Matching Points: Supplementing Instruments with Covariates in Triangular Models

Junlong Feng [1]

[1]Hong Kong University of Science and Technology

Motivation & Objective

A Common Problem in Practice. In economic applications, it is common to have a discrete endogenous variable (D) and an instrument (Z) taking on **fewer** values:

- Return to education. D: Multiple levels of education. Z: Whether lived near a college or not (binary).
- Program evaluation. D: Multiple training programs. Z: A lottery granting access to a certain program (binary).

No Sufficient Variation in IV, No Identification. Let $Y, \boldsymbol{X}$ and U be the outcome variable, covariates, and the unobservable. The model $Y = g(D, \boldsymbol{X}, U)$ is under-identified by the IV when the cardinality of the support of D is greater than that of Z ($|S(D)| > |S(Z)|$). **For example** (Newey and Powell, 2003):

- $S(D) = \{1, 2, 3\}$ and $S(Z) = \{0, 1\}$.
- Suppose $g(D, \boldsymbol{X}, U) = \sum_{d \in S(D)} \mathbb{1}(D = d) m_d^*(\boldsymbol{X}) + U$.
- For some covariates value $\boldsymbol{X} = \boldsymbol{x}_0$, if $\mathbb{E}_{U|\boldsymbol{X}Z}(\boldsymbol{x}_0, Z) = 0$, then for every $z \in S(Z)$,

$$\sum_{d \in S(D)} m_d^*(\boldsymbol{x}_0) p_d(x_0, z) = \mathbb{E}_{Y|\boldsymbol{X}Z}(\boldsymbol{x}_0, z) \tag{1}$$

  where $p_d(\boldsymbol{x}, z) \equiv \mathbb{P}(D = d | \boldsymbol{X} = \boldsymbol{x}_0, Z = z)$ is the generalized propensity score.
- **TWO linear equations** ($z = 0, 1$) **for THREE unknowns** $(m_1^*(\boldsymbol{x}_0), m_2^*(\boldsymbol{x}_0), m_3^*(\boldsymbol{x}_0))$: **Underidentified**.

Objective: Achieve point identification by supplementing instruments with covariates.

Model

A triangular model with $S(D) = \{1, 2, 3\}$ and $S(Z) = \{0, 1\}$:

$$\text{Outcome eq.: } Y = \sum_d \mathbb{1}(D = d) m_d^*(\boldsymbol{X}) + U \tag{SP}$$

$$\text{OR: } Y = \sum_d \mathbb{1}(D = d) g_d^*(\boldsymbol{X}, U) \tag{NSP}$$

$$\text{Selection eq.: } D = d \text{ iff } h_d(\boldsymbol{X}, Z, \boldsymbol{V}) = 1, \sum_d h_d = 1 \tag{SL}$$

- The outcome heterogeneity U is a scalar.
  - D-dependent outcome heterogeneity is allowed. See the paper for more details.
- The selection heterogeneity can be vector-valued.
- No notion of monotonicity is required for selection.

Main Idea

Take Model-SP as an example.

- Consider a different value of covariates $\boldsymbol{x}_m$ such that equation (1) also holds at $\boldsymbol{x}_m$.
- Four equations but *six* unknowns.
- Identification is possible if $\Delta_d(x_0, x_m) \equiv m_d^*(\boldsymbol{x}_m) - m_d^*(\boldsymbol{x}_0)$ is known for all $d \in S(D)$: One can substitute $m_d^*(\boldsymbol{x}_m) = m_d^*(\boldsymbol{x}_0) + \Delta_d(x_0, x_m)$ into the equations for $m_d^*(\boldsymbol{x}_m)$, and then:
- Only *three* unknowns.

Such special covariate values $\boldsymbol{x}_m$ are **matching points** of $\boldsymbol{x}_0$. The paper shows how to find them for given $\boldsymbol{x}_0$s and how identification is restored using them.

Identification of Model-SP

Definition. Matching point $\boldsymbol{x}_m$ of $\boldsymbol{x}_0$: for $z, z' \in S(Z)$ and for all $d \in S(D)$:

- $h_d(\boldsymbol{x}_0, z, \boldsymbol{V}) = h_d(\boldsymbol{x}_m, z', \boldsymbol{V})$ *a.s.*
- $\mathbb{E}_{U|\boldsymbol{V}\boldsymbol{X}Z}(\boldsymbol{V}, \boldsymbol{x}_m, Z) = \mathbb{E}_{U|\boldsymbol{V}\boldsymbol{X}Z}(\boldsymbol{V}, \boldsymbol{x}_0, Z)$ *a.s.*
- $(\boldsymbol{V}|\boldsymbol{x}_m, Z)$ and $(\boldsymbol{V}|\boldsymbol{x}_0, Z)$ are identically distributed.

