6 Estimation

6.1 Introduction

Macroeconometric models are typically nonlinear, simultaneous, and large. They also tend to have error terms that are serially correlated. The focus of this chapter is on models with these characteristics. The notation that will be used in this chapter and in Chapters 7–10 is as follows. Write the model as

$$f_i(y_t, x_t, \alpha_i) = u_{it}, \qquad i = 1, \ldots, n, \qquad t = 1, \ldots, T, \tag{6.1}$$

where y_t is an n-dimensional vector of endogenous variables, x_t is a vector of predetermined variables, α_i is a vector of unknown coefficients, and u_{it} is an error term. Assume that the first m equations are stochastic, with the remaining u_{it} ($i = m + 1, \ldots, n$) identically zero for all t.

Let J_t be the $n \times n$ Jacobian matrix whose ij element is $\partial f_i/\partial y_{jt}(i, j = 1, \ldots, n)$. Also, let u_i be the T-dimensional vector $(u_{i1}, \ldots, u_{iT})'$, and let u be the $m \cdot T$-dimensional vector $(u_{11}, \ldots, u_{1T}, \ldots, u_{m1}, \ldots, u_{mT})'$. Let α denote the k-dimensional vector $(\alpha_1', \ldots, \alpha_m')$ of all the unknown coefficients. Finally, let G_i' be the $k_i \times T$ matrix whose tth column is $\partial f_i(y_t, x_t, \alpha_i)/\partial \alpha_i$, where k_i is the dimension of α_i, and let G' be the $k \times m \cdot T$ matrix,

$$\begin{bmatrix} G_1' & 0 & \cdots & 0 \\ 0 & G_2' & & \\ \cdot & & \cdot & \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & & & G_m' \end{bmatrix}$$

where $k = \sum_{i=1}^{m} k_i$. These vectors and matrices will be used in the following sections.

6.2 Treatment of Serial Correlation

A convenient way of dealing with serially correlated error terms is to treat the serial correlation coefficients as structural coefficients and to transform the

equations into equations with serially uncorrelated error terms. This introduces nonlinear restrictions on the coefficients, but otherwise the equations are like any others with serially uncorrelated errors. It will be useful to consider this transformation first because once it has been done, little more needs to be said about serial correlation. Consider the ith equation of (6.1), and assume that u_{it} is first-order serially correlated:

$$u_{it} = \rho_i u_{it-1} + \epsilon_{it}, \qquad t = 2, \ldots, T, \tag{6.2}$$

where ϵ_{it} is not serially correlated. Lagging (6.1) one period, multiplying through by ρ_i, and subtracting the resulting expression from (6.1) yields

$$f_i(y_t, x_t, \alpha_i) - \rho_i f_i(y_{t-1}, x_{t-1}, \alpha_i) = u_{it} - \rho_i u_{it-1} = \epsilon_{it}, \qquad t = 2, \ldots, T, \tag{6.3}$$

or

$$f_i^*(y_t, x_t^*, \alpha_i^*) = \epsilon_{it}, \qquad t = 2, \ldots, T, \tag{6.4}$$

where x_t^* includes the variables in x_t, x_{t-1}, and y_{t-1}, and α_i^* includes both α_i and ρ_i. Equation (6.4) is no more general than (6.1), and thus one can deal directly with (6.1) under the assumption that serial correlation has been eliminated through transformation.

This procedure results in the "loss" of the first observation. This has no effect on the asymptotic properties of the estimators, and it is probably not a problem about which one needs to be concerned in practice. In many cases there are ways of using the first observation more efficiently, but at a considerable cost in complexity relative to the approach just presented.

This procedure can handle serial correlation of higher orders. If, for example, u_{it} is second-order serially correlated:

$$u_{it} = \rho_{1i} u_{it-1} + \rho_{2i} u_{it-2} + \epsilon_{it}, \qquad t = 3, \ldots, T, \tag{6.2}'$$

the transformation in (6.3) is:

$$f_i(y_t, x_t, \alpha_i) - \rho_{1i} f_i(y_{t-1}, x_{t-1}, \alpha_i) - \rho_{2i} f_i(y_{t-2}, x_{t-2}, \alpha_i) = \epsilon_{it}, \qquad t = 3, \ldots, T. \tag{6.3}'$$

In this case x_t^* in (6.4) includes the variables in x_t, x_{t-1}, x_{t-2}, y_{t-1}, and y_{t-2}, and α_i^* includes α_i, ρ_{1i}, and ρ_{2i}. Each additional order of the serial correlation process results in the "loss" of one more observation.

With respect to testing for serial correlation, it is well known that the Durbin-Watson (DW) test is biased toward accepting the null hypothesis of no serial correlation if there is a lagged dependent variable in the equation.

Since many equations in macroeconometric models have lagged dependent variables, the DW test is of limited use. My response to this problem is to estimate the equations initially under the assumption of serial correlation (usually first-order) by some consistent technique (usually 2SLS). From this, one can test the hypothesis that the serial correlation coefficients are zero, which is simply a t-test on each coefficient. This test is valid asymptotically if one has correctly estimated the asymptotic covariance matrix of the estimated coefficients, and it is not restricted to equations without lagged dependent variables. It also easily handles serial correlation of higher than first order, since all this requires is estimating the equation under the assumption of the particular order. If a test indicates that a serial correlation coefficient is zero, the equation can be reestimated without this coefficient being included.

Although this is the general procedure that I follow in handling serial correlation problems, I still include the DW statistic in the presentation of the results for a particular equation (see Chapter 4). Since the DW statistic is biased toward acceptance of the hypothesis of no serial correlation when there are lagged dependent variables, a value that rejects the hypothesis indicates that there are likely to be problems. The DW test is thus useful for testing in one direction, and this is the reason I tend to include it in the results.

6.3 Estimation Techniques

6.3.1 Ordinary Least Squares (OLS)

The OLS technique is a special case of the 2SLS technique, where D_i in (6.5) and (6.6) below is the identity matrix. It is thus unnecessary to consider this technique separately from the 2SLS technique.

6.3.2 Two-Stage Least Squares (2SLS)

General Case

2SLS estimates of α_i (say $\hat{\alpha}_i$) are obtained by minimizing

$$u_i'Z_i(Z_i'Z_i)^{-1}Z_i'u_i = u_i'D_iu_i \tag{6.5}$$

with respect to α_i, where Z_i is a $T \times K_i$ matrix of predetermined variables. Z_i and K_i can differ from equation to equation. An estimate of the covariance matrix of $\hat{\alpha}_i$ (say $\hat{V}_{2ii}$) is

$$\hat{V}_{2ii} = \hat{\sigma}_{ii}(\hat{G}_i'D_i\hat{G}_i)^{-1}, \tag{6.6}$$

where $\hat{G}_i$ is G_i evaluated at $\hat{\alpha}_i$ and $\hat{\sigma}_{ii} = T^{-1}\sum_{t=1}^{T}\hat{u}_{it}^2$, $\hat{u}_{it} = f_i(y_t, x_t, \hat{\alpha}_i)$.

The 2SLS estimator in this form is presented in Amemiya (1974). It handles the case of nonlinearity in both variables and coefficients. In earlier work, Kelejian (1971) considered the case of nonlinearity in variables only. Bierens (1981, p. 106) has pointed out that Amemiya's proof of consistency of this estimator is valid only in the case of linearity in the coefficients, that is, only in Kelejian's case. Bierens supplies a proof of consistency and asymptotic normality in the general case.

Linear-in-Coefficients Case

It will be useful to consider the special case in which the equation to be estimated is linear in coefficients. Write equation i in this case as

$$y_i = X_i\alpha_i + u_i, \tag{6.7}$$

where y_i is the T-dimensional vector $(y_{i1}, \ldots, y_{iT})'$ and X_i is a $T \times k_i$ matrix of observations on the explanatory variables in the equation. X_i includes both endogenous and predetermined variables. Both y_i and the variables in X_i can be nonlinear functions of other variables, and thus (6.7) is much more general than the standard linear model. All that is required is that the equation be linear in α_i. Substituting $u_i = y_i - X_i\alpha_i$ into (6.5), differentiating with respect to α_i, and setting the derivatives equal to zero yields the following formula for $\hat{\alpha}_i$:

$$\hat{\alpha}_i = (X_i'D_iX_i)^{-1}X_i'D_iy_i = (\hat{X}_i'X_i)^{-1}\hat{X}_i'y_i, \tag{6.8}$$

where $\hat{X}_i = D_iX_i$ is the matrix of predicted values of the regression of X_i on Z_i. Since $D_i' = D_i$ and $D_iD_i = D_i$, $\hat{X}_i'\hat{X}_i = \hat{X}_i'D_iD_iX_i = \hat{X}_i'D_iX_i = \hat{X}_i'X_i$, and thus (6.8) can be written

$$\hat{\alpha}_i = (\hat{X}_i'\hat{X}_i)^{-1}\hat{X}_i'y_i, \tag{6.9}$$

which is the standard 2SLS formula in the linear-in-coefficients case. In this case G_i' is simply X_i', and the formula (6.6) for $\hat{V}_{2ii}$ reduces to

$$\hat{V}_{2ii} = \hat{\sigma}_{ii}(\hat{X}_i'\hat{X}_i)^{-1}. \tag{6.10}$$

Linear-in-Coefficients Case with Serial Correlation

It will also be useful to consider the linear-in-coefficients case with serially correlated errors. Assume that u_i in (6.7) is first-order serially correlated:

$$u_i = u_{i-1}\rho_i + \epsilon_i. \tag{6.11}$$

Transforming (6.7) in the manner discussed above yields

$$(6.12) \qquad y_i - y_{i-1}\rho_i = (X_i - X_{i-1}\rho_i)\alpha_i + \epsilon_i.$$

Minimizing $\epsilon_i' D_i \epsilon_i$ with respect to α_i and ρ_i results in the following first-order conditions:

$$(6.13) \qquad \hat{\alpha}_i = [\widehat{(X_i - X_{i-1}\hat{\rho}_i)}'(X_i - X_{i-1}\hat{\rho}_i)]^{-1}\widehat{(X_i - X_{i-1}\hat{\rho}_i)}'(y_i - y_{i-1}\hat{\rho}_i),$$

$$(6.14) \qquad \hat{\rho}_i = \frac{(\hat{y}_{i-1} - \hat{X}_{i-1}\hat{\alpha}_i)'(y_i - X_i\hat{\alpha}_i)}{(\hat{y}_{i-1} - \hat{X}_{i-1}\hat{\alpha}_i)'(y_{i-1} - X_{i-1}\hat{\alpha}_i)},$$

where $\widehat{X_i - X_{i-1}\hat{\rho}_i} = D_i(X_i - X_{i-1}\hat{\rho}_i)$, $\hat{y}_{i-1} = D_i y_{i-1}$, and $\hat{X}_{i-1} = D_i X_{i-1}$. If X_{i-1} is included in Z_i, then $\hat{X}_{i-1} = X_{i-1}$ (since $\hat{X}_{i-1}$ is merely the predicted values from a regression of X_{i-1} on itself and other variables), and therefore $\widehat{X_i - X_{i-1}\hat{\rho}_i} = \hat{X}_i - X_{i-1}\hat{\rho}_i$. If in addition y_{i-1} is included in Z_i, then $\hat{y}_{i-1} = y_{i-1}$, and (6.14) becomes

$$(6.14)' \qquad \hat{\rho}_i = \frac{\hat{u}_{i-1}'\hat{u}_i}{\hat{u}_{i-1}'\hat{u}_{i-1}},$$

where $\hat{u}_{i-1} = y_{i-1} - X_{i-1}\hat{\alpha}_i$ and $\hat{u}_i = y_i - X_i\hat{\alpha}_i$. This is merely the formula for the coefficient estimate of the regression of $\hat{u}_i$ on $\hat{u}_{i-1}$.

Equations (6.13) and (6.14) can easily be solved iteratively. Given an initial guess for $\hat{\rho}_i$, $\hat{\alpha}_i$ can be computed from (6.13), and then given $\hat{\alpha}_i$, $\hat{\rho}_i$ can be computed from (6.14). Given this new value of $\hat{\rho}_i$, a new value of $\hat{\alpha}_i$ can be computed from (6.13), and so on. If convergence is reached, which means that the values of $\hat{\alpha}_i$ and $\hat{\rho}_i$ on successive iterations are within some prescribed tolerance level, the first-order conditions have been solved.

