Lectures on Identification 2

Andrew Chesher

CeMMAP & UCL

April 16th 2008

Topics

1. Monday April 14th. Motivation, history, definitions, types of model, parametric and nonparametric IV models.
2. Today.
    1. Quantiles and a non-additive IV model.
    2. Triangular models and control function methods.
    3. Models with excess heterogeneity and index restrictions.
3. Monday April 21st. Discrete endogenous variables, set identification and control function methods. Set identification in IV models for binary data.
4. Wednesday April 23rd. Seminar on "Endogeneity and Discrete Outcomes" - set identification in IV models for discrete data.

Quantiles 1

- Random variable A with **distribution** function:

$$F_A(a) \equiv \Pr[A \leq a]$$

has **quantile** function

$$Q_A(p) \equiv \inf\{a : F_A(a) \geq p\}$$

- When A is **continuously** distributed

$$p = F_A(Q_A(p)) \qquad a = Q_A(F_A(a))$$

but not when A is **discrete**.

- For example: binary A has $\Pr[A = 0] = p_0 \in (0, 1)$

$$F_A(a) = \begin{cases} p_0 & , \quad a < 1 \\ 1 & , \quad a = 1 \end{cases} \qquad Q_A(p) = \begin{cases} 0 & , \quad p \leq p_0 \\ 1 & , \quad p > p_0 \end{cases}$$

- So e.g. if $p_0 = 0.5$,

$$F_A(Q_A(0.3)) = 0.5 \qquad F_A(Q_A(0.7)) = 1$$

- Generally

$$F_A(Q_A(p)) \geq p$$

Quantiles 2

- Random variable A has **distribution** and quantile function:

$$F_A(a) \equiv \Pr[A \leq a] \qquad Q_A(p) \equiv \inf\{a : F_A(a) \geq p\}$$

- Definition: $U \in [0,1]$ has a uniform distribution, $Unif(0,1)$ if $F_U(u) = u$ then $Q_U(p) = p$
- For **continuous** A,

$$B = F_A(A) \sim Unif(0,1)$$

- This because

$$\Pr[A \leq \alpha] \equiv F_A(\alpha)$$

$$\{a : a \leq \alpha\} = \{a : F_A(a) \leq F_A(\alpha)\}$$

$$\Pr[F_A(A) \leq F_A(\alpha)] = F_A(\alpha)$$

- So

$$\Pr[B \leq b] = b$$

Quantiles 3

- Random variable A has **distribution** and quantile function:

$$F_A(a) \equiv \Pr[A \leq a] \qquad Q_A(p) \equiv \inf\{a : F_A(a) \geq p\}$$

- Definition: $U \in [0,1]$ has a uniform distribution, $Unif(0,1)$ if $F_U(u) = u$ then $Q_U(p) = p$
- For continuous A,

$$B = F_A(A) \sim Unif(0,1).$$

- For continuous A it follows from

$$F_A(A) = U \Rightarrow Q_A(U) = A$$

- For discrete and continuous A, $Q_A(U)$ has the same distribution as A.

$$Q_A(U) = A$$

- Conditioning on $X = x$

$$Q_{A|X}(U|x) \text{ has the distribution } A \text{ has when } X = x$$

Normalization

- Continuous Y, scalar continuous U, h strictly monotonic (increasing):

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

- Can normalize $U \sim Unif(0, 1)$ under monotonicity - free choice of units of measurement.
- Consider for continuous W

$$Y = h^*(X, W) = h^*(X, Q_W(F_W(W))) = h(X, U)$$

where

$$U \equiv F_W(W)$$

and

$$h(X, \cdot) \equiv h^*(X, Q_W(\cdot))$$

The IV model nonadditive U

- Chernozhukov and Hansen (2005), Chernozhukov, Imbens, Newey (2007):

$$Y = h(X, U) \qquad U \perp\!\!\!\perp Z$$

  h strictly increasing in scalar $U \sim Unif(0, 1)$.

