10

NONSPHERICAL DISTURBANCES—THE GENERALIZED REGRESSION MODEL

10.1 INTRODUCTION

In Chapter 9, we extended the classical linear model to allow the conditional mean to be a nonlinear function.[1] But we retained the important assumptions about the disturbances: that they are uncorrelated with each other and that they have a constant variance, conditioned on the independent variables. In this and the next several chapters, we extend the multiple regression model to disturbances that violate these classical assumptions. The **generalized linear regression model** is

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\ E[\boldsymbol{\varepsilon} \mid \mathbf{X}] &= \mathbf{0}, \\ E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] &= \sigma^2\boldsymbol{\Omega} = \boldsymbol{\Sigma}, \end{aligned} \tag{10-1}$$

where $\boldsymbol{\Omega}$ is a positive definite matrix. (The covariance matrix is written in the form $\sigma^2\boldsymbol{\Omega}$ at several points so that we can obtain the classical model, $\sigma^2\mathbf{I}$, as a convenient special case.) As we will examine briefly below, the extension of the model to nonlinearity is relatively minor in comparison with the variants considered here. For present purposes, we will retain the linear specification and refer to our model simply as the **generalized regression model.**

Two cases we will consider in detail are **heteroscedasticity** and **autocorrelation.** Disturbances are heteroscedastic when they have different variances. Heteroscedasticity usually arises in volatile high frequency time-series data such as daily observations in financial markets and in cross-section data where the scale of the dependent variable and the explanatory power of the model tend to vary across observations. Microeconomic data such as expenditure surveys are typical. The disturbances are still assumed to be uncorrelated across observations, so $\sigma^2\boldsymbol{\Omega}$ would be

$$\sigma^2\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} \omega_{11} & 0 & \cdots & 0 \\ 0 & \omega_{22} & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \omega_{nn} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}.$$

[1] Recall that our definition of nonlinearity pertains to the estimation method required to obtain the parameter estimates, not to the way that they enter the regression function.

(The first mentioned situation involving financial data is more complex than this, and is examined in detail in Section 11.8.)

Autocorrelation is usually found in time-series data. Economic time series often display a "memory" in that variation around the regression function is not independent from one period to the next. The seasonally adjusted price and quantity series published by government agencies are examples. Time-series data are usually homoscedastic, so $\sigma^2\boldsymbol{\Omega}$ might be

$$\sigma^2\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ & & \vdots & \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{bmatrix}.$$

The values that appear off the diagonal depend on the model used for the disturbance. In most cases, consistent with the notion of a fading memory, the values decline as we move away from the diagonal.

Panel data sets, consisting of cross sections observed at several points in time, may exhibit both characteristics. We shall consider them in Chapter 14. This chapter presents some general results for this extended model. The next several chapters examine in detail specific types of generalized regression models.

Our earlier results for the classical model will have to be modified. We will take the same approach in this chapter on general results and in the next two on heteroscedasticity and serial correlation, respectively:

1. We first consider the consequences for the least squares estimator of the more general form of the regression model. This will include assessing the effect of ignoring the complication of the generalized model and of devising an appropriate estimation strategy, still based on least squares.
2. In subsequent sections, we will examine alternative estimation approaches that can make better use of the characteristics of the model. We begin with GMM estimation, which is **robust** and **semiparametric.** Minimal assumptions about $\boldsymbol{\Omega}$ are made at this point.
3. We then narrow the assumptions and begin to look for methods of detecting the failure of the classical model—that is, we formulate procedures for testing the specification of the classical model against the generalized regression.
4. The final step in the analysis is to formulate **parametric** models that make specific assumptions about $\boldsymbol{\Omega}$. Estimators in this setting are some form of generalized least squares or maximum likelihood.

The model is examined in general terms in this and the next two chapters. Major applications to panel data and multiple equation systems are considered in Chapters 13 and 14.

10.2 LEAST SQUARES AND INSTRUMENTAL VARIABLES ESTIMATION

The essential results for the classical model with **spherical** disturbances

$$E[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \mathbf{0}$$

and

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \sigma^2\mathbf{I} \tag{10-2}$$

are presented in Chapters 2 through 8. To reiterate, we found that the **ordinary least squares (OLS) estimator**

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \tag{10-3}$$

is best linear unbiased (BLU), consistent and asymptotically normally distributed (CAN), and if the disturbances are normally distributed, like other maximum likelihood estimators considered in Chapter 17, asymptotically efficient among all CAN estimators. We now consider which of these properties continue to hold in the model of (10-1).

To summarize, the least squares, nonlinear least squares, and instrumental variables estimators retain only some of their desirable properties in this model. Least squares remains unbiased, consistent, and asymptotically normally distributed. It will, however, no longer be efficient—this claim remains to be verified—and the usual inference procedures are no longer appropriate. Nonlinear least squares and instrumental variables likewise remain consistent, but once again, the extension of the model brings about some changes in our earlier results concerning the asymptotic distributions. We will consider these cases in detail.

10.2.1 FINITE-SAMPLE PROPERTIES OF ORDINARY LEAST SQUARES

By taking expectations on both sides of (10-3), we find that if $E[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \mathbf{0}$, then

$$E[\mathbf{b}] = E_{\mathbf{X}}[E[\mathbf{b} \mid \mathbf{X}]] = \boldsymbol{\beta}. \tag{10-4}$$

Therefore, we have the following theorem.

THEOREM 10.1 Finite Sample Properties of b in the Generalized Regression Model

If the regressors and disturbances are uncorrelated, then the unbiasedness of least squares is unaffected by violations of assumption (10-2). The least squares estimator is unbiased in the generalized regression model. With nonstochastic regressors, or conditional on $\mathbf{X}$*, the sampling variance of the least squares estimator is*

$$\begin{aligned}
\text{Var}[\mathbf{b} \mid \mathbf{X}] &= E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})' \mid \mathbf{X}] \\
&= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mid \mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\boldsymbol{\Omega})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \frac{\sigma^2}{n}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}.
\end{aligned} \tag{10-5}$$

If the regressors are stochastic, then the unconditional variance is $E_{\mathbf{X}}[\text{Var}[\mathbf{b} \mid \mathbf{X}]]$. *In (10-3),* $\mathbf{b}$ *is a linear function of* $\boldsymbol{\varepsilon}$. *Therefore, if* $\boldsymbol{\varepsilon}$ *is normally distributed, then*

$$\mathbf{b} \mid \mathbf{X} \sim N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}].$$

The end result is that **b** has properties that are similar to those in the classical regression case. Since the variance of the least squares estimator is not $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, however, statistical inference based on $s^2(\mathbf{X}'\mathbf{X})^{-1}$ may be misleading. Not only is this the wrong matrix to be used, but s^2 may be a biased estimator of σ^2. There is usually no way to know whether $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is larger or smaller than the true variance of **b**, so even with a good estimate of σ^2, the conventional estimator of Var[**b**] may not be particularly useful. Finally, since we have dispensed with the fundamental underlying assumption, the familiar inference procedures based on the F and t distributions will no longer be appropriate. One issue we will explore at several points below is how badly one is likely to go awry if the result in (10-5) is ignored and if the use of the familiar procedures based on $s^2(\mathbf{X}'\mathbf{X})^{-1}$ is continued.

10.2.2 ASYMPTOTIC PROPERTIES OF LEAST SQUARES

If Var[**b** | **X**] converges to zero, then **b** is mean square consistent. With well-behaved regressors, $(\mathbf{X}'\mathbf{X}/n)^{-1}$ will converge to a constant matrix. But $(\sigma^2/n)(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}/n)$ need not converge at all. By writing this product as

$$\frac{\sigma^2}{n}\left(\frac{\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}}{n}\right) = \left(\frac{\sigma^2}{n}\right)\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{ij}\mathbf{x}_i\mathbf{x}_j'}{n}\right) \tag{10-6}$$

we see that though the leading constant will, by itself, converge to zero, the matrix is a sum of n^2 terms, divided by n. Thus, the product is a scalar that is $O(1/n)$ times a matrix that is, at least at this juncture, $O(n)$, which is $O(1)$. So, it does appear at first blush that if the product in (10-6) does converge, it might converge to a matrix of nonzero constants. In this case, the covariance matrix of the least squares estimator would not converge to zero, and consistency would be difficult to establish. We will examine in some detail, the conditions under which the matrix in (10-6) converges to a constant matrix.[2] If it does, then since σ^2/n does vanish, ordinary least squares is consistent as well as unbiased.

