Volume Title: Annals of Economic and Social Measurement, Volume 2, number 1

Volume Author/Editor: NBER

Volume Publisher: NBER

Volume URL: http://www.nber.org/books/aesm73-1

Publication Date: 1973

Chapter Title: A Generalized Approach to Estimation As Implemented in the Troll/I System

Chapter Author: Mark Eisner, Robert S. Pindyck

Chapter URL: http://www.nber.org/chapters/c9837

Chapter pages in book: (p. 29 - 51)

Annals of Economic and Social Measurement, 2/1, 1973

A GENERALIZED APPROACH TO ESTIMATION AS IMPLEMENTED IN THE TROLL/1 SYSTEM*

BY MARK EISNER AND ROBERT S. PINDYCK†

This paper presents the theoretical background of the methods and algorithms used in the estimation capability in the TROLL/1 system. The TROLL/1 approach provides the ability to combine most state-of-the-art procedures into one estimation process. As a result, a consistent theoretical framework, which is presented in this paper, can also serve as a review of estimation techniques, taken from a generalized point of view.

I. INTRODUCTION

The estimation capability in TROLL/1 is designed to provide a complete set of regression techniques within an effective and usable framework. Over the last decade a number of advanced statistical techniques have been explored and accepted by the econometric community. However, it has often been difficult for the applied econometrician to readily obtain access to computer programs which provide these techniques. Even when programs existed for a particular procedure, the econometrician could not easily combine several different procedures into one estimation. We have tried to solve this problem in TROLL/1 by not only providing most major state-of-the-art procedures but also by presenting them as basic units which can be combined in any chosen manner.

Since TROLL/1 allows the combination of a variety of statistical procedures for the estimation of linear or nonlinear equations, it must, in a sense, be viewed as a "user-beware" system. Often combinations of procedures, particularly when used in nonlinear estimation, have questionable statistical properties, and the statistical interpretation of a great many other procedures has simply not been explored. Our approach to estimation does not attempt to answer all of these questions. Rather, we have attempted to devise a coherent and meaningful plan for combining regression techniques, and to present a consistent method for producing the statistical results of any of these estimation procedures. This paper

* The design, development, and maintenance of TROLL have been funded by research grants GS-2310 and GS-27353 from the National Science Foundation to the Massachusetts Institute of Technology. M.I.T. supported the development effort with considerable amounts of free computer time. Arrangements are being made for the maintenance of TROLL/1 by the National Bureau of Economic Research, Computer Research Center for Economics and Management Science. Inquiries about the system, comments about these documents, and requests for companion documents may be addressed to:

Support Staff Coordinator
NBER Computer Research Center
545 Technology Square
Cambridge, Massachusetts 02139
(telephone 617-661-8788)

The companion documents that are currently available or in advanced stages of preparation are: *A Researcher's Overview of the TROLL/1 System*; *Troll/1 Primer*; and *Reference Manual for TROLL/1 (preliminary edition).* © 1971 Massachusetts Institute of Technology.

† We would like to thank the many people who have contributed directly to this paper, and indirectly to it through their work on the development of TROLL/1. Specifically, Alfredo Pastor-Bodmer and Michael Lichstein were involved in the initial development of the approach, and many helpful comments were provided by Professors Franklin Fisher, Edwin Kuh, and J. Phillip Cooper.

will outline the integrated approach to estimation that we have applied to TROLL/1. We will also discuss the substantive algorithms and methods used in the TROLL/1 estimation capability so that a user can have an authoritative source from which he can interpret the results of regressions performed on the system.

II. General Approach

All of the estimation techniques that have been implemented in TROLL/1 can be divided into three basic operations: a least squares procedure, sets of data transformations, and statistical analyses. These operations, functionally organized as separate entities, are combined for each estimation technique. This involves determining the correct ordering of the operations as well as producing a scheme which would provide meaningful statistics. Procedures and transformations have been adopted ior estimating parameters that appear with nonlinearities.

The following procedures have been implemented:

(1) An ordinary least squares procedure for either linear or nonlinear equations
(2) distributed lag operators
(3) instrumental variable substitution
(4) single equation generalized least squares error correction
(5) a standard set of statistics based on the observed residuals as well as the coefficient covariance matrix

These procedures can be used as building blocks for almost any standard single-equation estimation technique. For example: (a) Two-stage least squares (2SLS) can be implemented by applying the instrumental variable transformation on all endogenous terms in a regression and then applying the ordinary least squares procedure. (b) A first-order autoregressive adjustment can be performed by applying the generalized least squares procedure with an appropriate error covariance matrix obtained by factoring.

