Econometrics Lecture 3: GLS Estimation and the SURE Model

R. G. Pierse

1 Partitioned Regression

In the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

$\mathbf{y}$ and $\mathbf{u}$ are $n \times 1$ column vectors, $\boldsymbol{\beta}$ is a $k \times 1$ column vector and $\mathbf{X}$ is an $n \times k$ matrix, each column corresponding to a different regressor. Often we are interested in a particular $k_1 \times 1$ subset of the vector of parameters $\boldsymbol{\beta}_1$corresponding to the first k_1 regressors. We can write

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}$$

where $\mathbf{X}_1$ is an $n \times k_1$ matrix, $\mathbf{X}_2$ is an $n \times k_2$ matrix, $\boldsymbol{\beta}_1$ is a $k_1 \times 1$ vector, $\boldsymbol{\beta}_2$ is a $k_2 \times 1$ vector, and $k = k_1 + k_2$, or, alternatively,

$$\mathbf{y} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \mathbf{u}.$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix}$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}.$$

are known as partitioned matrices.

1.1 A formula for the inverse of a partitioned matrix

It is useful to be able to derive an expression for the *OLS* estimators of $\boldsymbol{\beta}_1$ and $bbeta_2$ in terms of the partitioned matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. To do this we need a result for the inverse of a square nonsingular matrix $\mathbf{A}$ partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

where it is assumed that $\mathbf{A}_{22}$ is square and nonsingular. It follows that the matrix

$$\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$$

will also be square and nonsingular. Denoting its inverse as

$$\mathbf{A}^{11} = (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}$$

it can be shown that

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & -\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11} & \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{bmatrix}.$$

Symmetrically, if $\mathbf{A}_{11}$ is square and nonsingular, then

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}^{22}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}^{22} \\ -\mathbf{A}^{22}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{A}^{22} \end{bmatrix}$$

where $\mathbf{A}^{22} = (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}$. These results can be verified by showing that products $\mathbf{A}\mathbf{A}^{-1}$ and $\mathbf{A}^{-1}\mathbf{A}$ do in fact result in the identity matrix

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0}' & \mathbf{I} \end{bmatrix}.$$

Note that it will not always be the case that both $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are nonsingular.

1.2 The *OLS* estimator in partitioned form

Now we can use this result to derive the *OLS* estimator in partitioned form.

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

or

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\begin{bmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix} \tag{1.1}$$

where

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}$$

and, using the partitioned inversion result,

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \mathbf{X}^{11} & -\mathbf{X}^{11}\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1} \\ -(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1\mathbf{X}^{11} & (\mathbf{X}_2'\mathbf{X}_2)^{-1} + (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1\mathbf{X}^{11}\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1} \end{bmatrix} \tag{1.2}$$

where

$$\begin{aligned}\mathbf{X}^{11} &= (\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1)^{-1} \\ &= (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\end{aligned}$$

and

$$\mathbf{M}_2 = \mathbf{I}_n - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2' \tag{1.3}$$

is a square $n \times n$ symmetric idempotent matrix of rank $n - k_2$ with the property that $\mathbf{M}_2\mathbf{X}_2 = \mathbf{0}$.

Multiplying out the expressions (1.1) and (1.2) it can be shown that

$$\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{y}. \tag{1.4}$$

Similarly, using the other form of the partitioned inverse, it can be shown that

$$\widehat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y} \tag{1.5}$$

where

$$\mathbf{M}_1 = \mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1' \tag{1.6}$$

is a square $n \times n$ symmetric idempotent matrix of rank $n - k_1$ with the property that $\mathbf{M}_1\mathbf{X}_1 = \mathbf{0}$.