THEOREM 10.2 Consistency of OLS in the Generalized Regression Model

If $\mathbf{Q} = \text{plim}(\mathbf{X}'\mathbf{X}/n)$ *and* $\text{plim}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}/n)$ *are both finite positive definite matrices, then* **b** *is consistent for* $\boldsymbol{\beta}$. *Under the assumed conditions,*

$$\text{plim}\,\mathbf{b} = \boldsymbol{\beta}. \tag{10-7}$$

The conditions in Theorem 10.2 depend on both **X** and $\boldsymbol{\Omega}$. An alternative formula[3] that separates the two components is as follows. Ordinary least squares is consistent in the generalized regression model if:

1. The smallest characteristic root of $\mathbf{X}'\mathbf{X}$ increases without bound as $n \to \infty$, which implies that $\text{plim}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0}$. If the regressors satisfy the Grenander conditions **G1** through **G3** of Section 5.2, then they will meet this requirement.

[2] In order for the product in (10-6) to vanish, it would be sufficient for $(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}/n)$ to be $O(n^{\delta})$ where $\delta < 1$.

[3] Amemiya (1985, p. 184).

2. The largest characteristic root of $\boldsymbol{\Omega}$ is finite for all n. For the heteroscedastic model, the variances are the characteristic roots, which requires them to be finite. For models with autocorrelation, the requirements are that the elements of $\boldsymbol{\Omega}$ be finite and that the off-diagonal elements not be too large relative to the diagonal elements. We will examine this condition at several points below.

The least squares estimator is asymptotically normally distributed if the limiting distribution of

$$\sqrt{n}(\mathbf{b}-\boldsymbol{\beta}) = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\frac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\varepsilon} \tag{10-8}$$

is normal. If $\operatorname{plim}(\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}$, then the limiting distribution of the right-hand side is the same as that of

$$\mathbf{v}_{n,LS} = \mathbf{Q}^{-1}\frac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{Q}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i, \tag{10-9}$$

where $\mathbf{x}_i'$ is a row of $\mathbf{X}$ (assuming, of course, that the limiting distribution exists at all). The question now is whether a central limit theorem can be applied directly to $\mathbf{v}$. If the disturbances are merely heteroscedastic and still uncorrelated, then the answer is generally yes. In fact, we already showed this result in Section 5.5.2 when we invoked the Lindberg–Feller central limit theorem (D.19) or the Lyapounov Theorem (D.20). The theorems allow unequal variances in the sum. The exact variance of the sum is

$$E_{\mathbf{x}}\left[\operatorname{Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right]\middle|\mathbf{x}_i\right] = \frac{\sigma^2}{n}\sum_{i=1}^{n}\omega_i\mathbf{Q}_i,$$

which, for our purposes, we would require to converge to a positive definite matrix. In our analysis of the classical model, the heterogeneity of the variances arose because of the regressors, but we still achieved the limiting normal distribution in (5-7) through (5-14). All that has changed here is that the variance of ε varies across observations *as well*. Therefore, *the proof of asymptotic normality in Section 5.2.2 is general enough to include this model without modification*. As long as $\mathbf{X}$ is well behaved and the diagonal elements of $\boldsymbol{\Omega}$ are finite and well behaved, the least squares estimator is asymptotically normally distributed, with the covariance matrix given in (10-5). That is:

In the heteroscedastic case, if the variances of ε_i are finite and are not dominated by any single term, so that the conditions of the Lindberg–Feller central limit theorem apply to $\mathbf{v}_{n,LS}$ in (10-9), then the least squares estimator is asymptotically normally distributed with covariance matrix

$$\text{Asy. Var}[\mathbf{b}] = \frac{\sigma^2}{n}\mathbf{Q}^{-1}\operatorname{plim}\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\right)\mathbf{Q}^{-1}. \tag{10-10}$$

For the most general case, asymptotic normality is much more difficult to establish because the sums in (10-9) are not necessarily sums of independent or even uncorrelated random variables. Nonetheless, Amemiya (1985, p. 187) and Anderson (1971) have shown the asymptotic normality of $\mathbf{b}$ in a model of autocorrelated disturbances general enough to include most of the settings we are likely to meet in practice. We will revisit

this issue in Chapters 19 and 20 when we examine time series modeling. We can conclude that, except in particularly unfavorable cases, we have the following theorem.

THEOREM 10.3 Asymptotic Distribution of b in the GR Model
If the regressors are sufficiently well behaved and the off-diagonal terms in $\boldsymbol{\Omega}$ *diminish sufficiently rapidly, then the least squares estimator is asymptotically normally distributed with mean* $\boldsymbol{\beta}$ *and covariance matrix given in (10-10).*

There are two cases that remain to be considered, the nonlinear regression model and the instrumental variables estimator.

10.2.3 ASYMPTOTIC PROPERTIES OF NONLINEAR LEAST SQUARES

If the regression function is nonlinear, then the analysis of this section must be applied to the pseudoregressors $\mathbf{x}_i^0$ rather than the independent variables. Aside from this consideration, no new results are needed. We can just apply this discussion to the linearized regression model. Under most conditions, the results listed above apply to the **nonlinear least squares estimator** as well as the linear least squares estimator.[4]

10.2.4 ASYMPTOTIC PROPERTIES OF THE INSTRUMENTAL VARIABLES ESTIMATOR

The second estimator to be considered is the **instrumental variables estimator** that we considered in Sections 5.4 for the linear model and 9.5.1 for the nonlinear model. We will confine our attention to the linear model. The nonlinear case can be obtained by applying our results to the linearized regression. To review, we considered cases in which the regressors $\mathbf{X}$ are correlated with the disturbances $\boldsymbol{\varepsilon}$. If this is the case, as in the time-series models and the errors in variables models that we examined earlier, then $\mathbf{b}$ is neither unbiased nor consistent.[5] In the classical model, we constructed an estimator around a set of variables $\mathbf{Z}$ that were uncorrelated with $\boldsymbol{\varepsilon}$,

$$\begin{aligned}\mathbf{b}_{\mathrm{IV}} &= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \\ &= \boldsymbol{\beta} + [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}.\end{aligned} \tag{10-11}$$

Suppose that $\mathbf{X}$ and $\mathbf{Z}$ are well behaved in the sense discussed in Section 5.4. That is,

$$\begin{aligned}&\operatorname{plim}(1/n)\mathbf{Z}'\mathbf{Z} = \mathbf{Q}_{\mathbf{ZZ}}, \text{ a positive definite matrix,} \\ &\operatorname{plim}(1/n)\mathbf{Z}'\mathbf{X} = \mathbf{Q}_{\mathbf{ZX}} = \mathbf{Q}_{\mathbf{XZ}}', \text{ a nonzero matrix,} \\ &\operatorname{plim}(1/n)\mathbf{X}'\mathbf{X} = \mathbf{Q}_{\mathbf{XX}}, \text{ a positive definite matrix.}\end{aligned}$$

[4] Davidson and MacKinnon (1993) consider this case at length.

[5] It may be asymptotically normally distributed, but around a mean that differs from $\boldsymbol{\beta}$.

To avoid a string of matrix computations that may not fit on a single line, for convenience let

$$\begin{aligned}\mathbf{Q}_{\mathbf{XX.Z}} &= [\mathbf{Q}_{\mathbf{XZ}}\mathbf{Q}_{\mathbf{ZZ}}^{-1}\mathbf{Q}_{\mathbf{ZX}}]^{-1}\mathbf{Q}_{\mathbf{XZ}}\mathbf{Q}_{\mathbf{ZZ}}^{-1} \\ &= \text{plim}\left[\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)\right]^{-1}\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1}.\end{aligned}$$

If $\mathbf{Z}$ is a valid set of instrumental variables, that is, if the second term in (10-11) vanishes asymptotically, then

$$\text{plim}\,\mathbf{b}_{\text{IV}} = \boldsymbol{\beta} + \mathbf{Q}_{\mathbf{XX.Z}}\,\text{plim}\left(\frac{1}{n}\mathbf{Z}'\boldsymbol{\varepsilon}\right) = \boldsymbol{\beta}.$$

This result is exactly the same one we had before. We might note that at the several points where we have established unbiasedness or consistency of the least squares or instrumental variables estimator, the covariance matrix of the disturbance vector has played no role; unbiasedness is a property of the means. As such, this result should come as no surprise. The large sample behavior of $\mathbf{b}_{\text{IV}}$ depends on the behavior of

$$\mathbf{v}_{n,\text{IV}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i.$$

This result is exactly the one we analyzed in Section 5.4. If the sampling distribution of $\mathbf{v}_n$ converges to a normal distribution, then we will be able to construct the asymptotic distribution for $\mathbf{b}_{\text{IV}}$. This set of conditions is the same that was necessary for $\mathbf{X}$ when we considered $\mathbf{b}$ above, with $\mathbf{Z}$ in place of $\mathbf{X}$. We will once again rely on the results of Anderson (1971) or Amemiya (1985) that under very general conditions,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i \xrightarrow{d} \mathbf{N}\left[\mathbf{0}, \sigma^2\text{plim}\left(\frac{1}{n}\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}\right)\right].$$

With the other results already in hand, we now have the following.

