Generalised Least Squares

References

Davidson, Russell and James G. MacKinnon (2004). *Econometric Theory and Methods*. Oxford University Press.

Hayashi, Fumio (2000). *Econometrics*. Princeton University Press.

1 Introduction

If the parameters of a regression model are to be estimated efficiently by OLS, the error terms must be uncorrelated and have the same variance. These assumptions are needed to prove the Gauss-Markov Theorem and to show that the least squares estimator is asymptotically efficient. Moreover, the usual estimators of the covariance matrix of the OLS estimators are not valid when these assumptions do not hold, although alternative "sandwich" covariance matrix estimators that are asymptotically valid may be available.

Thus, it is clear that we need new estimation methods to handle regression models with error terms that are heteroskedastic, serially correlated, or both.

Since heteroskedasticity and serial correlation affect both linear and nonlinear regression models in the same way, there is no harm in limiting our attention to the simpler, linear case. We will be concerned with the model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},\ \mathbb{E}[\mathbf{u}\mathbf{u}^\top] = \boldsymbol{\Omega}, \tag{1}$$

where $\boldsymbol{\Omega}$, the covariance matrix of the error terms, is a positive definite $n \times n$ matrix. If $\boldsymbol{\Omega}$ is equal to $\sigma^2\mathbf{I}$, then (1) is just the linear regression model we have always looked at when constructing OLS estimators (error terms are uncorrelated and homoskedastic). If $\boldsymbol{\Omega}$ is diagonal with non-constant diagonal elements, then the error terms are still uncorrelated, but they are heteroskedastic. If $\boldsymbol{\Omega}$ is not diagonal, then u_i and u_j are correlated whenever Ω_{ij}, the ij^{th} element of $\boldsymbol{\Omega}$, is nonzero.

We will obtain an efficient estimator for the vector $\boldsymbol{\beta}$ in the model (1). This efficient estimator is called the General Least Squares (GLS) estimator.

2 The GLS estimator

In order to obtain the GLS estimator, we need to transform the model so that the transformed model satisfies the conditions of the Gauss-Markov theorem. Estimating the transformed model by OLS therefore yields efficient estimates. The

transformation is expressed in terms of an $n \times n$ matrix $\mathbf{\Psi}$, which is usually triangular, that satisfies the equation:

$$\mathbf{\Omega}^{-1} = \mathbf{\Psi}\mathbf{\Psi}^\top. \tag{2}$$

Such a matrix can always be found using Crout's algorithm. Premultiplying (1) by $\mathbf{\Psi}^\top$ gives:

$$\mathbf{\Psi}^\top\mathbf{y} = \mathbf{\Psi}^\top\mathbf{X}\boldsymbol{\beta} + \mathbf{\Psi}^\top\mathbf{u}. \tag{3}$$

Because the covariance matrix $\mathbf{\Omega}$ is nonsingular, the matrix $\mathbf{\Psi}$ must be as well, and so the transformed regression model (3) is perfectly equivalent to the original model (1). The OLS estimator of $\boldsymbol{\beta}$ from regression (3) is:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{GLS}} &= (\mathbf{X}^\top\mathbf{\Psi}\mathbf{\Psi}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Psi}\mathbf{\Psi}^\top\mathbf{y} \\ &= (\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{y}. \end{aligned} \tag{4}$$

This estimator is the called the GLS estimator of $\boldsymbol{\beta}$. The GLS estimator solves the following problem:

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_{\mathbf{\Omega}} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top\mathbf{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

and its first order condition is:

$$-2\mathbf{X}^\top\mathbf{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{GLS}}) = 0.$$

Substituting (1) into (4) yields:

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} - \boldsymbol{\beta} = (\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{u}. \tag{5}$$

Since the GLS estimator is simply the OLS estimator from (3), its covariance matrix can be found directly from the standard formula for the OLS covariance matrix:

$$\mathbb{E}\left[\left(\hat{\boldsymbol{\beta}}_{\text{GLS}} - \boldsymbol{\beta}\right)(\hat{\boldsymbol{\beta}}_{\text{GLS}} - \boldsymbol{\beta})^\top|\mathbf{X}\right] = (\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbb{E}[\mathbf{u}\mathbf{u}^\top|\mathbf{X}]\mathbf{\Omega}^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1},$$

which simplifies to:

$$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{GLS}}) = (\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1}.$$

2.1 Efficiency of the GLS estimator and unbiasedness

It's quite easy to prove that the GLS estimator is unbiased. From (3) we have the following model and assumptions:

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \widetilde{\mathbf{u}},\ \mathbb{E}[\widetilde{\mathbf{u}}|\mathbf{X}] = 0,\ \mathbb{E}(\widetilde{\mathbf{u}}\widetilde{\mathbf{u}}|\mathbf{X}) = \sigma^2\mathbf{I}, \tag{6}$$

where a tilde variable denotes that the variable has been premultiplied by $\boldsymbol{\Psi}^\top$, so the GLS estimator is the OLS estimator of the transformed regression (3):

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{y}}.$$

To show its unbiasedness:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{GLS}} &= \boldsymbol{\beta} + (\tilde{\mathbf{X}}^\top \tilde{\boldsymbol{X}})^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{u}} \\ \mathbb{E}(\hat{\boldsymbol{\beta}}_{\text{GLS}} | \mathbf{X}) &= \boldsymbol{\beta} + (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \underbrace{\mathbb{E}[\tilde{\mathbf{u}} | \mathbf{X}]}_{=0}. \end{aligned}$$

Thus, the GLS estimator is unbiased. To show its efficiency, consider (1) and the OLS estimator:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{OLS}} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \\ \implies \mathbb{E}\left[\left(\hat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta} \right) (\hat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta})^\top | \mathbf{X} \right] &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[\mathbf{u}\mathbf{u}^\top | \mathbf{X}] \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ \therefore \text{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}} | \mathbf{X}) &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Omega} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}. \end{aligned}$$

When $\boldsymbol{\Omega} = \sigma^2 \mathbf{I}$, as in the case for unconditional homoskedasticity, then the variance-covariance matrix of the OLS estimator is $\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$. Also, notice that if $\boldsymbol{\Omega} = \mathbb{E}[\mathbf{u}\mathbf{u}^\top | \mathbf{X}]$ (conditional homoskedasticity), then we have the sandwich estimator, which we looked at when we analysed heteroskedasticity-consistent variance-covariance matrices.

