

## Seemingly Unrelated Regression (SUR) models as a Generalized Least Squares (GLS) Solution to Path Analytic Models

### Abstract

Multivariate regression requires the design matrix for each of  $p$  dependent variables to be the same in its form and assumes that all the coefficients in the model are unknown and estimated from the data. Zellner (1962) formulated Seemingly Unrelated Regression (SUR) models as  $p$  correlated regression equations. Of particular interest, SUR allow each of the  $p$  dependent variables to have a different design matrix with some of the predictor variables being the same. Of particular relevance to path analysis, SUR models allow for a variable to be both in the  $\mathbf{Y}$  and  $\mathbf{X}$  matrices. SUR models are underutilized in educational research. We will explicate how SUR models can be used to solve path analytic models and in other situations.

### Introduction

Standard multivariate regression requires that the design matrix for each of  $p$  dependent variables to be exactly the same in its functional form such that:

$$\mathbf{Y}_{(N \times p)} = \mathbf{X}_{(N \times k)} \mathbf{B}_{(k \times p)} + \boldsymbol{\epsilon}_{(N \times p)}, \quad (1)$$

where  $\mathbf{Y}$  is a matrix of  $p$  dependent variables,  $\mathbf{X}$  is a  $k$ -dimensional design matrix, and  $\boldsymbol{\epsilon}$  is a error matrix, which is assumed to be distributed as  $N_{(N \times p)}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I}_N)$ . Multivariate regression theory using ordinary least squares (OLS) assumes that all of the  $\mathbf{B}$  coefficients in the model are unknown and to be estimated from the data as:

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}. \quad (2)$$

Zellner (1962) formulated the Seemingly Unrelated Regression (SUR) model as  $p$  correlated regression equations, which has also been referred to as multiple-design multivariate (MDM) models (Srivastava, 1967). The  $p$  regression equations are “seemingly unrelated” because taken separately the error terms would follow standard linear OLS linear model form. However, the standard OLS solutions ignore any correlation among the errors across equations; however, because the dependent variables are correlated and the design matrices may contain some of the same variables there may be “contemporaneous” correlation among the errors across the  $p$  equations.

Thus, SUR models are often applied when there may be several equations, which appear to be unrelated; however, they may be related by the fact that: (1) some coefficients are the same or assumed to be zero; (2) the disturbances are correlated across equations; and/or (3) a subset of right hand side variables are the same. This third condition is of particular interest because it allows each of the  $p$  dependent variables to have a different design matrix with some of the predictor variables being the same. In fact, SUR models allow for a variable to be both in the  $\mathbf{Y}$  and  $\mathbf{X}$  matrices, which has particular relevance to path analysis. SUR models are an underused multivariate technique. We will explicate how SUR models can be used to solve path analytic models.

### SUR Model

The SUR model is a generalization of multivariate regression using a vectorized parameter model. The  $\mathbf{Y}$  matrix is vectorized and vertically concatenated,  $\mathbf{y}_v$ . The design matrix,  $\mathbf{D}$ , is formed as a block diagonal with the  $j^{\text{th}}$  design matrix,  $\mathbf{X}_j$ , is on the  $jj^{\text{th}}$  block of the matrix. The model is then expressed as:

$$E[\mathbf{Y}_{(N \times p)}] = \{ \mathbf{X}_{1(N \times m_1)} \boldsymbol{\beta}_1_{(m_1 \times 1)}, \mathbf{X}_{2(N \times m_2)} \boldsymbol{\beta}_2_{(m_2 \times 1)}, \mathbf{X}_{j(N \times m_j)} \boldsymbol{\beta}_j_{(m_j \times 1)}, \mathbf{X}_{p(N \times m_p)} \boldsymbol{\beta}_p_{(m_p \times 1)} \}; \quad (3)$$

where  $m_j$  is the number of parameters estimated (columns) by the  $j^{\text{th}}$  design matrix,  $\mathbf{X}_j$ .

