Chapter 4

Generalized Least Squares Theory

In Section 3.6 we have seen that the classical conditions need not hold in practice. Although these conditions have no effect on the OLS method per se, they do affect the properties of the OLS estimators and resulting test statistics. In particular, when the elements of $\boldsymbol{y}$ have unequal variances and/or are correlated, $\mathrm{var}(\boldsymbol{y})$ is no longer a scalar variance-covariance matrix, and hence there is no guarantee that the OLS estimator is the most efficient within the class of linear unbiased (or the class of unbiased) estimators. Moreover, hypothesis testing based on the standard OLS estimator of the variance-covariance matrix becomes invalid. In practice, we hardly know the true properties of $\boldsymbol{y}$. It is therefore important to consider estimation that is valid when $\mathrm{var}(\boldsymbol{y})$ has a more general form.

In this chapter, the method of *generalized least squares* (GLS) is introduced to improve upon estimation efficiency when $\mathrm{var}(\boldsymbol{y})$ is not a scalar variance-covariance matrix. A drawback of the GLS method is that it is difficult to implement. In practice, certain structures (assumptions) must be imposed on $\mathrm{var}(\boldsymbol{y})$ so that a feasible GLS estimator can be computed. This approach results in two further difficulties, however. First, the postulated structures on $\mathrm{var}(\boldsymbol{y})$ need not be correctly specified. Consequently, the resulting feasible GLS estimator may not be as efficient as one would like. Second, the finite-sample properties of feasible GLS estimators are not easy to establish. Consequently, exact tests based on the feasible GLS estimation results are not readily available. More detailed discussion of the GLS theory can also be found in e.g., Amemiya (1985) and Greene (2000).

4.1   The Method of Generalized Least Squares

4.1.1   When $\boldsymbol{y}$ Does Not Have a Scalar Covariance Matrix

Given the linear specification (3.1):

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

suppose that, in addition to the conditions [A1] and [A2](i),

$$\text{var}(\boldsymbol{y}) = \boldsymbol{\Sigma}_o,$$

where $\boldsymbol{\Sigma}_o$ is a positive definite matrix but cannot be written as $\sigma_o^2 \boldsymbol{I}_T$ for any positive number σ_o^2. That is, the elements of $\boldsymbol{y}$ may not have a constant variance, nor are they required to be uncorrelated. As [A1] and [A2](i) remains valid, the OLS estimator $\hat{\boldsymbol{\beta}}_T$ is still unbiased by Theorem 3.4(a). It is also straightforward to verify that, in contrast with Theorem 3.4(c), the variance-covariance matrix of the OLS estimator is

$$\text{var}(\hat{\boldsymbol{\beta}}_T) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_o\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{4.1}$$

In view of Theorem 3.5, there is no guarantee that the OLS estimator is the BLUE for $\boldsymbol{\beta}_o$. Similarly, when [A3] fails such that

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}_o, \boldsymbol{\Sigma}_o),$$

we have

$$\hat{\boldsymbol{\beta}}_T \sim \mathcal{N}\big(\boldsymbol{\beta}_o, \, (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_o\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\big);$$

cf. Theorem 3.7(a). In this case, the OLS estimator $\hat{\boldsymbol{\beta}}_T$ need not be the BUE for $\boldsymbol{\beta}_o$.

Apart from efficiency, a more serious consequence of the failure of [A3] is that the statistical tests based on the standard OLS estimation results become invalid. Recall that the OLS estimator for $\text{var}(\hat{\boldsymbol{\beta}}_T)$ is

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}_T) = \hat{\sigma}_T^2 (\boldsymbol{X}'\boldsymbol{X})^{-1},$$

which is, in general, a biased estimator for (4.1). As the t and F statistics depend on the elements of $\widehat{\text{var}}(\hat{\boldsymbol{\beta}}_T)$, they no longer have the desired t and F distributions under the null hypothesis. Consequently, the inferences based on these tests become invalid.

4.1.2   The GLS Estimator

The GLS method focuses on the efficiency issue resulted from the failure of the classical condition [A2](ii). Let $\boldsymbol{G}$ be a $T \times T$ non-stochastic matrix. Consider the "transformed" specification

$$\boldsymbol{G}\boldsymbol{y} = \boldsymbol{G}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{G}\boldsymbol{e},$$

where $\boldsymbol{G}\boldsymbol{y}$ denotes the transformed dependent variable and $\boldsymbol{G}\boldsymbol{X}$ is the matrix of transformed explanatory variables. It can be seen that $\boldsymbol{G}\boldsymbol{X}$ also has full column rank k provided that $\boldsymbol{G}$

is nonsingular. Thus, the identification requirement for the specification (3.1) carries over under nonsingular transformations. It follows that $\boldsymbol{\beta}$ can still be estimated by the OLS method using these transformed variables. The resulting OLS estimator is

$$\boldsymbol{b}(\boldsymbol{G}) = (\boldsymbol{X}'\boldsymbol{G}'\boldsymbol{G}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{G}'\boldsymbol{G}\boldsymbol{y}, \tag{4.2}$$

where the notation $\boldsymbol{b}(\boldsymbol{G})$ indicates that this estimator is a function of $\boldsymbol{G}$.

Given that the original variables $\boldsymbol{y}$ and $\boldsymbol{X}$ satisfy [A1] and [A2](i), it is easily seen that the transformed variables $\boldsymbol{G}\boldsymbol{y}$ and $\boldsymbol{G}\boldsymbol{X}$ also satisfy these two conditions because $\boldsymbol{G}\boldsymbol{X}$ is still non-stochastic and $\mathbb{E}(\boldsymbol{G}\boldsymbol{y}) = \boldsymbol{G}\boldsymbol{X}\boldsymbol{\beta}_o$. Thus, the estimator $\boldsymbol{b}(\boldsymbol{G})$ must be unbiased for any nonstochastic and nonsingular $\boldsymbol{G}$. A natural question then is: Can we find a transformation matrix that yields the most efficient estimator among all linear unbiased estimators? Observe that when $\mathrm{var}(\boldsymbol{y}) = \boldsymbol{\Sigma}_o$,

$$\mathrm{var}(\boldsymbol{G}\boldsymbol{y}) = \boldsymbol{G}\boldsymbol{\Sigma}_o\boldsymbol{G}'.$$

If $\boldsymbol{G}$ is such that $\boldsymbol{G}\boldsymbol{\Sigma}_o\boldsymbol{G}' = \sigma_o^2\boldsymbol{I}_T$ for some positive number σ_o^2, the condition [A2](ii) would also hold for the transformed specification. Given this $\boldsymbol{G}$, it is now readily seen that the OLS estimator (4.2) is the BLUE for $\boldsymbol{\beta}_o$ by Theorem 3.5. This shows that, as far as efficiency is concerned, one should choose $\boldsymbol{G}$ as a nonstochastic and nonsingular matrix such that $\boldsymbol{G}\boldsymbol{\Sigma}_o\boldsymbol{G}' = \sigma_o^2\boldsymbol{I}_T$.

To find the desired transformation matrix G, note that $\boldsymbol{\Sigma}_o$ is symmetric and positive definite so that it can be orthogonally diagonalized as $\boldsymbol{C}'\boldsymbol{\Sigma}_o\boldsymbol{C} = \boldsymbol{\Lambda}$, where $\boldsymbol{C}$ is the matrix of eigenvectors corresponding to the matrix of eigenvalues $\boldsymbol{\Lambda}$. For $\boldsymbol{\Sigma}_o^{-1/2} = \boldsymbol{C}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}'$ (or $\boldsymbol{\Sigma}_o^{-1/2} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}'$), we have

$$\boldsymbol{\Sigma}_o^{-1/2}\boldsymbol{\Sigma}_o\boldsymbol{\Sigma}_o^{-1/2\prime} = \boldsymbol{I}_T.$$

This result immediately suggests that $\boldsymbol{G}$ should be proportional to $\boldsymbol{\Sigma}_o^{-1/2}$, i.e., $\boldsymbol{G} = c\,\boldsymbol{\Sigma}_o^{-1/2}$ for some constant c. Given this choice of G, we have

$$\mathrm{var}(\boldsymbol{G}\boldsymbol{y}) = \boldsymbol{G}\boldsymbol{\Sigma}_o\boldsymbol{G}' = c^2\boldsymbol{I}_T,$$

a scalar covariance matrix, so that [A2](ii) holds. The estimator (4.2) with $\boldsymbol{G} = c\,\boldsymbol{\Sigma}_o^{-1/2}$ is known as the GLS estimator and can be expressed as

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = (c^2\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}(c^2\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{y}) = (\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{y}. \tag{4.3}$$

This estimator is, by construction, the BLUE for $\boldsymbol{\beta}_o$ under [A1] and [A2](i). The GLS and OLS estimators are not equivalent in general, except in some exceptional cases; see, e.g., Exercise 4.1.

As the GLS estimator does not depend on c, it is without loss of generality to set $\boldsymbol{G} = \boldsymbol{\Sigma}_o^{-1/2}$. Given this choice of $\boldsymbol{G}$, let $\boldsymbol{y}^* = \boldsymbol{G}\boldsymbol{y}$, $\boldsymbol{X}^* = \boldsymbol{G}\boldsymbol{X}$, and $e^* = \boldsymbol{G}e$. The transformed specification is

$$\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{e}^*, \tag{4.4}$$

and the OLS estimator for this specification is the GLS estimator (4.3). Clearly, the GLS estimator is a minimizer of the following GLS criterion function:

$$Q(\boldsymbol{\beta}; \boldsymbol{\Sigma}_o) = \frac{1}{T}(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta})'(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}) = \frac{1}{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{4.5}$$

This criterion function is the average of a *weighted* sum of squared errors and hence a generalized version of the standard OLS criterion function (3.2).

Similar to the OLS method, define the vector of GLS fitted values as

$$\hat{\boldsymbol{y}}_{\mathrm{GLS}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{y}.$$

The vector of GLS residuals is

$$\hat{\boldsymbol{e}}_{\mathrm{GLS}} = \boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathrm{GLS}}.$$

The fact that $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}$ is idempotent but not symmetric immediately implies that $\hat{\boldsymbol{y}}_{\mathrm{GLS}}$ is an *oblique* (but not orthogonal) projection of $\boldsymbol{y}$ onto span$(\boldsymbol{X})$. It can also be verified that the vector of GLS residuals is not orthogonal to $\boldsymbol{X}$ or any linear combination of the column vectors of $\boldsymbol{X}$; i.e.,

$$\hat{\boldsymbol{e}}_{\mathrm{GLS}}'\boldsymbol{X} = \boldsymbol{y}'[\boldsymbol{I}_T - \boldsymbol{\Sigma}_o^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{X} \neq \boldsymbol{0}.$$

In fact, $\hat{\boldsymbol{e}}_{\mathrm{GLS}}$ is orthogonal to span$(\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})$. It follows that

$$\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} \leq \hat{\boldsymbol{e}}_{\mathrm{GLS}}'\hat{\boldsymbol{e}}_{\mathrm{GLS}}.$$

That is, the OLS method still yields a better fit of original data.

4.1.3   Properties of the GLS Estimator

We have seen that the GLS estimator is, by construction, the BLUE for $\boldsymbol{\beta}_o$ under [A1] and [A2](i). Its variance-covariance matrix is

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}) = \mathrm{var}\big((\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{y}\big) = (\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}. \tag{4.6}$$

These results are summarized below.

Theorem 4.1 (Aitken) *Given the specification* (3.1), *suppose that* [A1] *and* [A2](i) *hold and that* $\mathrm{var}(\boldsymbol{y}) = \boldsymbol{\Sigma}_o$ *is a positive definite matrix. Then* $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ *is the BLUE for* $\boldsymbol{\beta}_o$ *with the variance-covariance matrix* $(\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}$.

It follows from Theorem 4.1 that $\mathrm{var}(\hat{\boldsymbol{\beta}}_T) - \mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}})$ must be a positive semi-definite matrix. This result can also be verified directly; see Exercise 4.3.

For convenience, we introduce the following condition.

