Identification of Parameters in Econometric Models

Introduction

- An inference problem can be separated in 1) statistical and 2) **identification component.**
- Identification involves being able to learn about the object of interest using a infinite amount of data.
- Identification should come before inference as it doesn't make sense to learn about something in finite samples if it cannot be identified in a infinite sample.
- Identification in economics relates assumptions (often derived from economic theory) and the distribution of the observable variables to the parameters of interest.
- Identification questions whether it is possible to uniquely deduce the parameter of interest from the distribution of the observed variables in the data.

Point Identification

- Suppose a parameter is used to define the structure under which the data is generated.
- Let F be the distribution of the observed data and characterized by a parameter theta:

↳ Obs. dbn $F \in \tilde{\mathcal{F}} = \{\text{Collection of } F\} = \{F_\theta : \theta \in (H)\}$

- For a **parameter to be identified** that means there exists a **unique mapping** from the distribution of observed variables to the parameter of interest.

↳ θ is identified if there exists a unique $g : \mathcal{F} \to (H)$ such that $\theta = g(F)$ for all $F \in \mathcal{F}$

- Using this definition we can argue that the moments of the observed data are identified.

Ex. $\mathcal{F} = \{F : E(y) < \infty, y \sim F\} \Rightarrow \mu$ identified since $\mu = E(y)$, where $y \sim F$ for all $F \in \mathcal{F}$

- We can also use this definition to show that a one parameter regression model is identified.

Ex. $Y_i = \beta_0 + \varepsilon_i$, $\varepsilon_i \sim iid(0, \sigma^2) \Rightarrow E(Y_i) = \beta_0$

$\Rightarrow \beta_0$ is identified

- In general we can show under certain assumptions the parameters in linear regression model are identified:

Ex. $Y = X\beta + \varepsilon$ and assume ① $E(x'\varepsilon) = 0$ and ② $E(x'x)$ is invertible $\Rightarrow \beta = [E(x'x)]^{-1} E(x'y) \Rightarrow \beta$ identified

- Note in the above regression all variables are treated as random variables rather than fixed data.
- Also in the above regression example we essentially have a set of moment conditions (implied by independence assumption) that uniquely determined our parameter of interest to prove identification. This is a common approach for arguing identification.

Point Identification Discussion

- If a parameter is not identified from the distribution of the observed data then there is no hope for consistent estimation.
- However if a parameter is identified then there may exists a consistent estimator.
- An informal definition of identification is that a parameter is identified if it is consistently estimated. This is because consistency implies if you have infinite amounts of data, then you can determine the parameter of interest. However consistent estimators are not unique.
- Suppose parameter theta cannot be differentiated from theta' from knowing the distribution of data alone. Then this set of parameters will result in the **"identified set"**:

↳ we say θ, θ' are "obs. equivalent" if $\theta, \theta' \in \mathcal{H}(F)$

F obs. dbn & $\mathcal{H}(F) = \{\theta \in \mathcal{H} : F_\theta = F\}$ is the "identified set"

- The identified set consists of all the parameter values that are consistent with the distribution of the observed data (cannot be rejected by the model).
- If a parameter is **"point identified"** than the **identified set above will be a singleton.**
- Parametric identification relates to making parametric assumptions on the distribution of observable variables to identify the parameter of interest.
- Suppose we are working with censored data, for example income in surveys is often top coded at 100,000 per year.

↳ $\underbrace{Y_i = \text{True income}}_{\text{Not Censored}}$ but $\underbrace{Y_i^{\text{observe}}}_{\text{Censored}} = \begin{cases} Y_i, & Y_i^{obs.} < 100{,}000 \\ 100{,}000, & Y_i^{obs.} \geq 100{,}000 \end{cases}$

- Note that the mean is not identified in censored data as we don't observe the entire uncensored distribution. However we can make a parametric assumption on the uncensored distribution for identification:

↳ Assume $Y_i \sim N(\mu, \sigma^2) \Rightarrow Y_i^{obs} \sim$ Truncated normal $\Rightarrow$

$E[Y_i^{obs}]$ & $V(Y_i^{obs})$ are well Known $\Rightarrow$ identify μ, σ^2 from $Y_i^{obs.}$

- The first two moments of the truncated normal distribution (2 equations and 2 unknowns) can be used to identify the mean and variance of the uncensored distribution.
- To summarize, point identification requires us to find a unique mapping from the distribution of the data to the parameter(s) of interest. We can also equivalently show the "identified set" is a singleton (for example doing a proof by contradiction). Many times identification can be proven by constructing a set of equations that uniquely determine the parameters of interest.
- Even if the parameter is not identified, its possible for a function of the parameter to identified. Binary Probit model is an example where coefficient and variance are not separately identified:

↳ $Y = I(X\beta \geq \varepsilon) \Rightarrow$ Assume $\varepsilon | X \sim N(0, \sigma^2)$ & $X'X$ is invertible. Let $\theta = (\beta, \sigma)$ and assume $F_\theta = F_{\theta'} \Rightarrow$

$\Phi_\theta\left(\frac{X\beta}{\sigma}\right) = \Phi_{\theta'}\left(\frac{X\beta'}{\sigma'}\right) \Rightarrow \frac{\beta}{\sigma} = \frac{\beta'}{\sigma'} \Rightarrow$ $g(\theta) = \frac{\beta}{\sigma}$ identified but not θ since $\theta \neq \theta'$ is possible

Partial Identification

- Identification is not an "all-or-nothing" concept. It is possible for an parametric economic model to contain information about the parameter of interest even if it is "not identified".
- For **partially identified** models the parameter of interest is not uniquely identified from the distribution of the data. However the **"identified set" is a subset of the parameter space.**
- Intuitively speaking, partial identification addresses the question of what can be learned about the parameter of interest given an infinite data set under various assumptions.
- **Partial identification** has applications in **1) measurement error, 2) missing data, 3) treatment effects.** Each of these applications is considered below.
- Let us consider the standard classical **<u>measurement error in the regressor</u>** problem:

$$Y = \beta_0 + \beta_1 X^* + \eta, \quad E(\varepsilon) = 0, \quad E(X\varepsilon) = 0,$$

X^* measured with error $\Rightarrow X = X^* + \varepsilon_X$

$$\Rightarrow Y = \beta_0 + \beta_1 X + \underbrace{(\eta - \varepsilon_X \beta_1)}_{u}, \quad Cov(X, u) \neq 0$$

$$\Rightarrow \hat{\beta}_1 \xrightarrow{p} \lambda \cdot \beta_1, \ \lambda < 1 \Rightarrow \beta_1 \text{ not identified}$$

- Although beta1 is not point identified under measurement error, it is partially identified:

① $V(Y) = \beta_1^2 V(X^*) + \sigma^2_{\varepsilon_y}$, ② $V(X) = V(X^*) + \sigma^2_{\varepsilon_x}$

③ $Cov(X, Y) = \beta_1 V(X^*) \Rightarrow$ ④ $V(Y) = \beta_1 Cov(X, Y) + \sigma^2_{\varepsilon_Y}$

<u>Lower bound</u>: Since $\sigma^2_{\varepsilon_x} \geq 0 \Rightarrow V(X) \geq V(X^*)$ ②

$\Rightarrow Cov(X, Y) \leq \beta_1 V(X) \Rightarrow \beta_1 \geq \frac{Cov(X,Y)}{V(X)}$ ③

<u>Upper bound</u>: Since $\sigma^2_{\varepsilon_y} \geq 0 \Rightarrow V(y) \geq \beta_1 Cov(X, Y)$ ④

$\Rightarrow \beta_1 \leq \frac{V(y)}{Cov(X,y)}$

↳ Identified set for $\beta_1 = \left[\frac{Cov(X,Y)}{V(X)}, \frac{V(Y)}{Cov(X,Y)}\right]$

- The **smallest bounds** that contain our parameter of interest under the assumptions are known as **"sharp"**. The **identified set is sharp** by definition because without further assumptions it cannot be made smaller.
- Let us now consider an **example with missing data** for a bounded outcome:

Let $Y_i \in [\underline{Y}, \overline{Y}]$ where $Y_i = \begin{cases} Y_i, & D_i = 1 \\ \text{unobs.}, & D_i = 0 \end{cases}$ and D_i always observed.

- Suppose our parameter of interest is the population mean of Y, but Y is not always observed:

$$\mu = E(Y_i) = E(Y_i \mid D_i = 1)\, P(D_i = 1) + P(D_i = 0)\, E(Y_i \mid D_i = 0)$$

↳ since $E(Y_i \mid D_i = 0)$ unobs. $\Rightarrow$ μ not identified

Lower bound Let $p = P(D_i = 1)$

$$\text{Since } \begin{cases} Y_i \geq \underline{Y} \Rightarrow \mu \geq E(Y_i \mid D_i = 1)\, p + (1-p)\, \underline{Y} \\ Y_i \leq \overline{Y} \Rightarrow \mu \leq E(Y_i \mid D_i = 1)\, p + (1-p)\, \overline{Y} \end{cases}$$

$$\Rightarrow \mu \in \left[E(Y_i \mid D_i = 1)\, p + (1-p)\, \underline{Y},\ E(Y_i \mid D_i = 1)\, p + (1-p)\, \overline{Y} \right]$$

