Lecture 9 – nonlinear instrumental variables

Economics 2123
George Washington University

Instructor: Prof. Ben Williams

Example 2

- Suppose $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$ and X_i is endogenous.
  - Need 2 instruments for identification.

Example 2

- Suppose $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$ and X_i is endogenous.
  - Need 2 instruments for identification.
  - You can use Z_i and Z_i^2 as instruments if $E(u_i \mid Z_i) = 0$.

Example 2

- Suppose $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$ and X_i is endogenous.
  - Need 2 instruments for identification.
  - You can use Z_i and Z_i^2 as instruments if $E(u_i \mid Z_i) = 0$.
  - 2SLS:
    - Regress X_i on Z_i and Z_i^2 to get $\hat{X}_i$
    - Regress X_i^2 on Z_i and Z_i^2 to get $\widehat{X^2}_i$
    - Regress Y_i on $\hat{X}_i$ and $\widehat{X^2}_i$

Example 2

- Suppose $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$ and X_i is endogenous.
  - Need 2 instruments for identification.
  - You can use Z_i and Z_i^2 as instruments if $E(u_i \mid Z_i) = 0$.
  - 2SLS:
    - Regress X_i on Z_i and Z_i^2 to get $\hat{X}_i$
    - Regress X_i^2 on Z_i and Z_i^2 to get $\widehat{X^2}_i$
    - Regress Y_i on $\hat{X}_i$ and $\widehat{X^2}_i$
    - Note: do not use $\hat{X}_i^2$!

Example 2

- Suppose $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$ and X_i is endogenous.
  - Need 2 instruments for identification.
  - You can use Z_i and Z_i^2 as instruments if $E(u_i \mid Z_i) = 0$.
  - 2SLS:
    - Regress X_i on Z_i and Z_i^2 to get $\hat{X}_i$
    - Regress X_i^2 on Z_i and Z_i^2 to get $\widehat{X^2}_i$
    - Regress Y_i on $\hat{X}_i$ and $\widehat{X^2}_i$
    - Note: do not use $\hat{X}_i^2$!
- Similar for models with an interacted endogenous regressor.

Example 2

- Suppose $Y_i = g(X_i; \beta) + u_i$ but $E(u_i \mid X_i) \neq 0$.
  - A straightforward application of GMM if there are instruments Z_i such that $E(u_i Z_i) = 0$.

Example 2

- Suppose $Y_i = g(X_i; \beta) + u_i$ but $E(u_i \mid X_i) \neq 0$.
  - A straightforward application of GMM if there are instruments Z_i such that $E(u_i Z_i) = 0$.
  - The GMM objective function is

$$\left(\sum_{i=1}^{n} (Y_i - g(X_i; \beta)) Z_i' \right) W \left(\sum_{i=1}^{n} (Y_i - g(X_i; \beta)) Z_i' \right)$$

Example 3

- Suppose $Y_i = g(X_i, u_i; \beta)$.
  - If u_i is independent of X_i then estimation is possible via GMM or MSM if the distribution of u_i is specified.

Example 3

- Suppose $Y_i = g(X_i, u_i; \beta)$.
  - If u_i is independent of X_i then estimation is possible via GMM or MSM if the distribution of u_i is specified.
  - What if X_i is endogenous but Z_i is not?
    - If g is an invertible function then you can construct moments $E(m(Y_i, X_i; \beta)Z_i) = 0$.

Example 3

- Suppose $Y_i = g(X_i, u_i; \beta)$.
  - Control functions: Suppose we can come up with a function $\nu(X_i, Z_i)$ such that X_i is independent of u_i conditional on $\nu(X_i, Z_i)$.

Example 3

- Suppose $Y_i = g(X_i, u_i; \beta)$.
  - Control functions: Suppose we can come up with a function $\nu(X_i, Z_i)$ such that X_i is independent of u_i conditional on $\nu(X_i, Z_i)$.
    - example: if $X_i = \gamma' Z_i + V_i$

Example 3

- Suppose $Y_i = g(X_i, u_i; \beta)$.
  - Control functions: Suppose we can come up with a function $\nu(X_i, Z_i)$ such that X_i is independent of u_i conditional on $\nu(X_i, Z_i)$.
    - example: if $X_i = \gamma' Z_i + V_i$ use $\nu(X_i, Z_i) = X_i - \gamma' Z_i = V_i$

Example 3

- Suppose $Y_i = g(X_i, u_i; \beta)$.
  - Control functions: Suppose we can come up with a function $\nu(X_i, Z_i)$ such that X_i is independent of u_i conditional on $\nu(X_i, Z_i)$.
    - example: if $X_i = \gamma' Z_i + V_i$ use $\nu(X_i, Z_i) = X_i - \gamma' Z_i = V_i$
  - Then

$$E(Y_i \mid X_i, \nu_i) = E(g(X_i, u_i; \beta) \mid X_i, \nu_i) = \int g(x, u; \beta) f_{u_i|\nu_i} du_i$$

