Regression

Contents:

1. GENERAL LINEAR MODEL (GLM)
2. MULTIVARIATE LINEAR REGRESSION MODEL
3. SEEMINGLY UNRELATED REGRESSION (SUR) MODELS

The General Linear Model (GLM)

Consider the regression:

$$y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_n X_{in} + \varepsilon_i,$$

A basic assumption of the standard linear regression is that $\mathrm{E}(\varepsilon_i) = 0$,

$$\mathrm{Var}(\varepsilon_i) = \mathrm{E}(\varepsilon_i^2) = \sigma^2 \quad \text{and} \quad \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \mathrm{E}(\varepsilon_i \varepsilon_j) = 0.$$

This is described as the assumption of *spherical disturbances*. It implies that disturbance variances is constant at each observations point and that the disturbance covariances at all possible pairs of observation points are zero.

We intend to consider some important cases where there are no spherical disturbances and develop estimation procedures for such models. That is,

$$\mathrm{E}(\varepsilon_i) = 0 \quad \text{and} \quad \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \mathrm{E}(\varepsilon_i \varepsilon_j) = \sigma^2 \omega_{ij}.$$

In this case the regression model is called the General Linear Model (GLM) and is the parent model of econometrics.

Matrix form of the GLM

In matrix form the GLM can be written as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix},$$

or

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2\Omega).$$

That is,

$$\mathrm{E}(\varepsilon) = 0 \quad \text{and} \quad \mathrm{Var}(\varepsilon) = \mathrm{E}(\varepsilon\varepsilon^T) = \sigma^2\Omega.$$

The symmetric and non-singular matrix Ω is given by:

$$\Omega = \begin{pmatrix} \omega_{11} & \omega_{12} & \dots & \omega_{1m} \\ \omega_{21} & \omega_{22} & \dots & \omega_{2m} \\ \vdots & \vdots & & \vdots \\ \omega_{m1} & \omega_{m2} & \dots & \omega_{mm} \end{pmatrix}$$

Ordinary Least-Squares (OLS) Estimation

The OLS (ordinary least squares) estimator of the GLM ignores that the disturbances are correlated. That is, it assumes $\Omega = I_m$, or $\omega_{ii} = 1$ and $\omega_{ij} = 0$ $(i, j = 1, \ldots, m)$.

Thus, the OLS estimator of the GLM is given (as before) by:

$$\beta_o = (X^T X)^{-1} X^T y = \beta + (X^T X)^{-1} X^T \varepsilon.$$

This implies that

- $\mathrm{E}(\beta_o) = \beta$, i.e. the β_o is unbiased estimator.
- $\mathrm{Var}(\beta_o) = \sigma^2 (X^T X)^{-1} X^T \Omega X (X^T X)^{-1}$.

Recall that in the standard regression the variance of the OLS estimator is given by:

$$\mathrm{Var}(\widehat{\beta}) = \sigma^2 (X^T X)^{-1}.$$

This indicates that the analyses made based on the Var$(\widehat{\beta})$ *do not hold for the General Linear Model.*

Aside

Note that

$$\beta_o - \mathrm{E}(\beta_o) = \beta_o - \beta = (X^T X)^{-1} X^T \varepsilon.$$

Thus, the variance of β_o is given by:

$$\begin{aligned}
\mathrm{Var}(\beta_o) &= \mathrm{E}\Big((\beta_o - \mathrm{E}(\beta_o))(\beta_o - \mathrm{E}(\beta_o))^T\Big) \\
&= \mathrm{E}\Big((X^T X)^{-1} X^T \varepsilon\varepsilon^T X (X^T X)^{-1}\Big) \\
&= (X^T X)^{-1} X^T \mathrm{E}(\varepsilon\varepsilon^T) X (X^T X)^{-1} \\
&= (X^T X)^{-1} X^T \sigma^2 \Omega X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1} X^T \Omega X (X^T X)^{-1}.
\end{aligned}$$

Generalized Least-Squares (GLS) Estimation

Since Ω is symmetric then it can be decomposed as

$$\Omega = LL^T.$$

Premultiplying the GLM by L^{-1} it gives:

$$L^{-1}y = L^{-1}X\beta + L^{-1}\varepsilon,$$

or the transformed model:

$$y_* = X_*\beta + \varepsilon_*,$$

where $y_* = L^{-1}y$, $X_* = L^{-1}X$ and $\varepsilon_* = L^{-1}\varepsilon$.