Procedures may also be combined in a less stylized manner. As an extreme example, one could perform a regression on a non-linear equation, in which a polynomial distributed lag operator is applied to a term, instrumental variables are applied to some or all endogenous terms, and a second order autoregressive correction is performed on the whole equation.

TROLL has been designed so that additional procedures can be added in an efficient manner. These could include other single-equation estimation techniques besides least squares regression, such as generalized maximum likelihood, special small-sample estimators, etc. Plans are also underway to expand the design to include multi-equation regression techniques such as full information maximum likelihood and three-stage least squares.

III. The Ordinary Least Squares Algorithm

A. Choice of Method

The ordinary least squares algorithm implemented in TROLL/1 is generalized to handle either linear or non-linear equation forms. The method used is based on a procedure by Marquardt [13].

Given an equation of the form

$$0 = f(x, B) + \varepsilon$$

or,

(1) $$-f(x, B) = \varepsilon, \qquad \text{where } x = (x_1, \ldots, x_m) \text{ and } B = (B_1, \ldots, B_n)$$

in which f is a function of a set of data vectors (x) and a set of parameters (B) with an implicit additive error vector (ε), the problem is to estimate the coefficients so that the observed sum of squared residuals is minimized. Letting e represent the observed residuals, and $\hat{B}$ the estimated set of parameters, we must minimize the objective function:

(2) $$e'e = f(x, \hat{B})'f(x, \hat{B})$$

There are three standard approaches to the solution of this problem:

(1) A direct search technique may be used in which the objective function is evaluated for sets of values, and that set which results in a minimum is chosen as the estimated parameters. However, this method is often inordinately expensive in computation.

(2) A second approach is to form the first c.der conditions of the objective function and solve them. However, this may result in complicated nonlinear equations which will have to be solved. This approach can also be computationally expensive.

(3) A third approach, and the one which we have chosen for TROLL/1, involves linearizing the given equation so that it fits the form $(Y - XB) = \varepsilon$. This results in a quadratic objective function

(3) $$e'e = (Y - X\hat{B})'(Y - X\hat{B})$$

whose first order conditions can be expressed as an explicit solution for $\hat{B}$

(4) $$\hat{B} = (X'X)^{-1}X'Y$$

This approach has a number of advantages, the first of which is computational efficiency. If the equation to be estimated is linear to begin with, the procedure reduces directly to ordinary linear least squares. In the nonlinear case the process provides a clear guideline for incorporating statistical techniques which are usually only applied to linear regression.

B. The Iterative Process

1. *The Linearization.* We begin by expanding the function

(5) $$-f(x, B) = \varepsilon; \qquad B = (B_1, \ldots, B_n)$$

where B is a set of parameters, in a Taylor Series expansion around an arbitrary set of points B^0:

(6) $$-f(B^0) - \sum_{i=1}^{n} (B - B^0)_i \frac{\partial f}{\partial B_i}\bigg|_{B^0} + \ldots = \varepsilon$$

Now, using just the first two terms of the expansion, we can rewrite the expanded equation so that all of the known (given B^0 and $\partial f/\partial B_i|_{B^0}$) values are on the left hand side.

$$\text{(7)} \qquad \sum_{i=1}^{n} B_i^0 \frac{\partial f}{\partial B_i}\bigg|_{B^0} - f(B^0) = \sum_{i=1}^{n} B_i \frac{\partial f}{\partial B_i}\bigg|_{B^0} + \varepsilon$$

The original equation has now been approximated by a linear equation form in which the unknown B can be estimated using an ordinary linear regression procedure.

2. *Performing a linear regression.* The partial derivative of $f(B)$ with respect to a particular B_i has been named the "co-term" of that coefficient. After performing the Taylor expansion on an equation, we can examine the resulting set of co-terms $\partial f/\partial B_i$; $i = 1, \ldots, n$. If no co-term contains a coefficient a linear regression is indicated. The initial values (B^0) are automatically set to zero, and a single regression is performed. For example, given the linear equation

$$y = ax + b + \varepsilon$$

or,

$$y - ax - b = \varepsilon$$

the expansion results in the following form:

$$y - a^0x - b^0 + a^0x + b^0 = ax + b + \varepsilon$$

which is equivalent to a linear regression on the original equation.

