Identification in models with discrete variables

Joint PhD Workshop in Economics

Lukáš Lafférs

NHH Norwegian School of Economics

January 9, 2012

Introduction to Partial Identification

Econometricians typically work with point-identified models, e.g. $Y_i = X_i' \beta + U_i \quad E(U_i|X_i) = 0$, elements of X_i not perfectly correlated
there exists **only one** β that satisfies these assumptions and is compatible with the distribution of (Y_i, X_i) which is revealed by the data.

In certain situations our assumptions are not strong enough to determine a unique value of a parameter but there is a set of **observationally equivalent models**.
Meaning that no amount of data would ever help me to distinguish between these models.

surveys Manski(1995,2003), Tamer(2010)

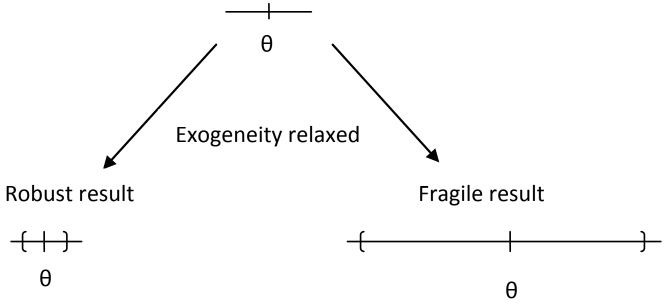
Throughout this presentation I will discuss **Identification** not Inference.
It is assumed that we know the true data generating process of observable variables.

An Example

Example 1 - Manski (1990) - Missing data
We are interested in $\theta = E(Y)$, it is only observed when $D = 1$.
$\theta = E(Y) = E(Y|D = 1)P(D = 1) + E(Y|D = 0)P(D = 0)$
$\theta = p.\mu_1 + (1 - p).\mu_0$
Additional assumptions needed, if e.g. $Y_i \in \{0, 1\}$ then
$\theta \in [\theta_{low}, \theta_{high}] = [p.\mu_1, p.\mu_1 + (1 - p)]$.

Motivation - Exogeneity assumption relaxed

- To see the strength of the assumption that cannot be tested
- Sensitivity analysis



Galichon and Henry Framework (simplified)

Galichon and Henry (2006, 2009, 2010, 2011)

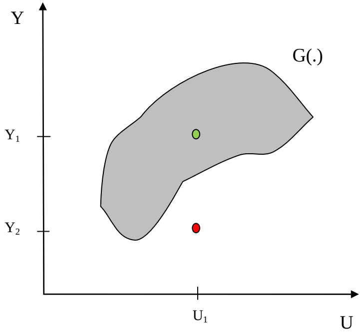
Two types of variables:
Y - **Observable** variables ($Y \in \mathcal{Y}$ with density p)
U - **Unobservable** variables ($U \in \mathcal{U}$ with density ν_θ)

Economic restrictions take the form of
G_θ - many-to-many mapping ($G_\theta : \mathcal{U} \mapsto \mathcal{Y}$)

θ - parameter of interest

Galichon and Henry Framework (simplified) (2)

Not all pairs (Y, U) are compatible with economic restrictions



(Y_1, U_1) is compatible $(Y_1 \in G_\theta(U_1))$
(Y_2, U_1) is not $(Y_2 \notin G_\theta(U_1))$

Galichon and Henry Framework (simplified) (3)

- Parameter θ is included in the **Identified set** if and only if there exists a joint distribution π of (Y, U) on $\mathcal{Y} \times \mathcal{U}$ with marginals p and ν_θ such that $\pi(\{Y \in G_\theta(U)\}) = 1$
- It means that the model is compatible with data at hand and satisfies economic restrictions almost surely

My Extension of GH Framework

Economics enters the model via G_θ only.

I extend the GH framework to entertain additional distributional restrictions.

My Extension of GH Framework

Economics enters the model via G_θ only.

I extend the GH framework to entertain additional distributional restrictions.

$$E(\phi(Y, U)) = 0$$
$$|cov(Y, U)| \leq 0.1$$
U is independent of a component of Y

My Extension of GH Framework (2)

What can be done using this extension ?

