Zellner's Seemingly Unrelated Regressions Model

James L. Powell
Department of Economics
University of California, Berkeley

Overview

The *seemingly unrelated regressions (SUR)* model, proposed by Zellner, can be viewed as a special case of the generalized regression model $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{V}(\mathbf{y}) = \sigma^2\boldsymbol{\Omega}$; however, it does not share all of the features or problems of other leading special cases (e.g., models of heteroskedasticity or serial correlation). While, like those models, the matrix $\boldsymbol{\Omega}$ generally involves unknown parameters which must be estimated, the usual estimators for the covariance matrix of the least squares estimator $\hat{\boldsymbol{\beta}}_{LS}$ are valid, so that the usual inference procedures based on normal theory are valid if the dependent variable $\mathbf{y}$ is multinormal or if the sample size N is large and suitable limit theorems are applicable. Also, unlike those other models, there is little reason to test the null hypothesis $H_0 : \boldsymbol{\Omega} = \mathbf{I}$; the form of $\boldsymbol{\Omega}$ is straightforward and its parameters are easy to estimate consistently, so a feasible version of Aitken's GLS estimator is an attractive alternative to the asymptotically-inefficient LS estimator.

The basic SUR model assumes that, for each individual observation i, there are M dependent variables $y_{i1}, ..., y_{ij}, ..., y_{iM}$ available, each with its own linear regression model:

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}_j + \varepsilon_{ij}, \qquad i = 1, ..., N,$$

or, with the usual stacking of observations over i,

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j$$

for $j = 1, ..., M$, where $\mathbf{y}_j$ and $\boldsymbol{\varepsilon}_j$ are N-vectors and $\mathbf{X}_j$ is an $N \times K_j$ matrix, where

$$K_j = \dim(\boldsymbol{\beta}_j)$$

is the number of regressors for the j^{th} regression.

The standard conditions for the classical regression model are assumed to hold for each j : namely,

$$\begin{aligned} E(\mathbf{y}_j) &= \mathbf{X}_j\boldsymbol{\beta}_j, \\ \mathbf{V}(\mathbf{y}_j) &= \sigma_{jj}\mathbf{I}_N, \end{aligned}$$

with $\mathbf{X}_j$ nonstochastic and $rank(\mathbf{X}_j) = K_j$. Under these conditions, and the additional condition of multinormality of $\mathbf{y}_j$, the usual inference theory is valid for the classical LS estimator of $\boldsymbol{\beta}_j$, applied separately to each equation.

However, the SUR model permits nonzero covariance between the error terms ε_{ij} and ε_{ij} for a given individual i across equations j and k, i.e.,

$$Cov(\varepsilon_{ij}, \varepsilon_{ik}) = \sigma_{ij}$$

while assuming

$$Cov(\varepsilon_{ij}, \varepsilon_{i'k}) = 0$$

if $i \neq i'$. This can be expressed more compactly in matrix form:

$$\mathbf{C}(\boldsymbol{\varepsilon}_j, \boldsymbol{\varepsilon}_k) = \sigma_{jk}\mathbf{I}_{N.}$$

It is the potential nonzero covariance across equations j and k that allows for an improvement in efficiency of GLS relative to the classical LS estimator of each $\boldsymbol{\beta}_j$.

Kronecker Product Notation

Zellner's insight was that, like the usual stacking of the individual dependent variables y_{ij} into an N-vector $\mathbf{y}_j$, those latter vectors can themselves be stacked into an MN-dimensional vector $\mathbf{y}$, with a corresponding arrangement for the error terms, coefficient vectors, and regressors:

$$\underset{(MN\times 1)}{\mathbf{y}} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ ... \\ \mathbf{y}_M \end{pmatrix}, \qquad \underset{(MN\times 1)}{\boldsymbol{\varepsilon}} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ ... \\ \boldsymbol{\varepsilon}_M \end{pmatrix}, \qquad \underset{(K\times 1)}{\boldsymbol{\beta}} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ ... \\ \boldsymbol{\beta}_M \end{pmatrix},$$

and

$$\underset{(MN\times K)}{\mathbf{X}} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & ... & ... \\ ... & ... & ... & \mathbf{0} \\ \mathbf{0} & ... & \mathbf{0} & \mathbf{X}_M \end{pmatrix},$$

with

$$K \equiv \sum_{j=1}^{M} K_j.$$

With this notation, and the individual assumptions for each equation j, it follows that