Now, $\mathrm{E}(\varepsilon_*) = \mathrm{E}(L^{-1}\varepsilon) = L^{-1}\mathrm{E}(\varepsilon) = 0$.

Similarly,

$$\begin{aligned}\mathrm{Var}(\varepsilon_*) &= \mathrm{Var}(L^{-1}\varepsilon) = L^{-1}\mathrm{Var}(\varepsilon)(L^{-1})^T \\ &= L^{-1}\sigma^2\Omega(L^{-1})^T \\ &= \sigma^2 L^{-1}LL^T(L^{-1})^T \\ &= \sigma^2(L^{-1}L)(L^{-1}L)^T \\ &= \sigma^2 I_m.\end{aligned}$$

- *This implies that the transformed GLM model has become a standard regression model:*

$$y_* = X_* \beta + \varepsilon_*, \quad \varepsilon_* \sim N(0, \sigma^2 I_m).$$

- That is, the Ordinary Least-Squares (OLS) can be used in the transformed model.

- This estimator is the *Generalized Least-Squares* (GLS) estimator of the original model and will be denoted by β_G.

- The GLS estimator and its variance matrix are given by:

$$\beta_G = (X_*^T X_*)^{-1} X_*^T y_* \quad \text{and} \quad \text{Var}(\beta_G) = \sigma^2 (X_*^T X_*)^{-1}.$$

- Equivalently β_G and $\text{Var}(\beta_G)$ are given by:

$$\beta_G = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y.$$

and

$$\text{Var}(\beta_G) = \sigma^2 (X^T \Omega^{-1} X)^{-1}.$$

- The GLS estimator β_G is the best linear unbiased estimator (BLUE), i.e. the $\text{Var}(\beta_G)$ is smaller than $\text{Var}(\beta_o)$.

Aside

Note that, $X_* = L^{-1}X$ and $\Omega^{-1} = (L^{-1})^T L^{-1}$.

Thus,

$$X_*^T X_* = X^T (L^{-1})^T L^{-1} X = X^T (LL^T)^{-1} X = X^T \Omega^{-1} X$$

and

$$X_*^T y_* = X^T (L^{-1})^T L^{-1} y = X^T (LL^T)^{-1} y = X^T \Omega^{-1} y.$$

Therefore,

$$\beta_G = (X_*^T X_*)^{-1} X_*^T y_* = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y$$

and

$$\mathrm{Var}(\beta_G) = \sigma^2 (X_*^T X_*)^{-1} = (X^T \Omega^{-1} X)^{-1}.$$

The residuals are given by:

$$e_* = y_* - X_* \beta_G = L^{-1}(y - X\beta_G).$$

Thus,

$$\begin{aligned} e_*^T e_* &= (y - X\beta_G)^T (L^{-1})^T L^{-1} (y - X\beta_G) \\ &= (y - X\beta_G)^T \Omega^{-1} (y - X\beta_G). \end{aligned}$$

An unbiased estimator of σ^2 is given by:

$$\boxed{s^2 = \frac{(y - X\beta_G)^T \Omega^{-1} (y - X\beta_G)}{m - n - 1}.}$$

Weighted linear regression

- If Ω is diagonal, i.e. $\Omega = \operatorname{diag}(w_1^2, w_2^2, \ldots, w_m^2)$, or

$$\Omega = \begin{pmatrix} w_1^2 & & \\ & \ddots & \\ & & w_m^2 \end{pmatrix},$$

  then the GLM is called *Weighted Linear Model* (WLM).

