

A FAMILY OF ESTIMATORS  
FOR SIMULTANEOUS EQUATION SYSTEMS

Gregory C. Chow

ECONOMETRIC RESEARCH PROGRAM  
Research Memorandum No. 155  
October 1973

The research described in this paper  
was supported by NSF Grant GS 32003X.

Econometric Research Program  
PRINCETON UNIVERSITY  
207 Dickinson Hall  
Princeton, New Jersey

# A FAMILY OF ESTIMATORS FOR SIMULTANEOUS EQUATION SYSTEMS

Gregory C. Chow\*

## 1. INTRODUCTION

The purpose of this paper is to set forth, in a systematic fashion, a family of estimators for systems of simultaneous stochastic equations based on modifications and simplifications of the method of maximum likelihood. It is hoped that this family will satisfy the need in most practical applications involving linear as well as non-linear structural equations. Most of these estimators are well known, but a few are newly suggested. They are developed by introducing the complicating statistical assumptions of simultaneous equations one at a time and by simplifying the computations involved in maximizing the likelihood function of the model parameters. Their development may clarify the relationships between estimation methods and the underlying statistical assumptions. They are arranged from the simple to the complicated so that the user may choose to apply them sequentially in practical applications. The development based on the method of maximum likelihood can also serve as a convenient framework for introducing robust estimation techniques for simultaneous equation systems.

In contrast with the model of multiple linear regression, a system of simultaneous econometric equations involves several complications. First, there may be several dependent variables in one equation. Since these variables are correlated with the

residual, consistent estimates of the parameters of the equation cannot be obtained by ordinary least squares. Second, the residuals in different equations may be mutually correlated. This assumption suggests the desirability of pooling different equations together in estimation in order to improve efficiency, but pooling will increase the size of the computation problem. This complication should be distinguished from the first, for it can exist even when, as in a multivariate regression system, there is only one dependent variable in each equation.<sup>1</sup> Third, the equations may be non-linear. Fourth, the residuals may be serially correlated.

It is the approach of this paper to deal with these complications one by one, not only for the purpose of understanding what each complication entails as far as estimation is concerned, but for the purposes of providing useful intermediate estimates and of solving the difficult problem of computing the final estimates step by step. Specifically, one may consider using ordinary least squares to estimate each equation separately, ignoring for the moment the above-mentioned complications. He then introduces only the first complication. This would involve revising the ordinary least squares estimates. Next, he would introduce the second complication, further revising his estimates, etc., until the final set of estimates are obtained.

This approach has been motivated by three related considerations. First, the practice of econometric model building does frequently proceed along the lines here suggested. Econometricians often apply ordinary least-squares to see how well an equation fits before introducing more sophisticated methods. Two stage least squares is often performed before three-stage least squares, if the latter is the method adopted. This paper will provide a systematic approach to step-by-step estimation. The progression from the simple to the complicated enables the research worker to learn something in the process. Second, for an understanding of the nature of simultaneous econometric systems, it is desirable to pinpoint the modification to the estimation procedure which is necessitated by each complication in statistical modeling. When the estimation of systems of linear econometric equations is studied in most of the existing literature, the first two complications have been introduced together. This has given rise to complicated estimation methods, but the modifications to the procedures are often not attributed to their respective sources. This paper will attempt to assign each complication in the procedure to a specific assumption made about the structure. Third, in a recent paper, Chow (1972), I have provided a method to obtain full-information maximum likelihood estimates for non-linear equation

systems with autoregressive residuals. The present study can be considered as a step-by-step implementation of that method. Although full-information maximum likelihood estimation of non-linear systems is feasible, few would attempt it as the first step in constructing an econometric model. One would prefer to build up to it gradually, and the approach here suggested is designed precisely for doing so. Furthermore, one might not wish to go all the way if, in the process, he discovers that certain complications are not worth introducing, or, if, partly because of the high cost of computations, he is prepared to accept a linearized version of the estimates after one or two iterations of the non-linear estimating equations. The family of estimates here set forth are available for him to pick and choose in order to satisfy his needs.

The methods suggested in this paper are based primarily on variations of the method of maximum likelihood, rather than the methods of two-stage and three-stage least squares. For non-linear systems, generalizations of two-stage and three-stage least squares remain to be further investigated. For linear systems, as it will be pointed out in this paper, the method of maximum likelihood is computationally not more difficult and in some cases easier than the method of two-stage or three-stage least squares to be applied under a comparable set

of assumptions. Thus, based on the method of maximum likelihood, this paper provides, systematically and fairly comprehensively, a set of estimators that can be used for both linear and non-linear systems of econometric equations under various simplifying assumptions.

Section 2 deals with linear systems with serially uncorrelated residuals by providing a set of estimators, from the simple to the complicated, which are based on various simplifications of the statistical model or the computational procedure. The same set of estimators will be set forth for non-linear systems in section 3. The subject of residuals satisfying an autoregressive scheme will be the concern of section 4. Section 5 provides a set of robust estimators for simultaneous equations, as generalizations of the techniques of robust estimation using the framework of this paper.

