14: Simultaneous Equations Models and Instrumental Variables
ECON 837

Brian Krauth (adapted from notes by Simon Woodcock), Spring 2010

The big picture

Most of the time, applied econometricians are not satisfied with a simply statistical description of the DGP for a given data set. In addition we want to make inferences about the economic mechanisms (preferences, constraints, and markets) underlying that DGP.

More generally, we are often interested in using the record of what *did* happen (i.e., the data) to make counterfactual predictions about what *would* happen under different circumstances. Doing so is a necessary condition for making any policy conclusions. For example, if we want to know the consequences of an increase in the minimum wage, we need to estimate the demand and supply elasticity for unskilled labor. These elasticities are at root statements about counterfactual quantities: what would workers and firms do if the wage were different?

In order to answer questions about counterfactuals, we will need to write down an explicit structural (or causal) model whose unknown parameters (usually called the structural parameters) happen to coincide with features of the DGP that we can estimate. When we can do this, we say that these structural parameters are *identified* from our data.

An example: Supply and demand

The classic example is the textbook supply and demand model. Suppose that we are interested in understanding the supply and demand for some particular good. We have n independent observations (i.e., different locations or points in time) on the market for this good, and for each observation we have data on the (log of the) good's market price p_t, the (log of the) quantity sold q_t and some k-vector of additional factors X_t (weather, income, taxes, etc.) that may affect supply and/or demand. We have the economic model:

$$\begin{aligned} q_t^D &= \eta_d p_t + X_t\Gamma + \epsilon_{dt} \\ q_t^S &= \eta_s p_t + X_t\Lambda + \epsilon_{st} \\ q_t &= q_t^S = q_t^D \end{aligned}$$

where the first equation is the demand curve, the second is the supply curve and the third is the market equilibrium condition that supply equals demand. The structural parameter η_d is the elasticity of demand, and the structural parameter η_s is the elasticity of supply.

We think of $(X_t, \epsilon_{dt}, \epsilon_{st})$ as exogenous, and (q_t, p_t) as endogenous. That is, our model is a model of the determination of (q_t, p_t).

We can solve this system of two equations for the two endogenous variables, and we get:

$$\begin{aligned} p_t &= X_t\left(\frac{\Gamma-\Lambda}{\eta_s-\eta_d}\right)+\frac{\epsilon_{dt}-\epsilon_{st}}{\eta_s-\eta_d} \\ &= X_t\Pi_1+v_{pt} \\ q_t &= X_t\left(\frac{\eta_s\Gamma-\eta_d\Lambda}{\eta_s-\eta_d}\right)+\frac{\eta_s\epsilon_{dt}-\eta_d\epsilon_{st}}{\eta_s-\eta_d} \\ &= X_t\Pi_2+v_{qt} \end{aligned}$$

This is called the *reduced form* of the model. The reduced form expresses the endogenous variables as functions of the exogenous variables.

Now:

- If we make the standard assumption that $E(\epsilon_{dt}|X_t) = E(\epsilon_{st}|X_t) = 0$, then we can consistently estimate the reduced form coefficients (π_1, π_2) by OLS.
- Since the reduced form residuals are correlated with each other, maybe we can use the information in that correlation to improve our estimates (using GLS).
- We cannot directly estimate the original structural equations by OLS, because the reduced form implies that p_t is necessarily correlated with both ϵ_{st} and ϵ_{dt}. So we cannot assume that $E(\epsilon_{dt}|X_t, p_t) = 0$ or that $E(\epsilon_{st}|X_t, p_t) = 0$.
- However, the $2k$ reduced form coefficients impose $2k$ nonlinear constraints on the values of the $2k+2$ unknown structural parameters. If we impose additional restrictions, we may be able to identify some of the structural parameters.

Identification in the SEM

Now, without any additional assumptions we cannot recover the structural parameters. We will need to impose additional identifying assumptions.

The most common kind of identifying assumption used in the SEM model is an exclusion restriction. Suppose that we can partition X_t into $[X_{st}X_{dt}X_{bt}]$ where X_{st} is a k_s-vector of "supply shifters" (variables that affect supply but not demand) and X_{dt} is a k_d-vector of "demand shifters" (variables that affect demand but not supply). For example, if we are analyzing the market for oranges, we might use the weather in Florida as our supply shifter. Apply the same partition to Γ, Λ, Π_1 and Π_2, and our exclusion restrictions can be written as:

$$\begin{aligned} \Gamma_{dt} &\neq 0 \\ \Gamma_{st} &= 0 \\ \Lambda_{dt} &= 0 \\ \Lambda_{st} &\neq 0 \end{aligned}$$