[**A3'**] $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}_o, \boldsymbol{\Sigma}_o)$, where $\boldsymbol{\Sigma}_o$ is a positive definite matrix.

The following result is an immediate consequence of Theorem 3.7(a).

Theorem 4.2 *Given the specification (3.1), suppose that* [A1] *and* [A3'] *hold. Then*

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} \sim \mathcal{N}\big(\boldsymbol{\beta}_o, (\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}\big).$$

Moreover, if we believe that [A3'] is true, the log-likelihood function is

$$\log L(\boldsymbol{\beta}; \boldsymbol{\Sigma}_o) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\log(\det(\boldsymbol{\Sigma}_o)) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \qquad (4.7)$$

The first order condition of maximizing this log-likelihood function with respect to $\boldsymbol{\beta}$ is

$$\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0},$$

so that the MLE is

$$\tilde{\boldsymbol{\beta}}_T = (\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{y},$$

which is exactly the same as the GLS estimator. The information matrix is

$$\mathbb{E}[\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X}]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_o} = \boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X},$$

and its inverse (the Crámer-Rao lower bound) is the variance-covariance matrix of the GLS estimator. We have established the following result.

Theorem 4.3 *Given the specification (3.1), suppose that* [A1] *and* [A3'] *hold. Then* $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ *is the BUE for* $\boldsymbol{\beta}_o$.

Under the null hypothesis $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r}$, it is readily seen from Theorem 4.2 that

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} - \boldsymbol{r}) \sim \chi^2(q).$$

The left-hand side above can serve as a test statistic for the linear hypothesis $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r}$.

4.1.4   FGLS Estimator

In practice, $\boldsymbol{\Sigma}_o$ is typically unknown so that the GLS estimator is not available. Substituting an estimator $\widehat{\boldsymbol{\Sigma}}_T$ for $\boldsymbol{\Sigma}_o$ in (4.3) yields the *feasible generalized least squares* (FGLS) estimator

$$\hat{\boldsymbol{\beta}}_{\mathrm{FGLS}} = (\boldsymbol{X}'\widehat{\boldsymbol{\Sigma}}_T^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{\Sigma}}_T^{-1}\boldsymbol{y}.$$

which is readily computed from data. Note, however, that $\boldsymbol{\Sigma}_o$ contains too many ($T(T + 1)/2$) parameters. Proper estimation of $\boldsymbol{\Sigma}_o$ would not be possible unless further restrictions on the elements of $\boldsymbol{\Sigma}_o$ are imposed. Under different assumptions on $\mathrm{var}(\boldsymbol{y})$, $\boldsymbol{\Sigma}_o$ has a simpler structure with much fewer (say, $p \ll T$) unknown parameters and may be properly estimated; see Sections 4.2 and 4.3. The properties of FGLS estimation crucially depend on these assumptions.

A clear disadvantage of the FGLS estimator is that its finite sample properties are usually unknown. Note that $\widehat{\boldsymbol{\Sigma}}_T$ is, in general, a function of $\boldsymbol{y}$, so that $\hat{\boldsymbol{\beta}}_{\mathrm{FGLS}}$ is a complex function of the elements of $\boldsymbol{y}$. It is therefore difficult, if not impossible, to derive the finite-sample properties, such as expectation and variance, of $\hat{\boldsymbol{\beta}}_{\mathrm{FGLS}}$. Consequently, the efficiency gain of an FGLS estimator is not at all clear. Deriving exact tests based on the FGLS estimator is also a formidable job. One must rely on the asymptotic properties of $\hat{\boldsymbol{\beta}}_{\mathrm{FGLS}}$ to draw statistical inferences.

4.2   Heteroskedasticity

In this section, we consider a simpler structure of $\boldsymbol{\Sigma}_o$ such that $\boldsymbol{\Sigma}_o$ is diagonal with possibly different diagonal elements:

$$\boldsymbol{\Sigma}_o = \mathrm{diag}[\sigma_1^2,\ldots,\sigma_T^2] = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_T^2 \end{bmatrix}, \tag{4.8}$$

where diag is the operator that puts its arguments on the main diagonal of a matrix. Given this $\boldsymbol{\Sigma}_o$, the elements of $\boldsymbol{y}$ are uncorrelated but may have different variances. When y_t, $t = 1,\ldots,T$, have a constant variance, they are said to be *homoskedastic*; otherwise, they are *heteroskedastic*.

To compute the GLS estimator, the desired transformation matrix is

$$\boldsymbol{\Sigma}_o^{-1/2} = \begin{bmatrix} \sigma_1^{-1} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_T^{-1} \end{bmatrix}.$$

As $\boldsymbol{\Sigma}_o$ still contains T unknown parameters, an even simpler structure of $\boldsymbol{\Sigma}_o$ is needed to ensure proper FGLS estimation.

4.2.1 Tests for Heteroskedasticity

It is clear that the OLS method would prevail unless there is evidence that $\boldsymbol{\Sigma}_o \neq \sigma_o^2 \boldsymbol{I}_T$. We therefore first study the tests of the null hypothesis of *homoskedasticity* against some form of *heteroskedasticity*. Such tests are usually based on simplified parametric specifications of $\mathrm{var}(y_t)$.

The simplest possible form of heteroskedastic y_t is *groupwise heteroskedasticity*. Suppose that data can be classified into two groups: group one contains T_1 observations with the constant variance σ_1^2, and group two contains T_2 observations with the constant variance σ_2^2. This assumption simplifies $\boldsymbol{\Sigma}_o$ in (4.8) to a matrix of only two unknown parameters:

$$\boldsymbol{\Sigma}_o = \begin{bmatrix} \sigma_1^2 \boldsymbol{I}_{T_1} & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \boldsymbol{I}_{T_2} \end{bmatrix}, \tag{4.9}$$

The null hypothesis of homoskedasticity is $\sigma_1^2 = \sigma_2^2 = \sigma_o^2$; the alternative hypothesis is, without loss of generality, $\sigma_1^2 > \sigma_2^2$.

Consider now two regressions based on the observations of the group one and group two, respectively. Let $\hat{\sigma}_{T_1}^2$ and $\hat{\sigma}_{T_2}^2$ denote the resulting OLS variance estimates. Intuitively, whether $\hat{\sigma}_{T_1}^2$ is "close" to $\hat{\sigma}_{T_2}^2$ constitutes an evidence for or against the null hypothesis. Under [A1] and [A3$'$] with (4.9),

$$(T_1 - k)\hat{\sigma}_{T_1}^2/\sigma_1^2 \sim \chi^2(T_1 - k),$$

$$(T_2 - k)\hat{\sigma}_{T_2}^2/\sigma_2^2 \sim \chi^2(T_2 - k),$$

by Theorem 3.7(b). When y_t are independent, these two χ^2 random variables are also mutually independent. Note that $\hat{\sigma}_{T_1}^2$ and $\hat{\sigma}_{T_2}^2$ must be computed from separate regressions so as to ensure independence. Under the null hypothesis,

$$\varphi := \frac{\hat{\sigma}_{T_1}^2}{\hat{\sigma}_{T_2}^2} = \frac{(T_1 - k)\hat{\sigma}_{T_1}^2}{\sigma_o^2(T_1 - k)} \bigg/ \frac{(T_2 - k)\hat{\sigma}_{T_2}^2}{\sigma_o^2(T_2 - k)} \sim F(T_1 - k, T_2 - k).$$

Thus, $\varphi = \hat{\sigma}_{T_1}^2/\hat{\sigma}_{T_2}^2$ is the F test for groupwise heteroskedasticity.

More generally, the variances of y_t may be changing with the values of a particular explanatory variable, say x_j. That is, for some constant $c > 0$,

$$\sigma_t^2 = c\, x_{tj}^2;$$

the larger the magnitude of x_{tj}, the greater is σ_t^2. An interesting feature of this specification is that σ_t^2 may take distinct values for every t, yet $\boldsymbol{\Sigma}_o$ contains only one unknown parameter c. The null hypothesis is then $\sigma_t^2 = \sigma_o^2$ for all t, and the alternative hypothesis is, without loss of generality,

$$\sigma_{(1)}^2 \geq \sigma_{(2)}^2 \geq \ldots \sigma_{(T)}^2,$$

where $\sigma_{(i)}^2$ denotes the i th largest variance. The so-called *Goldfeld-Quandt test* is of the same form as the F test for groupwise heteroskedasticity but with the following data grouping procedure.

(1) Rearrange observations according to the values of some explanatory variable x_j in a descending order.

(2) Divide the rearranged data set into three groups with T_1, T_m, and T_2 observations, respectively.

(3) Drop the T_m observations in the middle group and perform separate OLS regressions using the data in the first and third groups.

(4) The statistic is the ratio of the variance estimates:

$$\hat{\sigma}_{T_1}^2 / \hat{\sigma}_{T_2}^2.$$

This test is again distributed as $F(T_1 - k, T_2 - k)$ under the null hypothesis. If the data are rearranged according to the values of x_j in an ascending order, the resulting statistic should be computed as

$$\hat{\sigma}_{T_2}^2 / \hat{\sigma}_{T_1}^2.$$

In a time-series study, if the variances are conjectured to be decreasing (increasing) over time, data rearrangement would not be needed.

Note that in computing the Goldfeld-Quandt test, dropping observations in the middle group helps to enhance the test's ability of discriminating variances in the first and third groups. As for the classification of these groups, a rule of thumb is that no more than one third of the observations should be dropped. It is also typical to set $T_1 \approx T_2$. Clearly, this test would be powerful provided that one can correctly identify the source of heteroskedasticity (i.e., the explanatory variable that determines variances). On the other hand, such a variable may not be readily available in practice.

An even more general form of heteroskedastic covariance matrix is such that the diagonal elements

$$\sigma_t^2 = h(\alpha_0 + \boldsymbol{z}_t' \boldsymbol{\alpha}_1),$$

© Chung-Ming Kuan, 2007, 2009

where h is some function and z_t is a $p \times 1$ vector of exogenous variables affecting the variances of y_t. This assumption simplifies Σ_o to a matrix of $p + 1$ unknown parameters. Tests against this class of alternatives can be derived under the likelihood framework, and their distributions can only be analyzed asymptotically. This will not be discussed until Chapter **??**.

4.2.2 GLS Estimation

If the test for groupwise heteroskedasticity rejects the null hypothesis, one might believe that Σ_o is given by (4.9). Accordingly, the specified linear specification may be written as:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix},$$

where y_1 is $T_1 \times 1$, y_2 is $T_2 \times 1$, X_1 is $T_1 \times k$, and X_2 is $T_2 \times k$. The transformed specification based on $\Sigma_o^{-1/2}$ is

$$\begin{bmatrix} y_1/\sigma_1 \\ y_2/\sigma_2 \end{bmatrix} = \begin{bmatrix} X_1/\sigma_1 \\ X_2/\sigma_2 \end{bmatrix} \beta + \begin{bmatrix} e_1/\sigma_1 \\ e_2/\sigma_2 \end{bmatrix},$$

where the transformed y_t, $t = 1, \ldots, T$, have constant variance one. It follows that the GLS and FGLS estimators are, respectively,

$$\hat{\beta}_{\mathrm{GLS}} = \left[\frac{X_1' X_1}{\sigma_1^2} + \frac{X_2' X_2}{\sigma_2^2} \right]^{-1} \left[\frac{X_1' y_1}{\sigma_1^2} + \frac{X_2' y_2}{\sigma_2^2} \right],$$

$$\hat{\beta}_{\mathrm{FGLS}} = \left[\frac{X_1' X_1}{\hat{\sigma}_1^2} + \frac{X_2' X_2}{\hat{\sigma}_2^2} \right]^{-1} \left[\frac{X_1' y_1}{\hat{\sigma}_1^2} + \frac{X_2' y_2}{\hat{\sigma}_2^2} \right],$$

where $\hat{\sigma}_{T_1}^2$ and $\hat{\sigma}_{T_2}^2$ are, again, the OLS variance estimates obtained from separate regressions using T_1 and T_2 observations, respectively. Observe that $\hat{\beta}_{\mathrm{FGLS}}$ is not a linear estimator in y so that its finite-sample properties are not known.

