Lectures on Identification 1

Andrew Chesher

CeMMAP & UCL

April 14th 2008

Identification analysis

- An economic process generates data.
  - What can be known of the process from the data it yields?
  - Require restrictions that limit admissible process. What is the force of various restrictions?
  - What minimal restrictions are required for knowledge of particular features of a process?
  - How does the extent and nature of measurement affect the answers to these questions?
  - What are implications for design of surveys, experiments etc?
  - What is the nature of the knowledge that can be obtained - ambiguity? Point or set identification?

History: statistical laws, data, data, data

- Henry Ludwell Moore's book (1914) "**Economic Cycles their Law and Causes**".
  - Moore's "new type" of demand curve.
- Henry Schulz (1925) "The statistical law of demand".
- Holbrook Working "The statistical determination of demand curves" QJE 1925.
- Some references:
  - "Measurement, quantification and economic analysis: numeracy in economics" ed. Ingrid Rima, Routledge, 1995.
  - "The history of econometric ideas" Mary Morgan, Cambridge University Press, 1990.
  - "Foundations of Econometric Analysis" ed. David Hendry and Mary Morgan, Cambridge University Press, 1995.

History: early econometric thinking

- Moore's book reviewed by Robert Lehfeldt (1914), Philip Wright (1915). The "new demand curve" is a supply curve.
- Understanding that economics could/should illuminate statistical analysis.
  - Marcel Lenoir "Etudes sur la Formation et le Mouvement de Prix", 1913.
  - Robert Lehfeldt, Elasticity of demand for wheat, EJ, 1914.
  - Elmer Working "What do statistical 'Demand Curves' show?" QJE 1927.
  - Discussion of relative magnitudes of variation in demand and supply schedules. See Fisher (1965), "Near Identifiability and the Variance of Disturbance Terms".
- Landmark: Philip Wright "The Tarrif on Animal and Vegetable Oils' 1928.
  - Appendix B– developed and applied IV.
  - Sewall Wright's contribution? - see Stock and Trebbi (JEP, 2003).
- Landmark: Jan Tinbergen studies information content of reduced forms + economic theory - ILS estimator.
  - "Determination and Interpretation of Supply Curves: An Example, *Zeitschrift für Nationalökonomie* 1930. Potato flour.

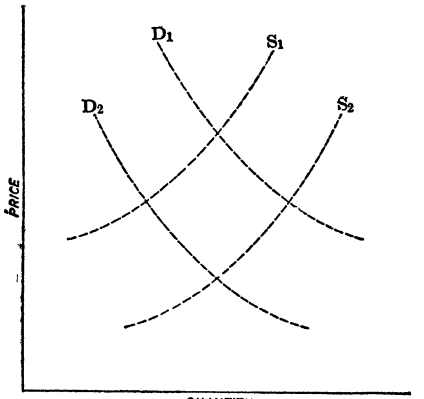


FIGURE II

Under such conditions there will result a series of prices which may be graphically represented by Figure III. It is from data such as those represented by the dots that we are to construct a demand curve, but evidently no satisfactory fit can be obtained. A line of one slope will give substantially as good a fit as will a line of any other slope.

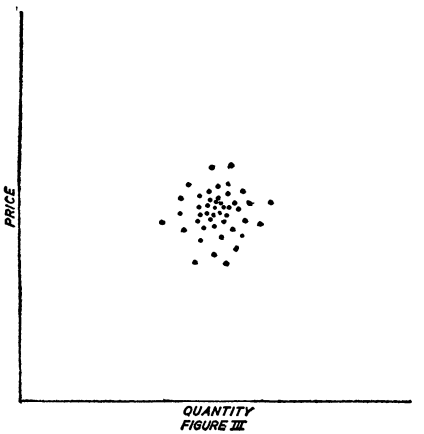


FIGURE III

But what happens if we alter our assumptions as to the relative shifting of the demand and supply curves? Suppose the supply curve shifts in some such manner as is indicated by Figure IV, that is, so that the shifting of the supply curve is greater than the shifting of the de-

