Chapter 9

The Generalized Method of Moments

9.1 Introduction

The models we have considered in earlier chapters have all been regression models of one sort or another. In this chapter and the next, we introduce more general types of models, along with a general method for performing estimation and inference on them. This technique is called the **generalized method of moments**, or **GMM**, and it includes as special cases all the methods we have so far developed for regression models.

As we explained in Section 3.1, a model is represented by a set of DGPs. Each DGP in the model is characterized by a parameter vector, which we will normally denote by $\boldsymbol{\beta}$ in the case of regression functions and by $\boldsymbol{\theta}$ in the general case. The starting point for GMM estimation is to specify functions, which, for any DGP in the model, depend both on the data generated by that DGP and on the model parameters. When these functions are evaluated at the parameters that correspond to the DGP that generated the data, their expectation must be zero.

As a simple example, consider the linear regression model $y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t$. An important part of the model specification is that the error terms have mean zero. These error terms are unobservable, because the parameters $\boldsymbol{\beta}$ of the regression function are unknown. But we can define the residuals $u_t(\boldsymbol{\beta}) \equiv y_t - \boldsymbol{X}_t\boldsymbol{\beta}$ as functions of the observed data and the unknown model parameters, and these functions provide what we need for GMM estimation. If the residuals are evaluated at the parameter vector $\boldsymbol{\beta}_0$ associated with the true DGP, they have mean zero under that DGP, but if they are evaluated at some $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, they do not have mean zero. In Chapter 1, we used this fact to develop a method-of-moments (MM) estimator for the parameter vector $\boldsymbol{\beta}$ of the regression function. As we will see in the next section, the various GMM estimators of $\boldsymbol{\beta}$ include as a special case the MM (or OLS) estimator developed in Chapter 1.

In Chapter 6, when we dealt with nonlinear regression models, and again in Chapter 8, we used instrumental variables along with residuals in order to develop MM estimators. The use of instrumental variables is also an essential aspect of GMM, and in this chapter we will once again make use of the various kinds of optimal instruments that were useful in Chapters 6 and 8 in order to develop a wide variety of estimators that are asymptotically efficient for a wide variety of models.

We begin by considering, in the next section, a linear regression model with endogenous explanatory variables and an error covariance matrix that is not proportional to the identity matrix. Such a model requires us to combine the insights of both Chapters 7 and 8 in order to obtain asymptotically efficient estimates. In the process of doing so, we will see how GMM estimation works more generally, and we will be led to develop ways to estimate models with both heteroskedasticity and serial correlation of unknown form. In Section 9.3, we study in some detail the **heteroskedasticity and autocorrelation consistent**, or **HAC**, covariance matrix estimators that we briefly mentioned in Section 5.5. Then, in Section 9.4, we introduce a set of tests, based on **GMM criterion functions**, that are widely used for inference in conjunction with GMM estimation. In Section 9.5, we move beyond regression models to give a more formal and advanced presentation of GMM, and we postpone to this section most of the proofs of consistency, asymptotic normality, and asymptotic efficiency for GMM estimators. In Section 9.6, which depends heavily on the more advanced treatment of the preceding section, we consider the **Method of Simulated Moments**, or **MSM**. This method allows us to obtain GMM estimates by simulation even when we cannot analytically evaluate the functions that play the same role as residuals for a regression model.

9.2 GMM Estimators for Linear Regression Models

Consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{uu}^\top) = \boldsymbol{\Omega}, \tag{9.01}$$

where there are n observations, and $\boldsymbol{\Omega}$ is an $n \times n$ covariance matrix. As in the previous chapter, some of the explanatory variables that form the $n \times k$ matrix $\boldsymbol{X}$ may not be predetermined with respect to the error terms $\boldsymbol{u}$. However, there is assumed to exist an $n \times l$ matrix of predetermined instrumental variables, $\boldsymbol{W}$, with $n > l$ and $l \geq k$, satisfying the condition $\mathrm{E}(u_t \,|\, \boldsymbol{W}_t) = 0$ for each row $\boldsymbol{W}_t$ of $\boldsymbol{W}$, $t = 1, \ldots, n$. Any column of $\boldsymbol{X}$ that is predetermined must also be a column of $\boldsymbol{W}$. In addition, we assume that, for all $t, s = 1, \ldots, n$, $\mathrm{E}(u_t u_s \,|\, \boldsymbol{W}_t, \boldsymbol{W}_s) = \omega_{ts}$, where ω_{ts} is the ts^{th} element of $\boldsymbol{\Omega}$. We will need this assumption later, because it allows us to see that

$$\begin{aligned}\mathrm{Var}(n^{-1/2}\boldsymbol{W}^\top\boldsymbol{u}) &= \tfrac{1}{n}\mathrm{E}(\boldsymbol{W}^\top\boldsymbol{uu}^\top\boldsymbol{W}) = \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathrm{E}(u_t u_s \boldsymbol{W}_t^\top\boldsymbol{W}_s)\\ &= \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathrm{E}\big(\mathrm{E}(u_t u_s \boldsymbol{W}_t^\top\boldsymbol{W}_s \,|\, \boldsymbol{W}_t, \boldsymbol{W}_s)\big)\end{aligned}$$

$$= \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathrm{E}(\omega_{ts}\boldsymbol{W}_t^{\top}\boldsymbol{W}_s) = \frac{1}{n}\mathrm{E}(\boldsymbol{W}^{\top}\boldsymbol{\Omega}\boldsymbol{W}). \tag{9.02}$$

The assumption that $\mathrm{E}(u_t \,|\, \boldsymbol{W}_t) = 0$ implies that, for all $t = 1, \ldots, n$,

$$\mathrm{E}\big(\boldsymbol{W}_t^{\top}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})\big) = \boldsymbol{0}. \tag{9.03}$$

These equations form a set of what we may call **theoretical moment conditions**. They were used in Chapter 8 as the starting point for MM estimation of the regression model (9.01). Each theoretical moment condition corresponds to a sample moment, or **empirical moment**, of the form

$$\frac{1}{n}\sum_{t=1}^{n} w_{ti}^{\top}(y_t - \boldsymbol{X}_t\boldsymbol{\beta}) = \frac{1}{n}\boldsymbol{w}_i^{\top}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{9.04}$$

where $\boldsymbol{w}_i$, $i = 1, \ldots, l$, is the i^{th} column of $\boldsymbol{W}$, and w_{ti} is the ti^{th} element. When $l = k$, we can set these sample moments equal to zero and solve the resulting k equations to obtain the simple IV estimator (8.12). When $l > k$, we must do as we did in Chapter 8 and select k independent linear combinations of the sample moments (9.04) in order to obtain an estimator.

Now let $\boldsymbol{J}$ be an $l \times k$ matrix with full column rank k, and consider the MM estimator obtained by using the k columns of $\boldsymbol{WJ}$ as instruments. This estimator solves the k equations

$$\boldsymbol{J}^{\top}\boldsymbol{W}^{\top}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}, \tag{9.05}$$

which are referred to as **sample moment conditions**, or just **moment conditions** when there is no ambiguity. They are also sometimes called **orthogonality conditions**, since they require that the vector of residuals should be orthogonal to the columns of $\boldsymbol{WJ}$. Let us assume that the data are generated by a DGP which belongs to the model (9.01), with coefficient vector $\boldsymbol{\beta}_0$ and covariance matrix $\boldsymbol{\Omega}_0$. Under this assumption, we have the following explicit expression, suitable for asymptotic analysis, for the estimator $\hat{\boldsymbol{\beta}}$ that solves (9.05):

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \big(n^{-1}\boldsymbol{J}^{\top}\boldsymbol{W}^{\top}\boldsymbol{X}\big)^{-1} n^{-1/2}\boldsymbol{J}^{\top}\boldsymbol{W}^{\top}\boldsymbol{u}. \tag{9.06}$$

From this, recalling (9.02), we find that the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$, that is, the covariance matrix of the plim of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, is

$$\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{J}^{\top}\boldsymbol{W}^{\top}\boldsymbol{X}\Big)^{-1}\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{J}^{\top}\boldsymbol{W}^{\top}\boldsymbol{\Omega}_0\boldsymbol{W}\boldsymbol{J}\Big)\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{W}\boldsymbol{J}\Big)^{-1}. \tag{9.07}$$

This matrix has the familiar sandwich form that we expect to see when an estimator is not asymptotically efficient.

The next step, as in Section 8.3, is to choose $\boldsymbol{J}$ so as to minimize the covariance matrix (9.07). We may reasonably expect that, with such a choice of $\boldsymbol{J}$, the covariance matrix would no longer have the form of a sandwich. The simplest choice of $\boldsymbol{J}$ that eliminates the sandwich in (9.07) is

$$\boldsymbol{J} = (\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{X}; \tag{9.08}$$

notice that, in the special case in which $\boldsymbol{\Omega}_0$ is proportional to $\mathbf{I}$, this expression reduces to the result (8.24) that we found in Section 8.3 as the solution for that special case. We can see, therefore, that (9.08) is the appropriate generalization of (8.24) when $\boldsymbol{\Omega}$ is not proportional to an identity matrix. With $\boldsymbol{J}$ defined by (9.08), the covariance matrix (9.07) becomes

$$\operatorname*{plim}_{n\to\infty}\Big(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{X}\Big)^{-1}, \tag{9.09}$$

and the **efficient GMM estimator** is

$$\hat{\beta}_{\text{GMM}} = \big(\boldsymbol{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{y}. \tag{9.10}$$

When $\boldsymbol{\Omega}_0 = \sigma^2\mathbf{I}$, this estimator reduces to the generalized IV estimator (8.29). In Exercise 9.1, readers are invited to show that the difference between the covariance matrices (9.07) and (9.09) is a positive semidefinite matrix, thereby confirming (9.08) as the optimal choice for $\boldsymbol{J}$.

The GMM Criterion Function

With both GLS and IV estimation, we showed that the efficient estimators could also be derived by minimizing an appropriate criterion function; this function was (7.06) for GLS and (8.30) for IV. Similarly, the efficient GMM estimator (9.10) minimizes the **GMM criterion function**

$$Q(\boldsymbol{\beta},\boldsymbol{y}) \equiv (\boldsymbol{y}-\boldsymbol{X\beta})^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W})^{-1}\boldsymbol{W}^\top(\boldsymbol{y}-\boldsymbol{X\beta}), \tag{9.11}$$

as can be seen at once by noting that the first-order conditions for minimizing (9.11) are

$$\boldsymbol{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W})^{-1}\boldsymbol{W}^\top(\boldsymbol{y}-\boldsymbol{X\beta}) = \mathbf{0}.$$

If $\boldsymbol{\Omega}_0 = \sigma_0^2\mathbf{I}$, (9.11) reduces to the IV criterion function (8.30), divided by σ_0^2. In Section 8.6, we saw that the minimized value of the IV criterion function, divided by an estimate of σ^2, serves as the statistic for the Sargan test for overidentification. We will see in Section 9.4 that the GMM criterion function (9.11), with the usually unknown matrix $\boldsymbol{\Omega}_0$ replaced by a suitable estimate, can also be used as a test statistic for overidentification.

The criterion function (9.11) is a quadratic form in the vector $\boldsymbol{W}^\top(\boldsymbol{y}-\boldsymbol{X\beta})$ of sample moments and the inverse of the matrix $\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W}$. Equivalently, it is a quadratic form in $n^{-1/2}\boldsymbol{W}^\top(\boldsymbol{y}-\boldsymbol{X\beta})$ and the inverse of $n^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W}$, since

Copyright © 2003, Russell Davidson and James G. MacKinnon

the powers of n cancel. Under the sort of regularity conditions we have used in earlier chapters, $n^{-1/2}\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{X\beta}_0)$ satisfies a central limit theorem, and so tends, as $n \to \infty$, to a normal random variable, with mean vector $\mathbf{0}$ and covariance matrix the limit of $n^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W}$. It follows that (9.11) evaluated using the true $\boldsymbol{\beta}_0$ and the true $\boldsymbol{\Omega}_0$ is asymptotically distributed as χ^2 with l degrees of freedom; recall Theorem 4.1, and see Exercise 9.2.

This property of the GMM criterion function is simply a consequence of its structure as a quadratic form in the sample moments used for estimation and the inverse of the asymptotic covariance matrix of these moments evaluated at the true parameters. As we will see in Section 9.4, this property is what makes the GMM criterion function useful for testing. The argument leading to (9.10) shows that this same property of the GMM criterion function leads to the asymptotic efficiency of the estimator that minimizes it.

Provided the instruments are predetermined, so that they satisfy the condition that $\mathrm{E}(u_t \,|\, \boldsymbol{W}_t) = 0$, we still obtain a consistent estimator, even when the matrix $\boldsymbol{J}$ used to select linear combinations of the instruments is different from (9.08). Such a consistent, but in general inefficient, estimator can also be obtained by minimizing a quadratic criterion function of the form

$$(\boldsymbol{y} - \boldsymbol{X\beta})^\top\boldsymbol{W\Lambda W}^\top(\boldsymbol{y} - \boldsymbol{X\beta}), \tag{9.12}$$

where the **weighting matrix** $\boldsymbol{\Lambda}$ is $l \times l$, positive definite, and must be at least asymptotically nonrandom. Without loss of generality, $\boldsymbol{\Lambda}$ can be taken to be symmetric; see Exercise 9.3. The inefficient GMM estimator is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{W\Lambda W}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{W\Lambda W}^\top\boldsymbol{y}, \tag{9.13}$$

from which it can be seen that the use of the weighting matrix $\boldsymbol{\Lambda}$ corresponds to the implicit choice $\boldsymbol{J} = \boldsymbol{\Lambda W}^\top\boldsymbol{X}$. For a given choice of $\boldsymbol{J}$, there are various possible choices of $\boldsymbol{\Lambda}$ that give rise to the same estimator; see Exercise 9.4.

When $l = k$, the model is exactly identified, and $\boldsymbol{J}$ is a nonsingular square matrix which has no effect on the estimator. This is most easily seen by looking at the moment conditions (9.05), which are equivalent, when $l = k$, to those obtained by premultiplying them by $(\boldsymbol{J}^\top)^{-1}$. Similarly, if the estimator is defined by minimizing a quadratic form, it does not depend on the choice of $\boldsymbol{\Lambda}$ whenever $l = k$. To see this, consider the first-order conditions for minimizing (9.12), which, up to a scalar factor, are

$$\boldsymbol{X}^\top\boldsymbol{W\Lambda W}^\top(\boldsymbol{y} - \boldsymbol{X\beta}) = \mathbf{0}.$$

If $l = k$, $\boldsymbol{X}^\top\boldsymbol{W}$ is a square matrix, and the first-order conditions can be premultiplied by $\boldsymbol{\Lambda}^{-1}(\boldsymbol{X}^\top\boldsymbol{W})^{-1}$. Therefore, the estimator is the solution to the equations $\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{X\beta}) = \mathbf{0}$, independently of $\boldsymbol{\Lambda}$. This solution is just the simple IV estimator defined in (8.12).

Copyright © 2003, Russell Davidson and James G. MacKinnon

When $l > k$, the model is overidentified, and the estimator (9.13) depends on the choice of $\boldsymbol{J}$ or $\boldsymbol{\Lambda}$. The efficient GMM estimator, for a given set of instruments, is defined in terms of the true covariance matrix $\boldsymbol{\Omega}_0$, which is usually unknown. If $\boldsymbol{\Omega}_0$ is known up to a scalar multiplicative factor, so that $\boldsymbol{\Omega}_0 = \sigma^2\boldsymbol{\Delta}_0$, with σ^2 unknown and $\boldsymbol{\Delta}_0$ known, then $\boldsymbol{\Delta}_0$ can be used in place of $\boldsymbol{\Omega}_0$ in either (9.10) or (9.11). This is true because multiplying $\boldsymbol{\Omega}_0$ by a scalar leaves (9.10) invariant, and it also leaves invariant the $\boldsymbol{\beta}$ that minimizes (9.11).

GMM Estimation with Heteroskedasticity of Unknown Form

The assumption that $\boldsymbol{\Omega}_0$ is known, even up to a scalar factor, is often too strong. What makes GMM estimation practical more generally is that, in both (9.10) and (9.11), $\boldsymbol{\Omega}_0$ appears only through the $l \times l$ matrix product $\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W}$. As we saw first in Section 5.5, in the context of heteroskedasticity consistent covariance matrix estimation, n^{-1} times such a matrix can be estimated consistently if $\boldsymbol{\Omega}_0$ is a diagonal matrix. What is needed is a preliminary consistent estimate of the parameter vector $\boldsymbol{\beta}$, which furnishes residuals that are consistent estimates of the error terms.

The preliminary estimates of $\boldsymbol{\beta}$ must be consistent, but they need not be asymptotically efficient, and so we can obtain them by using any convenient choice of $\boldsymbol{J}$ or $\boldsymbol{\Lambda}$. One choice that is often convenient is $\boldsymbol{\Lambda} = (\boldsymbol{W}^\top\boldsymbol{W})^{-1}$, in which case the preliminary estimator is the generalized IV estimator (8.29). We then use the preliminary estimates $\hat{\boldsymbol{\beta}}$ to calculate the residuals $\hat{u}_t \equiv y_t - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. A typical element of the matrix $n^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W}$ can then be estimated by

$$\frac{1}{n}\sum_{t=1}^{n}\hat{u}_t^2 w_{ti} w_{tj}. \tag{9.14}$$

This estimator is very similar to (5.36), and the estimator (9.14) can be proved to be consistent by using arguments just like those employed in Section 5.5.

The matrix with typical element (9.14) can be written as $n^{-1}\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W}$, where $\hat{\boldsymbol{\Omega}}$ is an $n \times n$ diagonal matrix with typical diagonal element $\hat{u}_t^2$. Then the **feasible efficient GMM estimator** is

$$\hat{\beta}_{\rm FGMM} = \big(\boldsymbol{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{y}, \tag{9.15}$$

which is just (9.10) with $\boldsymbol{\Omega}_0$ replaced by $\hat{\boldsymbol{\Omega}}$. Since $n^{-1}\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W}$ consistently estimates $n^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W}$, it follows that $\hat{\boldsymbol{\beta}}_{\rm FGMM}$ is asymptotically equivalent to (9.10). It should be noted that, in calling (9.15) efficient, we mean that it is asymptotically efficient within the class of estimators that use the given instrument set $\boldsymbol{W}$.

Like other procedures that start from a preliminary estimate, this one can be iterated. The GMM residuals $y_t - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\rm FGMM}$ can be used to calculate a new estimate of $\boldsymbol{\Omega}$, which can then be used to obtain second-round GMM estimates, which can then be used to calculate yet another estimate of $\boldsymbol{\Omega}$, and so

Copyright © 2003, Russell Davidson and James G. MacKinnon

on. This iterative procedure was investigated by Hansen, Heaton, and Yaron (1996), who called it **continuously updated GMM**. Whether we stop after one round or continue until the procedure converges, the estimates have the same asymptotic distribution if the model is correctly specified. However, there is evidence that performing more iterations improves finite-sample performance. In practice, the covariance matrix is estimated by

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}_{\rm FGMM}) = \big(\boldsymbol{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{X}\big)^{-1}. \tag{9.16}$$

It is not hard to see that n times the estimator (9.16) tends to the asymptotic covariance matrix (9.09) as $n \to \infty$.

Fully Efficient GMM Estimation

In choosing to use a particular matrix of instrumental variables $\boldsymbol{W}$, we are choosing a particular representation of the information sets Ω_t appropriate for each observation in the sample. It is required that $\boldsymbol{W}_t \in \Omega_t$ for all t, and it follows from this that any deterministic function, linear or nonlinear, of the elements of $\boldsymbol{W}_t$ also belongs to Ω_t. It is quite clearly impossible to use all such deterministic functions as actual instrumental variables, and so the econometrician must make a choice. What we have established so far is that, once the choice of $\boldsymbol{W}$ is made, (9.08) gives the optimal set of linear combinations of the columns of $\boldsymbol{W}$ to use for estimation. What remains to be seen is how best to choose $\boldsymbol{W}$ out of all the possible valid instruments, given the information sets Ω_t.

In Section 8.3, we saw that, for the model (9.01) with $\boldsymbol{\Omega} = \sigma^2\mathbf{I}$, the best choice, by the criterion of the asymptotic covariance matrix, is the matrix $\bar{\boldsymbol{X}}$ given in (8.18) by the defining condition that $\mathrm{E}(\boldsymbol{X}_t \,|\, \Omega_t) = \bar{\boldsymbol{X}}_t$, where $\boldsymbol{X}_t$ and $\bar{\boldsymbol{X}}_t$ are the $t^{\rm th}$ rows of $\boldsymbol{X}$ and $\bar{\boldsymbol{X}}$, respectively. However, it is easy to see that this result does not hold unmodified when $\boldsymbol{\Omega}$ is not proportional to an identity matrix. Consider the GMM estimator (9.10), of which (9.15) is the feasible version, in the special case of exogenous explanatory variables, for which the obvious choice of instruments is $\boldsymbol{W} = \boldsymbol{X}$. If, for notational ease, we write $\boldsymbol{\Omega}$ for the true covariance matrix $\boldsymbol{\Omega}_0$, (9.10) becomes

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\rm GMM} &= \big(\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}\\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}\\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}\\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \hat{\boldsymbol{\beta}}_{\rm OLS}.
\end{aligned}$$

However, we know from the results of Section 7.2 that the efficient estimator is actually the GLS estimator

$$\hat{\boldsymbol{\beta}}_{\rm GLS} = (\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{y}, \tag{9.17}$$

which, except in special cases, is different from $\hat{\boldsymbol{\beta}}_{\rm OLS}$.