1.3 Interpretation of partitioned regression

The formula (1.4) for the **OLS** estimator of the parameter vector $\boldsymbol{\beta}_1$offers a very interesting interpretation. Note that the vector $\mathbf{M}_2\mathbf{y}$ is the familiar expression for the *OLS* residuals from a regression of $\mathbf{y}$ on the set of regressors $\mathbf{X}_2$. Similarly, the matrix $\mathbf{M}_2\mathbf{X}_1$ is the set of **OLS** residuals formed from the regression of each of the columns of $\mathbf{X}_1$ on the set of regressors $\mathbf{X}_2$. Then the expression (1.4) can be seen as equivalent to a regression of $\mathbf{M}_2\mathbf{y}$ on $\mathbf{M}_2\mathbf{X}_1$.

This gives an interpretation of (1.4) as a two-step regression procedure where one set of regressors, $\mathbf{X}_1$, is regarded as being the variables of interest, while the other set of regressors, $\mathbf{X}_2$ is regarded as a set of 'nuisance' variables. The first step is a preliminary regression of both $\mathbf{y}$ and $\mathbf{X}_1$ on the nuisance regressors $\mathbf{X}_2$. The residuals from these preliminary regressions are no longer influenced by the effect of $\mathbf{X}_2$ since $\mathbf{M}_2\mathbf{X}_2 = \mathbf{0}$. Having removed the effect of the variables $\mathbf{X}_2$, the second step is a regression of $\mathbf{M}_2\mathbf{y}$ on $\mathbf{M}_2\mathbf{X}_1$. This gives the effect of $\mathbf{X}_1$ on $\mathbf{y}$, having taken into account the effect of the nuisance variables $\mathbf{X}_2$. Of course the algebra is entirely symmetric so that (1.5) could equally well be interpreted as a two-step procedure where $\mathbf{X}_2$ are the variables of interest and $\mathbf{X}_1$ the nuisance variables.

1.4 Example: demeaning data

As a practical example of the usefulness of this algebraic result, consider the case where $\mathbf{X}_1$ has a single column consisting of an intercept term. This is an $n \times 1$ vector of ones which is conventionally written algebraically as ι. Note that

$$\mathbf{X}_1'\mathbf{X}_1 = \iota'\iota = n$$

and, for any $n \times 1$ vector $\mathbf{z}$,

$$\begin{aligned} \mathbf{M}_1\mathbf{z} &= (\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{z} \\ &= (\mathbf{I}_n - \frac{1}{n}\iota\iota')\mathbf{z} = \mathbf{z} - \overline{\mathbf{z}} \end{aligned}$$

where $\overline{\mathbf{z}}$ is a vector with each element equal to the mean $\frac{1}{n}\iota'\mathbf{z} = \frac{1}{n}\sum z_i$. Thus the effect of the idempotent matrix $\mathbf{M}_1$ is to transform a variable by subtracting its mean. It follows that the *OLS* estimator of the remaining $k-1$ coefficients

$$\widehat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y}$$

can be rewritten as

$$\widehat{\boldsymbol{\beta}}_2 = (\overline{\mathbf{X}}_2'\overline{\mathbf{X}}_2)^{-1}\overline{\mathbf{X}}_2'\overline{\mathbf{y}}$$

where $\overline{\mathbf{X}}_2$ and $\overline{\mathbf{y}}$ are the mean deviations of the regressors and regressand respectively.

This result demonstrates that in a regression with an intercept, the *OLS* estimator of the slope coefficients can be derived by first demeaning the data and then estimating the equation without intercept using the demeaned data. This result was used many times in last semester's course.

1.5 The effect of omitted and superfluous variables

The results of partitioned regression can be used to compare the effects of omitting relevant variables from a regression with the effect of including superfluous variables.