- Define

$$a(\tau, x, z) \equiv \Pr[U \leq \tau | X = x, Z = z]$$

  Independence $U \perp\!\!\!\perp Z$ requires for all τ, z.

$$E_{X|Z}[a(\tau, X, z)|Z = z] = \Pr[U \leq \tau | Z = z] = \tau$$

- h strictly increasing with respect to U implies for each x (and z)

$$\{u : u \leq \tau\} = \{u : h(x, u) \leq h(x, \tau)\}$$

- So

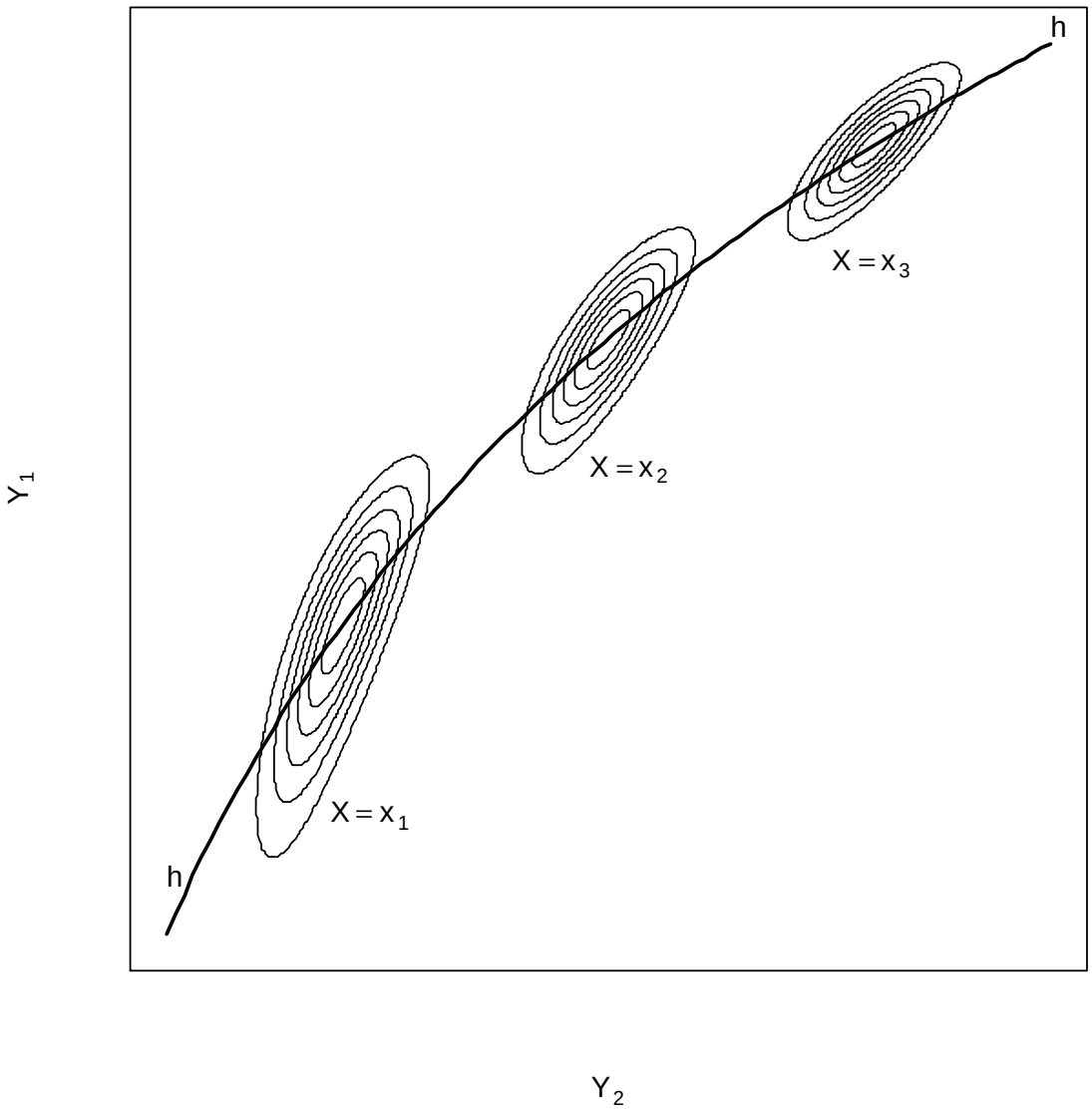
$$\Pr[U \leq \tau | X = x, Z = z] = \Pr[h(X, U) \leq h(X, \tau) | X = x, Z = z]$$