One can also easily see that

$$\left(\text{Var}(\hat{\boldsymbol{\beta}}_{\text{GLS}}) \right)^{-1} - \left(\text{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) \right)^{-1}$$

results in a positive semidefinite matrix. i.e.,

$$\left(\text{Var}(\hat{\boldsymbol{\beta}}_{\text{GLS}}) \right)^{-1} - \left(\text{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) \right)^{-1} = (\mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{X}) - \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \boldsymbol{\Omega} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}$$

is positive semidefinite, implying that the GLS estimator is more efficient than the OLS estimator when we no longer can assume homoskedasticity.

2.2 Weighted least squares

It is particularly easy to obtain GLS estimates when the error terms are heteroskedastic but uncorrelated. This implies that the matrix $\boldsymbol{\Omega}$ is diagonal. Let ω_t^2 denote the t^{th} diagonal element of $\boldsymbol{\Omega}$. Then $\boldsymbol{\Omega}^{-1}$ is a diagonal matrix with t^{th} diagonal element ω_t^{-2}, and $\boldsymbol{\Psi}$ can be chosen as the diagonal matrix with t^{th} diagonal element ω_t^{-1}. Thus, we see that, for a typical observation, regression (3) can be written as:

$$\frac{1}{\omega_t} y_t = \frac{1}{\omega_t} \mathbf{X}_t \boldsymbol{\beta} + \frac{1}{\omega_t} u_t \tag{7}$$

where we are trying to "weigh the observations". Also, as said:

$$\boldsymbol{\Omega} = \begin{bmatrix} \omega_1^2 & 0 & \cdots & 0 \\ 0 & \omega_2^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \omega_n^2 \end{bmatrix} = \text{diag}(\{\omega_t^2\}_{t=1}^n),$$

which implies that

$$\boldsymbol{\Psi} = \begin{bmatrix} \omega_1^{-1} & 0 & \cdots & 0 \\ 0 & \omega_2^{-1} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \omega_n^{-1} \end{bmatrix}.$$

In practice, we do not know ω_t^2. One choice of skedastic function is

$$\omega_t^2 = \mathbb{E}[u_t^2] = \exp(\delta + Z_t\gamma).$$

Note that for any values of δ and γ, $\exp(\delta + Z_t\gamma) > 0$. Hence why many economists like to express the function as an exponential function, this makes later evaluation, using logs, much easier. Also, some or all elements of Z_t may well belong to X_t. The function $\exp(\delta + Z_t\gamma)$ is an example of a skedastic function. In the same way that a regression function determines the conditional mean of a random variable, a skedastic function determines its conditional variance. The skedastic function has the property that it is positive for any parameter γ. This is desirable, since any negative estimated variances would be highly inconvenient. The estimation procedure for ω_t^2 and $\boldsymbol{\beta}$ is as follows:

1. Estimate $\boldsymbol{\beta}$ by OLS, and compute $\hat{\mathbf{u}} = \mathbf{M}_{\mathbf{X}}\mathbf{y}$, the OLS residuals.
2. Estimate δ and γ from the auxiliary regression

$$\ln \hat{u}_t^2 = \delta + Z_t\gamma + v_t.$$

3. Compute and estimate ω_t from

$$\hat{\omega}_t = \left(\exp(\hat{\delta} + Z_t\hat{\gamma})\right)^{\frac{1}{2}}.$$

4. Estimate $\boldsymbol{\beta}$ using $\hat{\omega}_t$ in place of ω_t. This is also known as the feasible GLS (FGLS):

$$\hat{\boldsymbol{\beta}}_{\text{FGLS}} = \left(\mathbf{X}^\top \hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top \hat{\boldsymbol{\Omega}}^{-1}\mathbf{y},$$

where

$$\hat{\boldsymbol{\Omega}} = \boldsymbol{\Omega}(\hat{\delta}, \hat{\gamma}).$$

2.3 Asymptotic distribution of GLS and FGLS

The GLS estimator can be written as

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \boldsymbol{\beta} + (\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{u},$$

and

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{GLS}} - \boldsymbol{\beta}) = \left(n^{-1}\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1} \frac{1}{\sqrt{n}}\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{u}.$$

Now, using the central limit theorem, we have

$$\frac{1}{\sqrt{n}}\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{u} \overset{a}{\sim} N(0, \mathbf{S}_{\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{X}}),$$
$$\operatorname*{plim}_{n\to\infty}\left(n^{-1}\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{X}\right) = \mathbf{S}_{\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{X}}.$$

Combining the above results we derive

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{GLS}} - \boldsymbol{\beta}) \overset{a}{\sim} N(0, \mathbf{S}^{-1}_{\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{X}}).$$