Example 3

- Suppose $Y_i = g(X_i, u_i; \beta)$.
  - Control functions: Suppose we can come up with a function $\nu(X_i, Z_i)$ such that X_i is independent of u_i conditional on $\nu(X_i, Z_i)$.
    - example: if $X_i = \gamma' Z_i + V_i$ use $\nu(X_i, Z_i) = X_i - \gamma' Z_i = V_i$
  - Then

$$E(Y_i \mid X_i, \nu_i) = E(g(X_i, u_i; \beta) \mid X_i, \nu_i) = \int g(x, u; \beta) f_{u_i|\nu_i} du_i$$

  - Average over values of ν_i to get

$$\int \left(\int g(x, u; \beta) f_{u_i|\nu_i} du_i \right) f_{\nu_i} d\nu_i = \int g(x, u; \beta) f_{u_i}(u) du_i$$

Example 3

- Suppose $Y_i = g(X_i, u_i; \beta)$.
  - The likelihood approach
    - Suppose $X_i = h(Z_i, v_i; \gamma)$ and (u_i, v_i) are independent of Z_i.
    - Suppose the density $f_{u_i, v_i; \alpha}$ is known.
    - Let $\tilde{Y}_i = (Y_i, X_i)$. Then the likelihood

$$\mathcal{L}(\beta, \gamma, \alpha) = \sum_i \log(f_{\tilde{Y}|Z}(\tilde{Y}_i \mid Z_i; \beta, \gamma, \alpha))$$

can be constructed.

IV in nonlinear models

Binary Outcomes

Heterogeneity

Linear probability model

- Suppose Y_i is a binary outcome, X_i is an endogenous regressor, and Z_i is an exogenous instrument.
  - The usual 2SLS formula treats the second stage as a linear probability model.

Linear probability model

- Suppose Y_i is a binary outcome, X_i is an endogenous regressor, and Z_i is an exogenous instrument.
  - The usual 2SLS formula treats the second stage as a linear probability model.
  - When X_i is binary also, 2SLS produces an estimate of an average of the treatment effect, $Y_1 - Y_0$, over a certain subset of the population.

Random utility model

- A random utility/threshold crossing/ linear index model:

$$y_i = \mathbf{1}(\beta_0 + \beta_1 x_i + u_i \geq 0)$$

- In this model, the treatment effect is given by

$$\mathbf{1}(\beta_0 + \beta_1 + u_i \geq 0) - \mathbf{1}(\beta_0 + u_i \geq 0)$$

- And the ATE is

$$Pr(\beta_0 + \beta_1 + u_i \geq 0) - Pr(\beta_0 + u_i \geq 0)$$

Random utility model

- Suppose that $X_i = \mathbf{1}(\gamma_0 + \gamma_1 Z_i + v_i \geq 0)$ where Z_i is binary.
- 2SLS provides an estimate of

$$
\begin{aligned}
&Pr(\beta_0 + \beta_1 + u_i \geq 0 \mid -\gamma_0 - \gamma_1 \leq v_i \leq -\gamma_0) \\
&\quad - Pr(\beta_0 + u_i \geq 0 \mid -\gamma_0 - \gamma_1 \leq v_i \leq -\gamma_0)
\end{aligned}
$$

- We will derive this later. For now, let's think about estimating β directly instead of "treatment effects".

Triangular model with probit second stage

- The two equations are

$$x_i = \gamma_0 + \gamma' Z_i + \sigma_\nu \nu_i$$
$$y_i = \mathbf{1}(\beta_0 + \beta_1 x_i + u_i \geq 0)$$

where

$$(u_i, \nu_i) \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

- This can be estimated via maximum likelihood.

Triangular model with probit second stage

- This model imposes some strong restrictions:
  - normality
  - homoskedasticity
  - full independence of Z_i
- Generally get misleading estimates from a probit that uses predicted values from first stage.
- `ivprobit` implements this in Stata (`biprobit` if X_i is also binary)

Triangular model with probit second stage

- Control function approach
  - assume that $u_i \mid x_i, \nu_i \sim u_i \mid \nu_i$
  - under this assumption:
    - estimate $\hat{\nu}_i$ from first stage
    - then estimate
      $Pr(y_i = 1 \mid x_i = x, \hat{\nu}_i = \nu) = F_{u_i|\nu_i}(\beta_0 + \beta_1 x \mid \nu)$
    - If $u_i, \nu_i \sim N(0, \Sigma)$ then the right hand side here can be derived analytically.
    - A semiparametric approach can be used to avoid specifiying the distribution $F_{u_i|\nu_i}$.