Copyright © 2003, Russell Davidson and James G. MacKinnon

The GLS estimator (9.17) can be interpreted as an IV estimator, in which the instruments are the columns of $\boldsymbol{\Omega}^{-1}\boldsymbol{X}$. Thus it appears that, when $\boldsymbol{\Omega}$ is not a multiple of the identity matrix, the optimal instruments are no longer the explanatory variables $\boldsymbol{X}$, but rather the columns of $\boldsymbol{\Omega}^{-1}\boldsymbol{X}$. This suggests that, when at least some of the explanatory variables in the matrix $\boldsymbol{X}$ are not predetermined, the optimal choice of instruments is given by $\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}$. This choice combines the result of Chapter 7 about the optimality of the GLS estimator with that of Chapter 8 about the best instruments to use in place of explanatory variables that are not predetermined. It leads to the theoretical moment conditions

$$\mathrm{E}\big(\bar{\boldsymbol{X}}^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})\big)=\mathbf{0}. \tag{9.18}$$

Unfortunately, this solution to the optimal instruments problem does not always work, because the moment conditions in (9.18) may not be correct. To see why not, suppose that the error terms are serially correlated, and that $\boldsymbol{\Omega}$ is consequently not a diagonal matrix. The i^{th} element of the matrix product in (9.18) can be expanded as

$$\sum_{t=1}^{n}\sum_{s=1}^{n}\bar{\boldsymbol{X}}_{ti}\,\omega^{ts}(y_s-\boldsymbol{X}_s\boldsymbol{\beta}), \tag{9.19}$$

where ω^{ts} is the ts^{th} element of $\boldsymbol{\Omega}^{-1}$. If we evaluate at the true parameter vector $\boldsymbol{\beta}_0$, we find that $y_s-\boldsymbol{X}_s\boldsymbol{\beta}_0=u_s$. But, unless the columns of the matrix $\bar{\boldsymbol{X}}$ are exogenous, it is not in general the case that $\mathrm{E}(u_s\,|\,\bar{\boldsymbol{X}}_t)=0$ for $s\neq t$, and, if this condition is not satisfied, the expectation of (9.19) is not zero in general. This issue was discussed at the end of Section 7.3, and in more detail in Section 7.8, in connection with the use of GLS when one of the explanatory variables is a lagged dependent variable.

Choosing Valid Instruments

As in Section 7.2, we can construct an $n\times n$ matrix $\boldsymbol{\Psi}$, usually triangular, that satisfies the equation $\boldsymbol{\Omega}^{-1}=\boldsymbol{\Psi\Psi}^\top$. As in equation (7.03) of Section 7.2, we can premultiply regression (9.01) by $\boldsymbol{\Psi}^\top$ to get

$$\boldsymbol{\Psi}^\top\boldsymbol{y}=\boldsymbol{\Psi}^\top\boldsymbol{X\beta}+\boldsymbol{\Psi}^\top\boldsymbol{u}, \tag{9.20}$$

with the result that the covariance matrix of the transformed error vector, $\boldsymbol{\Psi}^\top\boldsymbol{u}$, is just the identity matrix. Suppose that we propose to use a matrix $\boldsymbol{Z}$ of instruments in order to estimate the transformed model, so that we are led to consider the theoretical moment conditions

$$\mathrm{E}\big(\boldsymbol{Z}^\top\boldsymbol{\Psi}^\top(\boldsymbol{y}-\boldsymbol{X\beta})\big)=\mathbf{0}. \tag{9.21}$$

If these conditions are to be correct, then what we need is that, for each t, $\mathrm{E}\big((\boldsymbol{\Psi}^\top\boldsymbol{u})_t\,|\,\boldsymbol{Z}_t\big)=0$, where the subscript t is used to select the t^{th} row of the corresponding vector or matrix.

If $\boldsymbol{X}$ is exogenous, the optimal instruments are given by the matrix $\boldsymbol{\Omega}^{-1}\boldsymbol{X}$, and the moment conditions for efficient estimation are $\mathrm{E}\big(\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})\big)=\mathbf{0}$, which can also be written as

$$\mathrm{E}\big(\boldsymbol{X}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top(\boldsymbol{y}-\boldsymbol{X\beta})\big)=\mathbf{0}. \tag{9.22}$$

Comparison with (9.21) shows that the optimal choice of $\boldsymbol{Z}$ is $\boldsymbol{\Psi}^\top\boldsymbol{X}$. Even if $\boldsymbol{X}$ is not exogenous, (9.22) is a correct set of moment conditions if

$$\mathrm{E}\big((\boldsymbol{\Psi}^\top\boldsymbol{u})_t\,|\,(\boldsymbol{\Psi}^\top\boldsymbol{X})_t\big)=0. \tag{9.23}$$

But this is not true in general when $\boldsymbol{X}$ is not exogenous. Consequently, we seek a new definition for $\bar{\boldsymbol{X}}$, such that (9.23) becomes true when $\boldsymbol{X}$ is replaced by $\bar{\boldsymbol{X}}$.

In most cases, it is possible to choose $\boldsymbol{\Psi}$ so that $(\boldsymbol{\Psi}^\top\boldsymbol{u})_t$ is an innovation in the sense of Section 4.5, that is, so that $\mathrm{E}\big((\boldsymbol{\Psi}^\top\boldsymbol{u})_t\,|\,\Omega_t\big)=0$. As an example, see the analysis of models with AR(1) errors in Section 7.8, especially the discussion surrounding (7.58). What is then required for condition (9.23) is that $(\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}})_t$ should be predetermined in period t. If $\boldsymbol{\Omega}$ is diagonal, and so also $\boldsymbol{\Psi}$, the old definition of $\bar{\boldsymbol{X}}$ works, because $(\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}})_t=\Psi_{tt}\bar{\boldsymbol{X}}_t$, where Ψ_{tt} is the $t^{\rm th}$ diagonal element of $\boldsymbol{\Psi}$, and this belongs to Ω_t by construction. If $\boldsymbol{\Omega}$ contains off-diagonal elements, however, the old definition of $\bar{\boldsymbol{X}}$ no longer works in general. Since what we need is that $(\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}})_t$ should belong to Ω_t, we instead define $\bar{\boldsymbol{X}}$ implicitly by the equation

$$\mathrm{E}\big((\boldsymbol{\Psi}^\top\boldsymbol{X})_t\,|\,\Omega_t\big)=(\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}})_t. \tag{9.24}$$

This implicit definition must be implemented on a case-by-case basis. One example is given in Exercise 9.5.

By setting $\boldsymbol{Z}=\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}}$, we find that the moment conditions (9.21) become

$$\mathrm{E}\big(\bar{\boldsymbol{X}}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top(\boldsymbol{y}-\boldsymbol{X\beta})\big)=\mathrm{E}\big(\bar{\boldsymbol{X}}^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})\big)=\mathbf{0}. \tag{9.25}$$

These conditions do indeed use $\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}$ as instruments, albeit with a possibly redefined $\bar{\boldsymbol{X}}$. The estimator based on (9.25) is

$$\hat{\boldsymbol{\beta}}_{\rm EGMM}\equiv(\bar{\boldsymbol{X}}^\top\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{y}, \tag{9.26}$$

where EGMM denotes “efficient GMM.” The asymptotic covariance matrix of (9.26) can be computed using (9.09), in which, on the basis of (9.25), we see that $\boldsymbol{W}$ is to be replaced by $\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}}$, $\boldsymbol{X}$ by $\boldsymbol{\Psi}^\top\boldsymbol{X}$, and $\boldsymbol{\Omega}$ by $\mathbf{I}$. We cannot apply (9.09) directly with instruments $\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}$, because there is no reason to suppose that the result (9.02) holds for the untransformed error terms $\boldsymbol{u}$ and the instruments $\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}$. The result is

$$\operatorname*{plim}_{n\to\infty}\Big(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}\Big(\frac{1}{n}\bar{\boldsymbol{X}}^\top\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}\Big)^{-1}\frac{1}{n}\bar{\boldsymbol{X}}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X}\Big)^{-1}. \tag{9.27}$$

Copyright © 2003, Russell Davidson and James G. MacKinnon

By exactly the same argument as that used in (8.20), we find that, for any matrix $\boldsymbol{Z}$ that satisfies $\boldsymbol{Z}_t \in \Omega_t$,

$$\underset{n\to\infty}{\text{plim}}\,\frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{\Psi}^\top\boldsymbol{X} = \underset{n\to\infty}{\text{plim}}\,\frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}}. \tag{9.28}$$

Since $(\boldsymbol{\Psi}^\top\boldsymbol{X})_t \in \Omega_t$, this implies that

$$\begin{aligned}\underset{n\to\infty}{\text{plim}}\,\frac{1}{n}\bar{\boldsymbol{X}}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X} &= \underset{n\to\infty}{\text{plim}}\,\frac{1}{n}\bar{\boldsymbol{X}}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\boldsymbol{X}\\ &= \underset{n\to\infty}{\text{plim}}\,\frac{1}{n}\bar{\boldsymbol{X}}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}} = \underset{n\to\infty}{\text{plim}}\,\frac{1}{n}\bar{\boldsymbol{X}}^\top\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}.\end{aligned}$$

Therefore, the asymptotic covariance matrix (9.27) simplifies to

$$\underset{n\to\infty}{\text{plim}}\left(\frac{1}{n}\bar{\boldsymbol{X}}^\top\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}\right)^{-1}. \tag{9.29}$$

Although the matrix (9.09) is less of a sandwich than (9.07), the matrix (9.29) is still less of one than (9.09). This is a clear indication of the fact that the instruments $\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}$, which yield the estimator $\hat{\boldsymbol{\beta}}_{\text{EGMM}}$, are indeed optimal. Readers are asked to check this formally in Exercise 9.7.

In most cases, $\bar{\boldsymbol{X}}$ is not observed, but it can often be estimated consistently. The usual state of affairs is that we have an $n \times l$ matrix $\boldsymbol{W}$ of instruments, such that $\mathcal{S}(\bar{\boldsymbol{X}}) \subseteq \mathcal{S}(\boldsymbol{W})$ and

$$(\boldsymbol{\Psi}^\top\boldsymbol{W})_t \in \Omega_t. \tag{9.30}$$

This last condition is the form taken by the **predeterminedness condition** when $\boldsymbol{\Omega}$ is not proportional to the identity matrix. The theoretical moment conditions used for (overidentified) estimation are then

$$\text{E}\big(\boldsymbol{W}^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\big) = \text{E}\big(\boldsymbol{W}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\big) = \boldsymbol{0}, \tag{9.31}$$

from which it can be seen that what we are in fact doing is estimating the transformed model (9.20) using the transformed instruments $\boldsymbol{\Psi}^\top\boldsymbol{W}$. The result of Exercise 9.8 shows that, if indeed $\mathcal{S}(\bar{\boldsymbol{X}}) \subseteq \mathcal{S}(\boldsymbol{W})$, the asymptotic covariance matrix of the resulting estimator is still (9.29). Exercise 9.9 investigates what happens if this condition is not satisfied.

The main obstacle to the use of the efficient estimator $\hat{\boldsymbol{\beta}}_{\text{EGMM}}$ is thus not the difficulty of estimating $\bar{\boldsymbol{X}}$, but rather the fact that $\boldsymbol{\Omega}$ is usually not known. As with the GLS estimators we studied in Chapter 7, $\hat{\boldsymbol{\beta}}_{\text{EGMM}}$ cannot be calculated unless we either know $\boldsymbol{\Omega}$ or can estimate it consistently, usually by knowing the form of $\boldsymbol{\Omega}$ as a function of parameters that can be estimated consistently. But whenever there is heteroskedasticity or serial correlation of unknown form, this is impossible. The best we can then do, asymptotically, is to use the feasible efficient GMM estimator (9.15). Therefore, when we later refer to GMM estimators without further qualification, we will normally mean feasible efficient ones.

Copyright © 2003, Russell Davidson and James G. MacKinnon

9.3 HAC Covariance Matrix Estimation

Up to this point, we have seen how to obtain feasible efficient GMM estimates only when the matrix $\boldsymbol{\Omega}$ is known to be diagonal, in which case we can use the estimator (9.15). In this section, we also allow for the possibility of serial correlation of unknown form, which causes $\boldsymbol{\Omega}$ to have nonzero off-diagonal elements. When the pattern of the serial correlation is unknown, we can still, under fairly weak regularity conditions, estimate the covariance matrix of the sample moments by using a **heteroskedasticity and autocorrelation consistent**, or **HAC**, estimator of the matrix $n^{-1}\boldsymbol{W}^{\top}\boldsymbol{\Omega}\boldsymbol{W}$. This estimator, multiplied by n, can then be used in place of $\boldsymbol{W}^{\top}\hat{\boldsymbol{\Omega}}\boldsymbol{W}$ in the feasible efficient GMM estimator (9.15).

The asymptotic covariance matrix of the vector $n^{-1/2}\boldsymbol{W}^{\top}(\boldsymbol{y}-\boldsymbol{X\beta})$ of sample moments, evaluated at $\boldsymbol{\beta}=\boldsymbol{\beta}_0$, is defined as follows:

$$\boldsymbol{\Sigma} \equiv \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{W}^{\top}(\boldsymbol{y}-\boldsymbol{X\beta}_0)(\boldsymbol{y}-\boldsymbol{X\beta}_0)^{\top}\boldsymbol{W} = \operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{W}^{\top}\boldsymbol{\Omega}\boldsymbol{W}. \tag{9.32}$$

A HAC estimator of $\boldsymbol{\Sigma}$ is a matrix $\hat{\boldsymbol{\Sigma}}$ constructed so that $\hat{\boldsymbol{\Sigma}}$ consistently estimates $\boldsymbol{\Sigma}$ when the error terms u_t display any pattern of heteroskedasticity and/or autocorrelation that satisfies certain, generally quite weak, conditions. In order to derive such an estimator, we begin by rewriting the definition of $\boldsymbol{\Sigma}$ in an alternative way:

$$\boldsymbol{\Sigma} = \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathrm{E}\big(u_t u_s \boldsymbol{W}_t{}^{\top}\boldsymbol{W}_s\big), \tag{9.33}$$

in which we assume that a law of large numbers can be used to justify replacing the probability limit in (9.32) by the expectations in (9.33).

For regression models with heteroskedasticity but no autocorrelation, only the terms with $t=s$ contribute to (9.33). Therefore, for such models, we can estimate $\boldsymbol{\Sigma}$ consistently by simply ignoring the expectation operator and replacing the error terms u_t by least-squares residuals $\hat{u}_t$, possibly with a modification designed to offset the tendency for such residuals to be too small. The obvious way to estimate (9.33) when there may be serial correlation is again simply to drop the expectations operator and replace $u_t u_s$ by $\hat{u}_t\hat{u}_s$, where $\hat{u}_t$ denotes the t^{th} residual from some consistent but inefficient estimation procedure, such as generalized IV. Unfortunately, this approach does not work. To see why not, we need to rewrite (9.33) in yet another way. Let us define the **autocovariance matrices** of the $\boldsymbol{W}_t{}^{\top}u_t$ as follows:

$$\boldsymbol{\Gamma}(j) \equiv \begin{cases} \dfrac{1}{n}\displaystyle\sum_{t=j+1}^{n}\mathrm{E}(u_t u_{t-j}\boldsymbol{W}_t{}^{\top}\boldsymbol{W}_{t-j}) & \text{for } j\ge 0,\\[2ex] \dfrac{1}{n}\displaystyle\sum_{t=-j+1}^{n}\mathrm{E}(u_{t+j}u_t\boldsymbol{W}_{t+j}^{\top}\boldsymbol{W}_t) & \text{for } j<0. \end{cases} \tag{9.34}$$

Copyright © 2003, Russell Davidson and James G. MacKinnon

Because there are l moment conditions, these are $l \times l$ matrices. It is easy to check that $\boldsymbol{\Gamma}(j) = \boldsymbol{\Gamma}^\top(-j)$. Then, in terms of the matrices $\boldsymbol{\Gamma}(j)$, expression (9.33) becomes

$$\boldsymbol{\Sigma} = \lim_{n\to\infty} \sum_{j=-n+1}^{n-1} \boldsymbol{\Gamma}(j) = \lim_{n\to\infty} \Big(\boldsymbol{\Gamma}(0) + \sum_{j=1}^{n-1} \big(\boldsymbol{\Gamma}(j) + \boldsymbol{\Gamma}^\top(j)\big)\Big). \tag{9.35}$$

Therefore, in order to estimate $\boldsymbol{\Sigma}$, we apparently need to estimate all of the autocovariance matrices for $j = 0, \ldots, n-1$.

If $\hat{u}_t$ denotes a typical residual from some preliminary estimator, the **sample autocovariance matrix** of order j, $\hat{\boldsymbol{\Gamma}}(j)$, is just the appropriate expression in (9.34), without the expectation operator, and with the random variables u_t and u_{t-j} replaced by $\hat{u}_t$ and $\hat{u}_{t-j}$, respectively. For any $j \geq 0$, this is

$$\hat{\boldsymbol{\Gamma}}(j) = \frac{1}{n} \sum_{t=j+1}^{n} \hat{u}_t \hat{u}_{t-j} \boldsymbol{W}_t{}^\top \boldsymbol{W}_{t-j}. \tag{9.36}$$

Unfortunately, the sample autocovariance matrix $\hat{\boldsymbol{\Gamma}}(j)$ of order j is not a consistent estimator of the true autocovariance matrix for arbitrary j. Suppose, for instance, that $j = n-2$. Then, from (9.36), we see that $\hat{\boldsymbol{\Gamma}}(j)$ has only two terms, and no conceivable law of large numbers can apply to only two terms. In fact, $\hat{\boldsymbol{\Gamma}}(n-2)$ must tend to zero as $n \to \infty$ because of the factor of n^{-1} in its definition.

The solution to this problem is to restrict our attention to models for which the actual autocovariances mimic the behavior of the sample autocovariances, and for which therefore the actual autocovariance of order j tends to zero as $j \to \infty$. A great many stochastic processes generate error terms for which the $\boldsymbol{\Gamma}(j)$ do have this property. In such cases, we can drop most of the sample autocovariance matrices that appear in the sample analog of (9.35) by eliminating ones for which $|j|$ is greater than some chosen threshold, say p. This yields the following estimator for $\boldsymbol{\Sigma}$:

$$\hat{\boldsymbol{\Sigma}}_{\text{HW}} = \hat{\boldsymbol{\Gamma}}(0) + \sum_{j=1}^{p} \big(\hat{\boldsymbol{\Gamma}}(j) + \hat{\boldsymbol{\Gamma}}^\top(j)\big), \tag{9.37}$$

We refer to this estimator as the **Hansen-White estimator**, because it was originally proposed by Hansen (1982) and White and Domowitz (1984); see also White (2000).

For the purposes of asymptotic theory, it is necessary to let the parameter p, which is called the **lag truncation parameter**, go to infinity in (9.37) at some suitable rate as the sample size goes to infinity. A typical rate would be $n^{1/4}$. This ensures that, for large enough n, all the nonzero $\boldsymbol{\Gamma}(j)$ are estimated consistently. Unfortunately, this type of result does not say how large p should be in practice. In most cases, we have a given, finite, sample size, and we need to choose a specific value of p.

The Hansen-White estimator (9.37) suffers from one very serious deficiency: In finite samples, it need not be positive definite or even positive semidefinite. If one happens to encounter a data set that yields a nondefinite $\hat{\boldsymbol{\Sigma}}_{\rm HW}$, then, since the weighting matrix for GMM must be positive definite, (9.37) is unusable. Luckily, there are numerous ways out of this difficulty. The one that is most widely used was suggested by Newey and West (1987). The **Newey-West estimator** they propose is

$$\hat{\boldsymbol{\Sigma}}_{\rm NW} = \hat{\boldsymbol{\Gamma}}(0) + \sum_{j=1}^{p}\Big(1 - \frac{j}{p+1}\Big)\Big(\hat{\boldsymbol{\Gamma}}(j) + \hat{\boldsymbol{\Gamma}}^{\top}(j)\Big), \tag{9.38}$$

in which each sample autocovariance matrix $\hat{\boldsymbol{\Gamma}}(j)$ is multiplied by a weight $1 - j/(p+1)$ that decreases linearly as j increases. The weight is $p/(p+1)$ for $j = 1$, and it then decreases by steps of $1/(p+1)$ down to a value of $1/(p+1)$ for $j = p$. This estimator evidently tends to underestimate the autocovariance matrices, especially for larger values of j. Therefore, p should almost certainly be larger for (9.38) than for (9.37). As with the Hansen-White estimator, p must increase as n does, and the appropriate rate is $n^{1/3}$. A procedure for selecting p automatically was proposed by Newey and West (1994), but it is too complicated to discuss here.