$$a(\tau, x, z) = \Pr[Y \leq h(X, \tau) | X = x, Z = z]$$

- So for all τ, z

$$\Pr[Y \leq h(X, \tau) | Z = z] = \tau$$

Figure 1: Contours of a joint density function of Y_1 and Y_2 conditional on X at 3 values of X. The line marked hh is the structural function $h(Y_2, X, q_\tau)$, not varying across $X \in \{x_1, x_2, x_3\}$, drawn with $q_\tau = 0.5$.



The IV model nonadditive U

- h monotonic (normalized increasing) in scalar U.

$$Y = h(X, U) \qquad U \perp\!\!\!\perp Z$$

For all τ, z

$$\Pr[Y \leq h(X, \tau) | Z = z] = \tau$$

  - identification of h if restrictions on h and F_{UXZ} guarantee a unique solution.
  - rich support for Z; all or nothing identification.
  - set identification is a possibility.
  - local independence: could just have $Q_{U|Z}(\tau|z) = \tau$ then identify $h(x, \tau)$

Example: triangular Gaussian model

- Suppose Y and X are generated by a triangular Gaussian model:

$$\begin{aligned} Y &= \alpha_0 + \alpha_1 X + U \\ X &= g(Z) + V \end{aligned}$$

$$\begin{bmatrix} U \\ V \end{bmatrix} | Z = z \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{UU} & \sigma_{UV} \\ \sigma_{UV} & \sigma_{VV} \end{bmatrix} \right)$$

  Note that it could be $\sigma_{UV}(z)$ and still $U \perp\!\!\!\perp Z$.
- The distribution of $(Y - a_0 - a_1 X)$ given $Z = z$ is:

$$N\left((\alpha_0 - a_0) + (\alpha_1 - a_1)\, g(z), \sigma_{UU} + 2(\alpha_1 - a_1)\sigma_{UV} + (\alpha_1 - a_1)^2 \sigma_{VV} \right)$$

- With $\tau = 0.5$, we seek a_0 and a_1 such that $P[Y \leq a_0 + a_1 X | z] = 0.5$ for all z where

$$P[Y \leq a_0 + a_1 X | z] = \Phi\left(\frac{-(\alpha_0 - a_0) - (\alpha_1 - a_1)\, g(z)}{\left(\sigma_{UU} + 2(\alpha_1 - a_1)\sigma_{UV} + (\alpha_1 - a_1)^2 \sigma_{VV}\right)^{1/2}} \right)$$

Triangular models

- Consider **complete** models for outcomes X and Y with triangular structure

$$\text{A:} \quad \begin{array}{l} Y = h(X, U) \\ X = g(Z, V) \end{array}$$

and "incomplete models"

$$\text{B: } Y = h(X, U)$$

- Consider the impact of **discrete** vs **continuous** variation in (Y, X, Z) on the nature of identification.
- We find:

	X discrete	**X continuous**
Y discrete	A:set B: set	A: point B: set
Y continuous	A: set B: point	A:point B: point

Triangular Gaussian model

- Y and X are generated by a triangular Gaussian model:

$$\begin{aligned} Y &= \alpha_0 + \alpha_1 X + U \\ X &= g(Z) + V \end{aligned}$$

$$\begin{bmatrix} U \\ V \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{uu} & \sigma_{uv} \\ \sigma_{uv} & \sigma_{vv} \end{bmatrix} \right) \qquad U \perp\!\!\!\perp Z$$

- Conditional distribution of U given V and Z

$$N\left(\frac{\sigma_{uv}}{\sigma_{vv}} V, \sigma_{uu} - \frac{\sigma_{uv}^2}{\sigma_{vv}} \right)$$

