"New Developments in Econometrics"

Lecture 10

Partial Identification

Guido Imbens

Cemmap Lectures, UCL, June 2009

Outline

1. Introduction
2. Example I: Missing Data
3. Example II: Returns to Schooling
4. Example III: Initial Conditions Problems in Panel Data
5. Example IV: Auction Data
6. Example V: Entry Models
7. Estimation and Inference

1. Introduction

Traditionally in constructing statistical or econometric models researchers look for models that are *(point-)identified*: given a large (infinite) data set, one can infer without uncertainty what the values are of the objects of interest.

It would appear that a model where we cannot learn the parameter values even in infinitely large samples would not be very useful.

However, it turns out that even in cases where we cannot learn the value of the estimand *exactly* in large samples, in many cases we can still learn a fair amount, even in finite samples. A research agenda initiated by Manski has taken this perspective.

Here we discuss a number of examples to show how this approach can lead to interesting answers in settings where previously were viewed as intractable.

We also discuss some results on inference.

1. Are we interested in confidence sets for parameters or for identified sets?

2. Concern about uniformity of inferences (confidence cant be better in partially identified case than in point-identified case).

2. I: Missing Data

If $D_i = 1$, we observe Y_i, and if $D_i = 0$ we do not observe Y_i. We always observe the missing data indicator D_i. We assume the quantity of interest is the population mean $\theta = \mathbb{E}[Y_i]$.

In large samples we can learn $p = \mathbb{E}[D_i]$ and $\mu_1 = \mathbb{E}[Y_i|D_i = 1]$, but nothing about $\mu_0 = \mathbb{E}[Y_i|D_i = 0]$. We can write:

$$\theta = p \cdot \mu_1 + (1 - p) \cdot \mu_0.$$

Since even in large samples we learn nothing about μ_0, it follows that without additional information there is no limit on the range of possible values for θ.

Even if p is very close to 1, the small probability that $D_i = 0$ combined with the possibility that μ_0 is very large or very small allows for a wide range of values for θ.

Now suppose we know that the variable of interest is binary: $Y_i \in \{0, 1\}$. Then natural (not data-informed) lower and upper bounds for μ_0 are 0 and 1 respectively. This implies bounds on θ:

$$\theta \in [\theta_{\mathsf{LB}}, \theta_{\mathsf{UB}}] = [p \cdot \mu_1, p \cdot \mu_1 + (1 - p)] \,.$$

These bounds are *sharp*, in the sense that without additional information we can not improve on them.

Formally, for all values θ in $[\theta_{\mathsf{LB}}, \theta_{\mathsf{UB}}]$, we can find a joint distribution of (Y_i, W_i) that is consistent with the joint distribution of the observed data and with θ.

We can also obtain informative bounds if we modify the object of interest a little bit.

Suppose we are interested in the median of Y_i, $\theta_{0.5} = \text{med}(Y_i)$.

Define $q_\tau(Y_i)$ to be the τ quantile of the conditional distribution of Y_i given $D_i = 1$. Then the median cannot be larger than $q_{1/(2p)}(Y_i)$ because even if all the missing values were large, we know that at least $p \cdot (1/(2p)) = 1/2$ of the units have a value less than or equal to $q_{1/(2p)}(Y_i)$.

Then, if $p > 1/2$, we can infer that the median must satisfy

$$\theta_{0.5} \in [\theta_{\text{LB}}, \theta_{\text{UB}}] = \left[q_{(2p-1)/(2p)}(Y_i), q_{1/(2p)}(Y_i)\right],$$

and we end up with a well defined, and, depending on the data, more or less informative identified interval for the median.

If fewer than 50% of the values are observed, or $p < 1/2$, then we cannot learn anything about the median of Y_i without additional information (for example, a bound on the values of Y_i), and the interval is $(-\infty, \infty)$.

More generally, we can obtain bounds on the τ quantile of the distribution of Y_i, equal to

$$\theta_\tau \in [\theta_{\text{LB}}, \theta_{\text{UB}}] = \left[q_{(\tau-(1-p))/p}(Y_i|D_i = 1), q_{\tau/p}(Y_i|D_i = 1) \right].$$

which is bounded if the probability of Y_i being missing is less than $\min(\tau, 1-\tau)$.