If the Goldfeld-Quandt test rejects the null hypothesis, one might believe that $\sigma_t^2 = c \, x_{tj}^2$ for some variable x_j. A transformed specification is then

$$\frac{y_t}{x_{tj}} = \beta_j + \beta_1 \frac{1}{x_{tj}} + \cdots + \beta_{j-1} \frac{x_{t,j-1}}{x_{tj}} + \beta_{j+1} \frac{x_{t,j+1}}{x_{tj}} + \cdots + \beta_k \frac{x_{tk}}{x_{tj}} + \frac{e_t}{x_{tj}},$$

where $\mathrm{var}(y_t/x_{tj}) = c := \sigma_o^2$. The GLS estimator is thus the OLS estimator for this transformed specification. Note that this is a very special case in which the GLS estimator can be computed without estimating the covariance matrix Σ_o. Clearly, the validity of the GLS method crucially depends on whether the explanatory variable x_j can be correctly identified.

When $\sigma_t^2 = h(\alpha_0 + \boldsymbol{z}_t'\boldsymbol{\alpha}_1)$, it is typically difficult to implement an FGLS estimator, especially when h is nonlinear. If h is the identity function, one may regress the squared OLS residuals $\hat{e}_t^2$ on $\boldsymbol{z}_t$ to obtain estimates for α_0 and $\boldsymbol{\alpha}_1$. Of course, certain constraint must be imposed to ensure the fitted values are non-negative. The finite-sample properties of this estimator are, again, difficult to analyze.

Remarks:

1. When a test for heteroskedasticity rejects the null hypothesis, there is *no* guarantee that the alternative hypothesis (say, groupwise heteroskedasticity) must be a correct description of $\mathrm{var}(y_t)$.

2. When a form of heteroskedasticity is incorrectly specified, it is also possible that the resulting FGLS estimator is less efficient than the OLS estimator.

3. As discussed in Section 4.1.3, the finite-sample properties of FGLS estimators and hence the exact tests are usually not available. One may appeal to asymptotic theory to construct proper tests.

4.3   Serial Correlation

Another leading example that $\mathrm{var}(\boldsymbol{y}) \neq \sigma_o^2 \boldsymbol{I}_T$ is when the elements of $\boldsymbol{y}$ are correlated so that the off-diagonal elements of $\boldsymbol{\Sigma}_o$ are non-zero. This phenomenon is more common in time series data, though it is not necessary so. When time series data y_t are correlated over time, they are said to exhibit *serial correlation*. For cross-section data, the correlations of y_t are usually referred to as *spatial correlation*. Our discussion in this section will concentrate on serial correlation only.

4.3.1   A Simple Model of Serial Correlation

Consider time series y_t, $t = 1, \ldots, T$, with the constant variance σ_o^2. The correlation coefficient between y_t and y_{t-i} is

$$\mathrm{corr}(y_t, y_{t-i}) = \frac{\mathrm{cov}(y_t, y_{t-i})}{\sqrt{\mathrm{var}(y_t)}\,\sqrt{\mathrm{var}(y_{t-i})}} = \frac{\mathrm{cov}(y_t, y_{t-i})}{\sigma_o^2},$$

for $i = 0, 1, 2, \ldots, t-1$; in particular, $\mathrm{corr}(y_t, y_t) = 1$. Such correlations are also known as the *autocorrelations* of y_t. Similarly, $\mathrm{cov}(y_t, y_{t-i})$, $i = 0, 1, 2, \ldots, t-1$, are known as the *autocovariances* of y_t.

A very simple specification of autocorrelation is

$$\mathrm{corr}(y_t, y_{t-i}) = \mathrm{corr}(y_t, y_{t+i}) = c^i,$$

where c is a constant such that $|c| < 1$. Hence, $\mathrm{cov}(y_t, y_{t-i}) = c^i \sigma_o^2$. In this specification, the autocorrelations (and also autocorrelations) depend only i, the time periods between two observations, but not on t, such that they decay exponentially when i increases. Hence, closer observations have higher correlations; two observations would have little correlation if they are separated by a sufficiently long period. Letting $\mathrm{corr}(y_t, y_{t-i}) = \rho_i$, we have

$$\rho_i = c\,\rho_{i-1}, \tag{4.10}$$

with $c = \rho_1$, which must be bounded between -1 and 1. It follows that $\mathrm{var}(\boldsymbol{y})$ is

$$\boldsymbol{\Sigma}_o = \sigma_o^2 \begin{bmatrix} 1 & \rho_1 & \rho_1^2 & \cdots & \rho_1^{T-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_1^{T-2} \\ \rho_1^2 & \rho_1 & 1 & \cdots & \rho_1^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_1^{T-1} & \rho_1^{T-2} & \rho_1^{T-3} & \cdots & 1 \end{bmatrix}. \tag{4.11}$$

To avoid singularity, ρ_1 cannot be ± 1.

A novel feature of this specification is that, while permitting non-zero off-diagonal elements of $\boldsymbol{\Sigma}_o$, it involves only two unknown parameters: σ_o^2 and ρ_1. The transformation matrix is then

$$\boldsymbol{\Sigma}_o^{-1/2} = \frac{1}{\sigma_o} \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -\frac{\rho_1}{\sqrt{1-\rho_1^2}} & \frac{1}{\sqrt{1-\rho_1^2}} & 0 & \cdots & 0 & 0 \\ 0 & -\frac{\rho_1}{\sqrt{1-\rho_1^2}} & \frac{1}{\sqrt{1-\rho_1^2}} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\sqrt{1-\rho_1^2}} & 0 \\ 0 & 0 & 0 & \cdots & -\frac{\rho_1}{\sqrt{1-\rho_1^2}} & \frac{1}{\sqrt{1-\rho_1^2}} \end{bmatrix}.$$

Note that this choice of $\boldsymbol{\Sigma}_o^{-1/2}$ is not symmetric. As any matrix that is a constant proportion to $\boldsymbol{\Sigma}_o^{-1/2}$ can also serve as a transformation matrix for GLS estimation, the so-called *Cochrane-Orcutt Transformation* is based on the following transformation matrix:

$$\boldsymbol{V}_o^{-1/2} = \sigma_o \sqrt{1-\rho_1^2}\, \boldsymbol{\Sigma}_o^{-1/2} = \begin{bmatrix} \sqrt{1-\rho_1^2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho_1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho_1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -\rho_1 & 1 \end{bmatrix},$$

which depends only on ρ_1.

The data from the Cochrane-Orcutt transformation are $\boldsymbol{y}^* = \boldsymbol{V}_o^{-1/2}\boldsymbol{y}$ and $\boldsymbol{X}^* = \boldsymbol{V}_o^{-1/2}\boldsymbol{X}$ with

$$y_1^* = (1 - \rho_1^2)^{1/2}y_1, \qquad \boldsymbol{x}_1^* = (1 - \rho_1^2)^{1/2}\boldsymbol{x}_1,$$
$$y_t^* = y_t - \rho_1 y_{t-1}, \qquad \boldsymbol{x}_t^* = \boldsymbol{x}_t - \rho_1 \boldsymbol{x}_{t-1}, \qquad t = 2, \cdots, T,$$

where $\boldsymbol{x}_t$ is the tth column of $\boldsymbol{X}'$. It is then clear that $\mathrm{var}(\boldsymbol{y}^*) = \sigma_o^2(1 - \rho_1^2)\boldsymbol{I}_T$. Then provided that ρ_1 is known, regressing y_t^* on $\boldsymbol{x}_t^*$ yields the GLS estimator for $\boldsymbol{\beta}_o$.

4.3.2   An Alternative View

There is an alternative approach to generating the variance-covariance matrix (4.11). Under [A2](i), let

$$\boldsymbol{\epsilon} := \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o$$

denote the deviations of $\boldsymbol{y}$ from its mean vector. This vector is usually referred to as the vector of *disturbances*. Note that $\boldsymbol{\epsilon}$ is not the same as the residual vector $\hat{\boldsymbol{e}}$. While the former is not observable because $\boldsymbol{\beta}_o$ is unknown, the later is obtained from OLS estimation and hence observable. Under [A2], $\mathbb{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and

$$\mathrm{var}(\boldsymbol{y}) = \mathrm{var}(\boldsymbol{\epsilon}) = \mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}').$$

The variance and covariance structure of $\boldsymbol{y}$ is thus the same as that of $\boldsymbol{\epsilon}$.

A time series is said to be *weakly stationary* if its mean, variance, and autocovariances are all independent of the time index t. Thus, a weakly stationary series can not exhibit trending behavior and/or large fluctuations. In particular, a time series with zero mean, a constant variance, and zero autocovariances is weakly stationary; such a process is also known as a *white noise*. Let $\{u_t\}$ be a white noise with $\mathbb{E}(u_t) = 0$, $\mathbb{E}(u_t^2) = \sigma_u^2$, and $\mathbb{E}(u_t u_\tau) = 0$ for $t \neq \tau$. Now suppose that the elements of $\boldsymbol{\epsilon}$ is generated as a weakly stationary AR(1) process (*autoregressive process of order 1*):

$$\epsilon_t = \alpha_1 \epsilon_{t-1} + u_t, \tag{4.12}$$

where the process can be traced back to infinite past. Then, recursive substitution of (4.12) yields

$$\epsilon_t = \sum_{i=0}^{\infty} \alpha_1^i u_{t-i}, \tag{4.13}$$

a weighted sum of the current and all past random innovations (shocks).

It follows from (4.13) that $\mathbb{E}(\epsilon_t) = 0$,

$$\text{var}(\epsilon_t) = \sum_{i=0}^{\infty} \alpha_1^{2i} \sigma_u^2 = \sigma_u^2/(1 - \alpha_1^2).$$

Clearly, the right-hand side would not be meaningful unless $|\alpha_1| < 1$. The autocovariance of ϵ_t and ϵ_{t-1} is, by weak stationarity,

$$\mathbb{E}(\epsilon_t \epsilon_{t-1}) = \alpha_1 \, \mathbb{E}(\epsilon_{t-1}^2) = \alpha_1 \frac{\sigma_u^2}{1 - \alpha_1^2}.$$

This shows that

$$\alpha_1 = \text{corr}(\epsilon_t, \epsilon_{t-1}) = \text{corr}(y_t, y_{t-1}) = \rho_1.$$

Similarly,

$$\mathbb{E}(\epsilon_t \epsilon_{t-2}) = \alpha_1 \, \mathbb{E}(\epsilon_{t-1} \epsilon_{t-2}) = \alpha_1^2 \frac{\sigma_u^2}{1 - \alpha_1^2},$$

so that $\text{corr}(\epsilon_t, \epsilon_{t-2}) = \alpha_1 \, \text{corr}(\epsilon_t, \epsilon_{t-1}) = \rho_1^2$. More generally, we can write for $i = 1, 2, \ldots$,

$$\text{corr}(\epsilon_t, \epsilon_{t-i}) = \rho_1 \, \text{corr}(\epsilon_t, \epsilon_{t-i+1}) = \rho_1^i,$$

which depend only on i, the time difference between two ϵ's, but not on t. This is precisely what we postulated in (4.10). The variance-covariance matrix $\boldsymbol{\Sigma}_o$ under this structure is also (4.11), with $\sigma_o^2 = \sigma_u^2/(1 - \rho_1^2)$.

The AR(1) structure of disturbances permits a straightforward extension. Consider the disturbances that are generated as an AR(p) process (autoregressive process of order p):

$$\epsilon_t = \alpha_1 \epsilon_{t-1} + \cdots + \alpha_p \epsilon_{t-p} + u_t, \tag{4.14}$$

where the coefficients $\alpha_1, \ldots, \alpha_p$ should be restricted to ensure weak stationarity; we omit the details. Of course, ϵ_t may follow different structures; for example, ϵ_t may be generated as an MA(1) process (*moving average process of order 1*):

$$\epsilon_t = u_t + \alpha_1 u_{t-1}, \qquad |\alpha_1| < 1,$$

where $\{u_t\}$ is a white noise. The autocorrelations of this process can be easily derived; see Exercise 4.6.

