Yet More on the Scope of Application of IV Models

Conference in Honour of Grant Hillier

Andrew Chesher and Adam Rosen

CeMMAP & UCL

May 21st 2015

Homage

- Our title: "Yet More on the Scope of Application of IV Models".
- Of course this is a homage to Grant's beautifully executed series of papers on the exact properties of IV estimators.
- Amongst these landmark papers in ET 2009 was "Yet More on the Exact Properties of IV Estimators".
- The talk today reports on research into the application of the IV model in novel situations.
- Our work started in 2008 producing 6 published papers so far - so indeed this is "Yet More....."

IV models: incompleteness

- Typically IV models are **incomplete**. For example:

$$Y_1 = \alpha Y_2 + \beta Z + U \qquad E[U|Z] = 0$$

with endogenous $Y = (Y_1, Y_2)$, exogenous Z and unobserved U.

  - The IV model allows endogenous variables Y to be **set-valued** functions of Z and U.
  - Includes "single equation" models, models with multiple equilibria or inequality restrictions.

- **Independence**: IV models restrict dependence among exogenous Z and unobserved U, e.g. $E[U|Z] = 0$ or $U \perp\!\!\!\perp Z$.
- **Exclusion**: IV models typically restrict the direct dependence of outcomes on observed exogenous variables, e.g. above restrictions in β.

Classical IV models have point-valued residuals

- From the start IV models had **point-valued** residuals.
- By this I mean unobservable(s) were restricted to be **single-valued** functions of observable variables.
  - Linear models - Hillier (1984, 1985, 2006, 2009a, 2009b)

$$Y_1 = \alpha Y_2 + \beta Z + U$$

  - Nonparametric models - Newey and Powell (1988, 2003)

$$Y_1 = g(Y_2, Z) + U \qquad \forall z \quad E[U|Z = z] = 0$$

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

$$Y_1 = g(Y_2, Z, \underbrace{U}_{\text{scalar}}) \qquad U \perp\!\!\!\perp Z$$

with g strictly monotone in scalar U so $U = g^{-1}(Y_2, Z, Y_1)$.

Models with set-valued residuals: examples

- Point-valued residual restriction is limiting, excludes consideration of many cases.
  - Ordered and binary outcomes, e.g.

$$Y_1 = 1[\alpha_0 + \alpha_1 Y_2 + \beta Z + U > 0]$$

  - J alternative multiple discrete choice.

$$Y_1 = \arg\max_j \{\alpha_j Y_2 + \beta_j Z + U_j : j \in \{1, 2, \ldots, J\}\}$$

  - Censored endogenous variable Y_2^* with (Y_1, Y_{2l}, Y_{2u}, Z) observed.

$$Y_1 = g(Y_2^*, Z, U) \qquad P[Y_{2l} \leq Y_2^* \leq Y_{2u}] = 1$$

  - Random coefficients

$$Y_1 = U_1 + U_2 Y_2$$

  - Models with inequality restrictions, e.g. auction models.

Models with set valued residuals

- In incomplete models with set-valued residuals, applied practice has been to:
  - ignore endogeneity, or
  - use complete models, or
  - assume conditional exogeneity, e.g. $.U \perp\!\!\!\perp Y_2 | c(Y_2, Z) = c^*$, for some control function $c(Y_2, Z)$ and values c^*.
- We show that the incomplete IV model **can** be employed - it is typically partially identifying.
  - We set up a framework for analysis of incomplete IV models with set-valued residuals.
  - We characterize the (partial) identifying power of these generalized IV models,.
  - Apply to the random coefficients linear model with endogenous explanatory variable.

Economic processes, data

- A process delivers values of endogenous outcomes given values of exogenous variables.
  - Y: a list of observed endogenous random variables,
  - Z: a list of observed exogenous random variables,
  - U: a list of *unobserved* exogenous random variables.

- Point-valued (Y, Z, U) are random vectors defined on a suitable probability space. The support of (Y, Z, U), denoted $\mathcal{R}_{YZU}$, is a subset of Euclidean space.
  - Marginal and conditional support is denoted e.g. $\mathcal{R}_Z$, $\mathcal{R}_{Y|Z=z}$.
  - UPPER case font for random variables, lower case for realizations, $\mathcal{SCRIPT}$ font for sets.

Let's talk about sets

- Our models restrict a structural function $h : \mathcal{R}_{YZU} \to \mathbb{R}$ that delivers **sets** of outcomes.

$$\mathcal{Y}(u, z; h) \equiv \{y : h(y, z, u) = 0\}$$

  - $\mathcal{Y}(u, z; h)$ contains the values of Y the model allows when $Z = z$ and $U = u$.
  - in a complete model the set $\mathcal{Y}(u, z; h)$ is always *singleton*.
  - for example in the linear model with

$$y = z'\beta + u$$

$$h(y, z, u) = y - z'\beta - u$$

$$\mathcal{Y}(u, z; h) = \{z'\beta + u\}$$

Let's talk about sets

- Consider models that restrict a function $h : \mathcal{R}_{YZU} \to \mathbb{R}$ that delivers **sets** of outcomes.