3. Example II: Returns to Schooling

Manski-Pepper are interested in estimating returns to schooling. They start with an individual level response function $Y_i(w)$.

$$\Delta(s,t) = \mathbb{E}[Y_i(t) - Y_i(s)],$$

is the difference in average outcomes (log earnings) given t rather than s years of schooling. Values of $\Delta(s,t)$ are the object of interest.

W_i is the actual years of school, and $Y_i = Y_i(W_i)$ be the actual log earnings.

If one makes an unconfoundedness/exogeneity assumption that

$$Y_i(w) \perp\!\!\!\perp W_i \mid X_i,$$

for some set of covariates, one can estimate $\Delta(s,t)$ consistently given some support conditions. MP relax this assumption.

Alternative Assumptions considered by MP

Increasing education does not lower earnings:

Assumption 1 (Monotone Treatment Response)
If $w' \geq w$, *then* $Y_i(w') \geq Y_i(w)$.

On average, individuals who choose higher levels of education would have higher earnings at each level of education than individuals who choose lower levels of education.

Assumption 2 (Monotone Treatment Selection)
If $w'' \geq w'$, *then for all* w, $\mathbb{E}[Y_i(w)|W_i = w''] \geq \mathbb{E}[Y_i(w)|W_i = w']$.

Under these two assumptions, bound on $\mathbb{E}[Y_i(w)]$ and $\Delta(s,t)$:

$$\mathbb{E}[Y_i|W_i = w] \cdot \Pr(W_i \geq w) + \sum_{v<w} \mathbb{E}[Y_i|W_i = v] \cdot \Pr(W_i = v)$$

$$\leq \mathbb{E}[Y_i(w)] \leq$$

$$\mathbb{E}[Y_i|W_i = w] \cdot \Pr(W_i \leq w) + \sum_{v>w} \mathbb{E}[Y_i|W_i = v] \cdot \Pr(W_i = v).$$

Using NLS data MP estimate the upper bound on the the returns to four years of college, $\Delta(12,16)$ to be 0.397.

Translated into average yearly returns this gives us 0.099, which is in fact lower than some estimates that have been reported in the literature.

This analysis suggests that the upper bound is in this case reasonably informative, given a remarkably weaker set of assumptions.

4. Example III: Initial Conditions Problems in Panel Data (Honoré and Tamer)

$$Y_{it} = 1\{X_{it}'\beta + Y_{it-1} \cdot \gamma + \alpha_i + \epsilon_{it} \geq 0\},$$

with the ϵ_{it} independent $\mathcal{N}(0,1)$ over time and individuals. Focus on γ.

Suppose we also postulate a parametric model for the random effects α_i:

$$\alpha | X_{i1}, \ldots, X_{iT} \sim G(\alpha|\theta)$$

Then the model is almost complete.

All that is missing is:

$$p(Y_{i1}|\alpha_i, X_{i1}, \ldots, X_{iT}).$$

HT assume a discrete distribution for α, with a finite and known set of support points. They fix the support to be $-3, -2.8, \ldots, 2.8, 3$, with unknown probabilities.

In the case with $T = 3$ they find that the range of values for γ consistent with the data generating process (the identified set) is very narrow.

If γ is in fact equal to zero, the width of the set is zero. If the true value is $\gamma = 1$, then the width of the interval is approximately 0.1. (It is largest for γ close to, but not equal to, -1.) See Figure 1, taken from HT.

The HT analysis shows nicely the power of the partial identification approach: A problem that had been viewed as essentially intractable, with many non-identification results, was shown to admit potentially precise inferences. Point identification is not a big issue here.