- The expected value of Y given V and Z.

$$E[Y|V = v, Z = z] = \alpha_0 + \alpha_1 \left(g(z) + v \right) + \frac{\sigma_{uv}}{\sigma_{vv}} v$$

- Conditioning on $V = v$, $Z = z$ is the same as conditioning on $X = x \equiv g(z) + v$, $Z = z$ so

$$E[Y|X = x, Z = z] = \alpha_0 + \alpha_1 x + \frac{\sigma_{uv}}{\sigma_{vv}} \left(x - g(z) \right)$$

Additive latent variable model

- X and Y are generated by

$$\begin{aligned} Y &= h(X) + U \\ X &= g(Z) + V \end{aligned}$$

$$E[U|V = v, Z = z] = c(v) \qquad E[V|Z = z] = d$$

- Does **not** imply $E[U|Z = z]$ is free of z.
  - distinct from the IV model
  - If $c(\cdot)$ is linear then $E[U|Z = z]$ is free of z.
  - If $V \perp\!\!\!\perp Z$ then $E[U|Z = z]$ is free of z.

Additive latent variable model

- X and Y are generated by

$$\begin{aligned} Y &= h(X) + U \\ X &= g(Z) + V \end{aligned}$$

$$E[U|V = v, Z = z] = c(v)$$

- Condition on $V = v$, $Z = z$.

$$E[Y|V = v, Z = z] = h(g(z) + v) + c(v)$$

- Conditioning on $V = v$, $Z = z$ is the same as conditioning on $X = x \equiv g(z) + v$, $Z = z$ so

$$E[Y|X = x, Z = z] = h(x) + c(x - g(z))$$

- If e.g. $E[V|Z = z] = d$ then $E[X|Z = z] = g(z) + d$ and $g(z)$ is identified up to an additive constant.

Additive latent variable model: partial differences

- X and Y are generated by

$$\begin{aligned} Y &= h(X) + U \\ X &= g(Z) + V \end{aligned}$$

$$E[U|V = v, Z = z] = c(v) \qquad E[V|Z = z] = d$$

- Condition on $X = x$, $Z = z$.

$$E[Y|X = x, Z = z] = h(x) + c(x - g(z))$$

- Consider two values of Z: z' and z'' and let

$$x' \equiv g(z') + d = E[X|Z = z']$$
$$x'' \equiv g(z'') + d = E[X|Z = z'']$$

- Then

$$\begin{aligned} E[Y|X &= x', Z = z'] = h(x') + c(d) \\ E[Y|X &= x'', Z = z''] = h(x'') + c(d) \end{aligned}$$

$$E[Y|X = x', Z = z'] - E[Y|X = x'', Z = z''] = h(x') - x(x'')$$

Additive latent variable model: partial derivatives

- X and Y are generated by

$$\begin{aligned} Y &= h(X) + U \\ X &= g(Z) + V \end{aligned}$$

$$E[U|V = v, Z = z] = c(v) \qquad E[V|Z = z] = d$$

- Condition on $X = x$, $Z = z$.

$$E[Y|X = x, Z = z] = h(x) + c(x - g(z))$$

$$E[X|Z = z] = g(z) + d$$

- Consider partial derivatives with respect to x and z

$$\nabla_x E[Y|X = x, Z = z] = \nabla_x h(x) + c'(x - g(z))$$

$$\nabla_z E[Y|X = x, Z = z] = \nabla_z h(x) - c'(x - g(z))\nabla_z (g(z)$$

$$\nabla_z E[X|Z = z] = \nabla_z g(z)$$

- Since $\nabla_z h(x) = 0$ (exclusion restriction), if $\nabla_z g(z) \neq 0$

$$\nabla_x E[Y|X = x, Z = z] + \frac{\nabla_z E[Y|X = x, Z = z]}{\nabla_z E[X|Z = z]} = \nabla_x h(x)$$