Suppose the true regression model is

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u} \tag{1.7}$$

with $\beta_2 \neq \mathbf{0}$ but that the researcher mistakenly estimates the regression

$$\mathbf{y} = \mathbf{X}_1\mathbf{b} + \mathbf{u} \tag{1.8}$$

omitting the relevant variables $\mathbf{X}_2$. What is the effect of this mistake on the properties of the *OLS* estimator? Firstly, note that the estimated parameter vector

$$\begin{aligned} \widehat{\mathbf{b}} &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} \\ &= \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{u} \end{aligned} \tag{1.9}$$

and

$$\begin{aligned} \mathrm{E}(\widehat{\mathbf{b}}) &= \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\,\mathrm{E}(\mathbf{u}) \\ &= \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2. \end{aligned}$$

In general $\widehat{\mathbf{b}}$ will be *biased* except in the special case where $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$. The correct *OLS* estimator is given by

$$\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{y}$$

which is equal to (1.9) only in the case where $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$ in which case $\mathbf{M}_2\mathbf{X}_1 = \mathbf{X}_1$.

The variance of $\widehat{\mathbf{b}}$ is given by

$$\begin{aligned} \mathrm{var}(\widehat{\mathbf{b}}) &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\,\mathrm{var}(\mathbf{u})\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1} \\ &= \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1} \end{aligned}$$

which is the same as it would be were (1.8) the correct model. Thus omitting the variables $\mathbf{X}_2$ does not affect the variance of the estimator. Furthermore, the correct estimator $\widehat{\boldsymbol{\beta}}_1$ has variance

$$\begin{aligned} \mathrm{var}(\widehat{\boldsymbol{\beta}}_1) &= \sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1} \\ &= \sigma^2(\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1)^{-1} \\ &\geq \mathrm{var}(\widehat{\mathbf{b}}) \end{aligned}$$

where the final inequality comes from the result that if $\mathbf{A}$ is positive definite and $\mathbf{B}$ is positive semi-definite then $(\mathbf{A} - \mathbf{B})^{-1} - (\mathbf{A})^{-1}$ is a positive semi-definite matrix. This is an example of how a biased estimator can have a smaller variance than a *BLU* estimator.

Now consider the opposite case where the true model is (1.8) but the researcher mistakenly estimates the model (1.7) including superfluous variables $\mathbf{X}_2$. In this case the estimated coefficient parameter is

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_1 &= (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{y} \\ &= \mathbf{b} + (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{u} \end{aligned}$$

and, by assumption (A1),

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}_1) = \mathbf{b},$$

so that the estimator is *unbiased*.

On the other hand it has variance given by

$$\begin{aligned}\mathrm{var}(\widehat{\boldsymbol{\beta}}_1) &= (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2 \,\mathrm{var}(\mathbf{u})\mathbf{M}_2\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1} \\ &= \sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1} \\ &\geq \mathrm{var}(\widehat{\mathbf{b}}).\end{aligned}$$

Thus including superfluous regressors leads to *unbiased* but *inefficient* estimators since they have a variance higher than that of the *BLU* estimator $\widehat{\mathbf{b}}$.

2 Relaxing the Assumptions of the Classical Model

Consider again the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{2.1}$$

The two classical assumptions on the $n \times 1$ error process $\mathbf{u}$ in this model are:

$$\mathrm{E}(\mathbf{u}) = \mathbf{0} \tag{A1}$$

and

$$\mathrm{var}(\mathbf{u}) = \mathrm{E}(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I}_n. \tag{A2}$$

We now consider relaxing the second of these assumptions and replacing it by the more general specification

$$\mathrm{var}(\mathbf{u}) = \mathrm{E}(\mathbf{u}\mathbf{u}') = \boldsymbol{\Sigma} \tag{A2$'$}$$

where $\boldsymbol{\Sigma}$ is a positive definite matrix. Note that assumption (A2′) allows both for *serial correlation* in the u process (non-zero off-diagonal terms) and *heteroscedasticity* (non-identical diagonal terms). The assumption of positive definiteness is necessary for $\boldsymbol{\Sigma}$ to be a variance-covariance matrix. If $\boldsymbol{\Sigma}$ were not assumed to be positive definite, then some linear combinations of the u's would have variances which are negative or zero which is not allowed.

