Econ 620

Why GLS?

- Recall the assumptions of the classical multiple regression model - especially the assumption on the distribution of the disturbance terms;

$$y = X\beta + \varepsilon \tag{1}$$

$$E(\varepsilon) = \mathbf{0} \qquad E(\varepsilon\varepsilon') = \sigma^2 I \tag{2}$$

  The zero mean assumption is not so severe that we can easily accommodate the non-zero mean by defining the constant term differently. However, the assumption on the second moment matrix of the disturbance terms are very restrictive; the homoskedasticity & uncorrelatedness assumption (or, indeed, sometimes the stronger i.i.d. assumption) represented by (2) is too stringent to be applied to most economic data.

- Alternative specification of the error term is given by;

$$E(\varepsilon) = \mathbf{0} \qquad E(\varepsilon\varepsilon') = V \tag{3}$$

  where V is an arbitrary positive definite symmetric matrix. The specification can nest both heteroskedasticity and serial correlation in disturbance terms. To see the argument in detail, consider the explicit form of the matrix V;

$$V = \begin{bmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \cdots & E(\varepsilon_1\varepsilon_{N-1}) & E(\varepsilon_1\varepsilon_N) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & \cdots & E(\varepsilon_2\varepsilon_{N-1}) & E(\varepsilon_2\varepsilon_N) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ E(\varepsilon_{N-1}\varepsilon_1) & E(\varepsilon_{N-1}\varepsilon_2) & \cdots & E(\varepsilon_{N-1}^2) & E(\varepsilon_{N-1}\varepsilon_N) \\ E(\varepsilon_N\varepsilon_1) & E(\varepsilon_N\varepsilon_2) & \cdots & E(\varepsilon_N\varepsilon_{N-1}) & E(\varepsilon_N^2) \end{bmatrix} \tag{4}$$

- What is the consequence of the OLS estimation with error structure (3)?
  - It is still unbiased;

$$\widehat{\beta}_{OLS} = (X'X)^{-1} X'y = \beta + (X'X)^{-1} X'\varepsilon$$
$$E\left(\widehat{\beta}_{OLS}\right) = \beta + (X'X)^{-1} X'E(\varepsilon) = \beta$$

  - It has different variance matrix;

$$\begin{aligned} Var\left(\widehat{\beta}_{OLS}\right) &= E\left[\left(\widehat{\beta}_{OLS} - \beta\right)\left(\widehat{\beta}_{OLS} - \beta\right)'\right] \\ &= E\left((X'X)^{-1} X'\varepsilon\varepsilon'X(X'X)^{-1}\right) = (X'X)^{-1} X'VX(X'X)^{-1} \end{aligned} \tag{5}$$

    Note that under classical assumptions; $Var\left(\widehat{\beta}_{OLS}\right) = \sigma^2 (X'X)^{-1}$.
  - It is not BLUE. - immediate consequence of Gauss-Markov theorem.
  - Since we have different variance formula as in (5), the usual t-test and $F-$test statistics are **invalid**.
  - It is **still consistent** as long as $\text{plim}\frac{X'X}{N} = Q$ and $\text{plim}\frac{X'\varepsilon}{N} = \mathbf{0}$;

$$\begin{aligned} \text{plim}\widehat{\beta}_{OLS} &= \text{plim}\left[\beta + (X'X)^{-1} X'\varepsilon\right] \\ &= \beta + \left(\text{plim}\frac{X'X}{N}\right)^{-1} \text{plim}\frac{X'\varepsilon}{N} = \beta + Q^{-1}\mathbf{0} = \beta \end{aligned} \tag{6}$$

  - The asymptotic variance matrix is different from what we used to have in classical cases. The asymptotic distribution of $\widehat{\beta}_{OLS}$ is now given by;

$$\sqrt{N}\left(\widehat{\beta}_{OLS} - \beta\right) \xrightarrow{d} N\left(\mathbf{0}, \left(\frac{X'X}{N}\right)^{-1}\left(\frac{X'VX}{N}\right)\left(\frac{X'X}{N}\right)^{-1}\right) \tag{7}$$

    as long as the probability limits of three arguments of the asymptotic variance matrix exist.