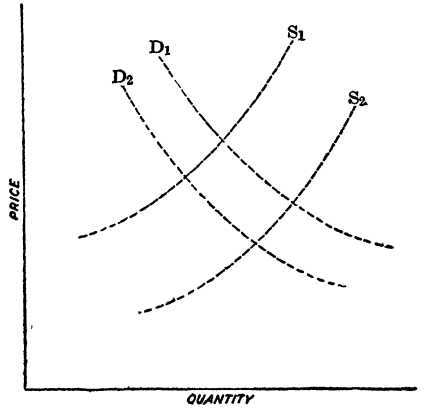


FIGURE IV

mand curve. We shall then obtain a very different set of observations — a set which may be represented by the dots of Figure V. To these points we may fit a curve

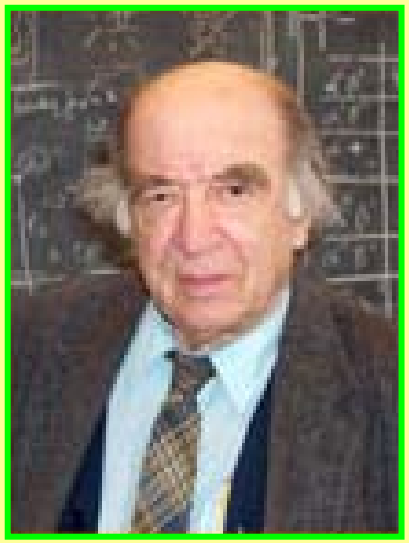
FIGURE V

Flowering of econometrics

- Frisch (1933) "Pitfalls in the Statistical Construction of Demand and Supply Curves", and "More Pitfalls ...' QJE 1934. **Errors in equations**.
  - Frisch co-founded the Econometric Society, 1930 from which Econometrica 1933.
  - Frisch and Waugh (1933) coin the term "structural" relationship.
- Cowles Commission founded 1932, Colorado Springs - to Chicago 1939, to Yale in 1955.
- Haavelmo (a student of Frisch) - "The **Probability Approach** in Econometrics" 1941 (published 1944). Errors in equations. "Cowles" models.
- The Cowles Commision research programme of 1943-47 - Cowles Monograph 10, 1950 (really 1946)
  - Remarkable paper by Hurwicz "Generalization of the Concept of Identification" (and other papers).
  - Koopmans, Rubin and Leipnik - order and rank conditions.
  - Koopmans "Identification problems in economic model construction" Econometrica, (1949).
  - Koopmans and Reiersøl *Annals of Mathematical Statistics*(1950) "The Identification of Structural Characteristics".

Jan Tinbergen: 1903 - 1994

Tjalling Koopmans: 1910 - 1985



Leonid Hurwicz: 1917 -

Modern era: towards nonlinear and nonparametric models

- Nonlinear restrictions and models
  - Franklin Fisher (1959, 1961, 1966), Leon Wegge (1965) and Rothenberg (1971), Roger Bowden (1973), B. Brown (1983)
- Nonparametric restrictions and models
  - Roehrig (1988), see also Benkard and Berry (2006) and Matzkin (2005).
  - Many new results.
- Nonparametric identification - is an important topic.
  - Suited to the qualitative restrictions of economics
  - Can see the force of particular parametric and semiparametric restrictions.
  - Can understand which restrictions of models are testable.
- Set identification - exciting area
  - Interesting to set identify a deep "parameter" rather than point identify a derived parameter - e.g. some average.
  - Could any result be got using particular parametrice restrictions?
  - Guides design of data collection and experiments.
  - Recognition of ambiguity.

These lectures

- Focus on nonparametric identification in various extensions of the "Cowles model".
  - models built around structural functions/equations delivering observables given observable and unobservabnle variables.
  - there is a role for economics in informing about restrictions on functions.
  - there are limits on the degree of heterogeneity
  - by contrast there are "treatment effects" models with typically a minor role for economics, various incomplete data problems.
  - enough common structure to allow a coherent (?) presentation
  - allows to address some important issues - discreteness, set identification

Topics

1. **Today.**
   1. Structures, structural functions, models, point and set identification.
   2. Linear and parametric models, order and rank conditions.
   3. Parametric and nonparametric IV models - discrete and continuous explanatory variables.
2. Wednesday April 16th. Triangular models and control function methods. models with excess heterogeneity and index restrictions.
3. Monday April 21st. Discrete endogenous variables, set identification and control function methods. Introduction to set identification in IV models for binary data.
4. Wednesday April 23rd. Seminar on "Endogeneity and Discrete Outcomes".