THEOREM 10.4 Asymptotic Distribution of the IV Estimator in the Generalized Regression Model

If the regressors and the instrumental variables are well behaved in the fashions discussed above, then

$$\mathbf{b}_{\text{IV}} \overset{a}{\sim} N[\boldsymbol{\beta}, \mathbf{V}_{\text{IV}}],$$

where **(10-12)**

$$\mathbf{V}_{\text{IV}} = \frac{\sigma^2}{n}(\mathbf{Q}_{\mathbf{XX.Z}})\,\text{plim}\left(\frac{1}{n}\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}\right)(\mathbf{Q}'_{\mathbf{XX.Z}}).$$

10.3 ROBUST ESTIMATION OF ASYMPTOTIC COVARIANCE MATRICES

There is a remaining question regarding all the preceding. In view of (10-5), is it necessary to discard ordinary least squares as an estimator? Certainly if $\boldsymbol{\Omega}$ is known, then, as shown in Section 10.5, there is a simple and efficient estimator available based on it, and the answer is yes. If $\boldsymbol{\Omega}$ is unknown but its structure is known and we can estimate $\boldsymbol{\Omega}$ using sample information, then the answer is less clear-cut. In many cases, basing estimation of $\boldsymbol{\beta}$ on some alternative procedure that uses an $\hat{\boldsymbol{\Omega}}$ will be preferable to ordinary least squares. This subject is covered in Chapters 11 to 14. The third possibility is that $\boldsymbol{\Omega}$ is completely unknown, both as to its structure and the specific values of its elements. In this situation, least squares or instrumental variables may be the only estimator available, and as such, the only available strategy is to try to devise an estimator for the appropriate asymptotic covariance matrix of $\mathbf{b}$.

If $\sigma^2\boldsymbol{\Omega}$ were known, then the *estimator* of the asymptotic covariance matrix of $\mathbf{b}$ in (10-10) would be

$$\mathbf{V}_{\text{OLS}} = \frac{1}{n}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'[\sigma^2\boldsymbol{\Omega}]\mathbf{X}\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}.$$

For the nonlinear least squares estimator, we replace $\mathbf{X}$ with $\mathbf{X}^0$. For the instrumental variables estimator, the left- and right-side matrices are replaced with this sample estimates of $\mathbf{Q}_{\mathbf{XX}.\mathbf{Z}}$ and its transpose (using $\mathbf{X}^0$ again for the nonlinear instrumental variables estimator), and $\mathbf{Z}$ replaces $\mathbf{X}$ in the center matrix. In all these cases, the matrices of sums of squares and cross products in the left and right matrices are sample data that are readily estimable, and the problem is the center matrix that involves the unknown $\sigma^2\boldsymbol{\Omega}$. For estimation purposes, note that σ^2 is not a separate unknown parameter. Since $\boldsymbol{\Omega}$ is an unknown matrix, it can be scaled arbitrarily, say by κ, and with σ^2 scaled by $1/\kappa$, the same product remains. In our applications, we will remove the indeterminacy by assuming that $\text{tr}(\boldsymbol{\Omega}) = n$, as it is when $\sigma^2\boldsymbol{\Omega} = \sigma^2\mathbf{I}$ in the classical model. For now, just let $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Omega}$. It might seem that to estimate $(1/n)\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}$, an estimator of $\boldsymbol{\Sigma}$, which contains $n(n+1)/2$ unknown parameters, is required. But fortunately (since with n observations, this method is going to be hopeless), this observation is not quite right. What is required is an estimator of the $K(K+1)/2$ unknown elements in the matrix

$$\text{plim}\,\mathbf{Q}_* = \text{plim}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}\mathbf{x}_i\mathbf{x}_j'.$$

The point is that $\mathbf{Q}_*$ is a matrix of sums of squares and cross products that involves σ_{ij} *and* the rows of $\mathbf{X}$ (or $\mathbf{Z}$ or $\mathbf{X}^0$). The least squares estimator $\mathbf{b}$ is a consistent estimator of $\boldsymbol{\beta}$, which implies that the least squares residuals e_i are "pointwise" consistent estimators of their population counterparts ε_i. The general approach, then, will be to use $\mathbf{X}$ and $\mathbf{e}$ to devise an estimator of $\mathbf{Q}_*$.

Consider the heteroscedasticity case first. We seek an estimator of

$$\mathbf{Q}_* = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\mathbf{x}_i\mathbf{x}_i'.$$

White (1980) has shown that under very general conditions, the estimator

$$\mathbf{S}_0 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \mathbf{x}_i \mathbf{x}_i' \tag{10-13}$$

has

$$\text{plim}\, \mathbf{S}_0 = \text{plim}\, \mathbf{Q}_*.^6$$

We can sketch a proof of this result using the results we obtained in Section 5.2.[7] Note first that $\mathbf{Q}_*$ is not a parameter matrix in itself. It is a weighted sum of the outer products of the rows of $\mathbf{X}$ (or $\mathbf{Z}$ for the instrumental variables case). Thus, we seek not to "estimate" $\mathbf{Q}_*$, but to find a function of the sample data that will be arbitrarily close to this function of the population parameters as the sample size grows large. The distinction is important. We are not estimating the middle matrix in (10-10) or (10-12); we are attempting to construct a matrix from the sample data that will behave the same way that this matrix behaves. In essence, if $\mathbf{Q}_*$ converges to a finite positive matrix, then we would be looking for a function of the sample data that converges to the same matrix. Suppose that the true disturbances ε_i could be observed. Then each term in $\mathbf{Q}_*$ would equal $E[\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i' \mid \mathbf{x}_i]$. With some fairly mild assumptions about $\mathbf{x}_i$, then, we could invoke a law of large numbers (see Theorems D.2 through D.4.) to state that if $\mathbf{Q}_*$ has a probability limit, then

$$\text{plim} = \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 \mathbf{x}_i \mathbf{x}_i' = \text{plim}\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i'.$$

The final detail is to justify the replacement of ε_i with e_i in $\mathbf{S}_0$. The consistency of $\mathbf{b}$ for $\boldsymbol{\beta}$ is sufficient for the argument. (Actually, residuals based on *any* consistent estimator of $\boldsymbol{\beta}$ would suffice for this estimator, but as of now, $\mathbf{b}$ or $\mathbf{b}_{\text{IV}}$ is the only one in hand.) The end result is that the **White heteroscedasticity consistent estimator**

$$\begin{aligned}\text{Est.Asy. Var}[\mathbf{b}] &= \frac{1}{n}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 \mathbf{x}_i \mathbf{x}_i'\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \\ &= n(\mathbf{X}'\mathbf{X})^{-1}\mathbf{S}_0(\mathbf{X}'\mathbf{X})^{-1}\end{aligned} \tag{10-14}$$

can be used to estimate the asymptotic covariance matrix of $\mathbf{b}$.

This result is extremely important and useful.[8] It implies that without actually specifying the type of heteroscedasticity, we can still make appropriate inferences based on the results of least squares. This implication is especially useful if we are unsure of the precise nature of the heteroscedasticity (which is probably most of the time). We will pursue some examples in Chapter 11.

[6] See also Eicker (1967), Horn, Horn, and Duncan (1975), and MacKinnon and White (1985).

[7] We will give only a broad sketch of the proof. Formal results appear in White (1980) and (2001).

[8] Further discussion and some refinements may be found in Cragg (1982). Cragg shows how White's observation can be extended to devise an estimator that improves on the efficiency of ordinary least squares.

The extension of White's result to the more general case of autocorrelation is much more difficult. The natural counterpart for estimating

$$\mathbf{Q}_* = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}\mathbf{x}_i\mathbf{x}_j'$$

would be

$$\hat{\mathbf{Q}}_* = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}e_ie_j\mathbf{x}_i\mathbf{x}_j'. \tag{10-15}$$

But there are two problems with this estimator, one theoretical, which applies to $\mathbf{Q}_*$ as well, and one practical, which is specific to the latter.