Then:

$$\begin{aligned}
\frac{\Pi_2}{\Pi_1} &= \frac{\left(\frac{\eta_s\Gamma-\eta_d\Lambda}{\eta_s-\eta_d}\right)}{\left(\frac{\Gamma-\Lambda}{\eta_s-\eta_d}\right)} \\
&= \frac{\eta_s\Gamma-\eta_d\Lambda}{\Gamma-\Lambda} \\
&= \left[\begin{array}{ccc} \frac{\eta_s\Gamma_s-\eta_d\Lambda_s}{\Gamma_s-\Lambda_s} & \frac{\eta_s\Gamma_d-\eta_d\Lambda_d}{\Gamma_d-\Lambda_d} & \frac{\eta_s\Gamma_b-\eta_d\Lambda_b}{\Gamma_b-\Lambda_b} \end{array}\right] \\
&= \left[\begin{array}{ccc} \frac{\eta_s\Gamma_s-\eta_d\Lambda_s}{\Gamma_s-\Lambda_s} & \frac{\eta_s\Gamma_d-\eta_d\Lambda_d}{\Gamma_d-\Lambda_d} & \frac{\eta_s\Gamma_b-\eta_d\Lambda_b}{\Gamma_b-\Lambda_b} \end{array}\right] \\
&= \left[\begin{array}{ccc} i_{ks}\eta_d & i_{kd}\eta_s & \frac{\eta_s\Gamma_b-\eta_d\Lambda_b}{\Gamma_b-\Lambda_b} \end{array}\right]
\end{aligned}$$

where $\frac{A}{B}$ just means element-by-element division of conformable matrices.

Now, we can analyze identification in the SEM. There are three cases to consider:

- Underidentification ($k_s = 0$): If we have no exogenous supply shifters, then there will be no consistent estimator of η_d.
- Exact identification ($k_s = 1$): If we have exactly one exogenous supply shifter then we have a consistent estimator of η_d, and our exclusion restriction will not be testable.
- Overidentification ($k_s > 1$): If we have more than one exogenous supply shifter then we have more than one value for η_d.
  - From an estimation point of view, this means we have multiple consistent estimators. Is one better than the others? Or maybe some average of them is better than any individual estimator.
  - From a specification testing point of view, this means we can test our set of exclusion restrictions.

Our identification analysis suggests an estimation method: estimate the reduced form coefficients (Π_1, Π_2) by equation-by-equation OLS and then calculate $\frac{\Pi_2}{\Pi_1}$. This method is known as *indirect least squares* and will yield CAN estimates. But maybe we can improve on it.

General Notation for Linear SEM

We will consider the general case with M equations, M endogenous variables, k exogenous variables, and T observations. We write the system of equations as

$$\begin{aligned}
\gamma_{11}y_{1t} + \gamma_{21}y_{2t} + \cdots + \gamma_{M1}y_{Mt} + \beta_{11}x_{1t} + \cdots + \beta_{k1}x_{kt} &= \varepsilon_{1t} \\
\gamma_{12}y_{1t} + \gamma_{22}y_{2t} + \cdots + \gamma_{M2}y_{Mt} + \beta_{12}x_{1t} + \cdots + \beta_{k2}x_{kt} &= \varepsilon_{2t} \\
&\vdots \\
\gamma_{1M}y_{1t} + \gamma_{2M}y_{2t} + \cdots + \gamma_{MM}y_{Mt} + \beta_{1M}x_{1t} + \cdots + \beta_{kM}x_{kt} &= \varepsilon_{Mt}.
\end{aligned}$$

In matrix notation, the structural form is

$$\mathbf{y}_t'\mathbf{\Gamma} + \mathbf{x}_t'\mathbf{B} = \varepsilon_t'$$

where

$$\begin{aligned}
\mathbf{y}_t &= \begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{Mt} \end{bmatrix}, \ \mathbf{\Gamma} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1M} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{M1} & \gamma_{M2} & \cdots & \gamma_{MM} \end{bmatrix}, \\
\mathbf{x}_t &= \begin{bmatrix} x_{1t} \\ x_{2t} \\ \vdots \\ x_{kt} \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1M} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{k1} & \beta_{k2} & \cdots & \beta_{kM} \end{bmatrix}, \ \varepsilon_t = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \vdots \\ \varepsilon_{Mt} \end{bmatrix}.
\end{aligned}$$

The m^{th} column of $\mathbf{\Gamma}$ and $\mathbf{B}$ is the vector of coefficients in the m^{th} equation. The m^{th} row of $\mathbf{\Gamma}$ is the vector of coefficients on the m^{th} endogenous variable in each of the M equations. Likewise the k^{th} row of $\mathbf{B}$ is the vector of coefficients on the k^{th} exogenous variable in each of the M equations.

We normalize one endogenous variable in each equation to be the **dependent variable**. The dependent variable in each equation has a coefficient of one. Thus each column of $\mathbf{\Gamma}$ will contain at least one entry of one. In many applications each endogenous variable will be the dependent variable in exactly one equation (so that the diagonal elements of $\mathbf{\Gamma}$ are all equal to one), but as we saw with our example this isn't necessary.

The reduced form of the model is

$$\begin{aligned}
\mathbf{y}_t' &= -\mathbf{x}_t'\mathbf{B}\mathbf{\Gamma}^{-1} + \varepsilon_t'\mathbf{\Gamma}^{-1} \\
&= \mathbf{x}_t'\mathbf{\Pi} + \nu_t'.
\end{aligned}$$

We see that the reduced form only exists if $\mathbf{\Gamma}$ has full rank.