The FGLS can be written as

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{FGLS}} - \boldsymbol{\beta}) = \left(n^{-1}\mathbf{X}^\top \hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right)^{-1} \frac{1}{\sqrt{n}}\mathbf{X}^\top \hat{\boldsymbol{\Omega}}^{-1}\mathbf{u},$$

which is very similar to the expression derived for the GLS estimator. One can show that

$$\frac{1}{\sqrt{n}}\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{u} \overset{a}{=} \frac{1}{\sqrt{n}}\mathbf{X}^\top \hat{\boldsymbol{\Omega}}^{-1}\mathbf{u}, \text{ and,}$$
$$\operatorname*{plim}_{n\to\infty}\left(n^{-1}\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{X}\right) = \operatorname*{plim}_{n\to\infty}\left(n^{-1}\mathbf{X}^\top \hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right),$$

i.e. the FGLS and GLS are asymptotically equivalent. This result holds for the case where $\boldsymbol{\Omega}$ is diagonal and $\mathbf{X}$ is exogenous or predetermined. However, if $\boldsymbol{\Omega}$ is not diagonal then the OLS estimator is not consistent whenever any element of $\mathbf{X}_t$ is a lagged dependent variable.

2.4 Heteroskedasticity

There are two situations in which the error terms are heteroskedastic but serially uncorrelated. In the first, the form of the heteroskedasticity is completely unknown, while, in the second, the heteroskedastic function is known except for the value of the parameters that can be estimated consistently. The fact that these HCCME's are sandwich covariance matrices makes it clear that, although they are consistent under standard regularity conditions, OLS is inefficient when the error

terms are heteroskedastic.

If the variances of all the error terms are known, at least up to a scale factor, then efficient estimates can be obtained by weighted least squares. For a linear model, we need to multiply all of the variables by ω_t^{-1}, the inverse of the standard error of u_t, and then use OLS. The usual OLS covariance matrix is perfectly valid, although it is desirable to replace s^2 by 1 if the variances are completely known, since in that case $s^2 \to 1$ as $n \to \infty$.

If the form of the heteroskedasticity is known, but the skedastic function depends on unknown parameters, then we can use feasible weighted least squares and still achieve asymptotic efficiency. An example of such a procedure was discussed in the previous lecture. As we have seen, it makes no difference asymptotically whether the ω_t are known or merely estimated consistently, although it can certainly make a difference in finite samples. Asymptotically, at least, the usual OLS covariance matrix is just as valid with feasible WLS as with WLS.

2.4.1 Testing for Heteroskedasticity

In some cases, it may be clear from the specification of the model that the error terms must exhibit a particular pattern of heteroskedasticity. In many cases, however, we may hope that the error terms are homoskedastic but be prepared to admit the possibility that they are not. In such cases, if we have no information on the form of the skedastic function, it may be prudent to employ an HCCME, especially if the sample size is large.

If we have information on the form of the skedastic function, we might well wish to use WLS. Before doing so, it is advisable to perform a specification test of the null hypothesis that the error terms are homoskedastic against whatever heteroskedastic alternatives may seem reasonable. There are many ways to perform this type of specification test. The simplest approach that is widely applicable, and the only one that we will discuss, involves running an artificial regression in which the regressand is the vector of squared residuals from the model or test.

A reasonably general model of conditional heteroskedasticity is:

$$\mathbb{E}[u_t^2|\Omega_t] = h(\delta + Z_t\gamma) \tag{8}$$

where the skedastic function $h(\cdot)$ is a non-linear function that can take on only positive values, $\boldsymbol{Z}_t$ is a $1 \times r$ vector of observations on exogenous or predetermined variables that belong to the information set Ω_t, δ is a scalar parameter, and $\boldsymbol{\gamma}$ is an

r-vector of parameters. Under the null, $\boldsymbol{\gamma} = 0$ and the function collapses to $h(\delta)$, a constant. A very simple procedure to detect the presence of heteroskedasticity is to estimate:

$$u_t^2 = \alpha + \boldsymbol{Z}_t\eta + \text{residuals}, \tag{9}$$

and test $\eta = 0$ by an F-test, asymptotically distributed as $F(r, \infty)$. This expression is obtained from a Taylor expansion of $h(\delta + Z_t\gamma)$ and imposing restrictions from the null assumption. Alternatively, we could perform an nR^2 test, where we use the centered R^2 from the above regression, which is asymptotically distributed as $\chi^2(r)$. In practice, we don't actually observe u_t^2, so we replace it with $\hat{u}_t^2$.

2.5 Time Series Models

The error terms for nearby observations may be correlated, or may appear to be correlated, in any sort of regression model, but this phenomenon is most commonly encountered in models estimated with time-series data, where it is known as serial correlation or autocorrelation. In practice, what appears to be serial correlation may instead just be a misspecified model. In some circumstances, though, it is natural to model the serial correlation by assuming the error terms follow some sort of stochastic process.

If there is reason to believe that serial correlation may be present, the first step is usually to test the null hypothesis that the errors are serially uncorrelated against a plausible alternative that involves serial correlation.

2.5.1 Autoregressive errors

One of the simplest and most commonly used stochastic processes is the first-order autoregressive progress, or AR(1) process. We have already encountered regression models with error terms that follow such a progress. Recall from Davidson & MacKinnon Exercise (6.04) that the AR(1) process can be written as:

$$u_t = \rho u_{t-1} + \varepsilon,\ \varepsilon_t \sim \text{IID}(0, \sigma_\varepsilon^2),\ |\rho| < 1, \tag{10}$$

where

$$y_t = \mathbf{X}_t\boldsymbol{\beta} + u_t.$$

The error at time t is equal to some fraction ρ of the error at time $t-1$, with the sign changed if $\rho < 0$, plus the innovation ε_t.
The condition in equation (10) that $|\rho| < 1$ is called a stationary condition, because it is necessary for the AR(1) process to be stationary. There are several

definitions of stationarity in time series analysis.

