Identification of Unobservables in Observations

Yingyao Hu

Department of Economics
Johns Hopkins University

February 15, 2023

Latent variables in microeconomic models

empirical models	unobservables	observables
measurement error	true earnings	self-reported earnings
consumption function	permanent income	observed income
production function	productivity	output, input
wage function	ability	test scores
learning model	belief	choices, proxy
auction model	unobserved heterogeneity	bids
contract model	effort, type	outcome, state var.
...	...	...

Identification in observation

- Most identification results focus on parameters in a model, which includes latent variables X^*
- Can we pin down values of X^* in each observation?
- Why this is interesting?
  - imputation
  - consumer unobserved heterogeneity: Marketing
  - measurement error correction: true GDP
- New techniques make it possible: Deep neural networks
- Identification at the population level. (not at the sample level)

Continuity

- For a variable with a distinct value in each observation in a sample, researchers usually consider it as a continuous variable in the population
- Such continuity only exists in assumptions given the discrete nature of a sample.
- It is observationally equivalent to assume that the population is a collection of a large but finite number of elements.

"No two leaves are alike."

- Each leaf i has observed traits x_i and unobserved heterogeneity x_i^*.

$$(x_i, x_i^*)$$

- only x_i is observed, we want to pin down x_i^*

Definition

A population $\mathcal{P}_{X,X^*}$ satisfies **the property of leaves** if it is a collection of ordered pairs (x_i, x_i^*) for $i = 1, 2, ..., N; N < \infty$ such that $x_i \neq x_j$ for any $i \neq j$. That is

$$\mathcal{P}_{X,X^*} = \{(x_i, x_i^*) : x_i \neq x_j \textit{ for } i \neq j \textit{ and } i, j = 1, 2, ..., N.\} \tag{1}$$

Population and distribution

- F_{X,X^*} denote the cumulative distribution function of random variables (X, X^*) randomly drawn from population $\mathcal{P}_{X,X^*}$ with probability
$$Pr(\{(X, X^*) = (x_i, x_i^*)\}) = p_i > 0 \tag{2}$$
with $\sum_{i=1}^{N} p_i = 1$.
- The population of observed traits x is
$$\mathcal{P}_X = \{x_i : (x_i, x_i^*) \in \mathcal{P}_{X,X^*} \; for\, some\, x_i^*\} \tag{3}$$
with a distribution function F_X. In fact, its probability function is
$$Pr(\{X = x_i\}) = p_i \tag{4}$$
because x_i is distinct for all $i = 1, 2, ..., N$, i.e., in the whole population.
-
$$Pr(\{X^* = x_i^*\}|\{X = x_i\}) = 1. \tag{5}$$

Identification in observation

Theorem

Suppose that Conditions 1 and 2 hold as follows:
– Population $\mathcal{P}_{X,X^*}$, with distribution function F_{X,X^*}, satisfies the property of leaves in Equations (1)
– F_X uniquely determines F_{X,X^*}, where distribution function F_X corresponds to population $\mathcal{P}_X$ in Equations (3).
Then, $\mathcal{P}_X$ and F_X uniquely determine $\mathcal{P}_{X,X^*}$ and F_{X,X^*}, i.e., each x_i in $\mathcal{P}_X$ uniquely determines its corresponding x_i^* through $\mathcal{P}_{X,X^*}$.

-

Examples

- linear regression: $X = (Y, Z)$ and $X^* = e$

$$Y = Z\beta + e$$

Solution:

$$e = Y - Z \times E(Z'Z)^{-1}E(Z'Y)$$

$F_{Y,Z}$ uniquely determines $F_{Y,Z,e}$

- nonseparable model: $X = (Y, Z)$ and $X^* = U$

$$Y = h(Z, U)$$

Solution:

$$U = F^{-1}_{Y|Z}(Y|Z)$$

$F_{Y,Z}$ uniquely determines $F_{Y,Z,U}$

From F_X to F_{X,X^*}

- A key assumption is that F_X uniquely determines F_{X,X^*}
- Here are two examples:
  - Kotlarski (1966): $X = (X_1, X_2)$
  - Hu (2008): $X = (X_1, X_2, X_3)$

2-measurement model: Kotlarski's identity

- $X = (X_1, X_2)$ satisfies the simplest factor model

$$\begin{aligned} X_1 &= X^* + \eta \\ X_2 &= X^* + \varepsilon \end{aligned}$$

- distribution function & characteristic function of X^* $(i = \sqrt{-1})$

$$f_{X^*}(x^*) = \frac{1}{2\pi} \int e^{-ix^* t} \Phi_{X^*}(t) dt \qquad \Phi_{X^*} = E\left[e^{itX^*}\right]$$

- Kotlarski's identity (1966)

$$\Phi_{X^*}(t) = \exp\left[\int_0^t \frac{iE\left[X_1 e^{isX_2}\right]}{Ee^{isX_2}} ds\right]$$