Both the Hansen-White and the Newey-West HAC estimators of $\boldsymbol{\Sigma}$ can be written in the form

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{W}^{\top}\hat{\boldsymbol{\Omega}}\boldsymbol{W} \tag{9.39}$$

for an appropriate choice of $\hat{\boldsymbol{\Omega}}$. This fact, which we will exploit in the next section, follows from the observation that there exist $n \times n$ matrices $\boldsymbol{U}(j)$ such that the $\hat{\boldsymbol{\Gamma}}(j)$ can be expressed in the form $n^{-1}\boldsymbol{W}^{\top}\boldsymbol{U}(j)\boldsymbol{W}$, as readers are asked to check in Exercise 9.10.

The Newey-West estimator is by no means the only HAC estimator that is guaranteed to be positive definite. Andrews (1991) provides a detailed treatment of HAC estimation, suggests some alternatives to the Newey-West estimator, and shows that, in some circumstances, they may perform better than it does in finite samples. A different approach to HAC estimation is suggested by Andrews and Monahan (1992). Since this material is relatively advanced and specialized, we will not pursue it further here. Interested readers may wish to consult Hamilton (1994, Chapter 10) as well as the references already given.

Feasible Efficient GMM Estimation

In practice, efficient GMM estimation in the presence of heteroskedasticity and serial correlation of unknown form works as follows. As in the case with only

Copyright © 2003, Russell Davidson and James G. MacKinnon

heteroskedasticity that was discussed in Section 9.2, we first obtain consistent but inefficient estimates, probably by using generalized IV. These estimates yield residuals $\hat{u}_t$, from which we next calculate a matrix $\hat{\boldsymbol{\Sigma}}$ that estimates $\boldsymbol{\Sigma}$ consistently, using (9.37), (9.38), or some other HAC estimator. The feasible efficient GMM estimator, which generalizes (9.15), is then

$$\hat{\boldsymbol{\beta}}_{\rm FGMM} = (\boldsymbol{X}^\top\boldsymbol{W}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{W}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}^\top\boldsymbol{y}. \qquad (9.40)$$

As before, this procedure may be iterated. The first-round GMM residuals may be used to obtain a new estimate of $\boldsymbol{\Sigma}$, which may be used to obtain second-round GMM estimates, and so on. For a correctly specified model, iteration should not affect the asymptotic properties of the estimates.

We can estimate the covariance matrix of (9.40) by

$$\widehat{\rm Var}(\hat{\boldsymbol{\beta}}_{\rm FGMM}) = n(\boldsymbol{X}^\top\boldsymbol{W}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}^\top\boldsymbol{X})^{-1}, \qquad (9.41)$$

which is the analog of (9.16). The factor of n here is needed to offset the factor of n^{-1} in the definition of $\hat{\boldsymbol{\Sigma}}$. We do not need to include such a factor in (9.40), because the two factors of n^{-1} cancel out. As usual, the covariance matrix estimator (9.41) can be used to construct pseudo-t tests and other Wald tests, and asymptotic confidence intervals and confidence regions may also be based on it. The GMM criterion function that corresponds to (9.40) is

$$\frac{1}{n}(\boldsymbol{y}-\boldsymbol{X\beta})^\top\boldsymbol{W}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}^\top(\boldsymbol{y}-\boldsymbol{X\beta}). \qquad (9.42)$$

Once again, we need a factor of n^{-1} here to offset the one in $\hat{\boldsymbol{\Sigma}}$.

The feasible efficient GMM estimator (9.40) can be used even when all the columns of $\boldsymbol{X}$ are valid instruments and OLS would be the estimator of choice if the error terms were not heteroskedastic and/or serially correlated. In this case, $\boldsymbol{W}$ typically consists of $\boldsymbol{X}$ augmented by a number of functions of the columns of $\boldsymbol{X}$, such as squares and cross-products, and $\hat{\boldsymbol{\Omega}}$ has squared OLS residuals on the diagonal. This estimator, which was proposed by Cragg (1983) for models with heteroskedastic error terms, is asymptotically more efficient than OLS whenever $\boldsymbol{\Omega}$ is not proportional to an identity matrix.

9.4 Tests Based on the GMM Criterion Function

For models estimated by instrumental variables, we saw in Section 8.5 that any set of r equality restrictions can be tested by taking the difference between the minimized values of the IV criterion function for the restricted and unrestricted models, and then dividing it by a consistent estimate of the error variance. The resulting test statistic is asymptotically distributed as $\chi^2(r)$. For models estimated by (feasible) efficient GMM, a very similar testing procedure

Copyright © 2003, Russell Davidson and James G. MacKinnon

is available. In this case, as we will see, the difference between the constrained and unconstrained minima of the GMM criterion function is asymptotically distributed as $\chi^2(r)$. There is no need to divide by an estimate of σ^2, because the GMM criterion function already takes account of the covariance matrix of the error terms.

Tests of Overidentifying Restrictions

Whenever $l > k$, a model estimated by GMM involves $l - k$ overidentifying restrictions. As in the IV case, tests of these restrictions are even easier to perform than tests of other restrictions, because the minimized value of the optimal GMM criterion function (9.11), with $n^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W}$ replaced by a HAC estimate, provides an asymptotically valid test statistic. When the HAC estimate $\hat{\boldsymbol{\Sigma}}$ is expressed as in (9.39), the GMM criterion function (9.42) can be written as

$$Q(\boldsymbol{\beta}, \boldsymbol{y}) \equiv (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{W}(\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W})^{-1}\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{9.43}$$

Since HAC estimators are consistent, the asymptotic distribution of (9.43), for given $\boldsymbol{\beta}$, is the same whether we use the unknown true $\boldsymbol{\Omega}_0$ or a matrix $\hat{\boldsymbol{\Omega}}$ that provides a HAC estimate. For simplicity, we therefore use the true $\boldsymbol{\Omega}_0$, omitting the subscript 0 for ease of notation. The asymptotic equivalence of the $\hat{\boldsymbol{\beta}}_{\rm FGMM}$ of (9.15) or (9.40) and the $\hat{\boldsymbol{\beta}}_{\rm GMM}$ of (9.10) further implies that what we will prove for the criterion function (9.43) evaluated at $\hat{\boldsymbol{\beta}}_{\rm GMM}$, with $\hat{\boldsymbol{\Omega}}$ replaced by $\boldsymbol{\Omega}$, is equally true for (9.43) evaluated at $\hat{\boldsymbol{\beta}}_{\rm FGMM}$.

We remarked in Section 9.2 that $Q(\boldsymbol{\beta}_0, \boldsymbol{y})$, where $\boldsymbol{\beta}_0$ is the true parameter vector, is asymptotically distributed as $\chi^2(l)$. In contrast, the minimized criterion function $Q(\hat{\boldsymbol{\beta}}_{\rm GMM}, \boldsymbol{y})$ is distributed as $\chi^2(l - k)$, because we lose k degrees of freedom as a consequence of having estimated k parameters. In order to demonstrate this result, we first express (9.43) in terms of an orthogonal projection matrix. This allows us to reuse many of the calculations performed in Chapter 8.

As in Section 9.2, we make use of a possibly triangular matrix $\boldsymbol{\Psi}$ that satisfies the equation $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top$, or, equivalently,

$$\boldsymbol{\Omega} = (\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}^{-1}. \tag{9.44}$$

If the $n \times l$ matrix $\boldsymbol{A}$ is defined as $\boldsymbol{\Psi}^{-1}\boldsymbol{W}$, and $\boldsymbol{P}_{\!\boldsymbol{A}} \equiv \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top$, then

$$\begin{aligned} Q(\boldsymbol{\beta}, \boldsymbol{y}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^{-1}\boldsymbol{W}\big(\boldsymbol{W}^\top(\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}^{-1}\boldsymbol{W}\big)^{-1}\boldsymbol{W}^\top(\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\Psi}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \end{aligned} \tag{9.45}$$

Since $\hat{\boldsymbol{\beta}}_{\rm GMM}$ minimizes (9.45), we see that one way to write it is

$$\hat{\boldsymbol{\beta}}_{\rm GMM} = (\boldsymbol{X}^\top\boldsymbol{\Psi}\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{y}; \tag{9.46}$$

compare (9.10). Expression (9.46) makes it clear that $\hat{\boldsymbol{\beta}}_{\rm GMM}$ can be thought of as a GIV estimator for the regression of $\boldsymbol{\Psi}^\top\boldsymbol{y}$ on $\boldsymbol{\Psi}^\top\boldsymbol{X}$ using instruments $\boldsymbol{A} \equiv \boldsymbol{\Psi}^{-1}\boldsymbol{W}$. As in (8.61), it can be shown that

$$\boldsymbol{P_A\Psi}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\rm GMM}) = \boldsymbol{P_A}(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{P_A\Psi}^\top\boldsymbol{X}})\boldsymbol{\Psi}^\top\boldsymbol{y},$$

where $\boldsymbol{P}_{\boldsymbol{P_A\Psi}^\top\boldsymbol{X}}$ is the orthogonal projection on to the subspace $\mathcal{S}(\boldsymbol{P_A\Psi}^\top\boldsymbol{X})$. It follows that

$$Q(\hat{\boldsymbol{\beta}}_{\rm GMM}, \boldsymbol{y}) = \boldsymbol{y}^\top\boldsymbol{\Psi}(\boldsymbol{P_A} - \boldsymbol{P}_{\boldsymbol{P_A\Psi}^\top\boldsymbol{X}})\boldsymbol{\Psi}^\top\boldsymbol{y}, \tag{9.47}$$

which is the analog for GMM estimation of expression (8.61) for generalized IV estimation.

Now notice that

$$\begin{aligned}(\boldsymbol{P_A} &- \boldsymbol{P}_{\boldsymbol{P_A\Psi}^\top\boldsymbol{X}})\boldsymbol{\Psi}^\top\boldsymbol{X}\\ &= \boldsymbol{P_A\Psi}^\top\boldsymbol{X} - \boldsymbol{P_A\Psi}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi P_A\Psi}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi P_A\Psi}^\top\boldsymbol{X}\\ &= \boldsymbol{P_A\Psi}^\top\boldsymbol{X} - \boldsymbol{P_A\Psi}^\top\boldsymbol{X} = \mathbf{O}.\end{aligned}$$

Since $\boldsymbol{y} = \boldsymbol{X\beta}_0 + \boldsymbol{u}$ if the model we are estimating is correctly specified, this implies that (9.47) is equal to

$$Q(\hat{\boldsymbol{\beta}}_{\rm GMM}, \boldsymbol{y}) = \boldsymbol{u}^\top\boldsymbol{\Psi}(\boldsymbol{P_A} - \boldsymbol{P}_{\boldsymbol{P_A\Psi}^\top\boldsymbol{X}})\boldsymbol{\Psi}^\top\boldsymbol{u}. \tag{9.48}$$

This expression can be compared with the value of the criterion function evaluated at $\boldsymbol{\beta}_0$, which can be obtained directly from (9.45):

$$Q(\boldsymbol{\beta}_0, \boldsymbol{y}) = \boldsymbol{u}^\top\boldsymbol{\Psi P_A\Psi}^\top\boldsymbol{u}. \tag{9.49}$$

The two expressions (9.48) and (9.49) show clearly where the k degrees of freedom are lost when we estimate $\boldsymbol{\beta}$. We know that $\mathrm{E}(\boldsymbol{\Psi}^\top\boldsymbol{u}) = \mathbf{0}$ and that $\mathrm{E}(\boldsymbol{\Psi}^\top\boldsymbol{uu}^\top\boldsymbol{\Psi}) = \boldsymbol{\Psi}^\top\boldsymbol{\Omega\Psi} = \mathbf{I}$, by (9.44). The dimension of the space $\mathcal{S}(\boldsymbol{A})$ is equal to l. Therefore, the extension of Theorem 4.1 treated in Exercise 9.2 allows us to conclude that (9.49) is asymptotically distributed as $\chi^2(l)$. Since $\mathcal{S}(\boldsymbol{P_A\Psi}^\top\boldsymbol{X})$ is a k–dimensional subspace of $\mathcal{S}(\boldsymbol{A})$, it follows (see Exercise 2.18) that $\boldsymbol{P_A} - \boldsymbol{P}_{\boldsymbol{P_A\Psi}^\top\boldsymbol{X}}$ is an orthogonal projection on to a space of dimension $l - k$, from which we see that (9.48) is asymptotically distributed as $\chi^2(l-k)$. Replacing $\boldsymbol{\beta}_0$ by $\hat{\boldsymbol{\beta}}_{\rm GMM}$ in (9.48) thus leads to the loss of the k dimensions of the space $\mathcal{S}(\boldsymbol{P_A\Psi}^\top\boldsymbol{X})$, which are “used up” when we obtain $\hat{\boldsymbol{\beta}}_{\rm GMM}$.

The statistic $Q(\hat{\boldsymbol{\beta}}_{\rm GMM}, \boldsymbol{y})$ is the analog, for efficient GMM estimation, of the Sargan test statistic that was discussed in Section 8.6. This statistic was suggested by Hansen (1982) in the famous paper that first proposed GMM estimation under that name. It is often called **Hansen's overidentification statistic** or **Hansen's *J* statistic**. However, we prefer to call it the **Hansen-Sargan**

statistic to stress its close relationship with the Sargan test of overidentifying restrictions in the context of generalized IV estimation.

As in the case of IV estimation, a Hansen-Sargan test may reject the null hypothesis for more than one reason. Perhaps the model is misspecified, either because one or more of the instruments should have been included among the regressors, or for some other reason. Perhaps one or more of the instruments is invalid because it is correlated with the error terms. Or perhaps the finite-sample distribution of the test statistic just happens to differ substantially from its asymptotic distribution. In the case of feasible GMM estimation, especially involving HAC covariance matrices, this last possibility should not be discounted. See, among others, Hansen, Heaton, and Yaron (1996) and West and Wilcox (1996).

Tests of Linear Restrictions

Just as in the case of generalized IV, both linear and nonlinear restrictions on regression models can be tested by using the difference between the constrained and unconstrained minima of the GMM criterion function as a test statistic. Under weak conditions, this test statistic is asymptotically distributed as χ^2 with as many degrees of freedom as there are restrictions to be tested. For simplicity, we restrict our attention to zero restrictions on the linear regression model (9.01). This model can be rewritten as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{\Omega}, \tag{9.50}$$

where $\boldsymbol{\beta}_1$ is a k_1–vector and $\boldsymbol{\beta}_2$ is a k_2–vector, with $k = k_1 + k_2$. We wish to test the restrictions $\boldsymbol{\beta}_2 = \mathbf{0}$.

If we estimate (9.50) by feasible efficient GMM using $\boldsymbol{W}$ as the matrix of instruments, subject to the restriction that $\boldsymbol{\beta}_2 = \mathbf{0}$, we obtain the restricted estimates $\tilde{\boldsymbol{\beta}}_{\text{FGMM}} = [\tilde{\boldsymbol{\beta}}_1 \;\vdots\; \mathbf{0}]$. By the reasoning that leads to (9.48), we see that, if indeed $\boldsymbol{\beta}_2 = \mathbf{0}$, the constrained minimum of the criterion function is

$$\begin{aligned} Q(\tilde{\boldsymbol{\beta}}_{\text{FGMM}}, \boldsymbol{y}) &= (\boldsymbol{y} - \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1)^\top \boldsymbol{W}(\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W})^{-1}\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1) \\ &= \boldsymbol{u}^\top\boldsymbol{\Psi}(\boldsymbol{P_A} - \boldsymbol{P}_{\boldsymbol{P_A}\boldsymbol{\Psi}^\top\boldsymbol{X}_1})\boldsymbol{\Psi}^\top\boldsymbol{u}. \end{aligned} \tag{9.51}$$

If we subtract (9.48) from (9.51), we find that the difference between the constrained and unconstrained minima of the criterion function is

$$Q(\tilde{\boldsymbol{\beta}}_{\text{FGMM}}, \boldsymbol{y}) - Q(\hat{\boldsymbol{\beta}}_{\text{FGMM}}, \boldsymbol{y}) = \boldsymbol{u}^\top\boldsymbol{\Psi}(\boldsymbol{P}_{\boldsymbol{P_A}\boldsymbol{\Psi}^\top\boldsymbol{X}} - \boldsymbol{P}_{\boldsymbol{P_A}\boldsymbol{\Psi}^\top\boldsymbol{X}_1})\boldsymbol{\Psi}^\top\boldsymbol{u}. \tag{9.52}$$

Since $\mathcal{S}(\boldsymbol{P_A}\boldsymbol{\Psi}^\top\boldsymbol{X}_1) \subseteq \mathcal{S}(\boldsymbol{P_A}\boldsymbol{\Psi}^\top\boldsymbol{X})$, we see that $\boldsymbol{P}_{\boldsymbol{P_A}\boldsymbol{\Psi}^\top\boldsymbol{X}} - \boldsymbol{P}_{\boldsymbol{P_A}\boldsymbol{\Psi}^\top\boldsymbol{X}_1}$ is an orthogonal projection matrix of which the image is of dimension $k - k_1 = k_2$. Once again, the result of Exercise 9.2 shows that the test statistic (9.52) is asymptotically distributed as $\chi^2(k_2)$ if the null hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$ is true. This result continues to hold if the restrictions are nonlinear, as we will see in Section 9.5.

Copyright © 2003, Russell Davidson and James G. MacKinnon

The result that the statistic $Q(\tilde{\boldsymbol{\beta}}_{\rm FGMM}, \boldsymbol{y}) - Q(\hat{\boldsymbol{\beta}}_{\rm FGMM}, \boldsymbol{y})$ is asymptotically distributed as $\chi^2(k_2)$ depends on two critical features of the construction of the statistic. The first is that the same matrix of instruments $\boldsymbol{W}$ is used for estimating both the restricted and unrestricted models. This was also required in Section 8.5, when we discussed testing restrictions on linear regression models estimated by generalized IV. The second essential feature is that the same weighting matrix $(\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W})^{-1}$ is used when estimating both models. If, as is usually the case, this matrix has to be estimated, it is important that the *same* estimate be used in both criterion functions. If different instruments or different weighting matrices are used for the two models, (9.52) is no longer in general asymptotically distributed as $\chi^2(k_2)$.

One interesting consequence of the form of (9.52) is that we do not always need to bother estimating the unrestricted model. The test statistic (9.52) must always be less than the constrained minimum $Q(\tilde{\boldsymbol{\beta}}_{\rm FGMM}, \boldsymbol{y})$. Therefore, if $Q(\tilde{\boldsymbol{\beta}}_{\rm FGMM}, \boldsymbol{y})$ is less than the critical value for the $\chi^2(k_2)$ distribution at our chosen significance level, we can be sure that the actual test statistic is even smaller and would not lead us to reject the null.

The result that tests of restrictions may be based on the difference between the constrained and unconstrained minima of the GMM criterion function holds only for efficient GMM estimation. It is not true for nonoptimal criterion functions like (9.12), which do not use an estimate of the inverse of the covariance matrix of the sample moments as a weighting matrix. When the GMM estimates minimize a nonoptimal criterion function, the easiest way to test restrictions is probably to use a Wald test; see Sections 6.7 and 8.5. However, we do not recommend performing inference on the basis of nonoptimal GMM estimation.

9.5 GMM Estimators for Nonlinear Models

The principles underlying GMM estimation of nonlinear models are the same as those we have developed for GMM estimation of linear regression models. For every result that we have discussed in the previous three sections, there is an analogous result for nonlinear models. In order to develop these results, we will take a somewhat more general and abstract approach than we have done up to this point. This approach, which is based on the theory of **estimating functions**, was originally developed by Godambe (1960); see also Godambe and Thompson (1978).

The method of estimating functions employs the concept of an **elementary zero function**. Such a function plays the same role as a residual in the estimation of a regression model. It depends on observed variables, at least one of which must be endogenous, and on a k--vector of parameters, $\boldsymbol{\theta}$. As with a residual, the expectation of an elementary zero function must vanish if it is evaluated at the true value of $\boldsymbol{\theta}$, but not in general otherwise.

We let $f_t(\boldsymbol{\theta}, y_t)$ denote an elementary zero function for observation t. It is called "elementary" because it applies to a single observation. In the linear regression case that we have been studying up to this point, $\boldsymbol{\theta}$ would be replaced by $\boldsymbol{\beta}$ and we would have $f_t(\boldsymbol{\beta}, y_t) \equiv y_t - \boldsymbol{X}_t\boldsymbol{\beta}$. In general, we may well have more than one elementary zero function for each observation.