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$$

and

$$\underset{(MN\times MN)}{\mathbf{V}(\mathbf{y})} = \begin{pmatrix} \sigma_{11}\mathbf{I}_N & \sigma_{12}\mathbf{I}_N & ... & \sigma_{1M}\mathbf{I}_N \\ \sigma_{21}\mathbf{I}_N & \sigma_{22}\mathbf{I}_N & ... & ... \\ ... & ... & ... & ... \\ \sigma_{M1}\mathbf{I}_N & ... & ... & \sigma_{MM}\mathbf{I}_N \end{pmatrix}.$$

This nonscalar covariance matrix is a particular mixture of the matrix

$$\underset{(M\times M)}{\mathbf{\Sigma}} \equiv \begin{pmatrix} \sigma_{11} & \sigma_{12} & ... & \sigma_{1M} \\ \sigma_{21} & \sigma_{22} & ... & ... \\ ... & ... & ... & ... \\ \sigma_{M1} & ... & ... & \sigma_{MM} \end{pmatrix}$$

and the $(N \times N)$ identity matrix $\mathbf{I}_N$. A notation system for such combinations was proposed by the otherwise-despicable mathematician Kronecker, the so-called *Kronecker product* notation; for two matrices $\mathbf{A} \equiv [a_{ij}]$ $(i = 1, ..., L,\ j = 1, ...M)$ and $\mathbf{B}$, the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ is defined as

$$\mathbf{A} \otimes \mathbf{B} \equiv \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & ... & a_{1M}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & ... & ... \\ ... & ... & ... & ... \\ a_{L1}\mathbf{B} & ... & ... & a_{LM}\mathbf{B} \end{pmatrix}.$$

With this notation, clearly

$$\mathbf{V}(\mathbf{y}) = \mathbf{\Sigma} \otimes \mathbf{I}_N$$

for the stacked SUR model.

Kronecker products satisfy a distributive rule, which will come in handy later:

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD},$$

assuming all matrix products are well defined. From this rule follows another for inverses of Kronecker products:

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1},$$

assuming both $\mathbf{A}$ and $\mathbf{B}$ are invertible.

Least Squares and Generalized Least Squares

With the foregoing notation, the classical least squares estimator for the vector $\boldsymbol{\beta}$ can be expressed as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{LS} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &= \begin{pmatrix} \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{y}_1 \\ \left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{y}_2 \\ \cdots \\ \left(\mathbf{X}_M'\mathbf{X}_M\right)^{-1}\mathbf{X}_M'\mathbf{y}_M \end{pmatrix}.\end{aligned}$$

In contrast, the GLS estimator of $\boldsymbol{\beta}$ (assuming $\boldsymbol{\Sigma}$ is known) is

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{GLS} &= \left(\mathbf{X}'\left(\boldsymbol{\Sigma}\otimes\mathbf{I}_N\right)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\left(\boldsymbol{\Sigma}\otimes\mathbf{I}_N\right)^{-1}\mathbf{y} \\ &= \left(\mathbf{X}'\left(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_N\right)\mathbf{X}\right)^{-1}\mathbf{X}'\left(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_N\right)\mathbf{y} \\ &= \begin{pmatrix} \sigma^{11}\left(\mathbf{X}_1'\mathbf{X}_1\right) & \sigma^{12}\left(\mathbf{X}_1'\mathbf{X}_2\right) & \cdots & \sigma^{1M}\left(\mathbf{X}_1'\mathbf{X}_M\right) \\ \sigma^{21}\left(\mathbf{X}_2'\mathbf{X}_1\right) & \sigma^{22}\left(\mathbf{X}_2'\mathbf{X}_2\right) & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \sigma^{M1}\left(\mathbf{X}_M'\mathbf{X}_1\right) & \cdots & \cdots & \sigma^{MM}\left(\mathbf{X}_M'\mathbf{X}_M\right) \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_1'\left(\sum_j \sigma^{1j}\mathbf{y}_j\right) \\ \mathbf{X}_2'\left(\sum_j \sigma^{2j}\mathbf{y}_j\right) \\ \cdots \\ \mathbf{X}_M'\left(\sum_j \sigma^{Mj}\mathbf{y}_j\right) \end{pmatrix},\end{aligned}$$

where σ^{ij} is defined to be the element in the i^{th} row and j^{th} column of $\boldsymbol{\Sigma}^{-1}$, i.e., $\boldsymbol{\Sigma}^{-1} \equiv \left[\sigma^{ij}\right]$.

