ECON0108 2022-23 Part 1

Slides for Lecture 5

Andrew Chesher

CeMMAP & UCL

A general approach

- I set out an approach to determining identified sets of structures in a wide class of models.

- This delivers all the results presented so far and also results for models of strategic interaction e.g. firm entry, product choice, English auctions.

- Formal exposition in Chesher and Rosen (Ecta, 2017) (CR),
  - informal exposition in Ch1 of Vol 7A of the *Handbook of Econometrics* (2020) including an application,
  - application to a dynamic IO process in Berry and Compiani, (Annual Review of Economics, 2021) and (REStud, forthcoming).

Notation and structures

- Notation:
  - Y observed endogenous outcomes
  - $Z \in \mathcal{R}_Z$ observed exogenous variables
  - U unobserved variables.

- A model defines *admissible* structures $(h, \mathcal{G}_{U|Z})$ where $h : \mathcal{R}_{YZU} \to \mathbb{R}$ determines the values of (Y, Z, U) that can occur

$$\mathbb{P}[h(Y, Z, U) = 0] = 1$$

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

$$G_{U|Z=z}(\mathcal{S}) \equiv \mathbb{P}[U \in \mathcal{S}|Z = z]$$

Y sets and U sets

- Define:

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

the values of Y that can occur when $Z = z$ and $U = u$.

- Define:

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

the values of U that can deliver $Y = y$ when $Z = z$ - **residual sets**.

- Complete models have singleton sets $\mathcal{Y}(u, z; h)$, non-intersecting $\mathcal{U}(y, z; h)$.

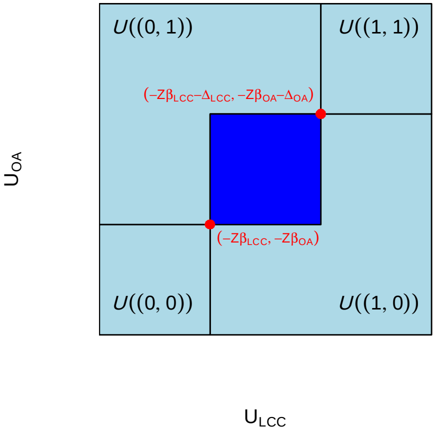
Example: Kline and Tamer (2016) (KT16)

- Data on 7882 markets - air routes with two airline types, Low Cost Carriers(LCC) and Other Airlines (OA).
- Binary Y_{LCC} and Y_{OA} indicate the presence of respectively LCC and OA operating on an air route in the USA. There are exogenous variables listed in vector $Z \in \mathcal{R}_Z$, structural equations

$$\begin{aligned} Y_{LCC} &= 1\left[Z\beta_{LCC} + Y_{OA}\Delta_{LCC} + U_{LCC} > 0\right] \\ Y_{OA} &= 1\left[Z\beta_{OA} + Y_{LCC}\Delta_{OA} + U_{OA} > 0\right] \end{aligned}$$

- This type of model is studied in many papers including Heckman (1978), Bresnahan and Reiss (1990,1991), and Tamer (2003).
- In KT16 and most other applications of this model U is restricted to be normally distributed independent of Z.

$\Delta_{LCC} < 0$ and $\Delta_{OA} < 0$



Partially identifying incomplete models

- **Incomplete** models are generically **partially identifying** when U sets are *non-singleton*.
  - occurs with discrete Y_1, e.g.

$$Y_1 = 1[\alpha_0 + \alpha_1 Y_2 < U]$$

  - occurs in random coefficient models:

$$Y_1 = (\alpha_0 + U_0) + (\alpha_1 + U_1) Y_2$$

  - occurs in models involving inequality restrictions e.g. Kline & Tamer (QE, 2016), Mazzeo (Rand, 2002).

A binary outcome example

- Define $y = (y_1, y_2)$

$$y_1 = 1[\alpha_0 + \alpha_1 y_2 + \beta' z < u]$$

so

$$h(y, z, u) = y_1 - 1[\alpha_0 + \alpha_1 y_2 + \beta' z < u]$$

- There are **non-singleton** U sets.

$$\mathcal{U}(y, z; h) = \begin{cases} (-\infty, \alpha_0 + \alpha_1 y_2 + \beta' z] & \text{when} \quad y_1 = 0 \\ (\alpha_0 + \alpha_1 y_2 + \beta' z, +\infty) & \text{when} \quad y_1 = 1 \end{cases}$$

Plan

- Now, give a characterization of identified sets for complete or **incomplete** models with singleton or **nonsingleton U sets**.
- Then study an example using the data employed in the Angrist and Evans (AER, 1998) analysis of labour force participation..

What data tell us

- Data informs about distributions of observable variables:

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

where for a set $\mathcal{T} \subset \mathcal{R}_Y$:

$$F_{Y|Z=z}(\mathcal{T}) \equiv \mathbb{P}[Y \in \mathcal{T}|Z = z]$$

- The identified set of structures delivered by a model and distributions $\mathcal{F}_{Y|Z}$ comprises the collection of admissible structures that **can deliver** the distributions $\mathcal{F}_{Y|Z}$.

- The analysis uses methods drawn from the theory of **random sets**, Molchanov (2005). I now *sketch* the development.

Observational equivalence

- A structure delivers a **set** of values of outcome Y at $U = u$ when $Z = z$.

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

Observational equivalence

- A structure delivers a **set** of values of outcome Y at $U = u$ when $Z = z$.

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

- When $U \sim G_{U|Z=z}$ the set $\mathcal{Y}(U, z; h)$ is a **random set** - characterized by its **selections**.

Observational equivalence

- A structure delivers a **set** of values of outcome Y at $U = u$ when $Z = z$.

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

- When $U \sim G_{U|Z=z}$ the set $\mathcal{Y}(U, z; h)$ is a **random set** - characterized by its **selections**.
    - The selections of a random set are the random variables that lie in the set with probability 1.

Observational equivalence

- A structure delivers a **set** of values of outcome Y at $U = u$ when $Z = z$.

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

- When $U \sim G_{U|Z=z}$ the set $\mathcal{Y}(U, z; h)$ is a **random set** - characterized by its **selections**.
  - The selections of a random set are the random variables that lie in the set with probability 1.

- $F_{Y|Z=z}$ is said to be **selectionable** with respect to the distribution of $\mathcal{Y}(U, z; h)$ when $F_{Y|Z=z}$ is the distribution of a selection of $\mathcal{Y}(U, z; h)$.

Observational equivalence

- A structure delivers a **set** of values of outcome Y at $U = u$ when $Z = z$.

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

- When $U \sim G_{U|Z=z}$ the set $\mathcal{Y}(U, z; h)$ is a **random set** - characterized by its **selections**.
  - The selections of a random set are the random variables that lie in the set with probability 1.

- $F_{Y|Z=z}$ is said to be **selectionable** with respect to the distribution of $\mathcal{Y}(U, z; h)$ when $F_{Y|Z=z}$ is the distribution of a selection of $\mathcal{Y}(U, z; h)$.
  - Notation:

$$F_{Y|Z=z} \sqsubset \mathcal{Y}(U, z; h)$$

Observational equivalence

- A structure delivers a **set** of values of outcome Y at $U = u$ when $Z = z$.

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

- When $U \sim G_{U|Z=z}$ the set $\mathcal{Y}(U, z; h)$ is a **random set** - characterized by its **selections**.
    - The selections of a random set are the random variables that lie in the set with probability 1.