Equations with RHS endogenous variables and serially correlated errors (that is, Eqs. 6.7 and 6.11) occur frequently in practice, and the 2SLS estimator for this case has been widely used. This estimator was discussed in Fair (1970), and I programmed it into the TSP regression package in 1968 under the name TSCORC. ("CORC" refers to the fact that the iterative procedure used to solve Eqs. 6.13 and 6.14 is like the Cochrane-Orcutt [1949] iterative procedure in the nonsimultaneous equations case.) There is an important difference between (6.13) and the formula for $\hat{\alpha}_i$ proposed in Fair (1970), and given the widespread use of the TSCORC command, this difference should be noted. Let $X_i = (Y_i \; X_{2i})$, where Y_i is the matrix of RHS endogenous variables in (6.7) and X_{2i} is the matrix of predetermined variables. Let $\hat{Y}_i = D_i Y_i$ and $\hat{\hat{X}}_i = (\hat{Y}_i \; X_{2i})$. The formula proposed for $\hat{\alpha}_i$ was

$$(6.13)' \qquad \hat{\alpha}_i = [(\hat{\hat{X}}_i - X_{i-1}\hat{\rho}_i)'(\hat{\hat{X}}_i - X_{i-1}\hat{\rho}_i)]^{-1}(\hat{\hat{X}}_i - X_{i-1}\hat{\rho}_i)'(y_i - y_{i-1}\hat{\rho}_i).$$

This is the formula for the coefficient estimates of the regression of $y_i - y_{i-1}\hat{\rho}_i$ on $\hat{X}_i - X_{i-1}\hat{\rho}_i$. Equation (6.13) reduces to (6.13)′ when X_{2i} and $X_{i-1} = (Y_{i-1}\ X_{2i-1})$ are included in Z_i, that is, when the exogenous, lagged endogenous, and lagged exogenous variables in the equation being estimated are included among the first-stage regressors. The inclusion of X_{2i} means that $\hat{\hat{X}}_i = \hat{X}_i$, and, as noted earlier, the inclusion of X_{i-1} means that $\widehat{X_i - X_{i-1}\hat{\rho}_i} = \hat{X}_i - X_{i-1}\hat{\rho}_i$. The proposed formula for $\hat{\rho}_i$ was (6.14)′, which, as noted above, is the same as (6.14) only if X_{i-1} and y_{i-1} are included in Z_i. Solving (6.13)′ and (6.14)′ is thus not the same as solving (6.13) and (6.14) unless X_{2i}, X_{i-1}, and y_{i-1} are included in Z_i. It can be shown that if this is not done, solving (6.13)′ and (6.14)′ does not result in consistent estimates. The need to include X_{2i}, X_{i-1}, and y_{i-1} among the first-stage regressors was stressed in Fair (1970), but one should keep in mind that this is not absolutely necessary if the formulas (6.13) and (6.14) are used. In general, however, X_{2i}, X_{i-1}, and y_{i-1} are obvious variables to include among the first-stage regressors, and for most problems this should probably be done even if one is using a program that solves (6.13) and (6.14) rather than (6.13)′ and (6.14)′.

In the case of linearity in the coefficients and first-order serial correlation, $G_i = (X_i - X_{i-1}\rho_i \quad y_{i-1} - X_{i-1}\alpha_i)$, and the formula (6.6) for $\hat{V}_{2ii}$ can be written

$$(6.15) \qquad \hat{V}_{2ii} = \hat{\sigma}_{ii}\begin{bmatrix} (\hat{X}_i - \hat{X}_{i-1}\hat{\rho}_i)'(\hat{X}_i - \hat{X}_{i-1}\hat{\rho}_i) & (\hat{X}_i - \hat{X}_{i-1}\hat{\rho}_i)'(\hat{y}_{i-1} - \hat{X}_{i-1}\hat{\rho}_i) \\ (\hat{y}_{i-1} - \hat{X}_{i-1}\hat{\alpha}_i)'(\hat{X}_i - \hat{X}_{i-1}\hat{\rho}_i) & (\hat{y}_{i-1} - \hat{X}_{i-1}\hat{\alpha}_i)'(\hat{y}_{i-1} - \hat{X}_{i-1}\hat{\alpha}_i) \end{bmatrix}^{-1}.$$

If X_{2i}, X_{i-1}, and y_{i-1} are included in Z_i, then (6.15) becomes

$$(6.15)' \qquad \hat{V}_{2ii} = \hat{\sigma}_{ii}\begin{bmatrix} (\hat{\hat{X}}_i - X_{i-1}\hat{\rho}_i)'(\hat{\hat{X}}_i - X_{i-1}\hat{\rho}_i) & (\hat{\hat{X}}_i - X_{i-1}\hat{\rho}_i)'\hat{u}_{i-1} \\ \hat{u}'_{i-1}(\hat{\hat{X}}_i - X_{i-1}\hat{\rho}_i) & \hat{u}'_{i-1}\hat{u}_{i-1} \end{bmatrix}^{-1},$$

where, as above, $\hat{u}_{i-1} = y_{i-1} - X_{i-1}\hat{\alpha}_i$. This is the formula presented in Fair (1970). Remember that $\hat{V}_{2ii}$ in this case is the covariance matrix for $(\hat{\alpha}_i \quad \hat{\rho}_i)$, not $\hat{\alpha}_i$ alone. It was suggested in Fair (1970, p. 514) that the off-diagonal terms in (6.15)′ be ignored (that is, set to zero) when computing $\hat{V}_{2ii}$, and this was initially done for the TSCORC option in TSP. This is not, however, a good idea, as Fisher, Cootner, and Baily (1972, p. 575, n. 6) first pointed out. The saving in computational costs from ignoring the off-diagonal terms is small, and in general one should not ignore the correlation between $\hat{\alpha}_i$ and $\hat{\rho}_i$ in

computing $\hat{V}_{2ii}$. In later versions of TSP the TSCORC option was changed to compute $\hat{V}_{2ii}$ according to (6.15)′, but many copies were distributed before this change was made.

The generalization of the preceding discussion to higher-order serial correlation is straightforward, and this will not be done here except to make one point. As the order of the serial correlation increases, the number of variables that must be included among the first-stage regressors to ensure consistent estimates increases if the higher-order equivalents of (6.13)′ and (6.14)′ are used. In going from first to second, for example, the new variables that must be included are X_{i-2} and y_{i-2}. At some point it may not be sensible, given the number of observations, to include all these variables, in which case the higher-order equivalents of (6.13) and (6.14) should be used for the estimates.

Restrictions on the Coefficients

In the general nonlinear case in which (6.5) is minimized using an algorithm like DFP, restrictions on the coefficients are easy to handle. Minimization is merely over the set of unrestricted coefficients. For each set of unrestricted coefficients tried by the algorithm, the restricted coefficients are first calculated and then the objective function (6.5) is computed. Except for calculating the restricted coefficients given the unrestricted ones, no extra work is involved in accounting for the restrictions.

In the case in which the restrictions are linear and the model is otherwise only nonlinear in variables, an alternative procedure is available for handling the restrictions. To see this, assume that a restriction is

$$R\alpha_i = r, \tag{6.16}$$

where R is $1 \times k_i$, α_i is $k_i \times 1$, and r is a scalar. R and r are assumed to be known. Let α_{1i} denote the first element of α_i, and assume without loss of generality that the first element of R is nonzero. Given this assumption, (6.16) can be solved for α_{1i}:

$$\alpha_{1i} = R^*\alpha_i^* + r^*, \tag{6.17}$$

where R^* is $1 \times k_i - 1$ and α_i^* is $k_i - 1 \times 1$. The vector α_i^* excludes α_{1i}.

Given (6.17), (6.7) can be written

$$y_i = X_{1i}\alpha_{1i} + X_{2i}\alpha_i^* + u_i = X_{1i}(R^*\alpha_i^* + r^*) + X_{2i}\alpha_i^* + u_i \tag{6.18}$$

or

$$y_i^* = X_i^* \alpha_i^* + u_i, \tag{6.19}$$

where $y_i^* = y_i - X_{1i} r^*$ and $X_i^* = X_{1i} R^* + X_{2i}$. The vector X_{1i} is a $T \times 1$ vector of observations on the variable corresponding to α_{1i}, and X_{2i} is a $T \times k_i - 1$ matrix of observations on the other explanatory variables. Given that R^* and r^* are known, y_i^* and X_i^* are known, and therefore (6.19) can be estimated in the usual way. The original equation has been transformed into one that is linear in the unrestricted coefficients. The extra work in this case is merely to create the transformed variables.

The coefficient restriction in the US model that is represented by (4.20) is a linear restriction on the coefficients of the wage equation (γ_1, γ_2, and γ_3) if the coefficients of the price equation (β_1 and β_2) are given. For all the limited information estimation techniques (that is, all the techniques except 3SLS and FIML), the variables in the wage equation were transformed into an equation like (6.19) before estimation. This required that the price equation be estimated first to get the estimates of β_1 and β_2 to be used in the transformation. This procedure was not followed for the 3SLS and FIML estimates, since the restriction (4.20) is not linear within the context of all the equations of the model.

Choice of First-Stage Regressors

Before estimating an equation by 2SLS, the first-stage regressors (FSRs) must be chosen. Since analytic expressions for the reduced form equations are not available for most nonlinear models, they cannot be used to guide the choice of FSRs. One must choose, given knowledge of the model, FSRs that seem likely to be important explanatory variables in the (unknown) reduced form equations for the RHS endogenous variables in the equation being estimated.

There is considerable judgment involved in the choice of FSRs for a particular equation, and there are only a few rules of thumb that can be given. Consider estimating an equation with y_{2t} and y_{3t} as RHS endogenous variables. Assume that the structural equations that determine y_{2t} and y_{3t} have y_{4t} and y_{5t} as RHS endogenous variables. One obvious choice of FSRs is to use predetermined variables that are in the structural equations that explain y_{2t} and y_{3t}. Another choice is predetermined variables that are in the structural equations that explain y_{4t} and y_{5t}. One can continue this procedure through further layers as desired. (This rule of thumb is discussed in Fisher 1965.)

A rule of thumb about functional forms is to use mostly logarithms of variables if the RHS endogenous variables are in logarithms and to use mostly linear variables if the RHS endogenous variables are linear. Sometimes squares and cubes of variables are used, and sometimes variables multiplied by each other are used. There is no requirement that the same set of FSRs be used for different equations (although the same set must be used for all the RHS endogenous variables in a particular equation), and thus one may want to use different sets across equations, each set depending on the particular RHS endogenous variables in the equation.

The predetermined variables in the equation being estimated should also be included among the FSRs. Not doing so means treating these variables as endogenous. There is, however, an exception to this in the linear-in-coefficients case, which should be explained to avoid possible confusion. Consider (6.7) and let $X_i = (Y_i \quad X_{2i})$, where Y_i is the matrix of RHS endogenous variables and X_{2i} is the matrix of predetermined variables. If $\hat{X}_i$ is defined to be $(\hat{Y}_i \quad X_{2i})$, where $\hat{Y}_i = D_i Y_i$, rather than $D_i X_i$, and if formula (6.8) is used to compute $\hat{\alpha}_i$, then X_{2i} is treated as exogenous even if it is not included in Z_i. Equation (6.8) is the instrumental variables formula for $\hat{\alpha}_i$, and when $(\hat{Y}_i \quad X_{2i})$ is used for $\hat{X}_i$, X_{2i} is serving as its own instrument. When $(\hat{Y}_i \quad X_{2i})$ is used for $\hat{X}_i$, and X_{2i} is not included in X_i, (6.8) and (6.9) are not the same, and (6.9) does not produce consistent estimates. (See McCarthy 1971.) Equations (6.8) and (6.9) are the same only if X_{2i} is included in Z_i.

Covariance Matrix of All the Estimated Coefficients

Some of the stochastic simulation work in Chapters 7, 8, and 9 requires the covariance matrix of all the coefficients estimates, that is, the $k \times k$ covariance matrix of $\hat{\alpha}$, where $\hat{\alpha} = (\hat{\alpha}_1', \ldots, \hat{\alpha}_m')'$. For the completely linear case (linear in both variables and coefficients), this covariance matrix is presented in Theil (1971, pp. 499–500) for the case in which the same set of FSRs is used for each equation. For the more general case of a nonlinear model and a different set of FSRs for each equation, it is straightforward to show that the covariance matrix (say V_2) is

$$V_2 = \begin{bmatrix} V_{211} & \cdots & V_{21m} \\ \vdots & & \vdots \\ V_{2m1} & \cdots & V_{2mm} \end{bmatrix}, \tag{6.20}$$

where

$$V_{2ii} = \sigma_{ii}\left[\operatorname{plim}\frac{1}{T}G_i'D_iG_i\right]^{-1}, \tag{6.21}$$

$$V_{2ij} = \sigma_{ij}\left[\operatorname{plim}\frac{1}{T}G_i'D_iG_i\right]^{-1}\left[\operatorname{plim}\frac{1}{T}G_i'D_iD_jG_j\right]\left[\operatorname{plim}\frac{1}{T}G_j'D_jG_j\right]^{-1}. \tag{6.22}$$

An estimate of V_{2ii} is $\hat{V}_{2ii}$ in (6.6). An estimate of V_{2ij} (say $\hat{V}_{2ij}$) is

$$\hat{V}_{2ij} = \hat{\sigma}_{ij}(\hat{G}_i'D_i\hat{G}_i)^{-1}(\hat{G}_i'D_iD_j\hat{G}_j)(\hat{G}_j'D_j\hat{G}_j)^{-1}, \tag{6.23}$$

where $\hat{\sigma}_{ij} = T^{-1}\sum_{t=1}^{T}\hat{u}_{it}\hat{u}_{jt}$.