$$\mathcal{Y}(u, z; h) \equiv \{y : h(y, z, u) = 0\}$$

- Function h also delivers a set: $\mathcal{U}(y, z; h)$:

$$\mathcal{U}(y, z; h) \equiv \{u : h(y, z, u) = 0\}$$

  - $\mathcal{U}(y, z; h)$ contains all values of U that can deliver $Y = y$ when $Z = z$.
  - in a classical IV model $\mathcal{U}(y, z; h)$ is a *singleton*, a residual
  - for example in the linear IV model

$$y_1 = \alpha y_2 + u$$

$$h(y, z, u) = y_1 - \alpha y_2 - u$$

$$\mathcal{U}(y, z; h) = \{y_1 - \alpha y_2\}$$

Structural functions: example

- In the binary threshold crossing IV model with

$$Y_1 = 1[\alpha_1 Y_2 + U > 0] = \begin{cases} 1 & , \quad \alpha_1 Y_2 + U > 0 \\ 0 & , \quad \alpha_1 Y_2 + U \leq 0 \end{cases}$$

$$\mathcal{Y}(u, z; h) = \{(y_1, y_2) : y_2 \in \text{ support}(Y_2) \ \wedge \ y_1 = 1[\alpha_1 y_2 + u > 0]\}$$

$$\mathcal{U}(y, z; h) = \begin{cases} (-\infty, -\alpha_1 y_2] & , \quad y_1 = 0 \\ [-\alpha_1 y_2, \infty) & , \quad y_1 = 1 \end{cases}$$

Models, structures and data

- Models restrict **structural functions**, h, and conditional **probability distributions**:

$$\mathcal{G}_{U|Z} \equiv \{G_{U|Z=z} : z \in \mathcal{R}_Z\}$$

  - $G_{U|Z=z}(\mathcal{S})$ is the probability mass placed on a set $\mathcal{S}$ when $Z = z$.

- **Models** define which **structures** $(h, \mathcal{G}_{U|Z})$ are admissible.

- **Data** informs about distributions of observable variables:

$$\mathcal{F}_{Y|Z} \equiv \{F_{Y|Z=z} : z \in \mathcal{R}_Z\}$$

  - $F_{Y|Z=z}(\mathcal{T})$ is the "observed" probability mass placed on a set $\mathcal{T}$ when $Z = z$.

Identified set

- The **identified set** of structures delivered by a model and collection of distributions $\mathcal{F}_{Y|Z}$ is the collection of structures $(h, \mathcal{G}_{U|Z})$ that
  - (1) are admitted by the model,
  - (2) can deliver the distributions in the collection: $\mathcal{F}_{Y|Z}$.
- We provide characterizations of this identified set.
  - our development and proofs use results from random set theory, Artstein (1983), Molchanov (2005), Norberg(1992).
  - random set methods are **not needed** to understand and apply the results.

Characterizing sharp identified sets

- Recall

$$\mathcal{U}(y, z; h) \equiv \{u : h(y, z, u) = 0\}$$

and define $\mathcal{J}(\mathcal{S}, z; h)$:

$$\mathcal{J}(\mathcal{S}, z; h) \equiv \{y : \mathcal{U}(y, z; h) \subseteq \mathcal{S}\}$$

Characterizing sharp identified sets

- Recall

$$\mathcal{U}(y, z; h) \equiv \{u : h(y, z, u) = 0\}$$

and define $\mathcal{J}(\mathcal{S}, z; h)$:

$$\mathcal{J}(\mathcal{S}, z; h) \equiv \{y : \mathcal{U}(y, z; h) \subseteq \mathcal{S}\}$$

  - values of Y which function h says **can** only occur if $U \in \mathcal{S}$ when $Z = z$.

Characterizing sharp identified sets

- Recall

$$\mathcal{U}(y, z; h) \equiv \{u : h(y, z, u) = 0\}$$

and define $\mathcal{J}(\mathcal{S}, z; h)$:

$$\mathcal{J}(\mathcal{S}, z; h) \equiv \{y : \mathcal{U}(y, z; h) \subseteq \mathcal{S}\}$$

  - values of Y which function h says **can** only occur if $U \in \mathcal{S}$ when $Z = z$.

- **Theorem**: The identified set delivered by $\mathcal{F}_{Y|Z}$ and a model $\mathcal{M}$, comprises the structures $(h, \mathcal{G}_{U|Z})$ admitted by $\mathcal{M}$ that satisfy:

$$G_{U|Z=z}(\mathcal{S}) \geq F_{Y|Z=z}(\mathcal{J}(\mathcal{S}, z; h))$$

for all z and $\forall\ \mathcal{S}$ in a collection of closed sets, $\mathrm{Q}(h, z)$, on the support of U.

Characterizing sharp identified sets

- Recall

$$\mathcal{U}(y, z; h) \equiv \{u : h(y, z, u) = 0\}$$

and define $\mathcal{J}(\mathcal{S}, z; h)$:

$$\mathcal{J}(\mathcal{S}, z; h) \equiv \{y : \mathcal{U}(y, z; h) \subseteq \mathcal{S}\}$$

  - values of Y which function h says **can** only occur if $U \in \mathcal{S}$ when $Z = z$.