## 2. LINEAR SYSTEMS

Let  $T$  observations on  $G$  linear simultaneous stochastic equations be written as

$$(2.1) \quad YB + Z\Gamma = U$$

where  $Y$  is a  $T \times G$  matrix of dependent variables,  $Z$  is a  $T \times K$  matrix of predetermined variables, each row of  $U$  is

G-variate normal with covariance matrix  $\Sigma$  and is uncorrelated with any other row, and the  $i^{\text{th}}$  columns of  $\mathbf{B}$  and  $\Gamma$  are coefficients of the  $i^{\text{th}}$  structural equations. Adopting the normalization  $\beta_{ii} = -1$  ( $i=1, \dots, G$ ), and letting  $\beta_i$  be the  $i^{\text{th}}$  column of  $\mathbf{B}$  omitting  $\beta_{ii}$  and the zero elements,  $\gamma_i$  be the  $i^{\text{th}}$  column of  $\Gamma$  omitting zero elements,  $Y_i$  be a matrix consisting of the columns of  $Y$  which correspond to the elements of  $\beta_i$ , and  $Z_i$  be a matrix consisting of the columns of  $Z$  which correspond to the elements of  $\gamma_i$ , one can write the  $T$  observations on the  $i^{\text{th}}$  equation as

$$(2.2) \quad -Y_i + Y_i \beta_i + Z_i \gamma_i = u_i \quad (i=1, \dots, G)$$

The symbol  $u_i$  will denote, according to context, either a vector of random residuals or a vector of functions of the variables  $\beta_i$  and  $\gamma_i$  as defined by (2.2). Let  $S$  be a  $G \times G$  symmetric matrix whose  $g$ - $h$  element is  $(u_g' u_h)/T$ .

The concentrated log-likelihood function, after the unknown matrix  $\Sigma$  is set equal to  $S$ , is

$$(2.3) \quad L = \text{const} - \frac{1}{2} T \log |S| + T \log |B|$$

and its gradient is<sup>2</sup>

$$(2.4a) \quad \frac{\partial L}{\partial \beta_i} = -Y_i' \sum_{h=1}^G s^{hi} u_h + T \beta^{i(i)}$$

$$(2.4b) \quad \frac{\partial L}{\partial \gamma_i} = - Z_i' \sum_{h=1}^G s^{hi} u_h$$

where  $s^{hi}$  is the  $h$ - $i$  element of  $S^{-1}$ ,  $\beta^{i(k)}$  denotes a column vector consisting of those elements of the  $i^{\text{th}}$  row of  $B^{-1}$  which correspond to the unknown elements of  $\beta_k$ , and where use is made of the differentiation rule  $\partial \log |B| / \partial \beta_{ki} = \beta^{ik}$ . The matrix of second partials of  $L$  consists of submatrices:<sup>3</sup>

$$(2.5a) \quad \frac{\partial^2 L}{\partial \beta_i \partial \beta_j'} = - s^{ji} Y_i' Y_j + \frac{1}{T} Y_i' \sum_{h=1}^G u_h \sum_{n=1}^G (s^{hj} s^{ni} + s^{hn} s^{ji}) u_n' Y_j - T \beta^{j(i)} \beta^{i(j)}$$

$$(2.5b) \quad \frac{\partial^2 L}{\partial \beta_i \partial \gamma_j'} = - s^{ji} Y_i' Z_j + \frac{1}{T} Y_i' \sum_{h=1}^G u_h \sum_{n=1}^G (s^{hj} s^{ni} + s^{hn} s^{ji}) u_n' Z_j$$

If one wished to estimate the system by the method of full-information maximum likelihood without further simplifying assumptions he could use the gradient  $g(\alpha) = g(\beta_1 \dots \beta_G, \gamma_1 \dots \gamma_G)$  as given by (2.4) and the matrix of the Hessian  $H(\alpha)$  as given by (2.5) and iterate by Newton's method:

$$(2.6) \quad \alpha^{r+1} = \alpha^r - h_r H(\alpha^r)^{-1} g(\alpha^r)$$

where  $\alpha^r$  denotes the value of the vector of unknowns at the



$r^{\text{th}}$  iteration and  $h_r$  is a scalar chosen to promote convergence of the iterative process.<sup>4</sup>

In this paper, a family of estimators are suggested. They are obtained by modifying the statistical assumptions concerning the model (2.1) and/or simplifying the computations based on the solution of the likelihood equations  $g(\alpha) = 0$  by Newton's method. The simplifying statistical assumptions are that  $\Sigma$  may be diagonal or bloc-diagonal, and that  $B$  may be triangular or bloc-triangular. The simplifying computations consist of performing only one iteration or solving the iterative equations for one subset of variables at a time, holding the other variables constant.