4.3.3   Tests for AR(1) Disturbances

As the AR(1) structure of disturbances is one of the most commonly used specification of serial correlation, we now consider the tests of the null hypothesis of no serial correlation ($\alpha_1 = \rho_1 = 0$) against AR(1) disturbances. We discuss only the celebrated Durbin-Watson test and Durbin's h test; the discussion of other large-sample tests will be deferred to Chapter **??**.

In view of the AR(1) structure, a natural estimator of ρ_1 is the OLS estimator of regressing the OLS residual $\hat{e}_t$ on its immediate lag $\hat{e}_{t-1}$:

$$\hat{\rho}_T = \frac{\sum_{t=2}^{T} \hat{e}_t \hat{e}_{t-1}}{\sum_{t=2}^{T} \hat{e}_{t-1}^2}. \tag{4.15}$$

The Durbin-Watson statistic is

$$d = \frac{\sum_{t=2}^{T} (\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^{T} \hat{e}_t^2}.$$

When the sample size T is large, it can be seen that

$$d = 2 - 2\hat{\rho}_T \frac{\sum_{t=2}^{T} \hat{e}_{t-1}^2}{\sum_{t=1}^{T} \hat{e}_t^2} - \frac{\hat{e}_1^2 + \hat{e}_T^2}{\sum_{t=1}^{T} \hat{e}_t^2}$$

$$\approx 2(1 - \hat{\rho}_T).$$

For $0 < \hat{\rho}_T \leq 1$ ($-1 \leq \hat{\rho}_T < 0$), the Durbin-Watson statistic is such that $0 \leq d < 2$ ($2 < d \leq 4$), which suggests that there is some positive (negative) serial correlation. Hence, this test essentially checks whether $\hat{\rho}_T$ is sufficiently "close" to zero (i.e., d is close to 2).

A major difficulty of the Durbin-Watson test is that the exact null distribution of d depends on the matrix $\boldsymbol{X}$ and therefore varies with data. (Recall that the t and F tests discussed in Section 3.3.1 have, respectively, t and F distributions regardless of the data $\boldsymbol{X}$.) This drawback prevents us from tabulating the critical values of d. Nevertheless, it has been shown that the null distribution of d lies between the distributions of a lower bound (d_L) and an upper bound (d_U) in the sense that for any significance level α,

$$d_{L,\alpha}^* < d_{\alpha}^* < d_{U,\alpha}^*,$$

where let d_{α}^*, $d_{L,\alpha}^*$ and $d_{U,\alpha}^*$ denote, respectively, the critical values of d, d_L and d_U. Although the distribution of d is data dependent, the distributions of d_L and d_U are not. Thus, the critical values $d_{L,\alpha}^*$ and $d_{U,\alpha}^*$ can be tabulated. One may then rely on these critical values to construct a "conservative" decision rule.

Specifically, when the alternative hypothesis is $\rho_1 > 0$ ($\rho_1 < 0$), the decision rule of the Durbin-Watson test is:

(1) Reject the null if $d < d_{L,\alpha}^*$ $(d > 4 - d_{L,\alpha}^*)$.

(2) Do not reject the null if $d > d_{U,\alpha}^*$ $(d < 4 - d_{U,\alpha}^*)$.

(3) Test is inconclusive if $d_{L,\alpha}^* < d < d_{U,\alpha}^*$ $(4 - d_{L,\alpha}^* > d > 4 - d_{U,\alpha}^*)$.

This is not completely satisfactory because the test may yield no conclusion. Some econometric packages such as SHAZAM now compute the exact Durbin-Watson distribution for each regression and report the exact p-values. When such a program is available, this test does not have to rely on the critical values of d_L and d_U and hence must be conclusive. Note that the tabulated critical values of the Durbin-Watson statistic are for the specifications with a constant term; the critical values for the specifications without a constant term can be found in Farebrother (1980).

Another problem with the Durbin-Watson statistic is that its null distribution holds only under the classical conditions [A1] and [A3]. In the time series context, it is quite common to include a lagged dependent variable as a regressor so that [A1] is violated. A leading example is the specification

$$y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \gamma y_{t-1} + e_t.$$

This model can also be derived from certain behavioral assumptions; see Exercise 4.7. It has been shown that the Durbin-Watson statistic under this specification is biased toward 2. That is, this test would not reject the null hypothesis even when serial correlation is present. On the other hand, Durbin's h test is designed specifically for the specifications that contain a lagged dependent variable. Let $\hat{\gamma}_T$ be the OLS estimate of γ and $\widehat{\mathrm{var}}(\hat{\gamma}_T)$ be the OLS estimate of $\mathrm{var}(\hat{\gamma}_T)$. The h statistic is

$$h = \hat{\rho}_T \sqrt{\frac{T}{1 - T\widehat{\mathrm{var}}(\hat{\gamma}_T)}},$$

and its asymptotic null distribution is $\mathcal{N}(0,1)$. A clear disadvantage of Durbin's h test is that it can not be calculated when $\widehat{\mathrm{var}}(\hat{\gamma}_T) \geq 1/T$. This test can also be derived as a Lagrange Multiplier test under the likelihood framework; see Chapter **??**

If we have quarterly data and want to test for the fourth-order serial correlation, the statistic analogous to the Durbin-Watson statistic is

$$d_4 = \frac{\sum_{t=5}^{T}(\hat{e}_t - \hat{e}_{t-4})^2}{\sum_{t=1}^{T} \hat{e}_t^2};$$

see Wallis (1972) for corresponding critical values.

4.3.4   FGLS Estimation

Recall that $\boldsymbol{\Sigma}_o$ depends on two parameters σ_o^2 and ρ_1. We may use a generic notation $\boldsymbol{\Sigma}(\sigma^2, \rho)$ to denote this function of σ^2 and ρ. In particular, $\boldsymbol{\Sigma}_o = \boldsymbol{\Sigma}(\sigma_o^2, \rho_1)$. Similarly, we may also write $\boldsymbol{V}(\rho)$ such that $\boldsymbol{V}_o = \boldsymbol{V}(\rho_1)$. The transformed data based on $\boldsymbol{V}(\rho)^{-1/2}$ are

$$y_1(\rho) = (1 - \rho^2)^{1/2} y_1, \qquad \boldsymbol{x}_1(\rho) = (1 - \rho^2)^{1/2} \boldsymbol{x}_1,$$
$$y_t(\rho) = y_t - \rho y_{t-1}, \qquad \boldsymbol{x}_t(\rho) = \boldsymbol{x}_t - \rho \boldsymbol{x}_{t-1}, \qquad t = 2, \cdots, T.$$

Hence, $y_t^* = y_t(\rho_1)$ and $\boldsymbol{x}_t^* = \boldsymbol{x}_t(\rho_1)$.

To obtain an FGLS estimator, we must first estimate ρ_1 by some estimator $\hat{\rho}_T$ and then construct the transformation matrix as $\widehat{\boldsymbol{V}}_T^{-1/2} = \boldsymbol{V}(\hat{\rho}_T)^{-1/2}$. Here, $\hat{\rho}_T$ may be computed as in (4.15); other estimators for ρ_1 may also be used, e.g., $\check{\rho}_T = \hat{\rho}_T (T - k)/(T - 1)$. The transformed data are then $y_t(\hat{\rho}_T)$ and $\boldsymbol{x}_t(\hat{\rho}_T)$. An FGLS estimator is obtained by regressing $y_t(\hat{\rho}_T)$ on $\boldsymbol{x}_t(\hat{\rho}_T)$. Such an estimator is known as the Prais-Winsten estimator or the Cochrane-Orcutt estimator when the first observation is dropped in computation.

The following iterative procedure is also commonly employed in practice.

(1) Perform OLS estimation and compute $\hat{\rho}_T$ as in (4.15) using the OLS residuals $\hat{e}_t$.

(2) Perform the Cochrane-Orcutt transformation based on $\hat{\rho}_T$ and compute the resulting FGLS estimate $\hat{\boldsymbol{\beta}}_{\mathrm{FGLS}}$ by regressing $y_t(\hat{\rho}_T)$ on $\boldsymbol{x}_t(\hat{\rho}_T)$.

(3) Compute a new $\hat{\rho}_T$ as in (4.15) with $\hat{e}_t$ replaced by the FGLS residuals

$$\hat{e}_{t,\mathrm{FGLS}} = y_t - \boldsymbol{x}_t' \hat{\boldsymbol{\beta}}_{\mathrm{FGLS}}.$$

(4) Repeat steps (2) and (3) until $\hat{\rho}_T$ converges numerically, i.e., when $\hat{\rho}_T$ from two consecutive iterations differ by a value smaller than a pre-determined convergence criterion.

Note that steps (1) and (2) above already generate an FGLS estimator. More iterations do not improve the asymptotic properties of the resulting estimator but may have a significant effect in finite samples. This procedure can be extended easily to estimate the specification with higher-order AR disturbances.

Alternatively, the Hildreth-Lu procedure adopts *grid search* to find the $\rho_1 \in (-1, 1)$ that minimizes the sum of squared errors of the model. We first set the grid points in $(-1, 1)$. For each grid point ρ, we conduct the Cochrane-Orcutt transformation and compute the resulting FGLS estimate $\hat{\boldsymbol{\beta}}_{\mathrm{FGLS}}$ by regressing $y_t(\rho)$ on $\boldsymbol{x}_t(\rho)$. A ρ is chosen if it yields the smallest sum of squared errors based on FGLS residuals. This procedure is computationally

intensive, and the result depends on how many grid points we consider. Note also that this procedure is difficult to implement when ϵ_t have an AR(p) structure with $p > 2$.

In view of the log-likelihood function (4.7), we must compute $\det(\boldsymbol{\Sigma}_o)$. Clearly,

$$\det(\boldsymbol{\Sigma}_o) = \frac{1}{\det(\boldsymbol{\Sigma}_o^{-1})} = \frac{1}{[\det(\boldsymbol{\Sigma}_o^{-1/2})]^2}.$$

In terms of the notations in the AR(1) formulation, $\sigma_o^2 = \sigma_u^2/(1 - \rho_1^2)$, and

$$\boldsymbol{\Sigma}_o^{-1/2} = \frac{1}{\sigma_o\sqrt{1 - \rho_1^2}}\boldsymbol{V}_o^{-1/2} = \frac{1}{\sigma_u}\boldsymbol{V}_o^{-1/2}.$$

As $\det(\boldsymbol{V}_o^{-1/2}) = (1 - \rho_1^2)^{1/2}$, we then have

$$\det(\boldsymbol{\Sigma}_o) = (\sigma_u^2)^T(1 - \rho_1^2)^{-1}.$$

The log-likelihood function for given σ_u^2 and ρ_1 is

$$\log L(\boldsymbol{\beta}; \sigma_u^2, \rho_1)$$
$$= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma_u^2) + \frac{1}{2}\log(1 - \rho_1^2) - \frac{1}{2\sigma_u^2}(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta})'(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}).$$

Clearly, when σ_u^2 and ρ_1 are known, the MLE of $\boldsymbol{\beta}$ is just the GLS estimator.

If σ_u^2 and ρ_1 are unknown, the log-likelihood function reads:

$$\log L(\boldsymbol{\beta}, \sigma^2, \rho)$$
$$= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) + \frac{1}{2}\log(1 - \rho^2) - \frac{1}{2\sigma^2}(1 - \rho^2)(y_1 - \boldsymbol{x}_1'\boldsymbol{\beta})^2$$
$$- \frac{1}{2\sigma^2}\sum_{t=2}^{T}[(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}) - \rho(y_{t-1} - \boldsymbol{x}_{t-1}'\boldsymbol{\beta})]^2,$$

which is a nonlinear function of the parameters. Nonlinear optimization methods (see, e.g., Section ??) are therefore needed to compute the MLEs of $\boldsymbol{\beta}$, σ^2, and ρ. For a given $\boldsymbol{\beta}$, estimating ρ by regressing $e_t(\boldsymbol{\beta}) = y_t - \boldsymbol{x}_t'\boldsymbol{\beta}$ on $e_{t-1}(\boldsymbol{\beta})$ is equivalent to maximizing the last term of the log-likelihood function above. This does not yield an MLE because the other terms involving ρ, namely,

$$\frac{1}{2}\log(1 - \rho^2) - \frac{1}{2\sigma^2}(1 - \rho^2)(y_1 - \boldsymbol{x}_1'\boldsymbol{\beta})^2,$$

have been ignored. This shows that the aforementioned iterative procedure does not result in the MLEs.