Non-additive latent variable model

- Structural functions: X and Y are generated by

$$Y = h(X, U) \qquad X = g(Z, V)$$

  - h weakly monotonic (increasing) in U - allows discrete Y
  - g **strictly** monotonic (increasing) in V- requires continuous X.
- Unobservables: (U, V) independent of Z.
- Can normalize $V \sim Unif(0, 1)$ - then

$$g(z, v) = Q_{X|Z}(v|z)$$

Non-additive latent variable model

- Structural functions: X and Y are generated by

$$Y = h(X, U) \qquad X = g(Z, V) \qquad (U, V) \perp\!\!\!\perp Z$$

  - h weakly monotonic (increasing) in U - g strictly monotonic (increasing) in V.
- Express Y in terms of V and Z

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

$$Q_{Y|VZ}(\tau|v, z) = h(g(z, v), Q_{U|V}(\tau|v))$$

- Since $(V = v \cap Z = z) \Leftrightarrow (\ X = x \equiv g(z, v) \cap Z = z)$:

$$Q_{Y|XZ}(\tau|x, z) = h(x, Q_{U|V}(\tau|v))$$

$$x \equiv g(z, v) = Q_{X|Z}(v|z)$$

so

$$Q_{Y|XZ}(\tau|Q_{X|Z}(v|z), z) = h(x, Q_{U|V}(\tau|v))$$

Non-additive latent variable model: partial differences

- Structural functions: X and Y are generated by

$$Y = h(X, U) \qquad X = g(Z, V) \qquad (U, V) \perp\!\!\!\perp Z$$

h weakly monotonic (increasing) in U - g strictly monotonic (increasing) in V.

$$Q_{Y|XZ}(\tau|Q_{X|Z}(v|z), z) = h(x, Q_{U|V}(\tau|v))$$

- Consider $Z \in \{z', z''\}$ define

$$x' \equiv Q_{X|Z}(v|z') \qquad x'' \equiv Q_{X|Z}(v|z'')$$

- Then

$$h(x', Q_{U|V}(\tau|v)) - h(x'', Q_{U|V}(\tau|v))$$

is identified by

$$Q_{Y|XZ}(\tau|Q_{X|Z}(v|z'), z') - Q_{Y|XZ}(\tau|Q_{X|Z}(v|z''), z'')$$

Angrist-Krueger QJE (1991)

1930-39 cohort: W: log wage, S: years of schooling, B: quarter of birth

$$\begin{aligned} W &= h(S, U) \\ S &= g(B, V) \end{aligned}$$

Quantiles of years of schooling (S) by quarter of birth

$$Q_{S|B}(v|b) \qquad b \in \{1, 2, 3, 4\}$$

$v =$	.1	.2	.3	.4	.5
$b = 1$	8.42	10.58	11.66	11.93	12.20
$b = 2$	8.48	10.64	11.67	11.95	12.22
$b = 3$	8.66	10.95	11.71	11.95	12.24
$b = 4$	8.75	11.06	11.71	11.98	12.25

Angrist-Krueger QJE (1991)

Estimated returns to schooling for median earner

$$\frac{Q_{W|SB}(.5|Q_{S|B}(v|b'),b') - Q_{W|SB}(.5|Q_{S|B}(v|b''),b'')}{Q_{S|B}(v|b') - Q_{S|B}(v|b'')}$$

$v =$	.1	.2	.3	.4	.5
$b' = 1$	.065	.012	$-.508$	$-.210$	$-.176$
$b'' = 2$	(.121)	(.090)	(.393)	(.218)	(.228)
$b' = 1$	.070	.045	.048	.064	.060
$b'' = 3$	(.030)	(.014)	(.095)	(.106)	(.111)
$b' = 1$	.065	.060	.083	.068	.043
$b'' = 4$	(.022)	(.011)	(.079)	(.083)	(.089)

Non-additive latent variable model: partial derivatives

- Structural functions: X and Y are generated by

$$Y = h(X, U) \qquad X = g(Z, V) \qquad (U, V) \perp\!\!\!\perp Z$$

$$Q_{Y|XZ}(\tau|Q_{X|Z}(v|z), z) = h(x, Q_{U|V}(\tau|v)) \text{ where } x \equiv g(z, v)$$