- **Theorem**: The identified set delivered by $\mathcal{F}_{Y|Z}$ and a model $\mathcal{M}$, comprises the structures $(h, \mathcal{G}_{U|Z})$ admitted by $\mathcal{M}$ that satisfy:

$$G_{U|Z=z}(\mathcal{S}) \geq F_{Y|Z=z}(\mathcal{J}(\mathcal{S}, z; h))$$

for all z and $\forall\ \mathcal{S}$ in a collection of closed sets, $\mathrm{Q}(h, z)$, on the support of U.

- Test sets $\mathrm{Q}(h, z)$ contains unions of the sets $\mathcal{U}(y, z; h)$ for $y \in R_{Y|Z=z}$.

A characterization of the sharp identified set

- **Theorem**: The identified set delivered by $\mathcal{F}_{Y|Z}$ and a model $\mathcal{M}$, comprises the structures $(h, \mathcal{G}_{U|Z})$ admitted by $\mathcal{M}$ that satisfy:

$$G_{U|Z=z}(\mathcal{S}) \geq F_{Y|Z=z}(\mathcal{J}(\mathcal{S}, z; h)) \qquad (*)$$

  for all z and $\forall\ \mathcal{S}$ in a collection of closed sets, $\mathbb{Q}(h, z)$,on the support of U.

  - If a model is complete or has point-valued residuals then (*) are equalities.
  - Classical identification results are delivered in these cases.
  - If $U \perp\!\!\!\perp Z$ the inequalities become

$$G_U(\mathcal{S}) \geq \sup_{z \in \mathcal{R}_Z} F_{Y|Z=z}(\mathcal{J}(\mathcal{S}, z; h))$$