We consider a model $\mathbb{M}$, which, as usual, is to be thought of as a set of DGPs. To each DGP in $\mathbb{M}$, there corresponds a unique value of $\boldsymbol{\theta}$, which is what we often call the "true" value of $\boldsymbol{\theta}$ for that DGP. It is important to note that the uniqueness goes just one way here: A given parameter vector $\boldsymbol{\theta}$ may correspond to many DGPs, perhaps even to an infinite number of them, but each DGP corresponds to just one parameter vector. In order to express the key property of elementary zero functions, we must introduce a symbol for the DGPs of the model $\mathbb{M}$. It is conventional to use the Greek letter μ for this purpose, but then it is necessary to avoid confusion with the conventional use of μ to denote a population mean. It is usually not difficult to distinguish the two uses of the symbol.

The key property of elementary zero functions can now be written as

$$\mathrm{E}_\mu\bigl(f_t(\boldsymbol{\theta}_\mu, y_t)\bigr) = 0, \tag{9.53}$$

where $\mathrm{E}_\mu(\cdot)$ denotes the expectation under the DGP μ, and $\boldsymbol{\theta}_\mu$ is the (unique) parameter vector associated with μ. It is assumed that property (9.53) holds for all t and for all $\mu \in \mathbb{M}$.

If estimation based on elementary zero functions is to be possible, these functions must satisfy a number of conditions in addition to condition (9.53). Most importantly, we need to ensure that the model is asymptotically identified. We therefore assume that, for some observations, at least,

$$\mathrm{E}_\mu\bigl(f_t(\boldsymbol{\theta}, y_t)\bigr) \neq 0 \quad \text{for all } \boldsymbol{\theta} \neq \boldsymbol{\theta}_\mu. \tag{9.54}$$

This just says that, if we evaluate f_t at a $\boldsymbol{\theta}$ that is different from the $\boldsymbol{\theta}_\mu$ that corresponds to the DGP under which we take expectations, then the expectation of $f_t(\boldsymbol{\theta}, y_t)$ must be nonzero. Condition (9.54) does not have to hold for every observation, but it must hold for a fraction of the observations that does not tend to zero as $n \to \infty$.

In the case of the linear regression model, if we write $\boldsymbol{\beta}_0$ for the true parameter vector, condition (9.54) is satisfied for observation t if, for all $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$,

$$\mathrm{E}(y_t - \boldsymbol{X}_t\boldsymbol{\beta}) = \mathrm{E}\bigl(\boldsymbol{X}_t(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + u_t\bigr) = \mathrm{E}\bigl(\boldsymbol{X}_t(\boldsymbol{\beta}_0 - \boldsymbol{\beta})\bigr) \neq 0. \tag{9.55}$$

It is clear from (9.55) that condition (9.54) must be satisfied whenever the fitted values actually depend on all the components of the vector $\boldsymbol{\beta}$ for at least some fraction of the observations. This is equivalent to the more familiar condition that

$$\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}} \equiv \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}$$

is a positive definite matrix; see Section 6.2 for a discussion of similar asymptotic identification conditions.

We also need to make some assumption about the variances and covariances of the elementary zero functions. If there is just one elementary zero function per observation, we let $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{y})$ denote the n–vector with typical element $f_t(\boldsymbol{\theta}, y_t)$. If there are $m > 1$ elementary zero functions per observation, then we can group all of them into a vector $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{y})$ with nm elements. In either event, we then assume that

$$\mathrm{E}\big(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{y})\boldsymbol{f}^{\top}(\boldsymbol{\theta}, \boldsymbol{y})\big) = \boldsymbol{\Omega}, \tag{9.56}$$

where $\boldsymbol{\Omega}$, which implicitly depends on μ, is a finite, positive definite matrix. Thus we are assuming that, under every DGP $\mu \in \mathbb{M}$, each of the f_t has a finite variance and a finite covariance with every f_s for $s \neq t$.

Estimating Functions and Estimating Equations

Like every procedure that is based on the method of moments, the method of estimating functions replaces relationships like (9.53) that hold in expectation with their empirical, or sample, counterparts. Because $\boldsymbol{\theta}$ is a k–vector, we need k **estimating functions** in order to estimate it. In general, these are weighted averages of the elementary zero functions. Equating the estimating functions to zero yields k **estimating equations**, which must be solved in order to obtain the GMM estimator.

As for the linear regression model, the estimating equations are, in fact, just sample moment conditions which, in most cases, are based on instrumental variables. There are generally more instruments than parameters, and so we need to form linear combinations of the instruments in order to construct precisely k estimating equations. Let $\boldsymbol{W}$ be an $n \times l$ matrix of instruments, which are assumed to be predetermined. Usually, one column of $\boldsymbol{W}$ is a vector of 1s. Now define $\boldsymbol{Z} \equiv \boldsymbol{W}\boldsymbol{J}$, where $\boldsymbol{J}$ is an $l \times k$ matrix with full column rank k. Later, we will discuss how $\boldsymbol{J}$, and hence $\boldsymbol{Z}$, should optimally be chosen, but, for the moment, we take $\boldsymbol{Z}$ as given.

If $\boldsymbol{\theta}_\mu$ is the parameter vector for the DGP μ under which we take expectations, the theoretical moment conditions are

$$\mathrm{E}\big(\boldsymbol{Z}_t^{\top} f_t(\boldsymbol{\theta}_\mu, y_t)\big) = \boldsymbol{0}, \tag{9.57}$$

where $\boldsymbol{Z}_t$ is the t^{th} row of $\boldsymbol{Z}$. Later on, when we take explicit account of the covariance matrix $\boldsymbol{\Omega}$ in formulating the estimating equations, we will need to modify these conditions so that they take the form of conditions (9.31), but (9.57) is all that is required at this stage. In fact, even (9.57) is stronger than we really need. It is sufficient to assume that $\boldsymbol{Z}_t$ and $f_t(\boldsymbol{\theta})$ are asymptotically uncorrelated, which, together with some regularity conditions, implies that

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{Z}_t^{\top} f_t(\boldsymbol{\theta}_\mu, y_t) = \boldsymbol{0}. \tag{9.58}$$

The vector of estimating functions that corresponds to (9.57) or (9.58) is the k-vector $n^{-1}\boldsymbol{Z}^\top\boldsymbol{f}(\boldsymbol{\theta},\boldsymbol{y})$. Equating this vector to zero yields the system of estimating equations

$$\frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{f}(\boldsymbol{\theta},\boldsymbol{y}) = \boldsymbol{0}, \tag{9.59}$$

and solving this system yields $\hat{\boldsymbol{\theta}}$, the **nonlinear GMM estimator**.

Consistency

If we are to prove that the nonlinear GMM estimator is consistent, we must assume that a law of large numbers applies to the vector $n^{-1}\boldsymbol{Z}^\top\boldsymbol{f}(\boldsymbol{\theta},\boldsymbol{y})$. This allows us to define the k-vector of **limiting estimating functions**,

$$\boldsymbol{\alpha}(\boldsymbol{\theta};\mu) \equiv \operatorname*{plim}_{n\to\infty}{}_{\mu}\,\frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{f}(\boldsymbol{\theta},\boldsymbol{y}). \tag{9.60}$$

In words, $\boldsymbol{\alpha}(\boldsymbol{\theta};\mu)$ is the probability limit, under the DGP μ, of the vector of estimating functions. Setting $\boldsymbol{\alpha}(\boldsymbol{\theta};\mu)$ to $\boldsymbol{0}$ yields a set of **limiting estimating equations**.

Either (9.57) or the weaker condition (9.58) implies that $\boldsymbol{\alpha}(\boldsymbol{\theta}_\mu;\mu) = \boldsymbol{0}$ for all $\mu \in \mathbb{M}$. We then need an asymptotic identification condition strong enough to ensure that $\boldsymbol{\alpha}(\boldsymbol{\theta};\mu) \neq \boldsymbol{0}$ for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}_\mu$. In other words, we require that the vector $\boldsymbol{\theta}_\mu$ must be the unique solution to the system of limiting estimating equations. If we assume that such a condition holds, it is straightforward to prove consistency in the nonrigorous way we used in Sections 6.2 and 8.3. Evaluating equations (9.59) at their solution $\hat{\boldsymbol{\theta}}$, we find that

$$\frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{f}(\hat{\boldsymbol{\theta}},\boldsymbol{y}) = \boldsymbol{0}. \tag{9.61}$$

As $n \to \infty$, the left-hand side of this system of equations tends under μ to the vector $\boldsymbol{\alpha}(\operatorname{plim}_\mu\hat{\boldsymbol{\theta}};\mu)$, and the right-hand side remains a zero vector. Given the asymptotic identification condition, the equality in (9.61) can hold asymptotically only if

$$\operatorname*{plim}_{n\to\infty}{}_{\mu}\,\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_\mu.$$

Therefore, we conclude that the nonlinear GMM estimator $\hat{\boldsymbol{\theta}}$, which solves the system of estimating equations (9.59), consistently estimates the parameter vector $\boldsymbol{\theta}_\mu$, for all $\mu \in \mathbb{M}$, provided the asymptotic identification condition is satisfied.

Asymptotic Normality

For ease of notation, we now fix the DGP $\mu \in \mathbb{M}$ and write $\boldsymbol{\theta}_\mu = \boldsymbol{\theta}_0$. Thus $\boldsymbol{\theta}_0$ has its usual interpretation as the "true" parameter vector. In addition, we suppress the explicit mention of the data vector $\boldsymbol{y}$. As usual, the proof that $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normally distributed is based on a Taylor series approximation, a law of large numbers, and a central limit theorem. For the purposes of the first of these, we need to assume that the zero functions f_t are continuously differentiable in the neighborhood of $\boldsymbol{\theta}_0$. If we perform a first-order Taylor expansion of $n^{1/2}$ times (9.59) around $\boldsymbol{\theta}_0$ and introduce some appropriate factors of powers of n, we obtain the result that

$$n^{-1/2}\boldsymbol{Z}^\top\boldsymbol{f}(\boldsymbol{\theta}_0) + n^{-1}\boldsymbol{Z}^\top\boldsymbol{F}(\bar{\boldsymbol{\theta}})n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{0}, \tag{9.62}$$

where the $n \times k$ matrix $\boldsymbol{F}(\boldsymbol{\theta})$ has typical element

$$F_{ti}(\boldsymbol{\theta}) \equiv \frac{\partial f_t(\boldsymbol{\theta})}{\partial \theta_i}, \tag{9.63}$$

where θ_i is the i^{th} element of $\boldsymbol{\theta}$. This matrix, like $\boldsymbol{f}(\boldsymbol{\theta})$ itself, depends implicitly on the vector $\boldsymbol{y}$ and is therefore stochastic. The notation $\boldsymbol{F}(\bar{\boldsymbol{\theta}})$ in (9.62) is the convenient shorthand we introduced in Section 6.2: Row t of the matrix is the corresponding row of $\boldsymbol{F}(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}_t$, where the $\bar{\boldsymbol{\theta}}_t$ all satisfy the inequality

$$\left\|\bar{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0\right\| \leq \left\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0\right\|.$$

The consistency of $\hat{\boldsymbol{\theta}}$ then implies that the $\bar{\boldsymbol{\theta}}_t$ also tend to $\boldsymbol{\theta}_0$ as $n \to \infty$.

The consistency of the $\bar{\boldsymbol{\theta}}_t$ implies that

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{F}(\bar{\boldsymbol{\theta}}) = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{F}(\boldsymbol{\theta}_0). \tag{9.64}$$

Under reasonable regularity conditions, we can apply a law of large numbers to the right-hand side of (9.64), and the probability limit is then deterministic. For asymptotic normality, we also require that it should be nonsingular. This is a condition of **strong asymptotic identification**, of the sort used in Section 6.2. By a first-order Taylor expansion of $\boldsymbol{\alpha}(\boldsymbol{\theta}; \mu)$ around $\boldsymbol{\theta}_0$, where it is equal to $\mathbf{0}$, we see from the definition (9.60) that

$$\boldsymbol{\alpha}(\boldsymbol{\theta}; \mu) \overset{a}{=} \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{F}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0). \tag{9.65}$$

Therefore, the condition that the right-hand side of (9.64) is nonsingular is a strengthening of the condition that $\boldsymbol{\theta}$ is asymptotically identified. Because it is nonsingular, the system of equations

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{F}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = \mathbf{0}$$

has no solution other than $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. By (9.65), this implies that $\boldsymbol{\alpha}(\boldsymbol{\theta}; \mu) \neq \mathbf{0}$ for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, which is the asymptotic identification condition.

Applying the results just discussed to equation (9.62), we find that

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{a}{=} -\Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{F}(\boldsymbol{\theta}_0)\Big)^{-1} n^{-1/2}\boldsymbol{Z}^\top\boldsymbol{f}(\boldsymbol{\theta}_0). \tag{9.66}$$

Next, we apply a central limit theorem to the second factor on the right-hand side of (9.66). Doing so demonstrates that $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normally distributed. By (9.57), the vector $n^{-1/2}\boldsymbol{Z}^\top\boldsymbol{f}(\boldsymbol{\theta}_0)$ must have mean $\mathbf{0}$, and, by (9.56), its covariance matrix is $\operatorname{plim} n^{-1}\boldsymbol{Z}^\top\boldsymbol{\Omega}\boldsymbol{Z}$. In stating this result, we assume that (9.02) holds with the $\boldsymbol{f}(\boldsymbol{\theta}_0)$ in place of the error terms. Then (9.66) implies that the vector $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normally distributed with mean vector $\mathbf{0}$ and covariance matrix

$$\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{F}(\boldsymbol{\theta}_0)\Big)^{-1}\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{\Omega}\boldsymbol{Z}\Big)\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{F}^\top(\boldsymbol{\theta}_0)\boldsymbol{Z}\Big)^{-1}. \tag{9.67}$$

Since this is a sandwich covariance matrix, it is evident that the nonlinear GMM estimator $\hat{\boldsymbol{\theta}}$ is not, in general, an asymptotically efficient estimator.

Asymptotically Efficient Estimation

In order to obtain an asymptotically efficient nonlinear GMM estimator, we need to choose the estimating functions $n^{-1}\boldsymbol{Z}^\top\boldsymbol{f}(\boldsymbol{\theta})$ optimally. This is equivalent to choosing $\boldsymbol{Z}$ optimally. How we should do this depends on what assumptions we make about $\boldsymbol{F}(\boldsymbol{\theta})$ and $\boldsymbol{\Omega}$, the covariance matrix of $\boldsymbol{f}(\boldsymbol{\theta})$. Not surprisingly, we will obtain results very similar to the results for linear GMM estimation obtained in Section 9.2.

We begin with the simplest possible case, in which $\boldsymbol{\Omega} = \sigma^2\mathbf{I}$, and $\boldsymbol{F}(\boldsymbol{\theta}_0)$ is predetermined in the sense that

$$\mathrm{E}\big(\boldsymbol{F}_t(\boldsymbol{\theta}_0)f_t(\boldsymbol{\theta}_0)\big) = \mathbf{0}, \tag{9.68}$$

where $\boldsymbol{F}_t(\boldsymbol{\theta}_0)$ is the t^{th} row of $\boldsymbol{F}(\boldsymbol{\theta}_0)$. If we ignore the probability limits and the factors of n^{-1}, the sandwich covariance matrix (9.67) is in this case proportional to

$$(\boldsymbol{Z}^\top\boldsymbol{F}_0)^{-1}\boldsymbol{Z}^\top\boldsymbol{Z}(\boldsymbol{F}_0^\top\boldsymbol{Z})^{-1}, \tag{9.69}$$

where, for ease of notation, $\boldsymbol{F}_0 \equiv \boldsymbol{F}(\boldsymbol{\theta}_0)$. The inverse of (9.69), which is proportional to the asymptotic precision matrix of the estimator, is

$$\boldsymbol{F}_0^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{F}_0 = \boldsymbol{F}_0^\top\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{F}_0. \tag{9.70}$$

If we set $\boldsymbol{Z} = \boldsymbol{F}_0$, (9.69) is no longer a sandwich, and (9.70) simplifies to $\boldsymbol{F}_0^\top\boldsymbol{F}_0$. The difference between $\boldsymbol{F}_0^\top\boldsymbol{F}_0$ and the general expression (9.70) is

$$\boldsymbol{F}_0^\top\boldsymbol{F}_0 - \boldsymbol{F}_0^\top\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{F}_0 = \boldsymbol{F}_0^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{F}_0,$$

which is a positive semidefinite matrix because $\boldsymbol{M}_{\boldsymbol{Z}} \equiv \mathbf{I} - \boldsymbol{P}_{\boldsymbol{Z}}$ is an orthogonal projection matrix. Thus, in this simple case, the optimal instrument matrix is just $\boldsymbol{F}_0$.

Since we do not know $\boldsymbol{\theta}_0$, it is not feasible to use $\boldsymbol{F}_0$ directly as the matrix of instruments. Instead, we use the trick that leads to the moment conditions

(6.27) which define the NLS estimator. This leads us to solve the estimating equations

$$\frac{1}{n}\boldsymbol{F}^\top(\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}. \tag{9.71}$$

If $\boldsymbol{\Omega} = \sigma^2\mathbf{I}$, and $\boldsymbol{F}(\boldsymbol{\theta}_0)$ is predetermined, solving these equations yields an asymptotically efficient GMM estimator.

It is not valid to use the columns of $\boldsymbol{F}(\boldsymbol{\theta})$ as instruments if condition (9.68) is not satisfied. In that event, the analysis of Section 8.3, taken up again in Section 9.2, suggests that we should replace the rows of $\boldsymbol{F}_0$ by their expectations conditional on the information sets Ω_t generated by variables that are exogenous or predetermined for observation t. Let us define an $n \times k$ matrix $\bar{\boldsymbol{F}}$, in terms of its typical row $\bar{\boldsymbol{F}}_t$, and another $n \times k$ matrix $\boldsymbol{V}$, as follows:

$$\bar{\boldsymbol{F}}_t \equiv \mathrm{E}\big(\boldsymbol{F}_t(\boldsymbol{\theta}_0)\,|\,\Omega_t\big) \quad \text{and} \quad \boldsymbol{V} \equiv \boldsymbol{F}_0 - \bar{\boldsymbol{F}}. \tag{9.72}$$

The matrices $\bar{\boldsymbol{F}}$ and $\boldsymbol{V}$ are entirely analogous to the matrices $\bar{\boldsymbol{X}}$ and $\boldsymbol{V}$ used in Section 8.3. The definitions (9.72) imply that

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{F}}^\top\boldsymbol{F}_0 = \operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{F}}^\top(\bar{\boldsymbol{F}}+\boldsymbol{V}) = \operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{F}}^\top\bar{\boldsymbol{F}}. \tag{9.73}$$

The term $\operatorname{plim} n^{-1}\bar{\boldsymbol{F}}^\top\boldsymbol{V}$ equals $\mathbf{O}$ because (9.72) implies that $\mathrm{E}(\boldsymbol{V}_t\,|\,\Omega_t) = \mathbf{0}$, and the conditional expectation $\bar{\boldsymbol{F}}_t$ belongs to the information set Ω_t.

To find the asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ when $\bar{\boldsymbol{F}}$ is used in place of $\boldsymbol{Z}$ and the covariance matrix of $\boldsymbol{f}(\boldsymbol{\theta})$ is $\sigma^2\mathbf{I}$, we start from expression (9.67). Using (9.73), we obtain

$$\begin{aligned}&\sigma^2\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{F}}^\top\boldsymbol{F}_0\Big)^{-1}\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{F}}^\top\bar{\boldsymbol{F}}\Big)\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{F}_0^\top\bar{\boldsymbol{F}}\Big)^{-1}\\ &= \sigma^2\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{F}}^\top\bar{\boldsymbol{F}}\Big)^{-1}.\end{aligned} \tag{9.74}$$

For any other choice of instrument matrix $\boldsymbol{Z}$, the argument giving (9.73) shows that $\operatorname{plim} n^{-1}\boldsymbol{Z}^\top\boldsymbol{F}_0 = \operatorname{plim} n^{-1}\boldsymbol{Z}^\top\bar{\boldsymbol{F}}$, and so the covariance matrix (9.67) becomes

$$\sigma^2\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{Z}^\top\bar{\boldsymbol{F}}\Big)^{-1}\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{Z}\Big)\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{F}}^\top\boldsymbol{Z}\Big)^{-1}. \tag{9.75}$$

The inverse of (9.75) is $1/\sigma^2$ times the probability limit of

$$\frac{1}{n}\bar{\boldsymbol{F}}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\bar{\boldsymbol{F}} = \frac{1}{n}\bar{\boldsymbol{F}}^\top\boldsymbol{P}_{\boldsymbol{Z}}\bar{\boldsymbol{F}}. \tag{9.76}$$

This expression is analogous to expression (8.21) for the asymptotic precision of the IV estimator for linear regression models with endogenous explanatory variables. Since the difference between $n^{-1}\bar{\boldsymbol{F}}^\top\bar{\boldsymbol{F}}$ and (9.76) is the positive semidefinite matrix $n^{-1}\bar{\boldsymbol{F}}^\top\boldsymbol{M}_{\boldsymbol{Z}}\bar{\boldsymbol{F}}$, we conclude that (9.74) is indeed the asymptotic covariance matrix that corresponds to the optimal choice of $\boldsymbol{Z}$. Therefore, when $\boldsymbol{F}_t(\boldsymbol{\theta})$ is not predetermined, we should use its expectation conditional on Ω_t in the matrix of instruments.