If $\boldsymbol{\Sigma}$ is positive definite, then it is always possible to find an $n \times n$ matrix $\mathbf{H}$ such that

$$\mathbf{H}'\mathbf{H} = \boldsymbol{\Sigma}^{-1} \tag{2.2}$$

and

$$\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}' = \mathbf{I}_n.$$

The matrix $\mathbf{H}$ that satisfies this property is not unique. One particular $\mathbf{H}$ is defined by

$$\mathbf{H} = \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Q}$$

where $\mathbf{Q}$ is the matrix of the eigenvectors of $\mathbf{\Sigma}^{-1}$ and $\mathbf{\Lambda}^{\frac{1}{2}}$ is a diagonal matrix with the *square roots* of the eigenvalues of $\mathbf{\Sigma}^{-1}$ along the diagonal. Another possible choice for $\mathbf{H}$ is the *lower triangular matrix* from what is called the *Cholesky decomposition* which can always be defined for a positive definite matrix.

Sometimes it is convenient to take out a scale factor from $\mathbf{\Sigma}$ and write

$$\mathbf{\Sigma} = \sigma^2\mathbf{V}$$

where $\mathbf{V}$ is positive definite. The decomposition above can then be applied to $\mathbf{V}^{-1}$ rather than to $\mathbf{\Sigma}^{-1}$. Note that the classical assumption (A2) corresponds to the special case where $\mathbf{V} = \mathbf{I}_n$.

2.1 Properties of *OLS*

Consider the properties of the *OLS* estimator

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{2.3}$$

under assumptions (A1) and (A2′). Substituting from (2.1)

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$

and, by assumption (A1), the expected value of the second term is zero so that

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

and *OLS* is still unbiased. Next consider the variance of $\widehat{\boldsymbol{\beta}}$ which is is now given by

$$\begin{aligned} \mathrm{var}(\widehat{\boldsymbol{\beta}}) &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\,\mathrm{var}(\mathbf{u})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{2.4}$$

This means that the conventional least squares variance formula

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{2.5}$$

is now wrong. Standard statistical tests of the *OLS* coefficient estimates based on this formula will lead to *incorrect inference.*

3 Generalised Least Squares

The least squares principle can be extended to deal with models satisfying assumptions (A1) and (A2′). Firstly, let us assume that the matrix $\mathbf{V}$ is known. Let $\mathbf{L}$ be a matrix satisfying

$$\mathbf{L}'\mathbf{L} = \mathbf{V}^{-1}$$

and

$$\mathbf{L}\mathbf{V}\mathbf{L}' = \mathbf{I}_n.$$

Consider transforming the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

by premultiplying by the matrix $\mathbf{L}$. This gives the transformed model

$$\mathbf{L}\mathbf{y} = \mathbf{L}\mathbf{X}\boldsymbol{\beta} + \mathbf{L}\mathbf{u} \tag{3.1}$$

where the error term $\mathbf{Lu}$ satisfies the properties that

$$\mathrm{E}(\mathbf{Lu}) = \mathbf{L}\,\mathrm{E}(\mathbf{u}) = \mathbf{0}$$

and

$$\begin{aligned} \mathrm{var}(\mathbf{Lu}) &= \mathbf{L}\,\mathrm{var}(\mathbf{u})\mathbf{L}' = \mathbf{L}\boldsymbol{\Sigma}\mathbf{L}' \\ &= \sigma^2\mathbf{I}_n. \end{aligned}$$

Note that the matrix $\mathbf{L}$ transforms the model to one that satisfies the classical assumptions (A1) and (A2) under which *OLS* is the best linear unbiased estimator.