Remark: Exact tests based on FGLS estimation results are not available because the finite-sample distribution of the FGLS estimator is, again, unknown. Asymptotic theory is needed to construct proper tests.

4.4   Application: Linear Probability Model

In some applications researchers are interested in analyzing why consumers own a house or participate a particular event. The ownership or the choice of participation are typically represented by a *binary* variable that takes the values one and zero. If the dependent variable in a linear regression is binary, we will see below that both the OLS and FGLS methods are not appropriate.

Let $\boldsymbol{x}_t$ denote the t th column of $\boldsymbol{X}'$. The t th observation of the linear specification $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ can be expressed as

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t.$$

For the binary dependent variable y whose t th observation is $y_t = 1$ or 0, we know

$$\mathbb{E}(y_t) = \mathbb{P}(y_t = 1).$$

Thus, $\boldsymbol{x}_t'\boldsymbol{\beta}$ is just a specification of the probability that $y_t = 1$. As such, the linear specification of binary dependent variables is usually referred to as the *linear probability model*.

When [A1] and [A2](i) hold for a linear probability model,

$$\mathbb{E}(y_t) = \mathbb{P}(y_t = 1) = \boldsymbol{x}_t'\boldsymbol{\beta}_o,$$

and the OLS estimator is unbiased for $\boldsymbol{\beta}_o$. Note, however, that the variance of y_t is

$$\mathrm{var}(y_t) = \mathbb{P}(y_t = 1)[1 - \mathbb{P}(y_t = 1)].$$

Under [A1] and [A2](i),

$$\mathrm{var}(y_t) = \boldsymbol{x}_t'\boldsymbol{\beta}_o(1 - \boldsymbol{x}_t'\boldsymbol{\beta}_o),$$

which varies with $\boldsymbol{x}_t$. Thus, the linear probability model suffers from the problem of heteroskedasticity, and the OLS estimator is not the BLUE for $\boldsymbol{\beta}_o$. Apart from the efficiency issue, the OLS method is still not appropriate for the linear probability model because the OLS fitted values need not be bounded between zero and one. When $\boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_T$ is negative or greater than one, it can not be interpreted as a probability.

Although the GLS estimator is the BLUE, it is not available because $\boldsymbol{\beta}_o$, and hence $\mathrm{var}(y_t)$, is unknown. Nevertheless, if y_t are uncorrelated so that $\mathrm{var}(\boldsymbol{y})$ is diagonal, an FGLS estimator may be obtained using the transformation matrix

$$\widehat{\boldsymbol{\Sigma}}_T^{-1/2} = \mathrm{diag}\left[[\boldsymbol{x}_1'\hat{\boldsymbol{\beta}}_T(1 - \boldsymbol{x}_1'\hat{\boldsymbol{\beta}}_T)]^{-1/2}, [\boldsymbol{x}_2'\hat{\boldsymbol{\beta}}_T(1 - \boldsymbol{x}_2'\hat{\boldsymbol{\beta}}_T)]^{-1/2}, \dots,\right.$$

$$\left.[\boldsymbol{x}_T'\hat{\boldsymbol{\beta}}_T(1 - \boldsymbol{x}_T'\hat{\boldsymbol{\beta}}_T)]^{-1/2}\right],$$

where $\hat{\boldsymbol{\beta}}_T$ is the OLS estimator of $\boldsymbol{\beta}_o$. Such an estimator breaks down when $\widehat{\boldsymbol{\Sigma}}_T^{-1/2}$ cannot be computed (i.e., when $\boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_T$ is negative or greater than one). Even when $\widehat{\boldsymbol{\Sigma}}_T^{-1/2}$ is available, there is still no guarantee that the FGLS fitted values are bounded between zero and one. This shows that the FGLS method may not always be a solution when the OLS method fails.

This example also illustrates the importance of data characteristics in estimation and modeling. Without taking into account the binary nature of the dependent variable, even the FGLS method may be invalid. More appropriate methods for specifications with binary dependent variables will be discussed in Chapter **??**.

4.5 Seemingly Unrelated Regressions

In many econometric practices, it is important to study the joint behavior of several dependent variables. For example, the input demands of a firm may be described using a system of linear regression functions in which each regression represents the demand function of a particular input. It is intuitively clear that these input demands ought to be analyzed jointly because they are related to each other.

Consider the specification of a system of N equations, each with k_i explanatory variables and T observations. Specifically,

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{e}_i, \qquad i = 1, 2, \ldots, N, \tag{4.16}$$

where for each i, $\boldsymbol{y}_i$ is $T \times 1$, $\boldsymbol{X}_i$ is $T \times k_i$, and $\boldsymbol{\beta}_i$ is $k_i \times 1$. The system (4.16) is also known as a specification of *seemingly unrelated regressions* (SUR). Stacking the equations of (4.16) yields

$$\underbrace{\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{X}_N \end{bmatrix}}_{\boldsymbol{X}} \underbrace{\begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_N \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_N \end{bmatrix}}_{\boldsymbol{e}}. \tag{4.17}$$

This is a linear specification (3.1) with $k = \sum_{i=1}^{N} k_i$ explanatory variables and TN observations. It is not too hard to see that the whole system (4.17) satisfies the identification requirement whenever every specification of (4.16) does.

Suppose that the classical conditions [A1] and [A2] hold for each specified linear regression in the system. Then under [A2](i), there exists $\boldsymbol{\beta}_o = (\boldsymbol{\beta}_{o,1}' \ \ldots \ \boldsymbol{\beta}_{o,N}')'$ such that

$\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}_o$. The OLS estimator obtained from (4.17) is therefore unbiased. Note, however, that [A2](ii) for each linear regression ensures that, for each i,

$$\operatorname{var}(\boldsymbol{y}_i) = \sigma_i^2 \boldsymbol{I}_T;$$

there is no restriction on the correlations between $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$. The variance-covariance matrix of $\boldsymbol{y}$ is then

$$\operatorname{var}(\boldsymbol{y}) = \boldsymbol{\Sigma}_o = \begin{bmatrix} \sigma_1^2 \boldsymbol{I}_T & \operatorname{cov}(\boldsymbol{y}_1, \boldsymbol{y}_2) & \cdots & \operatorname{cov}(\boldsymbol{y}_1, \boldsymbol{y}_N) \\ \operatorname{cov}(\boldsymbol{y}_2, \boldsymbol{y}_1) & \sigma_2^2 \boldsymbol{I}_T & \cdots & \operatorname{cov}(\boldsymbol{y}_2, \boldsymbol{y}_N) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(\boldsymbol{y}_N, \boldsymbol{y}_1) & \operatorname{cov}(\boldsymbol{y}_N, \boldsymbol{y}_2) & \cdots & \sigma_N^2 \boldsymbol{I}_T \end{bmatrix}. \tag{4.18}$$

This shows that $\boldsymbol{y}$, the vector of stacked dependent variables, violates [A2](ii), even when each individual dependent variable has a scalar variance-covariance matrix. Consequently, the OLS estimator of the whole system, $\hat{\boldsymbol{\beta}}_{TN} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, is not the BLUE in general. In fact, owing to the block-diagonal structure of $\boldsymbol{X}$, $\hat{\boldsymbol{\beta}}_{TN}$ simply consists of N equation-by-equation OLS estimators and hence ignores the correlations between equations and heteroskedasticity across equations.

In practice, it is also typical to postulate that for $i \neq j$,

$$\operatorname{cov}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \sigma_{ij} \boldsymbol{I}_T,$$

that is, y_{it} and y_{jt} are contemporaneously correlated but y_{it} and $y_{j\tau}$, $t \neq \tau$, are serially uncorrelated. Under this condition, (4.18) simplifies to $\boldsymbol{\Sigma}_o = \boldsymbol{S}_o \otimes \boldsymbol{I}_T$ with

$$\boldsymbol{S}_o = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_N^2 \end{bmatrix}. \tag{4.19}$$

As $\boldsymbol{\Sigma}_o^{-1} = \boldsymbol{S}_o^{-1} \otimes \boldsymbol{I}_T$, the GLS estimator of (4.17) is

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = [\boldsymbol{X}'(\boldsymbol{S}_o^{-1} \otimes \boldsymbol{I}_T)\boldsymbol{X}]^{-1}\boldsymbol{X}'(\boldsymbol{S}_o^{-1} \otimes \boldsymbol{I}_T)\boldsymbol{y},$$

and its covariance matrix is $[\boldsymbol{X}'(\boldsymbol{S}_o^{-1} \otimes \boldsymbol{I}_T)\boldsymbol{X}]^{-1}$.

It is readily verified that when $\sigma_{ij} = 0$ for all $i \neq j$, $\boldsymbol{S}_o$ becomes a diagonal matrix, and so is $\boldsymbol{\Sigma}_o$. Then, the resulting GLS estimator for each $\boldsymbol{\beta}_i$ reduces to the corresponding OLS estimator. This should not be too surprising because estimating a SUR system would not be necessary if the dependent variables are in fact uncorrelated. (Note that the heteroskedasticity across equations does not affect this result.) If all equations in the system

have the same regressors, i.e., $\boldsymbol{X}_i = \boldsymbol{X}_0$ (say), the GLS estimator is also the same as the OLS estimator; see e.g., Exercise 4.8. More generally, it can be shown that there would not be much efficiency gain for GLS estimation if $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ are less correlated and/or $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are highly correlated; see e.g., Goldberger (1991, p. 328) for an illustrative example.

The FGLS estimator is

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = [\boldsymbol{X}'(\widehat{\boldsymbol{S}}_{TN}^{-1} \otimes \boldsymbol{I}_T)\boldsymbol{X}]^{-1}\boldsymbol{X}'(\widehat{\boldsymbol{S}}_{TN}^{-1} \otimes \boldsymbol{I}_T)\boldsymbol{y}.$$

with $\widehat{\boldsymbol{S}}_{TN}$ an $N \times N$ matrix:

$$\widehat{\boldsymbol{S}}_{TN} = \frac{1}{T}\begin{bmatrix} \hat{\boldsymbol{e}}_1' \\ \hat{\boldsymbol{e}}_2' \\ \vdots \\ \hat{\boldsymbol{e}}_N' \end{bmatrix}\begin{bmatrix} \hat{\boldsymbol{e}}_1 & \hat{\boldsymbol{e}}_2 & \dots & \hat{\boldsymbol{e}}_N \end{bmatrix},$$

where $\hat{\boldsymbol{e}}_i$ is the OLS residual vector of the ith equation. The elements of this matrix are

$$\hat{\sigma}_i^2 = \frac{\hat{\boldsymbol{e}}_i'\hat{\boldsymbol{e}}_i}{T}, \qquad i = 1, \dots, N,$$

$$\hat{\sigma}_{ij} = \frac{\hat{\boldsymbol{e}}_i'\hat{\boldsymbol{e}}_j}{T}, \qquad i \neq j, \ i, j = 1, \dots, N.$$

Note that $\widehat{\boldsymbol{S}}_{TN}$ is of an inner product form and hence a positive semi-definite matrix. When the denominator of $\hat{\sigma}_i^2$ is replaced with $T - k_i$ and the denominator of $\hat{\sigma}_{ij}$ replaced with $T - \max(k_i, k_j)$, the resulting estimator $\widehat{\boldsymbol{S}}_{TN}$ need not be positive semi-definite.

Remark: The estimator $\widehat{\boldsymbol{S}}_{TN}$ mentioned above is valid provided that $\mathrm{var}(\boldsymbol{y}_i) = \sigma_i^2 \boldsymbol{I}_T$ and $\mathrm{cov}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \sigma_{ij}\boldsymbol{I}_T$. If these assumptions do not hold, FGLS estimation would be much more complicated. This may happen when heteroskedasticity and serial correlations are present in each equation, or when $\mathrm{cov}(y_{it}, y_{jt})$ changes over time.