For the structural errors, we assume

$$\begin{aligned}
E\left[\varepsilon_t|\mathbf{x}_t\right] &= \mathbf{0} \\
E\left[\varepsilon_t\varepsilon_t'|\mathbf{x}_t\right] &= \mathbf{\Sigma}
\end{aligned}$$

so that the errors may be correlated across equations. We'll assume that

$$E\left[\varepsilon_t\varepsilon_s'|\mathbf{x}_t, \mathbf{x}_s\right] = \mathbf{0} \ \text{ for } t \neq s$$

though this can be relaxed to allow for serial correlation. These conditions imply similar conditions for the reduced form errors:

$$\begin{aligned}
E\left[\nu_t|\mathbf{x}_t\right] &= E\left[\left(\mathbf{\Gamma}^{-1}\right)'\varepsilon_t|\mathbf{x}_t\right] = \mathbf{0} \\
E\left[\nu_t\nu_t'|\mathbf{x}_t\right] &= E\left[\left(\mathbf{\Gamma}^{-1}\right)'\varepsilon_t\varepsilon_t'\mathbf{\Gamma}^{-1}\right] = \left(\mathbf{\Gamma}^{-1}\right)'\mathbf{\Sigma}\mathbf{\Gamma}^{-1} \equiv \mathbf{\Omega} \\
E\left[\nu_t\nu_s'|\mathbf{x}_t, \mathbf{x}_s\right] &= \mathbf{0} \ \text{ for } t \neq s
\end{aligned}$$

so that $\mathbf{\Sigma} = \mathbf{\Gamma}'\mathbf{\Omega}\mathbf{\Gamma}$.

Stacking up the observations, we write the structural form as

$$\mathbf{Y\Gamma} + \mathbf{XB} = \mathbf{E}$$

and the reduced form as

$$\mathbf{Y} = \mathbf{X\Pi} + \mathbf{V}.$$

Finally, we assume as usual that

$$\begin{aligned} \text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{T}\right) &= \mathbf{Q} \text{ positive definite} \\ \text{plim}\left(\frac{\mathbf{X}'\mathbf{E}}{T}\right) &= \mathbf{0}. \end{aligned}$$

This assumption is what statistically distinguishes the exogenous variables from the endogenous variables, and implies:

$$\text{plim}\left(\frac{\mathbf{X}'\mathbf{V}}{T}\right) = \mathbf{0}.$$

Identification in the general model

As in our example, we can always estimate the reduced form parameters. However, being able to identify the various structural parameters will require us to impose some restrictions on $\mathbf{\Gamma}$, $\mathbf{B}$, and/or $\mathbf{\Sigma}$.

Identification Strategies

The reduced form contains $kM + M(M+1)/2$ distinct coefficients: kM for the $k \times M$ matrix $\mathbf{\Pi}$ and $M(M+1)/2$ for the symmetric $M \times M$ matrix $\mathbf{\Omega}$. The structural form contains $M(M-1) + kM + M(M+1)/2$ unknown parameters: $M(M-1)$ for the $M \times M$ matrix (normalized to have ones on the diagonal) $\mathbf{\Gamma}$, kM for the $k \times M$ matrix $\mathbf{B}$ and $M(M+1)/2$ for the symmetric $M \times M$ matrix $\mathbf{\Sigma}$. So we need $M(M-1)$ additional restrictions.

Classical Identification: Rank and Order Conditions

How do we know whether our exclusion restrictions are "enough" to identify all the structural parameters in a given equation? By checking two related conditions: the **rank condition**, and the **order condition**.

If we isolate one equation (equation j) of our system, we can write:

$$\mathbf{Y\Gamma}_j + \mathbf{XB}_j = \varepsilon_j$$

where $\mathbf{\Gamma}_j$ is the j^{th} column of $\mathbf{\Gamma}$, and $\mathbf{B}_j$ is the j^{th} column of $\mathbf{B}$. Recall that we normalize the coefficient on one endogenous variable to 1 in each equation. That is, one element of $\mathbf{\Gamma}_j$

is normalized to 1. We call the corresponding endogenous variable the dependent variable in equation j. We can re-label and partition the variables in a convenient way:

$$\begin{array}{rcll}
\mathbf{Y}' & = & \left[\begin{array}{c} y_j \\ \mathbf{y}_j \\ \mathbf{y}_j^* \end{array}\right] & \begin{array}{c} \text{dependent variable, } 1 \times T \\ \text{included endogenous variables, } M_j \times T \\ \text{excluded endogenous variables, } M_j^* \times T \end{array} \\
\mathbf{X}' & = & \left[\begin{array}{c} \mathbf{x}_j \\ \mathbf{x}_j^* \end{array}\right] & \begin{array}{c} \text{included exogenous variables, } k_j \times T \\ \text{excluded exogenous variables, } k_j^* \times T \end{array} \\
\mathbf{\Gamma}_j & = & \left[\begin{array}{c} 1 \\ -\gamma_j \\ \mathbf{0} \end{array}\right] & \begin{array}{c} 1 \times 1 \\ M_j \times 1 \text{ (notice the sign convention)} \\ M_j^* \times 1 \end{array} \\
\mathbf{B}_j & = & \left[\begin{array}{c} \beta_j \\ \mathbf{0} \end{array}\right] & \begin{array}{c} k_j \times 1 \\ k_j^* \times 1 \end{array} \\
\mathbf{\Pi} & = & \left[\begin{array}{ccc} \pi_j & \mathbf{\Pi}_j & \bar{\mathbf{\Pi}}_j \\ \pi_j^* & \mathbf{\Pi}_j^* & \bar{\mathbf{\Pi}}_j^* \end{array}\right] & \begin{array}{ccc} k_j \times 1 & k_j \times M_j & k_j \times M_j^* \\ k_j^* \times 1 & k_j^* \times M_j & k_j^* \times M_j^* \end{array}
\end{array}$$

Note that for each equation j, $M_j + M_j^* + 1 = M$ and $k_j + k_j^* = k$.

We know the reduced form coefficient matrix is given by $\mathbf{\Pi} = -\mathbf{B}\mathbf{\Gamma}^{-1}$. Hence $\mathbf{\Pi}\mathbf{\Gamma} = -\mathbf{B}$. The j^{th} column of this matrix corresponds to the j^{th} equation of our system:

$$\mathbf{\Pi}\mathbf{\Gamma}_j = -\mathbf{B}_j.$$

Substituting in we get:

$$\left[\begin{array}{ccc} \pi_j & \mathbf{\Pi}_j & \bar{\mathbf{\Pi}}_j \\ \pi_j^* & \mathbf{\Pi}_j^* & \bar{\mathbf{\Pi}}_j^* \end{array}\right] \left[\begin{array}{c} 1 \\ -\gamma_j \\ \mathbf{0} \end{array}\right] = \left[\begin{array}{c} \beta_j \\ \mathbf{0} \end{array}\right] \tag{1}$$

We can rewrite (1) as $k_j + k_j^*$ equations:

$$\pi_j - \mathbf{\Pi}_j\gamma_j = \beta_j \tag{2}$$
$$\pi_j^* - \mathbf{\Pi}_j^*\gamma_j = \mathbf{0}. \tag{3}$$

It is from these two equation systems that we see the technical requirements that our exclusion restrictions must satisfy for the structural parameters (γ_j and β_j) to be identified. First note that (3) can be rewritten

$$\mathbf{\Pi}_j^*\gamma_j = \pi_j^*. \tag{4}$$

If we can solve (4) **for** γ_j, **we can substitute the solutions into** (2) **and obtain solutions for** β_j. So, we need to be able to solve (4) for γ_j. Note this is a system of k_j^* equations in M_j unknowns. A solution only exists if there are no more unknowns than equations. Hence we have a necessary condition for a solution to exist, and hence a necessary (but not sufficient) condition for the structural parameters in equation j to be identified:

Definition 1 (Order Condition for Equation j**)** $k_j^* \geq M_j$. *That is, the number of exogenous variables excluded from equation j must be at least as large as the number of included endogenous variables.*

The order condition guarantees that at least one solution for γ_j (and hence β_j) exists, but does not rule out multiple solutions. A purely technical, but sufficient, condition for (4) to have a unique solution is that $\mathbf{\Pi}_j^*$ have full column rank.

Definition 2 (Rank Condition for Equation j) *$rank\left(\mathbf{\Pi}_j^*\right) = M_j$. That is, $\mathbf{\Pi}_j^*$ has full column rank.*

A few things to note about the rank condition:

1. What is $\mathbf{\Pi}_j^*$? It appears in the reduced form for $\mathbf{y}_j$:

   $$\mathbf{y}_j = \mathbf{\Pi}_j \mathbf{x}_j + \mathbf{\Pi}_j^* \mathbf{x}_j^* + \mathbf{v}_j$$

   That is, when we estimate the OLS regression of the included endogenous variables on the exogenous variables, it is the matrix of coefficients on the excluded exogenous variables.

2. If you've used two-stage least squares before, you may recognize this regression as the first stage regression, and the rank condition as the "relevance" condition on the instruments.

3. Since it is a condition on a parameter matrix that is identified, the rank condition is testable.

   (a) General tests for the rank of a random matrix are not easy to implement, but are available and increasingly common in use.

   (b) If there is only one included endogenous variable (i.e., $M_j = 1$) then the rank condition reduces to $\mathbf{\Pi}_j^* \neq \mathbf{0}$. This can be tested using ordinary t or F tests.

4. Note that the order condition is necessary but not sufficient for the rank condition. To see why, note that $\mathbf{\Pi}_j^*$ is $k_j^* \times M_j$, so $rank\left(\mathbf{\Pi}_j^*\right) \leq \min(k_j^*, M_j)$.

Finally, note that even if the rank and order conditions are not satisfied, the model may be still be identified; for example, if we place linear restrictions on the coefficients or restrictions on the error covariance. See Greene p. 370 (pp. 394-395 in 5th ed.) for some examples.