Applying *OLS* to this transformed model (3.1) defines the *Generalised Least Squares* (*GLS*) estimator $\widetilde{\boldsymbol{\beta}}$

$$\begin{aligned} \widetilde{\boldsymbol{\beta}} &= ((\mathbf{LX})'(\mathbf{LX}))^{-1}(\mathbf{LX})'(\mathbf{Ly}) \\ &= (\mathbf{X}'\mathbf{L}'\mathbf{L}\mathbf{X})^{-1}\mathbf{X}'\mathbf{L}'\mathbf{L}\mathbf{y} \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \end{aligned} \tag{3.2}$$

Because the *GLS* estimator is equivalent to *OLS* applied to a transformed model that satisfies the classical assumptions, it follows that *GLS* is the *BLU* estimator in the model (2.1) satisfying the more general assumptions (A1) and (A2′). In particular it follows that the variance covariance matrix of the *GLS* estimator

$$\mathrm{var}(\widetilde{\boldsymbol{\beta}}) = \sigma^2((\mathbf{LX})'(\mathbf{LX}))^{-1} = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \tag{3.3}$$

is smaller than the variance covariance estimator of any other estimator in the class of linear unbiased estimators. This class of course includes the *OLS* estimator $\widehat{\boldsymbol{\beta}}$ with covariance matrix (2.4).

All the results proved in the last lecture for $\widehat{\boldsymbol{\beta}}$ in the classical model carry over to $\widetilde{\boldsymbol{\beta}}$ in the generalised least squares model. An unbiased estimator of σ^2 is therefore given by

$$\begin{aligned}\widetilde{\sigma}^2 &= (\mathbf{L}(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}))'(\mathbf{L}(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}))/(n-k) \\ &= (\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}})'\mathbf{L}'\mathbf{L}(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}})/(n-k) \\ &= \widetilde{\mathbf{e}}'\mathbf{V}^{-1}\widetilde{\mathbf{e}}/(n-k)\end{aligned}$$

where $\widetilde{\mathbf{e}} = (\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}})$ is the vector of *GLS* residuals.

4 Heteroscedasticity

Heteroscedasticity is the property that the diagonal elements of the variance covariance matrix $\boldsymbol{\Sigma}$ are not identical. In the absence of autocorrelation, $\boldsymbol{\Sigma}$ will still be diagonal, and will have the form

$$\mathrm{var}(\mathbf{u}) = \boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} k_1^2 & 0 & \cdots & 0 \\ 0 & k_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & k_n^2 \end{bmatrix}$$

where the condition that all the diagonal elements $k_i^2 > 0$ is necessary to ensure that $\boldsymbol{\Sigma}$ is positive definite. The matrix $\mathbf{L}$ that satisfies

$$\mathbf{L}'\mathbf{L} = \mathbf{V}^{-1}$$

is given by

$$\mathbf{L} = \begin{bmatrix} \frac{1}{k_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{k_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \frac{1}{k_n} \end{bmatrix}$$

where, for any vector $\mathbf{z}$, $\mathbf{Lz}$ has the property that

$$\mathbf{Lz} = \begin{bmatrix} z_1/k_1 \\ z_2/k_2 \\ \vdots \\ z_n/k_n \end{bmatrix}$$

so that the vector $\mathbf{z}$ is scaled by the weights $\frac{1}{k_1}, \frac{1}{k_2}, \cdots, \frac{1}{k_n}$.

The transformed model

$$\mathbf{Ly} = \mathbf{LX}\boldsymbol{\beta} + \mathbf{Lu}$$

is thus equivalent to a regression on scaled variables. The *GLS* estimator for this model is also known as the *Weighted Least Squares* or *WLS* estimator.

5 Autocorrelation

Autocorrelation or *serial correlation* ocurs when the disturbances $\mathbf{u}$ in one period are correlated with those from one or more preceding periods. In terms of the variance covariance matrix $\mathbf{\Sigma}$, it is the presence of any non-zero off-diagonal elements in the matrix. This is too general to deal with easily so that we normally need to have a more precise model of the form that the autocorrelation takes. Specifically, we consider the hypothesis that the error process $\mathbf{u}$ follows a *first-order autoregressive* or *AR(1)* scheme. Since autocorrelation is generally a time series property, we adopt the convention of using the notation T rather than n to represent the number of observations, and use suffices t and s rather than i and j to denote matrix elements.