A benefit of this approach when dealing with linear equations is that one is not restricted to expressing the equation in the traditional sum of products form. For example, consider the linear equation

$$y = (1 - a)x + b + \varepsilon$$

or,

$$y - (1 - a)x - b = \varepsilon$$

After expanding this we have:

$$y - (1 - a^0)x - b^0 - a^0x + b^0 = -ax + b + \varepsilon$$

Now, clearing terms, a linear regression on the transformed equation

$$y - x = -ax + b + \varepsilon$$

is actually performed.

3. *Performing a nonlinear estimation.* If after the expansion any co-term does contain a coefficient, a nonlinear regression is indicated. The initial values for B^0 are arbitrary, but typically a "reasonable guess" is used. (The TROLL/1 system sets each value of B^0 to 1 if no first guess is supplied.) The iterative process then involves the following steps:

(a) perform the regression on the linearized equation
(b) test to see if $\hat{B}$ is markedly different from B^0
(c) if it is set B^0 equal to $\hat{B}$ and go back to a step a.

Clearly, this process involves performing a linear regression on the residual values derived from subtracting the observed values from the function specified by B^0, selecting a better fitting function as specified by the new $\hat{B}$, and repeating the process until $\hat{B}$ and B^0 converge.[1]

The convergence criterion used to determine if $\hat{B}$ is close enough to B^0 is:

$$\max\left(\frac{|\hat{B} - B^0|}{\gamma + |B^0|}, \frac{|\hat{B} - B^0|}{\gamma + |\hat{B}|}\right) < \delta \tag{8}$$

This is a continuous function which acts as a percent change when $B \gg \gamma$ and a straight difference when $B \ll \gamma$. In order to insure that scale problems do not confuse the results, the ratios computed from both end-points $\hat{B}$ and B^0 must meet this requirement.

4. *Preventing round-off error*. Since the computational process can return only a limited number of digits of significance, round-off error can occur. This is especially true in the regression process due to the formation of a cross-product matrix.

For each matrix a "condition number" can be established which reflects the probability of generating round-off error in manipulations involving the matrix. The formation of the cross-product $(X'X)$ will result in a matrix whose condition number is the square of the condition number of the original data matrix X. Thus, if X is poorly conditioned in this sense, the conditioning of the cross-product matrix will be much worse.

Round-off errors can to a great extent be removed if the conditioning problem in the X matrix is reduced before the cross-product matrix is formed. This is accomplished in TROLL/1 by transforming the X matrix by a standard Gram–Schmidt orthonormalization process. The Gram–Schmidt orthonormalization process, although somewhat order-dependent, has been chosen for a number of important computational considerations.

5. *Convergence and damping*. The nonlinear estimation procedure described above has the property of converging rapidly to the solution point if it stays within the circle of convergence. However, at times the estimate $\hat{B}$ can overshoot or leave the circle of convergence completely during the iteration process. This can result in the method diverging instead of converging, or in oscillations around the correct solution. In order to solve this problem, a method for damping the computed step in B has been provided. Rather than taking the whole step $\hat{B} - B^0$ for the next iteration the following step is taken instead:

$$B_d = B^0 + \alpha(\hat{B} - B^0), \quad \text{where } 0 \leq \alpha < 1 \tag{9}$$

By applying the damping factor, the algorithm is in fact choosing a step somewhere between the one indicated by the Taylor expansion method and one which would be indicated if a steepest descent method were being employed.[2]

Various automatic and semi-automatic heuristics are included in TROLL to determine an appropriate damping strategy

[1] For a discussion and formal proof of the convergence of this algorithm, the reader is referred to Marquardt [13].

[2] For discussion, see Marquardt [13].

6. *Producing statistics for least squares regression.* The general equation form for least squares regression has been presented so far as:

$$-f(x, B) = \varepsilon$$

A more standard formulation would be

$$-f(x, B) = g(x, B) - h(x, B) = \varepsilon \tag{10}$$

or,

$$g(x, B) = h(x, B) + \varepsilon \tag{11}$$

where $h(x, B)$ or the right hand side (RHS) represents what are considered the "independent" terms of the equation and $g(x, B)$ or the LHS represents the "dependent" terms of the equation. Traditionally, the LHS contains no coefficients, but in our generalized approach this need not be the case.

To produce all the standard regression statistics, the following basic information must be available.