- The key assumption that allowed us to construct the identified set for the population mean is that Y is bounded. Otherwise the identified set would be the entire real line.
- If the **data is missing at random** than the above mean is **point identified** as E(Y) = E(Y|D = 1).
- Finally let us consider the application of **partial identification to treatment effects**:

$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$, (Y_{1i}, Y_{0i}) are potential outcomes, $D_i \in \{0, 1\}$

$$ATE = E(Y_{1i} - Y_{0i}) = E(Y_{1i} \mid D_i = 1)\, p + (1-p)\, E(Y_{1i} \mid D_i = 0) - E(Y_{0i} \mid D_i = 1)\, p - (1-p)\, E(Y_{0i} \mid D_i = 0)$$

- Suppose the **treatment participation** (represented by Di = 1 for treatment group, Di = 0 for control group) is **randomly assigned**:

$$\text{If } D_i \perp (Y_{0i}, Y_{1i}) \Rightarrow E(Y_{1i} \mid D_i = 0) = E(Y_{0i} \mid D_i = 1)$$

$$\Rightarrow ATE = E(Y_{1i} \mid D_i = 1)\, p + E(Y_{0i} \mid D_i = 0) \Rightarrow \text{ATE is identified}$$

- The **ATE is identified** (computed from observed quantitates) if treatment participation is randomized. Suppose **Di not randomized** but we have **bounded potential outcomes**:

Suppose $Y_{0i} \in [\underline{Y}^0, \overline{Y}_0]$, $Y_{1i} \in [\underline{Y}^1, \overline{Y}_1] \Rightarrow$

$$\begin{cases} E[Y_{1i}] = E[Y_{1i} \mid D_i = 1]p + E[Y_{1i} \mid D_i = 0](1-p) \\ E[Y_{0i}] = E[Y_{0i} \mid D_i = 1]p + E[Y_{0i} \mid D_i = 0](1-p) \end{cases}$$

$\Rightarrow$ ATE is not identified since $E[Y_{1i} \mid D_i = 0]$ & $E[Y_{0i} \mid D_i = 1]$ are not observed

- Let us use the fact that the **potential outcomes are bounded** to partially identify the ATE:

Since $Y_{1i} \in [\underline{Y}^1, \overline{Y}_1] \Rightarrow E[Y_{1i}] \in [\underline{Y}^1, \overline{Y}_1] \Rightarrow$

$$E(Y_{1i}) \in \left[E[Y_{1i} \mid D_i = 1]p + \underline{Y}^1(1-p),\ E[Y_{1i} \mid D_i = 1]p + \overline{Y}_1(1-p)\right]$$

Since $Y_{0i} \in [\underline{Y}^0, \overline{Y}_0] \Rightarrow E[Y_{0i}] \in [\underline{Y}^0, \overline{Y}_0] \Rightarrow$

$$E(Y_{0i}) \in \left[E[Y_{0i} \mid D_i = 0](1-p) + p\underline{Y}^0,\ E[Y_{0i} \mid D_i = 0](1-p) + p\overline{Y}_0\right]$$

↳ Note bounds are informative if $p < 1$

↳ $ATE = E[Y_{1i} - Y_{0i}] \in [C_1, C_2]$

- ATE with **binary outcome** has bounds [-1,1] by definition. It is straightforward to bound the ATE to 50% of its range using the above inequalities:

$$\begin{cases} ATE \geq E[Y_{1i} \mid D_i = 1]p - E[Y_{0i} \mid D_i = 0](1-p) - p = C_1 \\ ATE \leq E[Y_{1i} \mid D_i = 1]p - E[Y_{0i} \mid D_i = 0](1-p) + (1-p) = C_2 \end{cases}$$

↳ $C_2 - C_1 = 1$

- Instruments, monotonicity assumption, and covariates can be used to make bounds more precise. Let us consider the monotonicity assumption:

monotonicity: $D_i = 1 \iff Y_{1i} \geq Y_{0i}$ for all i

- **Monotonicity assumption** as written above says the potential outcome under treatment is at least as great as under the control. If the treatment is after school tutoring, than monotonicity says that no student should be worse off from receiving tutoring from this program.