SEM Estimation

Reduced Form Estimation

Suppose we estimate each of the reduced form equations by least squares. The least squares estimator of the $k \times M$ coefficient matrix $\mathbf{\Pi}$ is

$$\begin{aligned}
\hat{\mathbf{\Pi}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{X}\mathbf{\Pi} + \mathbf{V}\right) \\
&= \mathbf{\Pi} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}
\end{aligned}$$

We know that

$$\begin{aligned}\text{plim } \hat{\mathbf{\Pi}} &= \mathbf{\Pi} + \left[\text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)\right]^{-1} \text{plim}\left(\frac{\mathbf{X}'\mathbf{V}}{T}\right) \\ &= \mathbf{\Pi} + \mathbf{Q}^{-1}\mathbf{0} \\ &= \mathbf{\Pi}.\end{aligned}$$

The least squares residuals are:

$$\begin{aligned}\hat{V} &= \mathbf{Y} - \mathbf{X}\hat{\mathbf{\Pi}} \\ &= \mathbf{X}\mathbf{\Pi} + \mathbf{V} - \mathbf{X}\left(\mathbf{\Pi} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\right) \\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{V} \\ &= \mathbf{M}\mathbf{V}\end{aligned}$$

Consider the following estimator of $\mathbf{\Omega}$:

$$\begin{aligned}\hat{\mathbf{\Omega}} &= \frac{\hat{\mathbf{V}}'\hat{\mathbf{V}}}{T} \\ &= \frac{(\mathbf{M}\mathbf{V})'\mathbf{M}\mathbf{V}}{T} \\ &= \frac{\mathbf{V}'\mathbf{M}'\mathbf{M}\mathbf{V}}{T} \\ &= \frac{\mathbf{V}'\mathbf{M}\mathbf{V}}{T} \\ &= \frac{\mathbf{V}'\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\mathbf{V}}{T} \\ &= \frac{\mathbf{V}'\mathbf{V}}{T} - \frac{\mathbf{V}'\mathbf{X}}{T}\left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1}\frac{\mathbf{X}'\mathbf{V}}{T}.\end{aligned}$$

We know that

$$\begin{aligned}\text{plim}\hat{\mathbf{\Omega}} &= \text{plim}\left(\frac{\mathbf{V}'\mathbf{V}}{T}\right) - \text{plim}\left(\frac{\mathbf{V}'\mathbf{X}}{T}\right)\left[\text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)\right]^{-1}\text{plim}\left(\frac{\mathbf{X}'\mathbf{V}}{T}\right) \\ &= \left(\mathbf{\Gamma}^{-1}\right)'\text{plim}\left(\frac{\mathbf{E}'\mathbf{E}}{T}\right)\mathbf{\Gamma}^{-1} - \mathbf{0}\mathbf{Q}^{-1}\mathbf{0} \\ &= \left(\mathbf{\Gamma}^{-1}\right)'\mathbf{\Sigma}\mathbf{\Gamma}^{-1} \\ &= \mathbf{\Omega}.\end{aligned}$$

Thus we can always consistently estimate the parameters of the reduced form.

Estimation by Instrumental Variables

In motivating the SEM exercise, we saw the indirect least squares estimator. This turns out to be a special case of the most common SEM estimator, the **instrumental variables** (IV)

estimator. If we stack up the observations in equation j, we can write

$$\begin{aligned}\mathbf{y}_j &= \mathbf{Y}_j\gamma_j + \mathbf{X}_j\beta_j + \varepsilon_j \\ &= \mathbf{Z}_j\delta_j + \varepsilon_j.\end{aligned}$$

where $\mathbf{Z}_j$ is a $T \times (M_j + k_j)$ matrix.

The OLS estimator of δ_j is

$$\hat{\delta}_j = \left[\mathbf{Z}_j'\mathbf{Z}_j\right]^{-1}\mathbf{Z}_j'\mathbf{y}_j = \delta_j + \begin{bmatrix} \mathbf{Y}_j'\mathbf{Y}_j & \mathbf{Y}_j'\mathbf{X}_j \\ \mathbf{X}_j'\mathbf{Y}_j & \mathbf{X}_j'\mathbf{X}_j \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Y}_j'\varepsilon_j \\ \mathbf{X}_j'\varepsilon_j \end{bmatrix}$$

which is inconsistent for δ_j since $\text{plim}\frac{1}{T}\mathbf{Y}_j'\varepsilon_j \neq \mathbf{0}$.

The IV estimator is a consistent method of estimating δ_j. It is based on a $T \times (M_j + k_j)$ matrix $\mathbf{W}_j$, that we call the matrix of **instruments**, that satisfies the requirements

$$\begin{aligned}\text{plim}\left(\frac{\mathbf{W}_j'\mathbf{W}_j}{T}\right) &= \mathbf{P}_j \text{ positive definite.} \\ \text{plim}\left(\frac{\mathbf{W}_j'\mathbf{Z}_j}{T}\right) &= \mathbf{Q}_j \text{ finite and nonsingular} \\ \text{plim}\left(\frac{\mathbf{W}_j'\varepsilon_j}{T}\right) &= \mathbf{0}\end{aligned}$$

The first two of these conditions (taken together) are sometimes called the "relevance" condition, while the third condition is sometimes called the "exogeneity" condition. Note that relevance is testable while exogeneity is not.