Regarding the proof that V_2 in (6.20) is the correct covariance matrix, the derivation in Theil can easily be modified to incorporate the case of different sets of FSRs. Nonlinearity can be handled as in Amemiya (1974, appendix 1), that is, by a Taylor expansion of each equation. The formal proof that V_2 is as in (6.20), (6.21), and (6.22) is straightforward but lengthy, and it is omitted here. Jorgenson and Laffont (1974, p. 363) incorrectly assert that the off-diagonal blocks of V_2 are zero.

6.3.3 Three-Stage Least Squares (3SLS)

3SLS estimates of α (say $\hat{\alpha}$) are obtained by minimizing

$$u'[\hat{\Sigma}^{-1} \otimes Z(Z'Z)^{-1}Z']u = u'Du \tag{6.24}$$

with respect to α, where $\hat{\Sigma}$ is a consistent estimate of Σ and Z is a $T \times K$ matrix of predetermined variables. As estimate of the covariance matrix of $\hat{\alpha}$ (say $\hat{V}_3$) is

$$\hat{V}_3 = (\hat{G}'D\hat{G})^{-1}, \tag{6.25}$$

where $\hat{G}$ is G evaluated at $\hat{\alpha}$. Σ is usually estimated from the 2SLS estimated residuals. This estimator is presented in Jorgenson and Laffont (1974), and it is further discussed in Amemiya (1977). Both prove consistency and asymptotic normality of 3SLS.

The 3SLS estimator that is based on minimizing (6.24) uses the same Z matrix for each equation. In small samples this can be a disadvantage of 3SLS relative to 2SLS. It is possible to modify (6.24) to include the case of different

Z_i matrices for each equation, and although this modification is not in general practical for large models, it is of some interest to consider. This estimator is the one that minimizes

$$u'\left[\begin{pmatrix} Z_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & Z_m \end{pmatrix}\begin{pmatrix} \hat{\sigma}_{11}Z_1'Z_1 & \cdots & \hat{\sigma}_{1m}Z_1'Z_m \\ \vdots & & \vdots \\ \hat{\sigma}_{m1}Z_m'Z_1 & \cdots & \hat{\sigma}_{mm}Z_m'Z_m \end{pmatrix}^{-1}\begin{pmatrix} Z_1' & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & Z_m' \end{pmatrix}\right]u = u'\bar{D}u \tag{6.26}$$

with respect to α. An estimate of the covariance matrix of this estimator is $(\hat{G}'\bar{D}\hat{G})^{-1}$. (6.26) reduces to (6.24) when $Z_1 = \cdots = Z_m = Z$. The computational problem with this estimator is that it requires inverting the middle matrix in brackets. This matrix is of dimension $K^* = \sum_{i=1}^{m} K_i$, which is generally a large number. For small to moderate models, however, it may be feasible to invert this matrix. This estimator has the advantage of being the natural full-information extension of 2SLS when different sets of FSRs are used. This estimator is a special case of one of the 3SLS estimators in Amemiya (1977, p. 963), namely the estimator determined by his equation (5.4), where his S_2 is the first matrix in brackets in (6.26) above.

Choice of First-Stage Regressors

If the estimator that minimizes (6.26) is used, a different set of FSRs can be used for each equation, and the same considerations apply here as apply for the 2SLS estimator. If the estimator that minimizes (6.24) is used, the same set of FSRs must be used for all equations. This set should be roughly equal to the union of the sets that are used (or that would be used) for the 2SLS estimator. The actual set used may have to be smaller than the union if the union contains more variables than seem sensible given the number of observations. Also, some nonlinear functions of the basic variables may be highly collinear (say, x_{1t}, $\log x_{1t}$, and x_{1t}^2), and one or more of these may be

able to be excluded without much loss of explanatory power in the first-stage regressions.

6.3.4 Full Information Maximum Likelihood (FIML)

Under the assumption that $(u_{1t}, \ldots, u_{mt})$ is independently and identically distributed as multivariate $N(0,S)$, the density function for one observation is

$$(6.27) \qquad (2\pi)^{-\frac{m}{2}}|S^*|^{\frac{1}{2}}|J_t|\exp\left(-\frac{1}{2}\sum_{i,j} u_{it}s_{ij}^*u_{jt}\right),$$

where $S^* = S^{-1}$ and s_{ij}^* is the ij element of S^*. The Jacobian J_t is defined in Section 6.1. The likelihood function of the sample $t = 1, \ldots, T$ is

$$(6.28) \qquad L^* = (2\pi)^{-\frac{mT}{2}}|S^*|^{\frac{T}{2}}\prod_{t=1}^{T}|J_t|\exp\left(-\frac{1}{2}\sum_{i,j,t} u_{it}s_{ij}^*u_{jt}\right),$$

and the log of L^* is

$$(6.29) \qquad \log L^* = -\frac{mT}{2}\log 2\pi + \frac{T}{2}\log|S^*| + \sum_{t=1}^{T}\log|J_t| - \frac{1}{2}\sum_{i,j,t} u_{it}s_{ij}^*u_{jt}.$$

Since $\log L^*$ is a monotonic function of L^*, maximizing $\log L^*$ is equivalent to maximizing L^*.

The problem of maximizing $\log L^*$ can be broken up into two parts: the first is to maximize $\log L^*$ with respect to the elements of S^*, and the second is to substitute the resulting expression for S^* into (6.29) and to maximize this "concentrated" likelihood function with respect to α. The derivative of $\log L^*$ with respect to s_{ij}^* is

$$(6.30) \qquad \frac{\partial \log L^*}{\partial s_{ij}^*} = \frac{T}{2}s^{*ij} - \frac{1}{2}\sum_{t=1}^{T} u_{it}u_{jt},$$

where s^{*ij} is the ij element of S^{*-1}. This derivative uses the fact that $\dfrac{\partial \log|A|}{\partial a_{ij}} = a^{ij}$ for a matrix A. Setting (6.30) equal to zero and solving for s^{*ij} yields

$$(6.31) \qquad s^{*ij} = \frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt}.$$

Since $S^* = S^{-1}$, $s^{*ij} = s_{ij}$, and therefore $s_{ij} = \frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt}$. Substituting (6.31) into (6.29) yields

$$\log L^* = -\frac{mT}{2}\log 2\pi + \frac{T}{2}\log|S^*| + \sum_{t=1}^{T}\log|J_t| - \frac{Tm}{2}. \tag{6.32}$$

The $-\frac{Tm}{2}$ term comes from the fact that $-\frac{1}{2}\sum_{i,j,t} u_{it}s^*_{ij}u_{jt} = -\frac{1}{2}\sum_{i,j} s^*_{ij}\sum_{t=1}^{T} u_{it}u_{jt} = -\frac{1}{2}\sum_{i,j} s^*_{ij}Ts^{*ij} = -\frac{Tm}{2}$. The first and last terms on the RHS of (6.32) are constants, and thus the expression to be maximized with respect to α consists of just the middle two terms. Since $\log|S^*| = \log|S^{-1}| = -\log|S|$, the function to be maximized can be written

$$L = -\frac{T}{2}\log|S| + \sum_{t=1}^{T}\log|J_t|, \tag{6.33}$$

where, as noted earlier, the ij element of S, s_{ij}, is $\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt}$. FIML estimates of α are thus obtained by maximizing L with respect to α. An estimate of the covariance matrix of these estimates (say $\hat{V}_4$) is

$$\hat{V}_4 = -\left(\frac{\partial^2 L}{\partial\alpha\partial\alpha'}\right)^{-1}, \tag{6.34}$$

where the derivatives are evaluated at the optimum.

Phillips (1982) has pointed out that Amemiya's proof of consistency and asymptotic efficiency (1977) is based on an incorrect lemma. This is corrected in a later paper (Amemiya 1982). Amemiya's article (1977), as corrected, shows that in the nonlinear case FIML is asymptotically more efficient than 3SLS under the assumption of normality. In the linear case FIML is consistent even if the error terms are not normally distributed, where "FIML" means the full information maximum likelihood estimator derived under the assumption of normality. In the nonlinear case this is not in general true, although it sometimes is. Phillips (1982) presents an example of a nonlinear model for which FIML is consistent for a wide class of error distributions. He also proves a "possibility" theorem, which shows that when FIML is consistent under normality it is always possible to find a nonnormal error distribu-

tion for which consistency is maintained. The assumption of normality is not necessary for the consistency of 3SLS. Given that 3SLS is consistent under a broader class of error distributions than is FIML, it is in this sense a more robust estimator. There is thus a trade-off between more robustness for 3SLS and more efficiency for FIML if the error terms are normal.

In the linear case Hausman (1975) has shown that FIML can be interpreted as an instrumental variables estimator in which all the nonlinear restrictions on the reduced form coefficients are taken into account in forming the instruments. This is contrary to the case for 3SLS, which forms the instruments from unrestricted estimates of the reduced form equations. FIML thus uses more information about the model than does 3SLS. In the linear case this makes no difference asymptotically because both estimates of the reduced form coefficient matrix are consistent (assuming that 3SLS uses all the explanatory variables in the reduced form equations as first-stage regressors). In the nonlinear case, however, it does make a difference because 3SLS does not obtain consistent estimates of the reduced form equations. In general, analytic expressions for the reduced form equations are not available, and 3SLS must be based on approximations to the equations. No such approximations are involved for FIML, and this is the reason it is asymptotically more efficient.

Another interesting difference between FIML and 3SLS concerns the LHS variable in each equation. Chow (1964) has shown in the linear case that FIML is the natural generalization of least squares in the sense that it minimizes the generalized variance of linear combinations of the endogenous variables. This is not true of 3SLS, which follows the principle of generalized variance but not of linear combinations. What Chow's interpretation shows is that there is no natural LHS variable for FIML: because of the linear combination aspect, each variable in the equation is treated equally. For 3SLS, on the other hand, a LHS variable must be chosen ahead of time for each equation.

For macroeconometric work it is unclear whether the symmetrical treatment of the endogenous variables by FIML is desirable or not. If the equations that are estimated are decision equations, as is the case for the model in Chapter 4, there is a natural LHS variable for each equation. FIML ignores this restriction, whereas 3SLS does not, so this may be an argument in favor of 3SLS. Given this difference and given the fact that 3SLS is more robust to specification errors regarding the distribution of the error terms, the question of which estimator is likely to be better in practice is far from clear.

6.3.5 Least Absolute Deviations (LAD)

LAD estimates of α_i (say $\hat{\alpha}_i$) are obtained by minimizing

$$(6.35) \qquad \sum_{t=1}^{T} |u_{it}|$$

with respect to α_i. For the general nonlinear model the asymptotic distribution of $\hat{\alpha}_i$ is not known. For the standard regression model $y_i = X_i\alpha_i + u_i$, where X_i is a matrix of exogenous variables and u_{it} is independent and identically distributed with distribution function F, Bassett and Koenker (1978) have shown that the asymptotic distribution of $\hat{\alpha}_i$ is normal with mean α_i (thus $\hat{\alpha}_i$ is consistent) and covariance matrix $\omega^2 Q$, where $Q = \lim \frac{1}{T} X_i' X_i$ and ω^2 is the asymptotic variance of the sample median from random samples with distribution F. Amemiya (1982) supplies an alternative proof of this proposition.

The LAD estimator is an example of a robust estimator. An estimator is said to be more robust than another if its properties are less sensitive to changes in the assumptions about the model, particularly assumptions about the distribution of the error terms. In a number of cases the LAD estimator has been shown to be more robust that the OLS estimator to deviations of the error terms from normality. In particular, the LAD estimator seems well suited to cases in which the distribution of the error terms is fat-tailed.

The literature in statistics on robust estimation is now quite extensive, and there are many types of robust estimators. The estimators differ primarily in how error terms that are large in absolute value (that is, outliers) are weighted. These estimators have not been used very much in applied econometric work, so there is little experience to guide the choice of estimator. Since LAD is the simplest of the estimators, it seems to be the best one to start with. An interesting open question is how useful any of the robust estimators are for empirical work in economics.

