Y(0),gt}(y)$ denotes the distribution function of $Y(0)$ given $G = g$ and $T = t$.

IDENTIFICATION: TAKEAWAYS

- What makes identification questions interesting and answers compelling?
  - It sheds new light on an existing problem.
    - ★ local average treatment effect case
    - ★ bounds examples
  - It establishes identification in settings where it was unknown.
    - ★ Chamberlain's conditional logit.
    - ★ Elbers-Ridder duration example.
- Delicate interplay between assumptions that can be credible (conditional independence relations) and those that are not convincing (functional form assumptions).

REFERENCES

- Athey, Susan, and Guido W. Imbens. "Identification and inference in nonlinear difference-in-differences models." *Econometrica* 74, no. 2 (2006): 431-497.
- Bekker, Paul A. "Alternative approximations to the distributions of instrumental variable estimators." *Econometrica* (1994): 657-681.
- Elbers, Chris, and Geert Ridder. "True and spurious duration dependence: The identifiability of the proportional hazard model." *The Review of Economic Studies* 49, no. 3 (1982): 403-409.
- Heckman, James J. "Sample selection bias as a specification error." *Econometrica* (1979): 153-161.
- Heckman, James. "Varieties of selection bias." *The American Economic Review* 80, no. 2 (1990): 313-318.

REFERENCES (CTD)

- Heckman, James, and Burton Singer. "A method for minimizing the impact of distributional assumptions in econometric models for duration data." Econometrica: Journal of the Econometric Society (1984): 271-320.
- Honoré, Bo E. "Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects." *Econometrica* (1992): 533-565.
- Imbens, Guido W., and Joshua D. Angrist. "Identification and Estimation of Local Average Treatment Effects." Econometrica 62, no. 2 (1994): 467-475.
- Leamer, Edward E. "Let's take the con out of econometrics." *The American Economic Review* 73, no. 1 (1983): 31-43.
- Lewbel, Arthur. "The identification zoo: Meanings of identification in econometrics." *Journal of Economic Literature* 57, no. 4 (2019): 835-903.
- Manski, Charles F. "Nonparametric bounds on treatment effects." *The American Economic Review* 80, no. 2 (1990): 319-323.

REFERENCES (CTD)

- Staiger, Douglas, and James H. Stock. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65, no. 3 (1997): 557-586.