Now suppose we transform the model as

$$\mathbf{W}_j'\mathbf{y}_j = \mathbf{W}_j'\mathbf{Z}_j\delta_j + \mathbf{W}_j'\varepsilon_j.$$

The IV estimator can be derived by OLS regression of $\mathbf{W}_j'\mathbf{y}_j$ on $\mathbf{W}_j'\mathbf{Z}_j$:

$$\begin{aligned}\delta_j^{IV} &= \left(\mathbf{Z}_j'\mathbf{W}_j\mathbf{W}_j'\mathbf{Z}_j\right)^{-1}\mathbf{Z}_j'\mathbf{W}_j\mathbf{W}_j'\mathbf{y}_j \\ &= \left(\mathbf{W}_j'\mathbf{Z}_j\right)^{-1}\left(\mathbf{Z}_j'\mathbf{W}_j\right)^{-1}\mathbf{Z}_j'\mathbf{W}_j\mathbf{W}_j'\mathbf{y}_j \\ &= \left(\mathbf{W}_j'\mathbf{Z}_j\right)^{-1}\mathbf{W}_j'\mathbf{y}_j. \\ &= \left(\mathbf{W}_j'\mathbf{Z}_j\right)^{-1}\mathbf{W}_j'\left(\mathbf{Z}_j\delta_j + \varepsilon_j\right) \\ &= \delta_j + \left(\mathbf{W}_j'\mathbf{Z}_j\right)^{-1}\mathbf{W}_j'\varepsilon_j\end{aligned}$$

and we see the IV estimator is consistent for δ_j, since

$$\text{plim}\left(\delta_j^{IV} - \delta_j\right) = \text{plim}\left(\mathbf{W}_j'\mathbf{Z}_j\right)^{-1}\mathbf{W}_j'\varepsilon_j = \text{plim}\left(\frac{\mathbf{W}_j'\mathbf{Z}_j}{T}\right)^{-1}\text{plim}\left(\frac{\mathbf{W}_j'\varepsilon_j}{T}\right) = \mathbf{0}.$$

As with any estimator based on least squares, we know from the CLT that

$$\frac{\mathbf{W}_j'\varepsilon_j}{\sqrt{T}} \xrightarrow{d} N\left(\mathbf{0}, \sigma_{jj}\mathbf{P}_j\right)$$

where $E\left[\varepsilon_j \varepsilon_j'\right] = \sigma_{jj}\mathbf{I}_T$ is the error variance in equation j. Hence

$$\sqrt{T}\left(\delta_j^{IV} - \delta_j\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma_{jj}\mathbf{Q}_j^{-1}\mathbf{P}_j\mathbf{Q}_j^{-1}\right). \tag{5}$$

Consistently estimating σ_{jj} is straightforward. In particular,

$$\hat{\sigma}_{jj} = \frac{1}{T}\left(\mathbf{y}_j - \mathbf{Z}_j\delta_j^{IV}\right)'\left(\mathbf{y}_j - \mathbf{Z}_j\delta_j^{IV}\right) \tag{6}$$

is consistent. Frequently, a degrees of freedom correction is applied to the denominator ($T - M_j - k_j$ in place of T), but asymptotically they are equivalent (and neither is unbiased).

The Choice of Instruments

The obvious question is: what should we use for the instruments $\mathbf{W}_j$? One possibility is to use all k exogenous variables in the system, i.e., $\mathbf{X}$.. Remember that $\mathbf{X}$ and $\mathbf{X}_j$ are different.

Note however, that $\mathbf{X}$ is $T \times k$ while $\mathbf{W}_j$ needs to be $T \times (M_j + k_j)$. The order condition for identification requires that $k - k_j = k_j^* \geq M_j$. If we are exactly identified then $k = k_j + M_j$, and we can use $\mathbf{X}$. as our instruments if:

$$\begin{aligned}
\operatorname{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{T}\right) &= \mathbf{P}_j \text{ positive definite.} \\
\operatorname{plim}\left(\frac{\mathbf{X}'\mathbf{Z}_j}{T}\right) &= \mathbf{Q}_j \text{ finite and nonsingular} \\
\operatorname{plim}\left(\frac{\mathbf{X}'\varepsilon_j}{T}\right) &= \mathbf{0}
\end{aligned}$$

The first and third conditions have been assumed from the beginning, so the only additional requirement is the middle condition. The middle condition is closely related to the rank condition for identification, and essentially means that each element of $\mathbf{X}$ is useful in predicting some element of $\mathbf{Z}$.

Two-Stage Least Squares

In the overidentified case, $k > k_j + M_j$. In that case, maybe we can construct some linear combination of $\mathbf{X}$ that obeys the necessary conditions. One example that would work is one that simply uses the first $k_j + M_j$ elements of $\mathbf{X}$. Since there are many linear combinations that would work, this raises the question of whether there is an optimal linear combination.