Definition: A series of typical element u_t is stationary if the unconditional expectation $\mathbb{E}[u_t]$ and the unconditional variance $\text{Var}(u_t)$ exist and are independent of t, and if the covariance $\text{Cov}(u_t, u_{t-j})$ is also, for any given j, independent of t. This definition of stationarity is sometimes referred to as covariance stationarity, or wide sense stationarity.

Consider:

$$\mathbb{E}\left[\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}\right]$$

and

$$\mathbb{E}\begin{bmatrix} \mathbb{E}[u_1^2] & \mathbb{E}[u_1u_2] & \cdots & \mathbb{E}[u_1u_n] \\ \mathbb{E}[u_2u_1] & \mathbb{E}[u_2^2] & & \vdots \\ \vdots & & \ddots & \mathbb{E}[u_{n-1}u_n] \\ \mathbb{E}[u_nu_1] & \cdots & \mathbb{E}[u_nu_{n-1}] & \mathbb{E}[u_n^2] \end{bmatrix} = \boldsymbol{\Omega}.$$

We can see that:

$$u_t = \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \rho^3\varepsilon_{t-3} + \ldots,$$

and as the innovations are independent and uncorrelated:

$$\begin{aligned} \sigma_u^2 &= \text{Var}(u_t) = \text{Var}(\varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \rho^3\varepsilon_{t-3} + \ldots) \\ &= \text{Var}(\varepsilon_t) + \text{Var}(\rho\varepsilon_t) + \text{Var}(\rho^2\varepsilon_{t-2}) + \ldots \\ &= \text{Var}(\varepsilon_t) + \rho^2\text{Var}(\varepsilon_{t-1}) + \rho^4\text{Var}(\varepsilon_{t-2}) + \ldots \\ &= (1 + \rho^2 + \rho^4 + \ldots)\sigma_\varepsilon^2 \\ &= \frac{1}{1-\rho^2}\sigma_\varepsilon^2 = \frac{\sigma_\varepsilon^2}{1-\rho^2} = \sigma_u^2 \end{aligned}$$

We say that the residuals follow an AR(1) process. By imposing $|\rho| < 1$ we guarantee that the variance of u_t does not explode when $t \to \infty$.

In order to find our $\boldsymbol{\Psi}$ matrix, we need to compute the following results (for the off diagonals in our $\boldsymbol{\Omega}$ matrix):

$$\begin{aligned}
\mathbb{E}[u_t u_{t-1}] &= \mathbb{E}[(\rho u_{t-1} + \varepsilon_t) u_{t-1}] \\
&= \rho \mathbb{E}[u_{t-1}^2] + \mathbb{E}[\varepsilon_t u_{t-1}] \\
&= \rho \frac{\sigma_\varepsilon^2}{(1-\rho^2)}, \\
\mathbb{E}[u_t u_{t-2}] &= \mathbb{E}[(\rho^2 u_{t-2} + \rho\varepsilon_{t-1} + \varepsilon_t) u_{t-2}] \\
&= \rho^2 E(u_{t-2}^2) + \rho E(\varepsilon_{t-1} u_{t-2}) + E(\varepsilon_t u_{t-1}) \\
&= \rho^2 \left(\frac{\sigma_\varepsilon^2}{1-\rho^2} \right), \\
\implies E[u_t u_{t-s}] &= \rho^s \left(\frac{\sigma_\varepsilon^2}{1-\rho^2} \right).
\end{aligned}$$

Then we find:

$$\boldsymbol{\Omega} = \begin{bmatrix}
\frac{\sigma_\varepsilon^2}{1-\rho^2} & \rho\left(\frac{\sigma_\varepsilon^2}{1-\rho^2}\right) & \rho^2\left(\frac{\sigma_\varepsilon^2}{1-\rho^2}\right) & \cdots & \rho^{n-1}\left(\frac{\sigma_\varepsilon^2}{1-\rho^2}\right) \\
\rho\left(\frac{\sigma_\varepsilon^2}{1-\rho^2}\right) & \frac{\sigma_\varepsilon^2}{1-\rho^2} & & & \vdots \\
\rho^2\left(\frac{\sigma_\varepsilon^2}{1-\rho^2}\right) & & \ddots & & \\
\vdots & & & \ddots & \rho\left(\frac{\sigma_\varepsilon^2}{1-\rho^2}\right) \\
\rho^{n-1}\left(\frac{\sigma_\varepsilon^2}{1-\rho^2}\right) & \cdots & & \rho\left(\frac{\sigma_\varepsilon^2}{1-\rho^2}\right) & \frac{\sigma_\varepsilon^2}{1-\rho^2}
\end{bmatrix},$$

and

$$\boldsymbol{\Omega} = \frac{\sigma_\varepsilon^2}{1-\rho^2} \begin{bmatrix}
1 & \rho & \cdots & \rho^{n-1} \\
\rho & 1 & & \vdots \\
\vdots & & \ddots & \rho \\
\rho^{n-1} & \cdots & \rho & 1
\end{bmatrix},$$

and so:

$$\boldsymbol{\Psi}(\rho) = \begin{bmatrix}
(1-\rho)^{\frac{-1}{2}} & -\rho & 0 & \cdots & 0 \\
0 & 1 & -\rho & \cdots & 0 \\
\vdots & 0 & \ddots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \ddots & -\rho \\
0 & \cdots & \cdots & 0 & 1
\end{bmatrix},$$

thus we can yield:

$$\begin{array}{rcl} \boldsymbol{\Psi}^\top \mathbf{u} & = & \boldsymbol{\varepsilon} \\ \begin{bmatrix} (1-\rho)^{\frac{-1}{2}} & 0 & \cdots & \cdots & 0 \\ -\rho & 1 & 0 & \ddots & \vdots \\ 0 & -\rho & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & -\rho & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_n \end{bmatrix} & = & \begin{bmatrix} (1-\rho^2)u_1 \\ u_2 - \rho u_1 \\ u_3 - \rho u_2 \\ \vdots \\ u_n - \rho u_{n-1} \end{bmatrix} \begin{array}{l} \Rightarrow \varepsilon_1 \\ \Rightarrow \varepsilon_2 \\ \Rightarrow \varepsilon_3 \\ \vdots \\ \Rightarrow \varepsilon_n \end{array} \end{array}$$