To begin with, assume that  $\Sigma$  is diagonal, so that  $\log|S|$  in the log-likelihood (2.3) becomes  $\sum_i \log s_{ii}$ . The gradient and the matrix of the Hessian will then become, respectively,

$$(2.7a) \quad \frac{\partial L}{\partial \beta_i} = -s_{ii}^{-1} Y_i' u_i + T \beta^{i(i)}$$

$$(2.7b) \quad \frac{\partial L}{\partial \gamma_i} = -s_{ii}^{-1} Z_i' u_i$$

and, with  $\delta_{ij}$  denoting the Kronecker delta,

$$(2.8a) \quad \frac{\partial^2 L}{\partial \beta_i \partial \beta_j'} = -\delta_{ij} [s_{ii}^{-1} Y_i' Y_i + \frac{2}{T} s_{ii}^{-2} Y_i' u_i u_i' Y_i] - T \beta^{j(i)} \beta^{i(j)'} ,$$

$$(2.8b) \quad \frac{\partial^2 L}{\partial \beta_i \partial \gamma_j} = - \delta_{ij} [s_{ii}^{-1} Y_i' Z_i + \frac{2}{T} s_{ii}^{-2} Y_i' u_i u_i' Z_i]$$

$$(2.8c) \quad \frac{\partial^2 L}{\partial \gamma_i \partial \gamma_j} = - \delta_{ij} [s_{ii}^{-1} Z_i' Z_i + \frac{2}{T} s_{ii}^{-2} Z_i' u_i u_i' Z_i] .$$

Except for the last term of (2.8a), the Hessian matrix would be bloc-diagonal, and the solution to equation (2.6) could be obtained for each bloc separately.

If  $B$  were upper-triangular, the  $j^{\text{th}}$  row of  $B^{-1}$  would have zeros in its first  $j-1$  places. Since  $\beta^{j(i)}$  is a vector obtained from selecting those elements in the  $j^{\text{th}}$  row of  $B^{-1}$  which correspond to the unknown coefficients in equation  $i$ , and these coefficients are confined to the first  $i-1$  for an upper-triangular  $B$ , we have  $\beta^{j(i)} = 0$  for  $i \leq j$ . From (2.8a), the Hessian matrix would then be truly bloc-diagonal, each bloc for the coefficients of only one equation. In fact, the last term of the gradient (2.7a) would also vanish, and one would simply be solving the linear normal equations  $Y_i' u_i = 0$  and  $Z_i' u_i = 0$  for each equation  $i$  separately as specified by the method of least squares.

If  $B$  is bloc-triangular, similar arguments will show that  $\beta^{j(i)} = 0$  for equation  $i$  belonging to a lower-numbered bloc than equation  $j$ . Hence, the last term of (2.8a) vanishes for  $i$  and  $j$  belonging to different blocs, and the Hessian matrix will also be bloc-diagonal, each bloc consisting of

possibly several equations. It is therefore recommended that the structural equations be so arranged as to make  $\mathbb{B}$  bloc-triangular.<sup>5</sup> The solution of (2.6) will then be obtained for each bloc separately.

After taking full advantage of the bloc-triangularity of  $\mathbb{B}$ , one might still find some very large blocs in the Hessian matrix. As an option, or as an intermediate step toward using the Hessian matrix of (2.8), one may choose to ignore the last term of (2.8a) for  $i \neq j$ , and perform the iterations for the coefficients of each equation separately. In other words, letting  $\alpha_i^i$  denote the unknowns  $(\beta_i^i \gamma_i^i)$ ,  $g_i$  denote the gradient (2.7), and  $H_{ij}$  denote the Hessian matrix (2.8), one may apply the formula

$$(2.9) \quad \alpha_i^{r+1} = \alpha_i^r - h_{ir} H_{ii}(\alpha^r)^{-1} g_i(\alpha^r) .$$

This amounts to applying Newton's method after setting the off-diagonal blocs  $H_{ij}$  ( $i \neq j$ ) equal to zero. As it is well-known, if the gradient is linear as given by  $g(\alpha) = H\alpha - c$ , Newton's method gives the correct solution in one iteration:

$$(2.10) \quad \alpha^{r+1} = \alpha^r - H^{-1}(H\alpha^r - c) = H^{-1}c .$$

Rather than solving the system of (linearized) equations  $H\alpha = c$

simultaneously as in (2.10), the method of (2.9) solves each subset of equations  $\sum_j H_{ij} \alpha_j = c_i$  separately for each  $\alpha_i$  holding the values of  $\alpha_j$  ( $j \neq i$ ) constant in the process, as the following demonstrates:

$$(2.11) \quad \alpha_i^{r+1} = \alpha_i^r - H_{ii}^{-1} (\sum_j H_{ij} \alpha_j^r - c_i) = H_{ii}^{-1} (c_i - \sum_{j \neq i} H_{ij} \alpha_j^r) .$$