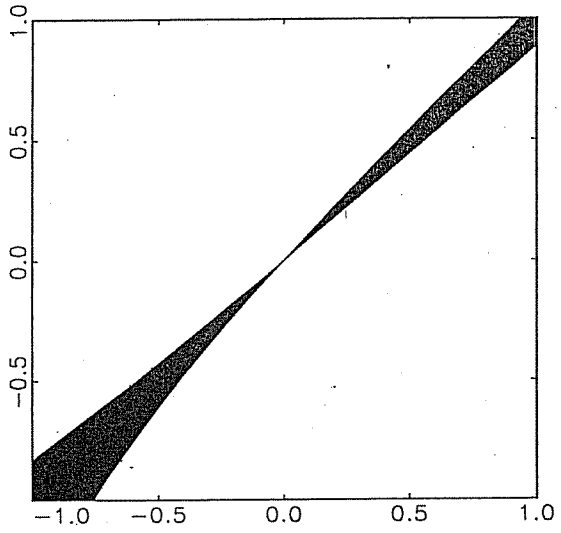


FIGURE 1.—Identified region for γ as a function of its true value.

5. Example IV: Auction Data

Haile and Tamer study English or oral ascending bid auctions. In such auctions bidders offer increasingly higher prices until only one bidder remains. HT focus on a symmetric independent private values model. In auction t, bidder i has a value ν_{it}, drawn independently from the value for bidder j, with cdf $F_\nu(v)$

HT are interested in the value distribution $F_\nu(v)$. This is assumed to be the same in each auction (after adjusting for observable auction characteristics).

One can imagine observing exactly when each bidder leaves the auction, thus directly observing their valuations. This is not what is typically observed. For each bidder we do not know at any point in time whether they are still participating unless they subsequently make a higher bid.

Haile-Tamer Assumptions

Assumption 3 *No bidder ever bids more than their valuation*

Assumption 4 *No bidder will walk away and let another bidder win the auction if the winning bid is lower than their own valuation*

Upper Bound on Value Distribution

Let the highest bid for participant i in auction t be b_{it}. We ignore variation in number of bidders per auction, and presence of covariates.

Let $F_b(b) = \Pr(b_{it} \leq b)$ be the distribution function of the bids (ignoring variation in the number of bidders by auction). This distribution can be estimated because the bids are observed.

Because no bidder ever bids more than their value, it follows that $b_{it} \leq \nu_{it}$. Hence, without additional assumptions,

$$F_\nu(v) \leq F_b(v), \quad \text{for all } v.$$

Lower Bound on Value Distribution

The second highest of the values among the n participants in auction t must be less than or equal to the winning bid. This follows from the assumption that no participant will let someone else win with a bid below their valuation.

Let $F_{\nu,m:n}(v)$ denote the mth order statistic in a random sample of size n from the value distribution, and let $F_{B,n:n}(b)$ denote the distribution of the winning bid in auctions with n participants. Then

$$F_{B,n:n}(v) \leq F_{\nu,n-1:n}(v).$$

The distribution of the any order statistic is monotonically related to the distribution of the parent distribution, and so a lower bound on $F_{\nu,n-1:n}(v)$ implies a lower bound on $F_{\nu}(v)$.

6. Example V: Entry Models (Cilberto & Tamer)

Suppose two firms, A and B, contest a set of markets. In market m, $m = 1, \ldots, M$, the profits for firms A and B are

$$\pi_{Am} = \alpha_A + \delta_A \cdot d_{Bm} + \varepsilon_{Am},\ \pi_{Bm} = \alpha_B + \delta_B \cdot d_{Am} + \varepsilon_{Bm}.$$

where $d_{Fm} = 1$ if firm F is present in market m, for $F \in \{A, B\}$, and zero otherwise.

