Topics in Simultaneous Equations Systems and Identification

Continuing with Greene pp 378-395. The structural form of a simultaneous equations model is

$$\begin{array}{ccccccccccc}
\gamma_{11}y_{t1} & + & \gamma_{21}y_{t2} & + & \cdots & \gamma_{M1}y_{tM} & + & \beta_{11}x_{t1} & + & \cdots & \beta_{K1}x_{tK} & = & \varepsilon_{t1} \\
\gamma_{12}y_{t1} & + & \gamma_{22}y_{t2} & + & \cdots & \gamma_{M2}y_{tM} & + & \beta_{12}x_{t1} & + & \cdots & \beta_{K2}x_{tK} & = & \varepsilon_{t2} \\
 & & & & & \vdots & & & & & & & \\
\gamma_{1M}y_{t1} & + & \gamma_{2M}y_{t2} & + & \cdots & \gamma_{MM}y_{tM} & + & \beta_{1M}x_{t1} & + & \cdots & \beta_{KM}x_{tK} & = & \varepsilon_{tM}
\end{array}$$

so there are M equations for M endogenous variables and K exogenous variables. In matrix notation this is

$$[y_1\ y_2 ... y_M]_t \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1M} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2M} \\ & & \vdots & \\ \gamma_{M1} & \gamma_{M2} & \cdots & \gamma_{MM} \end{bmatrix} + [x_1\ x_2 ... x_K]_t \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1M} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2M} \\ & & \vdots & \\ \beta_{K1} & \beta_{K2} & \cdots & \beta_{KM} \end{bmatrix}$$

$$= [\varepsilon_1\ \varepsilon_2 ... \varepsilon_M]_t$$

or

$$y_t^{\prime}\Gamma + x_t^{\prime}B = \varepsilon_t^{\prime}$$

Identification in the Linear Simultaneous Equations Model

Deepest possible view of identification (at least for me). Suppose we have a distribution of the 'data' $f(y, x)$. (The division into y and x is for expository purposes.) Then what models are compatible or not compatible with $f()$? On this view, identification is about the correspondence between models and the (true) joint distribution of the data.

When we restrict attention to the linear simultaneous equations model $y_t^{'}\Gamma + x_t^{'}B = \varepsilon_t^{'}$, we are looking at a very special situation.

Identification in the Linear Simultaneous Equations Model (2)

The *structural model* is written $y_t^{'}\Gamma + x_t^{'}B = \varepsilon_t^{'}$. 'Corresponding' to this is the *reduced form* obtained by postmultiplying by Γ^{-1} and rearranging

$$\begin{aligned} y_t^{'}\Gamma + x_t^{'}B &= \varepsilon_t^{'} \\ y_t^{'}\Gamma\Gamma^{-1} + x_t^{'}B\Gamma^{-1} &= \varepsilon_t^{'}\Gamma^{-1} \\ y_t &= -x_t^{'}B\Gamma^{-1} + \varepsilon_t^{'}\Gamma^{-1} \\ y_t &= x_t^{'}\Pi + \mathsf{v}_t^{'}, \end{aligned}$$

where $-B\Gamma^{-1} = \Pi$ and $\varepsilon_t^{'}\Gamma^{-1} =\mathsf{v}_t$. To do this the M x M matrix Γ must be invertible.

Now if we assumed that $E(\varepsilon_t|x_t) = 0$, $E(\varepsilon_t\varepsilon_t^{'}|x_t) = \mathbf{\Sigma}$, and $E(\varepsilon_t\varepsilon_s^{'}|x_t, x_s) = 0$ the reduced form of the model can be estimated by equation by equation OLS, as $E(\mathsf{v}_t|x_t) = (\Gamma^{-1})'0 = 0$. Define $E(\mathsf{v}_t\mathsf{v}_t') = E[(\Gamma^{-1})'\varepsilon_t\varepsilon_t'\Gamma^{-1}] = (\Gamma^{-1})\mathbf{\Sigma}\Gamma^{-1} = \Omega$, so $\mathbf{\Sigma} = \Gamma'\Omega\Gamma$.

Reduced and Structural Forms: A Roadmap

So $y_t'\Gamma + x_t'B = \varepsilon_t'$ is the structural form, $y_t = x_t'\Pi + \mathrm{v}_t'$ is the reduced form, and

1. $-B\Gamma^{-1} = \Pi$

2. $\boldsymbol{\Sigma} = \Gamma'\Omega\Gamma$

Before going on, let's do some econometric accountancy, and take an inventory of parameters. Γ is M x M nonsingular; B is K x M; $\boldsymbol{\Sigma}$ is M x M positive definite symmetric. So we are hoping to find $M^2 + K * M + \frac{1}{2}M(M+1)$ parameters.

We 'know' Π and Ω; these are K x M and M x M respectively; Ω is positive semidefinite symmetric. So we 'know' $K * M + \frac{1}{2}M(M+1)$ parameters. We want to 'know' M^2 more parameters than we have; these effectively correspond to Γ.