The idea is similar to the WLS estimates:
1. Estimate $\boldsymbol{\beta}$ by OLS, and compute $\hat{\mathbf{u}} = \mathbf{M}_{\mathbf{X}}\mathbf{y}$, the OLS residuals (Note: X_t cannot have lagged dependent variables!).

2. Estimate ρ from the auxiliary regression:

$$\hat{u}_t = \rho \hat{u}_{t-1} + \varepsilon_t^* \tag{11}$$

we have n-1 elements (because $\hat{u}_1$ is not included on the LHS).

3. An estimator for $\boldsymbol{\Psi}(\rho)$ is $\boldsymbol{\Psi}(\hat{\rho})$.
The feasible GLS estimator is:

$$\hat{\boldsymbol{\beta}}_{\text{FGLS}} = (\mathbf{X}^\top \triangle(\hat{\rho})^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \triangle(\hat{\rho})^{-1}\mathbf{y},$$

where we define $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top = \triangle^{-1}$, and it has the following variance estimator:

$$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{FGLS}}) = s^2(\mathbf{X}^\top \triangle(\hat{\rho})^{-1}\mathbf{X})^{-1}.$$

s^2 is an estimator of the variance of the residual $\boldsymbol{\Psi}^\top \mathbf{u}$ obtained from the transformed regression:

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

A simple test to verify if the residuals follow an AR(1) process is to test $H_0 : \rho = 0$ in the regression (11). However the best known test, albeit quite dated, is the d-test proposed by Durbin and Watson, also known as the DW statistic:

$$d = \frac{\sum_{t=2}^{n}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n}\hat{u}_t^2} = \frac{n^{-1}\hat{u}_1^2 + 2n^{-1}\hat{\mathbf{u}}^\top \hat{\mathbf{u}}_1}{n^{-1}\hat{\mathbf{u}}^\top \hat{\mathbf{u}}} \tag{12}$$

Asymptotically, one can show that the d-test tends to: $2 - 2\rho$. Hence a value of $d \approxeq 2$ corresponds to the absence of serial correlation in the residuals, while values

of d less than 2 corresponds to $\hat{\rho} > 0$, and values of greater than 2 correspond to $\hat{\rho} < 0$.
Note: The DW statistic is not valid when there are lagged dependent variables among the regressors.

2.5.2 Moving average errors

Autoregressive processes are not the only way to model stationary time series. Another type of stochastic process the moving average (MA) process. The simplest of these is the first order moving average (MA(1)) process

$$u_t = \epsilon_t + \alpha_1 \epsilon_{t-1}, \ \epsilon_t \sim \text{IID}(0, \sigma_\epsilon^2) \tag{13}$$

in which the error term u_t is a weighted average of two successive innovations, ϵ_t and ϵ_{t-1}. This is also a covariance-stationary process:

$$\begin{aligned}
\text{Var}(u_t) &= \text{Var}(\epsilon_t + \alpha_1 \epsilon_{t-1}) = \sigma_\epsilon^2 + \alpha_1^2 \sigma_\epsilon^2 = (1 + \alpha_1^2)\sigma_\epsilon^2 \\
\mathbb{E}[u_t u_{t-1}] &= \mathbb{E}[(\epsilon_t + \alpha_1 \epsilon_{t-1})(\epsilon_{t-1} + \alpha_1 \epsilon_{t-2})] = \alpha_1 \sigma_\epsilon^2 \\
\mathbb{E}[u_t u_{t-2}] &= \mathbb{E}[(\epsilon_t + \alpha_1 \epsilon_{t-1})(\epsilon_{t-2} + \alpha_1 \epsilon_{t-3})] = 0.
\end{aligned}$$

The variance-covariance matrix $\boldsymbol{\Omega}$ is

$$\boldsymbol{\Omega} = \sigma_\epsilon^2 \begin{bmatrix} (1+\alpha_1^2) & \alpha_1 & 0 & \cdots & 0 \\ \alpha_1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \alpha_1 \\ 0 & \cdots & 0 & \alpha_1 & (1+\alpha_1^2) \end{bmatrix},$$

where it is evident that there is no correlation between error terms which are more than one period apart.

2.6 Panel data

Many data sets are measured across two dimensions. One dimension is time, and the other is usually called the cross-section dimension. Data of this type are often referred to as panel data. It is likely that the error terms for a model using panel data will display certain types of dependence, which should be taken into account when we estimate such a model. Consider the following linear regression model:

$$y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + u_{it}, \ i = 1, ..., m, \ t = 1, ..., T, \tag{14}$$

where $\mathbf{X}_{it}$ is a $1 \times k$ vector of observations on explanatory variables. There is assumed to be m cross-sectional units and T time periods, for a total of $n = m \times T$

observations. If each u_{it} has expectation zero conditional on its corresponding $\mathbf{X}_{it}$, we can estimate (14) via OLS. But the OLS estimator is not efficient if the u_{it} are not IID, and the IID assumption is rarely realistic with panel data. For example, for (14) consider the following

$$u_{it} = v_i + e_t + \epsilon_{it},$$

$$\begin{aligned}
\mathbb{E}[v_i^2] = \sigma_v^2,\ \mathbb{E}[v_i v_j] = 0,\\
\mathbb{E}[e_t^2] = \sigma_e^2,\ \mathbb{E}[e_t e_{t-s}] = 0\\
\mathbb{E}[\epsilon_{it}] = \sigma_\epsilon^2,\ \mathbb{E}[\epsilon_{it}\epsilon_{jt-s}] = 0.
\end{aligned}$$