- $F_{Y|Z=z}$ is said to be **selectionable** with respect to the distribution of $\mathcal{Y}(U, z; h)$ when $F_{Y|Z=z}$ is the distribution of a selection of $\mathcal{Y}(U, z; h)$.
    - Notation:

$$F_{Y|Z=z} \sqsubset \mathcal{Y}(U, z; h)$$

- If for all z, $F_{Y|Z=z}$ is the distribution of a **selection** of $\mathcal{Y}(U, z; h)$ with $U \sim G_{U|Z=z}$ then $(h, \mathcal{G}_{U|Z})$ "can deliver" $\mathcal{F}_{Y|Z}$.

Observational equivalence

- The set of observationally equivalent structures delivering $\mathcal{F}_{Y|Z}$ is:

$$\Big\{(h, \mathcal{G}_{U|Z}) : \forall F_{Y|Z} \in \mathcal{F}_{Y|Z}, \quad \forall z \in \mathcal{R}_Z, \\ F_{Y|Z=z} \sqsubset \mathcal{Y}(U, z; h), \text{ when } U \sim G_{U|Z=z}\Big\}$$

Observational equivalence

- The set of observationally equivalent structures delivering $\mathcal{F}_{Y|Z}$ is:

$$\Big\{(h, \mathcal{G}_{U|Z}) : \forall F_{Y|Z} \in \mathcal{F}_{Y|Z}, \quad \forall z \in \mathcal{R}_Z, \\ F_{Y|Z=z} \sqsubset \mathcal{Y}(U, z; h), \text{ when } U \sim G_{U|Z=z}\Big\}$$

- The identified set of structures delivered by model $\mathcal{M}$ and distributions $\mathcal{F}_{Y|Z}$ comprises the members of this set that are admitted by $\mathcal{M}$.

$$\Big\{ S \ : \ S \in \mathcal{M} \text{ and } \forall F_{Y|Z} \in \mathcal{F}_{Y|Z}, \quad \forall z \in \mathcal{R}_Z, \\ F_{Y|Z=z} \sqsubset \mathcal{Y}(U, z; h^S) \text{ when } U \sim G^S_{U|Z=z}\Big\}$$

A duality property

- There is a duality property of Y and U level sets.

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

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

as follows.

  - For any z and any h:

$$y^* \in \mathcal{Y}(u^*, z; h) \quad \text{if and only if} \quad u^* \in \mathcal{U}(y^*, z; h)$$

because each inclusion occurs if and only if

$$h(y^*, z, u^*) = 0$$

Consequence of the duality property

- A consequence of this duality is (CR):

$$F_{Y|Z=z} \sqsubset \mathcal{Y}(U, z; h) \quad \text{when} \quad U \sim G_{U|Z=z}$$

$$\text{if and only if}$$

$$G_{U|Z=z} \sqsubset \mathcal{U}(Y, z; h) \quad \text{with} \quad Y \sim F_{Y|Z=z}$$

Consequence of the duality property

- A consequence of this duality is (CR):

$$F_{Y|Z=z} \sqsubset \mathcal{Y}(U,z;h) \quad \text{when} \quad U \sim G_{U|Z=z}$$
$$\text{if and only if}$$
$$G_{U|Z=z} \sqsubset \mathcal{U}(Y,z;h) \quad \text{with} \quad Y \sim F_{Y|Z=z}$$

- So, the set of observationally equivalent structures for $\mathcal{F}_{Y|Z}$ is:

$$\left\{(h, \mathcal{G}_{U|Z}) : \forall z \quad F_{Y|Z=z} \sqsubset \mathcal{Y}(U,z;h) \quad \text{when} \quad U \sim G_{U|Z=z}\right\}$$

is identical to the set

$$\left\{(h, \mathcal{G}_{U|Z}) : \forall z \quad G_{U|Z=z} \sqsubset \mathcal{U}(Y,z;h) \quad \text{when} \quad Y \sim F_{Y|Z=z}\right\} \qquad (*)$$

Consequence of the duality property

- A consequence of this duality is (CR):

$$F_{Y|Z=z} \sqsubset \mathcal{Y}(U, z; h) \quad \text{when} \quad U \sim G_{U|Z=z}$$
$$\text{if and only if}$$
$$G_{U|Z=z} \sqsubset \mathcal{U}(Y, z; h) \quad \text{with} \quad Y \sim F_{Y|Z=z}$$

- So, the set of observationally equivalent structures for $\mathcal{F}_{Y|Z}$ is:

$$\left\{(h, \mathcal{G}_{U|Z}) : \forall z \quad F_{Y|Z=z} \sqsubset \mathcal{Y}(U, z; h) \quad \text{when} \quad U \sim G_{U|Z=z}\right\}$$

  is identical to the set

$$\left\{(h, \mathcal{G}_{U|Z}) : \forall z \quad G_{U|Z=z} \sqsubset \mathcal{U}(Y, z; h) \quad \text{when} \quad Y \sim F_{Y|Z=z}\right\} \qquad (*)$$

- The structures in the set (*) admitted by a model $\mathcal{M}$ comprise the **identified set of structures** delivered by $\mathcal{M}$ and $\mathcal{F}_{Y|Z}$.

Characterizing selectionability

- **Artstein's inequality**. The distribution of random variable A is selectionable with respect to the distribution of random set $\mathcal{A}$ if and only if

$$\mathbb{P}[A \in \mathcal{S}] \geq \mathbb{P}[\mathcal{A} \subseteq \mathcal{S}]$$

for all closed sets $\mathcal{S}$, Artstein (Is J Math, 1983).

Characterizing selectionability

- **Artstein's inequality**. The distribution of random variable A is selectionable with respect to the distribution of random set $\mathcal{A}$ if and only if

$$\mathbb{P}[A \in \mathcal{S}] \geq \mathbb{P}[\mathcal{A} \subseteq \mathcal{S}]$$

for all closed sets $\mathcal{S}$, Artstein (Is J Math, 1983).

- Applying Artstein's inequality the identified set comprises the admissible structures $(h, \mathcal{G}_{U|Z})$ such that for all closed sets $\mathcal{S}$

$$\forall z : G_{U|Z=z}(\mathcal{S}) \geq \mathbb{P}[\mathcal{U}(Y, z; h) \subseteq \mathcal{S} | Z = z].$$

Characterizing selectionability

- **Artstein's inequality**. The distribution of random variable A is selectionable with respect to the distribution of random set $\mathcal{A}$ if and only if

$$\mathbb{P}[A \in \mathcal{S}] \geq \mathbb{P}[\mathcal{A} \subseteq \mathcal{S}]$$

  for all closed sets $\mathcal{S}$, Artstein (Is J Math, 1983).

- Applying Artstein's inequality the identified set comprises the admissible structures $(h, \mathcal{G}_{U|Z})$ such that for all closed sets $\mathcal{S}$

$$\forall z : G_{U|Z=z}(\mathcal{S}) \geq \mathbb{P}[\mathcal{U}(Y, z; h) \subseteq \mathcal{S} | Z = z].$$

  - Simplification: only need to consider sets $\mathcal{S}$ in a collection $\mathsf{Q}(h, z)$ comprising the connected unions of sets $\mathcal{U}(y, z; h)$.

Characterizing identified sets of structures

- The set
$$\mathcal{A}(\mathcal{S}, z; h) \equiv \{y : \mathcal{U}(y, z; h) \subseteq \mathcal{S}\}$$
is the set of values of Y that *only* occur when $U \in \mathcal{S}$, and
$$Y \in \mathcal{A}(\mathcal{S}, z; h) \implies U \in \mathcal{S}$$

- So, for all z sets $\mathcal{S}$, if $G_{U|Z=z} \sqsubset \mathcal{U}(Y, z; h)$ then
$$\begin{aligned} \mathbb{P}[Y \in \mathcal{A}(\mathcal{S}, z; h)|z] &\leq \mathbb{P}[U \in \mathcal{S}|z] \\ &= G_{U|Z=z}(\mathcal{S}) \end{aligned}$$
but if $G_{U|Z=z} \sqsubset \mathcal{U}(Y, z; h)$ *does not hold* $\exists \mathcal{S}$ such that the inequality is *violated*.