To recapitulate, we have outlined four methods of estimating simultaneous equations. The first is ordinary least squares and is equivalent to maximum likelihood for diagonal  $\Sigma$  and triangular  $B$  (assuming serially uncorrelated residuals). The second uses a diagonalized Hessian as indicated by (2.9), and is an iterative method of solving the likelihood equations for diagonal  $\Sigma$  and general  $B$ . The third uses the full Hessian of (2.8) and is equivalent to applying Newton's method of solving the likelihood equations for diagonal  $\Sigma$ . Note that, for the third method, the Hessian will still be bloc-diagonal provided that the bloc-triangularity of  $B$  is fully exploited. Finally, the fourth is full-information maximum likelihood (FIML) for general  $\Sigma$ , using the gradient (2.4) and the Hessian (2.5). Except for ordinary least squares, all other methods are iterative. However, one might choose to iterate only once, using as initial values the estimates from a computationally simpler method.

Several comments are now in order for comparing this family of estimators with the two-stage and three-stage least squares family. If the covariance matrix  $\Sigma$  is far from being diagonal, FIML for general  $\Sigma$  and three-stage least squares are to be recommended on the ground of asymptotic efficiency. Three-stage least squares is analogous to performing only one iteration of FIML because it replaces the true  $\Sigma$  (and its inverse) in the estimation equations by their estimates based on consistent estimates of the structural coefficients. If the same substitution is applied to the FIML estimation equations, only one set of linear equations with the same number of unknowns will have to be solved. Computationally, the two methods are of the same order of difficulty.

If  $\Sigma$  is diagonal or nearly so, the asymptotic efficiency argument in favor of FIML or 3SLS loses its force. Two-stage least squares is available for this situation, and should be compared with methods two and three suggested in this paper. Method three is truly FIML under the assumption of diagonal  $\Sigma$ . If the bloc-triangularity of the  $B$  matrix is fully exploited, this method may not require solving any larger set of linear equations than 2SLS as the first stage of 2SLS may involve solving a large regression problem (granted that approximation by principle components is possible). Note also that further

iterations on 2SLS are possible, using for the first stage revised estimates of the reduced form coefficients  $\Pi = -\Gamma B^{-1}$  based on improved estimates of the structural coefficients. Non-iterating 2SLS should perhaps be compared with our methods two and three terminated after only one iteration.

Our method two, if and when it converges, is also truly FIML for diagonal  $\Sigma$  for it solves the same set of likelihood equations (2.7) as method three. In contrast with method three, it solves a set of smaller problems (one for each structural equation) in each iteration. Each of these small problems has the same size as that of the second stage of 2SLS, but the probably larger problem of the first stage of 2SLS is avoided.<sup>6</sup>

Although methods two and three are based on the assumption of a diagonal  $\Sigma$ , they will have wider applicability than it may first appear. If the correlation between  $u_{1t}$  and  $u_{2t}$  is zero (small), it will be (or may be) more efficient not to combine the estimation of  $\beta_1$  and  $\beta_2$  since efficiency will increase by the imposition of a correct restriction on the parameter  $\sigma_{12}$ . A Monte Carlo study as reported in Mikhail (1972) for some illustrative two-equation models has concluded that  $\sigma_{12}$  has to be fairly large in order for the estimators that pool together the estimation of two equations to be more efficient in small samples. It may be better to impose a zero restriction on  $\sigma_{12}$  that is nearly correct than not to impose

any restriction at all. One can easily exploit the possibility of making  $\Sigma$  bloc diagonal, thus combining the coefficients of the equations in only the same bloc for estimation purposes. This would involve fairly straight-forward modifications of methods two and three and need not be detailed here.

For linear systems, the methods of two-stage and three-stage least squares are alternatives to the methods of this paper. For non-linear systems, however, generalizations of the former methods remain to be further investigated. We will turn to the subject of non-linear systems presently.

### 3. NON-LINEAR SYSTEMS

The same four estimators can be suggested for a system of non-linear structural equations as for a linear system. Let the  $t^{\text{th}}$  observation on the  $i^{\text{th}}$  equation be

$$(3.1) \quad f_i(y_{1t}, \dots, y_{Gt}, z_{1t}, \dots, z_{Kt}; \beta_i) = -y_{it} +$$

$$\phi_i(y_{1t}, \dots, y_{i-1,t}, y_{i+1,t}, \dots; \beta_i) = u_{it}$$

where  $u_{it}$  ( $i=1, \dots, G$ ) are multivariate normal with mean zero and covariance matrix  $\Sigma$ , and are serially uncorrelated, and where  $\beta_i$  is a vector of  $n_i$  unknown parameters in equation  $i$ . Define the matrix of the Jacobian as

$$(3.2) \quad J = (J_{ij}) = \left( \frac{\partial f_i}{\partial y_j} \right).$$

$J_t = (J_{ij,t})$  will denote the value of the function for the  $t^{\text{th}}$  observation. For most econometric models,  $J$  will be very sparse, consisting of probably two to three unknown elements in each row, on the average.