I replicate a few results from partial identification literature that were obtained by distinct approaches.

In addition: I show how to see the strength of the assumption of a strict exogeneity of instruments in a nonlinear model with discrete variables.

Single Equation Endogenous Binary Response Model

Model studied in Chesher (2010, ECTA).

- (Y, X, Z) - **Observable** variables (p_{ijk})
- U - **Unobservable** variables $(Unif(0,1))$

The economic restrictions are

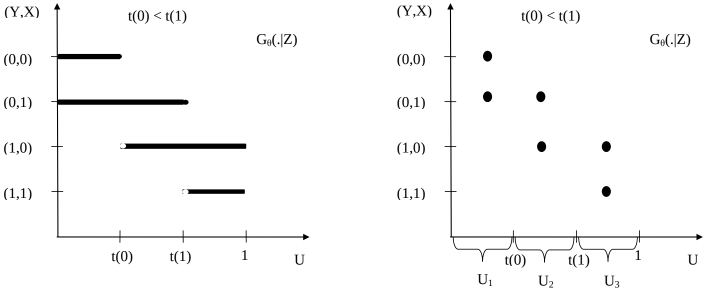
$$(Y, X, Z) \in G_\theta(U) \qquad \Leftrightarrow \qquad Y = \begin{cases} 0, & \text{if } U \leq \Phi(-\theta_0 - \theta_1 X), \\ 1, & \text{if } U > \Phi(-\theta_0 - \theta_1 X). \end{cases}$$

Further assumption
$U \perp Z$

What can we tell about (θ_0, θ_1) ?

Formulation in the extended GH framework

Support restrictions and Discretization



$$t(X) = \Phi(-\theta_0 - \theta_1 X)$$

Formulation in extended GH framework (2)

$$\pi_{ijkl} = Pr(Y = y_i, X = x_j, Z = z_k, U = u_l)$$

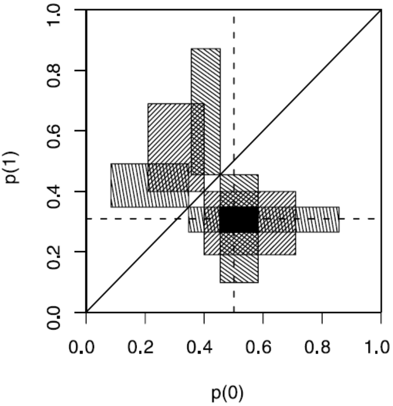
Penalty is given by

$$c_{ijkl} = \begin{cases} 0, (y_i, x_j, z_k) \in G_\theta(u_l), \\ 1, \text{ otherwise.} \end{cases}$$

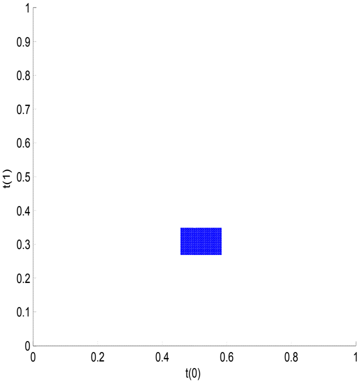
Problem is formulated as

$$\begin{array}{cc} \min_{(\pi)} \sum_{i,j,k,l} \pi_{ijkl} c_{ijkl} & \\ \text{s.t.} & \\ \sum_l \pi_{ijkl} = p_{ijk}, & \forall i,j,k \\ \sum_{i,j,k} \pi_{ijkl} = \nu_l, & \forall l \\ \sum_{i,j} \pi_{ijkl} = \sum_{i,j} p_{ijk} \nu_l, & \forall k,l \\ \pi_{ijkl} \geq 0, & \forall i,j,k,l. \end{array}$$

Comparison of Results (Identified set)



(Chesher 2010)



(my approach)

Exogeneity assumption relaxed (2)