6.3.6 Two-Stage Least Absolute Deviations (2SLAD)

There are two ways of interpreting the 2SLS estimator that is based on the minimization of (6.5), and these need to be discussed before considering the LAD analogue of 2SLS. For purposes of the discussion in this section and in Section 6.5.4, it will be assumed that the model (6.1) can be written

$$(6.1)' \qquad y_{it} = h_i(y_t, x_t, \alpha_i) + u_{it}, \qquad i = 1, \ldots, n. \qquad t = 1, \ldots, T,$$

where in the ith equation y_{it} appears only on the LHS. Given this and given that $D_i' = D_i$ and $D_iD_i = D_i$, (6.5) can be written

$$
\begin{aligned}
(6.36) \qquad u_i'D_iu_i &= u_i'D_iD_iu_i \\
&= (y_i' - h_i')D_iD_i(y_i - h_i) \\
&= (y_i'D_i - h_i'D_i)(D_iy_i - D_ih_i) \\
&= (\hat{y}_i' - \hat{h}_i')(\hat{y}_i - \hat{h}_i) \\
&= \hat{y}_i'\hat{y}_i - 2\hat{y}_i'\hat{h}_i + \hat{h}_i'\hat{h}_i,
\end{aligned}
$$

where $\hat{y}_i = D_iy_i$ and $\hat{h}_i = D_ih_i$. Instead of minimizing (6.36), consider minimizing

$$
(6.37) \qquad (y_i' - \hat{h}_i')(y_i - \hat{h}_i) = y_i'y_i - 2y_i'\hat{h}_i + \hat{h}_i'\hat{h}_i.
$$

Given that $\hat{y}_i'\hat{h}_i = y_i'D_iD_ih_i = y_i'D_ih_i = y_i'\hat{h}_i$ and given that $\hat{y}_i'\hat{y}_i$ and $y_i'y_i$ are not a function of α_i, minimizing (6.36) with respect to α_i is equivalent to minimizing (6.37). Therefore, the 2SLS estimator can be interpreted as minimizing either $(\hat{y}_i' - \hat{h}_i')(\hat{y}_i - \hat{h}_i)$ or $(y_i' - \hat{h}_i')(y_i - \hat{h}_i)$. The first interpretation is Basmann's (1957) and the second is Theil's (1953).

For the LAD analogue it is unclear which interpretation should be used. Using Basmann's one would minimize

$$
(6.38) \qquad \sum_{t=1}^{T} |\hat{y}_{it} - \hat{h}_{it}|,
$$

and using Theil's one would minimize

$$
(6.39) \qquad \sum_{t=1}^{T} |y_{it} - \hat{h}_{it}|.
$$

In this case the choice matters in that minimizing (6.38) and minimizing (6.39) lead to different estimates. Amemiya (1982) has proposed minimizing

$$
(6.40) \qquad \sum_{t=1}^{T} |qy_{it} + (1 - q)\hat{y}_{it} - \hat{h}_{it}|,
$$

where q is chosen ahead of time by the investigator. The estimator that is based on minimizing (6.40) will be called 2SLAD.

For the general nonlinear model the asymptotic distribution of 2SLAD is not known. For the linear model Amemiya (1982) has proved that 2SLAD is consistent. He has also in the linear case derived formulas for the asymptotic covariance matrix of the estimator for particular assumptions about the distributions of the error terms. If all the distributions are normal, he has proved that 2SLAD is asymptotically normal.

6.4 Sample Size Requirements for FIML and the Estimation of Subsets of Coefficients

6.4.1 Sample Size Requirements

For large models there may not be enough observations to estimate all the coefficients by FIML. For a linear model without identities, Sargan (1975) has shown that the FIML likelihood function has an infinite maximum if the number of observations is less than the number of endogenous and exogenous variables. With respect to more general models, Parke (1982b) has derived the FIML sample size requirement for models with identities, nonlinearity in variables, and serial correlation coefficients. It will be useful to consider Parke's main results.

Consider first the case of no identities and no serial correlation coefficients. If the model is only nonlinear in variables, it can be written

$$QA = U, \tag{6.41}$$

where Q is a $T \times q$ matrix of variables that are functions of the basic endogenous and exogenous variables, A is a $q \times m$ matrix of coefficients, and U is a $T \times m$ matrix of error terms. In general the variables in Q are nonlinear functions of the basic endogenous and exogenous variables, although many of them may simply be the basic variables. The total number of variables in the model is q. Under the assumption that each of these variables appears at least once in the model with a nonzero coefficient (a trivial assumption), Parke has shown that the sample size requirement for FIML is $T \geq q$.

Adding identities does not in general change this requirement. One need not include in Q variables that appear in identities but not in the structural equations when one is calculating the sample size requirement. When the identity is what Parke calls a "closed" identity, one that imposes a linear dependency on the columns of Q, the sample size requirement is less. For i closed identities the dependencies can be written

$$QP = 0, \tag{6.42}$$

where P is a $q \times i$ matrix of known coefficients. For i closed identities the sample size requirement is $T \geq q - i$.

An example of a model with a closed identity is the following:

$$Q_{1t} = \alpha_{11} + \alpha_{12}Q_{3t} + \alpha_{13}Q_{4t} + u_{1t}, \tag{6.43}$$

$$Q_{2t} = \alpha_{21} + \alpha_{22}Q_{3t} + \alpha_{23}Q_{5t} + u_{2t}, \tag{6.44}$$

$$Q_{3t} = Q_{1t} + Q_{2t}. \tag{6.45}$$

In this case Q_{3t} could be substituted out of the stochastic equations (6.43) and (6.44) without introducing any new variables, and therefore it is not a variable that needs to be counted against the sample size requirement. Identities of this type are likely to be rare. (There are, for example, no closed identities in the model in Chapter 4.) A much more common identity in the model just presented would be $Q_{3t} = Q_{1t} + Q_{2t} + Q_{6t}$, where Q_{6t} does not appear in the stochastic equations. In this case the identity is "open," and Q_{3t} does count against the sample size requirement.

The treatment of serial correlation is somewhat more involved. Assume that x_{jt} appears in equation i, where equation i has first-order serially correlated errors. After the equation is transformed, the variable appears as $x_{jt}^* = x_{jt} - \rho_i x_{jt-1}$. If x_{jt} and x_{jt-1} appear nowhere else in the model, x_{jt}^* can be counted as only one variable. Otherwise, both x_{jt} and x_{jt-1} must be counted. Even if x_{jt} appears in many equations with first-order serially correlated errors (and in general different serial correlation coefficients), the number of variables to be counted is still only two (x_{jt} and x_{jt-1}). What this says is that the introduction of first-order serial correlation to an equation at most increases the number of variables to be counted by the number of original variables in the equation. The increase is less than this if at least some of the original variables and their one-period-lagged values do not appear elsewhere in the model. If none of the original variables and their lagged values appear elsewhere in the model, the introduction of serial correlation to an equation does not increase the number of variables to be counted. Similar arguments apply to higher-order serial correlation. For example, the introduction of second-order serial correlation at most increases the number of variables to be counted by twice the number of original variables in the equation.

The introduction of a constraint across coefficients does not in general reduce the sample size requirement. If it does, it is sometimes possible to write the model with fewer variables after the constraint is imposed. Brown (1981) shows that this is always the case for a linear constraint across the coefficients in a single equation. As a general rule of thumb, if it is not obvious that a constraint can be used to write the model with fewer variables, it should be assumed that the constraint does not reduce the sample size requirement.

6.4.2 Estimation of Subsets of Coefficients

It is possible to reduce the sample size requirement of FIML by fixing some coefficients at, say, their 2SLS values (or some other consistently estimated values) and estimating the remaining coefficients by FIML. One can fix either all the coefficients in a given equation or only some of them. If all the

coefficients are fixed, the equation is still taken to be part of the estimation problem in the sense that the covariance matrix S in (6.33) is still $m \times m$, but none of the coefficients in the equation are estimated by FIML.

Consider the problem by estimating the free coefficients by FIML, and write the relevant subset of the model as

$$Q_1 A_1 = U_1, \tag{6.46}$$

where Q_1 is $T \times q_1$, A_1 is $q_1 \times m_1$, and U_1 is $T \times m_1$. The matrix A_1 is the matrix of free coefficients, and m_1 is the number of equations in which at least one coefficient is free. q_1, as will be seen, is the number of variables that count for purposes of calculating the sample size requirement. Its determination requires some explanation. Assume that x_{jt} and x_{kt} appear in equation i and that their coefficients (α_{i1} and α_{i2}) are fixed. Assume that $\log y_{it}$ is the LHS variable. This equation can be rewritten with $\log y_{it} - \hat{\alpha}_{i1} x_{jt} - \hat{\alpha}_{i2} x_{kt}$ on the LHS and x_{jt} and x_{kt} eliminated from the RHS. ($\hat{\alpha}_{i1}$ and $\hat{\alpha}_{i2}$ are the consistent estimates of α_{i1} and α_{i2}.) If $\log y_{it}$, x_{jt}, and x_{kt} do not appear elsewhere in the model, this fixing of the coefficients has eliminated two variables. If $\log y_{it}$ does appear elsewhere but x_{jt} and x_{kt} do not, only one variable has been eliminated because the new LHS variable and $\log y_{it}$ count as separate variables. If x_{jt} and x_{kt} appear elsewhere, no variables are eliminated. If all the coefficients in an equation are fixed, a variable in the equation is eliminated if it appears nowhere else in the model. q_1 is the number of variables that remain after all possible eliminations.

Parke has shown that the sample size requirement for this reduced problem is $T \geq q_1 + m_2 - i_1$, where $m_2 = m - m_1$ is the number of equations for which none of the coefficients are estimated and i_1 is the number of closed identities that pertain to the reduced set of equations (that is, the set of equations not counting the m_2 equations for which no coefficients are estimated). Note that one observation is needed for each of the m_2 equations that are not estimated.

Given this result, if the sample size requirement is not met for the complete model, the problem can be reduced by fixing various coefficients until it is met. An example of this procedure is presented in Section 6.5.2.

It should finally be noted that because of computational costs, one may want to restrict the size of the estimation problem even if the sample size requirement is met. The obvious way to do this is to fix some of the coefficients at their 2SLS estimates. This can be done for both the FIML and 3SLS estimators.

When only a subset of the coefficients is estimated by FIML or 3SLS, the easiest thing to do with regard to the estimation of the covariance matrix of all

the coefficient estimates is to assume that the coefficient estimates that are fixed with respect to the FIML or 3SLS estimation problem are uncorrelated with the FIML or 3SLS coefficient estimates. This allows the covariance matrix of all the coefficient estimates to be pieced together from the covariance matrix of the fixed estimates and the covariance matrix of the FIML or 3SLS estimates. Since correlation of coefficient estimates across equations is usually small relative to the correlation within an equation, the errors introduced by this procedure are likely to be fairly small in most applications. This is particularly true if the coefficient estimates that are fixed are of lesser importance than the others.

6.5 Computational Procedures and Results

6.5.1 OLS and 2SLS

For equations that are nonlinear in variables only, closed-form expressions exist for the OLS and 2SLS estimators. For 2SLS the expression is (6.9), and for OLS it is (6.9) with X_i replacing $\hat{X}_i$. If the nonlinearity in coefficients is due only to the presence of serially correlated error terms, the estimates can be obtained by solving (6.13) and (6.14) (or Eqs. 6.13′ and 6.14′) or higher-order versions of these iteratively. For general nonlinearities in coefficients, (6.5) must be minimized using some general-purpose algorithm like the DFP algorithm discussed in Section 2.5.

Results for the US Model

The 2SLS estimates of the US model are presented in Chapter 4. The first-stage regressors that were used for these estimates are given in Table 6-1. Two common sets are presented first in Table 6-1, one for equations in which the RHS endogenous variables are primarily linear and one for equations in which the RHS endogenous variables are primarily in logarithms. The additional FSRs that were used for each equation are presented second. These FSRs are primarily variables that appear as explanatory variables in the equation being estimated but that are not part of the common set. The common sets include 34 variables, and the number of additional variables ranges from 0 to 9. The equations that are estimated by OLS have no RHS endogenous variables.