Deriving the Structural Form from the Reduced Form

So we need a way to provide the M^2 pieces of information that we need. M of these are easy to come by: each structural equation can be normalized to put a -1 as the coefficient of one of the endogenous variables that appears in the equation; this corresponds to the 'natural' dependent variable. (Discussion: Wooldridge's notion of 'autonomy', pp. 209–211.)

This leaves us still with $M(M-1)$ more structural parameters than reduced form parameters. This is typically remedied by

1. Putting 0's in elements of Γ; this corresponds to excluding an endogenous variable from an equation.

2. Putting 0's in elements of B; this corresponds to excluding exogenous variables from an equation.

3. Putting linear restrictions (such as two coefficients must sum to unity) of elements of either B or Γ or both.

4. Putting restriction on Σ, the covariance of the structural disturbances.

From a certain perspective, (1) and (2) are just special cases of (3), but we want to emphasize the role of variable exclusions.

An important thing to realize is that sometimes we can put restrictions on e.g. Γ and B that are sufficient to 'identify' parts of Γ and B but not all their elements, where one part of special interest is a row of each of them, since this corresponds to a single structural equation.

Because there are lots of ways to restrict Γ, B and Σ, there are lots of ways that parts of total systems can be identified, and this generates lots of special cases.

Identification: Why Bother? The Big Picture

We want to return to: "identification is about the correspondence between models and the (true) joint distribution of the data."

A structural model is a 'structural interpretation' or 'structural representation' of the data. All we observe is the joint distribution of the data (actually this is false for two reasons: (1) we don't observe even $f(y|x)$, we just hope to estimate it; and (2) we observe more than the numbers, but never mind). If two structural representations of the data are possible, there is no evidence in favor of one over the other. This is not so bad if the number of possibilities is two and they are similar, but more typically they are infinite in number and include very different possibilities.

The Special Aspects of the Linear Model in the Big Picture

Let $z = (y, x)$. If we know $f(z)$, we can derive $f(y|x)$. In linear simultaneous equations models with disturbances of the type we have assumed, the models that we estimate are compatible with any marginal distribution for x, i.e. $f(x)$. These models, when ('just') identified, tell us that a particular $f(y|x)$ corresponds exactly to one structural model. The reduced form specifies $f(y|x)$ closely enough (it gives us the conditional means and variances of y) that it specifies the structural representations that are consistent with it.

In fact this correspondence between reduced form and structural models is reflected in the terms 'underidentified', 'just identified' and 'over identified'. Underidentified means more than one structural representation is consistent with the reduced form representation; just identified means there is a one-to-one correspondence; over-identified means that the structural model implies restrictions (which may be hopelessly complicated) on the reduced form parameters Π and Ω.

Simple Identification for the Practical Person

Identification conditions in the linear simultaneous equations model are stated in terms of rank and order conditions. Rank conditions are sufficient conditions for the reduced form to have a unique structural representation. These are a little abstract and sometimes not obviously verifiable. Order conditions are necessary conditions: the structural equation can not be identified if the order condition is violated. The order condition basically specifies that for each endogenous variable with a coefficient to be estimated, one exogenous variable must be *excluded* from the equation. (Discussion; overidentification can be tested.)

Maximum Likelihood Estimation

Following Greene pp. 468–482. The principle of maximum likelihood is a powerful and flexible tool for estimation and inference. We consider problems with independent sampling so that

$$\begin{aligned} f(y_1, y_2, \ldots y_n|\theta) &= \prod_{i=1}^{n} f(y_i|\theta) = L(\theta|y), \quad and \\ ln\ L(\theta|y) &= \sum_{i=1}^{n} ln\ f(y_i|\theta) \end{aligned}$$

The second form, the log–likelihood, is easier to work with. The basic idea is that we choose to estimate the unknown parameter θ by that value which gives the highest value of the density or log density for the sample. (Since the log is a monotonic function of its argument, max'ing the log maxes the original likelihood.)

Properties of the Maximum Likelihood Estimator

Here's the deal: when we estimate according to maximum likelihood, and the model is correctly specified, the resulting estimate is

1. consistent,

2. asymptotically normal,

3. efficient (any other consistent estimator has the same or greater variance), and

4. 'invariant': $c(\theta_0)$ is consistently and efficiently estimated by $c(\hat{\theta})$.

Regularity Conditions for ML–An Informal Statement

Following Definitition 17.3 on p.474 Greene

1. The first three derivatives of $ln\ f(y_i|\theta)$ with respect to θ are continuous and finite for almost all y_i and all θ. (This condition ensures the existence of Taylor series approximations and the finite variance of second derivatives of $ln\ L$.)

2. The conditions necessary to obtain the expectations of the first and second derivatives of $ln\ f(y_i|\theta)$ are met.

3. For all θ, the third derivative of $ln\ f(y_i|\theta)$ in θ is bounded by a function with a finite expectation. (This allows truncation of Taylor series.)