Unlike the heteroscedasticity case, the matrix in (10-15) is $1/n$ times a sum of n^2 terms, so it is difficult to conclude yet that it will converge to anything at all. This application is most likely to arise in a time-series setting. To obtain convergence, it is necessary to assume that the terms involving unequal subscripts in (10-15) diminish in importance as n grows. A sufficient condition is that terms with subscript pairs $|i - j|$ grow smaller as the distance between them grows larger. In practical terms, observation pairs are progressively less correlated as their separation in time grows. Intuitively, if one can think of weights with the diagonal elements getting a weight of 1.0, then in the sum, the weights in the sum grow smaller as we move away from the diagonal. If we think of the sum of the weights rather than just the number of terms, then this sum falls off sufficiently rapidly that as n grows large, the sum is of order n rather than n^2. Thus, we achieve convergence of $\mathbf{Q}_*$ by assuming that the rows of $\mathbf{X}$ are well behaved and that the correlations diminish with increasing separation in time. (See Sections 5.3, 12.5, and 20.5 for a more formal statement of this condition.)

The practical problem is that $\hat{\mathbf{Q}}_*$ need not be positive definite. Newey and West (1987a) have devised an estimator that overcomes this difficulty:

$$\hat{\mathbf{Q}}_* = \mathbf{S}_0 + \frac{1}{n}\sum_{l=1}^{L}\sum_{t=l+1}^{n}w_le_te_{t-l}(\mathbf{x}_t\mathbf{x}_{t-l}' + \mathbf{x}_{t-l}\mathbf{x}_t'),$$
$$w_l = 1 - \frac{l}{(L+1)}. \tag{10-16}$$

The **Newey–West autocorrelation consistent covariance estimator** is surprisingly simple and relatively easy to implement.[9] There is a final problem to be solved. It must be determined in advance how large L is to be. We will examine some special cases in Chapter 12, but in general, there is little theoretical guidance. Current practice specifies $L \approx T^{1/4}$. Unfortunately, the result is not quite as crisp as that for the heteroscedasticity consistent estimator.

We have the result that $\mathbf{b}$ and $\mathbf{b}_{IV}$ are asymptotically normally distributed, and we have an appropriate estimator for the asymptotic covariance matrix. We have not specified the distribution of the disturbances, however. Thus, for inference purposes, the F statistic is approximate at best. Moreover, for more involved hypotheses, the likelihood ratio and Lagrange multiplier tests are unavailable. That leaves the Wald

[9]Both estimators are now standard features in modern econometrics computer programs. Further results on different weighting schemes may be found in Hayashi (2000, pp. 406–410).

statistic, including asymptotic "t ratios," as the main tool for statistical inference. We will examine a number of applications in the chapters to follow.

The White and Newey–West estimators are standard in the econometrics literature. We will encounter them at many points in the discussion to follow.

10.4 GENERALIZED METHOD OF MOMENTS ESTIMATION

We will analyze this estimation technique in some detail in Chapter 18, so we will only sketch the important results here. It is useful to consider the instrumental variables case, as it is fairly general and we can easily specialize it to the simpler regression model if that is appropriate. Thus, we depart from the model specification in (10-1), but at this point, we no longer require that $E[\varepsilon_i \mid \mathbf{x}_i] = 0$. Instead, we adopt the instrumental variables formulation in Section 10.2.4. That is, our model is

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$
$$E[\varepsilon_i \mid \mathbf{z}_i] = 0$$

for K variables in $\mathbf{x}_i$ and for some set of L instrumental variables, $\mathbf{z}_i$, where $L \geq K$. The earlier case of the generalized regression model arises if $\mathbf{z}_i = \mathbf{x}_i$, and the classical regression form results if we add $\boldsymbol{\Omega} = \mathbf{I}$ as well, so this is a convenient encompassing model framework.

In the next section on generalized least squares estimation, we will consider two cases, first with a known $\boldsymbol{\Omega}$, then with an unknown $\boldsymbol{\Omega}$ that must be estimated. In estimation by the generalized method of moments neither of these approaches is relevant because we begin with much less (assumed) knowledge about the data generating process. In particular, we will consider three cases:

- Classical regression: $\text{Var}[\varepsilon_i \mid \mathbf{X}, \mathbf{Z}] = \sigma^2$,
- Heteroscedasticity: $\text{Var}[\varepsilon_i \mid \mathbf{X}, \mathbf{Z}] = \sigma_i^2$,
- Generalized model: $\text{Cov}[\varepsilon_t, \varepsilon_s \mid \mathbf{X}, \mathbf{Z}] = \sigma^2\omega_{ts}$,

where $\mathbf{Z}$ and $\mathbf{X}$ are the $n \times L$ and $n \times K$ observed data matrices. (We assume, as will often be true, that the fully general case will apply in a time series setting. Hence the change in the subscripts.) *No specific distribution is assumed for the disturbances, conditional or unconditional.*

The assumption $E[\varepsilon_i \mid \mathbf{z}_i] = 0$ implies the following **orthogonality condition:**

$$\text{Cov}[\mathbf{z}_i, \varepsilon_i,] = \mathbf{0}, \quad \text{or} \quad E[\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}.$$

By summing the terms, we find that this further implies the **population moment equation,**

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})\right] = E[\bar{\mathbf{m}}(\boldsymbol{\beta})] = \mathbf{0}. \tag{10-17}$$

This relationship suggests how we might now proceed to estimate $\boldsymbol{\beta}$. Note, in fact, that if $\mathbf{z}_i = \mathbf{x}_i$, then this is just the population counterpart to the least squares normal equations.

So, as a guide to estimation, this would return us to least squares. Suppose, we now translate this population expectation into a sample analog, and use that as our guide for estimation. That is, if the population relationship holds for the true parameter vector, $\boldsymbol{\beta}$, suppose we attempt to mimic this result with a sample counterpart, or **empirical moment equation,**

$$\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})\right] = \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_i(\hat{\boldsymbol{\beta}})\right] = \bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \mathbf{0}. \tag{10-18}$$

In the absence of other information about the data generating process, we can use the empirical moment equation as the basis of our estimation strategy.

The empirical moment condition is L equations (the number of variables in $\mathbf{Z}$) in K unknowns (the number of parameters we seek to estimate). There are three possibilities to consider:

1. Underidentified: $L < K$. If there are fewer moment equations than there are parameters, then it will not be possible to find a solution to the equation system in (10-18). With no other information, such as restrictions which would reduce the number of free parameters, there is no need to proceed any further with this case.

For the identified cases, it is convenient to write (10-18) as

$$\bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \left(\frac{1}{n}\mathbf{Z}'\mathbf{y}\right) - \left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)\hat{\boldsymbol{\beta}}. \tag{10-19}$$

2. Exactly identified. If $L = K$, then you can easily show (we leave it as an exercise) that the single solution to our equation system is the familiar instrumental variables estimator,

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}. \tag{10-20}$$

3. Overidentified. If $L > K$, then there is no unique solution to the equation system $\bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$. In this instance, we need to formulate some strategy to choose an estimator. One intuitively appealing possibility which has served well thus far is "least squares." In this instance, that would mean choosing the estimator based on the criterion function

$$\text{Min}_{\boldsymbol{\beta}}\, q = \bar{\mathbf{m}}(\hat{\boldsymbol{\beta}})'\bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}).$$

We do keep in mind, that we will only be able to minimize this at some positive value; there is no exact solution to (10-18) in the overidentified case. Also, you can verify that if we treat the exactly identified case as if it were overidentified, that is, use least squares anyway, we will still obtain the IV estimator shown in (10-20) for the solution to case (2). For the overidentified case, the first order conditions are

$$\begin{aligned}\frac{\partial q}{\partial \boldsymbol{\beta}} &= 2\left(\frac{\partial \bar{\mathbf{m}}'(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\right)\bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = 2\bar{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}) \\ &= 2\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{y} - \frac{1}{n}\mathbf{Z}'\mathbf{X}\hat{\boldsymbol{\beta}}\right) = \mathbf{0}.\end{aligned} \tag{10-21}$$

We leave as exercise to show that the solution in both cases (2) and (3) is now

$$\hat{\boldsymbol{\beta}} = [(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{y}). \tag{10-22}$$

The estimator in (10-22) is a hybrid that we have not encountered before, though if $L = K$, then it does reduce to the earlier one in (10–20). (In the overidentified case, (10-22) is not an IV estimator, it is, as we have sought, a **method of moments estimator.**)

It remains to establish consistency and to obtain the asymptotic distribution and an asymptotic covariance matrix for the estimator. These are analyzed in detail in Chapter 18. Our purpose here is only to sketch the formal result, so we will merely claim the intermediate results we need:

ASSUMPTION GMM1. **Convergence of the moments.** The population moment converges in probability to its population counterpart. That is, $\bar{\mathbf{m}}(\boldsymbol{\beta}) \rightarrow \mathbf{0}$. Different circumstances will produce different kinds of convergence, but we will require it in some form. For the simplest cases, such as a model of heteroscedasticity, this will be convergence in mean square. Certain time series models that involve correlated observations will necessitate some other form of convergence. But, in any of the cases we consider, we will require the general result, plim $\bar{\mathbf{m}}(\boldsymbol{\beta}) = \mathbf{0}$.

