SYSTEMS OF REGRESSION EQUATIONS

1. MULTIPLE EQUATIONS

$$y_{nt} = x_{nt}\beta_n + u_{nt},$$

$n = 1,...,N$, $t = 1,...,T$, x_{nt} is $1{\times}k$, and β_n is $k{\times}1$. This is a version of the standard regression model where the observations are indexed by the two indices n and t rather than by a single index. Applications are

- ◼ n indexes equations, with different dependent variables, and t indexes observation units. Example: $y_{1t},...,y_{nt}$ are the input demands of firm t. In this example, there are likely to be parameters in common across equations.
- ◼ n indexes observation units, t indexes time, and the data come from a time-series of cross-sections. Example: y_{nt} is the income of household n in the Census Public Use Sample in year t.
- ◼ n indexes observation units, t indexes time, and the data come from a longitudinal <u>panel</u> of time series observations on each observation unit. Examples: y_{nt} is hours supplied by the head of household n in year t in the Panel Study of Income Dynamics; <u>or</u> y_{nt} is the excess return on stock market asset n on day t in the CRISP financial database.

These problems may contain the usual litany of econometric problems: (1) a non-scalar covariance matrix due to heteroskedasticity across observation units, serial correlation over time, or covariance across equations within an observation unit; and (2) the potential for correlation of explanatory variables and disturbances when x includes lagged dependent variables. They also provide an opportunity for a richer analysis of covariance patterns, since observations across units can be used to identify covariance patterns over time, and observations across time can be used to identify heteroskedasticities across units.

2. STACKING THE DATA

For analysis (and computation), it is useful to organize the observations in vectors in which all the observations for n = 1 are stacked on top of all the observations for n = 2, etc. Use the notation:

$$
\mathbf{y_n} = \begin{bmatrix} y_{n1} \\ y_{n2} \\ \vdots \\ y_{nT} \end{bmatrix}, \mathbf{X_n} = \begin{bmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{nT} \end{bmatrix}, \mathbf{u_n} = \begin{bmatrix} u_{n1} \\ u_{n2} \\ \vdots \\ u_{nT} \end{bmatrix},
$$

$$
\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1T} \\ \vdots \\ y_{N1} \\ \vdots \\ y_{NT} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & X_N \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & 0 \\ \vdots & \cdots & \vdots \\ x_{1T} & \cdots & 0 \\ \vdots & \cdots & \vdots \\ 0 & \cdots & x_{N1} \\ \vdots & \cdots & \vdots \\ 0 & \cdots & x_{NT} \end{bmatrix} ,
$$

$$
\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} , \; \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}
$$

Then, the system can be written

$$ \mathbf{y}_n = \mathbf{x}_n \beta_n + \mathbf{u}_n \; , \; n = 1,...,N $$

or in stacked form,

(1) $$ \mathbf{y} = \mathbf{X}\beta + \mathbf{u} \; . $$

The vector y_n is of dimension T×1, the array X_n is of dimension T×k, the vector y is of dimension NT×1, the array X is of dimension NT×Nk. We wrote down the system assuming the number of parameters k was the

same in each equation, but this is not necessary. One could have X_n of dimension $T \times k_n$ and X of dimension $NT \times (k_1 + .. + k_n)$. If there are parameters in common across different equations, then the corresponding explanatory variables will be stacked in the same column rather than placed in different columns, and the overall number of columns in X reduced accordingly.

Suppose the observations are independent and identically distributed for different t, but the covariances $E(u_{nt}u_{mt}) = \sigma_{nm}$ are not necessarily zero. Let $\Sigma = (\sigma_{nm})$ be the $N \times N$ array of covariances of the observations for each t. The covariance matrix of the stacked disturbance vector u is then

$$
E(uu') = \begin{bmatrix} \sigma_{11}I_T & \sigma_{12}I_T & ... & \sigma_{1N}I_T \\ \sigma_{21}I_T & \sigma_{22}I_T & ... & \sigma_{2N}I_T \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{N1}I_T & \sigma_{N2}I_T & \cdots & \sigma_{NN}I_T \end{bmatrix},
$$

where I_T denotes a $T \times T$ identity matrix.