Hurwicz structures

- Outcome(s) Y are uniquely determined by structural equations

$$h(Y, Z, U) = 0$$

  given values of observed Z and unobserved U.
- There is a probability distribution F_{UZ} (or $F_{U|Z}$ and $Z \in \Omega_Z$).
- The pair $S \equiv \{h, F_{UZ}\}$ is a **structure**. Each S generates a distribution F_{YZ} (or $F_{Y|Z}$).
- **Observational equivalence**. S' and S'' such that $F_{YZ}^{S'} = F_{YZ}^{S''}$ are observationally equivalent.
- A **model** comprises restrictions on admissible structures.
- Let $\mathcal{M}_\Gamma$ be the set of admissible structures defined by a model, Γ. Consider $S^0 \in \mathcal{M}_\Gamma$.
  - The model Γ **identifies** S^0 if there is no $S^* \in \mathcal{M}_\Gamma$ such that $F_{YZ}^{S^*} = F_{YZ}^{S^0}$.
  - Γ is **uniformly** identifying if it identifies all $S \in \mathcal{M}_\Gamma$.

Identification of structural features

- Consider a **feature** of a structure, $\theta(S)$ and a model Γ.
- Γ **point** identifies $\theta(S^0)$ if $\theta(S)$ is constant across all structures admitted by Γ and observationally equivalent to S^0.
  - Uniform point identification if Γ point identifies $\theta(S)$ for all $S \in \mathcal{M}_\Gamma$.
- Γ **set** identifies $\theta(S^0)$ to within Θ_0 if for all admissible S observationally equivalent to S^0, $\theta(S) \in \Theta_0$.
- If there exists a functional $\mathcal{G}(F_{YZ})$ such that for all $S^* \in M_\Gamma$,

$$\theta(S^*) = a \Longrightarrow \mathcal{G}(F^*_{YZ}) = a$$

then Γ uniformly point identifies θ.

- *Proof.* If $\mathcal{G}$ exists and S' and S'' have $\theta = a'$ and a'', then

$$\mathcal{G}(F^{'}_{YZ}) = a' \qquad \mathcal{G}(F''_{YZ}) = a''$$

and if S' and S'' are observationally equivalent then

$$F^{'}_{YZ} = F''_{YZ} \Longrightarrow a' = a''.$$

Identification of structural features

- Consider a **feature** of a structure, $\theta(S)$ and a model Γ.
- Γ **point** identifies $\theta(S^0)$ if $\theta(S)$ is constant across all structures admitted by Γ and observationally equivalent to S^0.
  - Uniform point identification if Γ point identifies $\theta(S)$ for all $S \in \mathcal{M}_\Gamma$.
- Γ **set** identifies $\theta(S^0)$ to within Θ_0 if for all admissible S observationally equivalent to S^0, $\theta(S) \in \Theta_0$.
- If there exists a functional $\mathcal{G}(F_{YZ})$ such that for all $S^* \in M_\Gamma$,

$$\theta(S^*) = a \Longrightarrow \mathcal{G}(F^*_{YZ}) = a$$

  then Γ uniformly point identifies θ.
  - Overidentification.
  - Analogue estimation: $\mathcal{G}(\hat{F}_{YZ})$.

Example: linear models and IV

- The IV model

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

- The model identifies β if there are values $\{z_1, z_2\}$ in the support of Z and

$$E[Y_2|Z = z_1] \neq E[Y_2|Z = z_2]$$

  which exist.

$$\beta = \frac{E[Y_1|Z = z_1] - E[Y_1|Z = z_2]}{E[Y_2|Z = z_1] - E[Y_2|Z = z_2]}$$

- On the right is a functional of F_{YZ} which delivers the value of β.

Types of restrictions

- Notation: now use Y, X for outcomes.
- 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:

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

Incomplete in the sense that there is no restriction on the genesis of X.

- Model A: "control function" methods. Model B: "instrumental variable" methods.
- Consider various restrictions on structural functions - e.g. **additivity**, monotonicity.
- The **non-additive** cases permit discrete outcomes, censoring; many unobservables.
- Consider various restrictions on distributions of unobserved variables.

Types of restrictions

- 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

- Start with model, B and discrete or continuous X, continuous Y.