ASSUMPTION GMM2. **Identification.** The parameters are identified in terms of the moment equations. Identification means, essentially, that a large enough sample will contain sufficient information for us actually to estimate $\boldsymbol{\beta}$ consistently using the sample moments. There are two conditions which must be met—an **order condition,** which we have already assumed ($L \geq \boldsymbol{K}$), and a **rank condition,** which states that the moment equations are not redundant. The rank condition implies the order condition, so we need only formalize it:

Identification condition for GMM Estimation: The $L \times K$ matrix

$$\boldsymbol{\Gamma}(\boldsymbol{\beta}) = E[\bar{\mathbf{G}}(\boldsymbol{\beta})] = \text{plim}\, \bar{\mathbf{G}}(\boldsymbol{\beta}) = \text{plim} \frac{\partial \bar{\mathbf{m}}}{\partial \boldsymbol{\beta}'} = \text{plim} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \mathbf{m}_i}{\partial \boldsymbol{\beta}'}$$

must have (full) row rank equal to L.[10] Since this requires $L \geq K$, this implies the order condition. This assumption means that this derivative matrix converges in probability to its expectation. Note that we have assumed, in addition, that the derivatives, like the moments themselves, obey a law of large numbers—they converge in probability to their expectations.

ASSUMPTION GMM3. **Limiting Normal Distribution for the Sample Moments.** The population moment obeys a central limit theorem or some similar variant. Since we are studying a generalized regression model, Lindberg–Levy (D.19.) will be too narrow—the observations will have different variances. Lindberg–Feller (D.19.A) suffices in the heteroscedasticity case, but in the general case, we will ultimately require something more general. These theorems are discussed in Section 12.4 and invoked in Chapter 18.

[10]Strictly speaking, we only require that the row rank be at least as large as K, so there could be redundant, that is, functionally dependent, moments, so long as there are at least K that are functionally independent. The case of rank ($\boldsymbol{\Gamma}$) greater than or equal to K but less than L can be ignored.

It will follow from these assumptions (again, at this point we do this without proof) that the GMM estimators that we obtain are, in fact, consistent. By virtue of the Slutsky theorem, we can transfer our limiting results above to the empirical moment equations. A proof of consistency of the GMM estimator (pursued in Chapter 18) will be based on this result.

To obtain the asymptotic covariance matrix we will simply invoke a result we will obtain more formally in Chapter 18 for generalized method of moments estimators. That is,

$$\text{Asy. Var}[\hat{\boldsymbol{\beta}}] = \frac{1}{n}[\boldsymbol{\Gamma}'\boldsymbol{\Gamma}]^{-1}\boldsymbol{\Gamma}'\{\text{Asy. Var}[\sqrt{n}\,\bar{\mathbf{m}}(\boldsymbol{\beta})]\}\boldsymbol{\Gamma}[\boldsymbol{\Gamma}'\boldsymbol{\Gamma}]^{-1}.$$

For the particular model we are studying here,

$$\bar{\mathbf{m}}(\boldsymbol{\beta}) = (1/n)(\mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{X}\boldsymbol{\beta}),$$
$$\bar{\mathbf{G}}(\boldsymbol{\beta}) = (1/n)\mathbf{Z}'\mathbf{X},$$
$$\boldsymbol{\Gamma}(\boldsymbol{\beta}) = \mathbf{Q}_{\mathbf{ZX}} \text{ (from Section 10.2.4).}$$

(You should check in the preceding expression that the dimensions of the particular matrices and the dimensions of the various products produce the correctly configured matrix that we seek.) The remaining detail, which is the crucial one for the model we are examining, is for us to determine

$$\mathbf{V} = Asy.\ Var[\sqrt{n}\,\bar{\mathbf{m}}(\boldsymbol{\beta})].$$

Given the form of $\bar{\mathbf{m}}(\boldsymbol{\beta})$,

$$\mathbf{V} = \frac{1}{n}Var\left[\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma^2\omega_{ij}\mathbf{z}_i\mathbf{z}_j' = \sigma^2\frac{\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}}{n}$$

for the most general case. Note that this is precisely the expression that appears in (10-6), so the question that arose there arises here once again. That is, under what conditions will this converge to a constant matrix? We take the discussion there as given. The only remaining detail is how to estimate this matrix. The answer appears in Section 10.3, where we pursued this same question in connection with robust estimation of the asymptotic covariance matrix of the least squares estimator. To review then, what we have achieved to this point is to provide a theoretical foundation for the instrumental variables estimator. As noted earlier, this specializes to the least squares estimator. The estimators of $\mathbf{V}$ for our three cases will be

- Classical regression:

$$\hat{\mathbf{V}} = \frac{(\mathbf{e}'\mathbf{e}/n)}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i' = \frac{(\mathbf{e}'\mathbf{e}/n)}{n}\mathbf{Z}'\mathbf{Z}$$

- Heteroscedastic:

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n}e_i^2\mathbf{z}_i\mathbf{z}_i' \qquad \textbf{(10-23)}$$

- General:

$$\hat{\mathbf{V}} = \frac{1}{n}\left[\sum_{i=1}^{n} e_t^2 \mathbf{z}_t \mathbf{z}_t' + \sum_{l=1}^{L}\sum_{t=l+1}^{n}\left(1 - \frac{l}{(L+1)}\right) e_t e_{t-l}(\mathbf{z}_t \mathbf{z}_{t-l}' + \mathbf{z}_{t-l}\mathbf{z}_t')\right].$$

We should observe, that in each of these cases, we have actually used some information about the structure of $\mathbf{\Omega}$. If it is known only that the terms in $\bar{\mathbf{m}}(\boldsymbol{\beta})$ are uncorrelated, then there is a convenient estimator available,

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{m}_i(\hat{\boldsymbol{\beta}})\mathbf{m}_i(\hat{\boldsymbol{\beta}})'$$

that is, the natural, empirical variance estimator. Note that this is what is being used in the heteroscedasticity case directly above.

Collecting all the terms so far, then, we have

$$\begin{aligned} Est.Asy.\ Var[\hat{\boldsymbol{\beta}}] &= \frac{1}{n}[\bar{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\bar{\mathbf{G}}(\hat{\boldsymbol{\beta}})]^{-1}\bar{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\hat{\mathbf{V}}\bar{\mathbf{G}}(\hat{\boldsymbol{\beta}})[\bar{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\bar{\mathbf{G}}(\hat{\boldsymbol{\beta}})]^{-1} \\ &= n[(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})\hat{\mathbf{V}}(\mathbf{Z}'\mathbf{X})[(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{X})]^{-1}. \end{aligned} \tag{10-24}$$

The preceding would seem to endow the least squares or method of moments estimators with some degree of optimality, but that is not the case. We have only provided them with a different statistical motivation (and established consistency). We now consider the question of whether, since this is the generalized regression model, there is some better (more efficient) means of using the data. As before, we merely sketch the results.

The class of minimum distance estimators is defined by the solutions to the criterion function

$$\text{Min}_{\boldsymbol{\beta}}\ q = \bar{\mathbf{m}}(\boldsymbol{\beta})'\mathbf{W}\bar{\mathbf{m}}(\boldsymbol{\beta}),$$

where $\mathbf{W}$ is *any* positive definite **weighting matrix.** Based on the assumptions made above, we will have the following theorem, which we claim without proof at this point:

THEOREM 10.5 Minimum Distance Estimators
If $\text{plim}\ \bar{\mathbf{m}}(\boldsymbol{\beta}) = \mathbf{0}$ *and if* $\mathbf{W}$ *is a positive definite matrix, then* $\text{plim}\ \hat{\boldsymbol{\beta}} = \text{Argmin}[q = \bar{\mathbf{m}}(\boldsymbol{\beta})'\mathbf{W}\bar{\mathbf{m}}(\boldsymbol{\beta})] = \boldsymbol{\beta}$. *The minimum distance estimator is consistent. It is also asymptotically normally distributed and has asymptotic covariance matrix*

$$Asy.\ Var[\hat{\boldsymbol{\beta}}_{MD}] = \frac{1}{n}[\bar{\mathbf{G}}'\mathbf{W}\bar{\mathbf{G}}]^{-1}\bar{\mathbf{G}}'\mathbf{W}\mathbf{V}\mathbf{W}\bar{\mathbf{G}}[\bar{\mathbf{G}}'\mathbf{W}\bar{\mathbf{G}}]^{-1}.$$

Note that our entire preceding analysis was of the simplest minimum distance estimator, which has $\mathbf{W} = \mathbf{I}$. The obvious question now arises, if any $\mathbf{W}$ produces a consistent estimator, is any $\mathbf{W}$ better than any other one, or is it simply arbitrary? There is a firm answer, for which we have to consider two cases separately:

- Exactly identified case: If $L = K$; that is, if the number of moment conditions is the same as the number of parameters being estimated, then $\mathbf{W}$ is irrelevant to the solution, so on the basis of simplicity alone, the optimal $\mathbf{W}$ is $\mathbf{I}$.