- Differentiate **left** and **right** with respect to z

$$\nabla_x Q_{Y|XZ}(\tau|x, z)|_{x=Q_{X|Z}(v|z)} \nabla_z Q_{X|Z}(v|z) + \nabla_z Q_{Y|XZ}(\tau|x, z)|_{x=Q_{X|Z}(v|z)}$$

$$\nabla_x h(x, Q_{U|V}(\tau|v))|_{x=g(z,v)} \nabla_z g(z, v)$$

- Since $g(z, v)$ and $\nabla_z g(z, v)$ are identified by $Q_{X|Z}(v|z)$ and $\nabla_z Q_{X|Z}(v|z)$

$$\nabla_x Q_{Y|XZ}(\tau|x, z)|_{x=Q_{X|Z}(v|z)} + \frac{\nabla_z Q_{Y|XZ}(\tau|x, z)|_{x=Q_{X|Z}(v|z)}}{\nabla_z Q_{X|Z}(v|z)} =$$

$$\nabla_x h(x, Q_{U|V}(\tau|v))$$

Non-additive latent variable model: remarks

- Structural functions: X and Y are generated by

$$Y = h(X, U) \qquad X = g(Z, V) \qquad (U, V) \perp\!\!\!\perp Z$$

h weakly monotonic (increasing) in U - g strictly monotonic (increasing) in V.

  - could proceed with $U \perp\!\!\!\perp Z|V$ which implies $Q_{U|VZ}(\tau|v, z)$ is free of z.
  - but must be able to identify the function of X and Z that delivers V.
  - could have local independence, e.g. $Q_{U|VZ}(0.5|v, z)$ is free of z.
  - Chesher (2003), Ma and Koenker (2006), Lee (2007), Imbens and Newey (2003).
  - restrictions on number of latent variates.

Non-additive latent variable model: excess heterogeneity

- Scalar outcome W is determined by

$$Y = h(\theta(X, Z_1), U)$$

  where U is a latent vector random variable, $\theta(X, Z_1)$ is a scalar valued function of endogenous scalar X and vector Z_1.
- The continuous endogenous X is determined by

$$X = g(Z, V)$$

  where h is strictly increasing in scalar continuous latent variate V and $Z = (Z_1, Z_2)$.
- There is the covariation restriction: (U, V) independent of Z.
- Study identification and estimation of **index relative sensitivity**:

$$\frac{\partial\theta(x, z_1)/\partial x}{\partial\theta(x, z_1)/\partial z_{1j}}.$$

Non-additive latent variable model: excess heterogeneity

- Outcome Y is determined by

$$Y = h(\theta(X, Z_1), U)$$

- Continuous *endogenous* Y given by

$$X = g(Z, V) \qquad V = r(X, Z)$$

where r is the inverse function such that

$$X = g(Z, r(X, Z)).$$

- $(U, V) \perp\!\!\!\perp Z$. Normalize $V \sim Unif(0, 1)$. Then

$$r(x, z) = F_{X|Z}(x|z).$$

- The conditional distribution function $F_{Y|XZ}(y|x, z) \equiv P[Y \leq y|X = x, Z = z]$:

$$F_{Y|XZ}(y|x, z) = \int_{h(\theta(x,z_1),u)\leq y} \cdots \int dF_{U|V}(u|r(x, z)) \equiv s(\theta(x, z_1), r(x, z), y)$$

Non-additive latent variable model: excess heterogeneity

$$F_{Y|XZ}(y|x,z) = \int_{h(\theta(x,z_1),u)\leq y} \cdots \int dF_{U|V}(u|r(x,z)) \equiv s(\theta(x,z_1), r(x,z), y)$$

- To study identification of $\nabla_x \theta / \nabla_{z_1} \theta$ consider derivatives of $F_{Y|XZ}$.

