Economics 620, Lecture 12: Generalized Least Squares (GLS) II

Nicholas M. Kiefer

Cornell University

Example 1: Grouping (or data on averages)

Suppose the "true" model is $y = X\beta + \varepsilon$ where $E\varepsilon = 0$ and $V(\varepsilon) = \sigma^2 I$.

Suppose that the available data are $[\tilde{y}\ \tilde{X}]$ arranged in $M(> K)$ groups and the group means are the only givens (e.g. city or firm averages). Thus, we have to consider the model

$$\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon}$$

where $\tilde{y}$ is $M \times 1$ and $\tilde{X}$ is $M \times K$.

Now $\tilde{y} = Gy$ where G is $M \times N$. $E\tilde{\varepsilon} = 0$ and $V(\tilde{\varepsilon}) = \sigma^2 GG'$. Similarly, $\tilde{X} = GX$.

Then

$$\hat{\beta}_G = (\tilde{X}'(GG')^{-1}\tilde{X})^{-1}\tilde{X}'(GG')^{-1}\tilde{y}$$

with variance $\sigma^2(\tilde{X}'(GG')^{-1}\tilde{X})^{-1}$. In terms of the unobserved X and y,

$$\hat{\beta}_G = (X'QX)^{-1}X'Qy$$

with $Q = G'(GG')^{-1}G$. Q has an interesting structure.

Let's see what GG' looks like.

With 3 observations in the first group, 4 in the second and 2 in the third, we have

$$G = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

$$GG' = \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}.$$

If P denotes the factor that will be used to transform the model to the standard case, we have

$$P = \begin{bmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{4}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

since $PP' = GG'$. Note that P just divides each observation by its standard deviation.

Example 2: GLS Prediction

Let $Ey = X\beta$ and $V(y) = \sum$. Suppose we know X_{N+1}, and we want to predict y_{N+1}.

Note that $\hat{y}_{N+1} = X_{N+1}\hat{\beta}$ is unbiased. But ε_{N+1} is correlated with ε. How can this correlation be used to improve the forecast?

Let $E\varepsilon_{N+1}\varepsilon = w$ where ε_{N+1} is 1×1 and ε is $N \times 1$.
Let $\hat{y}_{N+1} = c'y$ be a linear predictor.

Unbiasedness implies that $E(\hat{y}_{N+1} - y_{N+1}) = 0 = (c'X - X_{N+1})\beta$. Thus $c'X = X_{N+1}$. (*Why*?)

The prediction error is $c'\varepsilon - \varepsilon_{N+1}$ with variance $c' \sum c + \sigma^2 - 2c'w$.

Minimize this variance subject to $c'X = X_{N+1}$ using the multiplier 2λ.

First order conditions yield

$$\begin{bmatrix} \sum & X \\ X' & 0 \end{bmatrix} \begin{bmatrix} c \\ -\lambda \end{bmatrix} = \begin{bmatrix} w \\ X_{N+1} \end{bmatrix}$$

Solving for c, we get

$$\begin{aligned} c^* &= \sum^{-1}[I - X(X'\sum^{-1}X)^{-1}X'\sum^{-1}]w \\ &\quad + \sum^{-1}X(X'\sum^{-1}X)^{-1}X_{N+1} \end{aligned}$$

using the first part of the partitioned inverse.

Note:

$$\hat{y}_{N+1} = c^{*\prime}y = X_{N+1}\hat{\beta}_G + w'\sum^{-1}(y - X\hat{\beta}_G)$$

which uses the correlation between ε_{N+1} and ε. The last term is the expectation of e_{N+1} given e. $\sum$ and w are usually modeled as functions of a small number of parameters.

SEEMINGLY UNRELATED REGRESSION EQUATIONS (SURE): (Zellner)

Let $y_j = X_j\beta_j + \varepsilon_j$ $j = 1, ..., M$ where y_j is $N \times 1$, X_j is $N \times K$, β_j is $K \times 1$, and ε_j is $N \times 1$.

Stacking up these M regressions yields $y = X\beta + \varepsilon$ where

$$y \text{ is } MN \times 1,$$

$$X \text{ is } MN \times MK,$$

$$\beta \text{ is } MK \times 1, \text{ and}$$

$$\varepsilon \text{ is } MN \times 1.$$

Note that

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix}; X = \begin{bmatrix} X_1 & 0 & \dots & 0 \\ 0 & X_2 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & X_M \end{bmatrix}; \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix}$$

and

$$E\varepsilon\varepsilon' = \begin{bmatrix} E\varepsilon_1\varepsilon_1' & E\varepsilon_1\varepsilon_2' & \dots & E\varepsilon_1\varepsilon_M' \\ E\varepsilon_2\varepsilon_1' & E\varepsilon_2\varepsilon_2' & \dots & E\varepsilon_2\varepsilon_M' \\ \bullet & \bullet & \dots & \bullet \\ E\varepsilon_M\varepsilon_1' & E\varepsilon_M\varepsilon_2' & \dots & E\varepsilon_M\varepsilon_M' \end{bmatrix}$$