Two-stage least squares will find this optimal combination. The 2SLS procedure (which no one actually follows directly) is to first estimate an OLS regression of $\mathbf{Z}_j$ on $\mathbf{X}$, and then an OLS regression of $\mathbf{y}_j$ on the fitted values $\hat{\mathbf{Z}}_j$ from the first stage . That is:

$$\hat{\mathbf{Z}}_j = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_j$$

and

$$\begin{aligned}
\hat{\delta}_j^{2SLS} &= (\hat{\mathbf{Z}}_j'\hat{\mathbf{Z}}_j)^{-1}\hat{\mathbf{Z}}_j\mathbf{y}_j \\
&= \left(\mathbf{Z}_j'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}_j\right)^{-1}\mathbf{Z}_j'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}_j
\end{aligned}$$

In the exactly identified case, where $\mathbf{X}'\mathbf{Z}_j$ is invertible, notice this estimator is

$$\begin{aligned}
\hat{\delta}_j^{2SLS} &= \left(\mathbf{Z}_j'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}_j\right)^{-1}\mathbf{Z}_j'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}_j \\
&= \left(\mathbf{X}'\mathbf{Z}_j\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{Z}_j'\mathbf{X}\right)^{-1}\mathbf{Z}_j'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}_j \\
&= \left(\mathbf{X}'\mathbf{Z}_j\right)^{-1}\mathbf{X}'\mathbf{y}_j \\
&= \delta_j^{IV}.
\end{aligned}$$

One note is required about estimating the error variance σ_{jj} in the two-stage least squares framework. Recall our consistent estimator:

$$\hat{\sigma}_{jj} = \frac{1}{T}\left(\mathbf{y}_j - \mathbf{Z}_j\delta_j^{IV}\right)'\left(\mathbf{y}_j - \mathbf{Z}_j\delta_j^{IV}\right) = \frac{1}{T}\left(\mathbf{y}_j - \mathbf{Z}_j\delta_j^{2SLS}\right)'\left(\mathbf{y}_j - \mathbf{Z}_j\delta_j^{2SLS}\right).$$

Notice that the actual data, $\mathbf{Z}_j$, appears here, and **not** $\hat{\mathbf{Z}}_j$. If you manually perform 2SLS estimation (i.e., collect predicted values from the first stage regression and regress $\mathbf{y}_j$ on these and $\mathbf{X}_j$), you will get the right coefficients but the wrong standard errors.

Limited Information vs. Full Information Methods

The estimators we have considered so far are called **limited information** estimators, since they are based on a single equation only. There are two sources of inefficiency inherent in this approach. First, they do not exploit information contained in the cross-equation error covariance. [Recall we assumed $E\left[\varepsilon_t\varepsilon_t'\right] = \mathbf{\Sigma}$, an $M \times M$ matrix.] Second, they impose no structure on the matrix of reduced form coefficients $\mathbf{\Pi}$. Why does this matter? Suppose we estimate each structural equation by 2SLS to obtain estimates of all the structural parameters. These imply a particular estimate of the reduced form parameters: $\mathbf{\Pi}^{2SLS} = -\mathbf{B}^{2SLS}\left(\mathbf{\Gamma}^{2SLS}\right)^{-1}$. This shows that the structural equations imply a set of nonlinear restrictions on the reduced form parameters. However, the first stage regressions are based on an unrestricted matrix of reduced form parameters ($\hat{\mathbf{Y}}_j = \mathbf{X}\hat{\mathbf{\Pi}}_j$), and hence does not exploit all available information.

There is another limited information estimator that we haven't discussed: limited information maximum likelihood (LIML). This is single equation maximum likelihood estimation under the assumption of normality. The LIML estimator of δ has the same asymptotic distribution as the 2SLS estimator.

The alternative to limited information methods is **full information** methods. These estimate the entire equation system simultaneously, and hence incorporate the cross-equation information ignored by limited information methods. The most commonly used full information estimator in this context is called three stage least squares. 3SLS is essentially 2SLS, but with system GLS estimation used to improve on the OLS estimates.

Appendix: Three-Stage Least Squares (3SLS)

The 3SLS estimator is a full information estimator. If we stack the M structural equations, we get

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{Z}_M \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_M \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_M \end{bmatrix}.$$

Write the equation system compactly as $\mathbf{y} = \mathbf{Z}\delta + \varepsilon$. Recall that for observation t, the cross-equation error covariance is $E[\varepsilon_t \varepsilon_t'] = \mathbf{\Sigma}$. Thus $E[\varepsilon\varepsilon'] = \mathbf{\Sigma} \otimes \mathbf{I}_T$. Notice that stacked this way, the equation looks like the SUR model. It is not the same as SUR, however, because the $\mathbf{Z}_j$ contain endogenous variables.