In practice, of course, the matrix $\bar{\boldsymbol{F}}$ is rarely observed. We therefore need to estimate it. The natural way to do so is to regress $\boldsymbol{F}(\boldsymbol{\theta})$ on an $n \times l$ matrix of instruments $\boldsymbol{W}$, where $l \geq k$, with the inequality holding strictly in most cases. This yields fitted values $\boldsymbol{P_W F}(\boldsymbol{\theta})$. If we estimate $\bar{\boldsymbol{F}}$ in this way, the optimal estimating equations become

$$\frac{1}{n}\boldsymbol{F}^{\top}(\boldsymbol{\theta})\boldsymbol{P_W f}(\boldsymbol{\theta}) = \boldsymbol{0}. \tag{9.77}$$

By reasoning like that which led to (8.27) and (9.73), it can be seen that these estimating equations are asymptotically equivalent to the same equations with $\bar{\boldsymbol{F}}$ in place of $\boldsymbol{F}(\boldsymbol{\theta})$. In particular, if $\mathcal{S}(\bar{\boldsymbol{F}}) \subseteq \mathcal{S}(\boldsymbol{W})$, the estimator obtained by solving (9.77) is asymptotically equivalent to the one obtained using the optimal instruments $\bar{\boldsymbol{F}}$.

The estimating equations (9.77) generalize the first-order conditions (8.28) for linear IV estimation and the moment conditions (8.84) for nonlinear IV estimation. As readers are asked to show in Exercise 9.14, the solution to (9.77) in the case of the linear regression model is simply the generalized IV estimator (8.29). As can be seen from (9.67), the asymptotic covariance matrix of the estimator $\hat{\boldsymbol{\theta}}$ defined by (9.77) can be estimated by

$$\hat{\sigma}^2(\hat{\boldsymbol{F}}^{\top}\boldsymbol{P_W}\hat{\boldsymbol{F}})^{-1},$$

where $\hat{\boldsymbol{F}} \equiv \boldsymbol{F}(\hat{\boldsymbol{\theta}})$, and $\hat{\sigma}^2 \equiv n^{-1}\sum_{t=1}^{n} f_t^2(\hat{\boldsymbol{\theta}})$, the average of the squares of the elementary zero functions evaluated at $\hat{\boldsymbol{\theta}}$, is a natural estimator of σ^2.

Efficient Estimation with an Unknown Covariance Matrix

When the covariance matrix $\boldsymbol{\Omega}$ is unknown, the GMM estimators defined by the estimating equations (9.71) or (9.77), according to whether or not $\boldsymbol{F}(\boldsymbol{\theta})$ is predetermined, are no longer asymptotically efficient in general. But, just as we did in Section 9.3 with regression models, we can obtain estimates that are efficient for a given set of instruments by using a heteroskedasticity-consistent or a HAC estimator.

Suppose there are $l > k$ instruments which form an $n \times l$ matrix $\boldsymbol{W}$. As in Section 9.2, we can construct estimating equations with instruments $\boldsymbol{Z} = \boldsymbol{WJ}$, using a full-rank $l \times k$ matrix $\boldsymbol{J}$ to select k linear combinations of the full set of instruments. The asymptotic covariance matrix of the estimator obtained by solving these equations is then, by (9.67),

$$\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{J}^{\top}\boldsymbol{W}^{\top}\boldsymbol{F}_0\Big)^{-1}\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{J}^{\top}\boldsymbol{W}^{\top}\boldsymbol{\Omega}\boldsymbol{WJ}\Big)\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{F}_0^{\top}\boldsymbol{WJ}\Big)^{-1}. \tag{9.78}$$

Copyright © 2003, Russell Davidson and James G. MacKinnon

This looks just like (9.07) with $\boldsymbol{F}_0$ in place of the regressor matrix $\boldsymbol{X}$. The optimal choice of $\boldsymbol{J}$ is therefore just (9.08) with $\boldsymbol{F}_0$ in place of $\boldsymbol{X}$. Since (9.08) depends on the unknown true $\boldsymbol{\Omega}$, we replace $n^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W}$ by an estimator $\hat{\boldsymbol{\Sigma}}$, which could be either a heteroskedasticity-consistent or a HAC estimator. This yields the estimating equations

$$\boldsymbol{F}^\top(\boldsymbol{\theta})\boldsymbol{W}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}^\top\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}, \tag{9.79}$$

and the asymptotic covariance matrix (9.78) simplifies to

$$\Big(\operatorname*{plim}_{n\to\infty} n^{-2}\boldsymbol{F}_0^\top\boldsymbol{W}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}^\top\boldsymbol{F}_0\Big)^{-1}, \tag{9.80}$$

in which, if $\boldsymbol{F}(\boldsymbol{\theta})$ is not predetermined, we may write $\bar{\boldsymbol{F}}$ instead of $\boldsymbol{F}_0$ without changing the limit. In practice, we can use

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\theta}}) = n(\hat{\boldsymbol{F}}^\top\boldsymbol{W}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}^\top\hat{\boldsymbol{F}})^{-1}, \tag{9.81}$$

where $\hat{\boldsymbol{F}} \equiv \boldsymbol{F}(\hat{\boldsymbol{\theta}})$, to estimate the covariance matrix of $\hat{\boldsymbol{\theta}}$. As with the estimator (9.41) for the linear regression case, the factor of n is needed to offset the factor of n^{-1} in $\hat{\boldsymbol{\Sigma}}$. The matrix (9.81) can be used to construct Wald tests and asymptotic confidence intervals in the usual way.

Efficient Estimation with a Known Covariance Matrix

When the covariance matrix $\boldsymbol{\Omega}$ is known, we can obtain a fully efficient GMM estimator. As before, we let $\boldsymbol{\Psi}$ denote an $n \times n$ matrix which satisfies the equation $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top$. The variance of the vector $\boldsymbol{\Psi}^\top\boldsymbol{f}(\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is the true parameter vector for the DGP that generates the data, is then

$$\mathrm{E}\big(\boldsymbol{\Psi}^\top\boldsymbol{f}(\boldsymbol{\theta}_0)\boldsymbol{f}^\top(\boldsymbol{\theta}_0)\boldsymbol{\Psi}\big) = \boldsymbol{\Psi}^\top\boldsymbol{\Omega}\boldsymbol{\Psi} = \mathbf{I}.$$

Thus the components of the vector $\boldsymbol{\Psi}^\top\boldsymbol{f}(\boldsymbol{\theta})$ form a set of zero functions that are homoskedastic and serially uncorrelated. As we mentioned in Section 9.2, it is often possible to choose $\boldsymbol{\Psi}$ in such a way that these components can be thought of as innovations in the sense of Section 4.5, and in this case $\boldsymbol{\Psi}$ is usually upper triangular.

The matrix $\boldsymbol{\Psi}$ does not depend on the parameters $\boldsymbol{\theta}$. Therefore, the matrix of derivatives of the transformed zero functions in the vector $\boldsymbol{\Psi}^\top\boldsymbol{f}(\boldsymbol{\theta})$ is just $\boldsymbol{\Psi}^\top\boldsymbol{F}(\boldsymbol{\theta})$. Consequently, if the t^{th} row of $\boldsymbol{\Psi}^\top\boldsymbol{F}(\boldsymbol{\theta})$ is predetermined with respect to the t^{th} component of $\boldsymbol{\Psi}^\top\boldsymbol{f}(\boldsymbol{\theta})$, the optimal estimating equations are constructed using the columns of $\boldsymbol{\Psi}^\top\boldsymbol{F}(\boldsymbol{\theta}_0)$ as instruments. Because $\boldsymbol{\theta}_0$ is not known, the optimal instruments are estimated along with the parameters by using the estimating equations

$$\frac{1}{n}\boldsymbol{F}^\top(\boldsymbol{\theta})\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\boldsymbol{f}(\boldsymbol{\theta}) = \frac{1}{n}\boldsymbol{F}^\top(\boldsymbol{\theta})\boldsymbol{\Omega}^{-1}\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}, \tag{9.82}$$

as in (9.71). The asymptotic covariance matrix of the resulting estimator is

$$\operatorname{Var}\bigl(\operatorname*{plim}_{n\to\infty} n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\bigr) = \operatorname*{plim}_{n\to\infty}\Bigl(\frac{1}{n}\boldsymbol{F}_0^{\top}\boldsymbol{\varOmega}^{-1}\boldsymbol{F}_0\Bigr)^{-1}, \tag{9.83}$$

where, as usual, $\boldsymbol{F}_0 \equiv \boldsymbol{F}(\boldsymbol{\theta}_0)$. The derivation of (9.83) from (9.67) is quite straightforward; see Exercise 9.15. In practice, the covariance matrix of $\hat{\boldsymbol{\theta}}$ is normally estimated by

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{F}}^{\top}\boldsymbol{\varOmega}^{-1}\hat{\boldsymbol{F}})^{-1}. \tag{9.84}$$

If the t^{th} row of $\boldsymbol{\varPsi}^{\top}\boldsymbol{F}(\boldsymbol{\theta})$ is not predetermined with respect to the t^{th} component of $\boldsymbol{\varPsi}^{\top}\boldsymbol{f}(\boldsymbol{\theta})$, and if this component is an innovation, then we can determine the optimal instruments just as we did in Section 9.2. By analogy with (9.24), we define the matrix $\bar{\boldsymbol{F}}(\boldsymbol{\theta})$ implicitly by the equation

$$\operatorname{E}\bigl((\boldsymbol{\varPsi}^{\top}\boldsymbol{F}(\boldsymbol{\theta}))_t \,|\, \Omega_t\bigr) = (\boldsymbol{\varPsi}^{\top}\bar{\boldsymbol{F}}(\boldsymbol{\theta}))_t. \tag{9.85}$$

As in Section 9.2, making this definition explicit depends on the details of the particular model under study. The moment conditions for fully efficient estimation are then given by (9.82) with $\boldsymbol{F}(\boldsymbol{\theta})$ replaced by $\bar{\boldsymbol{F}}(\boldsymbol{\theta})$. The asymptotic covariance matrix is (9.83) with $\boldsymbol{F}_0$ replaced by $\bar{\boldsymbol{F}}_0$, and the covariance matrix of $\hat{\boldsymbol{\theta}}$ can be estimated by (9.84) with $\hat{\boldsymbol{F}}$ replaced by $\bar{\boldsymbol{F}}(\hat{\boldsymbol{\theta}})$. All of these claims are proved in the same way as were the corresponding ones for linear regressions in Section 9.2.

When the matrix $\bar{\boldsymbol{F}}(\boldsymbol{\theta})$ is not observable, as is frequently the case, we can often find an $n \times l$ matrix of instruments $\boldsymbol{W}$, where usually $l > k$, such that $\boldsymbol{W}$ satisfies the predeterminedness condition in its form (9.30), and such that $\mathcal{S}(\boldsymbol{F}(\boldsymbol{\theta}_0)) \subseteq \mathcal{S}(\boldsymbol{W})$. In such cases, overidentified estimation that makes use of the transformed zero functions $\boldsymbol{\varPsi}^{\top}\boldsymbol{f}(\boldsymbol{\theta})$ and the transformed instruments $\boldsymbol{\varPsi}^{\top}\boldsymbol{W}$ yields asymptotically efficient estimates. The results of Exercises 9.8 and 9.9 can also be readily extended to the present nonlinear case.

Minimizing Criterion Functions

The nonlinear GMM estimators we have discussed in this section can all, like the ones for linear regression models, be obtained by minimizing appropriately chosen quadratic forms. We restrict our attention to cases in which $\operatorname{plim} n^{-1}\boldsymbol{F}^{\top}(\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{\theta}) \neq \boldsymbol{0}$, and we employ an $n \times l$ matrix of instruments, $\boldsymbol{W}$. When the covariance matrix $\boldsymbol{\varOmega}$ of the elementary zero functions is unknown, but a heteroskedasticity-consistent or HAC estimator $\hat{\boldsymbol{\varSigma}}$ is available, the appropriate GMM criterion function is

$$\frac{1}{n}\boldsymbol{f}^{\top}(\boldsymbol{\theta})\boldsymbol{W}\hat{\boldsymbol{\varSigma}}^{-1}\boldsymbol{W}^{\top}\boldsymbol{f}(\boldsymbol{\theta}). \tag{9.86}$$

Minimizing this function with respect to $\boldsymbol{\theta}$ is equivalent to solving the estimating equations (9.79).

In the case in which the matrix $\boldsymbol{\Omega}$ is known, or can be estimated consistently, the fully efficient estimators of the previous subsection can be obtained by minimizing the quadratic form

$$\boldsymbol{f}^\top(\boldsymbol{\theta})\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{\Psi}^\top\boldsymbol{W}}\boldsymbol{\Psi}^\top\boldsymbol{f}(\boldsymbol{\theta}), \tag{9.87}$$

where $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top = \boldsymbol{\Omega}^{-1}$, the components of $\boldsymbol{\Psi}^\top\boldsymbol{f}(\boldsymbol{\theta}_0)$ are innovations, and the matrix $\boldsymbol{W}$ satisfies the predeterminedness condition in the form (9.30). For full efficiency, the span $\mathcal{S}(\boldsymbol{W})$ of the instruments must (asymptotically) include as a subspace the span of the $\bar{\boldsymbol{F}}(\boldsymbol{\theta}_0)$, as defined in (9.85). In Exercise 9.16, readers are asked to check that minimizing (9.87) is asymptotically equivalent to solving the optimal estimating equations.

Fortunately, we need not treat (9.86) and (9.87) separately. As in Section 9.4, expression (9.86) is asymptotically unchanged if we replace $\hat{\boldsymbol{\Sigma}}$ by $n^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W}$, where $\boldsymbol{\Omega}$ is the true covariance matrix of the zero functions. Making this replacement, we see that both (9.86) and (9.87) can be written as

$$Q(\boldsymbol{\theta}, \boldsymbol{y}) \equiv \boldsymbol{f}^\top(\boldsymbol{\theta})\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{f}(\boldsymbol{\theta}), \tag{9.88}$$

where $\boldsymbol{A} = \boldsymbol{\Psi}^{-1}\boldsymbol{W}$ and $\boldsymbol{A} = \boldsymbol{\Psi}^\top\boldsymbol{W}$ for the criterion functions (9.86) and (9.87), respectively. Note how closely (9.88) resembles expression (9.45) for the linear regression case.

It is often more convenient to compute GMM estimators by minimizing a criterion function than by directly solving a set of estimating equations. One advantage is that algorithms for minimizing functions tend to be more stable numerically than algorithms for solving sets of nonlinear equations. Another advantage is that the criterion function may have more than one stationary point. In this event, the estimating equations are satisfied at each of these stationary points, although the criterion function may have a unique global minimum, which then corresponds to the solution of interest.

However, the main advantage of working with criterion functions is that the minimized value of a GMM criterion function can be used for testing, as we have already discussed for the linear regression case in Section 9.4. Notice that the factor of n^{-1} in (9.86), which does not matter for estimation, is essential when the criterion function is being used for testing. Its role is to offset the factor of n^{-1} in the definition of $\hat{\boldsymbol{\Sigma}}$.

Tests Based on the GMM Criterion Function

The Hansen-Sargan overidentification test statistic is $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y})$, the minimized value of the GMM criterion function. Up to an irrelevant scalar factor, the first-order conditions for the minimization of (9.88) are

$$\boldsymbol{F}^\top(\hat{\boldsymbol{\theta}})\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{f}(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}, \tag{9.89}$$

and it follows from this, either by a Taylor expansion or directly by using the result (9.66), that

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{a}{=} -\Big(\frac{1}{n}\boldsymbol{F}_0^{\top}\boldsymbol{\varPsi}\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{F}_0\Big)^{-1} n^{-1/2}\boldsymbol{F}_0^{\top}\boldsymbol{\varPsi}\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{f}_0,$$

where, as usual, $\boldsymbol{F}_0$ and $\boldsymbol{f}_0$ denote $\boldsymbol{F}(\boldsymbol{\theta}_0)$ and $\boldsymbol{f}(\boldsymbol{\theta}_0)$, respectively. We now follow quite closely the calculations of Section 9.4 in order to show that the minimized quadratic form $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y})$ is asymptotically distributed as $\chi^2(l-k)$. By a short Taylor expansion, we see that

$$\begin{aligned}\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{f}(\hat{\boldsymbol{\theta}}) &\overset{a}{=} \boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{f}_0 + n^{-1/2}\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{F}_0\, n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\\ &\overset{a}{=} \boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{f}_0 - n^{-1/2}\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{F}_0\big(\frac{1}{n}\boldsymbol{F}_0^{\top}\boldsymbol{\varPsi}\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{F}_0\big)^{-1} n^{-1/2}\boldsymbol{F}_0^{\top}\boldsymbol{\varPsi}\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{f}_0\\ &= (\mathbf{I} - \boldsymbol{P}_{\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{F}_0})\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{f}_0,\end{aligned}$$

where $\boldsymbol{P}_{\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{F}_0}$ projects orthogonally on to $\mathcal{S}(\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{F}_0)$. Thus $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y})$, the minimized value of the criterion function (9.88), is

$$\begin{aligned}\boldsymbol{f}^{\top}(\hat{\boldsymbol{\theta}})\boldsymbol{\varPsi}\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{f}(\hat{\boldsymbol{\theta}}) &\overset{a}{=} \boldsymbol{f}_0^{\top}\boldsymbol{\varPsi}\boldsymbol{P}_{\!\boldsymbol{A}}(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{F}_0})\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{f}_0\\ &= \boldsymbol{f}_0^{\top}\boldsymbol{\varPsi}\big(\boldsymbol{P}_{\!\boldsymbol{A}} - \boldsymbol{P}_{\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{F}_0}\big)\boldsymbol{\varPsi}^{\top}\boldsymbol{f}_0.\end{aligned} \tag{9.90}$$

Because $\mathcal{S}(\boldsymbol{P}_{\!\boldsymbol{A}}\boldsymbol{\varPsi}^{\top}\boldsymbol{F}_0) \subseteq \mathcal{S}(\boldsymbol{A})$, the difference of projection matrices in the last expression above is itself an orthogonal projection matrix, of which the image is of dimension $l-k$. As with (9.48), we see that estimating $\boldsymbol{\theta}$ uses up k degrees of freedom. By essentially the same argument as was used for (9.48), it can be shown that (9.90) is asymptotically distributed as $\chi^2(l-k)$. Thus, as expected, $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y})$ is the Hansen-Sargan test statistic for nonlinear GMM estimation.

As in the case of linear regression models, the difference between the GMM criterion function (9.88) evaluated at restricted estimates and evaluated at unrestricted estimates is asymptotically distributed as $\chi^2(r)$ when there are r equality restrictions. We will not prove this result, which was proved for the linear case in Section 9.3. However, we will present a very simple argument which provides an intuitive explanation.

Let $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ denote, respectively, the vectors of restricted and unrestricted (feasible) efficient GMM estimates. From the result for the Hansen-Sargan test that was just proved, we know that $Q(\tilde{\boldsymbol{\theta}}, \boldsymbol{y})$ and $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y})$ are asymptotically distributed as $\chi^2(l-k+r)$ and $\chi^2(l-k)$, respectively. Therefore, since a random variable that follows the $\chi^2(m)$ distribution is equal to the sum of m independent $\chi^2(1)$ variables,

$$Q(\tilde{\boldsymbol{\theta}}, \boldsymbol{y}) \overset{a}{=} \sum_{i=1}^{l-k+r} x_i^2 \quad\text{and}\quad Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y}) \overset{a}{=} \sum_{i=1}^{l-k} y_i^2, \tag{9.91}$$

Copyright © 2003, Russell Davidson and James G. MacKinnon

where the x_i and y_i are independent, standard normal random variables. Now suppose that the first $l-k$ of the x_i are equal to the corresponding y_i. If so, (9.91) implies that

$$Q(\tilde{\boldsymbol{\theta}}, \boldsymbol{y}) - Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y}) \overset{a}{=} \sum_{i=1}^{l-k+r} x_i^2 - \sum_{i=1}^{l-k} x_i^2 = \sum_{i=l-k+1}^{l-k+r} x_i^2. \tag{9.92}$$

Since the leftmost expression here is the test statistic we are interested in and the rightmost expression is evidently distributed as $\chi^2(r)$, we have apparently proved the result. The proof is not complete, of course, because we have not shown that the first $l-k$ of the x_i are, in fact, equal to the corresponding y_i. To prove this, we would need to show that, asymptotically, $Q(\tilde{\boldsymbol{\theta}}, \boldsymbol{y})$ is equal to $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y})$ plus another random variable independent of $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y})$. This other random variable would then be equal to the rightmost expression in (9.92).

Nonlinear GMM Estimators: Overview

We have discussed a large number of nonlinear GMM estimators, and it can be confusing to keep track of them all. We therefore conclude this section with a brief summary of the principal cases that are likely to be encountered in applied econometric work.