Recall exogenous case

$$
\begin{array}{cc}
\min_{(\pi)} \sum_{i,j,k,l} \pi_{ijkl} c_{ijkl} & \\
\text{s.t.} & \\
\sum_{l} \pi_{ijkl} = p_{ijk}, & \forall i,j,k \\
\sum_{i,j,k} \pi_{ijkl} = \nu_l, & \forall l \\
\sum_{i,j} \pi_{ijkl} = \sum_{i,j} p_{ijk} \nu_l, & \forall k,l \\
\pi_{ijkl} \geq 0, & \forall i,j,k,l.
\end{array}
\quad (1)
$$

Exogeneity assumption relaxed (3)

Recall exogenous case

$$\begin{aligned} \min_{(\pi)} \textstyle\sum_{i,j,k,l} \pi_{ijkl} c_{ijkl} \\ \text{s.t.} \\ \textstyle\sum_l \pi_{ijkl} = p_{ijk}, \quad & \forall i,j,k \\ \textstyle\sum_{i,j,k} \pi_{ijkl} = \nu_l, \quad & \forall l \\ \textstyle\sum_{i,j} \pi_{ijkl} = \sum_{i,j} p_{ijk} \nu_l, \quad & \forall k,l \\ \pi_{ijkl} \geq 0, \quad & \forall i,j,k,l. \end{aligned} \tag{2}$$

$$Pr(Z,U) = Pr(Z)Pr(U)$$

Exogeneity assumption relaxed (4)

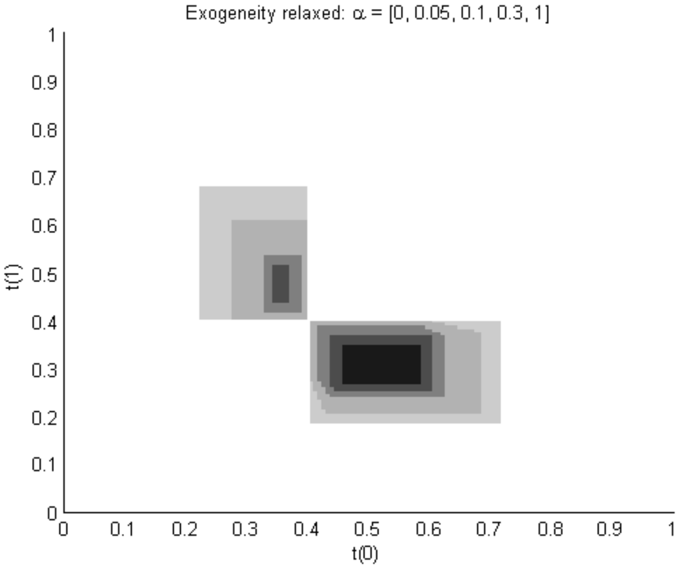
Now the Z and U are only "close" to being independent.

$$\begin{aligned}
&\min_{(\pi)} \textstyle\sum_{i,j,k,l} \pi_{ijkl} c_{ijkl} && (3)\\
&\text{s.t.} \\
&\textstyle\sum_l \pi_{ijkl} = p_{ijk}, && \forall i,j,k \\
&\textstyle\sum_{i,j,k} \pi_{ijkl} = \nu_l, && \forall l \\
&\textstyle\sum_{i,j} \pi_{ijkl} - \sum_{i,j} p_{ijk}\nu_l \leq \delta \sum_{i,j} p_{ijk}\nu_l, && \forall k,l \\
&-\textstyle\sum_{i,j} \pi_{ijkl} + \sum_{i,j} p_{ijk}\nu_l \leq \delta \sum_{i,j} p_{ijk}\nu_l, && \forall k,l \\
&\pi_{ijkl} \geq 0, && \forall i,j,k,l.
\end{aligned}$$

$$|Pr(Z,U) - Pr(Z)Pr(U)| \leq \delta Pr(Z)Pr(U)$$

Still a linear program - computationally feasible.

Exogeneity assumption relaxed (5)



Conclusions

- Extension of an existing framework for incompletely specified models with discrete variables
- Can replicate some existing results from partial identification literature in a straightforward manner
- It is possible to see the identification "strength" of the exogeneity of instruments in non-linear models with discrete variables

Thank you for your attention!