- Overidentified case: In this case, the "optimal" weighting matrix, that is, the $\mathbf{W}$ which produces the most efficient estimator is $\mathbf{W} = \mathbf{V}^{-1}$. That is, the best weighting matrix is the inverse of the asymptotic covariance of the moment vector.

THEOREM 10.6 Generalized Method of Moments Estimator
The Minimum Distance Estimator obtained by using $\mathbf{W} = \mathbf{V}^{-1}$ *is the Generalized Method of Moments, or GMM estimator. The* ***GMM estimator*** *is consistent, asymptotically normally distributed, and has asymptotic covariance matrix equal to*

$$Asy.\ Var[\hat{\boldsymbol{\beta}}_{GMM}] = \frac{1}{n}[\bar{\mathbf{G}}'\mathbf{V}^{-1}\bar{\mathbf{G}}]^{-1}.$$

For the generalized regression model, these are

$$\hat{\boldsymbol{\beta}}_{GMM} = [(\mathbf{X}'\mathbf{Z})\hat{\mathbf{V}}^{-1}(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})\hat{\mathbf{V}}^{-1}(\mathbf{Z}'\mathbf{y})$$

and

$$Asy.\ Var[\hat{\boldsymbol{\beta}}_{GMM}] = [(\mathbf{X}'\mathbf{Z})\hat{\mathbf{V}}(\mathbf{Z}'\mathbf{X})]^{-1}.$$

We conclude this discussion by tying together what should seem to be a loose end. The GMM estimator is computed as the solution to

$$\text{Min}_{\boldsymbol{\beta}}\ q = \bar{\mathbf{m}}(\boldsymbol{\beta})'\left\{Asy.\ Var[\sqrt{n}\,\bar{\mathbf{m}}(\boldsymbol{\beta})]\right\}^{-1}\bar{\mathbf{m}}(\boldsymbol{\beta}),$$

which suggests that the weighting matrix is a function of the thing we are trying to estimate. The process of GMM estimation will have to proceed in two steps: Step 1 is to obtain an estimate of $\mathbf{V}$, then Step 2 will consist of using the inverse of this $\mathbf{V}$ as the weighting matrix in computing the GMM estimator. We will return to this in Chapter 18, so we note directly, the following is a common strategy:

Step 1. Use $\mathbf{W} = \mathbf{I}$ to obtain a consistent estimator of $\boldsymbol{\beta}$. Then, estimate $\mathbf{V}$ with

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n} e_i^2\mathbf{z}_i\mathbf{z}_i'$$

in the heteroscedasticity case (i.e., the White estimator) or, for the more general case, the Newey–West estimator in (10-23).

Step 2. Use $\mathbf{W} = \hat{\mathbf{V}}^{-1}$ to compute the GMM estimator.

At this point, the observant reader should have noticed that in all of the preceding, we have never actually encountered the simple instrumental variables estimator that

we introduced in Section 5.4. In order to obtain this estimator, we must revert back to the classical, that is homoscedastic and nonautocorrelated disturbances case. In that instance, the weighting matrix in Theorem 10.5 will be $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$ and we will obtain the apparently missing result.

10.5 EFFICIENT ESTIMATION BY GENERALIZED LEAST SQUARES

Efficient estimation of $\boldsymbol{\beta}$ in the generalized regression model requires knowledge of $\boldsymbol{\Omega}$. To begin, it is useful to consider cases in which $\boldsymbol{\Omega}$ is a known, symmetric, positive definite matrix. This assumption will occasionally be true, but in most models, $\boldsymbol{\Omega}$ will contain unknown parameters that must also be estimated. We shall examine this case in Section 10.6.

10.5.1 GENERALIZED LEAST SQUARES (GLS)

Since $\boldsymbol{\Omega}$ is a positive definite symmetric matrix, it can be factored into

$$\boldsymbol{\Omega} = \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}',$$

where the columns of $\mathbf{C}$ are the characteristic vectors of $\boldsymbol{\Omega}$ and the characteristic roots of $\boldsymbol{\Omega}$ are arrayed in the diagonal matrix $\boldsymbol{\Lambda}$. Let $\boldsymbol{\Lambda}^{1/2}$ be the diagonal matrix with ith diagonal element $\sqrt{\lambda_i}$, and let $\mathbf{T} = \mathbf{C}\boldsymbol{\Lambda}^{1/2}$. Then $\boldsymbol{\Omega} = \mathbf{T}\mathbf{T}'$. Also, let $\mathbf{P}' = \mathbf{C}\boldsymbol{\Lambda}^{-1/2}$, so $\boldsymbol{\Omega}^{-1} = \mathbf{P}'\mathbf{P}$. Premultiply the model in (10-1) by $\mathbf{P}$ to obtain

$$\mathbf{P}\mathbf{y} = \mathbf{P}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\varepsilon}$$

or

$$\mathbf{y}_* = \mathbf{X}_*\boldsymbol{\beta} + \boldsymbol{\varepsilon}_*. \tag{10-25}$$

The variance of $\boldsymbol{\varepsilon}_*$ is

$$E[\boldsymbol{\varepsilon}_*\boldsymbol{\varepsilon}_*'] = \mathbf{P}\sigma^2\boldsymbol{\Omega}\mathbf{P}' = \sigma^2\mathbf{I},$$

so the classical regression model applies to this transformed model. Since $\boldsymbol{\Omega}$ is known, $\mathbf{y}_*$ and $\mathbf{X}_*$ are observed data. In the classical model, ordinary least squares is efficient; hence,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{y}_* \\ &= (\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{y} \\ &= (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}\end{aligned}$$

is the efficient estimator of $\boldsymbol{\beta}$. This estimator is the **generalized least squares (GLS)** or Aitken (1935) estimator of $\boldsymbol{\beta}$. This estimator is in contrast to the ordinary least squares (OLS) estimator, which uses a "weighting matrix," $\mathbf{I}$, instead of $\boldsymbol{\Omega}^{-1}$. By appealing to the classical regression model in (10-25), we have the following theorem, which includes the generalized regression model analogs to our results of Chapters 4 and 5.

THEOREM 10.7 Properties of the Generalized Least Squares Estimator

If $E[\boldsymbol{\varepsilon}_* \mid \mathbf{X}_*] = \mathbf{0}$, *then*

$$E[\hat{\boldsymbol{\beta}} \mid \mathbf{X}_*] = E[(\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{y}_* \mid \mathbf{X}_*] = \boldsymbol{\beta} + E[(\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\boldsymbol{\varepsilon}_* \mid \mathbf{X}_*] = \boldsymbol{\beta}$$

The GLS estimator $\hat{\boldsymbol{\beta}}$ *is unbiased. This result is equivalent to* $E[\mathbf{P}\boldsymbol{\varepsilon} \mid \mathbf{PX}] = \mathbf{0}$, *but since* $\mathbf{P}$ *is a matrix of known constants, we return to the familiar requirement* $E[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \mathbf{0}$. *The requirement that the regressors and disturbances be uncorrelated is unchanged.*

The GLS estimator is consistent if $\operatorname{plim}(1/n)\mathbf{X}_*'\mathbf{X}_* = \mathbf{Q}_*$, *where* $\mathbf{Q}_*$ *is a finite positive definite matrix. Making the substitution, we see that this implies*

$$\operatorname{plim}[(1/n)\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1} = \mathbf{Q}_*^{-1}. \tag{10-26}$$

We require the transformed data $\mathbf{X}_* = \mathbf{PX}$, *not the original data* $\mathbf{X}$, *to be well behaved.*[11] *Under the assumption in (10-1), the following hold:*

The GLS estimator is asymptotically normally distributed, with mean $\boldsymbol{\beta}$ *and sampling variance*

$$\operatorname{Var}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}_*] = \sigma^2(\mathbf{X}_*'\mathbf{X}_*)^{-1} = \sigma^2(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}. \tag{10-27}$$

The GLS estimator $\hat{\boldsymbol{\beta}}$ *is the minimum variance linear unbiased estimator in the generalized regression model. This statement follows by applying the Gauss–Markov theorem to the model in (10-25). The result in Theorem 10.7 is* ***Aitken's*** (1935) ***Theorem,*** *and* $\hat{\boldsymbol{\beta}}$ *is sometimes called the Aitken estimator. This broad result includes the Gauss–Markov theorem as a special case when* $\boldsymbol{\Omega} = \mathbf{I}$.