Main message of the definition: Selection patterns are the same at the covariates-IV combinations in a matching pair. Covariates need NOT to be exogeneous, but equal dependence of $(U, \boldsymbol{V})$ on $\boldsymbol{X}$ at $\boldsymbol{x}_0$ and $\boldsymbol{x}_m$ is required.

Key implication. Under exogeneity of Z, the definition of a matching point implies that:

$$\mathbb{E}_{U|D\boldsymbol{X}Z}(d, \boldsymbol{x}_0, z) = \mathbb{E}_{U|\boldsymbol{V}}\left(h_d(\boldsymbol{x}_0, z, \boldsymbol{V}) = 1\right) = \mathbb{E}_{U|\boldsymbol{V}}\left(h_d(\boldsymbol{x}_m, z', \boldsymbol{V}) = 1\right) = \mathbb{E}_{U|D\boldsymbol{X}Z}(d, \boldsymbol{x}_m, z')$$
$$\Downarrow$$
$$\mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_0, z) - m_d^*(\boldsymbol{x}_0) = \mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_m, z') - m_d^*(\boldsymbol{x}_m)$$

Identification. Substituting this into the equations in the form of (1) for $\boldsymbol{x}_m$ yields the following equation system for $\boldsymbol{m}^*(\boldsymbol{x}_0) \equiv (m_1^*(\boldsymbol{x}_0), m_2^*(\boldsymbol{x}_0), m_3^*(\boldsymbol{x}_0))$:

$$\begin{pmatrix} p_1(\boldsymbol{x}_0, z) & p_2(\boldsymbol{x}_0, z) & p_3(\boldsymbol{x}_0, z) \\ p_1(\boldsymbol{x}_0, z') & p_2(\boldsymbol{x}_0, z') & p_3(\boldsymbol{x}_0, z') \\ p_1(\boldsymbol{x}_m, z) & p_2(\boldsymbol{x}_m, z) & p_3(\boldsymbol{x}_m, z) \end{pmatrix} \cdot \boldsymbol{m}^*(\boldsymbol{x}_0) = \begin{pmatrix} \mathbb{E}_{Y|\boldsymbol{X}Z}(\boldsymbol{x}_0, z) \\ \mathbb{E}_{Y|\boldsymbol{X}Z}(\boldsymbol{x}_0, z') \\ \mathbb{E}_{Y|\boldsymbol{X}Z}(\boldsymbol{x}_m, z) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \sum_{d=1}^3 \left[\mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_0, z) - \mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_m, z')\right] p_d(\boldsymbol{x}_m, z) \end{pmatrix}$$

- The combination $(\boldsymbol{x}_m, z)$ is treated *as if* it was a third instrumental value.
- The term in red offsets the change caused by conditioning on a different value of the non-excluded covariates.
- Identification is achieved as long as the generalized propensity score matrix is full-rank.
- Over-identification is possible because multiple matching points may exist.

Finding the Matching Point

Two different approaches depending on how much we know about Model-SL.

If the selection model is known/specified. Find the matching point by solving $h_d(\boldsymbol{x}_0, z, \boldsymbol{v}) = h_d(\boldsymbol{x}_m, z', \boldsymbol{v})$.

Example. Ordered choice with linear cutoffs. $h_1(\boldsymbol{X}, Z, V) = \mathbb{1}(V < \kappa_1 + Z\alpha + \boldsymbol{X}'\beta)$ and $h_3(\boldsymbol{X}, Z, V) = \mathbb{1}(V \geq \kappa_2 + Z\alpha + \boldsymbol{X}'\beta)$ with $\kappa_1 < \kappa_2$. Then matching points can be obtained by solving $z\alpha + \boldsymbol{x}_0'\beta = z'\alpha + \boldsymbol{x}_m'\beta$.

If the selection model is unknown/unspecified. Obtain the matching points by matching the generalized propensity scores under the following assumption:

$$p_d(\boldsymbol{x}_0, z) = p_d(\boldsymbol{x}_m, z')\, \forall d \implies h_d(\boldsymbol{x}_0, z, \boldsymbol{V}) = h_d(\boldsymbol{x}_m, z', \boldsymbol{V})\ a.s.\ \forall d$$

The assumption holds in many widely used selection models like parametric/nonparametric ordered choice or discrete choice models.

Remark. These two approaches apply to both Model-SP and Model-NSP.

Identification of Model-NSP

- $U \sim$ Unif$[0, 1]$.
- $g_d^*(\boldsymbol{X}, \cdot) : [0, 1] \mapsto S(Y|d, \boldsymbol{X})$ strictly increasing for all realizations of $\boldsymbol{X}$ for all d.
- Matching point (strengthened): $(U, \boldsymbol{V}|\boldsymbol{x}_0, Z)$ and $(U, \boldsymbol{V}|\boldsymbol{x}_m, Z)$ are identically distributed.