Case 1. Scalar covariance matrix: $\boldsymbol{\Omega} = \sigma^2\mathbf{I}$.

When $\operatorname{plim} n^{-1}\boldsymbol{F}^\top(\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}$, we solve the estimating equations (9.71) to obtain an efficient estimator. This is equivalent to minimizing $\boldsymbol{f}^\top(\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{\theta})$. The estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\theta}}) = \hat{\sigma}^2(\hat{\boldsymbol{F}}^\top\hat{\boldsymbol{F}})^{-1},$$

where $\hat{\sigma}^2$ consistently estimates σ^2. If the model is a nonlinear regression model, then $\hat{\boldsymbol{\theta}}$ is really the nonlinear least-squares estimator discussed in Section 6.3.

When $\operatorname{plim} n^{-1}\boldsymbol{F}^\top(\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{\theta}) \neq \mathbf{0}$, we must replace $\boldsymbol{F}(\boldsymbol{\theta})$ by an estimate of its conditional expectation. This means that we solve the estimating equations (9.77), which is equivalent to minimizing $\boldsymbol{f}^\top(\boldsymbol{\theta})\boldsymbol{P_W}\boldsymbol{f}(\boldsymbol{\theta})$. The estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\theta}}) = \hat{\sigma}^2(\hat{\boldsymbol{F}}^\top\boldsymbol{P_W}\hat{\boldsymbol{F}})^{-1}.$$

If the model is a nonlinear regression model, then $\hat{\boldsymbol{\theta}}$ is really the nonlinear instrumental variables estimator discussed in Section 8.9.

Case 2. Covariance matrix known up to a scalar factor: $\boldsymbol{\Omega} = \sigma^2\boldsymbol{\Delta}$.

When $\operatorname{plim} n^{-1}\boldsymbol{F}^\top(\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}$, we solve the estimating equations (9.82), with $\boldsymbol{\Omega}$ replaced by $\boldsymbol{\Delta}$, to obtain an efficient estimator. This is equivalent to minimizing $\boldsymbol{f}^\top(\boldsymbol{\theta})\boldsymbol{\Delta}^{-1}\boldsymbol{f}(\boldsymbol{\theta})$. The estimated covariance matrix is

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\theta}}) = \hat{\sigma}^2(\hat{\boldsymbol{F}}^\top\boldsymbol{\Delta}^{-1}\hat{\boldsymbol{F}})^{-1},$$

Copyright © 2003, Russell Davidson and James G. MacKinnon

where $\hat{\sigma}^2$ consistently estimates σ^2. If the underlying model is a nonlinear regression model, then $\hat{\boldsymbol{\theta}}$ is really the nonlinear GLS estimator discussed in Section 7.3.

When $\operatorname{plim} n^{-1}\boldsymbol{F}^\top(\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{\theta}) \neq \mathbf{0}$, we again must replace $\boldsymbol{F}(\boldsymbol{\theta})$ by an estimate of its conditional expectation. This means that we should solve the estimating equations (9.89) with $\boldsymbol{A} = \boldsymbol{\Psi}^\top\boldsymbol{W}$, where $\boldsymbol{\Psi}$ satisfies $\boldsymbol{\Delta}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top$. This is equivalent to minimizing (9.88) with the same definition of $\boldsymbol{A}$. The estimated covariance matrix is

$$\hat{\sigma}^2(\hat{\boldsymbol{F}}^\top\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{\Psi}^\top\boldsymbol{W}}\boldsymbol{\Psi}^\top\hat{\boldsymbol{F}})^{-1}.$$

If the model is a linear regression model, then $\hat{\boldsymbol{\theta}}$ is the fully efficient GMM estimator (9.26) whenever the span of the instruments $\boldsymbol{W}$ includes the span of the optimal instruments $\bar{\boldsymbol{X}}$.

When the matrix $\boldsymbol{\Delta}$ is unknown but depends on a fixed number of parameters that can be estimated consistently, we can replace $\boldsymbol{\Delta}$ by a consistent estimator $\hat{\boldsymbol{\Delta}}$ and proceed as if it were known, as in feasible GLS estimation.

Case 3. Unknown diagonal or general covariance matrix.

This is the most commonly encountered case in which GMM estimation is explicitly used. Fully efficient estimation is no longer possible, but we can still obtain estimates that are efficient for a given set of instruments by using a consistent estimator $\hat{\boldsymbol{\Sigma}}$ of the matrix $\boldsymbol{\Sigma}$ defined in (9.33). This estimator is heteroskedasticity-consistent if $\boldsymbol{\Omega}$ is assumed to be diagonal and some sort of HAC estimator otherwise. Whether or not $\operatorname{plim} n^{-1}\boldsymbol{F}^\top(\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}$, we solve the estimating equations (9.79), which is equivalent to minimizing (9.86). The estimated covariance matrix is (9.81). If there is to be any gain in efficiency relative to NLS or nonlinear IV, it is essential that l, the number of columns of $\boldsymbol{W}$, be greater than k, the number of parameters to be estimated.

The consistent estimator $\hat{\boldsymbol{\Sigma}}$ is usually obtained from initial estimates that are consistent but inefficient. These may be NLS estimates, nonlinear IV estimates, or GMM estimates that do not use the optimal weighting matrix. The efficient GMM estimates are usually obtained by minimizing the criterion function (9.86), and the minimized value of this criterion function then serves as a Hansen-Sargan test statistic.

The first-round estimates $\hat{\boldsymbol{\theta}}$ can be used to obtain a new estimate of $\boldsymbol{\Sigma}$, which can then be used to obtain a second-round estimate of $\boldsymbol{\theta}$, which can be used to obtain yet another estimate of $\boldsymbol{\Sigma}$, and so on, until the process converges or the investigator loses patience. For a correctly specified model, all of these estimators have the same asymptotic distribution. However, performing more than one iteration often improves the finite-sample properties of the estimator. Thus, if computing cost is not a problem, it may well be best to use the continuously updated estimator that has been iterated to convergence.

For a more thorough treatment of the asymptotic theory of GMM estimation, see Newey and McFadden (1994).

Copyright © 2003, Russell Davidson and James G. MacKinnon

9.6 The Method of Simulated Moments

It is often possible to use GMM even when the elementary zero functions cannot be evaluated analytically. Suppose they take the form

$$f_t(y_t, \boldsymbol{\theta}) = h_t(y_t) - m_t(\boldsymbol{\theta}), \quad t = 1, \ldots, n, \tag{9.93}$$

where the function $h_t(y_t)$ depends only on y_t and, possibly, on exogenous or predetermined variables. The function $m_t(\boldsymbol{\theta})$ depends only on exogenous or predetermined variables and on the parameters. Like a regression function, it is the expectation of $h_t(y_t)$, conditional on the information set Ω_t, under a DGP characterized by the parameter vector $\boldsymbol{\theta}$. Estimating such a model by GMM presents no special difficulty if the form of $m_t(\boldsymbol{\theta})$ is known analytically, but this need not be the case.

There are numerous situations in which $m_t(\boldsymbol{\theta})$ may not be known analytically. In particular, it may well occur in models which involve **latent variables**, that is, variables which are not observable by an econometrician. The variables that actually are observed are related to the latent variables in such a way that knowing the former does not permit the values of the latter to be fully recovered. One example, which was discussed in Section 8.2, is economic variables that are observed with measurement error. Another example is variables that are **censored**, in the sense that they are observed only to a limited extent, for instance when only the sign of the variable is observed, or when all negative values are replaced by zeros. Even if the distributions of the latent variables are tractable, those of the observed variables may not be. In particular, it may not be possible to obtain analytic expressions for their expectations, or for the expectations of functions of them.

Even when analytic expressions are not available, it is often possible to obtain simulation-based estimates of the distributions of the observed variables. For example, suppose that an observed variable is equal to a latent variable plus a measurement error of some known distribution, possibly dependent on the parameter vector $\boldsymbol{\theta}$. Suppose further that, for a DGP characterized by $\boldsymbol{\theta}$, we can readily generate simulated values of the latent variable. Simulated values of the observed variable can then be generated by adding simulated measurement errors, drawn from their known distribution, to the simulated values of the latent variable. The mean of these drawings then provides an estimate of the expectation of the observed variable.

In general, an **unbiased simulator** for the unknown expectation $m_t(\boldsymbol{\theta})$ is any function $m_t^*(u_t^*, \boldsymbol{\theta})$ of the model parameters, variables in Ω_t, and a random variable u_t^*, which either has a known distribution or can be simulated, such that, for all $\boldsymbol{\theta}$ in the parameter space, $\mathrm{E}\bigl(m_t^*(u_t^*, \boldsymbol{\theta})\bigr) = m_t(\boldsymbol{\theta})$. To simplify notation, we write u_t^* as a scalar random variable, but it may well be a vector of random variables in practical situations of interest.

The conceptually simplest unbiased simulator can be implemented as follows. For given $\boldsymbol{\theta}$, we obtain S simulated values y_{ts}^* of the observed variable under the DGP characterized by $\boldsymbol{\theta}$, making use of S random numbers u^*_{ts}. Then we let $m^*(u^*_{ts}, \boldsymbol{\theta}) = h_t(y^*_{ts})$. If (9.93) is indeed a zero function, then $h_t(y^*_{ts})$ must have expectation $m_t(\boldsymbol{\theta})$, and it is obvious that the sample mean of the simulated values $h(y^*_{ts})$ is a simulation-based estimate of that expectation. This simple simulator, which is applicable whether or not the model involves any latent variables, is not the only possible simulator, and it may not be the most desirable one for some purposes. However, we will not consider more complicated simulators in this book.

If an unbiased simulator is available, the elementary zero functions (9.93) can be replaced by the functions

$$f^*_t(y_t, \boldsymbol{\theta}) = h_t(y_t) - \frac{1}{S}\sum_{s=1}^{S} m^*_t(u^*_{ts}, \boldsymbol{\theta}), \tag{9.94}$$

where the u^*_{ts}, $t = 1, \ldots, n$, $s = 1, \ldots, S$, are mutually independent draws. Since these draws are computer generated, they are evidently independent of the y_t. The functions (9.94) are legitimate elementary zero functions, even in the trivial case in which $S = 1$. If the true DGP is characterized by $\boldsymbol{\theta}_0$, then $\mathrm{E}\big(h_t(y_t)\big) = m_t(\boldsymbol{\theta}_0)$ by definition, and $\mathrm{E}\big(m^*_t(u^*_{ts}, \boldsymbol{\theta}_0)\big) = m_t(\boldsymbol{\theta}_0)$ for all s by construction. It follows that the expectation (9.94) is zero for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, but not in general for other values of $\boldsymbol{\theta}$.

The application of GMM to the zero functions (9.94) is called the **method of simulated moments**, or **MSM**. We can use an $n \times l$ matrix $\boldsymbol{W}$ of appropriate instruments, with $l \geq k$, in order to form the empirical moments

$$\boldsymbol{W}^\top \boldsymbol{f}^*(\boldsymbol{\theta}), \tag{9.95}$$

in which the n–vector of functions $\boldsymbol{f}^*(\boldsymbol{\theta})$ has typical element $f^*_t(y_t, \boldsymbol{\theta})$. A GMM estimator that is efficient relative to this set of empirical moments may be obtained by minimizing the quadratic form

$$Q(\boldsymbol{\theta}, \boldsymbol{y}) \equiv \frac{1}{n}\boldsymbol{f}^{*\top}(\boldsymbol{\theta})\boldsymbol{W}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}^\top\boldsymbol{f}^*(\boldsymbol{\theta}) \tag{9.96}$$

with respect to $\boldsymbol{\theta}$, where $\hat{\boldsymbol{\Sigma}}$ consistently estimates the covariance matrix of $n^{-1/2}\boldsymbol{W}^\top\boldsymbol{f}^*(\boldsymbol{\theta})$.

Minimizing the criterion function (9.96) with respect to $\boldsymbol{\theta}$ proceeds in the usual way, with one important proviso. Each evaluation of $\boldsymbol{f}^*(\boldsymbol{\theta})$ requires a large number of pseudo-random numbers (generally, at least nS of them). It is absolutely essential that the *same* set of random numbers be used every time $\boldsymbol{f}^*(\boldsymbol{\theta})$ is evaluated for a new value of the parameter vector $\boldsymbol{\theta}$. Otherwise, (9.96) would change not only as a result of changes in $\boldsymbol{\theta}$ but also as a result of changes in the random numbers used for the simulation. Therefore, if the algorithm happened to evaluate the criterion function twice at the same parameter vector, it would obtain two different values of $Q(\boldsymbol{\theta}, \boldsymbol{y})$, and it could not possibly tell where the minimum was located.

The details of the simulations, of course, differ from case to case. An important point is that, since we require a fully specified DGP in order to generate the simulated data, it is generally necessary to make stronger distributional assumptions for the purposes of MSM estimation than for the purposes of GMM estimation.

The Asymptotic Distribution of the MSM Estimator

Because the criterion function (9.96) is based on genuine zero functions, the estimator $\hat{\boldsymbol{\theta}}_{\text{MSM}}$ obtained by minimizing it is consistent whenever the parameters are identified. However, as we will see in a moment, using simulated quantities does affect the asymptotic covariance matrix of the estimator, although the effect is generally very small if S is a reasonably large number.

The first-order conditions for minimizing (9.96), ignoring a factor of $2/n$, are

$$\boldsymbol{F}^{*\top}(\boldsymbol{\theta})\boldsymbol{W}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}^\top\boldsymbol{f}^*(\boldsymbol{\theta}) = \mathbf{0}, \tag{9.97}$$

where $\boldsymbol{F}^*(\boldsymbol{\theta})$ is the $n \times k$ matrix of which the ti^{th} element is $\partial f_t^*(y_t, \boldsymbol{\theta})/\partial\theta_i$. The solution to these equations is $\hat{\boldsymbol{\theta}}_{\text{MSM}}$. Although conditions (9.97) look very similar to conditions (9.79), the covariance matrix is, in general, a good deal more complicated.

From (9.97), it can be seen that the instruments effectively used by the MSM estimator are $\boldsymbol{W}\hat{\boldsymbol{\Sigma}}^{-1}(n^{-1}\boldsymbol{W}^\top\boldsymbol{F}_0^*)$, where $\boldsymbol{F}_0^* \equiv \boldsymbol{F}^*(\boldsymbol{\theta}_0)$, and a factor of n^{-1} has been used to keep the expression of order unity as $n \to \infty$. If we think of the effective instruments as $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{J}$, then $\boldsymbol{J} = \hat{\boldsymbol{\Sigma}}^{-1}(n^{-1}\boldsymbol{W}^\top\boldsymbol{F}_0^*)$.

The asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}}_{\text{MSM}} - \boldsymbol{\theta}_0)$ can now be found by using the general formula (9.78) for the asymptotic covariance matrix of an efficient GMM estimator with unknown covariance matrix. This is a sandwich estimator of the form $\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}$, and we find that

$$\begin{aligned}
\boldsymbol{A} &= \operatorname*{plim}_{n\to\infty}(n^{-1}\boldsymbol{F}_0^{*\top}\boldsymbol{W})\hat{\boldsymbol{\Sigma}}^{-1}(n^{-1}\boldsymbol{W}^\top\boldsymbol{F}_0^*), \text{ and}\\
\boldsymbol{B} &= \operatorname*{plim}_{n\to\infty}(n^{-1}\boldsymbol{F}_0^{*\top}\boldsymbol{W})\hat{\boldsymbol{\Sigma}}^{-1}(n^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W})\hat{\boldsymbol{\Sigma}}^{-1}(n^{-1}\boldsymbol{W}^\top\boldsymbol{F}_0^*),
\end{aligned} \tag{9.98}$$

where $\boldsymbol{\Omega}$ is the $n \times n$ covariance matrix of $\boldsymbol{f}^*(\boldsymbol{\theta}_0)$.

The ti^{th} element of $\boldsymbol{F}^*(\boldsymbol{\theta})$ is, from (9.94),

$$F_{ti}^*(\boldsymbol{\theta}) = -\frac{1}{S}\sum_{s=1}^{S}\frac{\partial m_t^*(u_{ts}^*, \boldsymbol{\theta})}{\partial\theta_i}.$$

If m_t^* is differentiable with respect to $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}$, then we can

Copyright © 2003, Russell Davidson and James G. MacKinnon

differentiate the relation $\mathrm{E}\big(m_t^*(u_t^*, \boldsymbol{\theta})\big) = m_t(\boldsymbol{\theta})$ to find that

$$\mathrm{E}\Big(\frac{\partial m_t^*(u_t^*, \boldsymbol{\theta})}{\partial \theta_i}\Big) = \frac{\partial m_t(\boldsymbol{\theta})}{\partial \theta_i}.$$

We denote by $\boldsymbol{M}(\boldsymbol{\theta})$ the $n \times k$ matrix with typical element $\partial m_t / \partial \theta_i(\boldsymbol{\theta})$. By a law of large numbers, we then see that $\operatorname{plim} n^{-1}\boldsymbol{W}^\top\boldsymbol{F}_0^* = \operatorname{plim} n^{-1}\boldsymbol{W}^\top\boldsymbol{M}_0$, where $\boldsymbol{M}_0 \equiv \boldsymbol{M}(\boldsymbol{\theta}_0)$.

Consider next the covariance matrix $\boldsymbol{\Omega}$ of $\boldsymbol{f}^*(\boldsymbol{\theta}_0)$. The original data y_t are of course completely independent of the simulated u_{ts}^*, and the simulated data are independent across simulations. Thus, from (9.94), we see that

$$\boldsymbol{\Omega} = \mathrm{Var}\big(\boldsymbol{h}(\boldsymbol{y})\big) + \frac{1}{S}\mathrm{Var}\big(\boldsymbol{m}^*(\boldsymbol{\theta}_0)\big), \tag{9.99}$$

where $\boldsymbol{h}(\boldsymbol{y})$ and $\boldsymbol{m}^*(\boldsymbol{\theta})$ are the n–vectors with typical elements $h_t(y_t)$ and $m_t^*(u_t^*, \boldsymbol{\theta})$, respectively. We see that the covariance matrix $\boldsymbol{\Omega}$ has two components, one due to the randomness of the data and the other due to the randomness of the simulations. If the simulator $m_t^*(\cdot)$ is the simple one suggested above, then the simulated data $h_t(y_t^*)$ are generated from the DGP characterized by $\boldsymbol{\theta}$, which is also supposed to have generated the real data. Therefore, it is clear that $\mathrm{Var}\big(\boldsymbol{h}(\boldsymbol{y})\big) = \mathrm{Var}\big(\boldsymbol{m}^*(\boldsymbol{\theta}_0)\big)$, and we conclude that $\boldsymbol{\Omega} = (1 + 1/S)\mathrm{Var}\big(\boldsymbol{h}(\boldsymbol{y})\big)$.

In general, the $n \times n$ matrix $\boldsymbol{\Omega}$ cannot be estimated consistently, but an HCCME or HAC estimator can be used to provide a consistent estimate of $\boldsymbol{\Sigma}$, the covariance matrix of $n^{-1/2}\boldsymbol{W}^\top\boldsymbol{f}^*(\boldsymbol{\theta}_0)$. For the simple simulator we have been discussing, $\hat{\boldsymbol{\Sigma}}$ is just $1+1/S$ times whatever HAC estimator or HCCME would be appropriate if there were no simulation involved. For other simulators, it may be a little harder to estimate (9.99). In any case, once $\hat{\boldsymbol{\Sigma}}$ is available, we use it to replace $n^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W}$ in (9.98). We also replace $\operatorname{plim} n^{-1}\boldsymbol{W}^\top\boldsymbol{F}_0^*$ by $\operatorname{plim} n^{-1}\boldsymbol{W}^\top\boldsymbol{M}_0$. The sandwich estimator for the asymptotic covariance matrix then simplifies greatly, and we find that the asymptotic covariance matrix is just

$$\operatorname*{plim}_{n\to\infty}\Big((n^{-1}\boldsymbol{M}_0^\top\boldsymbol{W})\hat{\boldsymbol{\Sigma}}^{-1}(n^{-1}\boldsymbol{W}^\top\boldsymbol{M}_0)\Big)^{-1}.$$

In practice, $\boldsymbol{M}_0$ can be estimated using either analytical or numerical derivatives of $(1/S)\sum_{s=1}^{S} m_t^*(u_{ts}^*, \hat{\boldsymbol{\theta}})$, evaluated at $\hat{\boldsymbol{\theta}}_{\mathrm{MSM}}$. However, for this to be a reliable estimator, it is necessary for S to be reasonably large. If we let $\hat{\boldsymbol{M}}$ denote the estimate of $\boldsymbol{M}_0$, then in practice we use

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_{\mathrm{MSM}}) = n(\hat{\boldsymbol{M}}^\top\boldsymbol{W}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}^\top\hat{\boldsymbol{M}})^{-1}. \tag{9.100}$$

Notice that (9.100) has essentially the same form as (9.41) and (9.81), the estimated covariance matrices for the feasible efficient GMM estimators of linear regression and general nonlinear models, respectively. The most important new feature of (9.100) is the factor of $1 + 1/S$, which is buried in $\hat{\boldsymbol{\Sigma}}$.