The concentrated log-likelihood function is

$$(3.3) \quad L = \text{const} - \frac{1}{2} T \log |S| + \sum_{t=1}^T \log |J_t|$$

where  $s_{gh}$ , as before, denotes  $(u'_g u_h)/T$ . Its gradient is<sup>7</sup>

$$(3.4) \quad \frac{\partial L}{\partial \beta_i} = - Y_i^{O'} \sum_{h=1}^G s^{hi} u_h + \sum_{t=1}^T \sum_{h=1}^G J_t^{hi} \frac{\partial J_{ih,t}}{\partial \beta_i}$$

where  $Y_i^O$  denotes the  $T \times n_i$  matrix with  $\frac{\partial f_{i,t}}{\partial \beta_{ij}}$  as its  $t$ - $j$

element. If the  $i^{\text{th}}$  equation is linear,  $J_{ih,t} = \beta_{ih}$ ;  $\partial J_{ih,t} / \partial \beta_i$  will be a column vector of zeros if  $\beta_{ih} = 0$ , and it will be a column vector of zeros except for a unity in the same position as  $\beta_{ih}$  is located in the contracted (unknown) vector  $\beta_i$ . In this case,  $\sum_h J_t^{hi} \cdot \partial J_{ih,t} / \partial \beta_i$  would be a column vector formed by selecting those elements of the  $i^{\text{th}}$  column of  $J^{-1}$  which correspond to the unknown elements of  $\beta_i$ .<sup>8</sup> This vector was denoted by  $\beta^{i(i)}$  in equation (2.4a). The Hessian



matrix of the likelihood function (3.3) consists of submatrices

$$(3.5) \quad \frac{\partial^2 L}{\partial \beta_i \partial \beta_j'} = -s^{ij} Y_i^{O'} Y_j^O + \frac{1}{T} \sum_h u_h \sum_n (s^{hj} s^{ni} + s^{hn} s^{ji}) u_n' Y_j^O -$$

$$- \delta_{ij} \sum_t \frac{\partial^2 \hat{f}_{i,t}}{\partial \beta_i \partial \beta_i'} \sum_h s^{hi} u_{ht}$$

$$+ \sum_t \sum_h [\delta_{ij} J_t^{hi} \frac{\partial^2 J_{ih,t}}{\partial \beta_i \partial \beta_i'} - \frac{\partial J_{ih,t}}{\partial \beta_i} \sum_n (J_t^{hj} J_t^{ni}) \frac{\partial J_{jn,t}}{\partial \beta_j'}].$$

(3.4) and (3.5) can be used to obtain FIML estimates by Newton's method.

If  $\Sigma$  is diagonal, the gradient and the Hessian will be respectively

$$(3.6) \quad \frac{\partial L}{\partial \beta_i} = -s_{ii}^{-1} Y_i^{O'} u_i + \sum_{t=1}^T \sum_{h=1}^G J_t^{hi} \frac{\partial J_{ih,t}}{\partial \beta_i}$$

and

$$(3.7) \quad \frac{\partial^2 L}{\partial \beta_i \partial \beta_j'} = \delta_{ij} [-s_{ii}^{-1} Y_i^{O'} Y_j^O + \frac{2}{T} s_{ii}^{-2} Y_i^{O'} u_i u_i' Y_i^O - \sum_t \frac{\partial^2 \hat{f}_{i,t}}{\partial \beta_i \partial \beta_i'} s_{ii}^{-1} u_{it}]$$

$$+ \sum_t \sum_h [\delta_{ij} J_t^{hi} \frac{\partial^2 J_{ih,t}}{\partial \beta_i \partial \beta_i'} - \frac{\partial J_{ih,t}}{\partial \beta_i} \sum_n (J_t^{hj} J_t^{ni}) \frac{\partial J_{jn,t}}{\partial \beta_j'}].$$

If, in addition,  $J$  is (lower) triangular,<sup>9</sup>  $\frac{\partial J_{ih,t}}{\partial \beta_i} = 0$  for  $h \geq i$  and  $J_t^{hi} = 0$  for  $i > h$ , implying that the second term on the right-hand side of (3.6) vanishes. The normal equations for  $\beta_i$  will then be  $Y_i^{O'} u_i = 0$  and are simply the

estimation equations of the method of least squares applied to each equation separately. The gradient of  $\frac{1}{2} T s_{ii}$  is  $Y_i^{O'} u_i$ , and the Hessian is<sup>10</sup>

$$(3.8) \quad \frac{\partial^2}{\partial \beta_i \partial \beta_i'} \left( \frac{1}{2} T s_{ii} \right) = Y_i^{O'} Y_i^O + \sum_t \frac{\partial^2 f_{i,t}}{\partial \beta_i \partial \beta_i'} u_{it}.$$

Rather than using the entire Hessian matrix of (3.7), we propose as the second method following least squares, to use only the diagonal blocs and iterate by

$$(3.9) \quad \beta_i^{r+1} = \beta_i^r - h_{ir} \left[ \frac{\partial^2 L}{\partial \beta_i \partial \beta_i'} \right]^{-1} \left[ \frac{\partial L}{\partial \beta_i} \right]$$

where the gradient is defined by (3.6) and the Hessian by (3.7).