- This Implies that the WLM is given by:

$$y = X\beta + \varepsilon,$$

  where $\mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ when $i \neq j$ and

$$\mathrm{Var}(\varepsilon_i) = \sigma_i^2 = \sigma^2 w_i^2.$$

- If $\Omega = LL^T, \quad$ then $\quad L = \operatorname{diag}(w_1, \ldots, w_m)$.

- $\Omega^{-1} = \begin{pmatrix} \frac{1}{w_1^2} & & \\ & \ddots & \\ & & \frac{1}{w_m^2} \end{pmatrix}$ and $L^{-1} = \begin{pmatrix} \frac{1}{w_1} & & \\ & \ddots & \\ & & \frac{1}{w_m} \end{pmatrix}$.

- The transformation of the WLM to a standard regression model is given by premultiplying the regression by L^{-1}.

- That is:

$$L^{-1}y = L^{-1}X\beta + L^{-1}\varepsilon.$$

or

$$y_* = X_*\beta + \varepsilon_*.$$

- This is equivalent to multiplying the *i*th observation of the model by $1/w_i$. I.e.

$$\frac{1}{w_i}y_i = \beta_1\frac{1}{w_i}X_{i1} + \beta_2\frac{1}{w_i}X_{i2} + \cdots + \beta_n\frac{1}{w_i}X_{in} + \frac{1}{w_i}\varepsilon_i.$$

- The estimation of the WLM proceeds as in the case of the GLM. That is, the *Weighted Least-Squares* estimator is given by:

$$\beta_W = (X_*^T X_*)^{-1}X_*^T y_* \quad \text{and} \quad \text{Var}(\beta_W) = \sigma^2(X_*^T X_*)^{-1},$$

or equivalently

$$\beta_W = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}y$$

and

$$\text{Var}(\beta_W) = \sigma^2(X^T\Omega^{-1}X)^{-1}.$$

A Sampling experiment

- A sampling experiment was performed using the model:

$$y = X\beta + \varepsilon = 10.0x_1 + 0.4x_2 + 0.6x_3 + \varepsilon.$$

- The $\varepsilon \sim N(0, \sigma^2\Omega)$, where

$$\begin{aligned} \mathrm{E}(\varepsilon\varepsilon^T) &= \sigma^2\Omega \\ &= \frac{\sigma^2}{1-p^2}\begin{pmatrix} 1 & p & p^2 & \dots & p^{T-1} \\ p & 1 & p & \dots & p^{T-2} \\ p^2 & p & 1 & \dots & p^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p^{T-1} & p^{T-2} & p^{T-3} & \dots & 1 \end{pmatrix}, \end{aligned}$$

  with $p = 0.9$ and $\sigma^2 = 0.0625$.

- This covariance matrix Ω arises when we have a model with first-order autoregressive disturbances. That is,

$$\varepsilon_t = p\varepsilon_{t-1} + u_t, \quad (t = 1, \dots, T)$$

  and

$$\mathrm{E}(u_t) = 0, \quad \mathrm{E}(u_t^2) = \sigma^2, \quad \text{and} \quad \mathrm{E}(u_t u_s) = 0 \quad (t \neq s).$$

- The 20×3 matrix of explanatory variables used is:

$$X = \begin{pmatrix} 1 & 0.693 & 0.693 \\ 1 & 1.733 & 0.693 \\ 1 & 0.693 & 1.386 \\ 1 & 1.733 & 1.386 \\ 1 & 0.693 & 1.792 \\ 1 & 2.340 & 0.693 \\ 1 & 1.733 & 1.792 \\ \vdots & \vdots & \vdots \\ 1 & 1.733 & 1.386 \\ 1 & 0.693 & 1.693 \end{pmatrix}$$

- This provides sufficient information to calculate the true covariance matrices of the ordinary and generalized least squares estimators:

$$\beta^o = (X^T X)^{-1} X^T y \quad \text{and} \quad \beta^G = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y.$$