Suppose $E\varepsilon_i\varepsilon_i' = \sigma_{ii} I$. (Notation convenction: no square)

Then each equation satisfies the standard conditions but $E\varepsilon_i\varepsilon_j' = \sigma_{ij} I \neq 0$, that is, the equations are correlated with each other. Think of independent observtaions on a system of correlated equations.

Let the $M \times M$ matrix $\sum$ have elements σ_{ij} and $\sum^{-1}$ have elements σ^{ij}. (Note that $\sigma^{ij} \neq \sigma_{ij}^{-1}$.)

Then

$$E\varepsilon\varepsilon' = \sum \otimes I = \begin{bmatrix} \sigma_{11}I & \sigma_{12}I & \dots & \sigma_{1M}I \\ \sigma_{21}I & \sigma_{22}I & \dots & \sigma_{2M}I \\ \bullet & \bullet & \dots & \bullet \\ \sigma_{M1}I & \sigma_{M2}I & \dots & \sigma_{MM}I \end{bmatrix}$$

where $\otimes$ is the Kronecker product.

Note $(\sum \otimes I)^{-1} = \sum^{-1} \otimes I$. (Verify).

Thus $\hat{\beta}_G = (X'(\sum^{-1} \otimes I)X)^{-1}X'(\sum^{-1} \otimes I)y$. Here

$$\sum^{-1} \otimes I = \begin{bmatrix} \sigma^{11}I & \sigma^{12}I & \dots & \sigma^{1M}I \\ \sigma^{21}I & \sigma^{22}I & \dots & \sigma^{2M}I \\ \bullet & \bullet & \dots & \bullet \\ \sigma^{M1}I & \sigma^{M2}I & \dots & \sigma^{MM}I \end{bmatrix}$$

Writing out $\hat{\beta}_G$ gives

$$\hat{\beta}_G = \begin{bmatrix} \sigma^{11} X_1' X_1 & \sigma^{12} X_1' X_2 & \dots & \sigma^{1M} X_1' X_M \\ \bullet & \bullet & \dots & \bullet \\ \sigma^{M1} X_M' X_1 & \sigma^{M2} X_M' X_2 & \dots & \sigma^{MM} X_M' X_M \end{bmatrix}^{-1} \times \begin{bmatrix} \sum_{j=1}^{M} \sigma^{1j} X_1' y_j \\ \vdots \\ \sum_{j=1}^{M} \sigma^{Mj} X_M' y_j \end{bmatrix}$$

where the first term on the right hand side is $(X'(\sum^{-1} \otimes I)X)^{-1}$ and the second term is $X'(\sum^{-1} \otimes I)y$.

The "unbalanced" case, with different sample sizes for different equations, is straightforward with appropriate changes in the dimensions of the "stacked" matrices.

Proposition 1: If $\sigma_{ij} = 0$ for $i \neq j$, then $\hat{\beta}_G = \hat{\beta}$ obtained by estimating each equation by the LS method. In other words, when the equations are not correlated with each other, estimation of each quation by the LS method gives the GLS estimators. (*Why*?)

Proposition 2: If $X_1 = X_2 = ...X_M$, then $\hat{\beta}_G = \hat{\beta}$ obtained by estimating each equation by the LS method.

Proof: Let $X = I \otimes X_1$ where X is $MN \times MK$, I is $M \times M$, and X_1 is $N \times K$.

We know that $\hat{\beta}_G = (X'(\sum^{-1} \otimes I)X)^{-1}X'(\sum^{-1} \otimes I)$y. Substituting for X yields

$$\hat{\beta}_G = ((I \otimes X_1)'(\sum^{-1} \otimes I)(I \otimes X_1))^{-1}(I \otimes X_1)'(\sum^{-1} \otimes I)y.$$

Using the property

$$(A_1 \otimes B_1)(A_2 \otimes B_2) = A_1 A_2 \otimes B_1 B_2,$$

$$\begin{aligned} \hat{\beta}_G &= (\sum^{-1} \otimes X_1' X_1)^{-1} (\sum^{-1} \otimes X_1') y \\ &= (I \otimes (X_1' X_1)^{-1}) X_1' y. \end{aligned}$$

Note that $(X_1' X_1)^{-1} X_1' y$ gives the vector of the LS estimators for the first equation.