To get a better idea of what is going on with the GLS estimator, consider the special case $M = 2$, with

$$\hat{\boldsymbol{\beta}}_{LS} \equiv \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} \text{ and } \hat{\boldsymbol{\beta}}_{GLS} \equiv \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix};$$

then it can be shown that the GLS estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ satisfy the two equations

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= \mathbf{b}_1 - \left(\frac{\sigma_{21}}{\sigma_{22}}\right)\left(\mathbf{X}_1'\mathbf{X}_1\right)\mathbf{X}_1'\left(\mathbf{y}_2 - \mathbf{X}_2'\hat{\boldsymbol{\beta}}_2\right), \\ \hat{\boldsymbol{\beta}}_2 &= \mathbf{b}_2 - \left(\frac{\sigma_{12}}{\sigma_{11}}\right)\left(\mathbf{X}_2'\mathbf{X}_2\right)\mathbf{X}_2'\left(\mathbf{y}_1 - \mathbf{X}_1'\hat{\boldsymbol{\beta}}_1\right).\end{aligned}$$

Thus the GLS estimators can be viewed as "adjusted" versions of classical LS, where the adjustment involves the regression of the GLS residuals from the other equation on the regressors from each equation. As noted by Luce in JASA, 1964, the GLS estimator for this model can be calculated sequentially by including appropriately-reweighted residuals from all other equations as additional regressors for each equation.

Another important special case is when the matrix $\boldsymbol{\Sigma}$ is diagonal, i.e., $\sigma_{ij} = 0$ if $i \neq j$. In this case,

since $\Sigma^{-1} = diag[1/\sigma_{ii}]$, it follows that

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{GLS} &= \begin{pmatrix} \frac{1}{\sigma_{11}}\left(\mathbf{X}_1'\mathbf{X}_1\right) & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \frac{1}{\sigma_{22}}\left(\mathbf{X}_2'\mathbf{X}_2\right) & ... & ... \\ ... & ... & ... & \mathbf{0} \\ \mathbf{0} & ... & \mathbf{0} & \frac{1}{\sigma_{MM}}\left(\mathbf{X}_M'\mathbf{X}_M\right) \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{\sigma_{11}}\mathbf{X}_1'\mathbf{y}_1 \\ \frac{1}{\sigma_{22}}\mathbf{X}_2'\mathbf{y}_2 \\ ... \\ \frac{1}{\sigma_{MM}}\mathbf{X}_M'\mathbf{y}_M \end{pmatrix} \\
&= \begin{pmatrix} \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{y}_1 \\ \left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{y}_2 \\ ... \\ \left(\mathbf{X}_M'\mathbf{X}_M\right)^{-1}\mathbf{X}_M'\mathbf{y}_M \end{pmatrix} \\
&\equiv \hat{\boldsymbol{\beta}}_{LS}.
\end{aligned}$$

Not surprisingly, then, if there is no covariance across equations in the error terms, there is no prospect for an efficiency improvement in the GLS estimator relative to LS, applied equation by equation.

Still another important special case is when the matrix of regressors is identical for each equation, i.e., $\mathbf{X}_j \equiv \mathbf{X}_0$ for some $N \times K^*$ matrix $\mathbf{X}_0$, with $K^* = K/M$. Here the stacked matrix $\mathbf{X}$ takes the form

$$\begin{aligned}
\underset{(MN\times K)}{\mathbf{X}} &= \begin{pmatrix} \mathbf{X}_0 & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_0 & ... & ... \\ ... & ... & ... & \mathbf{0} \\ \mathbf{0} & ... & \mathbf{0} & \mathbf{X}_0 \end{pmatrix} \\
&= \left(\mathbf{I}_M \otimes \mathbf{X}_0\right),
\end{aligned}$$