Now, what to do?

- First of all, we will reparameterize the matrix V in slightly different way;

$$V = \sigma^2 \Omega$$

we lose no generality in this reparameterization. But the reparameterization will deliver a convenient comparison between OLS and GLS.

- Suppose that we know the complete structure of Ω, which, of course, is highly unlikely. Anyway, then we can always find a decomposition of Ω^{-1} such that

$$L'L = \Omega^{-1} \tag{8}$$

where L is an $(N \times N)$ non-singular matrix.

- Multiplying both sides of (1) with L, we have;

$$Ly = LX\beta + L\varepsilon \tag{9}$$

We can treat Ly as dependent variable, LX as independent variables, and $L\varepsilon$ as error terms. Then,

$$\begin{aligned} E\left(L\varepsilon\right) &= LE\left(\varepsilon\right) = 0 \\ Var\left(L\varepsilon\right) &= LVar\left(\varepsilon\right)L' = LVL' = L\sigma^2\Omega L' = \sigma^2 L\left(L'L\right)^{-1}L' = \sigma^2 I \end{aligned} \tag{10}$$

Note that the error terms now satisfies the assumptions of the classical regression model;

- Regressing Ly on LX gives;

$$\begin{aligned} \widehat{\beta}_{GLS} &= \left[\left(LX\right)'\left(LX\right)\right]^{-1}\left(LX\right)'Ly \\ &= \left[X'L'LX\right]^{-1}\left(X'L'Ly\right) = \left(X'\Omega^{-1}X\right)^{-1}\left(X'\Omega^{-1}y\right) \qquad (11) \\ &= \left(X'\left(\sigma^2\Omega\right)^{-1}X\right)^{-1}\left(X'\left(\sigma^2\Omega\right)^{-1}y\right) = \left(X'V^{-1}X\right)^{-1}\left(X'V^{-1}y\right) \qquad (12) \end{aligned}$$

- Let's check the characteristics of GLS estimator;

$$\begin{aligned} \widehat{\beta}_{GLS} &= \left(X'V^{-1}X\right)^{-1}\left(X'V^{-1}y\right) = \left(X'V^{-1}X\right)^{-1}X'V^{-1}\left[X\beta+\varepsilon\right] \\ &= \beta + \left(X'V^{-1}X\right)^{-1}X'V^{-1}\varepsilon \end{aligned}$$

Hence,

  - It is unbiased;

$$E\left(\widehat{\beta}_{GLS}\right) = \beta + \left(X'V^{-1}X\right)^{-1}X'V^{-1}E\left(\varepsilon\right) = \beta \tag{13}$$

  - Its variance is given by;

$$\begin{aligned} Var\left(\widehat{\beta}_{GLS}\right) &= E\left[\left(\widehat{\beta}_{GLS} - E\left(\widehat{\beta}_{GLS}\right)\right)\left(\widehat{\beta}_{GLS} - E\left(\widehat{\beta}_{GLS}\right)\right)'\right] \\ &= E\left[\left(\widehat{\beta}_{GLS} - \beta\right)\left(\widehat{\beta}_{GLS} - \beta\right)'\right] \\ &= E\left[\left(X'V^{-1}X\right)^{-1}X'V^{-1}\varepsilon\varepsilon'V^{-1}X\left(X'V^{-1}X\right)^{-1}\right] \\ &= \left(X'V^{-1}X\right)^{-1}X'V^{-1}E\left(\varepsilon\varepsilon'\right)V^{-1}X\left(X'V^{-1}X\right)^{-1} \\ &= \left(X'V^{-1}X\right)^{-1}X'V^{-1}VV^{-1}X\left(X'V^{-1}X\right)^{-1} = \left(X'V^{-1}X\right)^{-1} \\ &= \sigma^2\left(X'\Omega^{-1}X\right)^{-1} \end{aligned} \tag{14}$$