If  $J$  is bloc-triangular, the expressions  $\partial J_{ih} / \partial \beta_i$ ,  $J^{hj}$ ,  $\partial J_{jn} / \partial \beta_j'$ , and  $J^{ni}$  will be respectively non-zero only if  $h \leq i$ ,  $j \leq h$ ,  $n \leq j$  and  $i \leq n$ , where  $h \leq i$  means that equation  $h$  is in a lower or equal numbered bloc as compared with equation  $i$ . Hence the last term of (3.7) will be non-zero only if equation  $i$  and equation  $j$  belong to the same bloc. Thus, as in the linear case, bloc-triangularity of  $J$  implies bloc-diagonality of the Hessian (3.7), given a diagonal  $\Sigma$ . As before, the third method is to use (3.6) and (3.7) respectively as the gradient and the Hessian for the purpose of iteration, taking full advantage of the bloc-triangularity

of  $J$ . The fourth method is FIML for general  $\Sigma$ , using equations (3.4) and (3.5).

#### 4. AUTOREGRESSIVE RESIDUALS

Since the subject of autoregressive residuals is already treated in Chow (1972), only a few important points will be made here in connection with the family of estimators outlined in this paper. One essential point to note is that a system with autoregressive residuals can be converted into a system with serially uncorrelated residuals so that the methods of section 3 can be applied. For example, let the residuals of (3.1) satisfy a second-order autoregression

$$(4.1) \quad u_{.t} - R_1 u_{.t-1} - R_2 u_{.t-2} = e_{.t}$$

where  $u_{.t}$  is a column vector consisting of  $u_{1t}, \dots, u_{Gt}$  (or alternatively  $f_{1t}, \dots, f_{Gt}$ ) and  $e_{.t}$  are serially uncorrelated. Each row of (4.1) can be regarded as a new non-linear function. In the new set-up, the  $i^{\text{th}}$  row will contain not only  $\beta_i$  as parameters but also  $\beta_j$  if the  $j^{\text{th}}$  element of the  $i^{\text{th}}$  row of either  $R_1$  or  $R_2$  is non-zero. Since  $\beta_j$  are also parameters in equation  $j$ , there will be linear restrictions on the parameters across equations in the new set-up. A procedure to deal with these linear restrictions is described in Chow (1972) and

Chow and Fair (1973). It amounts to collapsing the gradient and the Hessian matrix by combining the first and second derivatives with respect to those parameters which are subject to linear restrictions. Since it does not disturb the bloc-diagonality of the Hessian matrix, method three remains applicable. Of course, method two can always be applied as it selects only the diagonal submatrices of the Hessian matrix for each iteration. However, each newly formed equation of (4.1) will contain more parameters and thus become more difficult to estimate from the computational point of view.

The simplest model, and the one to be recommended until it is invalidated by sufficient evidence, is the one assuming  $R_1$  and  $R_2$  to be diagonal. In this case, the new set-up due to autoregressive residuals will not introduce any common parameters in different equations. Each new equation  $i$  will simply have two extra parameters  $r_{1,ii}$  and  $r_{2,ii}$ . The Jacobian will be the same as before. One can apply the same four methods of section 3 to the new situation. Even if some off-diagonal elements of  $R_1$  and  $R_2$  are non-zero, this simple model can advantageously be used as an intermediate step in the estimation process.

5. ROBUST ESTIMATORS<sup>11</sup>

Recently there has been some interest in applying robust estimation techniques to the study of economic relationships. The basic idea is that the method of least squares, which is to a large extent based on a normal distribution of the residuals, may give too much weight to the large residuals if the distribution does have longer tails than the normal. A robust estimator, by giving less weight to the large residuals than least squares, will have reasonably good properties for a variety of possible distributions of the residuals, although it may not be quite as good as least squares for a truly normal situation.