The Lognormal Distribution: An Example

Since the implementation of MSM estimation typically involves several steps and can be rather tricky, we now work through a simple example in detail. The example is in fact sufficiently simple that there is no need for simulation at all; we can work out the "right answer" directly. This provides a benchmark with which to compare the various other estimators that we consider. In order to motivate these other estimators, we demonstrate how GMM can be used to **match moments** of distributions. Moment matching can be done quite easily when the moments to be matched can be expressed analytically as functions of the parameters to be estimated, and no simulation is needed in such cases. If analytic expressions are not available, moment matching can still be done whenever we can simulate the random variables of which the expectations are the moments to be matched.

A random variable is said to follow the **lognormal distribution** if its logarithm is normally distributed. The lognormal distribution for a scalar random variable y thus depends on just two parameters, the expectation and the variance of $\log y$. Formally, if $z \sim \mathrm{N}(\mu, \sigma^2)$, then the variable $y \equiv \exp(z)$ is lognormally distributed, with a distribution characterized by μ and σ^2.

Suppose we have an n–vector $\boldsymbol{y}$, of which the components y_t are IID, each lognormally distributed with unknown parameters μ and σ^2. The "right" way to estimate these unknown parameters is to take logs of each component of $\boldsymbol{y}$, thus obtaining an n–vector $\boldsymbol{z}$ with typical element z_t, and then to estimate μ and σ^2 by the sample mean and sample variance of the z_t. This can be done by regressing $\boldsymbol{z}$ on a constant.

The above estimation method implicitly **matches** the first and second moments of the log of y_t in order to estimate the parameters. It yields the parameter values that give theoretical moments equal to the corresponding moments in the sample. Since we have two parameters to estimate, we need at least two moments. But other sets of two moments could also be used in order to obtain MM estimators of μ and σ^2. So could sets of more than two moments, although the match could not be perfect, because there would implicitly be overidentifying restrictions.

We now consider precisely how we might estimate μ and σ^2 by matching the first moment of the y_t along with the first moment of the z_t. With this choice, it is once more possible to obtain an analytical answer, because, as the result of Exercise 9.19 shows, the expectation of y_t is $\exp(\mu + \frac{1}{2}\sigma^2)$. Thus, as before, we estimate μ by using $\bar{z}$, the sample mean of the z_t, and then estimate σ^2 by solving the equation

$$\log \bar{y} = \bar{z} + \tfrac{1}{2}\hat{\sigma}^2$$

for $\hat{\sigma}^2$, where $\bar{y}$ is the sample mean of the y_t. The estimate is

$$\hat{\sigma}^2 = 2(\log \bar{y} - \bar{z}). \tag{9.101}$$

Copyright © 2003, Russell Davidson and James G. MacKinnon

This estimate is not, except by random accident, numerically equal to the estimate obtained by regressing $\boldsymbol{z}$ on a constant, and in fact it has a higher variance; see Exercises 9.20 and 9.21.

Let us formalize the estimation procedure described above in terms of zero functions and GMM. The moments used are the first moments of the y_t and the z_t, for $t = 1, \ldots, n$. For each observation, then, there are two elementary zero functions, which serve to express the expectations of the y_t and the z_t in terms of the parameters μ and σ^2. We write these elementary zero functions as follows:

$$f_{t1}(z_t, \mu, \sigma^2) = z_t - \mu; \quad f_{t2}(y_t, \mu, \sigma^2) = y_t - \exp(\mu + \tfrac{1}{2}\sigma^2). \tag{9.102}$$

The derivatives of these functions with respect to the parameters are

$$\frac{\partial f_{t1}}{\partial \mu} = -1; \quad \frac{\partial f_{t1}}{\partial \sigma^2} = 0; \quad \frac{\partial f_{t2}}{\partial \mu} = -\,e^{\mu+\sigma^2/2}; \quad \frac{\partial f_{t2}}{\partial \sigma^2} = -\,\frac{1}{2}e^{\mu+\sigma^2/2}. \tag{9.103}$$

These derivatives, which are all deterministic, allow us to find the optimal instruments for the estimation of μ and σ^2 on the basis of the zero functions (9.102), provided that we can also obtain the covariance matrix $\boldsymbol{\Omega}$ of the zero functions.

Notice that, in contrast to many GMM estimation procedures, this one involves two elementary zero functions and no instruments. Nevertheless, we can set the problem up so that it looks like a standard one. Let $\boldsymbol{f}_1(\mu, \sigma^2)$ and $\boldsymbol{f}_2(\mu, \sigma^2)$ be two n–vectors with typical components $f_{t1}(z_t, \mu, \sigma^2)$ and $f_{t2}(y_t, \mu, \sigma^2)$, respectively. For notational simplicity, we suppress the explicit dependence of these vectors on the y_t and the z_t. The $2n$–vector $\boldsymbol{f}(\mu, \sigma^2)$ of the full set of elementary zero functions, and the $2n \times 2$ matrix $\boldsymbol{F}(\mu, \sigma^2)$ of the derivatives with respect to the parameters, can thus be written as

$$\boldsymbol{f}(\mu, \sigma^2) = \begin{bmatrix} \boldsymbol{f}_1(\mu, \sigma^2) \\ \boldsymbol{f}_2(\mu, \sigma^2) \end{bmatrix} \quad \text{and} \quad \boldsymbol{F}(\mu, \sigma^2) = -\begin{bmatrix} \boldsymbol{\iota} & \mathbf{0} \\ a\boldsymbol{\iota} & \frac{1}{2}a\boldsymbol{\iota} \end{bmatrix}, \tag{9.104}$$

where $a \equiv \exp(\mu + {}^1\!/_2\sigma^2)$. The constant vectors $\boldsymbol{\iota}$ in $\boldsymbol{F}(\mu, \sigma^2)$ arise because none of the derivatives in (9.103) depends on t, which is a consequence of the assumption that the data are IID.

Because $\boldsymbol{f}(\mu, \sigma^2)$ is a $2n$–vector, the covariance matrix $\boldsymbol{\Omega}$ is $2n \times 2n$. This matrix can be written as

$$\boldsymbol{\Omega} = E\left(\begin{bmatrix} \boldsymbol{f}_{10} \\ \boldsymbol{f}_{20} \end{bmatrix} [\,\boldsymbol{f}_{10}^{\top} \quad \boldsymbol{f}_{20}^{\top}\,]\right),$$

where $\boldsymbol{f}_{i0}$, $i = 1, 2$, is $\boldsymbol{f}_i$ evaluated at the true values μ_0 and σ_0^2. Since the data are IID, $\boldsymbol{\Omega}$ can be partitioned as follows into four $n \times n$ blocks, each of which is proportional to an identity matrix. The result is

$$\boldsymbol{\Omega} = \begin{bmatrix} \sigma_z^2\mathbf{I} & \sigma_{zy}\mathbf{I} \\ \sigma_{yz}\mathbf{I} & \sigma_y^2\mathbf{I} \end{bmatrix}, \tag{9.105}$$

where the coefficients of the identity matrices are the variances and covariances $\sigma_y^2 \equiv \mathrm{Var}(y_t)$, $\sigma_z^2 \equiv \mathrm{Var}(z_t)$, and $\sigma_{yz} = \sigma_{zy} \equiv \mathrm{Cov}(y_t, z_t)$.

We now have everything we need to set up the efficient estimating equations (9.82), which, ignoring the factor of n^{-1}, become

$$\boldsymbol{F}^\top(\mu, \sigma^2)\boldsymbol{\Omega}^{-1}\boldsymbol{f}(\mu, \sigma^2) = \mathbf{0}, \tag{9.106}$$

where $\boldsymbol{f}(\cdot)$ and $\boldsymbol{F}(\cdot)$ are given by (9.104), and $\boldsymbol{\Omega}$ is given by (9.105). By explicitly performing the multiplications of partitioned matrices in (9.106), inverting $\boldsymbol{\Omega}$, and ignoring irrelevant scalar factors, we obtain

$$\begin{bmatrix} \sigma_y^2 - a\sigma_{yz} & a\sigma_z^2 - \sigma_{yz} \\ -\frac{1}{2}a\sigma_{yz} & \frac{1}{2}a\sigma_z^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\iota}^\top & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\iota}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{f}_1(\mu, \sigma^2) \\ \boldsymbol{f}_2(\mu, \sigma^2) \end{bmatrix} = \mathbf{0}.$$

Since the leftmost factor above is a 2×2 nonsingular matrix, we see that these estimating equations are equivalent to

$$\boldsymbol{\iota}^\top\boldsymbol{f}_1(\mu, \sigma^2) = \mathbf{0} \quad \text{and} \quad \boldsymbol{\iota}^\top\boldsymbol{f}_2(\mu, \sigma^2) = \mathbf{0}. \tag{9.107}$$

The solution to these two equations is $\hat{\mu} = \bar{z}$ and $\hat{\sigma}^2$ given by (9.101). Curiously, it appears that the explicit expressions for $\boldsymbol{F}(\cdot)$ and $\boldsymbol{\Omega}$ are not needed in order to formulate the estimator. They are needed, however, for the evaluation of expression (9.67) for its asymptotic covariance matrix. This is left as an exercise for the reader; in particular, the same expression for the variance of $\hat{\sigma}^2$ should be found as in the answer to Exercise 9.21.

As we mentioned above, it is possible to use more than two moments. Suppose that, in addition to matching the first moments of the z_t and the y_t, we also wish to match the second moment of the y_t, or, equivalently, the first moment of the y_t^2. Since the log of y_t^2 is just $2z_t$, which is distributed as $\mathrm{N}(2\mu, 4\sigma^2)$, the expectation of y_t^2 is $\exp\bigl(2(\mu + \sigma^2)\bigr)$. We now have three elementary zero functions for each observation, the two given in (9.102) and

$$f_{t3}(y_t, \mu, \sigma^2) = y_t^2 - \exp\bigl(2(\mu + \sigma^2)\bigr).$$

The vector $\boldsymbol{f}(\cdot)$ and the matrix $\boldsymbol{F}(\cdot)$, originally defined in (9.104), now both have $3n$ rows. The latter still has two columns, both of which can be partitioned into three n–vectors, each proportional to $\boldsymbol{\iota}$. Further, the matrix $\boldsymbol{\Omega}$ of (9.105) grows to become a $3n \times 3n$ matrix. It is then a matter of taste whether to set up a just identified estimation problem using as optimal instruments the two columns of $\boldsymbol{\Omega}^{-1}\boldsymbol{F}(\mu, \sigma^2)$, or to use three instruments, which are the columns of the matrix

$$\boldsymbol{W} \equiv \begin{bmatrix} \boldsymbol{\iota} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\iota} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\iota} \end{bmatrix}, \tag{9.108}$$

Copyright © 2003, Russell Davidson and James G. MacKinnon

and to construct an optimal weighting matrix. Whichever choice is made, it is necessary to estimate $\boldsymbol{\Omega}$ in order to construct the optimal instruments for the first method, or the optimal weighting matrix for the second.

The procedures we have just described depend on the fact that we know the analytic forms of $\mathrm{E}(z_t)$, $\mathrm{E}(y_t)$, and $\mathrm{E}(y_t^2)$. In more complicated applications, comparable analytic expressions for the moments to be matched might not be available; see Exercise 9.24 for an example. In such cases, simulators can be used to replace such analytic expressions. We illustrate the method for the case of the lognormal distribution, matching the first moments of z_t and y_t, pretending that we do not know the analytic expressions for their expectations.

For any given values of μ and σ^2, we can draw from the lognormal distribution characterized by these values by first using a random number generator to give a drawing u^* from $\mathrm{N}(0,1)$ and then computing $y^* = \exp(\mu + \sigma u^*)$. Thus unbiased simulators for the expectations of $z \equiv \log y$ and of y itself are

$$m_1^*(u^*, \mu, \sigma^2) \equiv \mu + \sigma u^* \quad \text{and} \quad m_2^*(u^*, \mu, \sigma^2) \equiv \exp(\mu + \sigma u^*).$$

If we perform S simulations, the zero functions for MSM estimation can be written as

$$\begin{aligned} f_{t1}^*(z_t, \mu, \sigma^2) &= z_t - \frac{1}{S}\sum_{s=1}^{S} m_1^*(u_{ts}^*, \mu, \sigma^2) \quad \text{and} \\ f_{t2}^*(y_t, \mu, \sigma^2) &= y_t - \frac{1}{S}\sum_{s=1}^{S} m_2^*(u_{ts}^*, \mu, \sigma^2), \end{aligned}$$

where the u_{ts}^* are IID standard normal. Comparison with (9.102) shows clearly how we replace analytic expressions for the moments, assumed to be unknown, by simulation-based estimates.

Since the data are IID, it might appear tempting to use just one set of random numbers, u_s^*, $s = 1, \ldots, S$, for all t. However, doing this would introduce dependence among the zero functions, greatly complicating the computation of their covariance matrix. As S becomes large, of course, the law of large numbers ensures that this effect becomes less and less important. Using just one set of random numbers would in any case not affect the consistency of the MSM estimator, merely that of the covariance matrix estimate.

By analogy with (9.107), we can see that the MSM estimating equations are

$$\boldsymbol{\iota}^\top \boldsymbol{f}_1^*(\hat\mu, \hat\sigma^2) = \mathbf{0} \quad \text{and} \quad \boldsymbol{\iota}^\top \boldsymbol{f}_2^*(\hat\mu, \hat\sigma^2) = \mathbf{0}. \tag{9.109}$$

Here we have again grouped the elementary zero functions into two n−vectors $\boldsymbol{f}_1^*(\cdot)$ and $\boldsymbol{f}_2^*(\cdot)$. Recalling that the random numbers u_{ts}^* are drawn *only once*

Copyright © 2003, Russell Davidson and James G. MacKinnon

for the entire procedure, let us make the definitions

$$\begin{aligned} m_{t1}(\mu,\sigma^2) &\equiv \frac{1}{S}\sum_{s=1}^{S} m_1^*(u_{ts}^*,\mu,\sigma^2) = \mu + \sigma\,\frac{1}{S}\sum_{s=1}^{S} u_{ts}^*, \quad \text{and} \\ m_{t2}(\mu,\sigma^2) &\equiv \frac{1}{S}\sum_{s=1}^{S} m_2^*(u_{ts}^*,\mu,\sigma^2) = \frac{1}{S}\sum_{s=1}^{S} \exp(\mu+\sigma u_{ts}^*). \end{aligned} \tag{9.110}$$

It is clear that, as $S \to \infty$, these functions tend for all t to the limits of the expectations of z and y, respectively. It is also not hard to see that these limits are μ and $\exp(\mu + \frac{1}{2}\sigma^2)$.

On dividing by the sample size n and rearranging, the estimating equations (9.109) can be written as

$$\bar m_1(\mu,\sigma^2) = \bar z \quad \text{and} \quad \bar m_2(\mu,\sigma^2) = \bar y, \tag{9.111}$$

where $\bar z$ and $\bar y$ are the sample averages of the z_t and the y_t, respectively, and

$$\bar m_i(\mu,\sigma^2) \equiv \frac{1}{n}\sum_{t=1}^{n} m_{ti}(\mu,\sigma^2), \quad i = 1,2.$$

Equations (9.111) can be solved in various ways. One approach is to turn the problem of solving them into a minimization problem. Let

$$\boldsymbol{W} \equiv \begin{bmatrix} \boldsymbol{\iota} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\iota} \end{bmatrix}. \tag{9.112}$$

Then it is not difficult to see that minimizing the quadratic form

$$\begin{bmatrix} \boldsymbol{z} - \boldsymbol{m}_1(\mu,\sigma^2) \\ \boldsymbol{y} - \boldsymbol{m}_2(\mu,\sigma^2) \end{bmatrix}^{\top} \boldsymbol{W}\boldsymbol{W}^{\top} \begin{bmatrix} \boldsymbol{z} - \boldsymbol{m}_1(\mu,\sigma^2) \\ \boldsymbol{y} - \boldsymbol{m}_2(\mu,\sigma^2) \end{bmatrix} \tag{9.113}$$

also solves equations (9.111); see Exercise 9.23. Here the n–vectors $\boldsymbol{m}_1(\cdot)$ and $\boldsymbol{m}_2(\cdot)$ have typical elements $m_{t1}(\cdot)$ and $m_{t2}(\cdot)$, respectively.

Alternatively, we can use Newton's Method directly. We discussed this procedure in Section 6.4, in connection with minimizing a nonlinear function, but it can also be applied to sets of equations like (9.111). Suppose that we wish to solve a set of k equations of the form $\boldsymbol{g}(\boldsymbol{\theta}) = \mathbf{0}$ for a k–vector of unknowns $\boldsymbol{\theta}$, where $\boldsymbol{g}(\cdot)$ is also a k–vector. The iterative step analogous to (6.43) is

$$\boldsymbol{\theta}_{(j+1)} = \boldsymbol{\theta}_{(j)} - \boldsymbol{G}^{-1}(\boldsymbol{\theta}_{(j)})\boldsymbol{g}(\boldsymbol{\theta}_{(j)}), \tag{9.114}$$

where $\boldsymbol{G}(\boldsymbol{\theta})$ is the **Jacobian matrix** associated with $\boldsymbol{g}(\boldsymbol{\theta})$. This $k \times k$ matrix contains the derivatives of the components of $\boldsymbol{g}(\boldsymbol{\theta})$ with respect to the elements

Copyright © 2003, Russell Davidson and James G. MacKinnon

of $\boldsymbol{\theta}$. For the estimating equations (9.111), the iterative step (9.114) becomes

$$\begin{bmatrix} \mu_{(j+1)} \\ \sigma^2_{(j+1)} \end{bmatrix} = \begin{bmatrix} \mu_{(j)} \\ \sigma^2_{(j)} \end{bmatrix} - \begin{bmatrix} \dfrac{\partial \bar{m}_1}{\partial \mu} & \dfrac{\partial \bar{m}_1}{\partial \sigma^2} \\ \dfrac{\partial \bar{m}_2}{\partial \mu} & \dfrac{\partial \bar{m}_2}{\partial \sigma^2} \end{bmatrix} \begin{bmatrix} \bar{m}_1(\mu_{(j)}, \sigma^2_{(j)}) - \bar{z} \\ \bar{m}_2(\mu_{(j)}, \sigma^2_{(j)}) - \bar{y} \end{bmatrix},$$

where all the partial derivatives are evaluated at $(\mu_{(j)}, \sigma^2_{(j)})$. It should be noted that these partial derivatives *are* known analytically, as they can be calculated directly from (9.110).

To estimate the asymptotic covariance matrix of the MSM estimates, we can use any suitable estimator of (9.81), provided we remember to multiply the result by $1 + 1/S$ in order to account for the simulation randomness. The instrument matrix $\boldsymbol{W}$ of (9.81) is just the matrix $\boldsymbol{W}$ of (9.112). We are pretending that we do not know the analytic form of the matrix $\boldsymbol{F}(\mu, \sigma^2)$ given in (9.104), and so instead we use the matrix of partial derivatives of $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$, evaluated at $\hat{\mu}$ and $\hat{\sigma}^2$. This matrix is

$$\hat{\boldsymbol{F}} \equiv \begin{bmatrix} \dfrac{\partial \boldsymbol{m}_1}{\partial \mu}(\hat{\mu}, \hat{\sigma}^2) & \dfrac{\partial \boldsymbol{m}_1}{\partial \sigma^2}(\hat{\mu}, \hat{\sigma}^2) \\ \dfrac{\partial \boldsymbol{m}_2}{\partial \mu}(\hat{\mu}, \hat{\sigma}^2) & \dfrac{\partial \boldsymbol{m}_2}{\partial \sigma^2}(\hat{\mu}, \hat{\sigma}^2) \end{bmatrix}; \tag{9.115}$$

note that each block in $\hat{\boldsymbol{F}}$ is an n−vector. If we use Newton's Method for the estimation, then all the partial derivatives in this matrix have already been computed. Finally, the covariance matrix $\boldsymbol{\Omega}$ of the elementary zero functions can be estimated using (9.105), by replacing the unknown quantities σ_z^2, σ_y^2, and σ_{zy} with their sample analogs. If we denote the result of this by $\hat{\boldsymbol{\Omega}}$, then our estimate of the covariance matrix of $\hat{\mu}$ and $\hat{\sigma}^2$ is

$$\widehat{\mathrm{Var}} \begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} = (\boldsymbol{W}^\top \hat{\boldsymbol{F}})^{-1} \boldsymbol{W}^\top \hat{\boldsymbol{\Omega}} \boldsymbol{W} (\hat{\boldsymbol{F}}^\top \boldsymbol{W})^{-1}, \tag{9.116}$$

with $\boldsymbol{W}$ given by (9.112) and $\hat{\boldsymbol{F}}$ given by (9.115).