(a) the number of observations and number of co-terms
(b) a set of residuals generated from the estimated coefficients ($\hat{B}$)
(c) a covariance matrix for the estimated coefficients
(d) the variance of the LHS of the equation

For our generalized procedure this information is produced in a straight-forward manner. The covariance matrix for the estimated coefficients is, of course, based on the $(X'X)^{-1}$ matrix which is produced by the final iteration in the convergence process, and the variance of the LHS is determined by evaluating $g(x, B)$ and calculating the variance of the resulting vector.

The results calculated above are used to produce standard statistics, the meaning and properties of which are well-defined for the linear case. In the nonlinear case these statistics usually have meaning only in terms of the linearized equation at the solution point.

7. *Combining ordinary least-squares with other techniques.* Additional correction techniques such as instrumental variables or generalized least squares can be combined with a linear regression in the standard manner. If the regression is nonlinear, these techniques are applied at each iteration to the linear regression on the expanded equation form. In some cases, however, the statistical implication of these additional techniques is not clear for the nonlinear case. They should always be employed with this in mind, and then only after careful consideration of the particular regression to be performed.

IV. Distributed Lag Operators

A common equation form in estimation problems contains a weighted sum of co-terms lagged over a specified time range. This lag distribution for a co-term can be expressed in summation form as:

$$\sum_{\tau=0}^{n} p(a_\tau, x_{t-\tau})$$

Terms expressed in this form are referred to as distributed lag operators. Normal regression procedures can be applied directly to equations containing distributed lags but problems with collinearity between the lagged terms can often occur. Therefore, a number of methods have been derived to produce estimates of a_τ conditional on some constraint which reduces the problem of collinearity. The most popular of these procedures is the Almon process, which restricts the coefficients to values of a polynomial of degree less than or equal to the number of terms in the lag operator.

A. The Algorithm

Consider a lag operator within an equation:

$$- \sum_{\tau=0}^{n} p(a_\tau, x_{t-\tau}) - f(B, x_t) = \varepsilon \tag{12}$$

The Almon process constrains the estimates of a_τ to be on a polynomial of degree l such that

$$a_\tau = \sum_{j=0}^{l} w_j \tau^j \tag{13}$$

Linearizing equation (12) by the standard procedure yields

$$\sum_{\tau=0}^{n} a_\tau^0 \frac{\partial p}{\partial a_\tau}\bigg|_{a^0} + \sum_{i=1}^{m} B^0 \frac{\partial f}{\partial B_i}\bigg|_{B^0} - \sum_{\tau=0}^{n} p(a_\tau^0) - f(B^0) = \sum_{\tau=0}^{n} a_\tau \frac{\partial p}{\partial a}\bigg|_{a^0} + \sum_{i=1}^{m} B \frac{\partial f}{\partial B}\bigg|_{B^0} \tag{14}$$

now substituting for a_τ on the right-hand side:

$$\sum_{\tau=0}^{n} \left(\sum_{j=0}^{l} w_j \tau^j \right) \frac{\partial p}{\partial a}\bigg|_{a^0} + \sum_{i=1}^{m} B \frac{\partial f}{\partial B}\bigg|_{B^0} \tag{15}$$

let

$$X^{\mathrm{INC}} \equiv \frac{\partial p}{\partial a_1}, \ldots, \frac{\partial p}{\partial a_n} \tag{16}$$

$$X^{\mathrm{UN}} \equiv \frac{\partial f}{\partial B_1}, \ldots, \frac{\partial f}{\partial B_m} \tag{17}$$

Here, X^{INC} is the matrix of "included" (lagged) co-terms, and X^{UN} the matrix of "unincluded" (unlagged) co-terms. Then the right hand side of equation (14) can be rewritten as:

$$X^{\mathrm{INC}} a + X^{\mathrm{UN}} B \tag{18}$$

or from equation (15)

$$Zw + X^{\mathrm{UN}} B, \qquad \text{where } Z = X^{\mathrm{INC}} S$$

and S is an nxl "scrambling" matrix generated by the double sum

$$\sum_{\tau=0}^{n} \sum_{j=0}^{l} \tau^j \quad \text{i.e., } S = \begin{bmatrix} 0^0 & 0^1 \ldots 0^l \\ 1^0 & 1^1 \ldots 1^l \\ \vdots & \\ n^0 & n^1 \ldots n^l \end{bmatrix} \tag{19}$$