Consider the transformed model

$$(\mathbf{I}_T \otimes \mathbf{X}')\,\mathbf{y} = (\mathbf{I}_T \otimes \mathbf{X}')\,\mathbf{Z}\delta + (\mathbf{I}_T \otimes \mathbf{X}')\,\varepsilon.$$

If we write out the full equation system, we see it is

$$\begin{bmatrix} \mathbf{X}'\mathbf{y}_1 \\ \mathbf{X}'\mathbf{y}_2 \\ \vdots \\ \mathbf{X}'\mathbf{y}_M \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}'\mathbf{Z}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{X}'\mathbf{Z}_M \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_M \end{bmatrix} + \begin{bmatrix} \mathbf{X}'\varepsilon_1 \\ \mathbf{X}'\varepsilon_2 \\ \vdots \\ \mathbf{X}'\varepsilon_M \end{bmatrix}$$

which is just the M equations we used to derive the 2SLS estimator, neatly stacked. The error variance in the transformed model is given by

$$Var\left[(\mathbf{I}_T \otimes \mathbf{X}')\,\varepsilon\right] = (\mathbf{I}_T \otimes \mathbf{X}')\,Var\left[\varepsilon\right](\mathbf{I}_T \otimes \mathbf{X}')' = (\mathbf{I}_T \otimes \mathbf{X}')\,(\mathbf{\Sigma} \otimes \mathbf{I}_T)\,(\mathbf{I}_T \otimes \mathbf{X}')' = \mathbf{\Sigma} \otimes (\mathbf{X}'\mathbf{X})$$

(remember that $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$). The errors are nonspherical, so we know the GLS estimator is BLUE for this model. It is given by

$$\begin{aligned}
\delta_G &= \left[\mathbf{Z}'(\mathbf{I}_T \otimes \mathbf{X})\left(\mathbf{\Sigma} \otimes (\mathbf{X}'\mathbf{X})\right)^{-1}(\mathbf{I}_T \otimes \mathbf{X}')\,\mathbf{Z}\right]^{-1}\mathbf{Z}'(\mathbf{I}_T \otimes \mathbf{X})\left(\mathbf{\Sigma} \otimes (\mathbf{X}'\mathbf{X})\right)^{-1}(\mathbf{I}_T \otimes \mathbf{X}')\,\mathbf{y} \\
&= \left[\mathbf{Z}'(\mathbf{I}_T \otimes \mathbf{X})\left(\mathbf{\Sigma}^{-1} \otimes (\mathbf{X}'\mathbf{X})^{-1}\right)(\mathbf{I}_T \otimes \mathbf{X}')\,\mathbf{Z}\right]^{-1}\mathbf{Z}'(\mathbf{I}_T \otimes \mathbf{X})\left(\mathbf{\Sigma}^{-1} \otimes (\mathbf{X}'\mathbf{X})^{-1}\right)(\mathbf{I}_T \otimes \mathbf{X}')\,\mathbf{y} \\
&= \left[\mathbf{Z}'\left(\mathbf{\Sigma}^{-1} \otimes \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{Z}\right]^{-1}\mathbf{Z}'\left(\mathbf{\Sigma}^{-1} \otimes \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{y} \\
&= \left[\mathbf{Z}'\left(\mathbf{\Sigma}^{-1} \otimes (\mathbf{I}_T - \mathbf{M})\right)\mathbf{Z}\right]^{-1}\mathbf{Z}'\left(\mathbf{\Sigma}^{-1} \otimes (\mathbf{I}_T - \mathbf{M})\right)\mathbf{y}
\end{aligned}$$

(remember that $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$).

Feasible GLS requires a consistent estimate of $\mathbf{\Sigma}$. We can use the 2SLS residuals to construct one quite easily. Define $\hat{\mathbf{\Sigma}}$ as the matrix with (i,j) element

$$\hat{\sigma}_{ij} = \frac{1}{T}\left(\mathbf{y}_i - \mathbf{Z}_i\delta_i^{2SLS}\right)'\left(\mathbf{y}_j - \mathbf{Z}_j\delta_j^{2SLS}\right).$$

Then we call the estimator

$$\delta_{3SLS} = \left[\mathbf{Z}'\left(\hat{\mathbf{\Sigma}}^{-1} \otimes (\mathbf{I}_T - \mathbf{M})\right)\mathbf{Z}\right]^{-1}\mathbf{Z}'\left(\hat{\mathbf{\Sigma}}^{-1} \otimes (\mathbf{I}_T - \mathbf{M})\right)\mathbf{y}$$

the **three-stage least squares** estimator. It is so called because the first two stages are 2SLS, which we use to construct $\hat{\mathbf{\Sigma}}$, and then a GLS third stage. We can in fact rewrite the 3SLS estimator as (verify this yourself)

$$\delta_{3SLS} = \left[\hat{\mathbf{Z}}'\left(\hat{\mathbf{\Sigma}}^{-1} \otimes \mathbf{I}_T\right)\hat{\mathbf{Z}}\right]^{-1}\hat{\mathbf{Z}}'\left(\hat{\mathbf{\Sigma}}^{-1} \otimes \mathbf{I}_T\right)\mathbf{y}$$

and we see the GLS step is just SUR estimation using the 2SLS instruments.