  - It is BLUE;
  - It is consistent under the usual conditions; the crucial condition is again plim$\frac{X'\Omega^{-1}\varepsilon}{N} = 0$;
  - Asymptotic distribution is given by;

$$\sqrt{N}\left(\widehat{\beta}_{GLS} - \beta\right) \xrightarrow{d} N\left(0, \sigma^2\left(\frac{X'\Omega^{-1}X}{N}\right)^{-1}\right) \tag{15}$$

Feasible Generalized Least Squares (FGLS)

- The theory for GLS is nice. How useful is it? The answer is that it is virtually useless. The truth is that we don't know V or at least Ω. Then, what are we supposed to do? One universally true maxim in econometrics is that when you have something you don't know, estimate it!. There are a lot of way to estimate Ω depending on the model we consider. For the moment, just assume that we have a consistent estimator $\widehat{\Omega}$ of Ω. We can replace Ω with $\widehat{\Omega}$ in our procedure. The procedure is naturally called FGLS. We can derive the asymptotic distribution of FGLS estimator under some conditions.

- Suppose that
$$\text{plim}\frac{X'\widehat{\Omega}^{-1}X}{N} = Q \text{ where } Q \text{ is positive definite and finite}$$
$$\text{plim}\frac{X'\widehat{\Omega}^{-1}\varepsilon}{N} = \mathbf{0}$$
then, $\widehat{\beta}_{FGLS} = \left(X'\widehat{\Omega}^{-1}X\right)^{-1} X'\widehat{\Omega}^{-1}y$ is consistent. - prove it.

- Suppose that
$$\text{plim}\frac{X'\left(\widehat{\Omega}^{-1} - \Omega^{-1}\right)X}{N} = \mathbf{0}$$
$$\text{plim}\frac{X'\left(\widehat{\Omega}^{-1} - \Omega^{-1}\right)\varepsilon}{N} = \mathbf{0}$$
then,
$$\sqrt{N}\left(\widehat{\beta}_{FGLS} - \beta\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma^2\left(\frac{X'\Omega^{-1}X}{N}\right)^{-1}\right) \tag{16}$$
The proof is in the lecture note and you have to redo the exercise with your own pencil and paper. The above conditions are sufficient and they are satisfied when
$$\widehat{\Omega} \xrightarrow{p} \Omega$$

Examples

- Grouping of the observations; In some cases, statistical sources group observations and publish only average values for each group in order mainly to protect the identity of the survey subjects. However, most economic models are usually based on individual decision making. How can we solve the problem? Surely, we cannot solve the whole problem, but there is a lot better way to analyze the data set than simple OLS with grouped data. Suppose the "true" model is
$$y = X\beta + \varepsilon$$
$$E(\varepsilon) = 0 \qquad E(\varepsilon\varepsilon') = \sigma^2 I$$
But, we have G group-averaged observations on $\left(\widetilde{y}_i, \widetilde{X}_i\right)$ where $i = 1, 2, \cdots, G$. Suppose that we have n_i individuals in each group so that $n_1 + n_2 + \cdots + n_G = N$. Due to the data requirement, we have to consider the model;
$$\widetilde{y} = \widetilde{X}\beta + \widetilde{\varepsilon}$$