Define the <u>**Kronecker Product**</u> $A \otimes B$ of a $n \times m$ matrix A and a $p \times q$ matrix B:

$$
A \otimes B =
\begin{bmatrix}
a_{11}B & a_{12}B & \dots & a_{1m}B \\
a_{21}B & a_{22}B & \dots & a_{2m}B \\
\vdots & \vdots & \vdots & \vdots \\
a_{n1}B & a_{n2}B & \cdots & a_{nm}B
\end{bmatrix}.
$$

Then, $A \otimes B$ is $(np) \times (mq)$. Kronecker products have the following properties:

$(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ when the matrices are commensurate

$(A \otimes B)^{-1} = (A^{-1}) \otimes (B^{-1})$ when A and B are square and nonsingular

$(A \otimes B)' = (A') \otimes (B')$

$\text{trace}(A \otimes B) = (\text{trace}(A)) \cdot (\text{trace}(B))$ when A and B are square

$\det(A \otimes B) = (\det(A))^p (\det(B))^n$ when A is $n \times n$ and B is $p \times p$

Applying the Kronecker product notation to the covariance matrix of u, $E(uu') = \Sigma \otimes I_T$.

3. ESTIMATION

The problem of estimating the stacked model $y = X\beta + u$ when the covariance matrix of the disturbances is

$\Sigma \otimes I_T$ and Σ is known is a straightforward GLS problem, provided there are no additional complications of correlation of explanatory variables and disturbances. Using the rule for inverses of Kronecker products, the GLS estimator is

$$b = (X'(\Sigma^{-1} \otimes I_T)X)^{-1}X'(\Sigma^{-1} \otimes I_T)y \ .$$

Computationally, the most practical way to do this regression is to calculate a triangular Cholesky matrix L such that $L'L = \Sigma^{-1}$. Then, the transformed model

(2) $\qquad (L \otimes I_T)y = (L \otimes I_T)X\beta + (L \otimes I_T)u$

satisfies Gauss-Markov conditions (Verify), and the BLUE estimator of β is OLS applied to this equation. The data transformations can be carried out separately for each t, and recursively for n = 1,...,N.

When Σ is unknown, one can do FGLS estimation: First apply OLS to (1) and retrieve fitted residuals û. Then, estimate the elements σ_{nm} of Σ from the average (over T) of the squares and cross-products of the fitted residuals,

$$s_{nm} = \frac{1}{T} \sum_{t=1}^{T} \hat{u}_{nt}\hat{u}_{mt} \ .$$

Finally, apply OLS to (2), with L a Cholesky factor of the estimated Σ^{-1}.

The problem of estimating β in (1) when there are <u>no</u> cross-equation restrictions on the β_n is called the <u>seemingly</u> <u>unrelated</u> <u>regressions</u> <u>problem</u>. Summarizing, the β_n can be estimated consistently equation-by-equation using OLS; in most cases, this is inefficient compared to GLS; and FGLS is asymptotically fully efficient. There is one case in which there is no efficiency gain from use of GLS rather than OLS: Suppose no cross-equation restrictions on parameters and <u>common</u> explanatory variables across equations; i.e., $X_1 = X_2 = ... = X_N$. Then, $X = I_N \otimes X_1$, and the GLS estimator is

$$b = ((I_N \otimes X_1')(\Sigma^{-1} \otimes I_T)(I_N \otimes X_1))^{-1}(I_N \otimes X_1')(\Sigma^{-1} \otimes I_T)y \ .$$

As an exercise, use the Kronecker product rules to show that this formula reduces to the OLS estimator $b_n = (X_1'X_1)^{-1}X_1'y_n$ for each n. Intuitively, the reason OLS is efficient in this case is that the OLS residuals in, say, the first equation are automatically orthogonal to the (common) exogenous variables in each of the other equations, so that there is no additional information on the first equation parameters to be distilled from the cross-equation orthogonality conditions. Put another way, GLS can be interpreted as OLS applied to linear combinations of the original equations, with the linear combinations

obtained from the Cholesky factorization of the covariance matrix of the disturbances. But these linear combinations of the common exogenous variables leaves one with the same exogenous variables, and the orthogonality conditions satisfied by the GLS estimates are the same as the orthogonality conditions satisfied by OLS on the first equation in the original system.