The time taken to estimate the 30 equations by 2SLS was about 3.0 minutes on the IBM 4341 and about 8.4 minutes on the VAX. The estimation of the covariance matrix of all the coefficient estimates, V_2 in (6.20), took about 5.5

TABLE 6-1. First stage regressors for the US model for 2SLS

	Basic sets	
	Linear	Log
1	constant	constant
2	$(AA/POP)_{-1}$	$\log(AA/POP)_{-1}$
3	$C_g + C_s$	$\log(C_g + C_s)$
4	$(CD/POP)_{-1}$	$\log(CD/POP)_{-1}$
5	$(CN/POP)_{-1}$	$\log(CN/POP)_{-1}$
6	$(CS/POP)_{-1}$	$\log(CS/POP)_{-1}$
7	$(1 - d^M_{1g} - d^M_{1s} - d_{4g} - d_{4s})_{-1}$	$\log(1 - d^M_{1g} - d^M_{1s} - d_{4g} - d_{4s})_{-1}$
8	EX	log EX
9	H_{f-1}	$\log H_{f-1}$
10	$(IH_h/POP)_{-1}$	$\log(IH_h/POP)_{-1}$
11	$(IM/POP)_{-1}$	$\log(IM/POP)_{-1}$
12	$(J_f - JHMIN)_{-1}$	$\log(J_f/JHMIN)_{-1}$
13	$(J_gH_g + J_mH_m + J_sH_s)/POP$	$\log[(J_gH_g + J_mH_m + J_sH_s)/POP]$
14	$(KH/POP)_{-1}$	$\log(KH/POP)_{-1}$
15	$(KK - KKMIN)_{-1}$	$\log(KK/KKMIN)_{-1}$
16	$\dot{M1}_{-1}$	$\dot{M1}_{-1}$
17	$\dot{PD}_{-1}$	$\dot{PD}_{-1}$
18	P_{f-1}	$\log P_{f-1}$
19	PIM	log PIM
20	RB_{-1}	RB_{-1}
21	RS_{-1}	RS_{-1}
22	RS_{-2}	RS_{-2}
23	t	t
24	$(TR_{gh} + TR_{sh})/(POP \cdot P_{h-1})$	$\log[(TR_{gh} + TR_{sh})/(POP \cdot P_{h-1})]$
25	V_{-1}	$\log V_{-1}$
26	W_{f-1}	$\log W_{f-1}$
27	Y_{-1}	$\log Y_{-1}$
28	Y_{-2}	$\log Y_{-2}$
29	Y_{-3}	$\log Y_{-3}$
30	Y_{-4}	$\log Y_{-4}$
31	$YN/(POP \cdot P_{h-1})$	$\log[YN/(POP \cdot P_{h-1})]$
32	Z_{-1}	Z_{-1}
33	UR_{-1}	UR_{-1}
34	ZZ_{-1}	ZZ_{-1}

(continued)

TABLE 6-1 (continued)

Equation number	Additional first stage regressors for each equation
1	PCS_{-1}, WA_{-1}
2	PCN_{-1}, WA_{-1}
3	PCD_{-1}, RM_{-1}, WA_{-1}, $[YTR/(POP \cdot P_h)]_{-1}$
4	OLS estimation
5	$(L1/POP1)_{-1}$, P_{h-1}, WA_{-1}
6	$(L2/POP2)_{-1}$, P_{h-1}, WA_{-1}
7	$(L3/POP3)_{-1}$, P_{h-1}, WA_{-1}
8	$(LM/POP)_{-1}$, P_{h-1}, WA_{-1}
9 [a]	$\log[M_h/(POP\ P_h)]_{-1}$, $\log[YT/(POP\ P_h)]_{-1}$
10 [a]	$\log(1 + d_{5g} + d_{5s})$
11	D593, D594, D601, $D601_{-1}$, V_{-2}
12	$\delta_K KK_{-1}$, IK_{f-1}, RBA_{-1}
13 [a]	D593, D594, $D594_{-1}$, $\Delta \log J_{f-1}$, $\log(J_f/JHMIN)_{-2}$
14 [a]	$\log H_{f-2}$, $\log(J_f/JHMIN)_{-2}$
15	OLS estimation
16 [a]	$\log PX_{-1}$
17 [a]	$\log(M_f/PX)_{-1}$, $1 - d_{2g} - d_{2s}$
18	D_{f-1}, $(\pi_f - T_{fg} - T_{fs})_{-1}$, $d_{2g} + d_{2s}$
19	OLS estimation
20	OLS estimation
21	OLS estimation
22	$(BO/BR)_{-1}$, $(RS - RD)_{-1}$
23	no extra
24	RM_{-1}
25	$\Delta(CF - T_{fg} - T_{fs})_{-1}$, CF_{-1}, $d_{2g} + d_{2s}$, $(T_{fg} + T_{fs})_{-1}$
26 [a]	$\log[CUR/(POP \cdot PX)]_{-1}$, $\log(X/POP)_{-1}$
27	PIM_{-1}, PX_{-1}, RMA_{-1}, D651, D652, D691, D692, D714, D721
28 [a]	$\log U_{-1}$, $\log UB_{-1}$
29	OLS estimation
30	$DD793 \cdot \dot{M}1_{-1}$, JJ^*_{-1}

Note: a. Basic set is log.

minutes on the IBM 4341 and about 7.8 minutes on the VAX. The derivatives in the G_i matrices that are needed for the estimation of the covariance matrix were computed numerically.

Eight of the 30 equations were estimated under the assumption of first-order serial correlation of the error terms. The iterative procedure described above was used. The starting value of ρ was always zero, and the number of iterations required for convergence was 10, 7, 11, 4, 13, 6, 4, and 5 respectively. Convergence was defined to take place when successive estimates of ρ were within .001 of each other.

OLS estimation of the 30 equations took about .2 minutes on the IBM 4341 and about .5 minutes on the VAX, which compares to about 3.0 and 8.4 minutes respectively for 2SLS estimation. The number of coefficients estimated in any one equation is small compared to the number estimated in the first-stage regressions, and this is the reason for the considerably larger expense of the 2SLS estimates. The maximum number of coefficients estimated in an equation is 12, whereas the minimum number estimated in a first-stage regression is 34. Nevertheless, the cost of 2SLS estimation is small relative to many other costs reported below.

6.5.2 FIML

Until recently the estimation of large nonlinear models by FIML was not computationally feasible, but this has now changed. The computational problem can be separated into two main parts: the first is to find a fast way of computing L in (6.33) for a given value of α, and the second is to find an algorithm capable of maximizing L.

The main cost of computing L is computing the Jacobian term. Two savings can be made here. One is to exploit the sparseness of the Jacobian. The number of nonzero elements in J_t is usually much less than n^2. For the US model, for example, n is 128 (so $n^2 = 16{,}384$), whereas the number of nonzero elements is only 441. Considerable computer time is saved by using sparse matrix routines to calculate the determinant of J_t.

The second saving is based on an approximation. Consider approximating $\Sigma_{t=1}^{T} \log|J_t|$ by simply the average of the first and last terms in the summation multiplied by T: $\frac{T}{2}(\log|J_1| + \log|J_T|)$. Let S_0 denote the true summation, and let S_1 denote the approximation. It turns out in the applications I have dealt with that $S_0 - S_1$ does not change very much as the coefficients change from their starting values (usually the 2SLS estimates) to the values that maximize

the likelihood function. In other words, $S_0 - S_1$ is nearly a constant. This means that S_1 can be used instead of S_0 in computing L, and thus considerable computer time is saved since the determinant of the Jacobian only needs to be computed twice rather than T times for each evaluation of L. For the US model T is 115. Using S_1 in place of S_0 means, of course, that the coefficient values that maximize the likelihood function are not the exact FIML estimates. If one is concerned about the accuracy of the approximation, one can switch from S_1 to S_0 after finding the maximum using S_1. If the approximation is good, one should see little further change in the coefficients; otherwise additional iterations using the algorithm will be needed to find the true maximum.

The choice of algorithm turns out to be crucial in maximizing L for large nonlinear models. My experience is that general-purpose algorithms like DFP do not work, and in fact the only algorithm that does seem to work is the Parke algorithm (1982a), which is a special-purpose algorithm designed for FIML and 3SLS estimation. This algorithm exploits two key features of models. The first is that the mean of a particular equation's estimated residuals is approximately zero for the FIML and 3SLS estimates. For OLS this must be true, and empirically it turns out that it is approximately true for other estimators. The second feature is that the correlation of coefficient estimates within an equation is usually much greater than the correlation of coefficients across equations.

The problem with algorithms like DFP that require numerical first derivatives is that the computed gradients do not appear to be good guides regarding the directions to move in. Gradients are computed by perturbing one coefficient at a time. When a coefficient is changed without the constant term in the equation also being changed to preserve the mean of the residuals, a large change in L results (and thus a large derivative). This result can obviously be quite misleading. The Parke algorithm avoids this problem by spending most of its time perturbing two coefficients at once, namely a given coefficient and the constant term in the equation in which the coefficient appears. The constant term is perturbed to keep the mean of the residuals unchanged. (The algorithm does not, of course, do this all the time, since the means of the residuals must also be estimated). To take advantage of the generally larger correlation within an equation than between equations, the Parke algorithm spends more time searching within equations than between them. General-purpose algorithms do not do this, since they have no knowledge of the structure of the problem.

It should also be noted regarding the computational problem that if only a

few coefficients are changed before a new value of L is computed, considerable savings can be made by taking advantage of this fact. If, for example, the coefficients are not in the Jacobian, the Jacobian term does not have to be recomputed. If only a few equations are affected by the change in coefficients, only a few rows and columns in the S matrix have to be recomputed. Since the Parke algorithm spends much of its time perturbing two coefficients at a time, it is particularly suited for these kinds of savings.

The estimated covariance matrix for the FIML coefficient estimates, $\hat{V}_4$ in (6.34), is difficult to compute. It is not part of the output of the Parke algorithm, and thus extra work is involved in computing it once the algorithm has found the optimum. My experience is that simply trying to compute the second derivatives of L numerically does not result in a positive-definite matrix. Although the true second-derivative matrices at the optimum are undoubtedly positive-definite, they seem to be nearly singular. If this is true, small errors in the numerical approximations to the second derivatives may be sufficient to make the matrix not positive-definite.

Fortunately, there is an approach to computing $\hat{V}_4$ that does work, which is derived from Parke (1982a). Parke's results suggest that the inadequate numerical approximations may be due to the fact that the means of the RHS variables in the estimated equations are not zero. If so, the problem can be solved by subtracting the means from the RHS variables before taking numerical derivatives. Let β denote the coefficient vector that pertains to the model after the means have been subtracted, and let α denote the original coefficient vector. The relationship between α and β is

$$\alpha = M \cdot \beta, \tag{6.47}$$

where M is a $k \times k$ square matrix that is composed of the identity matrix plus additional nonzero elements that represent the means adjustments. Unless there are constraints across equations, M is block-diagonal. Assume, for example, that the first equation of the model is

$$y_{1t} = \beta_1 + \beta_2(y_{2t} - m_2) + \beta_3(y_{3t} - m_3) + u_{1t}, \qquad t = 1, \ldots, T, \tag{6.48}$$

where m_2 and m_3 are the sample means of y_{2t} and y_{3t} respectively. This equation can be written

$$\begin{aligned} y_{1t} &= \beta_1 - \beta_2 m_2 - \beta_3 m_3 + \beta_2 y_{2t} + \beta_3 y_{3t} + u_{1t} \\ &= \alpha_1 + \alpha_2 y_{2t} + \alpha_3 y_{3t} + u_{1t}, \qquad t = 1, \ldots, T. \end{aligned} \tag{6.49}$$

In this case the part of (6.47) that corresponds to the first equation is

$$(6.50) \qquad \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} 1 & -m_2 & -m_3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Parke found that the covariance matrix of β could easily be computed numerically. Let $\hat{V}_4(\beta)$ denote this matrix:

$$(6.51) \qquad \hat{V}_4(\beta) = -\left[\frac{\partial^2 L(M \cdot \beta)}{\partial\beta\partial\beta'}\right]^{-1}.$$

Given $\hat{V}_4(\beta)$, the covariance matrix of α is simply

$$(6.52) \qquad \hat{V}_4 = M \cdot \hat{V}_4(\beta) \cdot M'.$$

$\hat{V}_4$ can thus be obtained by first computing the covariance matrix of the coefficients of the transformed model (that is, the model in which the RHS variables have zero means) and then using (6.52) to get the covariance matrix of the original coefficients.

Results for the US Model

The solution of the FIML estimation problem for the US model is reported in Table 6-2. There are 169 unconstrained coefficients in the model; 107 of these were estimated by FIML, with the remaining fixed at their 2SLS estimates. The coefficients that were not estimated by FIML include the dummy variable coefficients in Eqs. 11, 13, and 27 and all the coefficients in Eqs. 5, 6, 7, 8, 15, 18, 19, 20, 21, 25, 28, and 29. These coefficients and equations were judged to be less important than the others, although this is obviously a subjective choice. The sample size requirement for this subset of coefficients is 99. There are 115 observations.

The starting values were the 2SLS estimates. The value of L in (6.34) at these estimates is 5098.66. The change in L after 70 iterations in Table 6-2 is 181.76. On the first iteration the Parke algorithm increased L by 67.07, and on the second and third iterations it increased L by 8.68 and 7.64 respectively. The change after three iterations was thus 83.39, which is 45.9 percent of the total change. This illustrates a general feature of the Parke algorithm: it climbs very quickly for the first few iterations and then slows down considerably for the rest.

TABLE 6-2. Solution of the FIML estimation problem for the US model

L = L in (6.33)
L at start (2SLS estimates) = 5098.66
L after 70 iterations = 5280.42
Total ΔL = 181.76

Iter. no.	ΔL	Iter. no.	ΔL	Iter. no.	ΔL	Iter. no.	ΔL	Iter. no.	ΔL
1	67.07	15	2.23	29	1.60	43	.43	57	.10
2	8.68	16	2.75	30	1.30	44	.31	58 [a]	.08
3	7.64	17	3.21	31	1.05	45	.42	59 [a]	.05
4	4.61	18	3.40	32	1.29	46	.39	60 [a]	.05
5	4.89	19	3.08	33	1.12	47	.30	61 [a]	.06
6	6.84	20	2.58	34	.53	48	.36	62 [a]	.06
7	5.51	21	3.19	35	.47	49	.20	63	.05
8	4.17	22	2.71	36	.70	50	.14	64	.04
9	4.10	23	1.38	37	.57	51	.20	65	.05
10	5.17	24	1.49	38	1.16	52	.23	66	.08
11	5.04	25	2.38	39	.99	53	.10	67	.11
12	2.54	26	1.20	40	.83	54	.20	68	.11
13	3.51	27	1.13	41	.41	55	.10	69	.10
14	3.15	28	1.18	42	.41	56	.10	70 [b]	.06

Notes: a. 13 Jacobians computed rather than 2. (Computations at observations 1, 10, 19, 28, 37, 46, 55, 64, 73, 82, 91, 100, 115.)

b. Between iterations 69 and 70, 26 coefficients changed by 1.0 percent or more and 4 changed by 5.0 percent or more. The largest 3 changes were 8.1, 12.6, and 18.4 percent.