Decisions assuming complete information satisfy Nash equilibrium condition

$$d_{Am} = 1\{\pi_{Am} \geq 0\}, \quad d_{Bm} = 1\{\pi_{Bm} \geq 0\}.$$

Incomplete Model

For pairs of values $(\varepsilon_{Am}, \varepsilon_{Bm})$ such that

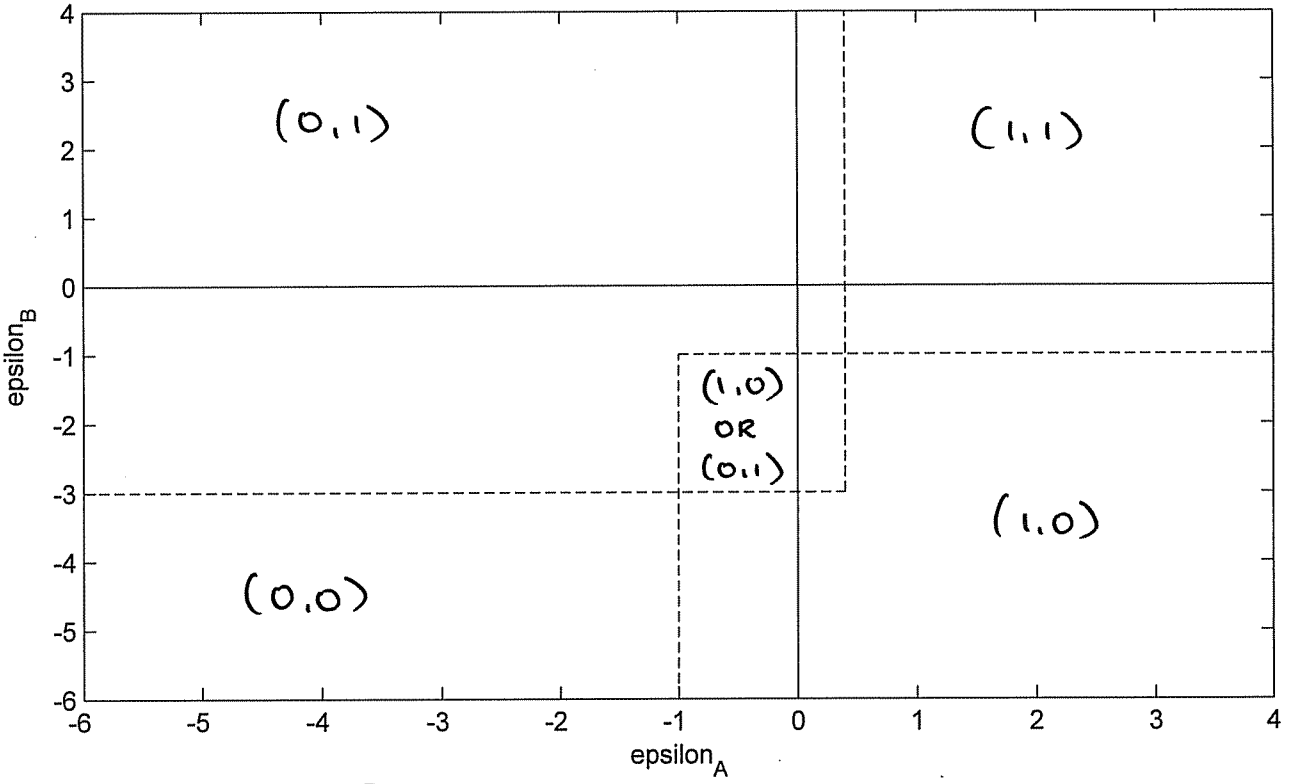
$$-\alpha_A < \varepsilon_A \leq -\alpha_A - \delta_A, \quad -\alpha_B < \varepsilon_B \leq -\alpha_B - \delta_B,$$

both $(d_A, d_B) = (0,1)$ and $(d_A, d_B) = (1,0)$ satisfy the profit maximization condition.

In the terminology of this literature, the model is *incomplete*. It does not specify the outcomes given the inputs. Missing is an equilibrium selection mechanism, which is typically difficult to justify.

Figure 1, adapted from CM, shows the different regions in the $(\varepsilon_{Am}, \varepsilon_{Bm})$ space.

Figure 1 (d_A, d_B)



$\alpha_A = 1$ $\delta_A = -1.4$

$\alpha_B = 3$ $\delta_B = -2$

Implication: Inequality Conditions

The implication of this is that the probability of the outcome $(d_{Am}, d_{Bm}) = (0,1)$ cannot be written as a function of the parameters of the model, $\theta = (\alpha_A, \delta_A, \alpha_B, \delta_B)$, even given distributional assumptions on $(\varepsilon_{Am}, \varepsilon_{Bm})$.