- The variance-covariance matrices of β_o and β_G are given by:

$$\Sigma_{\beta^o} = \sigma^2 (X^T X)^{-1} X^T \Omega X (X^T X)^{-1} \qquad \Sigma_{\beta^G} = \sigma^2 (X^T \Omega^{-1} X)^{-1}$$

$$= \begin{pmatrix} 0.239 & -0.027 & -0.029 \\ -0.027 & 0.017 & 0.010 \\ -0.029 & 0.010 & 0.019 \end{pmatrix} \qquad = \begin{pmatrix} 0.184 & -0.008 & -0.011 \\ -0.008 & 0.005 & 0.002 \\ -0.011 & 0.002 & 0.019 \end{pmatrix}$$

- The model was used to generate 500 samples of y, each having 20 elements. The table below presents β_3^G, β_3^o, $\mathrm{Var}(\beta_3^G)$ and $\mathrm{Var}(\beta_3^o)$ for 10 of these samples.

# Sample	β_3^G	β_3^o	$\mathrm{Var}(\beta_3^G)$	$\mathrm{Var}(\beta_3^o)$
1	0.644	0.731	0.0114	0.0460
2	0.589	0.745	0.0146	0.0642
3	0.641	0.557	0.0092	0.0603
4	0.534	0.497	0.0124	0.0956
5	0.749	0.753	0.0036	0.0092
6	0.584	0.660	0.0061	0.0140
7	0.329	0.293	0.0080	0.0268
8	0.685	0.646	0.0082	0.0531
9	0.634	0.707	0.0095	0.0264
10	0.382	0.450	0.0081	0.0139

- The range of β_3^o is from 0.329 to 0.749 and is more variable than that of β_3^G which is from 0.293 to 0.753.

- The β_3^G is better than β_3^o in terms of repeated samples. However, it is possible that β_3^o is closer to the true value of 0.6. For example in the sample number 8.

- The averages over 500 samples are $\overline{\beta}_3^o = \overline{\beta}_3^G = 0.596$.

- The estimates of $\mathrm{Var}(\beta_3^G)$ give us a reasonable indication of the true sampling variability in β_3^G, but the same is not true for β_3^o.

  This is supported by the averages from the 500 samples which are given by:

$$\overline{\mathrm{Var}(\beta_3^G)} = 0.0085 \quad \text{and} \quad \overline{\mathrm{Var}(\beta_3^o)} = 0.0389.$$

Summary: General Linear Model

- Consider the General Linear Model:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2\Omega).$$

- The ordinary and generalized least squares estimates of the GLM and their corresponding variance-covariances matrices are given by:

$$\beta_o = (X^T X)^{-1} X^T y \text{ with } \mathrm{Var}(\beta_o) = \sigma^2 (X^T X)^{-1} X^T \Omega X (X^T X)^{-1}.$$

  and

$$\beta_G = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y \text{ with } \mathrm{Var}(\beta_G) = \sigma^2 (X^T \Omega^{-1} X)^{-1}.$$

- Both β_o and β_G are unbiased estimators.

- The GLS estimator β_G has smaller variance than the OLS estimator β_o.

- Thus, β_G is the BLUE (best linear unbiased estimator) of the GLM.

Multivariate Multiple Regression

- Multivariate multiple regression is a logical extension of the multiple regression concept to allow for multiple response (dependent) variables.

- Multivariate regression estimates the same coefficients and standard errors as one would obtain using separate OLS regressions.

- In addition, multivariate regression, being a joint estimator, also estimates the between-equation covariances. This means that it is possible to test coefficient across equations.

- The multivariate linear regression model is the relationship between G responses $y_1, \ldots, y_G$ and a single set of variables $x_1, \ldots, x_n$.

  For example $x_1, \ldots, x_n$ are specific characteristics of students (IQ, age, etc.) and y_i is the grade of the student in the *i*th course.

  In marketing research the variables $x_1, \ldots, x_n$ are various characteristics of consumers (e.g. salary, education, family status, etc.) and and $y_1, \ldots, y_G$ denote their buying behavior in various products.

The multivariate multiple linear regression model between G responses $y_1,\ldots,y_G$ and a single set of variables $x_1,\ldots,x_n$ can be written as:

$$y_{tg} = \beta_{g1}X_{t1} + \beta_{g2}X_{t2} + \cdots + \beta_{gn}X_{tn} + \varepsilon_{tg},$$

where $i = 1,\ldots T$ and $g = 1,\ldots,G$. The number of observation of each regression is T.