and the GLS estimator also reduces to classical LS:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{GLS} &= \left(\mathbf{X}'\left(\boldsymbol{\Sigma}\otimes\mathbf{I}_N\right)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\left(\boldsymbol{\Sigma}\otimes\mathbf{I}_N\right)^{-1}\mathbf{y} \\
&= \left(\left(\mathbf{I}_M\otimes\mathbf{X}\right)'\left(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_N\right)\left(\mathbf{I}_M\otimes\mathbf{X}\right)\right)^{-1}\left(\mathbf{I}_M\otimes\mathbf{X}\right)'\left(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_N\right)\mathbf{y} \\
&= \left(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{X}'\mathbf{X}\right)^{-1}\left(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{X}'\right)\mathbf{y} \\
&= \left(I_M\otimes\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ ... \\ \mathbf{y}_M \end{pmatrix} \\
&= \begin{pmatrix} \left(\mathbf{X}_0'\mathbf{X}_0\right)^{-1}\mathbf{X}_0'\mathbf{y}_1 \\ \left(\mathbf{X}_0'\mathbf{X}_0\right)^{-1}\mathbf{X}_0'\mathbf{y}_2 \\ ... \\ \left(\mathbf{X}_0'\mathbf{X}_0\right)^{-1}\mathbf{X}_0'\mathbf{y}_M \end{pmatrix} \\
&= \hat{\boldsymbol{\beta}}_{LS}.
\end{aligned}$$

Some intuition for this reduction can be obtained by considering Luce's result that GLS can be obtained iteratively, by starting from classical LS estimators and including the residuals from other equations as

regressors for each equation. Since the LS residuals are, by construction, orthogonal to the common matrix of regressors $\mathbf{X}_0$, their inclusion in each equation will not affect the LS estimates of the $\boldsymbol{\beta}_j$ coefficients. An extension of this argument implies that, if each matrix of regressors $\mathbf{X}_j$ has a submatrix $\mathbf{X}_0$ in common – for example, if all equations have an intercept term – then the GLS coefficients corresponding to those common regressors $\mathbf{X}_0$ will be identical to their LS counterparts.

Feasible GLS

An obvious estimator of the unknown covariance matrix $\boldsymbol{\Sigma} = \mathbf{V}(\mathbf{y}) = [\sigma_{ij}]$ would be $\hat{\boldsymbol{\Sigma}} \equiv [\hat{\sigma}_{ij}]$, with

$$\hat{\sigma}_{jk} \equiv \frac{1}{N}\left(\mathbf{y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}_j\right)'\left(\mathbf{y}_k - \mathbf{X}_k\hat{\boldsymbol{\beta}}_k\right);$$

while these estimators are not unbiased for σ_{jk}, they are consistent under the usual conditions, and obtaining unbiased estimators for σ_{jk} when $j \neq k$ involves more than a simple "degrees of freedom" adjustment. Again imposing reasonable regularity conditions, it can be shown that the feasible GLS estimator

$$\hat{\boldsymbol{\beta}}_{FGLS} = \left(\mathbf{X}'\left(\hat{\boldsymbol{\Sigma}}^{-1}\otimes\mathbf{I}_N\right)\mathbf{X}\right)^{-1}\mathbf{X}'\left(\hat{\boldsymbol{\Sigma}}^{-1}\otimes\mathbf{I}_N\right)\mathbf{y}$$

is asymptotically equivalent to the infeasible GLS estimator which assumes $\boldsymbol{\Sigma}$ is known:

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{FGLS} - \hat{\boldsymbol{\beta}}_{GLS}\right) \xrightarrow{p} 0.$$

Thus

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{FGLS} - \boldsymbol{\beta}\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}),$$

where

$$\begin{aligned}
\mathbf{V} &= \text{plim}\left(\frac{1}{N}\mathbf{X}'\left(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_N\right)\mathbf{X}\right)^{-1} \\
&= \text{plim}\left(\frac{1}{N}\mathbf{X}'\left(\hat{\boldsymbol{\Sigma}}^{-1}\otimes\mathbf{I}_N\right)\mathbf{X}\right)^{-1} \\
&\equiv \text{plim}\ \hat{\mathbf{V}},
\end{aligned}$$

so inference on the parameter vector $\boldsymbol{\beta}$ can be carried out using the approximate normality of $\hat{\boldsymbol{\beta}}_{FGLS}$:

$$\hat{\boldsymbol{\beta}}_{FGLS} \overset{A}{\sim} N\left(\boldsymbol{\beta}, \hat{\mathbf{V}}\right).$$