MSM Estimation: Conclusion

Although it is very special, the example of the previous subsection illustrates most of the key features of MSM estimation. The example shows how to estimate two parameters by using two or more elementary zero functions, even when there are no genuine instruments. In econometric applications, it is more common for there to be as many elementary zero functions as there are dependent variables, just one in the case of univariate models, and for there to be more instruments than parameters. Also, in many applications, the data are not IID, but this complication generally does not require substantial changes to the methods illustrated above.

Inference in models estimated by MSM is almost always based on asymptotic theory, and it may therefore be quite unreliable in finite samples. Since MSM estimation makes sense only when a model is too intractable for less computationally demanding methods to be applicable, the cost of estimating such a model a large number of times, as would be needed to employ bootstrap methods, is likely to be prohibitive.

Not surprisingly, the literature on MSM is relatively recent. The two classic papers are McFadden (1989), who seems to have coined the name, and Pakes and Pollard (1989). Other important early papers include Lee and Ingram (1991), Keane (1994), McFadden and Ruud (1994), and Gallant and Tauchen (1996). An interesting early application of the method is Duffie and Singleton (1993). Useful references include Hajivassiliou and Ruud (1994), Gouriéroux and Monfort (1996), and van Dijk, Monfort, and Brown (1995), which is a collection of papers, both theoretical and applied.

9.7 Final Remarks

As its name implies, the generalized method of moments is a very general estimation method indeed, and numerous other methods can be thought of as special cases. These include all of the ones we have discussed so far: MM, OLS, NLS, GLS, and IV. Thus the number of techniques that can legitimately be given the label "GMM" is bewilderingly large. To avoid bewilderment, it is best not to attempt to enumerate all the possibilities, but simply to list some of the ways in which various GMM estimators differ:

- Methods for which the explanatory variables are exogenous or predetermined (including OLS, NLS, and GLS), and for which no extra instruments are required, versus methods that do require additional exogenous or predetermined instruments (including linear and nonlinear IV).
- Methods for linear models (including OLS, GLS, linear IV, and the GMM techniques discussed in Section 9.2) versus methods for nonlinear models (including NLS, GNLS, nonlinear IV, and the GMM techniques discussed in Section 9.5).
- Methods that are inefficient for a given set of moment conditions, which have sandwich covariance matrices, versus methods that are efficient for the same set of moment conditions, which do not.
- Methods that are fully efficient, because they are based on optimal instruments, versus methods that are not fully efficient.
- Methods based on a covariance matrix that is known, at least up to a finite number of parameters which can be estimated consistently, versus methods that require an HCCME or a HAC estimator. The latter can never be fully efficient.

- Methods that involve simulation, such as MSM, versus methods where the criterion function can be evaluated analytically.
- Univariate models versus multivariate models. We have not yet discussed any methods for estimating the latter, but we will do so in Chapter 12.

9.8 Exercises

9.1 Show that the difference between the matrix

$$(\boldsymbol{J}^\top\boldsymbol{W}^\top\boldsymbol{X})^{-1}\boldsymbol{J}^\top\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W}\boldsymbol{J}(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{J})^{-1}$$

and the matrix

$$(\boldsymbol{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{X})^{-1}$$

is a positive semidefinite matrix. **Hints:** Recall Exercise 3.8. Express the second of the two matrices in terms of the projection matrix $\boldsymbol{P}_{\boldsymbol{\Omega}^{1/2}\boldsymbol{W}}$, and then find a similar projection matrix for the first of them.

9.2 Let the $n-$vector $\boldsymbol{u}$ be such that $\mathrm{E}(\boldsymbol{u}) = \mathbf{0}$ and $\mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \mathbf{I}$, and let the $n \times l$ matrix $\boldsymbol{W}$ be such that $\mathrm{E}(\boldsymbol{W}_t u_t) = \mathbf{0}$ and that $\mathrm{E}(u_t u_s \,|\, \boldsymbol{W}_t, \boldsymbol{W}_s) = \delta_{ts}$, where δ_{ts} is the Kronecker delta introduced in Section 1.4. Assume that $\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{W}} \equiv \operatorname{plim} n^{-1}\boldsymbol{W}^\top\boldsymbol{W}$ is finite, deterministic, and positive definite. Explain why the quadratic form $\boldsymbol{u}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{u}$ must be asymptotically distributed as $\chi^2(l)$.

9.3 Consider the quadratic form $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x}$ is a $p \times 1$ vector and $\boldsymbol{A}$ is a $p \times p$ matrix, which may or may not be symmetric. Show that there exists a symmetric $p \times p$ matrix $\boldsymbol{B}$ such that $\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$ for all $p \times 1$ vectors $\boldsymbol{x}$, and give the explicit form of a suitable $\boldsymbol{B}$.

⋆9.4 For the model (9.01) and a specific choice of the $l \times k$ matrix $\boldsymbol{J}$, show that minimizing the quadratic form (9.12) with weighting matrix $\boldsymbol{\Lambda} = \boldsymbol{J}\boldsymbol{J}^\top$ gives the same estimator as solving the moment conditions (9.05) with the given $\boldsymbol{J}$. Assuming that these moment conditions have a unique solution for $\boldsymbol{\beta}$, show that the matrix $\boldsymbol{J}\boldsymbol{J}^\top$ is of rank k, and hence positive semidefinite without being positive definite.

Construct a symmetric, positive definite, $l \times l$ weighting matrix $\boldsymbol{\Lambda}$ such that minimizing (9.12) with this $\boldsymbol{\Lambda}$ leads once more to the same estimator as that given by solving conditions (9.05). It is convenient to take $\boldsymbol{\Lambda}$ in the form $\boldsymbol{J}\boldsymbol{J}^\top + \boldsymbol{N}\boldsymbol{N}^\top$. In the construction of $\boldsymbol{N}$, it may be useful to partition $\boldsymbol{W}$ as $[\boldsymbol{W}_1 \;\; \boldsymbol{W}_2]$, where the $n \times k$ matrix $\boldsymbol{W}_1$ is such that $\boldsymbol{W}_1^\top\boldsymbol{X}$ is nonsingular.

⋆9.5 Consider the linear regression model with serially correlated errors,

$$y_t = \beta_1 + \beta_2 x_t + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \tag{9.117}$$

where the ε_t are IID, and the autoregressive parameter ρ is assumed either to be known or to be estimated consistently. The explanatory variable x_t is assumed to be contemporaneously correlated with ε_t (see Section 8.4 for the definition of contemporaneous correlation).

Recall from Chapter 7 that the covariance matrix $\boldsymbol{\Omega}$ of the vector $\boldsymbol{u}$ with typical element u_t is given by (7.32), and that $\boldsymbol{\Omega}^{-1}$ can be expressed as $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top$,

Copyright © 2003, Russell Davidson and James G. MacKinnon

where $\boldsymbol{\Psi}$ is defined in (7.60). Express the model (9.117) in the form (9.20), without taking account of the first observation.

Let Ω_t be the information set for observation t with $\mathrm{E}(\varepsilon_t \,|\, \Omega_t) = 0$. Suppose that there exists a matrix $\boldsymbol{Z}$ of instrumental variables, with $\boldsymbol{Z}_t \in \Omega_t$, such that the explanatory vector $\boldsymbol{x}$ with typical element x_t is related to the instruments by the equation

$$\boldsymbol{x} = \boldsymbol{Z\pi} + \boldsymbol{v}, \tag{9.118}$$

where $\mathrm{E}(v_t \,|\, \Omega_t) = 0$. Derive the explicit form of the expression $(\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}})_t$ defined implicitly by equation (9.24) for the model (9.117). Find a matrix $\boldsymbol{W}$ of instruments that satisfy the predeterminedness condition in the form (9.30) and that lead to asymptotically efficient estimates of the parameters β_1 and β_2 computed on the basis of the theoretical moment conditions (9.31) with your choice of $\boldsymbol{W}$.

$\star$9.6 Consider the model (9.20), where the matrix $\boldsymbol{\Psi}$ is chosen in such a way that the transformed error terms, the $(\boldsymbol{\Psi}^\top\boldsymbol{u})_t$, are innovations with respect to the information sets Ω_t. In other words, $\mathrm{E}\big((\boldsymbol{\Psi}^\top\boldsymbol{u})_t \,|\, \Omega_t\big) = 0$. Suppose that the $n \times l$ matrix of instruments $\boldsymbol{W}$ is predetermined in the usual sense that $\boldsymbol{W}_t \in \Omega_t$. Show that these assumptions, along with the assumption that $\mathrm{E}\big((\boldsymbol{\Psi}^\top\boldsymbol{u})_t^2 \,|\, \Omega_t\big) = \mathrm{E}\big((\boldsymbol{\Psi}^\top\boldsymbol{u})_t^2\big) = 1$ for $t = 1, \ldots, n$, are enough to prove the analog of (9.02), that is, that

$$\mathrm{Var}(n^{-1/2}\boldsymbol{W}^\top\boldsymbol{\Psi}^\top\boldsymbol{u}) = n^{-1}\mathrm{E}(\boldsymbol{W}^\top\boldsymbol{W}).$$

In order to perform just-identified estimation, let the $n \times k$ matrix $\boldsymbol{Z} = \boldsymbol{WJ}$, for an $l \times k$ matrix $\boldsymbol{J}$ of full column rank. Compute the asymptotic covariance matrix of the estimator obtained by solving the moment conditions

$$\boldsymbol{Z}^\top\boldsymbol{\Psi}^\top(\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{J}^\top\boldsymbol{W}^\top\boldsymbol{\Psi}^\top(\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{0}. \tag{9.119}$$

The covariance matrix you have found should be a sandwich. Find the choice of $\boldsymbol{J}$ that eliminates the sandwich, and show that this choice leads to an asymptotic covariance matrix that is smaller, in the usual sense, than the asymptotic covariance matrix for any other choice of $\boldsymbol{J}$.

Compute the GMM criterion function for model (9.20) with instruments $\boldsymbol{W}$, and show that the estimator found by minimizing this criterion function is just the estimator obtained using the optimal choice of $\boldsymbol{J}$.

9.7 Compare the asymptotic covariance matrix found in the preceding question for the estimator of the parameters of model (9.20), obtained by minimizing the GMM criterion function for the $n \times l$ matrix of predetermined instruments $\boldsymbol{W}$, with the covariance matrix (9.29) that corresponds to estimation with instruments $\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}}$. In particular, show that the difference between the two is a positive semidefinite matrix.

9.8 Consider overidentified estimation based on the moment conditions

$$\mathrm{E}(\boldsymbol{W}^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})) = \boldsymbol{0},$$

which were given in (9.31), where the $n \times l$ matrix of instruments $\boldsymbol{W}$ satisfies the predeterminedness condition (9.30). Derive the GMM criterion function for these theoretical moment conditions, and show that the estimating equations that result from the minimization of this criterion function are

$$\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})=\boldsymbol{0}. \tag{9.120}$$

Suppose that $\mathcal{S}(\bar{\boldsymbol{X}})$, the span of the $n \times k$ matrix $\bar{\boldsymbol{X}}$ of optimal instruments defined by (9.24), is a linear subspace of $\mathcal{S}(\boldsymbol{W})$, the span of the transformed instruments. Show that, in this case, the estimating equations (9.120) are asymptotically equivalent to

$$\bar{\boldsymbol{X}}^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})=\boldsymbol{0},$$

of which the solution is the efficient estimator $\hat{\boldsymbol{\beta}}_{\rm EGMM}$ defined in (9.26).

9.9 Show that the asymptotic covariance matrix of the estimator obtained by solving the estimating equations (9.120) is

$$\operatorname*{plim}_{n\to\infty}\Big(\frac{1}{n}\bar{\boldsymbol{X}}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}\Big)^{-1}. \tag{9.121}$$

By expressing this asymptotic covariance matrix in terms of a matrix $\boldsymbol{\Psi}$ that satisfies the equation $\boldsymbol{\Omega}^{-1}=\boldsymbol{\Psi\Psi}^\top$, show that the difference between it and the asymptotic covariance matrix of the efficient estimator $\hat{\boldsymbol{\beta}}_{\rm EGMM}$ of (9.26) is a positive semidefinite matrix.

⋆9.10 Give the explicit form of the $n \times n$ matrix $\boldsymbol{U}(j)$ for which $\hat{\boldsymbol{\Gamma}}(j)$, defined in (9.36), takes the form $n^{-1}\boldsymbol{W}^\top\boldsymbol{U}(j)\boldsymbol{W}$.

9.11 This question uses data on daily returns for the period 1989–1998 from the file **daily-crsp.data**. These data are made available by courtesy of the Center for Research in Security Prices (CRSP); see the comments at the bottom of the file. Let r_t denote the daily return on shares of Mobil Corporation, and let v_t denote the daily return for the CRSP value-weighted index. Using all but the first four observations (to allow for lags), run the regression

$$r_t=\beta_1+\beta_2 v_t+u_t$$

by OLS. Report three different sets of standard errors: the usual OLS ones, ones based on the simplest HCCME, and ones based on a more advanced HCCME that corrects for the downward bias in the squared OLS residuals; see Section 5.5. Do the OLS standard errors appear to be reliable?

Assuming that the u_t are heteroskedastic but serially uncorrelated, obtain estimates of the β_i that are more efficient than the OLS ones. For this purpose, use r_{t-1}^2, v_t^2, v_{t-1}^2, and v_{t-2}^2 as additional instruments. Do these estimates appear to be more efficient than the OLS ones?

9.12 Using the data for consumption (C_t) and disposable income (Y_t) contained in the file **consumption.data**, construct the variables $c_t=\log C_t$, $\Delta c_t=c_t-c_{t-1}$, $y_t=\log Y_t$, and $\Delta y_t=y_t-y_{t-1}$. Then, for the period 1953:1 to 1996:4, run the regression

$$\Delta c_t=\beta_1+\beta_2\Delta y_t+\beta_3\Delta y_{t-1}+u_t \tag{9.122}$$

by OLS, and test the hypothesis that the u_t are serially uncorrelated against the alternative that they follow an AR(1) process.

Calculate eight sets of HAC estimates of the standard errors of the OLS parameter estimates from regression (9.122), using the Newey-West estimator with the lag truncation parameter set to the values $p = 1, 2, 3, 4, 5, 6, 7, 8$.

9.13 Using the squares of Δy_t, Δy_{t-1}, and Δc_{t-1} as additional instruments, obtain feasible efficient GMM estimates of the parameters of (9.122) by minimizing the criterion function (9.42), with $\hat{\boldsymbol{\Sigma}}$ given by the HAC estimators computed in the previous exercise. For $p = 6$, carry out the iterative procedure described in Section 9.3 by which new parameter estimates are used to update the HAC estimator, which is then used to update the parameter estimates. **Warning:** It may be necessary to rescale the instruments so as to avoid numerical problems.

9.14 Suppose that $f_t = y_t - \boldsymbol{X}_t\boldsymbol{\beta}$. Show that, in this special case, the estimating equations (9.77) yield the generalized IV estimator.

9.15 Starting from the asymptotic covariance matrix (9.67), show that, when $\boldsymbol{\Omega}^{-1}\boldsymbol{F}_0$ is used in place of $\boldsymbol{Z}$, the covariance matrix of the resulting estimator is given by (9.83). Then show that, for the linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ with exogenous explanatory variables $\boldsymbol{X}$, this estimator is the GLS estimator.

⋆9.16 The minimization of the GMM criterion function (9.87) yields the estimating equations (9.89) with $\boldsymbol{A} = \boldsymbol{\Psi}^\top\boldsymbol{W}$. Assuming that the $n \times l$ instrument matrix $\boldsymbol{W}$ satisfies the predeterminedness condition in the form (9.30), show that these estimating equations are asymptotically equivalent to the equations

$$\bar{\boldsymbol{F}}_0^\top\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{\Psi}^\top\boldsymbol{W}}\boldsymbol{\Psi}^\top\boldsymbol{f}(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}, \tag{9.123}$$

where, as usual, $\bar{\boldsymbol{F}}_0 \equiv \bar{\boldsymbol{F}}(\boldsymbol{\theta}_0)$, with $\boldsymbol{\theta}_0$ the true parameter vector. Next, derive the asymptotic covariance matrix of the estimator defined by these equations.

Show that the equations (9.123) are the optimal estimating equations for overidentified estimation based on the transformed zero functions $\boldsymbol{\Psi}^\top\boldsymbol{f}(\boldsymbol{\theta})$ and the transformed instruments $\boldsymbol{\Psi}^\top\boldsymbol{W}$. Show further that, if the condition $\mathcal{S}(\bar{\boldsymbol{F}}) \subseteq \mathcal{S}(\boldsymbol{W})$ is satisfied, the asymptotic covariance matrix of the estimator obtained by solving equations (9.123) coincides with the optimal asymptotic covariance matrix (9.83).

⋆9.17 Suppose the n–vector $\boldsymbol{f}(\boldsymbol{\theta})$ of elementary zero functions has a covariance matrix $\sigma^2\mathbf{I}$. Show that, if the instrumental variables used for GMM estimation are the columns of the $n \times l$ matrix $\boldsymbol{W}$, the GMM criterion function is

$$\frac{1}{\sigma^2}\boldsymbol{f}^\top(\boldsymbol{\theta})\boldsymbol{P_W}\boldsymbol{f}(\boldsymbol{\theta}). \tag{9.124}$$

Next, show that, whenever the instruments are predetermined, the artificial regression

$$\boldsymbol{f}(\boldsymbol{\theta}) = -\boldsymbol{P_W}\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{b} + \text{residuals}, \tag{9.125}$$

where $\boldsymbol{F}(\boldsymbol{\theta})$ is defined as usual by (9.63), satisfies all the requisite properties for hypothesis testing. These properties, which are spelled out in detail in Exercise 8.20 in the context of the IVGNR, are that the regressand should be orthogonal to the regressors when they are evaluated at the GMM estimator obtained by minimizing (9.124); that the OLS covariance matrix from (9.125)

Copyright © 2003, Russell Davidson and James G. MacKinnon

should be a consistent estimate of the asymptotic variance of that estimator; and that (9.125) should admit one-step estimation.

⋆9.18 Derive a heteroskedasticity robust version of the artificial regression (9.125), assuming that the covariance matrix of the vector $\boldsymbol{f}(\boldsymbol{\theta})$ of zero functions is diagonal, but otherwise arbitrary.

⋆9.19 If the scalar random variable z is distributed according to the $\mathrm{N}(\mu, \sigma^2)$ distribution, show that

$$\mathrm{E}(e^z) = \exp(\mu + \tfrac{1}{2}\sigma^2).$$

⋆9.20 Let the components z_t of the n−vector $\boldsymbol{z}$ be IID drawings from the $\mathrm{N}(\mu, \sigma^2)$ distribution, and let s^2 be the OLS estimate of the error variance from the regression of $\boldsymbol{z}$ on the constant vector $\boldsymbol{\iota}$. Show that the variance of s^2 is $2\sigma^4/(n-1)$.

Would this result still hold if the normality assumption were dropped? Without this assumption, what would you need to know about the distribution of the z_t in order to find the variance of s^2?

⋆9.21 Using the delta method, obtain an expression for the asymptotic variance of the estimator defined by (9.101) for the variance of the normal distribution underlying a lognormal distribution. Show that this asymptotic variance is greater than that of the sample variance of the normal variables themselves.

⋆9.22 Describe the two procedures by which the parameters μ and σ^2 of the lognormal distribution can be estimated by the method of simulated moments, matching the first and second moments of the lognormal variable itself, and the first moment of its log. The first procedure should use optimal instruments and be just identified; the second should use the simple instruments of (9.108) and be overidentified.

9.23 Show that minimizing the criterion function (9.113), when $\boldsymbol{W}$ is defined in (9.112), is equivalent to solving equations (9.111). Then show that it is also equivalent to minimizing the criterion function

$$\begin{bmatrix} \boldsymbol{z} - \boldsymbol{m}_1(\mu, \sigma^2) \\ \boldsymbol{y} - \boldsymbol{m}_2(\mu, \sigma^2) \end{bmatrix}^{\top} \boldsymbol{W}(\boldsymbol{W}^{\top}\boldsymbol{W})^{-1}\boldsymbol{W}^{\top} \begin{bmatrix} \boldsymbol{z} - \boldsymbol{m}_1(\mu, \sigma^2) \\ \boldsymbol{y} - \boldsymbol{m}_2(\mu, \sigma^2) \end{bmatrix}, \tag{9.126}$$

which is the criterion function for nonlinear IV estimation.

⋆9.24 The **Singh-Maddala** distribution is a three-parameter distribution which has been shown to give an acceptable account, up to scale, of the distributions of household income in many countries. It is characterized by the following CDF:

$$F(y) = 1 - \frac{1}{(1 + ay^b)^c}, \quad y > 0,\, a > 0,\, b > 0,\, c > 0. \tag{9.127}$$

Suppose that you have at your disposal the values of the incomes of a random sample of households from a given population. Describe in detail how to use this sample in order to estimate the parameters a, b, and c of (9.127) by the method of simulated moments, basing the estimates on the expectations of y, $\log y$, and $y \log y$. Describe how to construct a consistent estimate of the asymptotic covariance matrix of your estimator.