4.6 Models for Panel Data

A data set that contains a collection of cross-section units (individuals, families, firms, or countries), each with some time-series observations, is known as a *panel data* set. Well known panel data sets in the U.S. include the National Longitudinal Survey of Labor Market Experience and the Michigan Panel Study of Income Dynamics. Building such data sets is very costly because they are obtained by tracking thousands of individuals through time. Some panel data may be easier to establish; for example, the GDP data for all G7 countries over 50 years also form a panel data set.

Panel data contain richer information than pure cross-section or time-series data. On one hand, panel data offer a description of the dynamics of each cross-section unit. On the other hand, panel data are able to reveal the variations of dynamic patterns across individual units. Thus, panel data may render more precise parameter estimates and permit analysis of topics that could not be studied using only cross-section or time-series data. For example, to study whether each individual unit i has its own pattern, a specification with parameter(s) changing with i is needed. Such specifications can not be properly estimated using only cross-section data because the number of parameters must exceed the number of observations. The problem may be circumvented by using panel data. In what follows we will consider specifications for panel data that allow parameter to change across individual units.

4.6.1   Fixed-Effects Model

Given a panel data set with N cross-section units and T observations, the linear specification allowing for individual effects is

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta}_i + e_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T,$$

where $\boldsymbol{x}_{it}$ is $k \times 1$ and $\boldsymbol{\beta}_i$ is the parameter vector depending only on i but not on t. In this specification, individual effects are characterized by $\boldsymbol{\beta}_i$, and there is no time-specific effect. This may be reasonable when a short time series is observed for each individual unit. Analogous to the notations in the SUR system (4.16), we can express the specification above as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{e}_i, \qquad i = 1, 2, \ldots, N, \tag{4.20}$$

where $\boldsymbol{y}_i$ is $T \times 1$, $\boldsymbol{X}_i$ is $T \times k$, and $\boldsymbol{e}_i$ is $T \times 1$. This is a system of equations with $k \times N$ parameters. Here, the dependent variable y and explanatory variables X are the same across individual units such that $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$ are simply their observations for each individual i. For example, y may be the family consumption expenditure, and each $\boldsymbol{y}_i$ contains family i's annual consumption expenditures. By contrast, $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$ may be different variables in a SUR system.

When T is small (i.e., observed time series are short), estimating (4.20) is not feasible. A simpler form of (4.20) is such that only the intercept changes with i and the other parameters remain constant across i:

$$\boldsymbol{y}_i = \boldsymbol{\ell}_T a_i + \boldsymbol{Z}_i\boldsymbol{b} + \boldsymbol{e}_i, \qquad i = 1, 2, \ldots, N, \tag{4.21}$$

where $\boldsymbol{\ell}_T$ is the T-dimensional vector of ones, $[\boldsymbol{\ell}_T \ \boldsymbol{Z}_i] = \boldsymbol{X}_i$ and $[a_i \ \boldsymbol{b}']' = \boldsymbol{\beta}_i$. In (4.21), individual effects are completely captured by the intercept a_i. This specification simplifies

(4.20) from kN to $N + k - 1$ parameters and is known as the *fixed-effects model.* Stacking N equations in (4.21) together we obtain

$$
\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{\ell}_T & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\ell}_T & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{\ell}_T \end{bmatrix}}_{\boldsymbol{D}} \underbrace{\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}}_{\boldsymbol{a}} + \underbrace{\begin{bmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \\ \vdots \\ \boldsymbol{Z}_N \end{bmatrix}}_{\boldsymbol{Z}} \boldsymbol{b} + \underbrace{\begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_N \end{bmatrix}}_{\boldsymbol{e}}. \tag{4.22}
$$

Clearly, this is still a linear specification with $N + k - 1$ explanatory variables and TN observations. Note that each column of $\boldsymbol{D}$ is in effect a dummy variable for the i th individual unit. In what follows, an individual unit will be referred to as a "group."

The following notations will be used in the sequel. Let $\boldsymbol{Z}_i'$ $((k-1) \times T)$ be the i th block of $\boldsymbol{Z}'$ and $\boldsymbol{z}_{it}$ be its t th column. For $\boldsymbol{z}_{it}$, the i th group average over time is

$$
\bar{\boldsymbol{z}}_i = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{z}_{it} = \frac{1}{T} \boldsymbol{Z}_i' \boldsymbol{\ell}_T;
$$

the i th group average of y_{it} over time is

$$
\bar{y}_i = \frac{1}{T} \sum_{t=1}^{T} y_{it} = \frac{1}{T} \boldsymbol{y}_i' \boldsymbol{\ell}_T.
$$

The overall sample average of $\boldsymbol{z}_{it}$ (average over time and group) is

$$
\bar{\boldsymbol{z}} = \frac{1}{TN} \sum_{i=1}^{N} \sum_{t=1}^{T} \boldsymbol{z}_{it} = \frac{1}{TN} \boldsymbol{Z}' \boldsymbol{\ell}_{TN},
$$

and the overall sample average of y_{it} is

$$
\bar{y} = \frac{1}{TN} \sum_{i=1}^{N} \sum_{t=1}^{T} y_{it} = \frac{1}{TN} \boldsymbol{y}' \boldsymbol{\ell}_{TN}.
$$

Observe that the overall sample averages are

$$
\bar{\boldsymbol{z}} = \frac{1}{N} \sum_{i=1}^{N} \bar{\boldsymbol{z}}_i, \qquad \bar{y} = \frac{1}{N} \sum_{i=1}^{N} \bar{y}_i,
$$

which are the sample averages of group averages.

From (4.22) we can see that this specification satisfies the identification requirement [ID-1] provided that there is no time invariant regressor (i.e., no column of $\boldsymbol{Z}$ is a constant). Once this requirement is satisfied, the OLS estimator is readily computed. By Theorem 3.3, the OLS estimator for $\boldsymbol{b}$ is

$$
\hat{\boldsymbol{b}}_{TN} = [\boldsymbol{Z}'(\boldsymbol{I}_{TN} - \boldsymbol{P}_D)\boldsymbol{Z}]^{-1} \boldsymbol{Z}'(\boldsymbol{I}_{TN} - \boldsymbol{P}_D)\boldsymbol{y}, \tag{4.23}
$$

where $\boldsymbol{P}_D = \boldsymbol{D}(\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'$ is a projection matrix. Thus, $\hat{b}_{TN}$ can be obtained by regressing $(\boldsymbol{I}_{TN} - \boldsymbol{P}_D)\boldsymbol{y}$ on $(\boldsymbol{I}_{TN} - \boldsymbol{P}_D)\boldsymbol{Z}$. Let $\hat{\boldsymbol{a}}_{TN}$ denote the OLS estimator of the vector $\boldsymbol{a}$ of individual effects. By the facts that

$$\boldsymbol{D}'\hat{\boldsymbol{y}} = \boldsymbol{D}'\boldsymbol{D}\hat{\boldsymbol{a}}_{TN} + \boldsymbol{D}'\boldsymbol{Z}\hat{\boldsymbol{b}}_{TN},$$

and that the OLS residual vector is orthogonal to $\boldsymbol{D}$, $\hat{\boldsymbol{a}}_{TN}$ can be computed as

$$\hat{\boldsymbol{a}}_{TN} = (\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{b}}_{TN}). \tag{4.24}$$

We will present alternative expressions for these estimators which yield more intuitive interpretations.

Writing $\boldsymbol{D} = \boldsymbol{I}_N \otimes \boldsymbol{\ell}_T$, we have

$$\begin{aligned}
\boldsymbol{P}_D &= (\boldsymbol{I}_N \otimes \boldsymbol{\ell}_T)(\boldsymbol{I}_N \otimes \boldsymbol{\ell}_T'\boldsymbol{\ell}_T)^{-1}(\boldsymbol{I}_N \otimes \boldsymbol{\ell}_T') \\
&= (\boldsymbol{I}_N \otimes \boldsymbol{\ell}_T)[\boldsymbol{I}_N \otimes (\boldsymbol{\ell}_T'\boldsymbol{\ell}_T)^{-1}](\boldsymbol{I}_N \otimes \boldsymbol{\ell}_T') \\
&= \boldsymbol{I}_N \otimes [\boldsymbol{\ell}_T(\boldsymbol{\ell}_T'\boldsymbol{\ell}_T)^{-1}\boldsymbol{\ell}_T'] \\
&= \boldsymbol{I}_N \otimes \boldsymbol{\ell}_T\boldsymbol{\ell}_T'/T,
\end{aligned}$$

where $\boldsymbol{\ell}_T\boldsymbol{\ell}_T'/T$ is also a projection matrix. Thus,

$$\boldsymbol{I}_{TN} - \boldsymbol{P}_D = \boldsymbol{I}_N \otimes (\boldsymbol{I}_T - \boldsymbol{\ell}_T\boldsymbol{\ell}_T'/T),$$

and $(\boldsymbol{I}_T - \boldsymbol{\ell}_T\boldsymbol{\ell}_T'/T)\boldsymbol{y}_i = \boldsymbol{y}_i - \boldsymbol{\ell}_T\bar{\boldsymbol{y}}_i$ with the t th element being $y_{it} - \bar{\boldsymbol{y}}_i$. It follows that

$$(\boldsymbol{I}_{TN} - \boldsymbol{P}_D)\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{pmatrix} - \begin{pmatrix} \boldsymbol{\ell}_T\bar{\boldsymbol{y}}_1 \\ \boldsymbol{\ell}_T\bar{\boldsymbol{y}}_2 \\ \vdots \\ \boldsymbol{\ell}_T\bar{\boldsymbol{y}}_N \end{pmatrix},$$

which is the vector of all the deviations of y_{it} from the group averages $\bar{\boldsymbol{y}}_i$. Similarly,

$$(\boldsymbol{I}_{TN} - \boldsymbol{P}_D)\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \\ \vdots \\ \boldsymbol{Z}_N \end{pmatrix} - \begin{pmatrix} \boldsymbol{\ell}_T\bar{\boldsymbol{z}}_1' \\ \boldsymbol{\ell}_T\bar{\boldsymbol{z}}_2' \\ \vdots \\ \boldsymbol{\ell}_T\bar{\boldsymbol{z}}_N' \end{pmatrix},$$

with the t th observation in the i th block being $(\boldsymbol{z}_{it} - \bar{\boldsymbol{z}}_i)'$, the deviation of $\boldsymbol{z}_{it}$ from the group average $\bar{\boldsymbol{z}}_i$. This shows that the OLS estimator (4.23) can be obtained by regressing

$y_{it} - \bar{y}_i$ on $z_{it} - \bar{z}_i$ for $i = 1, \dots, N$, and $t = 1, \dots, T$. That is,

$$
\begin{aligned}
\hat{b}_{TN} &= \left(\sum_{i=1}^{N} (Z_i' - \bar{z}_i \ell_T')(Z_i - \ell_T \bar{z}_i') \right)^{-1} \left(\sum_{i=1}^{N} (Z_i' - \bar{z}_i \ell_T')(y_i - \ell_T \bar{y}_i) \right) \\
&= \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (z_{it} - \bar{z}_i)(z_{it} - \bar{z}_i)' \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (z_{it} - \bar{z}_i)(y_{it} - \bar{y}_i) \right).
\end{aligned}
\tag{4.25}
$$

The estimator $\hat{b}_{TN}$ will be referred to as the *within-groups estimator* because it is based on the observations that are deviations from their own group averages, as shown in (4.25). It is also easily seen that the ith element of $\hat{a}_{TN}$ is

$$
\hat{a}_{TN,i} = \frac{1}{T}(\ell_T' y_i - \ell_T' Z_i \hat{b}_{TN}) = \bar{y}_i - \bar{z}_i' \hat{b}_{TN},
$$

which involves only group averages and the within-groups estimator. To distinguish $\hat{a}_{TN}$ and $\hat{b}_{TN}$ from other estimators, we will suppress their subscript TN and denote them as $\hat{a}_w$ and $\hat{b}_w$.