One way to reduce the weight given to large residuals is to formulate the problem as one of weighted regression. For example, to minimize the sum of the absolute values of the residuals  $u_t$  is to minimize the sum  $u_t^2/|u_t|$ . The weight  $w_t$  to be applied to  $u_t$  in the weighted regression problem is simply  $|u_t|^{-\frac{1}{2}}$ . Although the weight depends on the value of  $u_t$  itself, one may choose to iterate by using a previous value of  $u_t$  in the weight for the weighted regression in the current iteration.<sup>12</sup> Other functions to truncate large residuals have been proposed in the literature on robust techniques, but there is no need to restate them here.<sup>13</sup>

Since a robust estimation method can be formulated as weighted least squares and since the method of maximum likelihood

as applied to a simultaneous-equation system can be viewed as a natural generalization of least squares,<sup>14</sup> the two ideas can be combined to form a robust estimator for simultaneous equations. Specifically, for a column vector  $u$  of residuals, a weighted sum of squares takes the form  $u'W'Wu$  where  $W$  is a diagonal weighting matrix consisting of  $w_1, \dots, w_T$  in the diagonal. When there are columns of residuals  $u_g$  and  $u_h$  (for equations  $g$  and  $h$ ), the sum of squares or cross products should accordingly be changed to  $u_g'W_gW_hu_h$ , with  $W_g$  and  $W_h$  again denoting diagonal weighting matrices. The method of maximum likelihood amounts to minimizing a generalized variance  $|S|$  of the residuals as in (3.3) subject to some normalization rules introduced by the Jacobian  $|J_t|$ . A natural combination of these two ideas is to maximize the function (3.3) with the elements of  $S$  changed to the weighted sum of squares or cross-products, i.e., to maximize

$$(5.1) \quad L^* = \text{const} - \frac{1}{2} T \log |S^*| + \sum_{t=1}^T \log |J_t|$$

where

$$(5.2) \quad S^* = (s_{gh}^*) = \left( \frac{1}{T} u_g'W_gW_hu_h \right).$$

Following the development of section 3, the gradient and the Hessian matrix of (5.1) can be simply stated. Since the derivatives of  $\sum_t \log |J_t|$  are already given in section 3, only

the derivatives of  $-\frac{1}{2} T \log|S^*|$  will be recorded below:

$$(5.3) \quad \frac{\partial}{\partial \beta_i} \left[ -\frac{1}{2} T \log|S^*| \right] = Y_i^{O'} W_i \sum_{h=1}^G s^{*hi} W_h u_h$$

$$(5.4) \quad \frac{\partial^2}{\partial \beta_i \partial \beta_j} \left[ -\frac{1}{2} T \log|S^*| \right] = -s^{*ji} Y_i^{O'} W_i W_j Y_j^{O'} +$$

$$+ \frac{1}{T} Y_i^{O'} W_i \sum_{h=1}^G W_h u_h \sum_n (s^{*hj} s^{*ni} + s^{*hn} s^{*ji}) u_n' W_n W_j Y_j^{O'} -$$

$$- \delta_{ij} \sum_{h=1}^G s^{*hi} \sum_t u_{ht} \frac{\partial^2 u_{it}}{\partial \beta_i \partial \beta_j} W_{it} W_{ht} .$$

Under appropriate assumptions, methods 2,3, and 4 of section 3 can be applied, using the modified gradient and Hessian of (5.3) and (5.4).

## REFERENCES

1. D.F. Andrews, et al., Robust Estimates of Location (Princeton: Princeton University Press, 1972).
2. G.C. Chow, "A Comparison of Alternative Estimators for Simultaneous Equations," Econometrica, Vol. 32, No. 4 (Oct. 1964), pp. 532-553.
3. \_\_\_\_\_, "Two Methods of Computing Full-Information Maximum Likelihood Estimates in Simultaneous Stochastic Equations," International Economic Review, Volume 9, No. 1 (February 1968), pp. 100-112.
4. \_\_\_\_\_, "On the Computation of Full-Information Maximum Likelihood Estimates for Non-linear Equation Systems," Econometric Research Program Research Memorandum No. 142, Princeton University (June 1972), a slightly abridged version in the Review of Economics and Statistics, Vol. 55, No. 1 (February, 1973), pp. 104-109.
5. \_\_\_\_\_, and R.C. Fair, "Maximum Likelihood Estimation of Linear Equation Systems with Auto-regressive Residuals," Annals of Economic and Social Measurement, Volume 2, Number 1 (January 1973), pp. 17-28.
6. F.M. Fisher, "Dynamic Structure and Estimation in Economy-Wide Econometric Models," in J.S. Duesenberry, ed., The Brookings Quarterly Econometric Model of the United States (Chicago: Rand McNally and Company, 1965), pp. 589-636.
7. W.M. Mikhail, "A Comparative Monte Carlo Study of the Properties of Econometric Estimators," (mimeographed 1972).
8. T. Rothenberg and C.T. Leenders, "Efficient Estimation of Simultaneous Equation Systems," Econometrica, Vol. 32, No. 1 and 2 (January-April, 1964), pp. 57-76.
9. A. Zellner, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," Journal of the American Statistical Association, Vol. 57, No. 298 (June 1962), pp. 348-368.