4. AN EXAMPLE

Suppose a firm t utilizes N = 3 inputs, and has a Diewert unit cost function,

$$C_t = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{ij} \sqrt{p_{it} p_{jt}} \quad,$$

where p_{it} is input i price, and the α's are nonnegative parameters with $\alpha_{ij} = \alpha_{ji}$. By Shephard's lemma, the unit input demand functions are given by the derivatives of the unit cost function with respect to the input prices:

$$z_{nt} = \sum_{j=1}^{N} \alpha_{nj} \sqrt{p_{jt}/p_{nt}} \quad.$$

Written in stacked form, these equations become

$$
\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix} =
$$

$$
\begin{bmatrix} 1_T & (\sqrt{p_2/p_1})_T & (\sqrt{p_3/p_1})_T & 0_T & 0_T & 0_T \\ 0_T & (\sqrt{p_1/p_2})_T & 0_T & 1_T & (\sqrt{p_3/p_2})_T & 0_T \\ 0_T & 0_T & (\sqrt{p_1/p_3})_T & 0_T & (\sqrt{p_2/p_3})_T & 1_T \end{bmatrix}
\begin{bmatrix} \alpha_{11} \\ \alpha_{12} \\ \alpha_{13} \\ \alpha_{22} \\ \alpha_{23} \\ \alpha_{33} \end{bmatrix} +
$$

$$
\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix},
$$

where 1_T denotes a T×1 vector of 1's and $\left(\sqrt{p_1/p_2}\right)_T$ denotes

a T×1 vector with components $\sqrt{p_{1t}/p_{2t}}$. Note that the

parameter restrictions across equations lead to variables
appearing stacked in the same column. The disturbances can
be interpreted as coming from random variations across firms

around the respective "average" parameters α_{11}, α_{22}, α_{33}. The interesting econometric feature of this setup is that even if there is considerable multicollinearity in prices so that OLS equation by equation is imprecise, this multicollinearity is broken when the data are stacked. Then, there is likely to be a substantial efficiency gain from estimating the equations in stacked form with the cross-equation restrictions imposed, even at the first OLS stage before the additional efficiency gain from the second-stage FGLS is achieved.

5. PANEL DATA

The application of systems of regressions equations to panel data, where n indexes observation units that are followed over time periods t, is very important in economics. A typical model for panel data is

$$y_{nt} = x_{nt}\beta + \alpha_n + u_{nt} \text{ for } n = 1,...,N \text{ and } t = 1,...,T .$$

In this model, the β parameters are not subscripted by n or t; this implies they are the same for every unit and every time period. (This is not as restrictive as it might appear, because variation in parameters over time or with some characteristics of the units can be reintroduced by including in the x's interactions with time dummies or with unit dummies.) The α_n are termed <u>individual</u> <u>effects</u>. They may be treated as

intercept terms that vary across units. The model with this interpretation is called a <u>fixed</u> <u>effects</u> (FE) model. Alternately, the α_n may be interpreted as components of the disturbance that vary randomly across units. The model with the second interpretation is called a <u>random</u> <u>effects</u> (RE) model. Often, the assumption is made that once the individual effects are isolated, the remaining disturbances u_{nt} are independent and identically distributed across n as well as t. Alternately, the u_{nt} could be serially correlated; this requires another layer of calculation for GLS.

The questions that arise in analysis of the panel data model are (a) under what conditions the model parameters can be estimated consistently, in either the fixed effects or the random effects interpretation; (b) what is the form of consistent or efficient estimators; and (c) whether the random effects or the fixed effects model is "better" in applications. I first analyze the fixed effects case, then the random effects case, and after this return to these questions to see what can be said.