Summary

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

$$G_{U|Z=z}(\mathcal{S}) \geq F_{Y|Z=z}(\mathcal{A}(\mathcal{S}, z; h)), \quad \text{for all } z \in \mathcal{R}_Z$$

for all sets $\mathcal{S}$ in a collection of sets $\mathrm{Q}(h, z)$, where

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

  - on the **left** - the probability that $U \in \mathcal{S}$ according to the structure's distribution of U,
  - on the **right** - the (identifiable) probability of the values of Y occurring only if $U \in \mathcal{S}$, according to the structure's function h, and $\mathcal{F}_{Y|Z}$.

Independence of U and Z

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

$$G_{U|Z=z}(\mathcal{S}) \geq F_{Y|Z=z}(\mathcal{A}(\mathcal{S}, z; h)), \quad \text{for all } z \in \mathcal{R}_Z$$

for all sets $\mathcal{S}$ in a core determining collection of sets in $\mathcal{R}_U$.

  - simplifies under **independence** $U \perp\!\!\!\perp Z$.

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

Application: binary endogenous variables

- The model specifies

$$Y_1 = s(Y_2, U) \equiv \begin{cases} 1 & , \quad g(Y_2) \leq U \\ 0 & , \quad g(Y_2) \geq U \end{cases} \quad \text{and} \quad U \perp\!\!\!\perp Z$$

with $U \sim Unif(0,1)$ and binary $Y_2 \in \{0,1\}$, so:

$$h(Y, Z, U) = Y_1 - s(Y_2, U)$$

Application: binary endogenous variables

- The model specifies

$$Y_1 = s(Y_2, U) \equiv \begin{cases} 1 & , \quad g(Y_2) \leq U \\ 0 & , \quad g(Y_2) \geq U \end{cases} \quad \text{and} \quad U \perp\!\!\!\perp Z$$

with $U \sim Unif(0,1)$ and binary $Y_2 \in \{0,1\}$, so:

$$h(Y, Z, U) = Y_1 - s(Y_2, U)$$

- There are residual sets, $\mathcal{U}(y, z; h) \equiv \{u : h(y, z, u) = 0)\}$

$$\begin{array}{ll} \mathcal{U}((0,0), z; h) = [0, g(0)] & \qquad \mathcal{U}((0,1), z; h) = [0, g(1)] \\ \mathcal{U}((1,0), z; h) = [g(0), 1] & \qquad \mathcal{U}((1,1), z; h) = [g(1), 1] \end{array}$$

Binary endogenous variables - probit IV

$$Y_1 = s(Y_2, U) \equiv \begin{cases} 1 & , \quad g(Y_2) \leq U \\ 0 & , \quad g(Y_2) \geq U \end{cases} \quad \text{and} \quad U \perp\!\!\!\perp Z \quad U \sim Unif(0,1)$$

- In a IV *probit* model with Φ denoting the standard Gaussian distribution function

$$g(Y_2) = \Phi(a + bY_2)$$

and

$$Y_1 = s(Y_2, U) \equiv \begin{cases} 1 & , \quad \Phi(a + bY_2) \leq U \\ 0 & , \quad \Phi(a + bY_2) \geq U \end{cases} \quad \text{and} \quad U \perp\!\!\!\perp Z$$

equivalently, with $\Phi^{-1}(U) \sim N(0,1)$

$$Y_1 = s(Y_2, U) \equiv \begin{cases} 1 & , \quad a + bY_2 \leq \Phi^{-1}(U) \\ 0 & , \quad a + bY_2 \geq \Phi^{-1}(U) \end{cases} \quad \text{and} \quad U \perp\!\!\!\perp Z$$

Application: binary endogenous variables

$$Y_1 = s(Y_2, U) \equiv \begin{cases} 1 & , \quad g(Y_2) \leq U \\ 0 & , \quad g(Y_2) \geq U \end{cases} \quad \text{and} \quad U \perp\!\!\!\perp Z$$

$$h(Y, Z, U) = Y_1 - s(Y_2, U)$$

$$\mathcal{U}((0,0), z; h) = [0, g(0)] \qquad \mathcal{U}((0,1), z; h) = [0, g(1)]$$
$$\mathcal{U}((1,0), z; h) = [g(0), 1] \qquad \mathcal{U}((1,1), z; h) = [g(1), 1]$$

Application: binary endogenous variables

$$Y_1 = s(Y_2, U) \equiv \begin{cases} 1 & , \quad g(Y_2) \leq U \\ 0 & , \quad g(Y_2) \geq U \end{cases} \quad \text{and} \quad U \perp\!\!\!\perp Z$$

$$h(Y, Z, U) = Y_1 - s(Y_2, U)$$

$$\begin{array}{ll} \mathcal{U}((0,0), z; h) = [0, g(0)] & \qquad \mathcal{U}((0,1), z; h) = [0, g(1)] \\ \mathcal{U}((1,0), z; h) = [g(0), 1] & \qquad \mathcal{U}((1,1), z; h) = [g(1), 1] \end{array}$$

- The set of values of Y that *only* occur when $U \in \mathcal{S}$ is

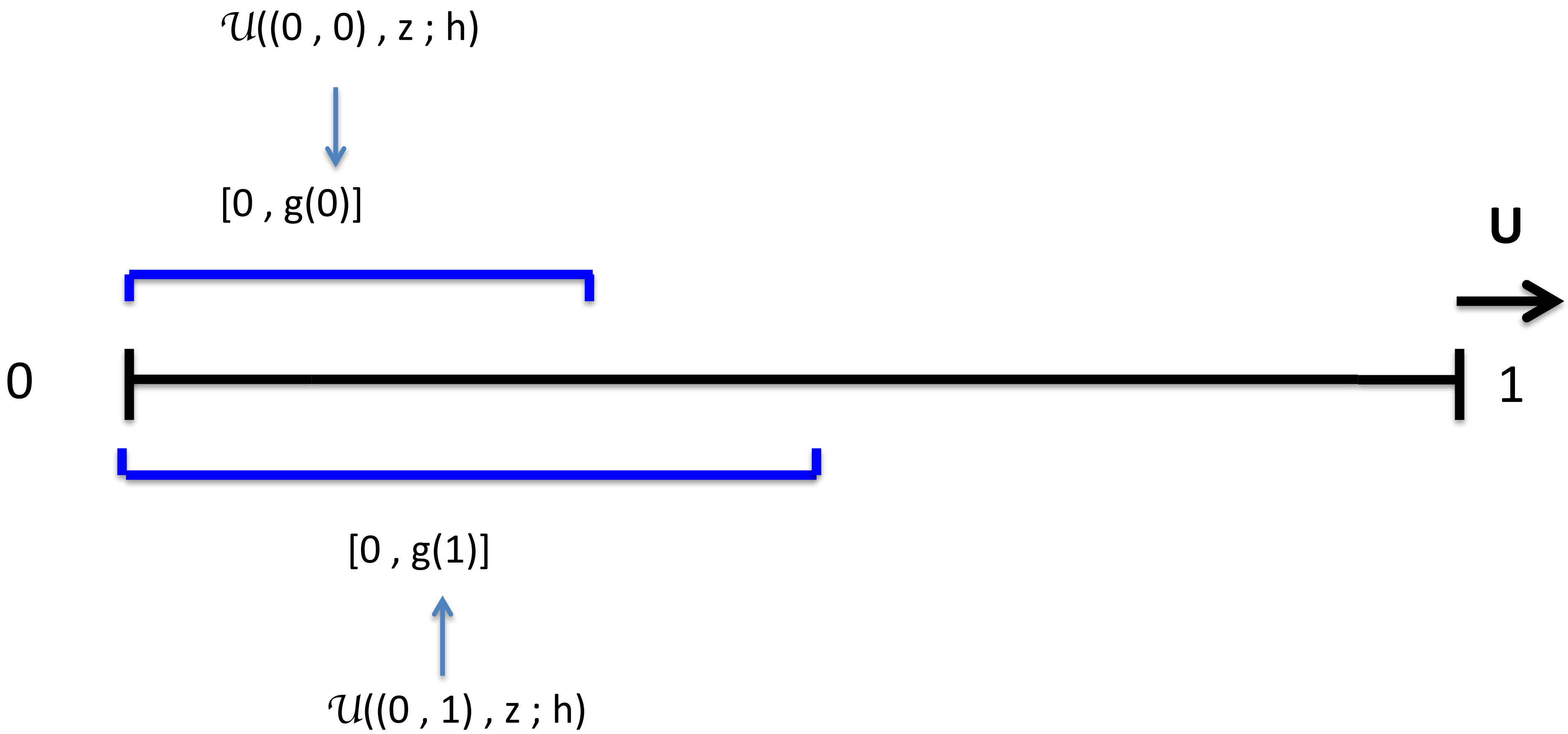
$$\mathcal{A}(\mathcal{S}, z; h) \equiv \{y : \mathcal{U}(y, z; h) \subseteq \mathcal{S}\}$$