$$\begin{aligned}
\nabla_x F_{Y|XZ} &= \boldsymbol{\nabla_\theta s} \times \boldsymbol{\nabla_x \theta} + \nabla_r s \times \nabla_x r \\
\nabla_{z_1} F_{Y|XZ} &= \boldsymbol{\nabla_\theta s} \times \boldsymbol{\nabla_{z_1} \theta} + \nabla_r s \times \nabla_{z_1} r \\
\nabla_{z_2} F_{Y|XZ} &= \nabla_r s \times \nabla_{z_2} r
\end{aligned}$$

$$\nabla_x F_{X|Z} = \nabla_x r \qquad \nabla_{z_1} F_{X|Z} = \nabla_{z_1} r \qquad \nabla_{z_2} F_{X|Z} = \nabla_{z_2} r$$

- There is identification of $\nabla_x \theta / \nabla_{z_1} \theta$ if $\nabla_{z_2} F_{X|Z} \neq 0$ because

$$\begin{aligned}
\boldsymbol{\nabla_\theta s} \times \boldsymbol{\nabla_x \theta} &= \nabla_x F_{Y|XZ} - \nabla_x F_{X|Z} \times \frac{\nabla_{z_2} F_{Y|XZ}}{\nabla_{z_2} F_{X|Z}} \\
\boldsymbol{\nabla_\theta s} \times \boldsymbol{\nabla_{z_1} \theta} &= \nabla_{z_1} F_{Y|XZ} - \nabla_{z_1} F_{X|Z} \times \frac{\nabla_{z_2} F_{Y|XZ}}{\nabla_{z_2} F_{X|Z}}
\end{aligned}$$

Non-additive latent variable model: excess heterogeneity

- With **continuous** Y

$$\nabla_x F_{X|Z}(x|z) = \frac{1}{\nabla_v Q_{X|Z}(v|z)} \qquad \nabla_{z_i} F_{X|Z}(x|z) = -\frac{\nabla_{z_i} Q_{X|Z}(v|z)}{\nabla_v Q_{X|Z}(v|z)}$$

  where $v \equiv F_{X|Z}(x|z)$
- There is identification of $\nabla_x \theta / \nabla_{z_1} \theta$ if $\nabla_{z_2} F_{X|Z} \neq 0$ because

$$\begin{aligned} \boldsymbol{\nabla_\theta s \times \nabla_x \theta} &= \nabla_x F_{Y|XZ} - \nabla_x F_{X|Z} \times \frac{\nabla_{z_2} F_{Y|XZ}}{\nabla_{z_2} F_{X|Z}} \\ \boldsymbol{\nabla_\theta s \times \nabla_{z_1} \theta} &= \nabla_{z_1} F_{Y|XZ} - \nabla_{z_1} F_{X|Z} \times \frac{\nabla_{z_2} F_{Y|XZ}}{\nabla_{z_2} F_{X|Z}} \end{aligned}$$

- In terms of quantile derivatives:

$$\begin{aligned} \boldsymbol{\nabla_\theta s \times \nabla_x \theta} &= -\frac{1}{\nabla_\tau Q_{Y|XZ}} \left(\nabla_x Q_{Y|XZ} + \frac{\nabla_{z_2} Q_{Y|XZ}}{\nabla_{z_2} Q_{X|Z}} \right) \\ \boldsymbol{\nabla_\theta s \times \nabla_{z_1} \theta} &= -\frac{1}{\nabla_\tau Q_{Y|XZ}} \left(\nabla_{z_1} Q_{Y|XZ} - \nabla_{z_1} Q_{X|Z} \frac{\nabla_{z_2} Q_{Y|XZ}}{\nabla_{z_2} Q_{X|Z}} \right) \end{aligned}$$

Types of restrictions

- The triangular model results require X continuous

$$\text{A:} \quad \begin{array}{l} Y = h(X, U) \\ X = g(Z, V) \end{array}$$

- The IV model results require Y continuous

$$\text{B: } Y = h(X, U)$$

- What is the nature of identification when these conditions fail?
- We find:

	X discrete	**X continuous**
Y discrete	A:set B: set	A: point B: set
Y continuous	A: set B: point	A:point B: point