Instead the model implies a lower and upper bound on this probability:

$$H_{L,01}(\theta) \leq \Pr\left((d_{Am}, d_{Bm}) = (0,1)\right) \leq H_{U,01}(\theta).$$

Thus in general we can write the information about the parameters in large samples as

$$\begin{pmatrix} H_{L,00}(\theta) \\ H_{L,01}(\theta) \\ H_{L,10}(\theta) \\ H_{L,11}(\theta) \end{pmatrix} \leq \begin{pmatrix} \Pr\left((d_{Am}, d_{Bm}) = (0,0)\right) \\ \Pr\left((d_{Am}, d_{Bm}) = (0,1)\right) \\ \Pr\left((d_{Am}, d_{Bm}) = (1,0)\right) \\ \Pr\left((d_{Am}, d_{Bm}) = (1,1)\right) \end{pmatrix} \leq \begin{pmatrix} H_{U,00}(\theta) \\ H_{U,01}(\theta) \\ H_{U,11}(\theta) \\ H_{U,11}(\theta) \end{pmatrix}.$$

7.A Estimation

Chernozhukov, Hong, and Tamer study Generalized Inequality Restriction (GIR) setting:

$$\mathbb{E}[\psi(Z, \theta)] \geq 0,$$

where $\psi(z, \theta)$ is known. Fits CT entry example

Define for a vector x the vector $(x)_+$ to be the component-wise non-negative part, and $(x)_-$ to be the component-wise non-positive part, so that for all x, $x = (x)_- + (x)_+$.

For a given $M \times M$ non-negative definite weight matrix W, CHT consider the population objective function

$$Q(\theta) = \mathbb{E}[\psi(Z, \theta)]'_{-} W \mathbb{E}[\psi(Z, \theta)]_{-}.$$

For all $\theta \in \Theta_I$, we have $Q(\theta) = 0$, and for $\theta \notin \Theta_I$, we have $Q(\theta) > 0$

The sample equivalent to this population objective function is

$$Q_N(\theta) = \left(\frac{1}{N} \sum_{i=1}^{N} \psi(Z_i, \theta) \right)'_{-} W \left(\frac{1}{N} \sum_{i=1}^{N} \psi(Z_i, \theta) \right)_{-}.$$

We cannot simply estimate the identified set as

$$\tilde{\Theta}_I = \{\theta \in \Theta \,|Q_N(\theta) = 0\}\,,$$

The reason is that even for θ in the identified set $Q_N(\theta)$ may be positive with high probability, and $\tilde{\Theta}_I$ can be empty when Θ_I is not, even in large samples.

A simple way to see that is to consider the standard GMM case with equalities and over-identification. If $\mathbb{E}[\psi(Z,\theta)] = 0$, the objective function will not be zero in finite samples in the case with over-identification.

This is the reason CHT suggest estimating the set Θ_I as

$$\hat{\Theta}_I = \{\theta \in \Theta \,|Q_N(\theta) \leq a_N\}\,,$$

where $a_N \to 0$ at the appropriate rate.

7.B Inference

Fast growing literature, Beresteanu and Molinari (2006), Chernozhukov, Hong, and Tamer (2007), Galichon and Henry (2006), Imbens and Manski (2004), Rosen (2006), and Romano and Shaikh (2007ab).

First issue: do we want a confidence set that includes each element of the identified set with fixed probability, or the entire identified set with that probability. First

$$\inf_{\theta \in [\theta_{\mathsf{LB}}, \theta_{\mathsf{UB}}]} \Pr\left(\theta \in \mathsf{CI}_{\alpha}^{\theta}\right) \geq \alpha.$$

Second

$$\Pr\left([\theta_{\mathsf{LB}}, \theta_{\mathsf{UB}}] \subset \mathsf{CI}_{\alpha}^{[\theta_{\mathsf{LB}}, \theta_{\mathsf{UB}}]}\right) \geq \alpha.$$

The second requirement is stronger than the first, and so generally $\mathsf{CI}_{\alpha}^{\theta} \subset \mathsf{CI}_{\alpha}^{[\theta_{\mathsf{LB}}, \theta_{\mathsf{UB}}]}$.