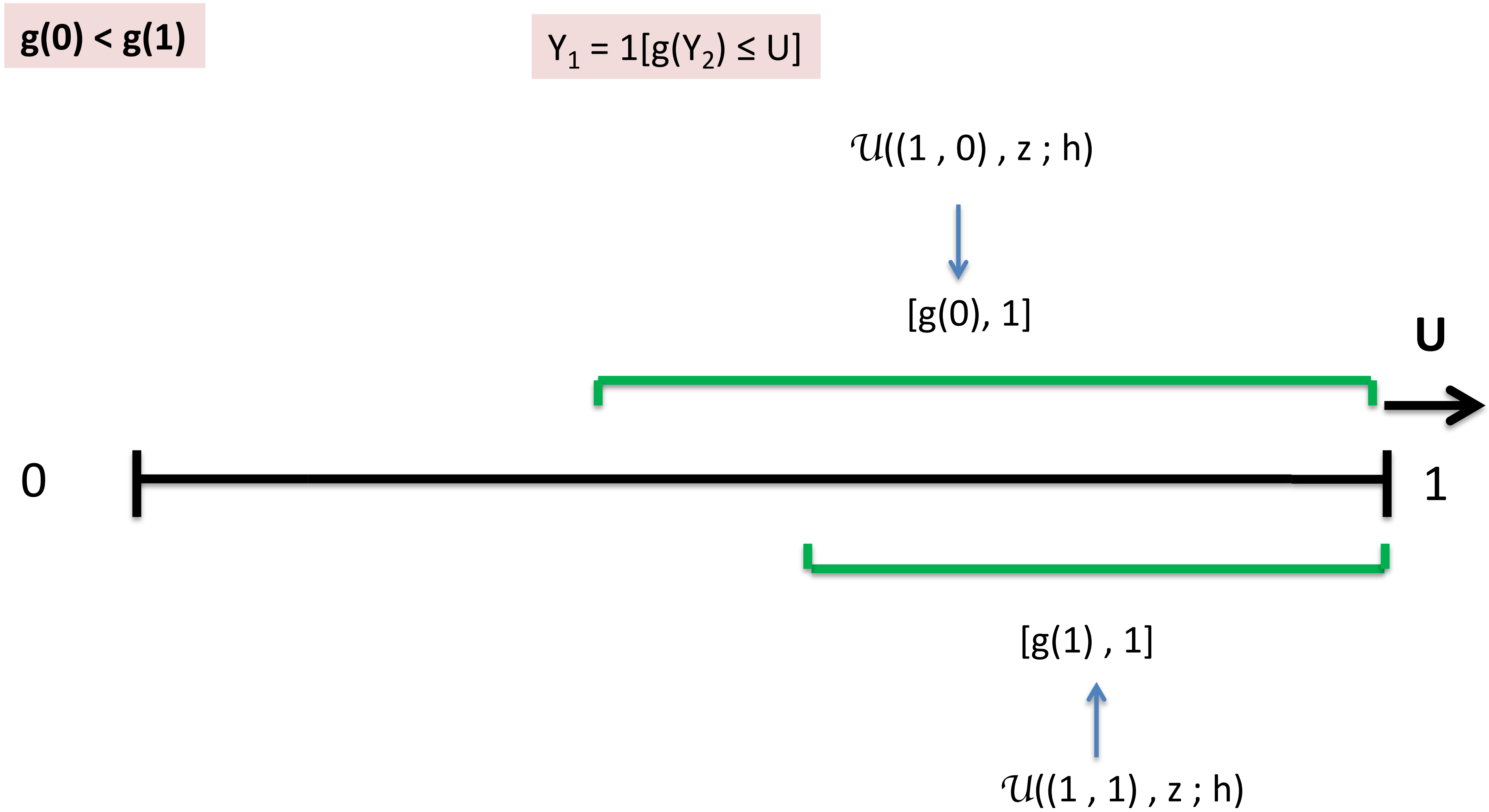
- The set $\mathcal{A}(\mathcal{S}, z; h)$ for $\mathcal{S} = [0, g(1)]$ is

$$\mathcal{A}([0, g(1)], z; h) = \begin{cases} \{(0,1)\} & \text{when} \quad g(0) > g(1) \\ \{(0,0), (0,1)\} & \text{when} \quad g(0) \leq g(1) \end{cases}$$

g(0) < g(1)

$Y_1 = 1[g(Y_2) \leq U]$





g(0) < g(1)
$Y_1 = 1[g(Y_2) \leq U]$
$\mathcal{U}((1, 0), z; h)$
[g(0), 1]
U
0
1
[g(1), 1]
$\mathcal{U}((1, 1), z; h)$