- Model consists of 169 unconstrained coefficients. 107 coefficients estimated by FIML. Sample period is 1954 I - 1982 III (115 observations).
- Each iteration requires about 462 function evaluations. The time per iteration when 2 Jacobians were computed was about 2.8 minutes on the IBM 4341 and about 7.3 minutes on the VAX. When 13 Jacobians were computed the respective times were 5.4 minutes and 12.3 minutes. The total time on the IBM 4341 for the 70 iterations was thus about 65×2.8 minutes $+ 5 \times 5.4$ minutes = 3.5 hours.
- The time taken to compute the FIML covariance matrix, $\hat{V}_4$ in (6.34), was about 53 minutes on the IBM 4341 and about 2.1 hours on the VAX.

Between iterations 58 and 62 the number of Jacobians computed to approximate the sum was increased from 2 to 13. When 13 Jacobians were used, the sum was approximated by interpolating between the points. As can be seen in the table, the change in L was little affected by this. If the use of 2 Jacobians in fact provided a poor approximation, it is likely that the Parke algorithm would have increased L by much more than it did on the first few iterations after the switch. That it did not is some evidence in favor of the approximation.

Another way of looking at the 2 versus 13 question is to consider how sensitive the difference in L computed the two ways is to changes in the coefficients. The following results help answer this:

Value of L	*2 Jacobians*	*13 Jacobians*	*Difference*
L at start (2SLS estimates)	5,098.66	5,284.49	−185.83
L after 59 iterations	5,279.53	5,464.04	−184.51
L after 62 iterations	5,279.82	5,464.34	−184.52
L after 70 iterations	5,280.42	5,464.96	−184.54

It is clear that the difference is little affected by the change in the coefficients from the 2SLS estimates to the estimates at the end of iteration 70. It thus seems that the use of 2 Jacobians is adequate. Note that this saves considerable time, since the cost of one iteration of the Parke algorithm increases from about 2.8 minutes to about 5.4 minutes on the IBM 4341 when 13 rather than 2 Jacobians are used.

As discussed earlier, when only one or two coefficients are being changed by the algorithm, many of the calculations involved in computing L do not have to be performed. In the present example, if these cost savings had not been used, the time taken for one iteration of the Parke algorithm would have increased by about a factor of 4.5, which is a considerable difference. As will be seen in the next section, this difference is even more pronounced in the 3SLS estimation problem.

It is a characteristic of the estimation problem that the likelihood function is fairly flat in the vicinity of the optimum. For example, the change in L on iteration 70 was only .06, and yet, as reported in note b in the table, 26 coefficients changed by 1.0 percent or more and 4 changed by 5.0 percent or more. The largest three changes were 8.1, 12.6, and 18.4 percent. The coefficients that change this much are obviously not significant, and they are not coefficients that are very important in the model. Nevertheless, these results do point out one of the reasons the FIML estimation problem is so hard to solve.

As noted in Table 6-2, the total time for the FIML estimation problem was about 3.5 hours on the IBM 4341. The time taken to compute the FIML covariance matrix after the coefficient estimates were obtained was about 53 minutes. The M transformation discussed earlier was used in the calculation of this matrix, and the second derivatives were obtained numerically.

6.5.3 3SLS

The 3SLS estimation problem is to minimize (6.24). The only cost saving to note for this problem is that the D matrix, which is $m \cdot T \times m \cdot T$, need not be calculated anew each time (6.24) is computed if only a few coefficients are changed.

TABLE 6-3. First stage regressors for the US model for 3SLS

From the basic sets for 2SLS	Additional first stage regressors
1. constant	35. WA_{-1}
2. $(AA/POP)_{-1}$	36. RM_{-1}
3. $C_g + C_s$	37. $\log(M_h/(POP \cdot P_h))_{-1}$
4. $(CD/POP)_{-1}$	38. $\log(1 + d_{5g} + d_{5s})$
5. $(CN/POP)_{-1}$	39. V_{-2}
6. $(CS/POP)_{-1}$	40. IK_{f-1}
7. $(1 - d^M_{1g} - d^M_{1s} - d_{4g} - d_{4s})_{-1}$	41. $\delta_K KK_{-1}$
8. EX	42. $\log(J_f/JHMIN)_{-2}$
9. $\log H_{f-1}$	43. $\log(M_f/PX)_{-1}$
10. $(IH_h/POP)_{-1}$	44. $(BO/BR)_{-1}$
11. $(IM/POP)_{-1}$	45. RD_{-1}
12. $\log(J_f/JHMIN)_{-1}$	46. $\log(CUR/(POP \cdot PX))_{-1}$
13. $(J_g H_g + J_m H_m + J_s H_s)/POP$	47. PIM_{-1}
14. $(KH/POP)_{-1}$	48. PX_{-1}
15. $(KK - KKMIN)_{-1}$	49. $DD793 \cdot \dot{M1}_{-1}$
16. $\dot{M1}_{-1}$	
17. $\dot{PD}_{-1}$	
18. $\log P_{f-1}$	
19. log PIM	
20. RB_{-1}	
21. RS_{-1}	
22. RS_{-2}	
23. t	
24. $(TR_{gh} + TR_{sh})/(POP \cdot P_{h-1})$	
25. V_{-1}	
26. $\log W_{f-1}$	
27. Y_{-1}	
28. Y_{-2}	
29. Y_{-3}	
30. Y_{-4}	
31. $(YN/(POP \cdot P_h))_{-1}$	
32. Z_{-1}	
33. UR_{-1}	
34. ZZ_{-1}	

TABLE 6-4. Solution of the 3SLS estimation problem for the US model

F = u'Du in (6.24)
F at start (2SLS estimates) = 1890.33
F after 26 iterations = 1843.78
Total |ΔF| = 46.55

Iteration number	|ΔF|	Iteration number	|ΔF|
1	23.90	14	.24
2	9.31	15	.16
3	6.60	16	.10
4	1.91	17	.13
5	.92	18	.12
6	.67	19	.11
7	.62	20	.08
8	.29	21	.06
9	.32	22	.08
10	.22	23	.05
11	.21	24	.07
12	.12	25	.08
13	.16	26 [a]	.05

Notes: a. Between iterations 25 and 26 eight coefficients changed by 1.0 percent or more. The largest three changes were 6.6, 10.5, and 26.7 percent.

- Model consists of 169 unconstrained coefficients. 107 coefficients estimated by 3SLS. Sample period is 1954 I - 1982 III (115 observations).
- Each iteration requires about 444 function evaluations. The time per iteration was about 4 minutes on the IBM 4341 and about 11 minutes on the VAX. The total time on the IBM 4341 was thus about 26 × 4 minutes = 1.7 hours.
- The time taken to compute the 3SLS covariance matrix, $\hat{V}_3$ in (6.25), was about 23 minutes on the IBM 4341 and about 11 minutes on the VAX.

Results for the US Model

The first-stage regressors for this problem are presented in Table 6-3. There are 49 variables in this set. A number of the variables in Table 6-1 that were used for the 2SLS estimates were not used for the 3SLS estimates because of the desire to keep the number relatively small. The 2SLS estimates of the

residuals were used to compute $\hat{\Sigma}$ in (6.24), which remained unchanged throughout the solution of the problem.

The same subset of coefficients was estimated by 3SLS as was estimated by FIML. The solution of the 3SLS problem is reported in Table 6-4. This problem was easier to solve than the FIML problem. Again, the 2SLS estimates were used as starting values. The total change in the objective function, F, after 26 iterations was 46.55, of which 39.81 was obtained by the Parke algorithm after 3 iterations. On iteration 26, eight coefficients changed by 1.0 percent or more, and the largest three changes were 6.6, 10.5, and 26.7 percent.

Each iteration requires about 4 minutes on the IBM 4341 and about 11 minutes on the VAX. The total time for the 26 iterations on the IBM 4341 was about 1.7 hours. The D matrix for the US model is $3{,}450 \times 3{,}450$ ($m = 30$, $T = 115$), and considerable time was saved by not computing this matrix from scratch any more times than were absolutely necessary. If the entire matrix had been computed each time that (6.24) was computed, the time per iteration would have increased by about a factor of 17, and thus the total time would have increased from 1.7 hours to 28.9 hours.

The time taken to compute the 3SLS covariance matrix, $\hat{V}_3$ in (6.25), was about 23 minutes on the IBM 4341 and about 11 minutes on the VAX. The derivative matrix $\hat{G}$ that is needed for this calculation was computed numerically. The reason the IBM 4341 time is large relative to the VAX time is that in the calculation of $\hat{V}_3$ much reading and writing from the disk is done, and the IBM 4341 is relatively slow at this.

6.5.4 LAD and 2SLAD

The LAD and 2SLAD computational problem is to minimize

$$(6.53) \qquad \sum_{t=1}^{T} |v_{it}|$$

with respect to α_i, where $v_{it} = u_{it} = y_{it} - h_{it}$ for LAD and $v_{it} = qy_{it} + (1-q)\hat{y}_{it} - \hat{h}_{it}$ for 2SLAD. This computational problem is not particularly easy, especially when v_{it} is a nonlinear function of α_i. I have had no success in trying to minimize (6.53) using the DFP algorithm and Powell's no-derivative algorithm (1964). (When the DFP algorithm was tried, the derivatives were computed numerically. The problem that they do not exist everywhere was ignored.) Both algorithms failed to get close to the optimum in most of the cases that I tried.

Because the standard algorithms do not work, other approaches must be tried. I have used two, one that worked well and one that did not. The one that worked well uses the fact that

$$\sum_{t=1}^{T} |v_{it}| = \sum_{t=1}^{T} \frac{v_{it}^2}{|v_{it}|} = \sum_{t=1}^{T} \frac{v_{it}^2}{w_{it}}, \tag{6.54}$$

where $w_{it} = |v_{it}|$. For a given set of values of w_{it} ($t = 1, \ldots, T$), minimizing (6.54) is simply a weighted least squares problem. If v_{it} is a linear function of α_i, closed-form expressions exist for $\hat{\alpha}_i$; otherwise a nonlinear optimization algorithm can be used. This suggests the following iterative procedure. (1) Pick an initial set of values of w_{it}. These can be the absolute values of the OLS or 2SLS estimated residuals. (2) Given these values, minimize (6.54). (3) Given the estimate of α_i from step 2, compute new values of v_{it} and thus new values of w_{it}. (4) With the new weights, go back to step 2 and minimize (6.54) again. Keep repeating steps 2 and 3 until successive estimates of α_i are within some prescribed tolerance level. If on any step some value of w_{it} is smaller than some small preassigned number (say ϵ), the value of w_{it} should be set equal to ϵ.

The accuracy of the estimates using this approach is a function of ϵ: the smaller is ϵ, the greater is the accuracy. If v_{it} is a linear function of α_i, the estimates will never be exact because the true estimates correspond to k_i values of w_{it} being exactly zero, where k_i is the number of elements of α_i.

In the case in which the equation to be estimated is linear in coefficients, the closed-form expression for $\hat{\alpha}_i$ for a given set of values of w_{it} is

$$\hat{\alpha}_i = (\hat{X}_i^{*\prime} \hat{X}_i^{*})^{-1} \hat{X}_i^{*\prime} \hat{y}_i^{*}. \tag{6.55}$$

$\hat{X}_i^{*}$ is the same as $\hat{X}_i$ in (6.9) except that each element in row t of $\hat{X}_i$ is divided by $\sqrt{w_{it}}$. The vector $\hat{y}_i^{*}$ equals $qy_i + (1 - q)\hat{y}_i$ except that row t is divided by $\sqrt{w_{it}}$. ($\hat{y}_i$ equals $D_i y_i$.)