Key implication. For the conditional CDFs $F_{U|D\boldsymbol{X}Z}$ and $F_{Y|D\boldsymbol{X}Z}$ for all $u \in [0, 1]$ and $d \in S(D)$:

$$F_{U|D\boldsymbol{X}Z}(u|d, \boldsymbol{x}_0, z) = F_{U|\boldsymbol{V}}\left(u|h_d(\boldsymbol{x}_0, z, \boldsymbol{V}) = 1\right) = F_{U|\boldsymbol{V}}\left(u|h_d(\boldsymbol{x}_m, z', \boldsymbol{V}) = 1\right) = F_{U|D\boldsymbol{X}Z}(u|d, \boldsymbol{x}_m, z')$$
$$\Downarrow$$
$$F_{Y|D\boldsymbol{X}Z}(g_d^*(\boldsymbol{x}_0, u)|d, \boldsymbol{x}_0, z) = F_{Y|D\boldsymbol{X}Z}(g_d^*(\boldsymbol{x}_m, u)|d, \boldsymbol{x}_m, z')$$

Inverting the CDF yields the relationship between $g_d^*(\boldsymbol{x}_0, u)$ and $g_d^*(\boldsymbol{x}_m, u)$ for all u and d. Then for each u, obtain three moment equations $\Psi(\boldsymbol{g}^*(\boldsymbol{x}_0, u)) = \boldsymbol{u}$.

- $\Psi(\cdot) : \mathbb{R}^3 \mapsto \mathbb{R}^3$ is continuous everywhere and strictly increasing on $\Pi_d S(Y|d, \boldsymbol{x}_0)$.
- A typical element in $\Psi(\boldsymbol{g}^*(\boldsymbol{x}_0, u))$ is

$$\sum_{d=1}^3 p_d(\boldsymbol{x}, z) \cdot F_{Y|D\boldsymbol{X}Z}(\varphi_d(g_d^*(\boldsymbol{x}_0, u); \boldsymbol{x})|d, \boldsymbol{x}, z)$$

- $(\boldsymbol{x}, z) = (\boldsymbol{x}_0, z), (\boldsymbol{x}_0, z'), (\boldsymbol{x}_m, z)$.
- $\varphi_d(y; \boldsymbol{x}) = F^{-1}_{Y|D\boldsymbol{X}Z}(F_{Y|D\boldsymbol{X}Z}(y|d, \boldsymbol{x}_0, z)|d, \boldsymbol{x}, z_1)$ where $(\boldsymbol{x}_0, z)$ and $(\boldsymbol{x}, z_1)$ are a matching pair.

Three moments conditions for three unknowns at each u: Under-identification solved!

A New Global Uniqueness Theorem. These moment conditions nest Chernozhukov and Hansen (2005). We prove a new global uniqueness theorem under weaker conditions.

Theorem. If Y is bounded at least from one side and the Jacobian of Ψ is full rank at $\boldsymbol{g}^*(\boldsymbol{x}_0, u)$ for all u, then $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ is the unique solution (path) to $\Psi(\boldsymbol{y}(u)) = \boldsymbol{u}$ for all u among all increasing functions on $[0, 1]$.

- Exploited monotonicity and continuity of $g_d^*(\boldsymbol{X}, \cdot)$ and Ψ.
- The condition is minimal: Only full-rankness is required without which local identification at some u is even not guaranteed.

Idea of Proof. Below is the key idea to show uniqueness among increasing functions whose range is $\Pi_d S(Y|d, \boldsymbol{x}_0)$. The general case is shown in the paper.

By boundedness of Y, all candidate solution paths start or end at the same point, so they must intersect with the true solution paths at least for once. There are then two cases.

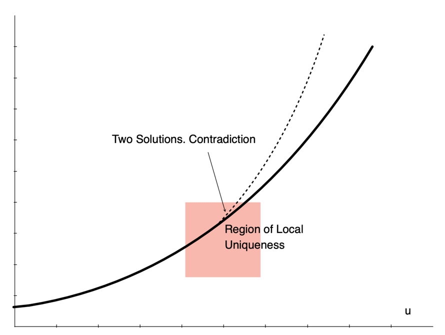

Figure 1. Case 1.

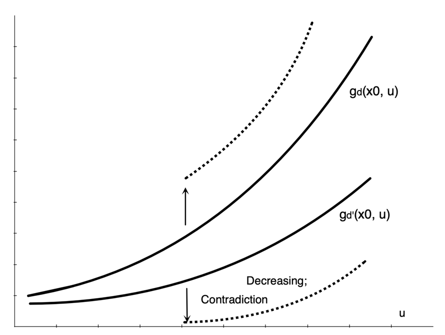

Figure 2. Case 2

- Case 1: The difference between two solution paths is continuous. Impossible because they must enter the local-uniqueness region by continuity.
- Case 2: The difference is discontinuous. Then the alternative solution path must jump up for some d. But by monotonicity and continuity of Ψ, there must exist a d' for which the alternative solution path jumps down to make the equation still hold, violating monotonicity.