## FOOTNOTES

\*

Research conducted while the author was consultant to the National Bureau of Economic Research Computer Research Center. It was also supported by National Science Foundation Grant No. 32003X1. Helpful comments from David Belsley, Marc Eisner, Ray Fair, Edwin Kuh, and Richard Quandt are much appreciated.

1. The improvement in efficiency by pooling together several regressions that contain different sets of explanatory variables was suggested by Zellner (1962).
2. Details of differentiation can be found in Rothenberg and Leenders (1964) and in Chow (1968). The latter reference uses a slightly modified form of the log-likelihood function.
3. See footnote 2 above.
4. See Chow (1968) and Chow and Fair (1973) for discussion of the choice of  $h_r$ .
5. The Stuart algorithm installed in the TROLL system at the NBER Computer Research Center can be used for this purpose. A theoretical treatment of bloc-triangular  $B$  in the context of estimation of simultaneous econometric equations can be found in Fisher (1965).
6. There is a modified version of method two which will provide an interpretation and justification of the method. It is obtained by replacing  $H_{ii}$  of (2.9) by only the leading term on the right-hand side of (2.8). Thus the estimation equation for  $\beta_i$  is

$$\begin{aligned}
 (2.12) \quad \beta_i^{r+1} &= \beta_i^r - (-s_{ii}^{-1} Y_i' Y_i)^{-1} (-s_{ii}^{-1} Y_i' u_i + T \beta^{i(i)}) \\
 &= \beta_i^r - (Y_i' Y_i)^{-1} Y_i' (-Y_i + Y_i \beta_i^r + Z_i \gamma_i^r) + s_{ii} (Y_i' Y_i)^{-1} T \beta^{i(i)r} \\
 &= (Y_i' Y_i)^{-1} (Y_i' Y_i - Y_i' Z_i \gamma_i^r + T s_{ii} \beta^{i(i)})
 \end{aligned}$$

Equation (2.12) can be interpreted as a correction to the method of ordinary least squares as applied to the model (2.2)

to adjust for the correlation between  $Y_i$  and  $u_i$ . Without the correction, the least squares estimate is

$$(2.13) \quad b_i = (Y_i' Y_i)^{-1} Y_i' (y_i - Z_i \gamma_i) = (Y_i' Y_i)^{-1} Y_i' (Y_i \beta_i - u_i) \\ = \beta_i - (Y_i' Y_i)^{-1} Y_i' u_i .$$

Hence, a consistent estimate of  $\beta_i$  can be obtained by adding to  $b_i$  the stochastic limit of  $(Y_i' Y_i)^{-1} Y_i' u_i$ . The latter can be obtained by post multiplying the transpose of the reduced form of (2.1) by  $\frac{1}{T} u_i$ ,

$$(2.14) \quad \frac{1}{T} Y_i' u_i = - \frac{1}{T} B_i'^{-1} \Gamma_i' Z_i' u_i + \frac{1}{T} B_i'^{-1} U_i' u_i \rightarrow B_i'^{-1} \begin{bmatrix} 0 \\ \vdots \\ \sigma_{ii} \\ \vdots \\ 0 \end{bmatrix} = \sigma_{ii} \begin{bmatrix} \beta_i^{i1} \\ \vdots \\ \beta_i^{iG} \end{bmatrix}$$

where the assumption of diagonal  $\Sigma$  is utilized. (2.14) implies that the limit of  $(Y_i' Y_i)^{-1} Y_i' u_i$  equals the limit of  $(Y_i' Y_i)^{-1} T \sigma_{ii} \beta_i^{i(i)}$ . Equation (2.12) results when this term is added to the least squares estimates of (2.13). The method of (2.12), when it converges, is also truly FIML for diagonal  $\Sigma$  since its solution will make the gradient (2.7) of the likelihood function vanish.

7. See Chow (1972) for the derivation of (3.4) and (3.5) below.
8. It should be noted that, if the system is linear, the  $B$  matrix of section 2 is the transpose of the matrix of the Jacobian and not the matrix of the Jacobian itself. The transposition is due to the adoption in section 2 of the

convention that the coefficients of each equation are contained in a column (not row) of the matrix  $B$ .

9. A lower triangular  $J$  corresponds to an upper triangular  $B$ . See footnote 8 on this point.
10. In using the method of least squares for a non-linear equation, one has the option of applying Newton's method to minimize  $\frac{1}{2} \sum s_{ii}$  or  $\frac{1}{2} \sum \log s_{ii}$ . The gradient and the Hessian for minimizing the latter expression are given respectively by (3.6) and (3.7). Which option will converge better remains an open question.
11. Ideas of this section have been developed by the author after attending a working conference on robust regression techniques sponsored by the NBER Computer Research Center in June, 1973. That conference has stimulated my interest in robust techniques and their possible application to the simultaneous-equation situation.
12. This method breaks down if a previous value of  $u_t$  is zero. In this case one may choose to replace it by a small number.
13. See, for example, Andrews et al. (1972).
14. See Chow (1964) for a development of this idea.