6. FIXED EFFECTS

The fixed effects model can be rewritten by stacking the T observations on unit n,

$$(3) \qquad\qquad y_n = x_n\beta + 1_T\alpha_n + u_n ,$$

where 1_T is a T×1 vector of ones. Equation (3) is a special case of a general system of regression equations, and can be approached in the same way. Stacking the unit data, first unit followed by second unit, etc., gives the stacked model

(4) $$y = X\beta + D\alpha + u,$$

where $D = [d_1\ d_2\ ...\ d_N]$ is a NT×N array whose columns are dummy variables such that d_m is one for observations from unit m, and zero otherwise, and α is a N×1 vector with components α_n. (Exercise: Verify that this setup follows from the general stacking pattern shown in Section 2.)

In (4), note first that any column of X that does not change over t, within the observations for a unit, is linearly dependent on the columns of D. Then, when there are fixed effects, there is no possibility of identifying the separate effects of X variables that are time-invariant. Suppose we remove any such columns from X, so that only time-varying variables are left. For good measure, we can also remove from X the within-unit means of the X variables, so that X now denotes deviations from within-unit means. The model (3) can be rewritten as a relationship in unit means plus relationships in deviations from within unit means:

(5) $$\bar{y}_n = \alpha_n + \bar{u}_n$$
(6) $$Y_n = X_n\beta + \tilde{u}_n,$$

where $\bar{y}_n$ and $\bar{u}_n$ are unit means, Y_n is a vector of deviations of the unit n observations from the unit mean, and X_n is an array of deviations that has zero unit means by construction. Stack these models further, with the unit one data followed by the unit two data, etc., to obtain

$$(7) \qquad \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} \beta + \begin{bmatrix} \tilde{u}_1 \\ \tilde{u}_2 \\ \vdots \\ \tilde{u}_N \end{bmatrix}.$$

The deviations in (7) eliminate the fixed effects. Then, (7) can be estimated by OLS, which is consistent for β as $N \to +\infty$ or $T \to +\infty$ or both. (Note that (7) has one redundant observation for each observation unit, since the within group deviations must sum to zero. One can eliminate any one of the observations in each unit, or alternately leave it in the regression and remember that the number of observations is really N(T-1) rather than NT.) The regression (7) is called the within regression. One can estimate the fixed effect for each unit n using the formula $\hat{\alpha}_n = \bar{y}_n$; this is called the between regression. The fixed effects are estimated consistently only if $T \to +\infty$.

The particularly simple formula above for the fixed effects estimates came from normalizing the x's to have zero within-unit means. In the general case where the x's can have

non-zero unit means, the fixed effect estimators become $\hat{\alpha}_n = \bar{y}_n - \bar{x}_n b$, where b is the vector of estimates from (7).

Exercise 1: Using the projection notation $Q_D = I - D(D'D)^{-1}D'$, note that the OLS estimator of β in (4) is $b = (X'Q_D X)^{-1}X'Q_D y$. Show that this is the same as the within estimator of β.

7. RANDOM EFFECTS

Suppose the α's in (3) are treated as components of the disturbance, so that (3) can be rewritten as $y = X\beta + v$, where $v_{nt} = \alpha_n + u_{nt}$. Then, an OLS regression of y on X yields a consistent estimator of β as $NT \to +\infty$, provided the x's and the disturbances are uncorrelated. The covariance matrix of the stacked disturbances is now $E(vv') = I_N \otimes \Omega$, where Ω is the T$\times$T matrix of covariances of the disturbances $\alpha_n + u_{nt}$ for given n, with the form

$$(8) \quad \Omega = \begin{bmatrix} \sigma_\alpha^2 + \sigma_u^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_u^2 & \cdots & \sigma_\alpha^2 \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 + \sigma_u^2 \end{bmatrix} \equiv \sigma_\alpha^2 1_T 1_T' + \sigma_u^2 I_T$$

.