To illustrate, the SUR model is laid out as:

$$E(\mathbf{y}_v) = \begin{bmatrix} \hat{\mathbf{y}}_1 & (N \times 1) \\ \hat{\mathbf{y}}_2 & (N \times 1) \\ \dots & \\ \hat{\mathbf{y}}_j & (N \times 1) \\ \dots & \\ \hat{\mathbf{y}}_p & (N \times 1) \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ (N \times m_1) & \mathbf{X}_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ & (N \times m_2) & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ & & (sym) & \mathbf{X}_j & \mathbf{0} & \mathbf{0} \\ & & & (N \times m_j) & \dots & \mathbf{0} \\ & & & & (N \times m_p) & \mathbf{X}_p \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 & (m_1 \times 1) \\ \boldsymbol{\beta}_2 & (m_2 \times 1) \\ \vdots & \\ \boldsymbol{\beta}_j & (m_j \times 1) \\ \vdots & \\ \boldsymbol{\beta}_p & (m_p \times 1) \end{bmatrix} ; \quad (4)$$

where  $M$  is the total number of parameters estimated over the  $p$  models,  $M = \sum_{j=1}^p m_j$ . The parameter estimates are solved as:

$$\hat{\mathbf{B}} = [ \underset{(M \times Np)}{\mathbf{D}'} \underset{(Np \times Np)}{\mathbf{Q}}^{-1} \underset{(Np \times M)}{\mathbf{D}} ] [ \underset{(M \times Np)}{\mathbf{D}'} \underset{(Np \times Np)}{\mathbf{Q}}^{-1} \underset{(Np \times 1)}{\mathbf{y}_v} ] . \quad (5)$$

$\mathbf{Q}$  is weight matrix based on the residual covariance matrix of the  $\mathbf{Y}$  variables and is formed as:

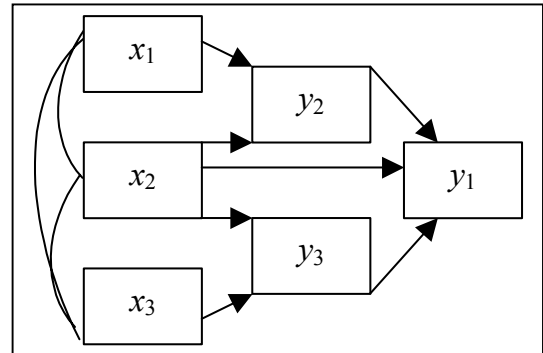
$\mathbf{Q} = \hat{\boldsymbol{\Sigma}} \otimes \mathbf{I}_N$ . The  $ij^{\text{th}}$  element of  $\hat{\boldsymbol{\Sigma}}$  is calculated as:

$$\hat{\sigma}_{ij} = \frac{1}{(N - df^*)} \mathbf{y}'_i [\mathbf{I}_N - \mathbf{H}_i] [\mathbf{I}_N - \mathbf{H}_j] \mathbf{y}_j ; \quad (6)$$

where  $\mathbf{H}_j = \mathbf{X}_j (\mathbf{X}'_j \mathbf{X}_j)^{-1} \mathbf{X}'_j$  is the hat matrix for the  $j^{\text{th}}$  design matrix and  $df^*$  is the average of the numerator degrees-of-freedom ( $df$ ) for the  $i^{\text{th}}$  and  $j^{\text{th}}$  models. Thus, a SUR model is an application of generalized least squares (GLS). In fact, because the residual covariance matrix is unknown and must be estimated from the data, this application is often called feasible generalized least squares (FGLS; see Timm, 2002).

### Path Analysis as a SUR Model

To demonstrate how SUR can be used to solve a path analysis problem, suppose the following path model.  $y_1$  is the “terminal variable and is directly influenced by  $y_2$ ,  $y_3$ , and  $x_2$ .  $x_2$  also has indirect effects on  $y_1$  through  $y_2$  and  $y_3$ .  $x_1$  only has an indirect effect on  $y_1$  through  $y_2$ .  $x_3$  only has an indirect effect on  $y_1$  through  $y_3$ . Assuming standardized variables so that all intercepts will be zero, the correctly specified regression models would be:



$$\begin{aligned} \hat{\mathbf{y}}_1 &= \beta_{1(y_2)} \mathbf{y}_2 + \beta_{1(y_3)} \mathbf{y}_3 + \mathbf{0} \mathbf{X}_1 + \beta_{1(x_2)} \mathbf{X}_2 + \mathbf{0} \mathbf{X}_3 \\ \hat{\mathbf{y}}_2 &= \beta_{2(x_1)} \mathbf{X}_1 + \beta_{2(x_2)} \mathbf{X}_2 + \mathbf{0} \mathbf{X}_3 \\ \hat{\mathbf{y}}_3 &= \mathbf{0} \mathbf{X}_1 + \beta_{3(x_2)} \mathbf{X}_2 + \beta_{3(x_3)} \mathbf{X}_3 \end{aligned} \quad (7)$$