Suppose that the classical conditions [A1] and [A2](i) hold for every equation i in (4.21) so that

$$
\mathbb{E}(y_i) = \ell_T a_{i,o} + Z_i b_o. \qquad i = 1, 2, \dots, N.
$$

Then, [A1] and [A2](i) also hold for the entire system (4.22) as

$$
\mathbb{E}(y) = D a_o + Z b_o,
$$

where the ith element of a_o is $a_{i,o}$. Theorem 3.4(a) now ensures that $\hat{a}_w$ and $\hat{b}_w$ are unbiased for a_o and b_o, respectively.

Suppose also that $\mathrm{var}(y_i) = \sigma_o^2 I_T$ for every equation i and that $\mathrm{cov}(y_i, y_j) = 0$ for every $i \neq j$. Under these assumptions, $\mathrm{var}(y)$ is the scalar variance-covariance matrix $\sigma_o^2 I_{TN}$. By the Gauss-Markov theorem, $\hat{a}_w$ and $\hat{b}_w$ are the BLUEs for a_o and b_o, respectively. The variance-covariance matrix of the within-groups estimator is

$$
\begin{aligned}
\mathrm{var}(\hat{b}_w) &= \sigma_o^2 [Z'(I_{TN} - P_D)Z]^{-1} \\
&= \sigma_o^2 \left[\sum_{i=1}^{N} \sum_{t=1}^{T} (z_{it} - \bar{z}_i)(z_{it} - \bar{z}_i)' \right]^{-1}.
\end{aligned}
$$

It is also easy to verify that the variance of the ith element of $\hat{a}_w$ is

$$
\mathrm{var}(\hat{a}_{w,i}) = \frac{1}{T}\sigma_o^2 + \bar{z}_i'[\mathrm{var}(\hat{b}_w)]\bar{z}_i;
\tag{4.26}
$$

see Exercise 4.9. The OLS estimator for the regression variance σ_o^2 in this case is

$$\hat{\sigma}_w^2 = \frac{1}{TN - N - k + 1} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \hat{a}_{w,i} - z_{it}' \hat{b}_w)^2,$$

which can be used to compute the estimators of $\mathrm{var}(\hat{a}_{w,i})$ and $\mathrm{var}(\hat{b}_w)$.

It should be emphasized that the conditions $\mathrm{var}(y_i) = \sigma_o^2 I_T$ for all i and $\mathrm{cov}(y_i, y_j) = 0$ for every $i \neq j$ may be much too strong in applications. When any one of these conditions fails, $\mathrm{var}(y)$ can not be written as $\sigma_o^2 I_{TN}$, and $\hat{a}_w$ and $\hat{b}_w$ are no longer the BLUEs. Despite that $\mathrm{var}(y)$ may not be a scalar variance-covariance matrix in practice, the fixed-effects model is typically estimated by the OLS method and hence also known as the *least squares dummy variable model*. For GLS and FGLS estimation see Exercise 4.10.

If [A3] holds for (4.22) such that

$$y \sim \mathcal{N}(Da_o + Zb_o, \ \sigma_o^2 I_{TN}),$$

the t and F tests discussed in Section 3.3 remain applicable. An interesting hypothesis for the fixed-effects model is whether fixed (individual) effects indeed exist. This amounts to applying an F test to the hypothesis

$$H_0 : a_{1,o} = a_{2,o} = \cdots = a_{N,o}.$$

The null distribution of this F test is $F(N - 1, TN - N - k + 1)$. In practice, it may be more convenient to estimate the following specification for the fixed-effects model:

$$
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}
=
\begin{bmatrix}
\ell_T & 0 & \cdots & 0 \\
\ell_T & \ell_T & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
\ell_T & 0 & \cdots & \ell_T
\end{bmatrix}
\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}
+
\begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_N \end{bmatrix} b
+
\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}.
\tag{4.27}
$$

This specification is virtually the same as (4.22), yet the parameters a_i, $i = 2, \ldots, N$, now denote the differences between the ith and first group effects. Testing the existence of fixed effects is then equivalent to testing

$$H_0 : a_{2,o} = \cdots = a_{N,o} = 0.$$

This can be easily done using an F test; see Exercise 4.11.

4.6.2   Random-Effects Model

Given the specification (4.21) that allows for individual effects:

$$y_i = \ell_T a_i + Z_i b + e_i, \qquad i = 1, 2, \ldots, N,$$

© Chung-Ming Kuan, 2007, 2009

we now treat a_i as random variables rather than parameters. Writing $a_i = a + u_i$ with $a = \mathbb{E}(a_i)$, the specification above can be expressed as

$$\boldsymbol{y}_i = \boldsymbol{\ell}_T a + \boldsymbol{Z}_i \boldsymbol{b} + (\boldsymbol{\ell}_T u_i + \boldsymbol{e}_i), \qquad i = 1, 2, \ldots, N. \tag{4.28}$$

where $\boldsymbol{\ell}_T u_i$ and $\boldsymbol{e}_i$ form the error term. This specification differs from the fixed-effects model in that the intercept does not vary across i. The presence of u_i also makes (4.28) different from the specification that does not allow for individual effects. Here, the group heterogeneity due to individual effects is characterized by the random variables u_i and absorbed into the error term. Thus, (4.28) is known as the *random-effects model*.

If we apply the OLS method to (4.28), the OLS estimator of $\boldsymbol{b}$ and a are, respectively,

$$\hat{\boldsymbol{b}}_p = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (\boldsymbol{z}_{it} - \bar{\boldsymbol{z}})(\boldsymbol{z}_{it} - \bar{\boldsymbol{z}})' \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (\boldsymbol{z}_{it} - \bar{\boldsymbol{z}})(y_{it} - \bar{\boldsymbol{y}}) \right), \tag{4.29}$$

and $\hat{a}_p = \bar{\boldsymbol{y}} - \bar{\boldsymbol{z}}' \hat{\boldsymbol{b}}_p$. Comparing with the within-groups estimator (4.25), $\hat{\boldsymbol{b}}_p$ is based on the deviations of y_{it} and $\boldsymbol{z}_{it}$ from their respective overall averages $\bar{\boldsymbol{y}}$ and $\bar{\boldsymbol{z}}$, whereas $\hat{\boldsymbol{b}}_w$ is based on the deviations from group averages $\bar{\boldsymbol{y}}_i$ and $\bar{\boldsymbol{z}}_i$. Note that we have suppressed the subscript TN for these two estimators. Alternatively, pre-multiplying $\boldsymbol{\ell}'/T$ through equation (4.28) yields

$$\bar{\boldsymbol{y}}_i = a + \bar{\boldsymbol{z}}_i' \boldsymbol{b} + (u_i + \bar{e}_i), \qquad i = 1, 2, \ldots, N, \tag{4.30}$$

where $\bar{e}_i = \sum_{t=1}^{T} e_{it}/T$. By noting that the sample averages of $\bar{\boldsymbol{y}}_i$ and $\bar{\boldsymbol{z}}_i$ are just $\bar{\boldsymbol{y}}$ and $\bar{\boldsymbol{z}}$, the OLS estimators for the specification (4.30) are

$$\hat{\boldsymbol{b}}_b = \left(\sum_{i=1}^{N} (\bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}})(\bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}})' \right)^{-1} \left(\sum_{i=1}^{N} (\bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}})(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}}) \right), \tag{4.31}$$

and $\hat{a}_b = \bar{\boldsymbol{y}} - \bar{\boldsymbol{z}}' \hat{\boldsymbol{b}}_b$. The estimator $\hat{\boldsymbol{b}}_b$ is known as the *between-groups estimator* because it is based on the deviations of group averages from the overall averages. Here, we suppress the subscript N for $\hat{a}_b$ and $\hat{\boldsymbol{b}}_b$. It can also be shown that the estimator (4.29) is a weighted sum of the within-groups estimator (4.25) and the between-group estimator (4.31); see Exercise 4.12. Thus, $\hat{\boldsymbol{b}}_p$ is known as the *pooled estimator*.

Suppose that the classical conditions [A1] and [A2](i) hold for every equation i such that

$$\mathbb{E}(\boldsymbol{y}_i) = \boldsymbol{\ell}_T a_o + \boldsymbol{Z}_i \boldsymbol{b}_o, \qquad i = 1, \ldots, N.$$

Then, [A1] and [A2](i) hold for the entire system of (4.28) as

$$\mathbb{E}(\boldsymbol{y}) = \boldsymbol{\ell}_{TN} a_o + \boldsymbol{Z} \boldsymbol{b}_o.$$

Moreover, they also hold for the specification (4.30) as

$$\mathbb{E}(\bar{\boldsymbol{y}}_i) = a_o + \bar{\boldsymbol{z}}_i' \boldsymbol{b}_o, \qquad i = 1, \ldots, N.$$

It follows that $\hat{a}_b$ and $\hat{\boldsymbol{b}}_b$, as well as $\hat{a}_p$ and $\hat{\boldsymbol{b}}_p$, are unbiased for a_o and $\boldsymbol{b}_o$ in the random-effects model. Moreover, write

$$\boldsymbol{y}_i = \boldsymbol{\ell}_T a_o + \boldsymbol{Z}_i \boldsymbol{b}_o + \boldsymbol{\epsilon}_i^*, \tag{4.32}$$

where $\boldsymbol{\epsilon}_i^*$ is the sum of two components: the random effects $\boldsymbol{\ell}_T u_i$ and the disturbance $\boldsymbol{\epsilon}_i$ for equation i. Then,

$$\mathrm{var}(\boldsymbol{y}_i) = \sigma_u^2 \boldsymbol{\ell}_T \boldsymbol{\ell}_T' + \mathrm{var}(\boldsymbol{\epsilon}_i) + 2\,\mathrm{cov}(\boldsymbol{\ell}_T u_i, \boldsymbol{\epsilon}_i),$$

where σ_u^2 is $\mathrm{var}(u_i)$. As the first term on the right-hand side above is a full matrix, $\mathrm{var}(\boldsymbol{y}_i)$ is not a scalar variance-covariance matrix in general. Consequently, $\hat{a}_p$ and $\hat{\boldsymbol{b}}_p$ are not the BLUEs. For the specification (4.30), $\hat{a}_b$ and $\hat{\boldsymbol{b}}_b$ are not the BLUEs unless more stringent conditions are imposed.

Remark: If the fixed-effects model (4.21) is correct, the random-effects model (4.28) can be viewed as a specification that omits $n - 1$ relevant dummy variables. This implies the pooled estimator $\hat{\boldsymbol{b}}_p$ and the between-groups estimator $\hat{\boldsymbol{b}}_b$ are biased for $\boldsymbol{b}_o$ in the fixed-effects model. This should not be too surprising because, while there are $N + k - 1$ parameters in the fixed-effects model, the specification (4.30) only permits estimation of k parameters. We therefore conclude that neither the between-groups estimator nor the pooled estimator is a proper choice for the fixed-effects model.

To perform FGLS estimation for the random-effects model, more conditions on $\mathrm{var}(\boldsymbol{y}_i)$ are needed. If $\mathrm{var}(\boldsymbol{\epsilon}_i) = \sigma_o^2 \boldsymbol{I}_T$ and $\mathbb{E}(u_i \boldsymbol{\epsilon}_i) = \boldsymbol{0}$, $\mathrm{var}(\boldsymbol{y}_i)$ has a simpler form:

$$\boldsymbol{S}_o := \mathrm{var}(\boldsymbol{y}_i) = \sigma_u^2 \boldsymbol{\ell}_T \boldsymbol{\ell}_T' + \sigma_o^2 \boldsymbol{I}_T.$$

Under the following additional conditions: $\mathbb{E}(u_i u_j) = 0$, $E(u_i \boldsymbol{\epsilon}_j) = \boldsymbol{0}$ and $\mathbb{E}(\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_j) = \boldsymbol{0}$ for all $i \neq j$, we have $\mathrm{cov}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \boldsymbol{0}$. Hence, $\mathrm{var}(\boldsymbol{y})$ simplifies to a block diagonal matrix:

$$\boldsymbol{\Sigma}_o := \mathrm{var}(\boldsymbol{y}) = \boldsymbol{I}_N \otimes \boldsymbol{S}_o,$$

which is not a scalar variance-covariance matrix unless $\sigma_u^2 = 0$. It can be verified that the desired transformation matrix for GLS estimation is $\boldsymbol{\Sigma}_o^{-1/2} = I_N \otimes \boldsymbol{S}_o^{-1/2}$, where

$$\boldsymbol{S}_o^{-1/2} = \boldsymbol{I}_T - \frac{c}{T} \boldsymbol{\ell}_T \boldsymbol{\ell}_T',$$

and $c = 1 - \sigma_o^2/(T\sigma_u^2 + \sigma_o^2)^{1/2}$. The transformed data are then $\boldsymbol{S}_o^{-1/2} \boldsymbol{y}_i$ and $\boldsymbol{S}_o^{-1/2} \boldsymbol{Z}_i$, $i = 1, \ldots, N$, and their tth elements are, respectively, $y_{it} - c\bar{y}_i$ and $\boldsymbol{z}_{it} - c\bar{\boldsymbol{z}}_i$. Regressing $y_{it} - c\bar{\boldsymbol{y}}_i$ on $\boldsymbol{z}_{it} - c\bar{\boldsymbol{z}}_i$ gives the desired GLS estimator.