Note that the elements in this matrix are in effect arbitrarily determined by the indexing scheme used to express the polynomial, and therefore, another scrambling matrix which represents the same summation range but which avoids problems with zero index values can be chosen. An appropriate S matrix can be formed by constructing $n + 1$ rows, the elements of which represent the coefficient of a Lagrangian interpolation polynomial. It has the following structure:[3]

$$S = \begin{bmatrix} 1 & \dfrac{1}{n+2} & \left[\dfrac{1}{n+2}\right]^2 & \cdots & \left[\dfrac{1}{n+2}\right]^l \\ 1 & \dfrac{2}{n+2} & \cdots & & \\ 1 & \dfrac{3}{n+2} & \cdots & & \\ \vdots & \vdots & & & \\ 1 & \dfrac{n+1}{n+2} & \left[\dfrac{n+1}{n+2}\right]^2 & \cdots & \left[\dfrac{n+1}{n+2}\right]^l \end{bmatrix}$$

B. The Process

The co-terms included in the lag operator, X^{INC}, are extracted and multiplied by the scrambling matrix S to produce a transformed data matrix Z.

$$Z = X^{\text{INC}}S \tag{20}$$

This constructed Z matrix is combined with the matrix of co-terms of the unincluded variables:

$$\tilde{X} = [Z \mid X^{\text{UN}}] \tag{21}$$

Ordinary least squares is then applied resulting in a coefficient vector which is composed of the polynomial weighting coefficients and the coefficients of the unincluded variables. This expanded vector is given by:

$$\begin{bmatrix} \hat{w} \\ \hat{B} \end{bmatrix}$$

The correct coefficients, $\hat{a}_\tau$, are produced from the estimated weights $\hat{w}$;

$$\hat{a} = S\hat{w} \tag{22}$$

The final coefficient vector is found by concatenating $\hat{a}$ and $\hat{B}$, i.e., is given by:

$$\begin{bmatrix} \hat{a} \\ \hat{B} \end{bmatrix}$$

It is often desirable to apply "zero-restrictions" to the polynomial terms of the transformation. It is possible to set any weight w_i to zero by dropping out the corresponding column from the scrambling matrix. For example, to set the first

[3] For an elaboration see Cooper [5].

term w_0 to 0, remove the first column from the S matrix. Similarly, one may force the tail of the polynomial to zero by the same procedure.

Equation forms containing more than one polynomial distributed lag operator can be estimated within our generalized framework, and there is, of course, no reason why the term within the lag operator need be linear.

For example, the following equation could be estimated:

$$Y = \sum_{\tau} a_\tau x_{t-\tau}^{b} + \sum_{\tau} c_\tau z_{t-\tau} + \varepsilon$$

Statistics consistent with ordinary least squares are produced for regressions containing a distributed lag. Residuals are generated by evaluating the original equation with the final coefficient estimates $\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix}$ and the original data. The covariance matrix produced by the regression, however, is generated from the datamatrix which contains the constructed variables; i.e.,

$$\tilde{C} = [(Z \quad X)' \quad (Z \quad X)]^{-1} \tag{23}$$

To transform this into the correct asymptotic covariance matrix, a transformation matrix D is formed from the original scrambling matrix and an identity matrix with the rank of the additional co-terms X^{UN}:

$$D = \left[\begin{array}{c|c} S & 0 \\ \hline 0 & I_{\mathrm{UN}} \end{array}\right] \tag{24}$$

The new covariance matrix is produced by

$$C = D\tilde{C}D' \tag{25}$$

V. Instrumental Variable Transformations

A basic assumption when performing a least squares regression is that the implicit error term is uncorrelated with any of the co-terms in the equation. If this assumption is violated for a particular co-term, an instrument can be created which is statistically independent of the error and which can then be used in place of the co-term. This instrument can be created by regressing the co-term on a set of variables which are assumed to be uncorrelated with the error term but are correlated with the co-term. Since this constructed variable is a linear combination of terms which are uncorrelated with the error, it too is uncorrelated with the error and can be used in the estimation in place of the original co-term.

A. The Process

Given an equation:

$$-f(x, B) = g(x, B) - h(x, B) = \varepsilon$$

a set of co-terms correlated with ε:

$$X_i = \frac{\partial f}{\partial B_i}; \qquad i = 1, \ldots, k$$

and a set of variables $W = (W_1, \ldots, W_s)$ which are assumed to be uncorrelated with ε but correlated with the X_i, perform the regression on the equations

$$X_i = WP_i + v_i \tag{26}$$

where P_i are a set of coefficients and v_i is an implicit error term. Then

$$\hat{P}_i = (W'W)^{-1}W'X_i \tag{27}$$

$$\hat{X}_i = W\hat{P}_i, \quad \text{and} \quad v_i = X_i - \hat{X}_i \tag{28}$$

where v_i are the observed residuals from the regression. This process is repeated for all X_i.