Heterogeneity

- When we talk about heterogeneity, usually we mean heterogeneity *in causal effects*.
  - The individual causal effect differs across individuals.
- James Heckman, among many others, has argued over the past 30-40 years that this type of heterogeneity is prevalent.

Review

- Recall the discussion from the first lecture:
  - $Y_{1i} - Y_{0i}$ represents the individual treatment effect
  - $\delta_x = E(Y_{1i} - Y_{0i} \mid X_i = x)$ is the average treatment effect conditional on x
  - observable heterogeneity is when δ_x varies with x

Review

- Recall the discussion from the first lecture:
  - $Y_{1i} - Y_{0i}$ represents the individual treatment effect
  - $\delta_x = E(Y_{1i} - Y_{0i} \mid X_i = x)$ is the average treatment effect conditional on x
  - observable heterogeneity is when δ_x varies with x
  - under the conditional independence assumption, OLS estimates a weighted average, $\sum_x w_x \delta_x$.

Review

- Recall the discussion from the first lecture:
  - $Y_{1i} - Y_{0i}$ represents the individual treatment effect
  - $\delta_x = E(Y_{1i} - Y_{0i} \mid X_i = x)$ is the average treatment effect conditional on x
  - observable heterogeneity is when δ_x varies with x
  - under the conditional independence assumption, OLS estimates a weighted average, $\sum_x w_x \delta_x$.
  - note that this result allows for unobserved heterogeneity too because we do not assume that $Y_{1i} - Y_{0i} = \delta_{X_i}$.

Heterogeneity+Endogeneity

- What if the conditional independence assumption fails?
  - we may use an instrumental variable strategy
  - if there is also heterogeneity, what does IV estimate?

Heterogeneity

- What if there is unobserved heterogeneity?
  - i.e., if $Y_{1i} - Y_{0i} \neq \delta_{X_i}$
  - this could be ok
  - a textbook example:
    - suppose $Y_i = \alpha + \beta_i D_i + u_i$ where $\beta_i = \beta + \eta_i$

Heterogeneity

- What if there is unobserved heterogeneity?
  - i.e., if $Y_{1i} - Y_{0i} \neq \delta_{X_i}$
  - this could be ok
  - a textbook example:
    - suppose $Y_i = \alpha + \beta_i D_i + u_i$ where $\beta_i = \beta + \eta_i$
    - then $Y_i = \alpha + \beta D_i + \varepsilon_i$ where $\varepsilon_i = u_i + \eta_i D_i$
    - if $E(u_i D_i) = 0$ and $E(\eta_i \mid D_i) = 0$ then OLS estimates $\beta = E(\beta_i)$

Heterogeneity

- The Roy model will be used to demonstrate a link between unobserved heterogeneity and endogeneity.
- The textbook example above is misleading because often η_i will be correlated with D_i.
- Moreover, even if Z_i is uncorrelated with u_i it will often not be uncorrelated with $\eta_i D_i$.

LATE

- Ignore X (exogenous control variables) and let D_z denote the (counteractual) value of D when Z is fixed at z.

LATE

- Ignore X (exogenous control variables) and let D_z denote the (counteractual) value of D when Z is fixed at z.
    - in the random utility/threshold crossing model, $D_z = \mathbf{1}(\gamma_2' z \geq V)$

LATE

- Ignore X (exogenous control variables) and let D_z denote the (counteractual) value of D when Z is fixed at z.
  - in the random utility/threshold crossing model, $D_z = \mathbf{1}(\gamma_2' z \geq V)$
- Imbens and Angrist consider a binary Z and show that

$$\frac{E(Y \mid Z=1) - E(Y \mid Z=0)}{E(D \mid Z=1) - E(D \mid Z=0)} = E(Y_1 - Y_0 \mid D_1 > D_0)$$

- Thus, IV (lhs) identifies the local average treatment effect (LATE; rhs), which is the average effect for those induced to “participate” by Z. This population is sometimes called the “compliers”.

LATE assumptions

- Let $Y_i(d, z)$ denote the counterfactual outcome.
- Theorem 4.4.1 in MHE.
  - Assumption 1. $(Y_i(D_{1i}, 1), Y_i(D_{0i}, 0), D_{1i}, D_{0i}) \perp\!\!\!\perp Z_i$
  - Assumption 2. $Y_i(d, 1) = Y_i(d, 0)$
  - Assumption 3. $E(D_{1i} - D_{0i}) \neq 0$
  - Assumption 4. $D_{1i} - D_{0i} \geq 0$ for all i, or vice versa

LATE assumptions

- Theorem 4.4.1 in MHE.
  - *monotonicity*: The ceteris paribus effect of changing Z on D has the same sign for everyone, i.e., either $D_{1i} \geq D_{0i}$ for all i or $D_{1i} \leq D_{0i}$ for all i.