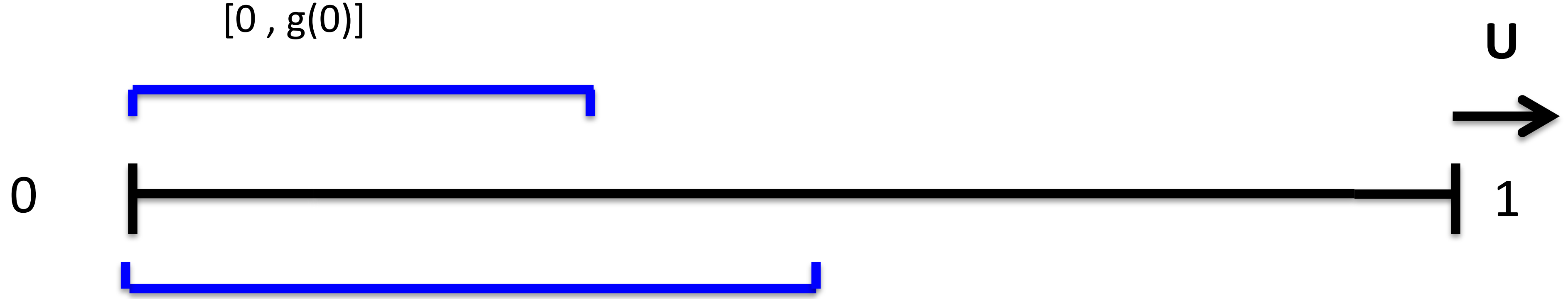
$g(0) < g(1)$

$Y_1 = 1[g(Y_2) \leq U]$

$\mathcal{U}((0\,,0)\,, z\,; h)$

$\downarrow$

$[0\,, g(0)]$



$[0\,, g(1)]$

$\uparrow$

$\mathcal{U}((0\,,1)\,, z\,; h)$

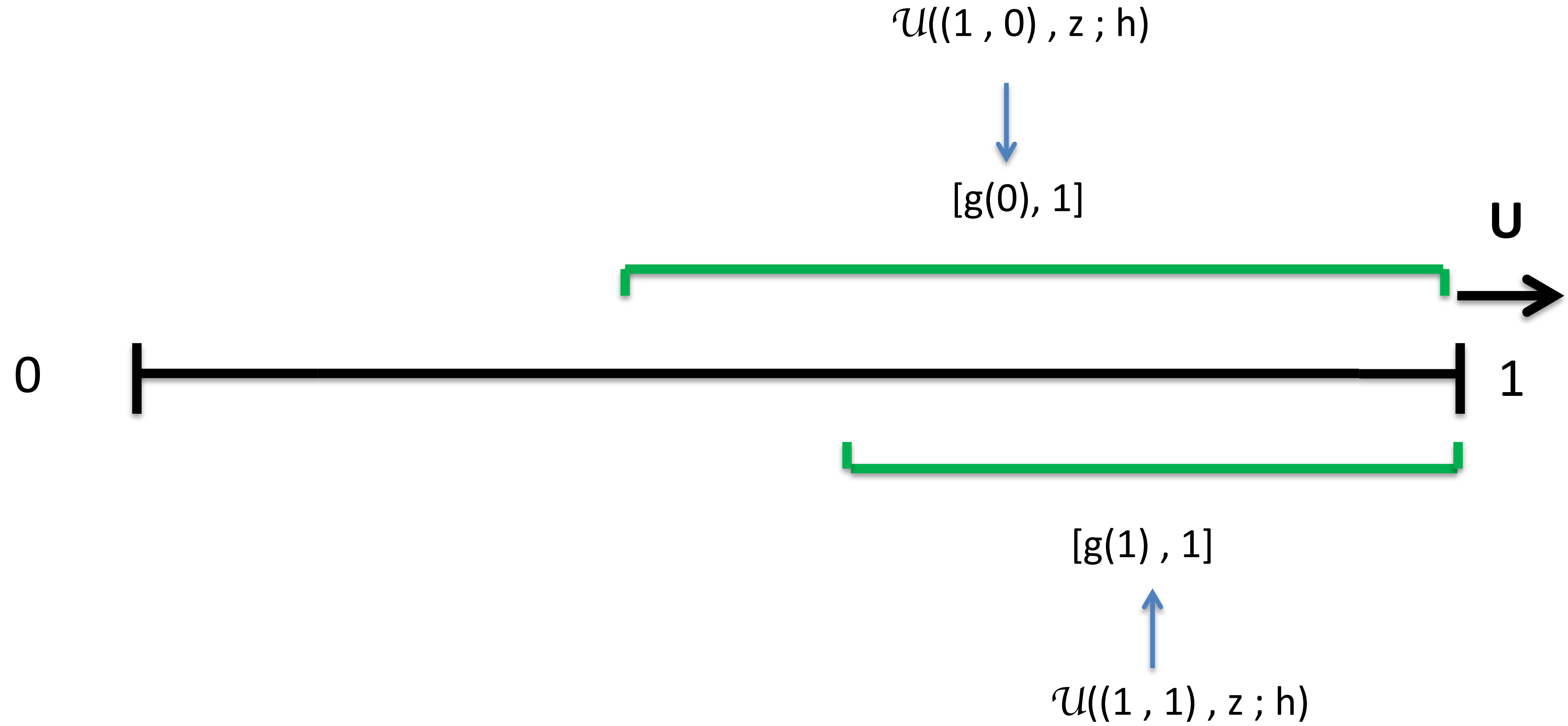
$\mathcal{A}(\mathcal{S}\,, z\,; h) = \{\, y : \mathcal{U}(y\,, z\,; h) \subseteq \mathcal{S} \,\}$

$\mathcal{A}(\,[0\,, g(0)]\,, z\,; h) = \{\,(0,0)\,\}$

$\mathcal{A}(\,[0\,, g(1)]\,, z\,; h) = \{\,(0,0)\,, (0,1)\,\}$

$g(0) < g(1)$

$Y_1 = 1[g(Y_2) \leq U]$



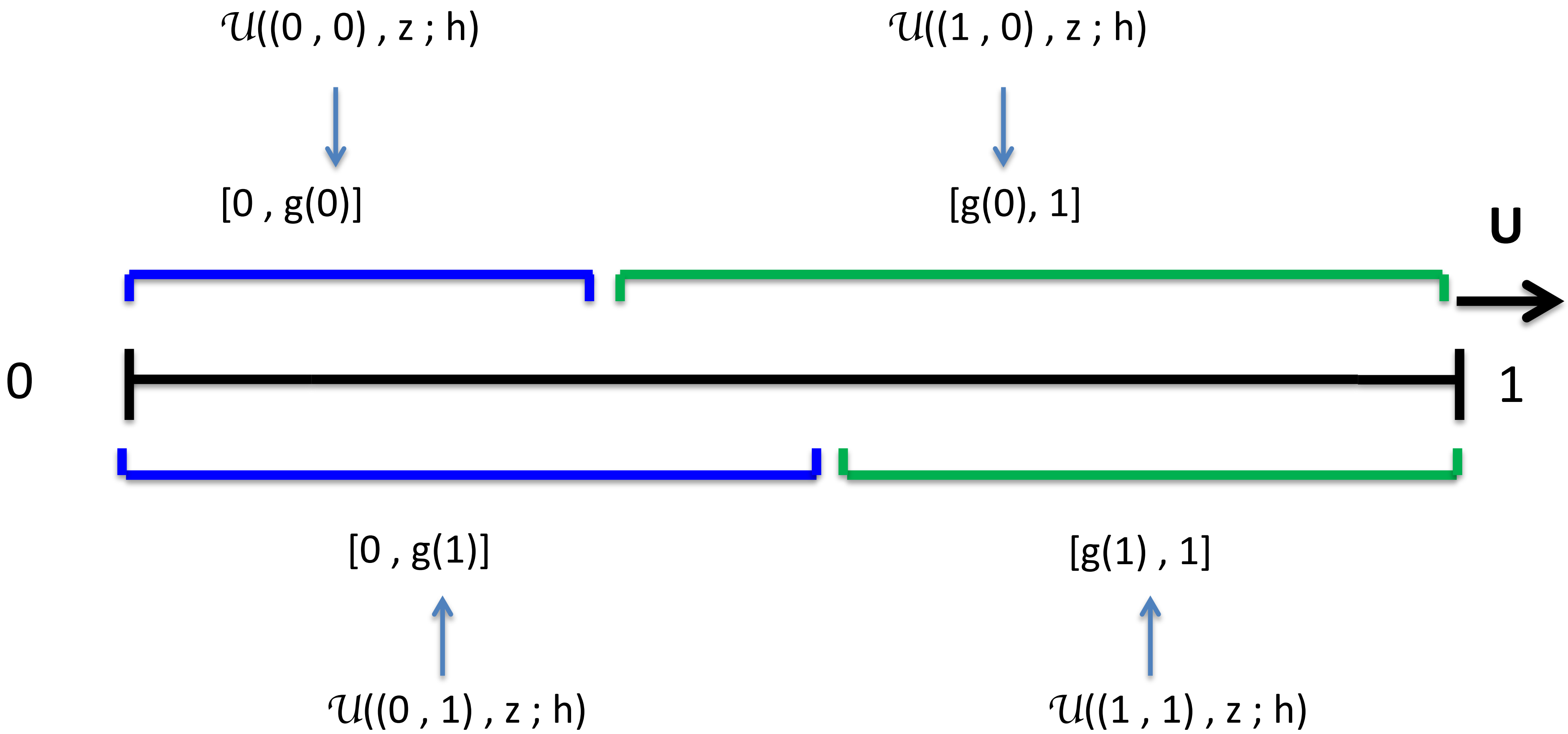
$\mathcal{A}(\mathcal{S}, z; h) = \{ y : \mathcal{U}(y, z; h) \subseteq \mathcal{S} \}$