Moments of the Derivatives of the Log–Likelihood

Following Greene's Theorem 17.2, p. 474.

1. $\ln f(y_i|\theta)$, $g_i = \partial \ln f(y_i|\theta)/\partial\theta$, and $H_i = \partial^2 \ln f(y_i|\theta)/\partial\theta\partial\theta'$ for $i = 1, ..., n$ are all random samples of random variables. The notation $g_i(\theta_0)$ and $H_i(\theta_0)$ indicates the derivative evaluated at θ_0.

2. $E_0[g_i(\theta_0)] = 0$

3. $Var[g_i(\theta_0)] = -E[H_i(\theta_0)]$

Demonstration of the Key Properties of ML

By definition of a density (with parameter dependent support or range, $[A(\theta_0), B(\theta_0)]$),

$$\int_{A(\theta_0)}^{B(\theta_0)} f(y_i|\theta_0)dy_i = 1$$

Differentiating w.r.t. θ

$$\begin{aligned}\frac{\partial \int_{A(\theta_0)}^{B(\theta_0)} f(y_i|\theta_0)dy_i}{\partial\theta} &= \int_{A(\theta_0)}^{B(\theta_0)} \frac{\partial f(y_i|\theta_0)dy_i}{\partial\theta} + f(B(\theta_0)|\theta_0)\frac{\partial B(\theta_0)}{\partial\theta} \\ &\quad - f(A(\theta_0)|\theta_0)\frac{\partial A(\theta_0)}{\partial\theta} \\ &= 0,\end{aligned}$$

where the second and third terms are zero if the density at the end points is zero or the support does not depend on θ. (This allows passing derivatives through the integral sign.) Assuming this, we proceed.

Demonstration of the Key Properties of ML (2)

Since, schematically $[\ln f]' = f'/f$, where $'$ indicates differentiation, (so $f' =$[ln $f]' *$ f) we have:

$$\int_{A(\theta_0)}^{B(\theta_0)} f(y_i|\theta_0)dy_i = 1$$

$$\begin{aligned}
\frac{\partial \int f(y_i|\theta_0)dy_i}{\partial\theta} &= \int \frac{\partial f(y_i|\theta_0)dy_i}{\partial\theta} = \int \frac{\partial \ln f(y_i|\theta_0)}{\partial\theta} f(y_i|\theta_0)dy_i \\
&= E_0\left[\frac{\partial \ln f(y_i|\theta_0)}{\partial\theta}\right] \\
&= 0
\end{aligned}$$

This is the ‘likelihood equation’: ‘the expectation of the score at the true parameter value is zero’. (The first derivative of the log likelihood is called the score.)

Demonstration of the Key Properties of ML: The Information Matrix Equality

Let's differentiate again:

$$\frac{\partial}{\partial\theta'}\int\frac{\partial \ln f(y_i|\theta_0)}{\partial\theta}f(y_i|\theta_0)dy_i$$
$$= \int\left\{\frac{\partial^2 \ln f(y_i|\theta_0)}{\partial\theta\partial\theta'}f(y_i|\theta_0)+\frac{\partial \ln f(y_i|\theta_0)}{\partial\theta}\frac{\partial f(y_i|\theta_0)}{\partial\theta'}\right\}dy_i = 0$$

But

$$\frac{\partial f(y_i|\theta_0)}{\partial\theta'}=\frac{\partial \ln f(y_i|\theta_0)}{\partial\theta'}f(y_i|\theta_0)$$

so

$$\int\left\{\frac{\partial^2 \ln f(y_i|\theta_0)}{\partial\theta\partial\theta'}f(y_i|\theta_0)+\frac{\partial \ln f(y_i|\theta_0)}{\partial\theta}\frac{\partial \ln f(y_i|\theta_0)}{\partial\theta'}f(y_i|\theta_0)\right\}=0, \quad thus$$

$$Var_0\left[\frac{\partial \ln f(y_i|\theta_0)}{\partial\theta}\right]=E_0\left[\frac{\partial \ln f(y_i|\theta_0)}{\partial\theta}\frac{\partial \ln f(y_i|\theta_0)}{\partial\theta'}\right]=-E_0\left[\frac{\partial^2 \ln f(y_i|\theta_0)}{\partial\theta\partial\theta'}\right]$$

The Information Matrix Equality: Why We Will Care

The expression

$$-E_0\left[\frac{\partial^2 \ln f(y_i|\theta_0)}{\partial\theta\partial\theta'}\right]$$

is known as the 'information matrix', and it turns out that $\sqrt{n}(\hat{\theta}-\theta_0)\sim N(0, E_0\left[\frac{\partial^2 \ln f(y_i|\theta_0)}{\partial\theta\partial\theta'}\right])$. So the information matrix equality will say 'that the variance of the score equals the expectation of the information matrix (Hessian of the log likelihood) at the true parameter value, and these both equal the asymptotic covariance matrix of the estimated parameters.'