The IV model

- A weakly restrictive version of model "B" has

$$\text{B: } Y = h(X, U) \qquad U \perp\!\!\!\perp Z$$

- This instrumental variables (IV) model is silent about the genesis of X.
- Written in terms of density functions:

$$p(y, x, u, z) = p(y|x, u) \times p(x|u, z) \times p(u) \times p(z)$$

  and $p(x|u, z)$ is unspecified.
- A Bayesian would have some trouble with this.

Solutions to simultaneous equations

- Consider the IV model

$$Y = \alpha + \beta X + \gamma X^2 + U \qquad E[U|Z = z] = c$$

- With 3 values of Z

$$\begin{aligned} E[Y|z_1] &= \alpha + \beta E[X|z_1] + \gamma E[X^2|z_1] + c \\ E[Y|z_2] &= \alpha + \beta E[X|z_2] + \gamma E[X^2|z_2] + c \\ E[Y|z_3] &= \alpha + \beta E[X|z_3] + \gamma E[X^2|z_3] + c \end{aligned}$$

  and $\alpha + c$, β, γ are solutions to 3 simultaneous equations.
- Require rich support for Z as polynomial (spline etc) becomes more flexible.

The nonparametric IV model: discrete X

- Discrete, $X \in \{x_1, x_2, \ldots, x_M\}$

$$Y = h(X) + U \qquad E[U|Z = z] = 0$$

$$\theta_m \equiv h(x_m) \qquad D_m(X) \equiv 1[X = x_m]$$

- Then - Das (2005)

$$Y = \sum_{m=1}^{M} \theta_m D_m(X) + U \qquad E[Y|z] = \sum_{m=1}^{M} \theta_m P[X = x_m|z]$$

- If $Z \in \{z_1, \ldots, z_K\}$ identifiability of h depends on the rank of the $K \times M$ matrix of probabilities;

$$\begin{bmatrix} P[X = x_1|z_1] & P[X = x_2|z_1] & \cdots & P[X = x_M|z_1] \\ \vdots & \vdots & & \vdots \\ P[X = x_1|z_K] & P[X = x_2|z_K] & \cdots & P[X = x_M|z_K] \end{bmatrix}$$

  - ill conditioned matrix when M large; smoothness; dimension of X and Z;
  - all or nothing identification unless Z can perfectly predict X; or set identification?

The nonparametric IV model: continuous X

- The IV model with continuous X

$$Y = h(X) + U \qquad E[U|Z = z] = 0$$

  See Newey & Powell (2003).

- Identification of h requires restrictions on h and F_{UXZ} guaranteeing unique solution for h to the integral equation:

$$a(z) \equiv E[Y|Z = z] = \int h(x) f_{X|Z}(x|z) dx \equiv b(z)$$

  - requires restrictions on the "incomplete" element of the model
  - requires rich support for Z,
  - all or nothing identification,
  - possibly set identification when point identification conditions fail
  - smoothness or parametric restrictions helpful.

Nonadditive U

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

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

- Can normalise $U \sim Unif(0, 1)$ under montonicity - free choice of units of measurement.
- Matzkin (2003). If $Y = h(X, U)$ and $U \perp\!\!\!\perp X$ and $U \sim Unif(0, 1)$ then

$$h(x, u) = Q_{Y|X}(u|x)$$

- Also true (for suitable definition of h) for **discrete** Y if we define quantiles for discrete variates as follows.

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

- Nonmonotonicity? Hoderlein and Mammen (2008).

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 $Unif(0, 1)$.

- Define

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

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

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

- Strict monotonicity of h with respect to U implies

$$\begin{aligned} a(x, z, \tau) &= P[h(X, U) \leq h(X, \tau) | X = x, Z = z] \\ &= P[Y \leq h(X, \tau) | X = x, Z = z] \end{aligned}$$

- So for all τ, z

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

The IV model nonadditive U

- h monotonic (normalised increasing) in scalar U.