Clearly, we can infer that

$$E\left(\widetilde{\varepsilon}\right) = E\left[\begin{array}{c} \widetilde{\varepsilon}_1 \\ \widetilde{\varepsilon}_2 \\ \cdots \\ \widetilde{\varepsilon}_G \end{array}\right] = \mathbf{0}$$

$$\begin{aligned} Var\left(\widetilde{\varepsilon}\widetilde{\varepsilon}'\right) &= E\left[\begin{array}{cccc} \widetilde{\varepsilon}_1^2 & \widetilde{\varepsilon}_1\widetilde{\varepsilon}_2 & \cdots & \widetilde{\varepsilon}_1\widetilde{\varepsilon}_G \\ \widetilde{\varepsilon}_1\widetilde{\varepsilon}_2 & \widetilde{\varepsilon}_2^2 & \cdots & \widetilde{\varepsilon}_2\widetilde{\varepsilon}_G \\ \cdots & \cdots & \cdots & \cdots \\ \widetilde{\varepsilon}_1\widetilde{\varepsilon}_G & \widetilde{\varepsilon}_2\widetilde{\varepsilon}_G & \cdots & \widetilde{\varepsilon}_G^2 \end{array}\right] = \left[\begin{array}{cccc} \frac{\sigma^2}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{\sigma^2}{n_2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{\sigma^2}{n_G} \end{array}\right] \\ &= \sigma^2\left[\begin{array}{cccc} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{1}{n_G} \end{array}\right] \end{aligned}$$

If we know the number of individuals in each group, which is usually available, we can construct $\sigma^2\Omega$. We know exact structure of Ω. The L matrix in this case is;

$$L = \left[\begin{array}{cccc} \sqrt{n_1} & 0 & \cdots & 0 \\ 0 & \sqrt{n_2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sqrt{n_G} \end{array}\right]$$

- It is sometimes not reasonable to assume that the type of heteroskadasticity depends on one or a combination of independent variables. Suppose that, for simplicity, the pattern of heteroskadasticity is determined by j's independent variable. Then;

  $$y = X\beta + \varepsilon \qquad E\left(\varepsilon\right) = 0$$

  and;

  $$Var\left(\varepsilon\right) = \sigma^2\left[\begin{array}{cccc} x_{1j}^2 & 0 & \cdots & 0 \\ 0 & x_{2j}^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & x_{Nj}^2 \end{array}\right]$$

  where x_{ij} is the i^{th} observation on the j^{th} independent variable. Then, the GLS estimate is obtained from;

  $$\begin{aligned} \frac{y_i}{x_{ij}} &= \beta_j + \beta_1\left(\frac{1}{x_{ij}}\right) + \beta_2\left(\frac{x_{i2}}{x_{ij}}\right) + \cdots \\ &+ \beta_{j-1}\left(\frac{x_{ij-1}}{x_{ij}}\right) + \beta_{j+1}\left(\frac{x_{ij+1}}{x_{ij}}\right) + \cdots + \beta_2\left(\frac{x_{iK}}{x_{ij}}\right) + \frac{\varepsilon_i}{x_{ij}} \end{aligned}$$

- We can also assume that the pattern of heteroskadasticity is governed by a combination of some variables - which may include independent variables or other variables-. The specification is then;

  $$y_i = \beta' x_i + \varepsilon_i \qquad i = 1, 2, \cdots, N$$

  where β is a $(k \times 1)$ vector of parameters.. We cam specify;

  $$E\left(\varepsilon_i^2\right) = \sigma^2\left(\alpha' z_i\right)^2 \text{ and } E\left(\varepsilon_i\varepsilon_j\ \right) = 0 \text{ when } i \neq j$$

  where z_i is an $(h \times 1)$ vector. We still hold the independence assumption but give up homoskadasticity. In the variance specification, both σ^2 and α are unknown parameters and z_i is the vector of observation on variables $z's$. We can estimate the model using GLS. The problem is that we don't know the parameter α so that we don't know Ω. If we can somehow consistently estimate α, therefore, Ω, we can do FGLS. More appealing approach is MLE. If we assume that

$$\varepsilon_i \sim N\left(0, \sigma^2 \left(\alpha' z_i\right)^2\right)$$

with serial independence. The the log likelihood function is;