It can be shown that the GLS estimator is also a weighted average of the within-groups and between-groups estimators. For the special case that $\sigma_o^2 = 0$, we have $c = 1$ and

$$\boldsymbol{\Sigma}_o^{-1/2} = \boldsymbol{I}_N \otimes (\boldsymbol{I}_T - \boldsymbol{\ell}_T \boldsymbol{\ell}_T'/T) = \boldsymbol{I}_{TN} - \boldsymbol{P}_D,$$

as in the fixed-effects model. In this case, the GLS estimator of $\boldsymbol{b}$ is nothing but the within-groups estimator $\hat{\boldsymbol{b}}_w$. When $c = 0$, the GLS estimator of $\boldsymbol{b}$ reduces to the pooled estimator $\hat{\boldsymbol{b}}_p$.

To compute the FGLS estimator, we must estimate the parameters σ_u^2 and σ_o^2 in $\boldsymbol{S}_o$. We first eliminate the random effects u_i by taking the difference of $\boldsymbol{y}_i$ and $\boldsymbol{\ell}_T \bar{y}_i$:

$$\boldsymbol{y}_i - \boldsymbol{\ell}_T \bar{y}_i = (\boldsymbol{Z}_i - \boldsymbol{\ell}_T \bar{\boldsymbol{z}}_i') \boldsymbol{b}_o + (\boldsymbol{\epsilon}_i - \boldsymbol{\ell}_T \bar{\epsilon}_i).$$

For this specification, the OLS estimator of $\boldsymbol{b}_o$ is just the within-groups estimator $\hat{\boldsymbol{b}}_w$. As u_i have been eliminated, we can estimate σ_o^2, the variance of ϵ_{it}, by

$$\hat{\sigma}_\epsilon^2 = \frac{1}{TN - N - k + 1} \sum_{i=1}^N \sum_{i=1}^T [(y_{it} - \bar{y}_i) - (\boldsymbol{z}_{it} - \bar{\boldsymbol{z}}_i)' \hat{\boldsymbol{b}}_w]^2,$$

which is also the variance estimator $\hat{\sigma}_w^2$ in the fixed-effects model. To estimate σ_u^2, note that under [A1] and [A2](i) we have

$$\bar{y}_i = a_o + \bar{\boldsymbol{z}}_i' \boldsymbol{b}_o + (u_i + \bar{\epsilon}_i), \qquad i = 1, 2, \ldots, N,$$

which corresponds to the specification (4.30) for computing the between-groups estimator. When $[A2](ii)$ also holds for every i,

$$\mathrm{var}(u_i + \bar{\epsilon}_i) = \sigma_u^2 + \sigma_o^2/T.$$

This variance may be estimated by $\sum_{i=1}^N \hat{e}_{b,i}^2/(N - k)$, where

$$\hat{e}_{b,i} = (\bar{y}_i - \bar{y}) - (\bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}})' \hat{\boldsymbol{b}}_b, \qquad i = 1, \ldots, N.$$

Consequently, the estimator for σ_u^2 is

$$\hat{\sigma}_u^2 = \frac{1}{N - k} \sum_{i=1}^N \hat{e}_{b,i}^2 - \frac{\hat{\sigma}_\epsilon^2}{T}.$$

The estimators $\hat{\sigma}_u^2$ and $\hat{\sigma}_\epsilon^2$ now can be used to construct the estimated transformation matrix $\hat{\boldsymbol{S}}^{-1/2}$. It is clear that the FGLS estimator is, again, a very complex function of $\boldsymbol{y}$.

4.7   Limitations of the FGLS Method

In this chapter we relax only the classical condition [A2](ii) while maintaining [A1] and [A2](i). The limitations of [A1] and [A2](i) discussed in Chapter 3.6 therefore still exist. In particular, stochastic regressors and nonlinear specifications are excluded in the present context.

Although the GLS and FGLS methods are designed to improve on estimation efficiency when there is a non-scalar covariance matrix $\mathbf{\Sigma}_o$, they also create further difficulties. First, the GLS estimator is usually not available, except in some exceptional cases. Second, a convenient FGLS estimator is available at the expense of more conditions on $\mathbf{\Sigma}_o$. If these simplifying conditions are incorrectly imposed, the resulting FGLS estimator may perform poorly. Third, the finite-sample properties of the FGLS estimator are typically unknown. In general, we do not know if an FGLS estimator is unbiased, nor do we know its efficiency relative to the OLS estimator and its exact distribution. It is therefore difficult to draw statistical inferences from FGLS estimation results.

Exercises

4.1 Given the linear specification $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, suppose that the conditions [A1] and [A2](i) hold and that $\mathrm{var}(\boldsymbol{y}) = \mathbf{\Sigma}_o$. If the matrix $\boldsymbol{X}$ contains k eigenvectors of $\mathbf{\Sigma}_o$ which are normalized to unit length. What are the resulting $\hat{\boldsymbol{\beta}}_T$ and $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$? Explain your result.

4.2 For the specification (3.1) estimated by the GLS method, a natural goodness-of-fit measure is

$$\text{Centered } R^2_{\mathrm{GLS}} = 1 - \frac{\hat{\boldsymbol{e}}'_{\mathrm{GLS}}\hat{\boldsymbol{e}}_{\mathrm{GLS}}}{\text{Centered TSS of } \boldsymbol{y}},$$

where the denominator is the centered TSS of the original dependent variable y. Show that R^2_{GLS} need not be bounded between zero and one.

4.3 Given the linear specification $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, suppose that the conditions [A1] and [A2](i) hold and that $\mathrm{var}(\boldsymbol{y}) = \mathbf{\Sigma}_o$. Show directly that

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_T) - \mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}})$$

is a positive semi-definite matrix.

4.4 Given the linear specification $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, suppose that the conditions [A1] and [A2](i) hold and that $\mathrm{var}(\boldsymbol{y}) = \mathbf{\Sigma}_o$. Show that

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}_T, \hat{\boldsymbol{\beta}}_{\mathrm{GLS}}) = \mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}).$$

Also find $\mathrm{cov}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}, \hat{\boldsymbol{\beta}}_{\mathrm{GLS}} - \hat{\boldsymbol{\beta}}_T)$.

4.5 Suppose that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_o + \boldsymbol{\epsilon}$ and the elements of $\boldsymbol{\epsilon}$ are $\epsilon_t = \alpha_1 \epsilon_{t-1} + u_t$, where $\alpha_1 = 1$ and $\{u_t\}$ is a white noise with mean zero and variance σ_u^2. What are the properties of ϵ_t? Is $\{\epsilon_t\}$ still weakly stationary?

4.6 Suppose that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_o + \boldsymbol{\epsilon}$ and the elements of $\boldsymbol{\epsilon}$ are $\epsilon_t = u_t + \alpha_1 u_{t-1}$, where $|\alpha_1| < 1$ and $\{u_t\}$ is a white noise with mean zero and variance σ_u^2. Calculate the variance, autocovariances, and autocorrelations of ϵ_t and compare them with those of AR(1) disturbances.

4.7 Let y_t denote investment expenditure that is determined by expected earning x_t^*:

$$y_t = a_o + b_o x_t^* + u_t.$$

When x_t^* is adjusted adaptively:

$$x_t^* = x_{t-1}^* + (1 - \lambda_o)(x_t - x_{t-1}^*), \qquad 0 < \lambda_o < 1,$$

show that y_t can be represented by a model with a lagged dependent variable and moving average disturbances.

4.8 Given the SUR specification (4.17), show that the GLS estimator is the same as the OLS estimator when $\boldsymbol{X}_i = \boldsymbol{X}_0$ for all i. Give an intuitive explanation of this result.

4.9 Given the fixed-effects model (4.21) for panel data, suppose that [A1] and [A2](i) hold for each group equation, $\mathrm{var}(\boldsymbol{y}_i) = \sigma_o^2 \boldsymbol{I}_T$ and $\mathrm{cov}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \boldsymbol{0}$ for $i \neq j$. Prove (4.26).

4.10 Given the fixed-effects model (4.21) for panel data, suppose that [A1] and [A2](i) hold for each group equation, $\mathrm{var}(\boldsymbol{y}_i) = \sigma_i^2 \boldsymbol{I}_T$ and $\mathrm{cov}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \sigma_{ij} \boldsymbol{I}_T$ for $i \neq j$. What is $\mathrm{var}(\boldsymbol{y})$? Find the GLS estimator and propose an FGLS estimator.

4.11 Given the fixed-effects model (4.27) for panel data, consider testing the null hypothesis of no fixed effects. Write down the F statistic that is based on constrained and unconstrained R^2 and explain clearly how this test should be implemented.

4.12 Consider the pooled estimator (4.29), the within-groups estimator (4.25) and the between-groups estimator (4.31). Define the total sum of squares (TSS), total sum of cross products (TSC), within-groups sum of squares (WSS), within-groups cross products (WSC), between-groups sum of squares (BSS) and between-groups sum of

cross products (BSC) as, respectively,

$$\text{TSS} = \sum_{i=1}^{N} \sum_{t=1}^{T} (\boldsymbol{z}_{it} - \bar{\boldsymbol{z}})(\boldsymbol{z}_{it} - \bar{\boldsymbol{z}})', \qquad \text{TSC} = \sum_{i=1}^{N} \sum_{t=1}^{T} (\boldsymbol{z}_{it} - \bar{\boldsymbol{z}})(y_{it} - \bar{\boldsymbol{y}})',$$

$$\text{WSS} = \sum_{i=1}^{N} \sum_{t=1}^{T} (\boldsymbol{z}_{it} - \bar{\boldsymbol{z}}_i)(\boldsymbol{z}_{it} - \bar{\boldsymbol{z}}_i)', \qquad \text{WSC} = \sum_{i=1}^{N} \sum_{t=1}^{T} (\boldsymbol{z}_{it} - \bar{\boldsymbol{z}}_i)(y_{it} - \bar{\boldsymbol{y}}_i)',$$

$$\text{BSS} = \sum_{i=1}^{N} T(\bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}})(\bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}})', \qquad \text{BSC} = \sum_{i=1}^{N} T(\bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}})(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}})'.$$

Prove that TSS = WSS + BSS and TSC = WSC + BSC. Based on these results, show that the pooled estimator (4.29) is a weighted sum of the within-groups estimator (4.25) and the between-groups estimator (4.31).

References

Amemiya, Takeshi (1985). *Advanced Econometrics*, Cambridge, MA: Harvard University Press.

Farebrother, R. W. (1980). The Durbin-Watson test for serial correlation when there is no intercept in the regression, *Econometrica*, **48**, 1553–1563.

Goldberger, Arthur S. (1991). *A Course in Econometrics*, Cambridge, MA: Harvard University Press.

Greene, William H. (2000). *Econometric Analysis*, Fourth ed., New York, NY: Macmillan.

Wallis, K. F. (1972). Testing for fourth order autocorrelation in quarterly regression equations, *Econometrica*, **40**, 617–636.