Efficiency of estimation can be improved by GLS. Verify as an exercise that $L = (I_T - \lambda 1_T 1_T{}')/\sigma_u$, with $\lambda = (1 - \sigma_u/(\sigma_u{}^2 + T\sigma_\alpha{}^2)^{1/2})/T$, satisfies $L\Omega L' = I_T$. Then, GLS is the same as OLS applied to the transformed data $(I_N \otimes L)y = (I_N \otimes L)X\beta + (I_N \otimes L)v$. In practice, Ω is unknown and FGLS must be used. Intuition for how to estimate $\sigma_\alpha{}^2$ and $\sigma_u{}^2$ can be obtained from an analogy to population moments. Let $v_n{}^*$ denote the unit mean of v_{nt}. We know that $Ev_{nt}{}^2 = \sigma_u{}^2 + \sigma_\alpha{}^2$ and that $Ev_n{}^{*2} = \sigma_u{}^2/T + \sigma_\alpha{}^2$. Solve these two equations for $\sigma_u{}^2$ and $\sigma_\alpha{}^2$:

(9) $\quad \sigma_u{}^2 = \dfrac{T}{T-1} (Ev_{nt}{}^2 - Ev_n{}^{*2})$ and $\sigma_\alpha{}^2 = (T\,Ev_n{}^{*2} - Ev_{nt}{}^2)/(T-1)$.

Then, substituting sample moments of fitted OLS disturbances in place of the population moments will give consistent estimates of the variance components. The steps to do FGLS are then to first regress y on X and retrieve the fitted residuals v_{nt}, and second, estimate $Ev_{nt}{}^2$ and $Ev_n{}^{*2}$ by the respective formulas

$$\frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} v_{nt}^2 \quad \text{and} \quad \frac{1}{N} \sum_{n=1}^{N} \left(\frac{1}{T} \sum_{t=1}^{T} v_{nt} \right)^2$$

Third, substitute these expressions in (9) to estimate the variance components and substitute the results into the L matrix, carry out the data transformations unit by unit, and run OLS on the transformed stacked data to get the FGLS estimates. The variance component estimates above are the

same as in Greene except for degrees of freedom adjustments. (Since only consistency of the estimates of σ_α^2 and σ_u^2 matter for the efficiency of the FGLS estimator, unbiasedness is no particular virtue. Finite sample monte carlo results on the value of degrees of freedom adjustments are not compelling. Thus, in most cases, it is probably not worth making these adjustments.) The estimator of σ_α^2 can go negative in finite samples. The usual recommendation in this case is to set the estimator to zero and assume there are no individual effects. Show as a (difficult) exercise that if the α's and u's are normal and uncorrelated with each other, then the estimators above are the maximum likelihood estimators for the variances.

Suppose that instead of starting from the original stacked data, we had started from the *within* regression model

(10) $$Y = X\beta + v^*,$$

which contains the stacked deviations from unit means, and constitutes N(T-1) observations if redundant observations are excluded; and the *between* regression model

(11) $$\overline{y} = \overline{x}\beta + \overline{v},$$

which contains the N stacked unit means. Provided the coefficients are identified (e.g., each variable is time-varying so that no columns of X are identically zero), one could estimate β consistently by applying OLS to either (10) or (11) separately. Greene shows that the OLS estimator can be

interpreted as a weighted combination of the within and between OLS estimators, and that the GLS estimator can be interpreted as a different weighted combination that gives less weight to the between model. For comparison, the fixed effects estimator of β was given by the within regression only.

8. FIXED EFFECTS VERSUS RANDOM EFFECTS

In the (unusual) case that you need estimates of the individual effects, you have no choice but to estimate the fixed effects model; even then, you need $T \to +\infty$ to estimate the α's consistently. The fixed effects model has the advantage that the estimates of β are consistent even if X is correlated with the individual effects, provided of course that X and the individual effects are uncorrelated with u. Its major drawbacks are that it uses up quite a few degrees of freedom, and makes it impossible to identify the effects of time-invariant explanatory variables. The random effects model economizes on degrees of freedom, and permits consistent estimation of the effects of all explanatory variables, including ones that are time-invariant, <u>provided</u> that all these explanatory variables are uncorrelated with the disturbances. (This is an advantage only if you have a convincing story to support the identifying assumption that there is zero correlation of these variables and the α's.)