$$L\left(\beta, \sigma^2, \alpha\right) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log \sigma^2 - \sum_{i=1}^{N} \alpha' z_i - \frac{1}{2\sigma^2}\sum_{i=1}^{N} \frac{\left(y_i - \beta' x_i\right)^2}{\left(\alpha' z_i\right)^2}$$

we can estimate β, σ^2,and α by differentiating the log-likelihood function. We know that the MLE are consistent and asymptotically efficient. The asymptotic variance matrix is obtained by the inverse of information matrix as usual. Another quite popular specification is that

$$\varepsilon_i \sim N\left(0, \sigma^2 \exp\left(\alpha' z_i\right)\right)$$

- We now turn to the example where we keep the homoskadasticity assumption but weaken dependence structure of error terms. If we allow some correlations in error terms, our variance matrix of error terms is not a diagonal matrix anymore. Do you see why? Look at the matrix (4). One of the most popular specification of disturbance terms with serial dependence is AR(1) model;

$$\begin{aligned} y_t &= \beta' x_t + u_t \\ u_t &= \rho u_{t-1} + \varepsilon_t \qquad |\rho| < 1 \\ E\left(\varepsilon_t\right) = 0, E\left(\varepsilon_t^2\right) &= \sigma_\varepsilon^2, E\left(\varepsilon_t \varepsilon_s\right) = 0 \text{ when } t \neq s \end{aligned}$$

Under the specification, we know that

$$\begin{aligned} E\left(u_t\right) = 0, Var\left(u_t\right) &= \frac{\sigma_\varepsilon^2}{1-\rho^2} \text{ for all } t = 1, 2, \cdots .T \\ Cov\left(u_t, u_{t+h}\right) &= \frac{\sigma_\varepsilon^2}{1-\rho^2}\rho^h,\ Corr\left(u_t, u_{t-h}\right) = \rho^h \end{aligned}$$

Hence, in vector notation, the variance matrix of error terms are;

$$\begin{aligned} Var\left(uu'\right) &= \frac{\sigma_\varepsilon^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \cdots & \rho^{T-2} & \rho^{T-1} \\ \rho & 1 & \cdots & \rho^{T-3} & \rho^{T-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{T-2} & \rho^{T-3} & \cdots & 1 & \rho \\ \rho^{T-1} & \rho^{T-1} & \cdots & \rho & 1 \end{bmatrix} \\ &= \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho^{T-2} & \rho^{T-1} \\ \rho & 1 & \cdots & \rho^{T-3} & \rho^{T-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{T-2} & \rho^{T-3} & \cdots & 1 & \rho \\ \rho^{T-1} & \rho^{T-1} & \cdots & \rho & 1 \end{bmatrix} = \sigma^2 \Omega \end{aligned}$$

where $\sigma^2 = \frac{\sigma_\varepsilon^2}{1-\rho^2}$. It is know that

$$\Omega^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho & 0 & 0 & 0 & \cdots & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & 0 & \cdots & \cdots & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & 0 & \cdots & \cdots & 0 \\ 0 & 0 & -\rho & 1+\rho^2 & -\rho & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & 0 & \cdots & 0 & -\rho & 1 \end{bmatrix}$$

and

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

The estimation of the model will be discussed later.

Seemingly Unrelated Regression Estimator (SURE)

- Consider a typical utility maximization problem a consumer solves;

$$\max\ U(x)$$
$$s.t.\ p\cdot x\leq w$$

where x is a $(M\times 1)$ vector of quantity demanded and p is the price vector. The solution to the max problem will be given as;

$$\begin{aligned}x_1&=f(p_1,p_2,\cdots,p_M,w)\\ x_2&=f(p_1,p_2,\cdots,p_M,w)\\ &\ldots\ldots\ldots\ldots\ldots..\\ x_M&=f(p_1,p_2,\cdots,p_M,w)\end{aligned}$$