Thus, the multivariate model can be written as:

$$\begin{pmatrix} y_{11} & y_{12} & \ldots & y_{1G} \\ y_{21} & y_{22} & \ldots & y_{2G} \\ \vdots & \vdots & \vdots & \vdots \\ y_{T1} & y_{T2} & \ldots & y_{TG} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{T1} & x_{T2} & \cdots & x_{Tn} \end{pmatrix} \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1G} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2G} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{n1} & \beta_{n2} & \cdots & \beta_{nG} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1G} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2G} \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon_{T1} & \varepsilon_{T2} & \cdots & \varepsilon_{TG} \end{pmatrix}$$

or

$$(y_1\ y_2\ \cdots\ y_G) = X\,(\beta_1\ \beta_2\ \cdots\ \beta_G) + (\varepsilon_1\ \varepsilon_2\ \cdots\ \varepsilon_G)$$

or

$$Y = XB + E.$$

It is assumed that

$$\mathrm{E}(\varepsilon_{ti}) = 0 \quad \text{and} \quad \mathrm{E}(\varepsilon_{ti}\varepsilon_{tj}) = \sigma_{ij} \quad \text{for} \quad i, j = 1,\ldots,T.$$

That is, the disturbances are contemporaneously correlated.

The model can also be written as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_G \end{pmatrix} = \begin{pmatrix} X & & & \\ & X & & \\ & & \ddots & \\ & & & X \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_G \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_G \end{pmatrix}$$

or in compact form

$$y = Xb + \varepsilon.$$

Note that the variance-covariance matrix of ε is given by:

$$\mathrm{Var}(\varepsilon) = \Sigma \otimes I_T = \begin{pmatrix} \sigma_{11}I_T & \sigma_{12}I_T & \dots & \sigma_{1G}I_T \\ \sigma_{21}I_T & \sigma_{22}I_T & \dots & \sigma_{2G}I_T \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{G1}I_T & \sigma_{G2}I_T & \dots & \sigma_{GG}I_T \end{pmatrix}.$$

The GLS and OLS estimators of the multivariate linear regression model are the same and given by:

$$\hat{b} = (X^T X)^{-1} X^T y$$

or in matrix form

$$\hat{B} = (X^T X)^{-1} X^T Y$$

Thus, *the ordinary Least-Squares estimator of the multivariate multiple regression model is the BLUE.*

Seemingly unrelated Regression (SUR) models

The SUR model is a special case of the GLM and has widely applicability in economics and finance. The SUR models frequently encountered in empirical work. It comprises a set of regression equations and assumes that contemporaneous disturbances are correlated. The regressions are linked statistically, even though not structurally, through the non-spherical covariance matrix of the disturbances.

The expression Seemingly Unrelated Regression is used to reflect the fact that the individual equations are in fact related to one another, even though superficially they may not seem to be.

Examples

- Empirical analysis in finance (cross-section);
- Insurance;
- Marketing;

Specifically, the SUR model comprising G regressions is given by:

$$y_i = X_i\beta_i + \varepsilon_i, \quad i = 1, \ldots, G,$$

For the ith equation we have

- y_i is the response m-element vector;
- X_i is the $m \times k_i$ exogenous matrix;
- β_i is the k_i-element vector of coefficients;
- ε_i is the m-element vector of errors.
- $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{E}(\varepsilon_i\varepsilon_i^T) = \sigma_{ii}^2 I_m$.

If the tth element of ε_i is denoted by ε_{tj}, then

$$\mathrm{E}(\varepsilon_{ti}\varepsilon_{tj}) = \sigma_{ij}.$$

Thats is *Contemporaneous disturbances are correlated*.