The above specification for the error term is an example of an 'error-components model', where the error term u_{it} consists of two or three independent shocks. In order to estimate an error-components model, the e_t and v_i can be regarded as being either fixed or random. If the e_t and v_i are thought of as fixed effects, then they are treated as parameters to be estimated, and it turns out that they can then be estimated by OLS using dummy variables.

If they are thought of as random effects, then we must figure out the covariance matrix of the u_{it} as functions of the variances of the e_t, v_i, and ϵ_{it}, and use FGLS. Each of these approaches can be appropriate in some circumstances but may be inappropriate in others.

For simplicity, let us eliminate e_t and consider $u_{it} = v_i + \epsilon_{it}$. In matrix notation, the linear panel data model is

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1T} \\ y_{21} \\ \vdots \\ y_{2T} \\ \vdots \\ y_{m1} \\ \vdots \\ y_{mT} \end{bmatrix}_{n\times 1} = \begin{bmatrix} \mathbf{X}_{11} \\ \vdots \\ \mathbf{X}_{1T} \\ \mathbf{X}_{21} \\ \vdots \\ \mathbf{X}_{2T} \\ \vdots \\ \mathbf{X}_{m1} \\ \vdots \\ \mathbf{X}_{mT} \end{bmatrix}_{n\times k} \boldsymbol{\beta} + \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix}_{n\times m} \boldsymbol{\eta} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1T} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2T} \\ \vdots \\ \epsilon_{m1} \\ \vdots \\ \epsilon_{mT} \end{bmatrix}_{n\times 1}, \quad \boldsymbol{\eta} = \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix},$$

$$\Leftrightarrow \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \underbrace{\mathbf{D}\boldsymbol{\eta} + \boldsymbol{\epsilon}}_{\mathbf{u}}, \quad \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \sigma_\epsilon^2 \mathbf{I}_n, \tag{15}$$

where $\mathbf{y}$ and $\boldsymbol{\epsilon}$ are n-vectors with typical elements y_{it} and ϵ_{it}, respectively, and $\mathbf{D}$ is an $n \times m$ matrix of dummy variables, constructed in such a way that the

element in the row corresponding to observation it, for $i = 1, ..., m$ and $t = 1, ..., T$, and column j, for $j = 1, ..., m$, is equal to 1 if $i = j$ and equal to 0 otherwise. The m-vector $\boldsymbol{\eta}$ has typical element v_i, and so it follows that the n-vector $\mathbf{D}\boldsymbol{\eta}$ has element v_i in the row corresponding to observation it. Note that there is exactly one element on $\mathbf{D}$ equal to 1 in each row, which implies that the n-vector $\boldsymbol{\iota}$ with each element equal to 1 i s a linear combination of the columns of $\mathbf{D}$. Consequently, in order to avoid collinear regressors, the matrix $\mathbf{X}$ should not contain a constant.

The vector $\boldsymbol{\eta}$ essentially plays the role of a parameter vector, and it is in this sense the the v_i are called fixed effects (though they can be random, as we will later show). The essential thing is that they must be independent of the error terms, ϵ_{it}. They may, however, be correlated with the explanatory variables, $\mathbf{X}_{it}$.

We are going to use this specification to investigate two cases:

1. Fixed-effect estimation.
2. Random-effect estimation.

2.6.1 Fixed effect estimation

Also known as the "within-group" estimator or dummy variable estimator, because we assume that $e_t = 0$ (no time component shock). Let us assume that the moment conditions

$$\begin{aligned}\mathbb{E}[y_{it} - \mathbf{X}_{it}\boldsymbol{\beta} - v_i] = 0,\\ \mathbb{E}[\mathbf{X}_{it}^\top(y_{it} - \mathbf{X}_{it}\boldsymbol{\beta} - v_i)] = 0,\end{aligned}$$

are satisfied, which is similar to the OLS regression model assumptions. Note that we are allowing that $\mathbb{E}[\boldsymbol{X}_{it}^\top v_i] \neq 0$. By applying the FWL theorem, we have

$$\hat{\boldsymbol{\beta}}_{\text{WG}} = (\mathbf{X}^\top \mathbf{M}_{\mathbf{D}} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{M}_{\mathbf{D}} \mathbf{y}, \tag{16}$$

where

$$\mathbf{M}_{\mathbf{D}} = \mathbf{I}_n - \mathbf{P}_{\mathbf{D}} = \mathbf{I}_n - \mathbf{D}(\mathbf{D}^\top \mathbf{D})^{-1} \mathbf{D}^\top.$$

One can show that (with some butchering of notation):

$$\mathbf{M_D X} = \begin{bmatrix} \mathbf{X}_{11} - \bar{\mathbf{X}}_1 \\ \vdots \\ \mathbf{X}_{1T} - \bar{\mathbf{X}}_1 \\ \mathbf{X}_{21} - \bar{\mathbf{X}}_2 \\ \vdots \\ \mathbf{X}_{2T} - \bar{\mathbf{X}}_1 \\ \vdots \\ \mathbf{X}_{m1} - \bar{\mathbf{X}}_m \\ \vdots \\ \mathbf{X}_{mT} - \bar{\mathbf{X}}_m \end{bmatrix}, \ \bar{\mathbf{X}}_i = \frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_{it}.$$

Since all the variables in (16) are premultiplied by $\mathbf{M_D}$, it follows that this estimator makes use only of the information in the variation around the mean for each of the m groups. Hence, $\hat{\boldsymbol{\beta}}_{\text{WG}}$ is called the within-groups estimator.