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

For all τ, z

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

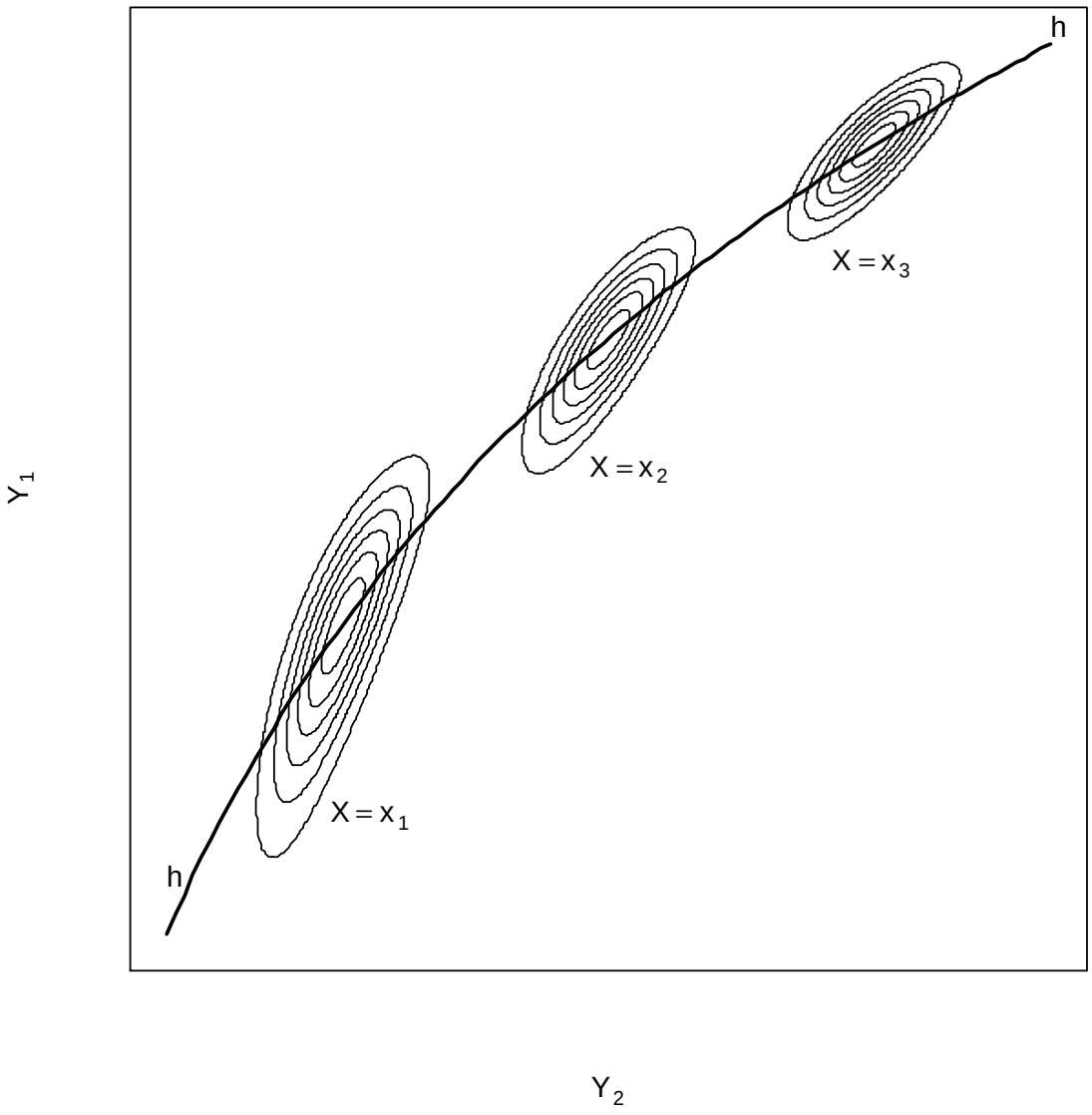
- identification of h if restrictions on h and F_{UXZ} guarantee a unique solution.
- rich support for Z; all or nothing identification.
- set identification is a possibility.
- local independence: could just have $Q_{U|Z}(\tau|z) = \tau$ then identify $h(x, \tau)$
- discrete $X \in \{x_1, \ldots, x_M\}$,

$$G(a, b|z) \equiv \Pr[Y \leq a \cap X = b|Z = z]$$

The values $\theta_m^\tau \equiv h(x_m, \tau)$ are solutions to nonlinear simultaneous equations generated as z varies.

$$\sum_{m=1}^{M} G(\theta_m^\tau, x_m|z) = \tau$$

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 $q_\tau = 0.5$.



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

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

Now

- 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"

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

- 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**

Next

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

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

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