$\mathcal{A}([g(0), 1], z; h) = \{ (1,0), (1,1) \}$

$\mathcal{A}([g(1), 1], z; h) = \{ (1,1) \}$

$g(0) < g(1)$

$Y_1 = 1[g(Y_2) \leq U]$



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

$\mathcal{A}([0, g(0)], z; h) = \{ (0,0) \}$

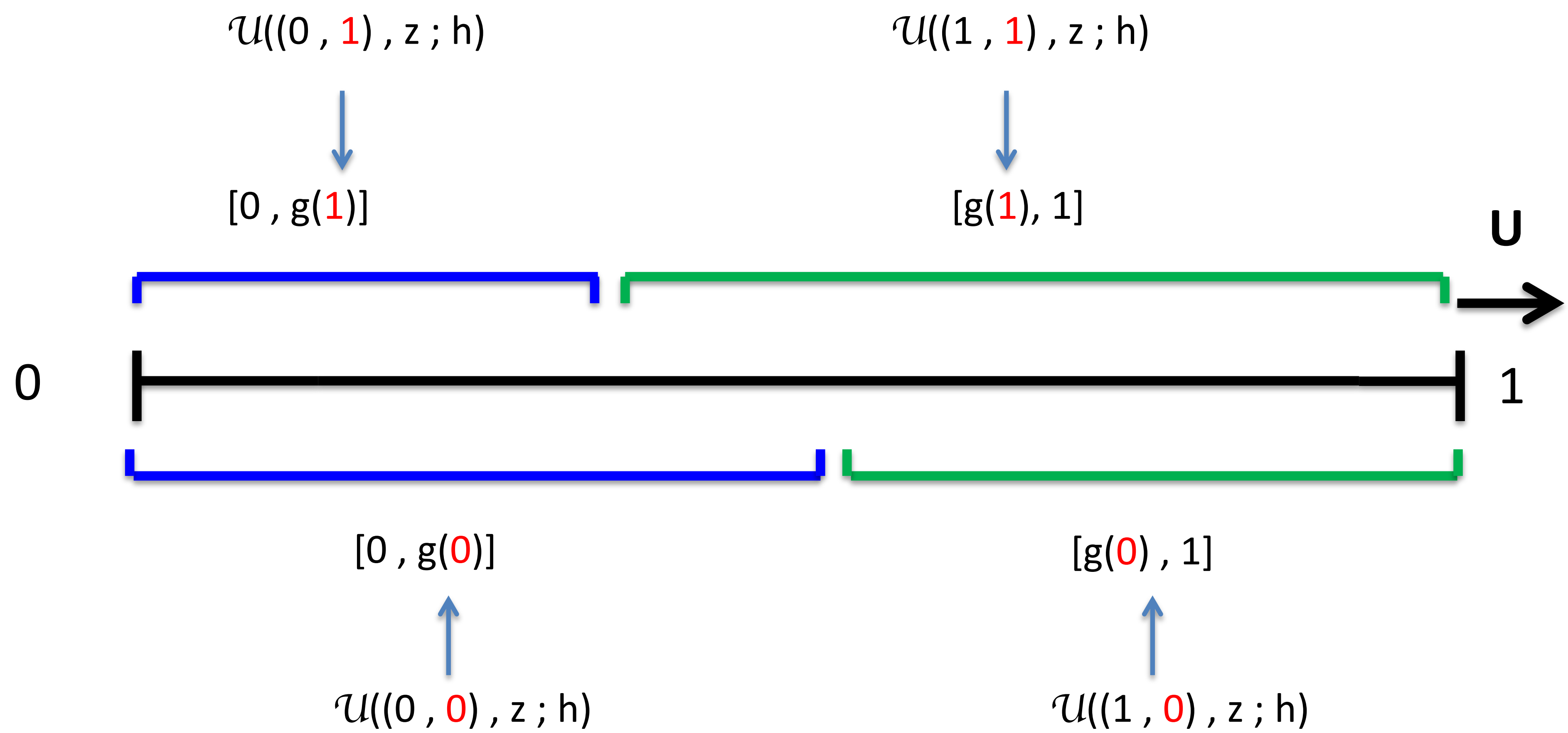
$\mathcal{A}([g(0), 1], z; h) = \{ (1,0), (1,1) \}$

$\mathcal{A}([0, g(1)], z; h) = \{ (0,0), (0,1) \}$

$\mathcal{A}([g(1), 1], z; h) = \{ (1,1) \}$

$g(0) > g(1)$

$Y_1 = 1[g(Y_2) \le U]$



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

$\mathcal{A}([0, g(1)], z; h) = \{ (0,1) \}$

$\mathcal{A}([g(1), 1], z; h) = \{ (1,0), (1,1) \}$

$\mathcal{A}([0, g(0)], z; h) = \{ (0,0), (0,1) \}$

$\mathcal{A}([g(0), 1], z; h) = \{ (1,0) \}$

Application: binary endogenous variables

- The set $\mathcal{A}(\mathcal{S}, z; h)$ for $\mathcal{S} = [0, g(1)]$ is

$$\mathcal{A}([0, g(1)], z; h) = \begin{cases} \{(0,1)\} & , \quad g(0) > g(1) \\ \{(0,0),(0,1)\} & , \quad g(0) \leq g(1) \end{cases}$$

- When $\mathcal{S} = [0, g(1)]$ the condition

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

is:

$$\left((g(0) > g(1)) \wedge \left(g(1) \geq \sup_{z \in \mathcal{R}_Z} \left(F_{Y|Z=z}((0,1) \right) \right) \right)$$

$$\vee \left((g(0) \leq g(1)) \wedge \left(g(1) \geq \sup_{z \in \mathcal{R}_Z} \left(F_{Y|Z=z}((0,0)) + F_{Y|Z=z}(\,(0,1)) \right) \right) \right)$$

Illustration: the effect of family size on female employment

- Angrist and Evans (AER, 1998), Angrist (JBES, 2001), Angrist and Pishke (Mostly Harmless, 2009).
- Sample: 1980 US Census Public Use Micro Samples
  - 254,654 married mothers aged 21-35 **with 2 or more children**, oldest < 18.
- Binary outcome:
  - $Y_1 = 1$ if worked for pay in 1979, $Y_1 = 0$ otherwise.
- Explanatory variables:
  - $Y_2 = 1$ if 3 or more children, $Y_2 = 0$ if 2 children.
  - $Z_1 = 1$ if more than 12 years of education, 0 otherwise.
- **Instrumental variables**:
  - $Z_2 = 1$ if first two children are same-sex, 0 otherwise.
  - $Z_3 = 1$ if at 2nd birth there are twins, 0 otherwise

An incomplete model

- I consider the incomplete model:

$$Y_1 = 1[\beta_0 + \alpha Y_2 + \beta_1 Z_1 < U_1]$$

$$U_1 \perp\!\!\!\perp Z = (Z_1, Z_2, Z_3) \qquad\qquad U_1 \sim N(0,1)$$

  When $\alpha > 0$, more children lowers the probability of working for pay.

- The values of $(\beta_0, \beta_1, \alpha)$ are **set** identified. We estimate the identified set.