Since the OLS and fixed-effect estimation assumptions are the same, the $\hat{\boldsymbol{\beta}}_{\text{WG}}$ is BLUE.

$\hat{\boldsymbol{\beta}}_{\text{WG}}$ has advantages and disadvantages. It is easy to compute, even when m is large, because it is never necessary to make direct use of the $n \times n$ matrix $\mathbf{M_D}$. We just need to compute the m group means for each variable. Also, the estimates of $\hat{\boldsymbol{\eta}}$ may themselves be of interest.

However, the estimator cannot be used with explanatory variables that take on the same value for all the observations of each group, because such a column would be collinear with the columns of $\mathbf{D}$. More generally, if the explanatory variables in $\mathbf{X}$ are well explained by the dummy variables in $\mathbf{D}$, then $\boldsymbol{\beta}$ is not estimated at all precisely.

2.6.2 Random effects estimation

The random-effects estimator uses the following assumption. $\mathbf{X}$ and the cross-sectional errors, v_i, should both be independent of the ϵ_{it}, but this does not rule out the possibility of a correlation between them. Thus,

$$\mathbb{E}[u_{it}|\mathbf{X}] = \mathbb{E}[v_i + \epsilon_{it}|\mathbf{X}] = 0, \tag{17}$$

i.e., both v_i and ϵ_{it} are independent of $\mathbf{X}$. Condition (17) is precisely the condition which ensures that OLS estimation of the model (14),

$$y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + u_{it},$$

yields unbiased estimates, rather than the model (15),

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\eta} + \boldsymbol{\epsilon}.$$

But the estimation of (14) is, in general, not efficient, because the u_{it} are not IID. We can calculate the variance-covariance matrix of the u_{it} by assuming that the v_i are IID $\sim (0, \sigma_v^2)$. This is why we call this "random effects" estimation.

Remember that we have $e_t = 0$ for all t, so:

$$\begin{aligned} \text{Var}(u_{it}) &= \text{Var}(v_i) + \text{Var}(\epsilon_{it}) = \sigma_v^2 + \sigma_\epsilon^2, \\ \text{Cov}(u_{it}, u_{it-s}) &= \sigma_v^2, \forall s, \\ \text{Cov}(u_{it}, u_{js}) &= 0, \forall i \neq j. \end{aligned}$$

Therefore the $n \times n$ variance-covariance matrix, $\mathbb{E}[\mathbf{u}\mathbf{u}^\top|\mathbf{X}] = \boldsymbol{\Omega}$, where the data is ordered by the cross-sectional units, is

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{\Sigma} \end{bmatrix},$$

where

$$\boldsymbol{\Sigma} = \sigma_v^2 \boldsymbol{\iota}_T \boldsymbol{\iota}_T^\top + \sigma_\epsilon^2 \mathbf{I}_T,$$

is a $T \times T$ matrix with $\sigma_v^2 + \sigma_\epsilon^2$ along its diagonal and σ_v^2 everywhere else. Here, $\boldsymbol{\iota}_T$ is the T-unit vector. One can show that

$$\boldsymbol{\Sigma}^{-\frac{1}{2}} = \frac{1}{\sigma_\epsilon}(\mathbf{I}_T - \lambda \mathbf{P}_{\boldsymbol{\iota}}),$$

where

$$\lambda = 1 - \left(T\frac{\sigma_v^2}{\sigma_\epsilon^2} + 1 \right)^{-\frac{1}{2}}.$$

Also, note that $\mathbf{P}_{\boldsymbol{\iota}}$ is the diagonal of $\mathbf{P}_{\mathbf{D}}$. Since $\frac{1}{\sigma_\epsilon}$ is a scalar, we need to just find an estimator for the transformation matrix

$$\mathbf{I}_n - \lambda \mathbf{P}_{\mathbf{D}},$$

which depends on σ_ϵ^2 and σ_v^2.

To get the GLS estimates of $\boldsymbol{\beta}$, we would need to know the values of σ_ϵ^2 and σ_v^2, or, at least, the value of their ratio. This is because GLS estimation

requires only that $\boldsymbol{\Omega}$ be specified up to a factor. To get FGLS estimates, we need a consistent estimate of that ratio. Notice that we haven't discussed too much about asymptotic concepts thus far. This is because in order to get definite results, we need to specify what happens to both m and T when $n = mT \to \infty$.

Consider the fixed-effects model, (15). If m remains fixed as $T \to \infty$, then the number of regressors also remains fixed as $n \to \infty$, and standard asymptotic theory applies. But if T remains fixed as $m \to \infty$, then the number of parameters to be estimated tends to infinity, and the m-vector, $\hat{\boldsymbol{\eta}}$, of estimates of the fixed effects is not consistent.[1]

But it is always possible to find a consistent estimate of σ_ϵ^2 by estimating the fixed-effects model, (15). This is because no matter how m and T may behave as $n \to \infty$, there are n residuals. Thus, if we divide the SSR from (15) by $n-m-k$, we obtain an unbiased and consistent estimate of σ_ϵ^2, since the error terms of this model are just the ϵ_{it}.