If the equation is linear in coefficients but has serially correlated errors, v_{it} is not a linear function of the coefficients inclusive of the serial correlation coefficients, and therefore a closed-form expression does not exist. It is possible in his case, however, to solve for the estimates by iteratively solving equations like (6.13) and (6.14). This avoids having to use a general-purpose algorithm like DFP. Assuming that X_{i-1} and y_{i-1} are included in Z_i, the two equations for the first-order serial correlation case are

$$\hat{\alpha}_i = (\hat{X}_i^{**\prime} \hat{X}_i^{**})^{-1} \hat{X}_i^{**\prime} \hat{y}_i^{**}, \tag{6.56}$$

$$\hat{\rho}_i = \frac{\hat{u}_{i-1}^{*\prime} \hat{u}_i^{*}}{\hat{u}_{i-1}^{*\prime} \hat{u}_{i-1}^{*}}. \tag{6.57}$$

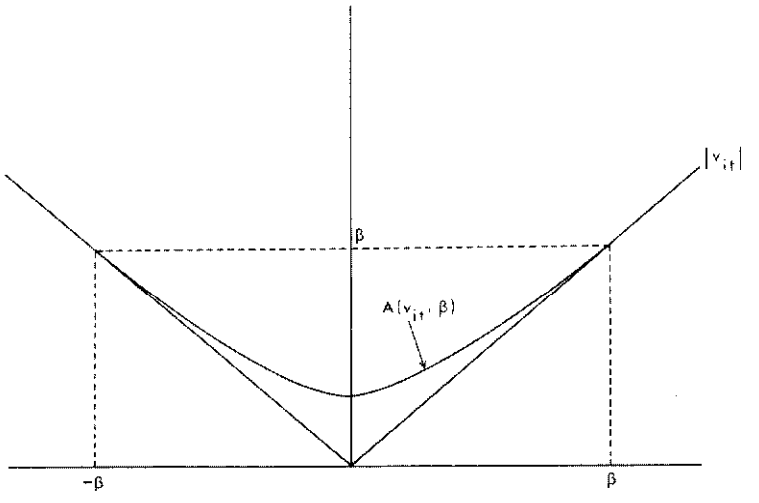

Figure 6-1 Approximation of $A(v_{it}, \beta)$ to $|v_{it}|$

$\hat{X}_i^{**}$ is the matrix $\hat{X}_i - X_{i-1}\hat{\rho}_i$ with each element in row t divided by $\sqrt{w_{it}}$; $\hat{y}_i^{**}$ is the vector $qy_i + (1-q)\hat{y}_i - y_{i-1}\hat{\rho}_i$ with row t divided by $\sqrt{w_{it}}$; $\hat{u}_{i-1}^{*}$ is the vector $y_{i-1} - X_{i-1}\hat{\alpha}_i$ with row t divided by $\sqrt{w_{it}}$; and $\hat{u}_i^{*}$ is the vector $qy_i + (1-q)\hat{y}_i - X_i\hat{\alpha}_i$ with row t divided by $\sqrt{w_{it}}$. For a given set of weights, (6.56) and (6.57) can be solved iteratively.

The second approach is derived from Tishler and Zang (1980). The problem of minimizing (6.53) is changed to a problem of minimizing

$$\sum_{t=1}^{T} A(v_{it}, \beta), \tag{6.58}$$

where

$$A(v_{it}, \beta) = \begin{cases} -v_{it} & \text{if } v_{it} \leqq -\beta \\ (v_{it}^2 + \beta^2)/2\beta & \text{if } -\beta < v_{it} < \beta \\ v_{it} & \text{if } v_{it} \geqq \beta \end{cases}. \tag{6.59}$$

The value of β is some small preassigned number. Since $\lim_{\beta \to 0} A(v_{it}, \beta) = |v_{it}|$, the smaller is β, the closer is (6.53) to (6.59). The approximation of $A(v_{it}, \beta)$ to $|v_{it}|$ is presented in Figure 6-1. Since $A(v_{it}, \beta)$ is once continuously differentiable, an optimization algorithm like DFP can be used to minimize (6.59) for a given value of β. The smaller is β, the more difficult the minimization problem is likely to be, and thus there is a trade-off between accuracy and ease of solution.

Results for the US Model

Four sets of estimates of the US model were obtained: LAD, 2SLAD using $q = 0.0$, 2SLAD using $q = 0.5$, and 2SLAD using $q = 1.0$. The method of Tishler and Zang did not work well, in the sense that the results were quite sensitive to the value of β chosen, and therefore it was dropped from further consideration fairly early in the calculations. For small values of β the DFP algorithm, which was the algorithm used, failed to converge, and for large values of β the algorithm converged to answers that implied values of the true objective function, (6.53), that were larger than those obtained by the first method. It was difficult to find in-between values of β that worked well.

The first method, on the other hand, worked extremely well. For 2SLAD using $q = 0.5$, for example, the number of iterations required for convergence for the 30 equations ranged from 4 to 145, with an average of 35.6. Convergence was taken to be achieved when successive estimates of each coefficient were within .002 percent of each other. The value used for ϵ was .0000001. The total time for estimating the model by LAD was about 2.2 minutes on the IBM 4341 and about 5.7 minutes on the VAX. The total time for each of the three 2SLAD estimation problems was about 6.5 minutes on the IBM 4341 and about 16.5 minutes on the VAX. Of the 120 equations estimated, none had a residual that was smaller than ϵ in absolute value at the time that convergence was achieved. These results are very encouraging, and they indicate that computational costs are not likely to be a serious problem in the future with respect to LAD and 2SLAD estimation.

6.6 Comparison of the OLS, 2SLS, 3SLS, FIML, LAD, and 2SLAD Results for the US Model

If the model is correctly specified and all the assumptions about the error terms are correct, all but the OLS and LAD estimates of the US model are consistent. They should thus differ from each other only because of a finite sample size. In practice the model is likely to be misspecified, and not all the assumptions about the error terms are likely to be correct. Given this, it is not obvious how the estimates should compare. In this section the quantitative differences among the estimates are examined. The consequences of these differences for the predictive accuracy of the model are discussed in Section 8.5.5, and the consequences for the properties of the model are discussed in Section 9.4.5.

Table 6-5 presents a comparison of the estimates for six equations: the three consumption equations, 1, 2, and 3; the price equation, 10; the production

TABLE 6-5. Comparison of coefficient estimates for selected equations of the US model

Eq. no.	Coeff. no.	2SLS a	FIML a	FIML b	3SLS a	3SLS b	2SLAD (q = 0.0) a	2SLAD (q = 0.0) b	2SLAD (q = 0.5) a	2SLAD (q = 0.5) b	2SLAD (q = 1.0) a	2SLAD (q = 1.0) b	LAD a	LAD b	OLS a	OLS b
1	1	.00019	-.00240	-0.80	-.00042	-0.19	-.00040	-0.18	-.00161	-0.56	.00123	0.32	-.00008	-0.08	.00154	0.42
	2	.98650	.99893	0.77	.99021	0.23	.98920	0.17	.99260	0.38	.97983	-0.42	.98125	-0.33	.98786	0.08
	3	.00055	.00066	0.47	.00058	0.12	.00057	0.07	.00063	0.31	.00062	0.28	.00058	0.12	.00039	-0.72
	4	.01979	.02922	0.99	.02143	0.17	.02051	0.07	.02094	0.12	.02052	0.08	.01183	-0.83	.00936	-1.09
	5	.00714	-.02427	-1.58	-.00118	-0.42	.00265	-0.23	-.00257	-0.49	.01021	0.15	.01490	0.39	.01277	0.28
	6	-.00126	-.00111	0.73	-.00124	0.12	-.00126	0.00	-.00123	0.16	-.00116	0.48	-.00082	2.08	-.00088	1.78
	7	.02312	.01107	-1.00	.01852	-0.38	.02094	-0.18	.01913	-0.33	.02280	-0.03	.00864	-1.21	.01969	-0.29
2	1	.10903	.23997	4.76	.11422	0.19	.10362	-0.20	.08105	-1.02	.07983	-1.06	.11491	0.21	.13478	0.94
	2	.66619	.41164	-3.83	.65620	-0.15	.67928	0.20	.71909	0.80	.73586	1.05	.65734	-0.13	.61356	-0.79
	3	.00227	.00181	-1.04	.00220	-0.15	.00231	0.08	.00235	0.17	.00197	-0.67	.00208	-0.44	.00218	-0.19
	4	.18547	.58420	5.34	.18727	0.02	.19026	0.06	.08513	-1.34	.08915	-1.29	.20575	0.27	.25481	0.93
	5	-.04689	-.16405	-5.40	-.04951	-0.12	-.04648	0.02	-.02047	1.22	-.02517	1.00	-.05775	-0.58	-.06867	-1.00
	6	.06369	.02956	-1.15	.06952	0.20	.05044	-0.45	.08090	0.58	.08117	0.59	.07367	0.34	.06468	0.03
	7	-.00061	.00207	4.59	-.00038	0.39	-.00057	0.06	-.00091	-0.51	-.00054	0.13	-.00005	0.96	-.00004	0.97
	8	.08291	.11018	1.17	.08212	-0.03	.06744	-0.66	.07383	-0.39	.06047	-0.96	.05669	-1.12	.08567	0.12
3	1	.07349	.20807	6.53	.07710	0.18	.05432	-0.93	.06464	-0.43	.07328	-0.01	.06771	-0.28	.06016	-0.65
	2	.45821	.07423	-4.98	.44448	-0.18	.48225	0.31	.51434	0.73	.48472	0.34	.53824	1.04	.49524	0.48
	3	.00235	.00247	0.32	.00222	-0.35	.00238	0.07	.00211	-0.65	.00211	-0.63	.00187	-1.28	.00221	-0.39
	4	.40468	1.03962	6.39	.39950	-0.05	.30037	-1.05	.35109	-0.54	.37741	-0.27	.34360	-0.62	.31159	-0.94
	5	-.10399	-.32751	-6.71	-.10491	-0.03	-.06856	1.06	-.08905	0.45	-.10485	-0.03	-.09862	0.16	-.07811	0.78
	6	.06682	-.11739	-3.27	.08347	0.30	.13503	1.21	.07585	0.16	.08615	0.34	.08378	0.30	.11408	0.84
	7	-.00617	-.00749	-1.70	-.00608	0.12	-.00602	0.19	-.00570	0.60	-.00519	1.27	-.00447	2.20	-.00537	1.04
	8	.12315	.17590	1.45	.12353	0.01	.13739	0.39	.11844	-0.13	.12722	0.11	.11679	-0.17	.13968	0.45
10	1	.18683	.19928	0.49	.18257	-0.17	.16517	-0.85	.18921	0.09	.20110	0.56	.20672	0.78	.18718	0.01
	2	.92214	.90850	-1.22	.91983	-0.21	.93113	0.81	.92353	0.12	.91721	-0.44	.91457	-0.68	.92200	-0.01
	3	.03394	.03672	0.57	.03326	-0.14	.02984	-0.84	.03438	0.09	.03665	0.55	.03771	0.77	.03401	0.01
	4	.03389	.04079	1.74	.03650	0.66	.03192	-0.50	.03210	-0.45	.03482	0.23	.03538	0.38	.03392	0.01
	5	-.08096	-.07402	0.36	-.08966	-0.45	-.07814	0.15	-.07365	0.38	-.07883	0.11	-.08155	-0.03	-.08094	0.00
11	1	11.36381	21.87884	4.04	10.69493	-0.26	8.12354	-1.24	11.64381	0.11	8.81205	-0.98	9.25990	-0.81	10.58937	-0.30
	2	.16209	-.03324	-4.43	.15484	-0.16	.15023	-0.27	.14886	-0.30	.18034	0.41	.17040	0.19	.18002	0.41
	3	1.01142	1.43204	8.15	1.01595	0.09	.98842	-0.45	1.03510	0.46	.97080	-0.79	.98684	-0.48	.98039	-0.60
	4	-.19265	-.43424	-5.57	-.18766	0.11	-.14986	0.99	-.20464	-0.28	-.16538	0.63	-.17028	0.52	-.17797	0.34
	8	.60491	.76119	1.74	.56992	-0.39	.59695	-0.09	.61413	0.10	.60371	-0.01	.58844	-0.18	.58023	-0.27
30	1	-9.45741	-5.60570	1.22	-7.66028	0.57	-9.52105	-0.02	-7.29281	0.68	-8.18864	0.40	-7.91120	0.49	-9.80375	-0.11
	2	.85812	.90573	1.42	.88586	0.83	.83247	-0.76	.89576	1.12	.93716	2.35	.92674	2.04	.86039	0.07
	3	.06872	-.01235	-2.49	.03783	-0.95	.06980	0.03	.05541	-0.47	.02104	-1.46	.03465	-1.05	.06709	-0.05
	4	.02962	.01761	-1.21	.02389	-0.58	.03043	0.08	.02282	-0.69	.02533	-0.43	.02446	-0.52	.03065	0.10
	5	.05974	.06527	0.27	.06069	0.05	.03105	-1.40	.04213	-0.86	.03831	-1.05	.03532	-1.19	.06343	0.18
	6	.03248	.05675	1.28	.04356	0.58	.03640	0.21	.02722	-0.28	.03938	0.36	.04258	0.53	.03166	-0.04
	7	.13149	.11772	-0.44	.10735	-0.77	.14851	0.54	.15053	0.61	.03880	-2.96	.03972	-2.93	.13201	0.02

Notes: a. Coefficient estimate; b. (Coefficient estimate - 2SLS coefficient estimate)/standard error of 2SLS coefficient estimate.

equation, 11; and the interest rate reaction function, 30. The 2SLS estimates are used as the basis of comparison. Each number in a "b" column in the table is the difference between the particular estimate and the 2SLS estimate divided by the standard error of the 2SLS estimate. These numbers thus indicate how many standard errors the estimates are from the 2SLS estimates, where the standard errors that are used are 2SLS standard errors. Table 6-6 provides summary measures for all the coefficient estimates.