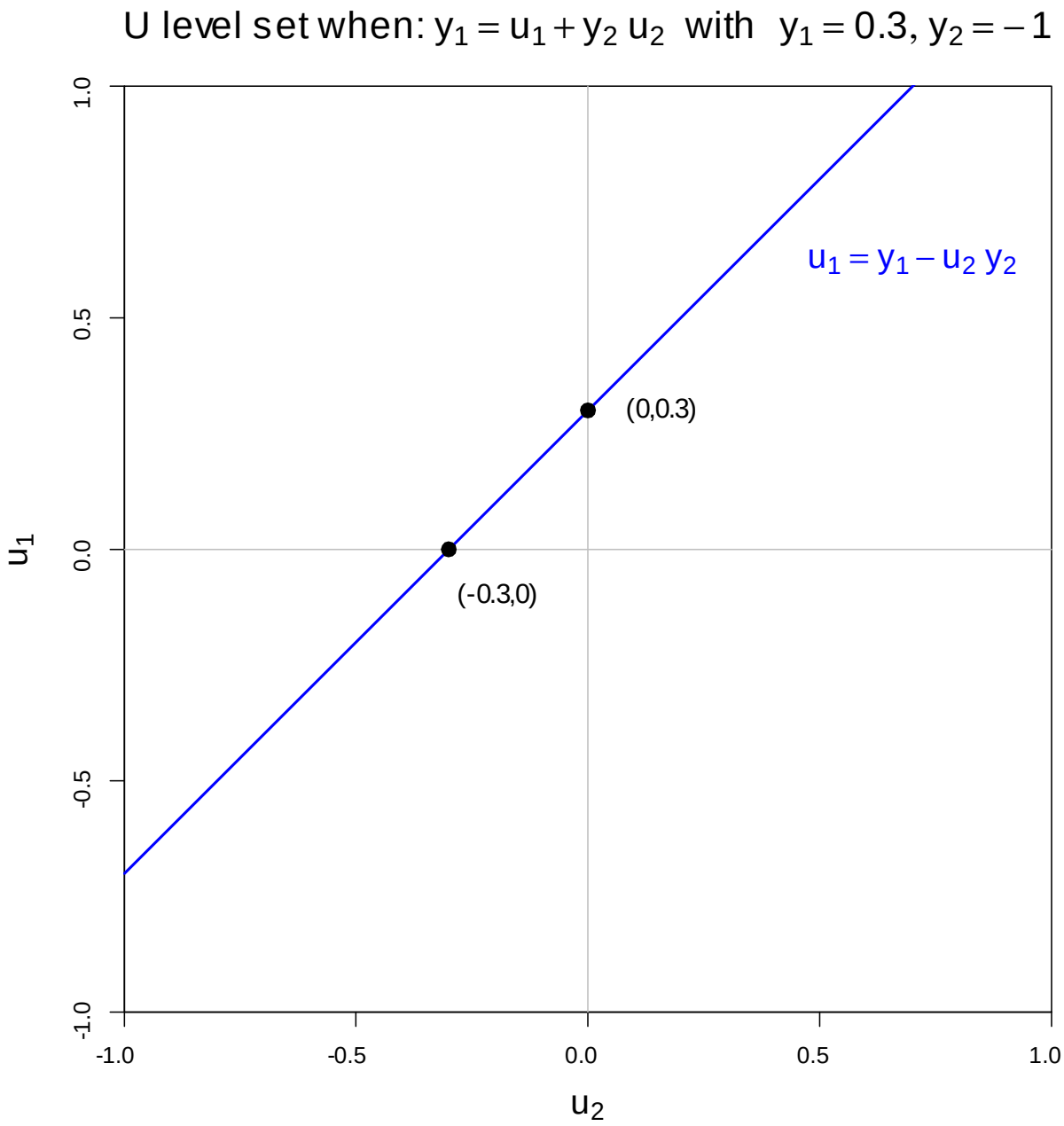
Random coefficients linear model

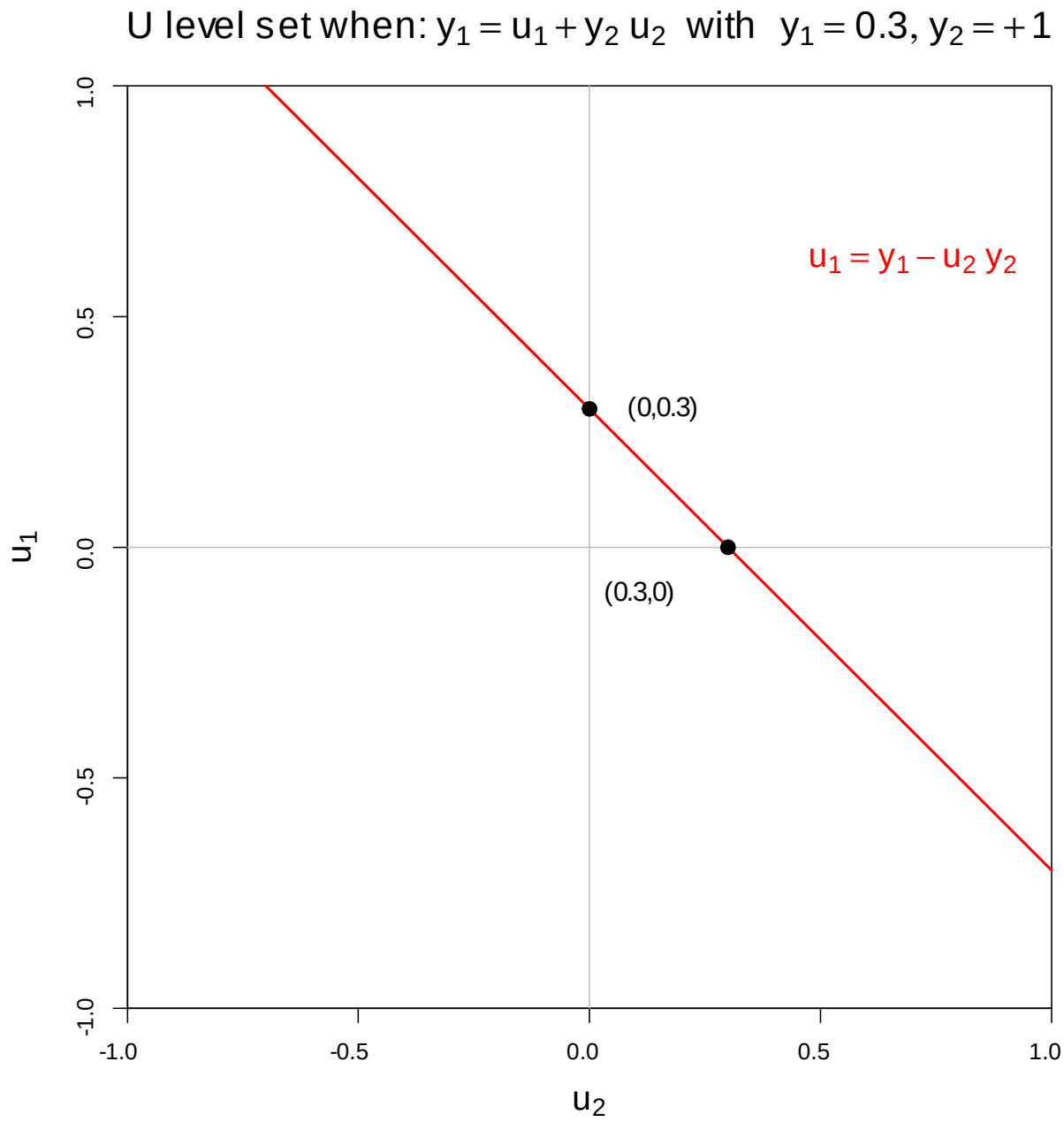
- A parametric IV random coefficients model:

$$Y_1 = U_1 + U_2 Y_2 \qquad U \equiv (U_1 . U_2) \sim N_2(\mu, \Sigma), \qquad U \perp\!\!\!\perp Z \in \mathcal{R}_Z$$

has:

$$\mathcal{U}(y, z; h) = \{u : u_1 = y_1 - u_2 y_2\}$$


U level set when: $y_1 = u_1 + y_2\, u_2$ with $y_1 = 0.3, y_2 = -1$
$u_1 = y_1 - u_2\, y_2$
(0,0.3)
(-0.3,0)
u_1
u_2
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0


U level set when: $y_1 = u_1 + y_2\, u_2$ with $y_1 = 0.3, y_2 = +1$
$u_1 = y_1 - u_2\, y_2$
(0,0.3)
(0.3,0)
u_1
u_2
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0

Random coefficients linear model

- A parametric IV random coefficients model:

$$Y_1 = U_1 + U_2 Y_2 \qquad U \equiv (U_1 . U_2) \sim N_2(\mu, \Sigma), \qquad U \perp\!\!\!\perp Z \in \mathcal{R}_Z$$