A complete model for comparison

- The **complete** model has

$$Y_1 = 1[\beta_0 + \alpha Y_2 + \beta_1 Z_1 < U_1]$$

AND a **second equation**

$$Y_2 = 1[\gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_X < U_2]$$

where Z_X is one of the IVs, and restrictions:

$$U \equiv \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} \perp\!\!\!\perp Z = (Z_1, Z_2, Z_3) \qquad U \sim N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$$

- This is a **point identifying** model. It features in Heckman (Ecta, 1978).

The power of the instruments

- The same-sex instrument has low predictive power for advancing beyond 2 children ($Y_2 = 1$).

$$\mathbb{P}[Y_2 = 1 | \text{same-sex}] = 0.41 \qquad \mathbb{P}[Y_2 = 1 | \text{not same-sex}] = 0.35$$

The power of the instruments

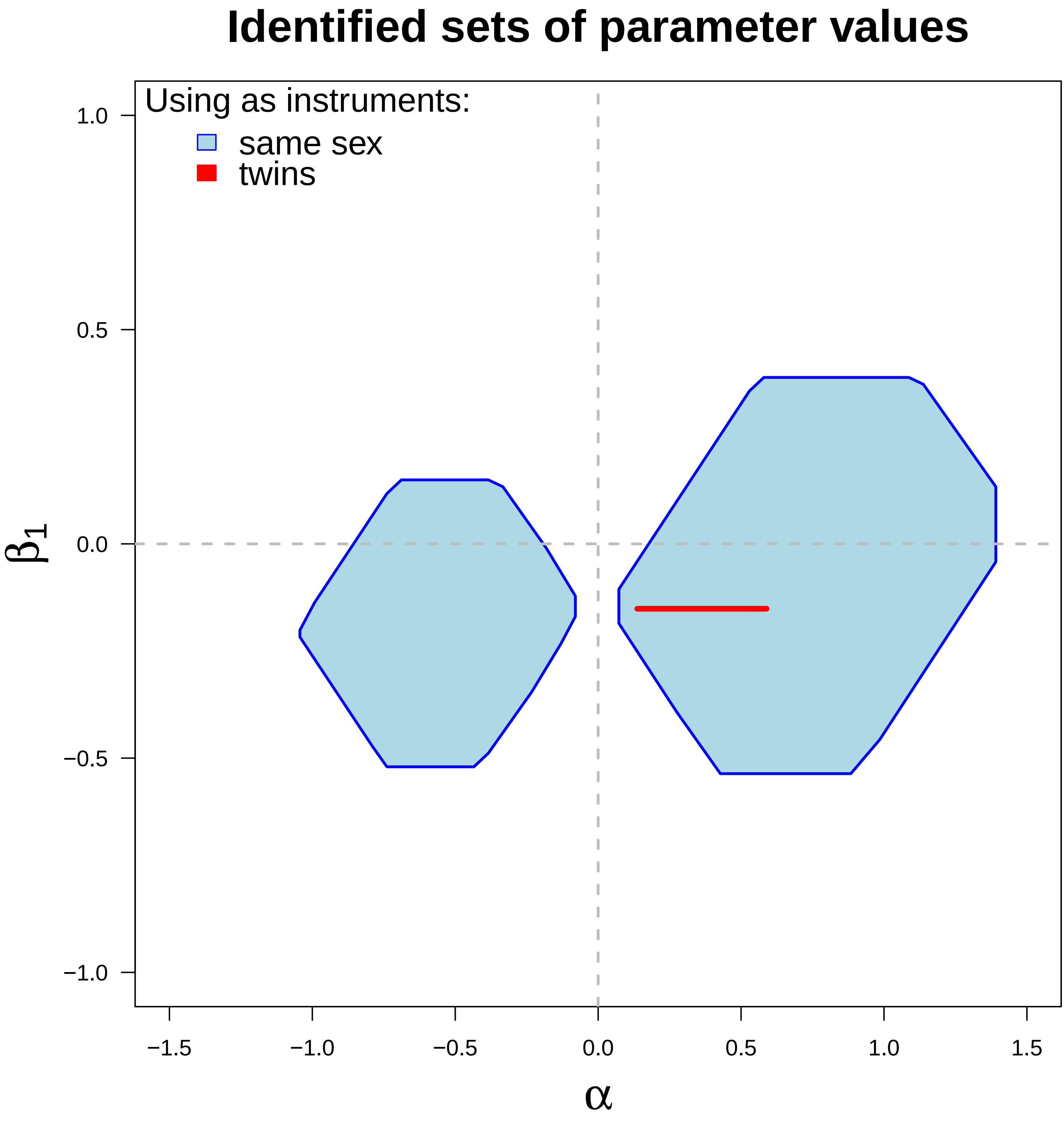
- The same-sex instrument has low predictive power for advancing beyond 2 children ($Y_2 = 1$).

$$\mathbb{P}[Y_2 = 1|\text{same-sex}] = 0.41 \qquad \mathbb{P}[Y_2 = 1|\text{not same-sex}] = 0.35$$

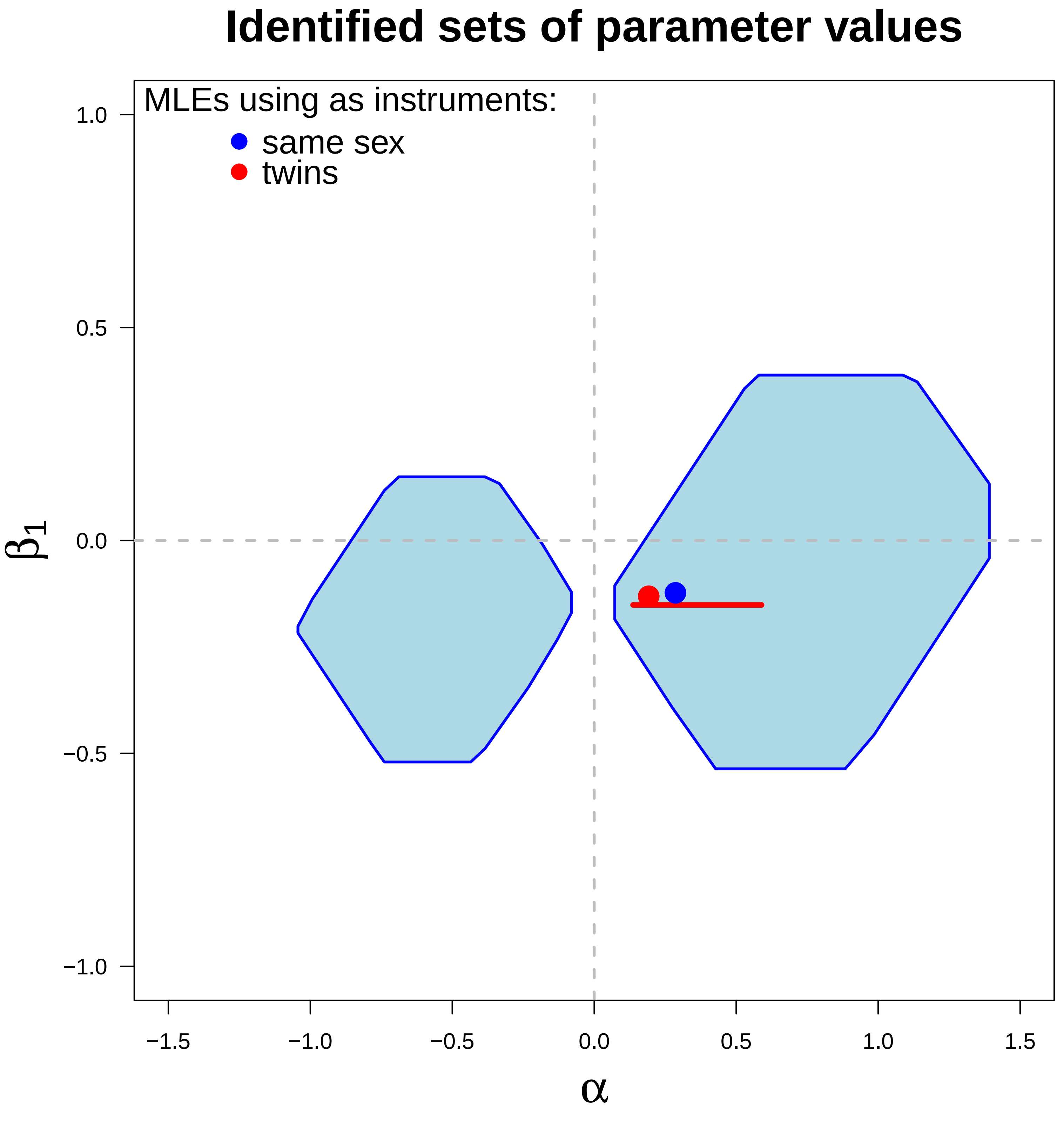
- The twins instrument is a **MUCH** better predictor