- Econometrically, we would specify the model as;

$$\begin{aligned}x_{1i}&=f(p_{1t},p_{2t},\cdots,p_{Mt},w_t)+\varepsilon_{1i}\\ x_{2i}&=f(p_{1t},p_{2t},\cdots,p_{Mt},w_t)+\varepsilon_{2i}\\ &\ldots\ldots\ldots\ldots\ldots..\\ x_{Mi}&=f(p_{1t},p_{2t},\cdots,p_{Mt},w_t)+\varepsilon_{Ni}\end{aligned}$$

where $i=1,2,\cdots,N$. We may estimate each equation by OLS to get the estimates of parameters. However, we may lose some information doing that. It is highly likely that the demand equations are interdpendent since consumers determine the quantity demanded simultaneously, not separately. In statistical notation, it is natural to assume that

$$E(\varepsilon_{ji}\varepsilon_{li})\neq 0 \text{ when } j\neq l \tag{17}$$

We can achieve some improvement in efficiency by incorporating the information on the inter-equation dependence into estimation procedure. The seemingly unrelated regression estimator will give us the answer to the question of how to do that.

- Suppose that we have M system of equations;

$$\begin{aligned}y_1&=X_1\beta_1+\varepsilon_1\\ y_2&=X_1\beta_1+\varepsilon_2\\ &\cdots\\ y_M&=X_1\beta_1+\varepsilon_M\end{aligned}$$

where y_j is $(N\times 1)$ matrix of observations on the dependent variable of the j^{th} equation, X_j is $(N\times K)$ is $(N\times K_j)$ matrix of observations on the independent variables of the j^{th} equation, and ε_j is $(N\times 1)$ matrix of the disturbances of the j^{th} equation. For the notational simplicity, we will assume that each equation has the same number of regressors, K, i.e. $K_1=K_2=\cdots=K_M=K$. We assume that error terms are independent across observations but dependent across equations;

$$\begin{aligned}E(\varepsilon_{ji}\varepsilon_{hi})&=\sigma_{jh}\text{ for all } i=1,2,\cdots,N\\ E(\varepsilon_{ji}\varepsilon_{jl})&=0\text{ when } i\neq l\\ E(\varepsilon_{jr}\varepsilon_{hs})&=0\text{ when } j\neq h, r\neq s\end{aligned} \tag{18}$$

- We stack the data as;

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_{M-1} \\ y_M \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \cdots & 0 & 0 \\ 0 & X_2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & X_{M-1} & 0 \\ 0 & 0 & \cdots & 0 & X_M \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_{M-1} \\ \beta_M \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_{M-1} \\ \varepsilon_M \end{bmatrix}$$

$$y = X\beta + \varepsilon \tag{19}$$

where y is now $(MN \times 1)$, X is $(MN \times MK)$, β is $(MK \times 1)$, and ε is $(MN \times 1) -$ remember each element in matrix above is either vector or matrix itself. Let' check the structure of the variance matrix of ε.

-

$$Var(\varepsilon) = E(\varepsilon\varepsilon') = E \begin{bmatrix} \varepsilon_1\varepsilon_1' & \varepsilon_1\varepsilon_2' & \cdot\cdot & \varepsilon_1\varepsilon_M' \\ \varepsilon_2\varepsilon_1' & \varepsilon_2\varepsilon_2' & \cdot\cdot & \varepsilon_2\varepsilon_M' \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ \varepsilon_M\varepsilon_1' & \varepsilon_M\varepsilon_2' & \cdot\cdot & \varepsilon_M\varepsilon_M' \end{bmatrix}$$

Remember that $\varepsilon_r\varepsilon_s'$ is $(N \times N)$ matrix and $Var(\varepsilon)$ is $(MN \times MN)$ matrix. Let's check what they are.