As T → +∞, the FE and RE estimators merge, and the FE estimator can be interpreted as estimation of the RE model by conditioning on the realized values of the α's. From this, one can see how to test the RE model specification by examining the correlation of α and X. One way to do this is to regress the fitted α on X, and carry out a conventional F test that the coefficients in this regression are all zero. Unless T is very large, or the assumption that α is uncorrelated with X particularly implausible, it is usually better to work with the RE model.

9. SPECIFICATION TESTING

Standard regression model hypothesis testing of linear hypotheses on model coefficients, using Wald, LR, or SSR test statistics, carries over to the case of systems of regressions. This is most transparent when the FGLS estimators are given by OLS applied to data that is transformed to give a (asymptotically) scalar covariance matrix. This setup allows one to test not only hypotheses about coefficients in one equation, but also hypotheses connecting coefficients across equations, or in the panel context, across time.

For tests on covariance parameters, such as a test for homoskedasticity across equations, or a test for serial correlation, two useful ways to get suitable test statistics are to proceed by analogy with single-indexed regression problems, and to derive LM statistics under the assumption that

disturbances are normal. One example is a Durbin-Watson like test for serial correlation in panel data, using the estimated coefficient from a regression of v_{nt} on $v_{n,t-1}$ for n = 1,...N and t = 2,...,T.

Exercise 2: Consider the panel data model in which T → +∞. If the disturbances are uncorrelated with the right-hand-side variables, then both the FE and RE model estimates will be consistent and the RE estimates will be efficient. On the other hand, if there is correlation between the disturbances and the right-hand-side variables, only the FE estimates will be consistent. From these observations, suggest a simple specification test for the hypothesis that the disturbances are uncorrelated with the right-hand-side variables. Use (10) and (11) to show that this test is equivalent to a test for over-identifying restrictions.

Exercise 3: One of the ways a panel data model might come about is from a regression model $y_{nt} = x_{nt}\gamma_{nt} + u_{nt}$, where the γnt are random coefficients that vary with n (or t). When does this model reduce to the standard panel data model with random n effects? What are the generalizations of the standard RE and FE estimators when $\gamma_{nt} = \delta + \kappa_n + \lambda_t$?

10. VECTOR AUTOREGRESSION

The generic systems of equations model (1) with n indexing dependent variables and t indexing time, and with the right-hand-side variables various lags of the dependent variables, is called a *vector autoregression* (VAR) model. The model may include current and lagged exogenous variables, but is often applied to macroeconomic data where all the variables in the analysis are treated as dependent variables. To write out the lag structure, form the date-t vectors

$$\mathbf{y}_t = \begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{Nt} \end{bmatrix}, \mathbf{X}_t = \begin{bmatrix} x_{1t} & 0 & \cdots & 0 \\ 0 & x_{2t} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & x_{Nt} \end{bmatrix}, \mathbf{u}_t = \begin{bmatrix} u_{1t} \\ u_{2t} \\ \vdots \\ u_{Nt} \end{bmatrix},$$

and then

(12) $\qquad \mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \mathbf{A}_1\mathbf{y}_{t-1} + \dots + \mathbf{A}_J\mathbf{y}_{t-J} + \mathbf{u}_t,$

where the $\mathbf{A}_j$ are N×N arrays of lag coefficients. The VAR assumption is that with inclusion of sufficient lags, the disturbances in (12) are i.i.d. innovations that are statistically independent of $\mathbf{X}_t, \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots$. In this case, the variables $\mathbf{X}_t, \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots$ are said to be *strongly predetermined* in (12). The $\mathbf{X}_t$ are often assumed, further, to be *strongly exogenous*; i.e., $\mathbf{u}_t$ is statistically independent of $\mathbf{X}_t$ and all leads and lags of $\mathbf{X}_t$.