7.B.I Well behaved Estimators for Bounds

Missing data example, (p, prob of missing data, known). Identified set:

$$\Theta_I = [p \cdot \mu_1, p \cdot \mu_1 + (1 - p)].$$

Standard interval for μ_1:

$$\text{CI}_\alpha^{\mu_1} = \left[\overline{Y} - 1.96 \cdot \sigma/\sqrt{N}_1, \overline{Y} + 1.96 \cdot \sigma/\sqrt{N}_1\right].$$

Three ways to construct 95% confidence intervals for θ.

$$\text{CI}_\alpha^\theta = \left[p \cdot \left(\overline{Y} - 1.96 \cdot \sigma/\sqrt{N_1}\right), p \cdot \left(\overline{Y} + 1.96 \cdot \sigma/\sqrt{N_1}\right) + 1 - p\right].$$

This is conservative. For each θ in the interior of Θ_I, the cov rate is 1. For $\theta \in \{\theta_{\text{LB}}, \theta_{\text{UB}}\}$, if $p < 1$, the cov rate is 0.975.

$$\text{CI}_\alpha^\theta = \left[p \cdot \left(\overline{Y} - 1.645 \cdot \sigma/\sqrt{N_1}\right), p \cdot \left(\overline{Y} + 1.645 \cdot \sigma/\sqrt{N_1}\right) + 1 - p\right].$$

This has the problem that if $p = 1$ (when θ is point-identified), the coverage is only 0.90. Imbens and Manski (2004) suggest modifying the confidence interval to

$$\text{CI}_\alpha^\theta = \left[p \cdot \left(\overline{Y} - C_N \cdot \sigma/\sqrt{N_1}\right), p \cdot \left(\overline{Y} + C_N \cdot \sigma/\sqrt{N_1}\right) + 1 - p\right],$$

where the critical value C_N satisfies

$$\Phi\left(C_N + \sqrt{N} \cdot \frac{1-p}{\sigma/\sqrt{p}}\right) - \Phi\left(-C_N\right) = 0.95$$

This confidence interval has asymptotic coverage 0.95, uniformly over p, for $p \in [p_0, 1]$.

7.B.II Irregular Estimators for Bounds

Simple example of Generalized Inequality Restrictions (GIR) set up.

$$\mathbb{E}[X] \geq \theta, \quad \text{and} \quad \mathbb{E}[Y] \geq \theta.$$

The parameter space is $\Theta = [0, \infty)$. Let $\mu_X = \mathbb{E}[X]$, and $\mu_Y = \mathbb{E}[Y]$. We have a random sample of size N of the pairs (X, Y). The identified set is

$$\Theta_I = [0, \min(\mu_X, \mu_Y)].$$

A naive 95% confidence interval would be

$$C_\alpha^\theta = [0, \min(\overline{X}, \overline{Y}) + 1.645 \cdot \sigma/N].$$

This confidence interval essentially ignores the moment inequality that is not binding in the sample. It has pointwise asymptotic 95% coverage for all values of μ_X, μ_Y, as long as $\min(\mu_X, \mu_Y) > 0$, and $\mu_X \neq \mu_Y$.

The first condition ($\min(\mu_X, \mu_Y) > 0$) is the same as the condition in the Imbens-Manski example. It can be dealt with in the same way by adjusting the critical value slightly based on an initial estimate of the width of the identified set.

The naive confidence interval essentially assumes that the researcher knows which moment conditions are binding. This is true in large samples, unless there is a tie.

However, in finite samples ignoring uncertainty regarding the set of binding moment inequalities may lead to a poor approximation, especially if there are many inequalities. One possibility is to construct conservative confidence intervals (e.g., Pakes, Porter, Ho, and Ishii, 2007). However, such intervals can be unnecessarily conservative if there are moment inequalities that are far from binding.

One would like construct confidence intervals that asymptotically ignore irrelevant inequalities, and at the same time are valid uniformly over the parameter space. Subsampling (but not bootstrapping) appears to work theoretically. See Romano and Shaikh (2007a), and Andrews and Guggenberger (2007). Little is known about finite sample properties in realistic settings.