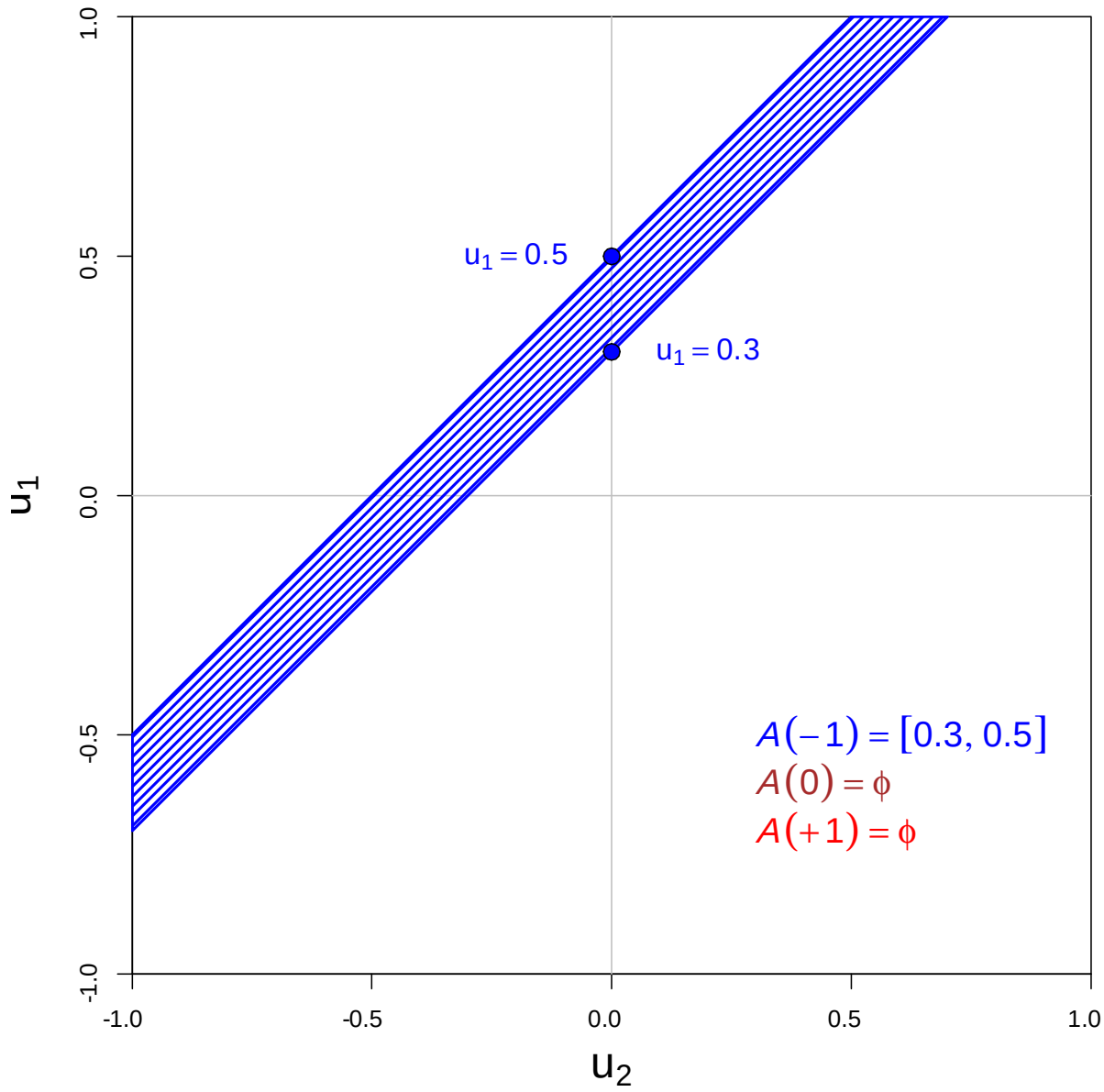
  has:

$$\mathcal{U}(y, z; h) = \{u : u_1 = y_1 - u_2 y_2\}$$

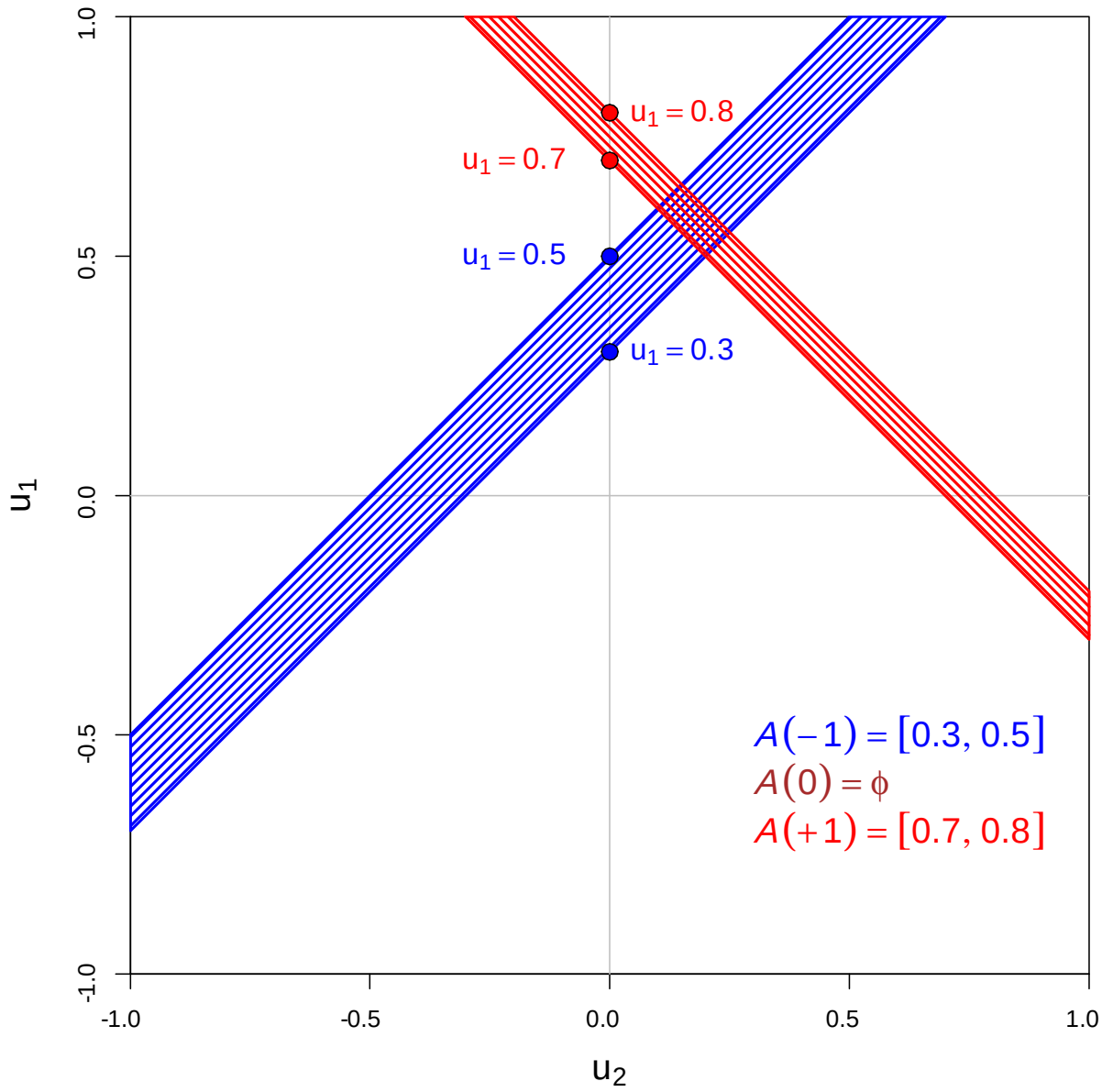
- Consider $Y_2 \in \{-1, 0, 1\}$. Let $\mathcal{A}(y_2)$ be a set of values of Y_1, specific to the value y_2 of Y_2, e.g. intervals, possibly empty.
- Test sets $\mathrm{Q}(h, z)$ are unions of sets $\mathcal{U}(y, z; h)$, thus.

$$\mathcal{S}(\mathcal{A}(-1), \mathcal{A}(0), \mathcal{A}(1)) \equiv \bigcup_{y_2 \in \{-1,0,1\}} \bigcup_{y_1 \in \mathcal{A}(y_2)} \{u : u_1 = y_1 - u_2 y_2\}$$