The dynamics of the system (12) are most easily analyzed by defining

$$\mathbf{y}_t = \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-J+1} \end{bmatrix} \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} A_1 & A_2 & \cdots & A_{J-1} & A_J \\ I_J & 0_J & \cdots & 0_J & 0_J \\ \vdots & \vdots & & \vdots & \vdots \\ 0_J & 0_J & \cdots & I_J & 0_J \end{bmatrix},$$

and rewriting the system in the form

$$\mathbf{y}_t = \begin{bmatrix} X_t\beta \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \mathbf{A}\mathbf{y}_{t-1} + \begin{bmatrix} u_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The system (12) with the strongly exogenous forcing variables X_t and the disturbances u_y omitted, is a *stable* difference equation if all the characteristic roots of A are less than one in modulus. The long-run dynamics of a stable system will be dominated by the largest (in modulus) characteristic root of A, and will have the feature that the impact on y_t of a shock in the disturbance in a specified period eventually damps out. Further, the most slowly decaying component in each variable in y_t will damp out at the same rate. (There is an exception if the characteristic vector associated with the largest characteristic root lies in a subspace spanned by a subset of the variables.) In the stable case, i.i.d. innovations, combined

with strongly exogenous variables that have a stationary distribution, will produce y_t with a stationary distribution. In particular, the covariance matrix of y_t will not vary with t, so that the y_t are homoskedastic. The estimation and hypothesis testing procedures discussed in Section 3 will then apply, with the predetermined and strongly exogenous variables treated the same. There will in general be contemporaneous correlation, so that (12) has the structure of a seemingly unrelated regressions problem for which GLS can be used to obtain BLUE estimates of the coefficients. If the strictly exogenous variables are the same in every equation, there are no exclusion restrictions in the lag coefficients, and no restrictions on coefficients across equations, GLS estimation reduces to OLS applied to each equation separately, as before.

If A has one or more roots of modulus one or greater, then the impact of past disturbances does not damp out, the system (12) is unstable, and the variance of y_t rises with t. The occurrence of modulus one (unit) roots seems to be fairly common in macroeonomic time series. Statistical inference in such systems is quite different than in stable systems. In particular, detection and testing for unit roots, and the corresponding characteristic roots that determine cointegrating relationships among the variables, require a special statistical analysis. The topic of testing for unit roots and cointegrating relationships is discussed extensively by Stock "Unit Roots, Structural Breaks, and Trends," and Watson "Vector Autoregression and Cointegration," both in

R. Engle and D. McFadden, eds., Handbook of Econometrics IV, 1994.

11. SYSTEMS OF NONLINEAR EQUATIONS

The systems of equations linear in variables and parameters, with additive disturbances, that were introduced at the beginning of this chapter, can be extended easily to systems that retain the assumption of additive disturbances, but are nonlinear in variables and/or parameters:

$$(13) \qquad y_{nt} = h_n(x_{nt}, \beta_n) + u_{nt},$$

where $n = 1,...,N$, $t = 1,...,T$, and β_n is $k_n \times 1$. Assume for the following discussion that the disturbances u_{nt} are independent for different t. If the x_{nt} are strongly predetermined, implying that $E(u_{nt} | x_{nt}) = 0$, then each equation in (13) can be estimated by nonlinear least squares. This can be interpreted as a "limited information" or "marginal" GMM estimation procedure in which information from the equations for the remaining variables is not used. Chapter 3 discusses the statistical properties of nonlinear least squares estimators.

In general, there will be an efficiency gain from taking into account the covariance structure of the disturbances u_{nt} for different n. This can be done practically in TSP by using the LSQ command applied to all the equations in the model. This procedure then applies nonlinear least squares to each

equation separately, retrieves fitted residuals, uses these residuals to estimate the covariance matrix of the disturbances at each t, and then does feasible generalized nonlinear least squares employing the estimated covariance matrix.