Consider the correlation of the error vectors ε_i and ε_j. That is, $\mathrm{E}(\varepsilon_i\varepsilon_j^T)$.

$$
\begin{aligned}
\mathrm{E}(\varepsilon_i\varepsilon_j^T) &= \mathrm{E}\left(\begin{pmatrix}\varepsilon_{1i}\\ \varepsilon_{2i}\\ \vdots\\ \varepsilon_{Gi}\end{pmatrix}\begin{pmatrix}\varepsilon_{1j} & \varepsilon_{2j} & \ldots & \varepsilon_{Gj}\end{pmatrix}\right)\\
&= \mathrm{E}\begin{pmatrix}\varepsilon_{1i}\varepsilon_{1j} & \varepsilon_{1i}\varepsilon_{2j} & \ldots & \varepsilon_{1i}\varepsilon_{mj}\\ \varepsilon_{2i}\varepsilon_{1j} & \varepsilon_{2i}\varepsilon_{2j} & \ldots & \varepsilon_{2i}\varepsilon_{mj}\\ \vdots & \vdots & \ddots & \vdots\\ \varepsilon_{mi}\varepsilon_{1j} & \varepsilon_{mi}\varepsilon_{2j} & \ldots & \varepsilon_{mi}\varepsilon_{mj}\end{pmatrix}\\
&= \begin{pmatrix}\sigma_{ij} & 0 & \ldots & 0\\ 0 & \sigma_{ij} & \ldots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \ldots & \sigma_{ij}\end{pmatrix}\\
&= \sigma_{ij}I_m.
\end{aligned}
$$

The SUR model

$$y_i = X_i\beta_i + \varepsilon_i, \quad i = 1, \ldots, G,$$

can be written as a single model:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_G \end{pmatrix} = \begin{pmatrix} X_1 & & & \\ & X_2 & & \\ & & \ddots & \\ & & & X_G \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_G \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_G \end{pmatrix}$$

or, in compact form

$$y = X\beta + \varepsilon,$$

where

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_G \end{pmatrix}, \; X = \begin{pmatrix} X_1 & & \\ & \ddots & \\ & & X_G \end{pmatrix}, \; \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_G \end{pmatrix} \text{ and } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_G \end{pmatrix}.$$

Note that:

- y is the Gm-element vector of responses,
- X is the block-diagonal comprising all exogenous data,
- β is the collection of all coefficients, and
- ε is the Gm-element of errors.

Now, $\mathrm{E}(\varepsilon) = 0$ and

$$\begin{aligned}
\mathrm{E}(\varepsilon\varepsilon^T) &= \mathrm{E}\left(\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_G\end{pmatrix}\begin{pmatrix}\varepsilon_1^T & \varepsilon_2^T & \dots & \varepsilon_G^T\end{pmatrix}\right)\\
&= \mathrm{E}\begin{pmatrix}\varepsilon_1\varepsilon_1^T & \varepsilon_1\varepsilon_2^T & \dots & \varepsilon_1\varepsilon_G^T\\ \varepsilon_2\varepsilon_1^T & \varepsilon_2\varepsilon_2^T & \dots & \varepsilon_2\varepsilon_G^T\\ \vdots & \vdots & \ddots & \vdots\\ \varepsilon_G\varepsilon_1^T & \varepsilon_G\varepsilon_2^T & \dots & \varepsilon_G\varepsilon_G^T\end{pmatrix}\\
&= \begin{pmatrix}\sigma_{11}I_m & \sigma_{12}I_m & \dots & \sigma_{1G}I_m\\ \sigma_{21}I_m & \sigma_{22}I_m & \dots & \sigma_{2G}I_m\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{G1}I_m & \sigma_{G2}I_m & \dots & \sigma_{GG}I_m\end{pmatrix}\\
&= \Sigma \otimes I_m\\
&= \Omega,
\end{aligned}$$

where

$$\Sigma = \begin{pmatrix}\sigma_{11} & \sigma_{12} & \dots & \sigma_{1G}\\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2G}\\ \dots & \dots & \dots & \dots\\ \sigma_{G1} & \sigma_{G2} & \dots & \sigma_{GG}\end{pmatrix}.$$

Thus, the SUR model can be written as:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim (0, (\Sigma \otimes I_m)).$$

This is a special case of the GLM, where $\Omega = (\Sigma \otimes I_m)$. In this case:

- $\Omega^{-1} = (\Sigma \otimes I_m)^{-1} = (\Sigma^{-1} \otimes I_m)$.
- If $\Sigma = LL^T$, then $(\Sigma \otimes I_m) = (LL^T \otimes I_m) = (L \otimes I_m)(L^T \otimes I_m)$
- $\Omega^{-1} = (L^{-T} \otimes I_m)(L^{-1} \otimes I_m)$.