For testing hypotheses, we can apply the full set of results in Chapter 6 to the transformed model in (10-25). For testing the J linear restrictions, $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, the appropriate statistic is

$$F[J, n-K] = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{R}\hat{\sigma}^2(\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})}{J} = \frac{(\hat{\boldsymbol{\varepsilon}}_c'\hat{\boldsymbol{\varepsilon}}_c - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})/J}{\hat{\sigma}^2},$$

where the residual vector is

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}}$$

and

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-K} = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n-K}. \tag{10-28}$$

The constrained GLS residuals, $\hat{\boldsymbol{\varepsilon}}_c = \mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}}_c$, are based on

$$\hat{\boldsymbol{\beta}}_c = \hat{\boldsymbol{\beta}} - [\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}).^{12}$$

[11] Once again, to allow a time trend, we could weaken this assumption a bit.

[12] Note that this estimator is the constrained OLS estimator using the transformed data.

To summarize, all the results for the classical model, including the usual inference procedures, apply to the transformed model in (10-25).

There is no precise counterpart to R^2 in the generalized regression model. Alternatives have been proposed, but care must be taken when using them. For example, one choice is the R^2 in the transformed regression, (10-25). But this regression need not have a constant term, so the R^2 is not bounded by zero and one. Even if there is a constant term, the transformed regression is a computational device, not the model of interest. That a good (or bad) fit is obtained in the "model" in (10-25) may be of no interest; the dependent variable in that model y_* is different from the one in the model as originally specified. The usual R^2 often suggests that the fit of the model is improved by a correction for heteroscedasticity and degraded by a correction for autocorrelation, but both changes can often be attributed to the computation of y_*. A more appealing fit measure might be based on the residuals from the original model once the GLS estimator is in hand, such as

$$R_G^2 = 1 - \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Like the earlier contender, however, this measure is not bounded in the unit interval. In addition, this measure cannot be reliably used to compare models. The generalized least squares estimator minimizes the **generalized sum of squares**

$$\boldsymbol{\varepsilon}_*'\boldsymbol{\varepsilon}_* = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

not $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$. As such, there is no assurance, for example, that dropping a variable from the model will result in a decrease in R_G^2, as it will in R^2. Other goodness-of-fit measures, designed primarily to be a function of the sum of squared residuals (raw or weighted by $\boldsymbol{\Omega}^{-1}$) and to be bounded by zero and one, have been proposed.[13] Unfortunately, they all suffer from at least one of the previously noted shortcomings. The R^2-like measures in this setting are purely descriptive.

10.5.2 FEASIBLE GENERALIZED LEAST SQUARES

To use the results of Section 10.5.1, $\boldsymbol{\Omega}$ must be known. If $\boldsymbol{\Omega}$ contains unknown parameters that must be estimated, then generalized least squares is not feasible. But with an unrestricted $\boldsymbol{\Omega}$, there are $n(n+1)/2$ additional parameters in $\sigma^2\boldsymbol{\Omega}$. This number is far too many to estimate with n observations. Obviously, some structure must be imposed on the model if we are to proceed.

The typical problem involves a small set of parameters $\boldsymbol{\theta}$ such that $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\theta})$. A commonly used formula in time series settings is

$$\boldsymbol{\Omega}(\rho) = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{n-2} \\ & & & & \vdots & \\ \rho^{n-1} & \rho^{n-2} & & \cdots & & 1 \end{bmatrix},$$

[13] See, example, Judge et al. (1985, p. 32) and Buse (1973).

which involves only one additional unknown parameter. A model of heteroscedasticity that also has only one new parameter is

$$\sigma_i^2 = \sigma^2 z_i^{\theta}. \tag{10-29}$$

Suppose, then, that $\hat{\theta}$ is a consistent estimator of θ. (We consider later how such an estimator might be obtained.) To make GLS estimation feasible, we shall use $\hat{\Omega} = \Omega(\hat{\theta})$ instead of the true Ω. The issue we consider here is whether using $\Omega(\hat{\theta})$ requires us to change any of the results of Section 10.5.1.

It would seem that if plim $\hat{\theta} = \theta$, then using $\hat{\Omega}$ is asymptotically equivalent to using the true Ω.[14] Let the **feasible generalized least squares (FGLS)** estimator be denoted

$$\hat{\hat{\beta}} = (\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}^{-1}\mathbf{y}.$$

Conditions that imply that $\hat{\hat{\beta}}$ is asymptotically equivalent to $\hat{\beta}$ are

$$\text{plim}\left[\left(\frac{1}{n}\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X}\right) - \left(\frac{1}{n}\mathbf{X}'\Omega^{-1}\mathbf{X}\right)\right] = \mathbf{0} \tag{10-30}$$

and

$$\text{plim}\left[\left(\frac{1}{\sqrt{n}}\mathbf{X}'\hat{\Omega}^{-1}\boldsymbol{\varepsilon}\right) - \left(\frac{1}{\sqrt{n}}\mathbf{X}'\Omega^{-1}\boldsymbol{\varepsilon}\right)\right] = \mathbf{0}. \tag{10-31}$$

The first of these equations states that if the weighted sum of squares matrix based on the true Ω converges to a positive definite matrix, then the one based on $\hat{\Omega}$ converges to the same matrix. We are assuming that this is true. In the second condition, if the *transformed* regressors are well behaved, then the right-hand side sum will have a limiting normal distribution. This condition is exactly the one we used in Chapter 5 to obtain the asymptotic distribution of the least squares estimator; here we are using the same results for $\mathbf{X}_*$ and $\boldsymbol{\varepsilon}_*$. Therefore, (10-31) requires the same condition to hold when Ω is replaced with $\hat{\Omega}$.[15]

These conditions, in principle, must be verified on a case-by-case basis. Fortunately, in most familiar settings, they are met. If we assume that they are, then the FGLS estimator based on $\hat{\theta}$ has the same asymptotic properties as the GLS estimator. This result is extremely useful. Note, especially, the following theorem.

THEOREM 10.8 Efficiency of the FGLS Estimator
An asymptotically efficient FGLS estimator does not require that we have an efficient estimator of θ; only a consistent one is required to achieve full efficiency for the FGLS estimator.

[14] This equation is sometimes denoted plim $\hat{\Omega} = \Omega$. Since Ω is $n \times n$, it cannot have a probability limit. We use this term to indicate convergence element by element.

[15] The condition actually requires only that if the right-hand sum has *any* limiting distribution, then the left-hand one has the same one. Conceivably, this distribution might not be the normal distribution, but that seems unlikely except in a specially constructed, theoretical case.

Except for the simplest cases, the finite-sample properties and exact distributions of FGLS estimators are unknown. The asymptotic efficiency of FGLS estimators may not carry over to small samples because of the variability introduced by the estimated $\boldsymbol{\Omega}$. Some analyses for the case of heteroscedasticity are given by Taylor (1977). A model of autocorrelation is analyzed by Griliches and Rao (1969). In both studies, the authors find that, over a broad range of parameters, FGLS is more efficient than least squares. But if the departure from the classical assumptions is not too severe, then least squares may be more efficient than FGLS in a small sample.

10.6 MAXIMUM LIKELIHOOD ESTIMATION

This section considers efficient estimation when the disturbances are normally distributed. As before, we consider two cases, first, to set the stage, the benchmark case of known $\boldsymbol{\Omega}$, and, second, the more common case of unknown $\boldsymbol{\Omega}$.[16]

If the disturbances are multivariate normally distributed, then the log-likelihood function for the sample is

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\ln|\boldsymbol{\Omega}|. \tag{10-32}$$

Since $\boldsymbol{\Omega}$ is a matrix of known constants, the maximum likelihood estimator of $\boldsymbol{\beta}$ is the vector that minimizes the **generalized sum of squares,**

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

(hence the name *generalized least squares*). The necessary conditions for maximizing L are

$$\begin{aligned}
\frac{\partial \ln L}{\partial \boldsymbol{\beta}} &= \frac{1}{\sigma^2}\mathbf{X}'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma^2}\mathbf{X}_*'(\mathbf{y}_*-\mathbf{X}_*\boldsymbol{\beta}) = \mathbf{0},\\
\frac{\partial \ln L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \qquad \text{(10-33)}\\
&= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y}_*-\mathbf{X}_*\boldsymbol{\beta})'(\mathbf{y}_*-\mathbf{X}_*\boldsymbol{\beta}) = 0.
\end{aligned}$$

The solutions are the OLS estimators using the transformed data:

$$\hat{\boldsymbol{\beta}}_{\mathrm{ML}} = (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{y}_* = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}, \tag{10-34}$$

$$\begin{aligned}
\hat{\sigma}^2_{\mathrm{ML}} &= \frac{1}{n}(\mathbf{y}_*-\mathbf{X}_*\hat{\boldsymbol{\beta}})'(\mathbf{y}_*-\mathbf{X}_*\hat{\boldsymbol{\beta}})\\
&= \frac{1}{n}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}),
\end{aligned} \tag{10-35}$$

which implies that with normally distributed disturbances, generalized least squares is

[16]The method of maximum likelihood estimation is developed in Chapter 17.

also maximum likelihood. As in the classical regression model, the maximum likelihood estimator of σ^2 is biased. An unbiased estimator is the one in (10-28). The conclusion, which would be expected, is that when $\boldsymbol{\Omega}$ is known, the maximum likelihood estimator is generalized least squares.