5.1 The AR(1) model

The *AR(1)* model for the $T \times 1$ vector $\mathbf{u}$ is defined by the equation:

$$\mathbf{u} = \rho \mathbf{u}_{-1} + \varepsilon$$

where the autoregressive coefficient ρ satisfies

$$|\rho| < 1$$

and $\mathbf{u}_{-1}$ represents the lagged vector

$$\mathbf{u}_{-1} = \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_{T-1} \end{bmatrix}$$

with initial value $u_0 = 0$. The error vector ε satisfies the classical assumptions that

$$\text{var}(\varepsilon) = \sigma_\varepsilon^2 \mathbf{I}_T.$$

The variance of each observation of this process is given by

$$\text{E}(u_t^2) = \frac{\sigma_\varepsilon^2}{1-\rho^2} \ .$$

Similarly, the first order covariance elements are given by

$$\begin{aligned} \text{E}(u_t u_{t-1}) &= \text{E}(\rho u_{t-1} + \varepsilon_t)\ u_{t-1} \\ &= \rho\,\text{E}(u_{t-1}^2) + \rho\,\text{E}(u_{t-1}\varepsilon_t) \\ &= \frac{\sigma_\varepsilon^2 \rho}{1-\rho^2} \end{aligned}$$

and, more generally,

$$\mathrm{E}(u_t u_{t-s}) = \frac{\sigma_\varepsilon^2 \rho^s}{1-\rho^2}, \quad s \geq 0 \ .$$

Representing these elements in the variance covariance matrix $\mathbf{\Sigma}$ we have

$$\mathbf{\Sigma} = \sigma_\varepsilon^2 \mathbf{V} = \sigma_u^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \rho \\ \rho^{T-1} & \rho^{T-2} & \cdots & \rho & 1 \end{bmatrix}.$$

where

$$\sigma_u^2 = \frac{\sigma_\varepsilon^2}{1-\rho^2}.$$

The form of the matrix $\mathbf{V}$ where all the diagonals are constant is known as a *band matrix*. The autocorrelation between u_t and u_{t-s} decreases as the distance between the observations, s, increases. If ρ is negative, then this autocorrelation alternates in sign.

The matrix $\mathbf{V}^{-1}$ can be decomposed as

$$\mathbf{V}^{-1} = \mathbf{L}'\mathbf{L}$$

where $\mathbf{L}$ is a lower triangular matrix that takes the form

$$\mathbf{L} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}$$

and which, (excepting the first observation which needs to be treated specially), has the property that

$$\mathbf{Lz} = \mathbf{z} - \rho \mathbf{z}_{-1} = \mathbf{z}^+$$

for any vector $\mathbf{z}$ where by assumption the initial condition $z_0 = 0$. This is known as a *quasi-difference* of the variable $\mathbf{z}$.

If ρ is known, then the *GLS* estimator is equivalent to running the regression in transformed variables

$$\mathbf{Ly} = \mathbf{LX}\boldsymbol{\beta} + \mathbf{Lu}$$

or

$$\mathbf{y}^+ = \mathbf{X}^+\boldsymbol{\beta} + \mathbf{u}^+ = \mathbf{X}^+\boldsymbol{\beta} + \varepsilon.$$

5.2 More general models

More generally, as long as the variance covariance matrix is known, the appropriate transformation matrix $\mathbf{L}$ can be computed and the model estimated by GLS. This can encompass models with both heteroscedasticity and autocorrelation of complex form.