- latent distribution f_{X^*} is uniquely determined by observed distribution f_{X_1,X_2} with a closed form. Thus, F_X uniquely determines F_{X,X^*}

Illustration

Table: An illustration of identification in observations

observation	observables		unobservables			probab
i	$X_1 = X^* + \epsilon_1$	$X_2 = X^* + \epsilon_2$	ϵ_1	X^*	ϵ_2	p_i
1	0	0	-1	1	-1	$f_{X_1,X_2}(0,0) = f_{\epsilon_1}(-$
2	0	1	-1	1	0	$f_{X_1,X_2}(0,1) = f_{\epsilon_1}(-$
3	0	2	-1	1	1	...
4	-1	-1	-1	0	-1	...
5	-1	0	-1	0	0	...
6	-1	1	-1	0	1	...
7	3	0	2	1	-1	...
8	3	1	2	1	0	...
9	3	2	2	1	1	...
10	2	-1	2	0	-1	...
11	2	0	2	0	0	...
12	2	1	2	0	1	...

Illustration

Table: A second example

observation	observables		unobservables			prob.
i	$X_1 = X^* + \epsilon_1$	$X_2 = X^* + \epsilon_2$	ϵ_1	X^*	ϵ_2	
1	0	-0.5	-1	1	-1.5	$f_{X_1,X_2}(0,-0.5) = f_\epsilon$
2	0	1.5	-1	1	0.5	$f_{X_1,X_2}(0,1.5) = f_\epsilon$
3	0	2	-1	1	1	$f_{X_1,X_2}(0,2) = f_\epsilon$
4	-1	-1.5	-1	0	-1.5	.
5	-1	0.5	-1	0	0.5	.
6	-1	1	-1	0	1	.
7	1	-0.5	0	1	-1.5	.
8	1	1.5	0	1	0.5	.
9	1	2	0	1	1	.
10	0	-1.5	0	0	-1.5	$f_{X_1,X_2}(0,-1,5) =$
11	0	0.5	0	0	0.5	$f_{X_1,X_2}(0,0.5) =$
12	0	1	0	0	1	$f_{X_1,X_2}(0,1) =$

Hu (2008)

- definition of 3-measurement model:
  $X = (X_1, X_2, X_3)$ satisfies

$$X_1 \perp X_2 \perp X_3 \mid X^*$$

- for y

$$f_{X_1,X_2,X_3}(x, y, z) = \sum_{x^* \in \mathcal{X}^*} f_{X_1|X^*}(x|x^*) f_{X_2|X^*}(y|x^*) f_{X_3|X^*}(z|x^*) f_{X^*}(x^*)$$

$$f_{X_1,X_2,X_3} \text{ uniquely determines } f_{X_1,X_2,X_3,X^*}$$
$$f_{X_1,X_2,X_3,X^*} = f_{X_1|X^*} f_{X_2|X^*} f_{X_3|X^*} f_{X^*}$$

- A global nonparametric point identification
- And identification in observation

Illustration

Table: An illustration of identification in observations

observation i	observables X_1	X_2	X_3	unobservables X^*	probability p_i
1	0	0	1	0	$f_{X_1,X_2,X_3}(0,0,1) = f_{X_1\|X^*}(0\|0) f_{X_2\|X^*}(0\|0) f_{X_3\|X^*}$
2	1	0	1	0	$f_{X_1,X_2,X_3}(1,0,1) = f_{X_1\|X^*}(1\|0) f_{X_2\|X^*}(0\|0) f_{X_3\|X^*}$
3	0	1	1	0	...
4	1	1	1	0	...
5	0	0	2	1	...
6	1	0	2	1	...
7	0	1	2	1	...
8	1	1	2	1	...
9	0	0	3	1	...
10	1	0	3	1	...
11	0	1	3	1	...
12	1	1	3	1	

Illustration

Table: A violation of the property of leaves in Equation 1

observation i	observables X_1	X_2	X_3	unobservables X^*	probability p_i
1	0	0	1	0	$f_{X_1,X_2,X_3}(0,0,1) = f_{X_1\|X^*}(0\|0) f_{X_2\|X^*}(0\|0) f_{X_3\|X^*}$
2	1	0	1	0	$f_{X_1,X_2,X_3}(1,0,1) = f_{X_1\|X^*}(1\|0) f_{X_2\|X^*}(0\|0) f_{X_3\|X^*}$
3	0	1	1	0	...
4	1	1	1	0	...
5	0	0	2	1	...
6	1	0	2	1	...
7	0	1	2	1	...
8	1	1	2	1	...
9	0	0	3	1	...
10	1	0	3	1	...
11	0	1	3	1	...
12	1	1	3	1	

Conclusion

- Identification in observation
- We can pin down values of X^* in each observation
- Estimation with a sample – What are the properties of the estimator?
- Example: Estimating GDP using deep neural networks