Consider the transformed SUR model:

$$(L^{-1} \otimes I_m)y = (L^{-1} \otimes I_m)X\beta + (L^{-1} \otimes I_m)\varepsilon,$$

or $\quad y_* = X_*\beta + \varepsilon_*, \quad$ where now

$$\begin{aligned}
\mathrm{Var}(\varepsilon_*) &= \mathrm{Var}\big((L^{-1} \otimes I_m)\varepsilon\big) \\
&= (L^{-1} \otimes I_m)\mathrm{Var}(\varepsilon)(L^{-T} \otimes I_m) \quad \text{since } \mathrm{Var}(A\varepsilon) = A\mathrm{Var}(\varepsilon)A^T \\
&= (L^{-1} \otimes I_m)(\Sigma \otimes I_m)(L^{-T} \otimes I_m) \quad \text{since } \mathrm{Var}(\varepsilon) = \Sigma \otimes I_m \\
&= (L^{-1}\Sigma L^{-T} \otimes I_m) \quad \text{since } (A \otimes B)(C \otimes D) = (AC \otimes BD) \\
&= (L^{-1}LL^TL^{-T} \otimes I_m) \quad \text{since } \Sigma = LL^T \\
&= (I_G \otimes I_m) \quad \text{since } L^{-1}L = I_G \text{ and } L^TL^{-T} = I_G \\
&= I_{Gm}.
\end{aligned}$$

The OLS estimator of the transformed model and its variance matrix are given, respectively, by:

$$\beta_G = (X_*^T X_*)^{-1} X_*^T y_* \quad \text{and} \quad \mathrm{Var}(\beta_G) = (X_*^T X_*)^{-1}.$$

Thus, as in the case of the general linear model[a], the generalized least squares estimator of β is given by:

$$\boxed{\beta_G = (X^T(\Sigma^{-1} \otimes I_m)X)^{-1} X^T(\Sigma^{-1} \otimes I_m)y.}$$

The Variance of β_G is given by:

$$\boxed{\mathrm{Var}(\beta_G) = (X^T(\Sigma^{-1} \otimes I_m)X)^{-1}.}$$

In most cases Σ is unknown and needs to be estimated by:

$$\widehat{\Sigma} = \begin{pmatrix} \widehat{\sigma}_{11} & \cdots & \widehat{\sigma}_{1G} \\ \widehat{\sigma}_{21} & \cdots & \widehat{\sigma}_{2G} \\ \cdots & \cdots & \cdots \\ \widehat{\sigma}_{G1} & \cdots & \widehat{\sigma}_{GG} \end{pmatrix}$$

In this case the *Feasible generalizes least squares estimator* of the SUR model and its variance matrix are given, respectively, by:

$$\boxed{\beta_F = (X^T(\widehat{\Sigma}^{-1} \otimes I_m)X)^{-1} X^T(\widehat{\Sigma}^{-1} \otimes I_m)y.}$$

and

$$\boxed{\mathrm{Var}(\beta_F) = (X^T(\widehat{\Sigma}^{-1} \otimes I_m)X)^{-1}.}$$

---

[a] Now $\Omega = \Sigma \otimes I_m$ and thus, $\Omega^{-1} = \Sigma^{-1} \otimes I_m$.