6 Feasible *GLS*

In practice of course, the matrix $\mathbf{V}$ is generally not known, but has to be estimated. The *feasible GLS* estimator is defined by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{y}, \tag{6.1}$$

where the estimated matrix $\widehat{\mathbf{V}}$ replaces the known matrix $\mathbf{V}$. Since the elements of $\widehat{\mathbf{V}}$ are estimated, the matrix is *random* as is the decomposition matrix $\widehat{\mathbf{L}}$. A consequence of this is that the assumption that the transformed regressors $\mathbf{LX}$ are *fixed in repeated samples* must be abandoned. The properties of the *feasible GLS* estimator are justified using asymptotic statistical theory. This will be covered in next week's lecture.

7 Seemingly Unrelated Regression (SURE)

We turn now to consider a simple model involving several equations. More sophisticated models (*simultaneous equation systems*) will be considered at a later stage in this course.

Suppose that the model consists of m equations of the form

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \mathbf{u}_i, \quad i = 1, \cdots, m \tag{7.1}$$

where $\mathbf{y}_i$ and $\mathbf{u}_i$ are $n\times 1$ column vectors, $\boldsymbol{\beta}_i$ is a $k\times 1$ vector of coefficient estimates and $\mathbf{X}_i$ is an $n \times k_i$ matrix of regressors for equation i. The set of regressors $\mathbf{X}_i$ can be completely different for each equation and each can have a different number of regressors k_i. Every equation satisfies the classical assumptions that

$$\mathrm{E}(\mathbf{u}_i) = \mathbf{0}$$

and

$$\mathrm{var}(\mathbf{u}_i) = \sigma_i^2\mathbf{I}_n.$$

These assumptions rule out heteroscedasticity and serial correlation within any individual equation but allow the variances to differ between equations.

7.1 Contemporaneous correlation

The classical assumptions also do not rule out a contemporaneous correlation between the disturbances from different equations of the form

$$\text{cov}(\mathbf{u}_i, \mathbf{u}_j) = \sigma_{ij}\mathbf{I}_n.$$

The assumption of diagonality of this matrix means that the covariance between disturbance $\mathbf{u}_i$ at time period s and disturbance $\mathbf{u}_j$ at time period t is zero for $t \neq s$ and is equal to σ_{ij} for $t = s$, or formally,

$$\begin{aligned} \text{cov}(u_{is}u_{jt}) &= 0, \quad t \neq s \\ &= \sigma_{ij}, \quad t = s. \end{aligned}$$

The possibility that the disturbance terms may be contemporaneously correlated can correspond to the economic notion that shocks to the economy can affect more than one variable simultaneously.

In statistical terms, contemporaneous correlation represents a hidden connection between the m equations in the model that otherwise would be unrelated. This is the explanation for the use of the term 'seemingly unrelated regression' or *SURE* to represent this model.

7.2 *SURE* as a *GLS* model

It is possible to stack the m equations into the single equation

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_m \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_m \end{bmatrix} \tag{7.2}$$

or

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{u}^*$$

where $\mathbf{y}^*$ is an $nm \times 1$ vector, $\mathbf{X}^*$ is an $nm \times \sum k_i$ matrix of regressors, $\boldsymbol{\beta}^*$ is an $\sum k_i \times 1$ vector of coefficients and $\mathbf{u}^*$ is an $nm \times 1$ vector of residuals.

The residual vector $\mathbf{u}^*$ has $\text{E}(\mathbf{u}^*) = \mathbf{0}$ and variance covariance matrix of the form

$$\begin{aligned} \text{var}(\mathbf{u}^*) &= \boldsymbol{\Sigma}^* = \begin{bmatrix} \sigma_{11}\mathbf{I}_n & \sigma_{12}\mathbf{I}_n & \cdots & \sigma_{1m}\mathbf{I}_n \\ \sigma_{21}\mathbf{I}_n & \sigma_{22}\mathbf{I}_n & \cdots & \sigma_{2m}\mathbf{I}_n \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1}\mathbf{I}_n & \sigma_{m2}\mathbf{I}_n & \cdots & \sigma_{mm}\mathbf{I}_n \end{bmatrix} \\ &= \boldsymbol{\Sigma} \otimes \mathbf{I}_n \end{aligned} \tag{7.3}$$

where

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{bmatrix}. \tag{7.4}$$

This covariance matrix is clearly not diagonal. Consequently, the model can be viewed as a special case of the *GLS* model with *theoretical GLS* estimator given by

$$\boldsymbol{\beta}^* = (\mathbf{X}^{*\prime}\mathbf{\Sigma}^{*-1}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{\Sigma}^{*-1}\mathbf{y}^*. \tag{7.5}$$