$$E[\varepsilon_r\varepsilon_r'] = E\left[\begin{pmatrix} \varepsilon_{r1} & \varepsilon_{r2} & \cdot\cdot & \varepsilon_{rN} \end{pmatrix} \begin{pmatrix} \varepsilon_{r1} \\ \varepsilon_{r2} \\ \cdot\cdot \\ \varepsilon_{rN} \end{pmatrix}\right]$$

$$= E \begin{bmatrix} \varepsilon_{r1}^2 & \varepsilon_{r1}\varepsilon_{r2} & \cdot\cdot & \varepsilon_{r1}\varepsilon_{rN} \\ \varepsilon_{r2}\varepsilon_{r1} & \varepsilon_{r2}^2 & \cdot\cdot & \varepsilon_{r2}\varepsilon_{rN} \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ \varepsilon_{rN}\varepsilon_{r1} & \varepsilon_{rN}\varepsilon_{r2} & & \varepsilon_{rN}^2 \end{bmatrix} = \begin{bmatrix} \sigma_{rr} & 0 & \cdot\cdot & 0 \\ 0 & \sigma_{rr} & \cdot\cdot & 0 \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ 0 & 0 & 0 & \sigma_{rr} \end{bmatrix}$$

and

$$E[\varepsilon_r\varepsilon_s'] = E\left[\begin{pmatrix} \varepsilon_{r1} & \varepsilon_{r2} & \cdot\cdot & \varepsilon_{rN} \end{pmatrix} \begin{pmatrix} \varepsilon_{s1} \\ \varepsilon_{s2} \\ \cdot\cdot \\ \varepsilon_{sN} \end{pmatrix}\right]$$

$$= E \begin{bmatrix} \varepsilon_{r1}\varepsilon_{s1} & \varepsilon_{r1}\varepsilon_{s2} & \cdot\cdot & \varepsilon_{r1}\varepsilon_{sN} \\ \varepsilon_{r2}\varepsilon_{s1} & \varepsilon_{r2}\varepsilon_{s2} & \cdot\cdot & \varepsilon_{r2}\varepsilon_{sN} \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ \varepsilon_{rN}\varepsilon_{s1} & \varepsilon_{rN}\varepsilon_{s2} & & \varepsilon_{rN}\varepsilon_{sN} \end{bmatrix} = \begin{bmatrix} \sigma_{rs} & 0 & \cdot\cdot & 0 \\ 0 & \sigma_{rs} & \cdot\cdot & 0 \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ 0 & 0 & 0 & \sigma_{rs} \end{bmatrix}$$

Hence, we have

$$Var(\varepsilon) = \begin{bmatrix} \sigma_{11}I_N & \sigma_{12}I_N & \cdots & \sigma_{1M}I_N \\ \sigma_{21}I_N & \sigma_{22}I_N & \cdots & \sigma_{2M}I_N \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{M1}I_N & \sigma_{M2}I_N & \cdots & \sigma_{MM}I_N \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{M1} & \sigma_{M2} & \cdots & \sigma_{MM} \end{bmatrix} \otimes I_N = \Sigma \otimes I_N$$

- We can apply the GLS technique to the stacked system of equations to get;

$$\widehat{\beta}_{GLS} = \left(X'(\Sigma \otimes I_N)^{-1} X\right)^{-1} \left(X'(\Sigma \otimes I_N)^{-1} y\right)$$

$$= \left(X'\left(\Sigma^{-1} \otimes I_N\right) X\right)^{-1} \left(X'\left(\Sigma^{-1} \otimes I_N\right) y\right) \text{ since } (A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$

Now, denote σ^{ij} as the (i,j) element of Σ^{-1}. Then,

$$\Sigma^{-1}\otimes I_N = \begin{bmatrix} \sigma^{11}I_N & \sigma^{12}I_N & \cdots & \sigma^{1M}I_N \\ \sigma^{21}I_N & \sigma^{22}I_N & \cdots & \sigma^{2M}I_N \\ \cdots & \cdots & \cdots & \cdots \\ \sigma^{M1}I_N & \sigma^{M2}I_N & \cdots & \sigma^{MM}I_N \end{bmatrix}$$