LATE assumptions

- Theorem 4.4.1 in MHE.
  - *monotonicity*: The ceteris paribus effect of changing Z on D has the same sign for everyone, i.e., either $D_{1i} \geq D_{0i}$ for all i or $D_{1i} \leq D_{0i}$ for all i.
  - Really this is a "uniformity" assumption. If Z takes more than two values there is no need for monotonicity, only that D changes in the same direction for everyone as Z changes.
  - The assumption is implied by the equation $D = \mathbf{1}(\gamma_1' X + \gamma_2' Z \geq V)$ but it would fail if γ_2 was a random coefficient.
  - MHE interpret the assumption as requiring no "defiers".

More on LATE

- When Z is continuous, we can estimate the MTE and various weighted averages of the MTE.
- The LATE framework is useful in understanding what we are able to learn when Z is discrete.
  - Cases where $LATE = TT$ or $LATE = TUT$
  - Characterizing compliers.
  - LATE with covariates

Special cases

- The TT can be written as a weighted average of LATE and the average effect for the always-takers.
- In some cases, D must be equal to 0 when $Z = 0$.
  - The Bloom example – Z is a random assignment and D a treatment and there is one-way noncompliance.
  - One-way noncompliance means that some with $Z = 1$ choose $D = 0$ (refuse treatment) but no one with $Z = 0$ can have $D = 1$.
- In these cases, IV estimates TT.

Special cases

- The TUT can be written as a weighted average of LATE and the average effect for the never-takers.
- In some cases, D must be equal to 1 when $Z = 1$.
    - Suppose D indicates having a third child (as opposed to only 2) and Z indicates whether the second birth was a multiple birth.
    - Then if $Z = 1$ we must have $D = 1$.
    - There are no “never-takers”.
- In these cases, IV estimates TUT.

Compliers

- A few results:
  - $Pr(D_1 > D_0) = E(D \mid Z = 1) - E(D \mid Z = 0)$
  - for any W such that (D_1, D_0) is independent of Z conditional on W, $E(W \mid D_1 > D_0) = \frac{E(\kappa W)}{E(\kappa)}$ where

$$\kappa = 1 - \frac{D(1-Z)}{1 - Pr(Z = 1 \mid W)} - \frac{(1-D)Z}{Pr(Z = 1 \mid W)}$$

  - and, more generally, $f_{W|D_1>D_0}(w)$ is equal to

$$\frac{E(D \mid Z = 1, W = w) - E(D \mid Z = 0, W = w)}{E(D \mid Z = 1) - E(D \mid Z = 0)} f_W(w)$$

LATE with covariates

- The LATE story gets quite a bit more complicated with covariates.
- Let $\lambda(x) = E(Y_1 - Y_0 \mid D_1 > D_0, X = x)$ denote the LATE conditional on X.
- We could estimate these directly using the Wald formula conditional on X.

LATE with covariates

- The LATE story gets quite a bit more complicated with covariates.
- Let $\lambda(x) = E(Y_1 - Y_0 \mid D_1 > D_0, X = x)$ denote the LATE conditional on X.
- We could estimate these directly using the Wald formula conditional on X.
- If we do 2SLS where the first stage is fully saturated and the second stage is saturated in X we get a weighted average of the $\lambda(x)$.
  - The weights are larger for values of x such that $Var(E(D \mid X = x, Z) \mid X = x)$ is larger.

LATE with covariates

- The LATE story gets quite a bit more complicated with covariates.
- Let $\lambda(x) = E(Y_1 - Y_0 \mid D_1 > D_0, X = x)$ denote the LATE conditional on X.
- We could estimate these directly using the Wald formula conditional on X.
- If we do 2SLS where the first stage is fully saturated and the second stage is saturated in X we get a weighted average of the $\lambda(x)$.
  - The weights are larger for values of x such that $Var(E(D \mid X = x, Z) \mid X = x)$ is larger.
- if $Pr(Z = 1 \mid X)$ is a linear function of X then 2SLS gives the minimum MSE approximation to $E(Y \mid D, X, D_1 > D_0)$.
  - This is useful because $E(Y \mid D = 1, X, D_1 > D_0) - E(Y \mid D = 0, X, D_1 > D_0) = \lambda(X)$.
  - Abadie (2003) proposes a way to estimate this same minimum MSE approximation when $Pr(Z = 1 \mid X)$ is not linear.