The main conclusion to be drawn from these results is that all the estimates are fairly close to each other except for the FIML estimates. Consider Table 6-6: only 3 of the 107 3SLS coefficient estimates are more than 1.5 standard errors away from the 2SLS estimates, whereas 38 of the FIML estimates are. Only 1 of the 169 OLS estimates is more than 1.5 standard errors away. Of the 2SLAD estimates, 7 are more than 1.5 standard errors away for $q = 0.0$, 12 are for $q = 0.5$, and 19 are for $q = 1.0$. For LAD the number is 15. Very few of the estimates changed signs, as can be seen in the bottom half of Table 6-6. Even for FIML, only 6 estimates changed sign.

With respect to the individual estimates in Table 6-5, one important difference between the FIML estimates and the others occurs in Eq. 11, the equation determining production, Y. Coefficient 3 in Eq. 11 is the coefficient for the sales variable, X. For all the estimates except FIML, this coefficient is around 1.0, whereas for FIML it is around 1.4. Also, coefficient 2 in Eq. 11, which is the coefficient of the lagged dependent variable, is around .15 for the other estimates and close to zero for FIML. The FIML estimates of the lagged dependent variable coefficients in two of the three consumption equations (Eqs. 2 and 3) are likewise quite different from the others. In both equations the lagged dependent variable coefficient is number 2. The FIML and 2SLS estimates in the two equations are, respectively, .66619 versus .41164 and .45821 versus .07423.

It should be stressed that the only reason for the present comparison is to get a general idea of how close the estimates are. Of more importance are the comparisons in Sections 8.5.5 and 9.4.5, which examine the estimates within the context of the overall model. What can be said so far is that the FIML estimates differ most from the others when the examination is coefficient by coefficient.

Comparison of Standard Errors

Table 6-7 presents a comparison of the 2SLS, 3SLS, and FIML estimated standard errors. As expected, the 2SLS standard errors are generally larger

TABLE 6-6. Comparison of coefficient estimates of the US model

	Number of coefficient estimates greater than .5, 1.0, 1.5, 2.0 2.5, and 3.0 standard errors away from the 2SLS estimates					
	.5	1.0	1.5	2.0	2.5	3.0
107 total coefficients:						
FIML	81	63	38	25	16	16
3SLS	34	8	3	2	0	0
169 total coefficients:						
2SLAD (q = 0.0)	64	21	7	5	3	1
2SLAD (q = 0.5)	77	33	12	8	3	1
2SLAD (q = 1.0)	98	53	19	12	6	3
LAD	91	40	15	11	4	1
OLS	28	9	1	0	0	0

	Number of sign changes from 2SLS estimates other than those for constant terms
FIML	6
3SLS	2
2SLAD (q = 0.0)	1
2SLAD (q = 0.5)	2
2SLAD (q = 1.0)	1
LAD	1
OLS	2

than the 3SLS standard errors, where the average of the ratios of the two is 1.27. This is not always the case, however, as can be seen for coefficients 1–6 and 8 in Eq. 4, where the 2SLS standard errors are smaller. This difference is due to the different first-stage regressors that are used by 2SLS and 3SLS. As discussed earlier, 2SLS uses different sets of FSRs for different equations, whereas 3SLS uses a common set that is smaller than the union of the 2SLS sets. This can cause the 2SLS standard errors to be smaller. In the present case, Eq. 4 has no RHS endogenous variables, and thus the 2SLS estimates are the OLS estimates. The FSRs in this case include all the explanatory variables in the equation. Not all of these explanatory variables were included in the common set of FSRs for the 3SLS estimates, and therefore some of the variables in the equation were treated as endogenous. This was enough to lead to larger 3SLS standard errors for some of the coefficients.

TABLE 6-7. Ratios of 3SLS and FIML standard errors and of 2SLS and 3SLS standard errors for the US model

Eq. no.	Coeff. no.	$\frac{SE_3}{SE_4}$	$\frac{SE_2}{SE_3}$	Eq. no.	Coeff. no.	$\frac{SE_3}{SE_4}$	$\frac{SE_2}{SE_3}$	Eq. no	Coeff. no.	$\frac{SE_3}{SE_4}$	$\frac{SE_2}{SE_3}$
1	1	.86	1.20	10	1	.75	1.17	17	1	.82	1.16
	2	.78	1.21		2	.72	1.19		2	.77	1.20
	3	.69	1.17		3	.74	1.17		3	.79	1.19
	4	.72	1.19		4	.75	1.18		4	.75	1.20
	5	.82	1.22		5	.75	1.15	22	1	1.03	1.18
	6	.77	1.23	11	1	.32	1.25		2	.67	1.37
	7	.85	1.19		2	.52	1.22		3	.68	1.20
2	1	.63	1.28		3	.40	1.21	23	1	.79	1.12
	2	.68	1.25		4	.27	1.24		2	.70	1.17
	3	.71	1.17		8	.78	1.18		3	.42	1.42
	4	.56	1.30	12	1	.91	1.10		4	.53	1.34
	5	.57	1.32		2	.77	1.22		5	.70	1.22
	6	.79	1.20		3	.66	1.22	24	1	.68	1.18
	7	.65	1.34		4	.74	1.25		2	.65	1.22
	8	.84	1.20		5	.85	1.27		3	.36	1.45
3	1	.26	1.39		6	.82	1.28		4	.49	1.39
	2	.28	1.34		7	.77	1.21		5	.65	1.26
	3	.48	1.25		8	.77	1.21	26	1	.76	1.15
	4	.24	1.41	13	1	.86	1.78		2	.81	1.13
	5	.25	1.40		2	.86	1.78		3	.76	1.15
	6	.43	1.38		3	.87	1.72		4	.82	1.13
	7	.33	1.29		4	.65	1.58		5	.74	1.16
	8	.61	1.33		5	.86	1.66	27	1	.85	1.37
4	1	1.47	.75		6	.83	1.56		2	.87	1.39
	2	1.26	.91		9	.78	1.57		3	.85	1.39
	3	1.09	.95	14	1	.75	1.38		4	.72	1.29
	4	1.32	.87		2	.75	1.45		5	.68	1.28
	5	1.32	.91		3	.77	1.53		6	.67	1.27
	6	1.43	.85		4	.75	1.38	30	1	.82	1.19
	7	1.03	1.17		5	.63	1.51		2	.80	1.22
	8	.99	.96	16	1	.68	1.40		3	.80	1.26
	9	.77	1.40		2	.69	1.41		4	.83	1.19
9	1	.78	1.21		3	.72	1.62		5	.76	1.32
	2	.81	1.20		4	.67	1.36		6	.75	1.29
	3	.77	1.24		5	.81	1.36		7	.73	1.27
	4	.82	1.19								
	5	.72	1.19						AVERAGE	.74	1.27

The more interesting result in Table 6-7 is that the 3SLS standard errors are generally smaller than the FIML standard errors. The average of the ratios of the two is .74. This result has also been obtained, but not discussed, by Hausman (1974). For 10 of the 12 estimated coefficients of Klein's model I that are reported in Hausman's table 1, p. 649, the FIML standard error is larger than the corresponding 3SLS standard error.

My conjecture as to why the 3SLS standard errors are generally smaller is the following. Given the large number of FSRs that are used by 3SLS, the predicted values of the endogenous variables from the first-stage regressions are fairly close to the actual values. For FIML, on the other hand, we know from Hausman's interpretation (1975) of the FIML estimator as an instru-

mental variables estimator that FIML takes into account the nonlinear restrictions on the reduced form coefficients in forming the instruments. This means that in small samples the instruments that FIML forms are likely to be based on worse first-stage fits of the endogenous variables than are the instruments that 3SLS forms. In a loose sense, this situation is analogous to the fact that in the 2SLS case the more variables that are used in the first-stage regressions, the better is the fit in the second-stage regression.

Possible Use of the Hausman Test

An interesting question is whether Hausman's m-statistic (1978) provides a useful way of examining the differences among the estimates. The m-statistic is as follows. Consider two estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$, where under some null hypothesis both estimators are consistent but only $\hat{\beta}_0$ is asymptotically efficient, while under the alternative hypothesis only $\hat{\beta}_1$ is consistent. Let $q = \hat{\beta}_1 - \hat{\beta}_0$, and let $\hat{V}_0$ and $\hat{V}_1$ denote consistent estimates of the asymptotic covariance matrices (V_0 and V_1) of $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. Hausman's m-statistic is $\hat{q}'(\hat{V}_1 - \hat{V}_0)^{-1}\hat{q}$, and he has shown that it is asymptotically distributed as χ^2 with k degrees of freedom, where k is the dimension of $\hat{q}$. Note that under the null hypothesis $V_1 - V_0$ is positive-definite.

Consider now comparing the FIML and 3SLS estimates. Under the null hypothesis of correct specification and normally distributed errors, both estimates are consistent, but only the FIML estimates are asymptotically efficient. On the other hand, 3SLS estimates are consistent for a broad class of error distributions, whereas for many distributions FIML estimates are inconsistent. If the alternative hypothesis is taken to be that the error distribution is one that leads to consistent 3SLS estimates but inconsistent FIML estimates, then in principle Hausman's m-statistic can be used to test the null hypothesis of normality against the alternative. Let $\hat{\alpha}^{(3)}$ and $\hat{\alpha}^{(4)}$ denote the 3SLS and FIML estimates of α respectively, and let $\hat{q} = \hat{\alpha}^{(3)} - \hat{\alpha}^{(4)}$. The m-statistic in this case is $\hat{q}'(\hat{V}_3 - \hat{V}_4)^{-1}\hat{q}$, where the estimated covariance matrices $\hat{V}_3$ and $\hat{V}_4$ are defined in (6.25) and (6.34) respectively.

In practice the test cannot be performed if $\hat{V}_3 - \hat{V}_4$ is not positive-definite. For the US model it is clear from Table 6-7 that $\hat{V}_3 - \hat{V}_4$ is not positive-definite, since most of the diagonal elements of $\hat{V}_3$ are smaller than the corresponding elements of $\hat{V}_4$. If anything, $\hat{V}_3 - \hat{V}_4$ is closer to being negative-definite, although this is not true either since some of the diagonal elements of $\hat{V}_4$ are smaller than the corresponding elements of $\hat{V}_3$. The matrix $\hat{V}_3 - \hat{V}_4$ is also not positive-definite for Klein's model I, since, as noted earlier, Hausman's

results (1974) show that 10 of the 12 estimated coefficients have larger FIML standard errors than 3SLS standard errors. It thus seems unlikely that $\hat{V}_3 - \hat{V}_4$ will be positive-definite in practice for most models, and therefore the m-statistic is not likely to be useful for testing the normality hypothesis. (If the model is linear, the test obviously has no power, since FIML, like 3SLS, is consistent for a broad class of error distributions.)

The m-statistic can also be used in principle to compare the FIML and 2SLS estimates. Under the null hypothesis of normally distributed errors and correct specification, both estimates are consistent, but only the FIML estimates are asymptotically efficient. Under the alternative hypothesis of normality and misspecification of some subset of the equations, all the FIML estimates are inconsistent, but only the 2SLS estimates of the misspecified subset are inconsistent. The m-statistic can thus be applied to one or more equations at a time to test the hypothesis that the rest of the model is correctly specified. If for some subset the m-statistic exceeds the critical value, the test would indicate that there is misspecification somewhere in the rest of the model.

In practice this test cannot be applied if $\hat{V}_2 - \hat{V}_4$ is not positive-definite, and for the US model, as is clear from Table 6-7, $\hat{V}_2 - \hat{V}_4$ is not positive-definite. Many of the diagonal elements of $\hat{V}_2$ are smaller than the corresponding elements of $\hat{V}_4$. It thus also seems unlikely that this test of misspecification will be useful in practice.

Finally, the specification hypothesis can be tested in certain circumstances using the m-statistic on the 2SLS and 3SLS estimates. If both estimators are members of a class of estimators for which 3SLS is asymptotically efficient, the test can be applied. The problem is that when the two estimators are based on different sets of FSRs, as is usually the case with large models, they are not members of the same class. One cannot argue, for example, that the 3SLS estimates given above for the US model are asymptotically efficient relative to the 2SLS estimates, and thus the Hausman test cannot be applied in this case.

In summary, the m-statistic does not seem useful for testing either the normality hypothesis or the correct specification hypothesis. Regarding the latter, my feeling is that it is better simply to assume that the model is misspecified (so that no test is needed) and to try to estimate the degree of misspecification. This is the procedure followed for the comparison method in Chapter 8.