Hence,

$$\begin{aligned}
\widehat{\beta}_{GLS} &= \left(\begin{bmatrix} X_1' & 0 & \cdot\cdot & 0 \\ 0 & X_2' & \cdot\cdot & 0 \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ 0 & 0 & \cdot\cdot & X_M' \end{bmatrix} \begin{bmatrix} \sigma^{11}I_N & \sigma^{12}I_N & \cdots & \sigma^{1M}I_N \\ \sigma^{21}I_N & \sigma^{22}I_N & \cdots & \sigma^{2M}I_N \\ \cdots & \cdots & \cdots & \cdots \\ \sigma^{M1}I_N & \sigma^{M2}I_N & \cdots & \sigma^{MM}I_N \end{bmatrix} \begin{bmatrix} X_1 & 0 & \cdot\cdot & 0 \\ 0 & X_2 & \cdot\cdot & 0 \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ 0 & 0 & \cdot\cdot & X_M \end{bmatrix} \right)^{-1} \\
&\times \begin{bmatrix} X_1' & 0 & \cdot\cdot & 0 \\ 0 & X_2' & \cdot\cdot & 0 \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ 0 & 0 & \cdot\cdot & X_M' \end{bmatrix} \begin{bmatrix} \sigma^{11}I_N & \sigma^{12}I_N & \cdots & \sigma^{1M}I_N \\ \sigma^{21}I_N & \sigma^{22}I_N & \cdots & \sigma^{2M}I_N \\ \cdots & \cdots & \cdots & \cdots \\ \sigma^{M1}I_N & \sigma^{M2}I_N & \cdots & \sigma^{MM}I_N \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_M \end{bmatrix} \\
&= \begin{bmatrix} \sigma^{11}X_1'X_1 & \sigma^{12}X_1'X_2 & \cdot\cdot & \sigma^{1M}X_1'X_M \\ \sigma^{21}X_2'X_1 & \sigma^{22}X_2'X_2 & \cdot\cdot & \sigma^{2M}X_2'X_M \\ \cdots\cdots & \cdots\cdots & \cdot\cdot & \cdots\cdots \\ \sigma^{M1}X_M'X_1 & \sigma^{M2}X_M'X_2 & \cdot\cdot & \sigma^{MM}X_M'X_M \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^{M}\sigma^{1j}X_1'y_j \\ \sum_{j=1}^{M}\sigma^{2j}X_M'y_j \\ \cdots\cdots \\ \sum_{j=1}^{M}\sigma^{Mj}X_M'y_j \end{bmatrix} \qquad (20)
\end{aligned}$$

- The formula given above is useless in the sense that $\sigma_{ij}'s$ are unknown. What do we do, then? Yes, we always estimate when we have something unknown. How can we estimate the stuff? Surely, it's got to be consistent for $\sigma_{ij}'s$. We can use the OLS residuals to estimate $\sigma_{ij}'s$ consistently. The procedure is called Feasible SURE. The procedure is;
  - Estimate each equation by OLS ignoring the inter-equation dependence.
  - Calculate the residual vectors $e_j, j = 1, 2, \cdots, M$ where each e_j is $(N \times 1)$ matrix.
  - Calculate cross product moments of residuals such as $e_i'e_j, i, j = 1, 2, \cdots, M$
  - Set $\widehat{\sigma}_{ij} = \frac{e_i'e_j}{N-K}$ or $\widehat{\sigma}_{ij} = \frac{e_i'e_j}{N}$.
  - Form the matrix $\widehat{\Sigma}$ with $\widehat{\sigma}_{ij}'s$.
  - Do FGLS with $\widehat{\Sigma}$.

$$\widehat{\beta}_{GLS} = \left(X'\left(\widehat{\Sigma}^{-1}\otimes I_N\right)X\right)^{-1}\left(X'\left(\widehat{\Sigma}^{-1}\otimes I_N\right)y\right)$$

- It is a **helpful exercise** to write down the likelihood function and find MLE for the model.