So, let us first estimate σ_ϵ^2. Our linear regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \underbrace{\mathbf{D}\boldsymbol{\eta} + \boldsymbol{\epsilon}}_{\mathbf{u}},$$

and via the FWL theorem the OLS residuals for $\boldsymbol{\epsilon}$ from the above regression and from

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

are numerically the same. So we can compute $\hat{\boldsymbol{\epsilon}}$ as

$$\hat{\boldsymbol{\epsilon}} = \mathbf{M_D y} - \mathbf{M_D X}\hat{\boldsymbol{\beta}}_{\text{FE}},$$

and the estimator for σ_ϵ^2 is:

$$\hat{\sigma}_\epsilon^2 = \frac{\hat{\boldsymbol{\epsilon}}^\top \hat{\boldsymbol{\epsilon}}}{n-(k+m)}. \tag{18}$$

But getting an estimate for σ_v^2 is not as simple. The natural estimator of σ_v^2, namely the sample variance of the m elements of $\hat{\boldsymbol{\eta}}$, is not consistent unless $m \to \infty$. Therefore, we should not use the random-effects estimator if m is small. For estimating σ_v^2 we will need the auxiliary regression

$$\mathbf{P_D y} = \mathbf{P_D X}\boldsymbol{\beta} + \underbrace{\mathbf{P_D u}}_{\mathbf{r}},$$

[1] Interestingly, one can show that $\hat{\boldsymbol{\beta}}$ is still consistent under these circumstances, however. Sufficient conditions for this are that $\text{plim}_{m\to\infty} m^{-1}\sum_{i=1}^m \mathbf{X}_i^\top \mathbf{M}_\iota \mathbf{X}_i$ is positive definite (for invertibility) and that $\text{plim}_{m\to\infty} m^{-1}\sum_{i=1}^m \mathbf{X}_i^\top \mathbf{M}_\iota \boldsymbol{\epsilon}_i = \mathbf{0}$.

where $\mathbf{P_D} \equiv \mathbf{I} - \mathbf{M_D}$, so as to obtain the "between-groups" estimator

$$\hat{\boldsymbol{\beta}}_{\text{BG}} = (\mathbf{X}^\top \mathbf{P_D} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P_D} \mathbf{y}. \tag{19}$$

Although $\mathbf{P_D u}$ is $n \times 1$ in dimension (remember, $n = mT$), we really only have m observations:

$$\mathbf{P_D y} = \begin{bmatrix} \bar{\mathbf{u}}_1 \\ \vdots \\ \bar{\mathbf{u}}_1 \\ \bar{\mathbf{u}}_2 \\ \vdots \\ \bar{\mathbf{u}}_2 \\ \vdots \\ \bar{\mathbf{u}}_m \\ \vdots \\ \bar{\mathbf{u}}_m \end{bmatrix}, \ \bar{\mathbf{u}}_i = \frac{1}{T} \sum_{t=1}^{T} u_{it},$$

i.e., this is because the regressand and all the regressors are the same for every observation in each group. The estimator $\hat{\boldsymbol{\beta}}_{\text{BG}}$ uses only the variation among the group means, hence why it is termed as the between-groups estimator. Note that if $m < k$, then (19) does not exist, since $\mathbf{X}^\top \mathbf{P_D} \mathbf{X}$ can have rank at most m.

Next,

$$\text{Var}(\bar{\mathbf{u}}_i) = \text{Var}\left(\frac{1}{T} \sum_{t=1}^{T} v_i + \frac{1}{T} \sum_{t=1}^{T} \epsilon_{it} \right) = \sigma_v^2 + \frac{1}{T} \sigma_\epsilon^2 = \sigma_r^2,$$

hence we can estimate σ_r^2 by

$$\hat{\sigma}_r^2 = \frac{\hat{\mathbf{r}}^\top \hat{\mathbf{r}}}{(m-k)},$$

where

$$\hat{\mathbf{r}} = \mathbf{P_D y} - \mathbf{P_D X} \hat{\boldsymbol{\beta}}_{\text{BG}}.$$

We note that in this case that the OLS estimator is a matrix-weighted average estimator:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{P_D} + \mathbf{M_D}) \mathbf{y} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P_D} \mathbf{X} \hat{\boldsymbol{\beta}}_{\text{BG}} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{M_D} \mathbf{X} \hat{\boldsymbol{\beta}}_{\text{WG}}. \end{aligned} \tag{20}$$

Finally, we estimate σ_v^2 as

$$\hat{\sigma}_v^2 = \hat{\sigma}_r^2 - \frac{1}{T} \hat{\sigma}_\epsilon^2,$$

and the FGLS "random effects" estimator is the OLS estimator of the following transformed regression

$$(\mathbf{I}_n - \hat{\boldsymbol{\lambda}}\mathbf{P_D})\mathbf{y} = (\mathbf{I}_n - \hat{\boldsymbol{\lambda}}\mathbf{P_D})\mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \hat{\boldsymbol{\lambda}}\mathbf{P_D})\mathbf{u}.$$

In other words

$$\hat{\boldsymbol{\beta}}_{\text{RE}} = \left(\mathbf{X}^\top(\mathbf{I}_n - \hat{\boldsymbol{\lambda}}\mathbf{P_D})^\top(\mathbf{I}_n - \hat{\boldsymbol{\lambda}}\mathbf{P_D})\mathbf{X}\right)^{-1}\mathbf{X}^\top(\mathbf{I}_n - \hat{\boldsymbol{\lambda}}\mathbf{P_D})^\top(\mathbf{I}_n - \hat{\boldsymbol{\lambda}}\mathbf{P_D})\mathbf{y}, \quad (21)$$

where

$$\hat{\lambda} = 1 - \left(T\frac{\hat{\sigma}_v^2}{\hat{\sigma}_\epsilon^2} + 1\right)^{-\frac{1}{2}}.$$