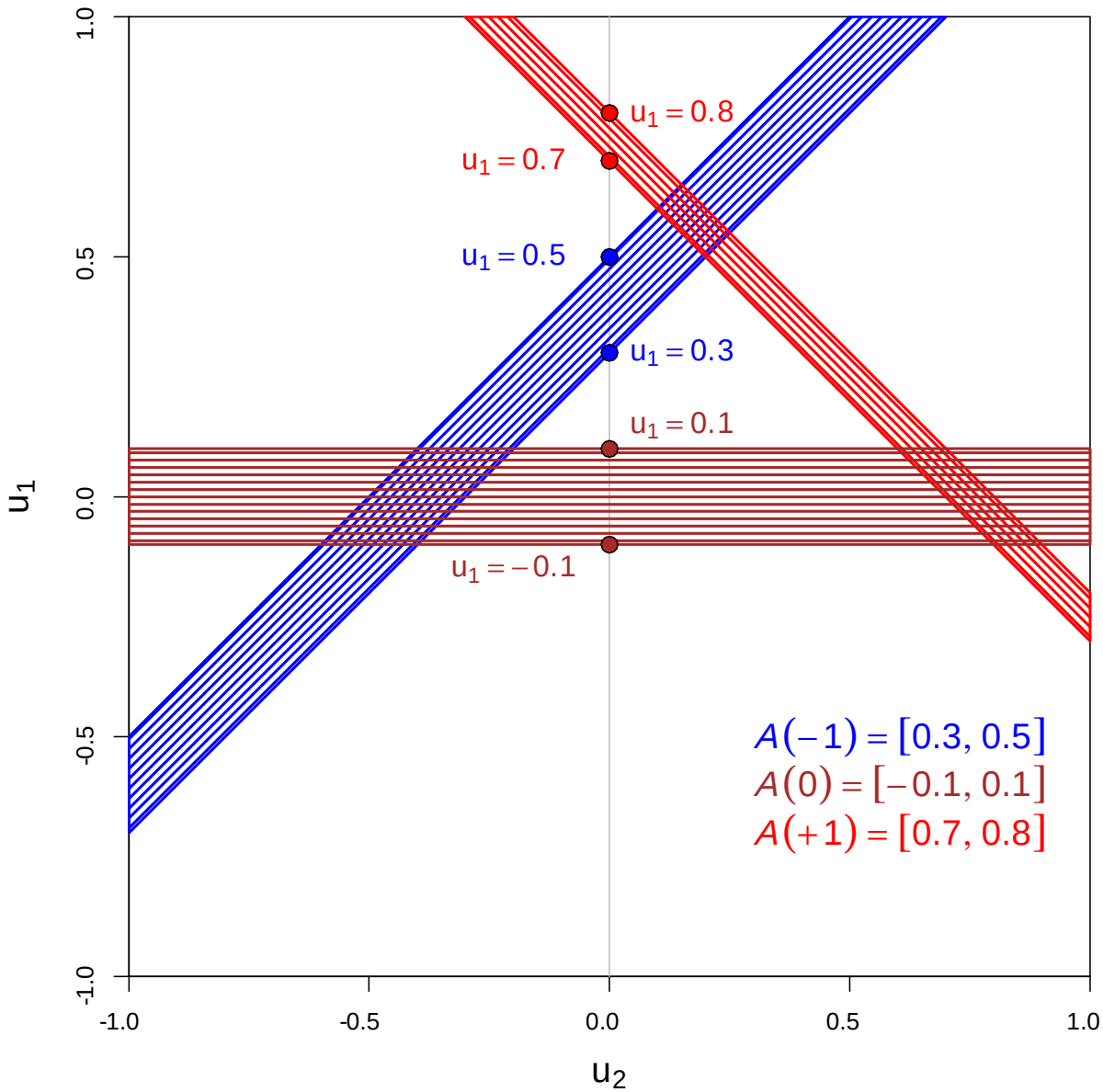
Union of U level sets: $S(A(-1), A(0), A(+1))$



Union of U level sets: $S(A(-1), A(0), A(+1))$



Union of U level sets: $S(A(-1), A(0), A(+1))$



Random coefficients linear model

- The identified set for $\theta = (\mu, \Sigma)$ comprises θ such that for all $\mathcal{A}(-1)$, $\mathcal{A}(0)$, $\mathcal{A}(1)$:

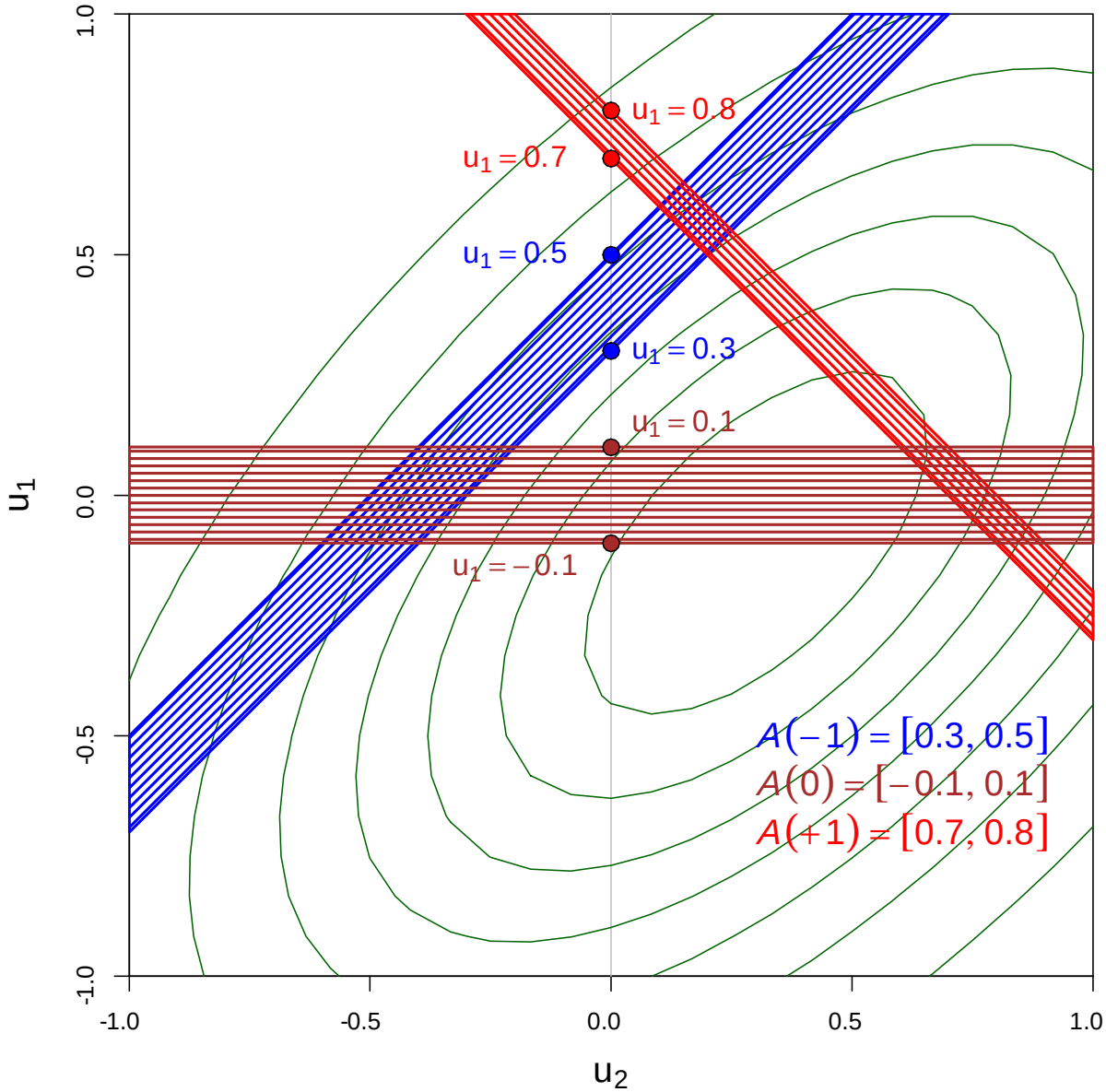
$$G_U(\mathcal{S}(\mathcal{A}(-1), \mathcal{A}(0), \mathcal{A}(1)), \theta) \geq \sup_{z \in \mathcal{R}_Z} (\mathbb{P}[\mathcal{U}(Y, Z; h) \subseteq G_U(\mathcal{S}(\mathcal{A}(-1), \mathcal{A}(0), \mathcal{A}(1)), \theta) | z])$$