$$E[Y_{1i} \mid D_i = 0] \underset{\text{mono.}}{=} E[Y_{1i} \mid Y_{1i} < Y_{0i}] \leq E[Y_{1i} \mid Y_{1i} \geq Y_{0i}] = E[Y_i \mid D_i = 1]$$

$$\Rightarrow \text{similarly} \quad E[Y_{0i} \mid D_i = 1] \leq E[Y_i \mid D_i = 0]$$

$$\Rightarrow \begin{cases} E(Y_{1i}) \leq E(Y_{1i} \mid D_i = 1) \\ E(Y_{0i}) \leq E(Y_{0i} \mid D_i = 0) \end{cases}$$

- We can use the above conditions (implied from monotonicity) and the general bounds we derived before for the ATE under binary outcome to **get smaller bounds**:

$$\begin{cases} ATE \geq E(Y_{1i} \mid D_i = 1)p - E(Y_{0i} \mid D_i = 0) = c_1 \\ ATE \leq E(Y_{1i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 0)(1-p) = c_2 \end{cases}$$

$$\hookrightarrow c_2 - c_1 = (1-p)E(Y_{1i} \mid D_i = 1) + pE(Y_{0i} \mid D_i = 0) \leq 1$$

- Now suppose we have an valid **binary instrumental variable** for treatment participation, we can use this additional information to **tighten the bounds** even further.

↳ let Z_i be a IV, $Z_i \in \{0, 1\}$

- Now the treatment participation can be written in terms of potential treatments:

$$D_i = Z_i \cdot D_{1i} + (1 - Z_i) D_{0i}, \quad (D_{0i}, D_{1i}) \text{ are potential treatments}$$

- Assume the standard assumptions for the instrument holds and that there are **no defiers**:

i) $Cov(Z_i, D_i) \neq 0$ (Relevance)

ii) $Z_i \perp (Y_{0i}, Y_{1i}, D_{0i}, D_{1i})$ (Z_i randomized)

iii) $D_{1i} \geq D_{0i}$ (no defiers / monotonicity)

- Note the no defiers or monotonicity assumption has to hold for all units of interest.

- Recall that all units can be categorized as always takers (AT), never takers (NT), compliers (C).
- E(Y1) and E(Y0) depend on AT, NT, and C. The bounds for the unidentified components are:

$$\text{Since } Y \in \{0,1\} \Rightarrow \begin{cases} E(Y_{1i} \mid D_{1i} = D_{0i} = 0) = E(Y_{1i} \mid NT) \in [0,1] \\ E(Y_{1i} \mid D_{1i} = D_{0i} = 1) = E(Y_{0i} \mid AT) \in [0,1] \end{cases}$$

- With a bit of algebra we can show that under all the above assumptions:

$$\begin{cases} ATE \geq E(Y_i D_i \mid Z_i = 1) - E(Y_i (1 - D_i) \mid Z_i = 0) - E(D_i \mid Z_i = 0) = c_1 \\ ATE \leq E(Y_i D_i \mid Z_i = 1) - E(Y_i (1 - D_i) \mid Z_i = 0) + E((1 - D_i) \mid Z_i = 1) = c_2 \end{cases}$$

$$\hookrightarrow c_2 - c_1 = 1 - \underbrace{[E(D_i \mid Z_i = 1) - E(D_i \mid Z_i = 0)]}_{\text{Prop. of compliers to } Z} < 1$$

- The **bounds tighten in the proportion of compliers present**. Note if proportion of compliers = 1 then D = Z, since Z randomizes implies D randomizes hence ATE is point identified.

Partial Identification Discussion

- Partial identification is about finding a set of values which are consistent with the distribution of the data (the values in the set cannot be rejected as being the true parameter of interest) under the given modelling assumptions.
- For all values (denoted by theta') in the identified set we can find a distribution that is parametrized by theta' and is consistent with the distribution of the observed data.
- Characterizing the identified set is a challenge of partial identification.
- If the identified set equals the entire parameter space than usually we say the parameter is unidentified. However this can be thought of as a special case of partial identification.
- Treatment effect bounds containing 0 are not informative as they don't identify the direction of treatment.
- Imposing **more structure and assumptions** generally **decreases the size of the identified set**. With the right assumptions the identified set shrinks to a singleton and we get point identification.

Conclusion

- The identified set represents a class of econometric models that are consistent with the data. The cardinality of this set can be used to determine whether the parameter of interest are 1) point identified, 2) partially identified, or 3) unidentified.
- Partial identification is when the identified set is a subset of the parameter space. Whereas for point identification this set reduces to a singleton.
- If a parameter is point identified than that means there exists a unique mapping from the distribution of the observed data to the parameter of interest.