Estimating Σ

Two methods are used to estimate Σ. The first method ignores the restrictions on the coefficients of the SUR model. Let Z denote the $m \times k$ exogenous matrix corresponding to all distinct regressors in the SUR model, where k is the total number of these distinct regressors. It derives the OLS estimates of

$$y_i = Z\widetilde{\beta}_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_{ii} I_m).$$

This gives the *unrestricted* residuals $\widetilde{e}_i = M_z y_i$, where $M_z = I_m - Z(Z^T Z)Z^T$. An unbiased estimator of σ_{ij} is obtained by:

$$\boxed{\widehat{\sigma}_{ij} = \frac{\widetilde{e}_i^T \widetilde{e}_j}{m} = \frac{y_i^T M_z y_i}{m}.}$$

The second method constructs $\widehat{\Sigma}$ by applying the OLS to each regression

$$y_i = X_i \beta_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_{ii} I_m)$$

and obtains the *restricted* residuals $\widehat{e}_i = M_i y_i$, where $M_i = I_m - X_i(X_i^T X_i)^{-1} X_i^T$. In this case an unbiased estimator of σ_{ij} is given by:

$$\boxed{\widehat{\sigma}_{ij} = \frac{\widehat{e}_i^T \widehat{e}_i}{m} = \frac{y_i^T M_i M_j y_i}{m}.}$$

Iterative feasible generalized least squares estimator

Consider the FGLS of the SUR model:

$$\beta_F = (X^T(\widehat{\Sigma}^{-1} \otimes I_m)X)^{-1}X^T(\widehat{\Sigma}^{-1} \otimes I_m)y. \qquad (1)$$

Suppose that $\widehat{\Sigma} = I_G$ so that the GLS and OLS are the same[b], i.e. $\beta_F^{(0)} = (X^TX)^{-1}X^Ty$. Base on the residuals of the OLS estimator Σ is estimated by $\widehat{\Sigma}^{(1)}$. The estimator $\widehat{\Sigma}^{(1)}$ of Σ can now be used to replace $\widehat{\Sigma}$ in (1) to provide another estimator of β_F and Σ, say $\beta_F^{(1)}$ and $\widehat{\Sigma}^{(2)}$. Repeating this process leads to an *Iterative* FGLS (IFGLS) estimator of β.

In general, the *p*th IFGLS estimator of β is given by:

$$\beta_F^{(p)} = (X^T((\widehat{\Sigma}^{(p)})^{-1} \otimes I_m)X)^{-1}X^T((\widehat{\Sigma}^{(p)})^{-1} \otimes I_m)y,$$

where $\widehat{\Sigma}^{(p)}$ is the estimator of Σ obtained by the residual of $\beta^{(p-1)}$ with $\widehat{\Sigma}^{(0)} = I_G$.

The iterative process continues until convergence, i.e. until

$$\widehat{\Sigma}^{(p)} = \widehat{\Sigma}^{(p+1)} \quad \text{and} \quad \beta_F^{(p)} = \beta_F^{(p+1)}.$$

---

[b] since $(\widehat{\Sigma}_{(0)} \otimes I_m) = (I_G \otimes I_m) = I_{Gm}$.

Summary for deriving the IFGLS of SUR models

Consider the SUR model:

$$y = X\beta + \varepsilon, \tag{2}$$

where $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{E}(\varepsilon\varepsilon^T) = \Sigma \otimes I_m$.

The Iterative feasible generalized least squares estimator $\beta_F^{(p)}$ is derived as follows:

1. Derive the OLS estimator: $\beta_F^{(0)} = (X^T X)^{-1} X^T y$.
2. For $p = 1, 2, \ldots$
   - (a) Calculate the residuals of the SUR model using the estimator $\beta_F^{(p-1)}$.
   - (b) Calculate the estimator $\widehat{\Sigma}^{(p)}$ using the restricted residuals.
   - (c) Derive the FGLS:

     $$\beta_F^{(p)} = (X^T((\widehat{\Sigma}^{(p)})^{-1} \otimes I_m)X)^{-1} X^T((\widehat{\Sigma}^{(p)})^{-1} \otimes I_m)y$$
   - (d) STOP if $\widehat{\Sigma}^{(p)} = \widehat{\Sigma}^{(p-1)}$ and $\beta_F^{(p)} = \beta_F^{(p-1)}$.