$$\mathbb{P}[Y_2 = 1|\text{twins}] = 1.00 \qquad \mathbb{P}[Y_2 = 1|\text{not twins}] = 0.38$$

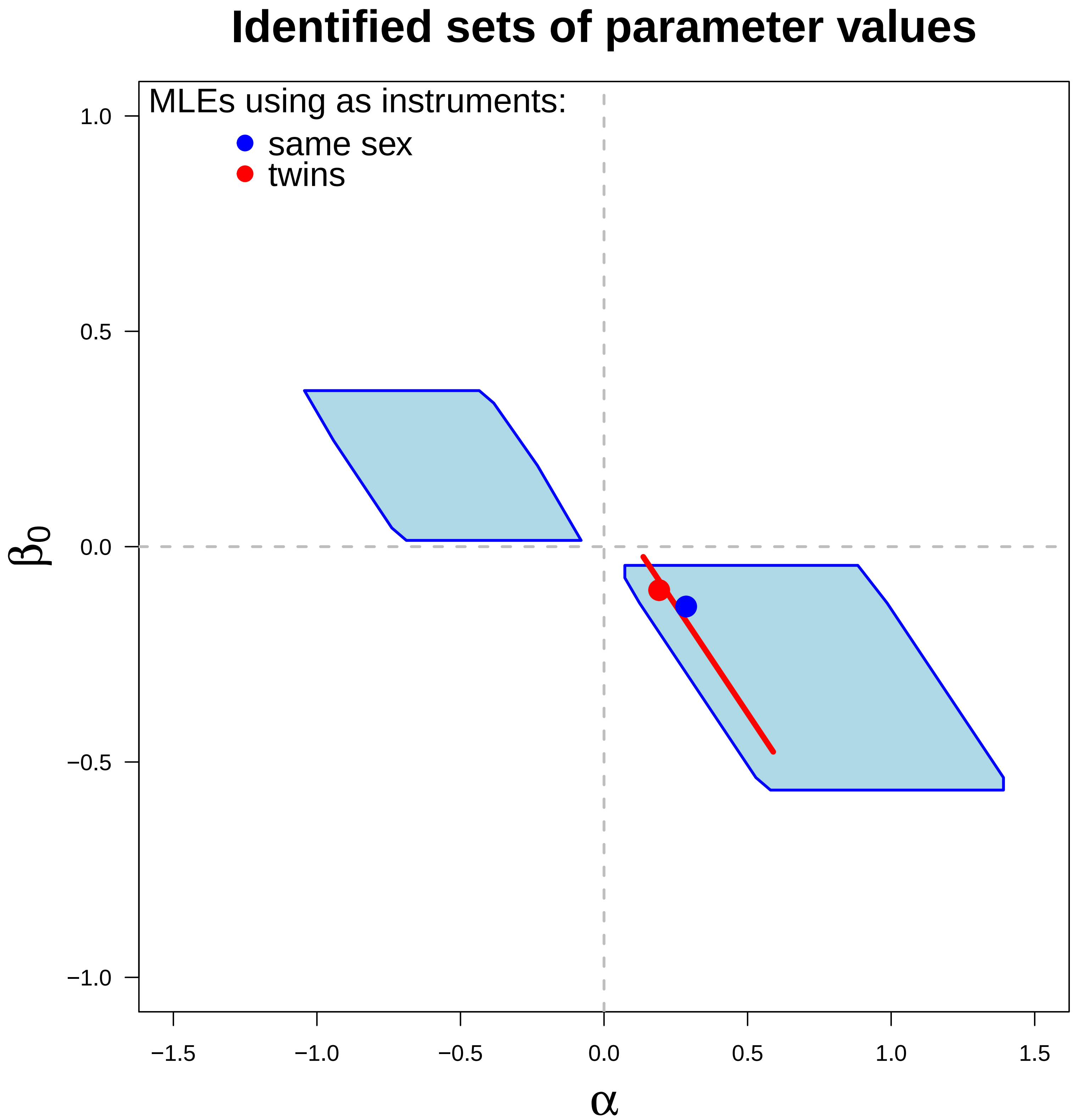
- if a second birth event is a multiple live birth there **MUST** be 3 or more children.



Identified sets of parameter values
Using as instruments:
same sex
twins
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
β_1
α



Identified sets of parameter values
MLEs using as instruments:
same sex
twins
1.0
0.5
0.0
−0.5
−1.0
β_1
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
α



Identified sets of parameter values
MLEs using as instruments:
same sex
twins
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
α
β0

Remarks

- **Incomplete** models are generically **partially identifying** when unobserved variables are **not single valued functions of observed variables**.
- Until our GIV work, applied work for these cases used point identifying **complete** models, control function methods, or other conditional independence restrictions.
  - but there are many ways to complete models and data cannot distinguish one from another.
  - the conditional independence restriction underlying the control function method is highly restrictive.
  - generally conditional independence restrictions can be difficult to justify - which variables to condition on?
  - control functions don't deliver when endogenous variables are discrete or affected by multidimensional heterogeneity.
- The sharp identified sets identified by incomplete models can be characterized and estimated.

Review

- Identification analysis underpins all econometric analysis.

- We have focused on structural econometric models in which knowledge of economic context and restrictions on economic behaviour motivate restrictions incorporated in models.

- Identification analysis tells us under what restrictions (i.e. models) data can be informative about a structural feature,

  - and what features of probability distributions are informative about structural features,

  - and hence how analog estimation can proceed.

  - and whether and how a model's restrictions can be tested.

Research challenges

- There remain many research challenges, for example:
  - understanding the identifying power of novel restrictions in new contexts e.g. in dynamic models,
  - developing estimation and inference procedures that perform well in practice,
  - understanding the consequences of misspecification, so far little studied in the context of partially identifying models,
  - bringing new understanding in practical applications until now studied using restrictive complete models.

Application: why alpha equal to zero is not in the identified set

- The incomplete model is as follows.

$$Y_1 = 1[\beta_0 + \alpha Y_2 + \beta_1 Z_1 < U_1]$$

$$U_1 \perp\!\!\!\perp Z = (Z_1, Z_2, Z_3) \qquad\qquad U_1 \sim N(0,1)$$

- If, for any z_1, $\mathbb{P}[Y_1|Z_1 = z_1, Z_2 = z_2]$ depends on z_2 then α cannot be zero.
  - The data suggests $\mathbb{P}[Y_1|Z_1 = z_1, Z_2 = z_2]$ does depend on z_2.