Example: Here is an example which shows the *efficiency* gain over the LS estimators.

Consider a two-equation system with $X_1'X_2 = 0$, that is, the regressors in the equations are orthogonal, $R[X_1] \cap R[X_2] = \varnothing$. Then

$$\hat{\beta}_G = \begin{bmatrix} \sigma^{11}X_1'X_1 & 0 \\ 0 & \sigma^{22}X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} \sigma^{11}X_1'y_1 & + & \sigma^{12}X_1'y_2 \\ \sigma^{21}X_2'y_1 & + & \sigma^{22}X_2'y_2 \end{bmatrix}$$

Let $\hat{\beta}_G^1$ represent the vector of coefficients estimates for the first equation. Then

$$\begin{aligned} \hat{\beta}_G^1 &= (X_1'X_1)^{-1}X_1'y_1 \\ &\quad +(\sigma^{12}/\sigma^{11})(X_1'X_1)^{-1}X_1'y_2. \end{aligned}$$

Note that $E\ \hat{\beta}_G^1 = \beta_1$.

We will show that $V(\ \hat{\beta}_G^1) = (\sigma_{11} - \sigma_{12}^2/\sigma_{22})(X_1'X_1)^{-1}$.

Using the expression for $\hat{\beta}_G^1$, we can get

$$\begin{aligned} V(\hat{\beta}_G^1) &= \sigma_{11}(X_1'X_1)^{-1} + (\sigma^{12}/\sigma^{11})^2(X_1'X_1)^{-1}\sigma_{22} \\ &\quad +2(\sigma^{12}/\sigma^{11})\sigma_{12}(X_1'X_1)^{-1}. \ (Why?) \end{aligned}$$

Recall that

$$\begin{bmatrix} \sigma^{11} & \sigma^{12} \\ \sigma^{21} & \sigma^{22} \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = I.$$

This implies that
$\sigma^{11} = \sigma_{22}/(\sigma_{11}\sigma_{22} - \sigma_{12}^2)$ and
$\sigma^{12} = -\sigma_{12}/(\sigma_{11}\sigma_{22} - \sigma_{12}^2)$.

This is a simple trick to remember or reobtain the partitioned inversion formula.

Thus,
$$V(\hat{\beta}_G^1) = [\sigma_{11} + (-\sigma_{12}/\sigma_{22})^2\sigma_{22} - 2(\sigma_{12}/\sigma_{22})\sigma_{12}]\ (X_1'X_1)^{-1}.$$

Hence
$$V(\hat{\beta}_G^1) = (\sigma_{11} - \sigma_{12}^2/\sigma_{22})(X_1'X_1)^{-1} < \sigma_{11}(X_1'X_1)^{-1}.$$

The second term in the variance gives the efficiency gain. It is useful to divide by the variance of the LS estimator, $\sigma_{11}(X_1'X_1)^{-1}$, to get relative efficiency $1 - \rho^2$ where ρ^2 is the correlation squared.

Case of unknown covariance matrix (Feasible SURE):

Let $e_j = y_j - X_j\hat{\beta}_j$ be the vector of LS residuals from equation j.

Define an $M \times M$ matrix S such that
$S = \{S_{jl}\}$ $j, l = 1, 2, ..., M$ where
$S_{jl} = e_j' e_1/(N - K)$.

Note that $ES = \sum$. (That is, S_{jl} is an unbiased estimator of $\sum_{j\ell} = \sigma_{j\ell}$).

(If K and/or N are different in each equation, degrees of freedom have to be adjusted.)

Using S instead of $\sum$ gives Zellner's estimator. This estimator will have the same asymptotic distribution as $\hat{\beta}_G$ calculated with known $\sum$.

Note that the LS estimator is the same as the ML estimator, even with σ^2 unknown. This is not ture in the case of a GLS estimator. (*Why not*?)

Suppose we iterate:

1. Calculate S^k using β^k (with β^1 the LS estimator $\hat{\beta}$).
2. Calculate β^{k+1} using S^k in the GLS formula.
3. If $|\beta^{k+1} - \beta^k| < \eta$ where η is a small number, then stop. Else, go to 1.

On convergence this gives the ML estimator.

Time-series of cross sections

Consider $y_t = X_t\beta_t + \varepsilon_t$ with $t = 1, 2, ..., T$ where y_t is $N \times 1$, X_t is $N \times K$, β_t is $K \times 1$, and ε_t is $N \times 1$.

Notes:

1. Each equation corresponds to a time period; usually T is small and N is large.
2. Allowing time correlation and estimating the $T \times T$ matrix $\sum$ is not difficult.
3. We often want to test $\beta_t = \beta_{t'}$ for all times or a subset.