In practice, the elements of the unknown matrix $\mathbf{\Sigma}$ can be estimated from the residuals from separate *OLS* estimation of each of the m equations, using the formula

$$\widehat{\sigma}_{ij} = \mathbf{e}_i'\mathbf{e}_j/n$$

where e_i represents the *OLS* residual vector for the ith equation.

7.3 *OLS* estimation of the *SURE* model

The model (7.1) can also be estimated by the separate *OLS* regression of each equation given by

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_i &= (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{y}_i, \quad i = 1, \cdots, m \\ &= \boldsymbol{\beta}_i + (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{u}_i. \end{aligned}$$

OLS estimates are clearly unbiased since $\mathrm{E}(\mathbf{u}_i) = \mathbf{0}$. Moreover, within the context of each separate equation, the conditions of the Gauss-Markov theorem are satisfied so that *OLS* is *BLU*.

However, considering the system as a whole, the complete covariance matrix (7.3) is non-diagonal so that *OLS* is not fully efficient. The efficient estimator which takes into account the form of the covariance matrix (7.3) is the *GLS* estimator.

7.4 Special cases

There are two special cases where the *GLS* estimator of the *SURE* model (7.5) reduces to the *OLS* estimator. The first is the case where the matrix (7.4) is diagonal so that there is no contemporaneous correlation. As a consequence, (7.5) will also be diagonal.

The second case is where the regressor matrices $\mathbf{X}_i$ are the same for all equations $i = 1, \cdots, m$. In this case

$$\begin{aligned}
\mathbf{X}^* &= \begin{bmatrix} \mathbf{X} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X} \end{bmatrix} \\
&= \mathbf{I}_m \otimes \mathbf{X}
\end{aligned}$$

so that, using the rules for manipulating Kronecker products, it can be shown that

$$\begin{aligned}
& \mathbf{X}^{*\prime}\boldsymbol{\Sigma}^{*-1}\mathbf{X}^* \\
= \; & (\mathbf{I}_m \otimes \mathbf{X})'(\boldsymbol{\Sigma} \otimes \mathbf{I}_n)^{-1}(\mathbf{I}_m \otimes \mathbf{X}) \\
= \; & (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}'\mathbf{X})
\end{aligned}$$

and

$$\begin{aligned}
& \mathbf{X}^{*\prime}\boldsymbol{\Sigma}^{*-1}\mathbf{y}^* \\
= \; & (\mathbf{I}_m \otimes \mathbf{X})'(\boldsymbol{\Sigma} \otimes \mathbf{I}_n)^{-1}\mathbf{y}^* \\
= \; & (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}')\mathbf{y}^*
\end{aligned}$$

so that the *GLS* estimator (7.5) becomes

$$\begin{aligned}
\boldsymbol{\beta}^* &= (\mathbf{X}^{*\prime}\boldsymbol{\Sigma}^{*-1}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\boldsymbol{\Sigma}^{*-1}\mathbf{y}^* \\
&= (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}'\mathbf{X})^{-1}(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}')\mathbf{y}^* \\
&= (\boldsymbol{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1})(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}')\mathbf{y}^* \\
&= (\mathbf{I}_m \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}^*
\end{aligned}$$

but the final line can be written as

$$\begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_1 \\ \vdots \\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_m \end{bmatrix} = \begin{bmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \vdots \\ \widehat{\boldsymbol{\beta}}_m \end{bmatrix}$$

which is just the stacked vector of *OLS* estimators for each equation.