Because the dependent variables are correlated and the design matrices contain some of the same variables there is “contemporaneous” correlation among the errors across the  $p$  equations. However, the standard OLS solutions will ignore any correlation among the errors across these

three equations. The correctly specified SUR model for this path analytic problem would be laid out as such:

$$\begin{array}{l}
 \mathbf{E}(\mathbf{y}_v) = \\
 \begin{array}{l}
 \hat{y}_{11} \\
 \hat{y}_{12} \\
 \dots \\
 \hat{y}_{1N} \\
 \hat{y}_{21} \\
 \hat{y}_{22} \\
 \dots \\
 \hat{y}_{2N} \\
 \hat{y}_{31} \\
 \hat{y}_{32} \\
 \dots \\
 \hat{y}_{3N}
 \end{array}
 \end{array}
 =
 \begin{array}{l}
 \mathbf{D} \\
 \begin{array}{l}
 y_{21} \quad y_{31} \quad x_{21} \\
 y_{22} \quad y_{32} \quad x_{22} \\
 \dots \quad \dots \quad \dots \\
 y_{2N} \quad y_{3N} \quad x_{2N} \\
 \text{(N x 3)} \\
 \text{(sym)} \\
 \text{(N x 7)}
 \end{array}
 \end{array}
 \begin{array}{l}
 \mathbf{B} \\
 \begin{array}{l}
 0 \quad 0 \quad 0 \quad 0 \\
 0 \quad 0 \quad 0 \quad 0 \\
 \dots \\
 0 \quad 0 \quad 0 \quad 0 \\
 x_{11} \quad x_{21} \quad 0 \quad 0 \\
 x_{12} \quad x_{22} \quad 0 \quad 0 \\
 \dots \quad \dots \\
 x_{1N} \quad x_{2N} \quad 0 \quad 0 \\
 \text{(N x 2)} \\
 x_{21} \quad x_{21} \\
 x_{22} \quad x_{22} \\
 \dots \quad \dots \\
 x_{2N} \quad x_{2N} \\
 \text{(N x 2)}
 \end{array}
 \end{array}
 \begin{array}{l}
 \beta_{1(y_2)} \\
 \beta_{1(y_3)} \\
 \beta_{1(x_2)} \\
 \beta_{2(x_1)} \\
 \beta_{2(x_2)} \\
 \beta_{3(x_2)} \\
 \beta_{3(x_3)} \\
 \text{(7 x 1)}
 \end{array}
 \quad (8)$$

Setting this path analysis model up as a SUR model allows for the simultaneous solution of the coefficients in closed form and will produce estimates of the standard errors that take the contemporaneous correlations into account. To develop robust standard error or more precise maximum likelihood (ML) estimates the FGLS solution can be iterated.

### Applications

A researcher interested in conducting a simulation study could compare the bias in the coefficients and standard errors of the correctly specified regression (7) and SUR (8) models and the results from structural equation modeling software (e.g. SAS PROC CALIS; AMOS). One could also assess power and Type I error of correctly specified and misspecified models. For example, one could analyze a model that incorrectly assumes a direct path from  $x_1$  to  $y_1$  and then investigate the Type I error rates produced by the different analytic approaches.

There are many situations in educational and behavioral research in which multiple dependent variables are of interest. Oftentimes these variables may take the pattern of path analytic model, but there are many other cases where they do not. However, it is commonplace for educational researchers to conduct separate analyses for multiple dependent variables even though they are likely to be correlated and have similar although not identical design matrices. For example, researchers in counseling often have multiple outcomes (measure of symptoms, coping, etc.) that are assumed to have some of the same predictors but to also have predictors that are unique to each measure. This is a situation that calls for a SUR model; however, a search of ERIC and PSYCHINFO located less than 10 applications of SUR model despite the enormous number of articles that analyze multiple dependent variables. We contend that SUR models are underutilized and should be given more consideration as an analytic technique. The issue begins with education and thus, we as statistics educators should devote more time to covering SUR models as a flexible analytic tool in our multivariate analyses courses.

### References

- Timm, N. H. (2002). *Applied multivariate analysis*. New York: Springer.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57, 348-368.