that is:

$$G_U(\mathcal{S}(\mathcal{A}(-1), \mathcal{A}(0), \mathcal{A}(1)), \theta) \geq \sup_{z \in \mathcal{R}_Z} \left(\sum_{y_2 \in \{-1,0,1\}} \mathbb{P}\left[Y_1 \in \mathcal{A}(y_2) \wedge Y_2 = y_2 | z \right] \right)$$

where $G_U(\mathcal{S}(\mathcal{A}(-1), \mathcal{A}(0), \mathcal{A}(1)), \theta)$ is the volume under a $N_2(\mu, \Sigma)$ density supported on $\mathcal{S}(\mathcal{A}(-1), \mathcal{A}(0), \mathcal{A}(1))$.

Union of U level sets: $S(A(-1), A(0), A(+1))$



Random coefficients linear model: calculations

- We generate probability distributions for Y given $Z = z$ using a triangular structure

$$\begin{aligned} Y_1 &= U_1 + U_2 Y_2 \\ Y_2 &= -1[d \times Z + U_3 < -0.5] + 1[d \times Z + U_3 > 0.5] \end{aligned}$$

$$U \sim N_3(0, \Sigma) \qquad U \perp\!\!\!\perp Z \in \{-2, -1, 0, 1, 2\}$$

$$\Sigma = \begin{bmatrix} 1 & 0 & 0.5 \\ 0 & 1 & -0.5 \\ 0.5 & -0.5 & 1 \end{bmatrix} \qquad d \in \{0.5, 1\}$$

and collections of intervals, $\mathcal{A}(-1)$, $\mathcal{A}(0)$, $\mathcal{A}(1)$. We calculate probabilities in the inequalities

$$G_U(\mathcal{S}(\mathcal{A}(-1), \mathcal{A}(0), \mathcal{A}(1)), \theta) \geq \sup_{z \in \mathcal{R}_Z} \left(\sum_{y_2 \in \{-1,0,1\}} \mathbb{P}\left[Y_1 \in \mathcal{A}(y_2) \wedge Y_2 = y_2 | z \right] \right)$$

and projections of the identified set for θ onto each parameter axis in turn.

Random coefficients linear model: projections

Parameter	Value in Structures	Instrument Strength	Lower bound	Upper bound
μ_1	0	$d = 1.0$	-0.130	0.007
		$d = 0.5$	-0.366	0.161
μ_2	0	$d = 1.0$	-0.028	0.050
		$d = 0.5$	-0.209	0.667
σ_{11}	1	$d = 1.0$	0.879	1.082
		$d = 0.5$	0.791	1.489
σ_{12}	0	$d = 1.0$	-0.021	0.095
		$d = 0.5$	-0.219	0.563
σ_{22}	1	$d = 1.0$	0.897	1.082
		$d = 0.5$	0.663	2.080

Summary

- We have extended the scope of application of incomplete IV models to cases in which unobservables are set-valued functions of observable variables.
  - Now incomplete models can be employed when outcomes are discrete or there is non-scalar heterogeneity or models are defined in terms of inequalities.
  - We have results for binary outcomes - Chesher (Ecta 2010, ET 2013), ordered outcomes - Chesher and Smolinski (JoE 2012), multiple discrete choice - Chesher and Rosen (QE 2013), random coefficients binary outcome - Chesher and Rosen (EctJ 2014), and in development, joint with Adam: auction models, models of games with mutiple equilibria, endogenous variables measured with error, endogenous censoring, ...
- Incomplete models with set-valued residuals are generically partially identifying
  - we can characterize identified sets as systems of moment inequalities.

Challenges

- Challenges include:
  - Learning from data about relevant identified-set-defining inequalities,
  - Calculation of and inference on projections when parameters are high dimensional.

Challenges

- Challenges include:
  - Calculating exact distributions of set IV estimators - that's one for .... Grant?