When $\boldsymbol{\Omega}$ is unknown and must be estimated, then it is necessary to maximize the log likelihood in (10-32) with respect to the full set of parameters $[\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Omega}]$ simultaneously. Since an unrestricted $\boldsymbol{\Omega}$ alone contains $n(n+1)/2 - 1$ parameters, it is clear that some restriction will have to be placed on the structure of $\boldsymbol{\Omega}$ in order for estimation to proceed. We will examine several applications in which $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\theta})$ for some smaller vector of parameters in the next two chapters, so we will note only a few general results at this point.

(a) For a given value of $\boldsymbol{\theta}$ the estimator of $\boldsymbol{\beta}$ would be feasible GLS and the estimator of σ^2 would be the estimator in (10-35).

(b) The likelihood equations for $\boldsymbol{\theta}$ will generally be complicated functions of $\boldsymbol{\beta}$ and σ^2, so joint estimation will be necessary. However, in many cases, for given values of $\boldsymbol{\beta}$ and σ^2, the estimator of $\boldsymbol{\theta}$ is straightforward. For example, in the model of (10-29), the iterated estimator of θ when $\boldsymbol{\beta}$ and σ^2 and a prior value of θ are given is the prior value plus the slope in the regression of $(e_i^2/\hat{\sigma}_i^2 - 1)$ on z_i.

The second step suggests a sort of back and forth iteration for this model that will work in many situations—starting with, say, OLS, iterating back and forth between (a) and (b) until convergence will produce the joint maximum likelihood estimator. This situation was examined by Oberhofer and Kmenta (1974), who showed that under some fairly weak requirements, most importantly that $\boldsymbol{\theta}$ not involve σ^2 or any of the parameters in $\boldsymbol{\beta}$, this procedure would produce the maximum likelihood estimator. Another implication of this formulation which is simple to show (we leave it as an exercise) is that under the Oberhofer and Kmenta assumption, the asymptotic covariance matrix of the estimator is the same as the GLS estimator. This is the same whether $\boldsymbol{\Omega}$ is known or estimated, which means that if $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ have no parameters in common, then *exact knowledge of $\boldsymbol{\Omega}$ brings no gain in asymptotic efficiency in the estimation of $\boldsymbol{\beta}$ over estimation of $\boldsymbol{\beta}$ with a consistent estimator of $\boldsymbol{\Omega}$.*

10.7 SUMMARY AND CONCLUSIONS

This chapter has introduced a major extension of the classical linear model. By allowing for heteroscedasticity and autocorrelation in the disturbances, we expand the range of models to a large array of frameworks. We will explore these in the next several chapters. The formal concepts introduced in this chapter include how this extension affects the properties of the least squares estimator, how an appropriate estimator of the asymptotic covariance matrix of the least squares estimator can be computed in this extended modeling framework, and, finally, how to use the information about the variances and covariances of the disturbances to obtain an estimator that is more efficient than ordinary least squares.

Key Terms and Concepts

- Aitken's Theorem
- Asymptotic properties
- Autocorrelation
- Efficient estimator
- Feasible GLS
- Finite sample properties
- Generalized least squares (GLS)
- Generalized regression model
- GMM estimator
- Heteroscedasticity
- Instrumental variables estimator
- Method of moments estimator
- Newey–West estimator
- Nonlinear least squares estimator
- Order condition
- Ordinary least squares (OLS)
- Orthogonality condition
- Panel data
- Parametric
- Population moment equation
- Rank condition
- Robust estimation
- Semiparametric
- Weighting matrix
- White estimator

Exercises

1. What is the covariance matrix, $\text{Cov}[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}} - \mathbf{b}]$, of the GLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$ and the difference between it and the OLS estimator, $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$? The result plays a pivotal role in the development of specification tests in Hausman (1978).
2. This and the next two exercises are based on the test statistic usually used to test a set of J linear restrictions in the generalized regression model:

$$F[J, n-K] = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})/J}{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(n-K)},$$

   where $\hat{\boldsymbol{\beta}}$ is the GLS estimator. Show that if $\boldsymbol{\Omega}$ is known, if the disturbances are normally distributed and if the null hypothesis, $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, is true, then this statistic is exactly distributed as F with J and $n - K$ degrees of freedom. What assumptions about the regressors are needed to reach this conclusion? Need they be nonstochastic?
3. Now suppose that the disturbances are not normally distributed, although $\boldsymbol{\Omega}$ is still known. Show that the limiting distribution of previous statistic is $(1/J)$ times a chi-squared variable with J degrees of freedom. (Hint: The denominator converges to σ^2.) Conclude that in the generalized regression model, the limiting distribution of the Wald statistic

$$W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'\{\mathbf{R}(\text{Est. Var}[\hat{\boldsymbol{\beta}}])\mathbf{R}'\}^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})$$

   is chi-squared with J degrees of freedom, regardless of the distribution of the disturbances, as long as the data are otherwise well behaved. Note that in a finite sample, the true distribution may be approximated with an $F[J, n-K]$ distribution. It is a bit ambiguous, however, to interpret this fact as implying that the statistic is asymptotically distributed as F with J and $n - K$ degrees of freedom, because the limiting distribution used to obtain our result is the chi-squared, not the F. In this instance, the $F[J, n-K]$ is a random variable that tends asymptotically to the chi-squared variate.
4. Finally, suppose that $\boldsymbol{\Omega}$ must be estimated, but that assumptions (10-27) and (10-31) are met by the estimator. What changes are required in the development of the previous problem?

5. In the generalized regression model, if the K columns of $\mathbf{X}$ are characteristic vectors of $\mathbf{\Omega}$, then ordinary least squares and generalized least squares are identical. (The result is actually a bit broader; $\mathbf{X}$ may be any linear combination of exactly K characteristic vectors. This result is **Kruskal's Theorem.**)
   a. Prove the result directly using matrix algebra.
   b. Prove that if $\mathbf{X}$ contains a constant term and if the remaining columns are in deviation form (so that the column sum is zero), then the model of Exercise 8 below is one of these cases. (The seemingly unrelated regressions model with identical regressor matrices, discussed in Chapter 14, is another.)
6. In the generalized regression model, suppose that $\mathbf{\Omega}$ is known.
   a. What is the covariance matrix of the OLS and GLS estimators of $\boldsymbol{\beta}$?
   b. What is the covariance matrix of the OLS residual vector $\mathbf{e} = \mathbf{y} - \mathbf{Xb}$?
   c. What is the covariance matrix of the GLS residual vector $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$?
   d. What is the covariance matrix of the OLS and GLS residual vectors?
7. Suppose that y has the pdf $f(y \mid \mathbf{x}) = (1/\mathbf{x}'\boldsymbol{\beta})e^{-y/(\boldsymbol{\beta}'\mathbf{x})}$, $y > 0$.

   Then $E[y \mid \mathbf{x}] = \boldsymbol{\beta}'\mathbf{x}$ and $\text{Var}[y \mid \mathbf{x}] = (\boldsymbol{\beta}'\mathbf{x})^2$. For this model, prove that GLS and MLE are the same, even though this distribution involves the same parameters in the conditional mean function and the disturbance variance.
8. Suppose that the regression model is $y = \mu + \varepsilon$, where ε has a zero mean, constant variance, and equal correlation ρ across observations. Then $\text{Cov}[\varepsilon_i, \varepsilon_j] = \sigma^2\rho$ if $i \neq j$. Prove that the least squares estimator of μ is inconsistent. Find the characteristic roots of $\mathbf{\Omega}$ and show that Condition 2. after Theorem 10.2 is violated.