Now to perform a regression on $-f(x, B)$, expand the equation:

$$\sum_i B^0 \frac{\partial f}{\partial B_i} - f(B^0) = \sum_i B \frac{\partial f}{\partial B_i} + \varepsilon \tag{29}$$

and substitute the constructed variables for those co-terms assumed to be correlated with ε

If $\hat{X}_i$ = the constructed variables

X_j = the other co-terms

and

$$Y = \sum_i B^0 \hat{X}_i + \sum_j B^0 X_j - f(B^0) \tag{30}$$

then the equation to be regressed becomes

$$Y = \sum_j B_j X_j + \sum_i B_i \hat{X}_i + \left(\varepsilon + \sum_i B_i v_i\right) \tag{31}$$

It should be noticed that the residuals $v_i = X_i - \hat{X}_i$ will appear in the error term of this equation. Estimates for B are given by:

$$\hat{B} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y \tag{32}$$

where

$$\hat{X} = [X_j \mid \hat{X}_i] \tag{33}$$

B. Statistics for Instrumental Variables

Residuals are obtained by evaluating the original function using the final estimates of the coefficients:

$$e = -f(x, \hat{B}) \tag{34}$$

Note that the original data and not the constructed co-terms should be used in calculating the residuals. This insures that the estimated variance of the regression corresponds to the structural disturbances not compounded by first-stage residuals.

Given the final estimates of the coefficients $\hat{B}$ the variance-covariance matrix is defined as

$$\text{cov} = E[(\hat{B} - B)(\hat{B} - B)'] \tag{35}$$

where B are the true values of the coefficients. From equations (31) and (32):

$$\begin{aligned} \hat{B} - B &= (\hat{X}'\hat{X})^{-1}\hat{X}'Y - B \\ &= (\hat{X}'\hat{X})^{-1}\hat{X}'[\hat{X}B + \varepsilon + vB] - B \\ &= (\hat{X}'\hat{X})^{-1}[\hat{X}'\hat{X}B + \hat{X}\varepsilon + \hat{X}vB] - B \end{aligned} \tag{36}$$

but $\hat{X}$ is by construction orthogonal to v, so

$$\begin{aligned} \hat{B} - B &= B + (\hat{X}'\hat{X})^{-1}\hat{X}\varepsilon - B \\ &= (\hat{X}'\hat{X})^{-1}\hat{X}\varepsilon \end{aligned} \tag{37}$$

and the variance-covariance matrix is

$$\text{cov} = \sigma_\varepsilon^2(\hat{X}'\hat{X})^{-1} \tag{38}$$

Notice that the variance σ_ε^2 is based on the residuals from the original equation, and that the $(\hat{X}'\hat{X})^{-1}$ is the inverse of the cross-product matrix of co-terms which results from the final regression procedure which includes the instrumental variable substitution.

C. *Using Instrumental Variables for Nonlinear Equations*

Instrumental variables should be used in a nonlinear regression if it is believed that one or more co-terms are correlated with the implicit error term.

For example, given the nonlinear equation

$$y = abx_1 + bx_2 + \varepsilon \tag{39}$$

the expanded form is

$$y - a^0b^0x_1 - b^0x_2 + a^0(b^0x_1) + b^0(a^0x_1 + x_2) = a(b^0x_1) + b(a^0x_1 + x_2) \tag{40}$$

or

$$y + a^0(b^0x_1) = a(b^0x_1) + b(a^0x_1 + x_2) \tag{41}$$

the co-terms are

$$\frac{\partial f}{\partial a} = b^0x_1 \tag{42}$$

$$\frac{\partial f}{\partial b} = (a^0x_1 + x_2) \tag{43}$$

It may certainly be assumed that the co-term of either a or b is correlated with ε and that instrumental variables should be used to correct this problem.

The value of a nonlinear co-term can change during each iteration of the solution process, i.e., the co-term b^0x_1 changes for each new value of b^0 obtained. Therefore, a new constructed variable for that co-term must be evaluated at each iteration of the solution process. This process can, of course, be computationally expensive.

D. Principal Component Transformation of a Set of Instruments

The set of variables which are used to construct instruments can often be quite large, especially when a two-stage least squares procedure is followed. For example, the entire set of exogenous variables as well as some lagged endogenous variables may be chosen. A large set of variables greatly increases the complexity and cost of the process, or can even result in a negative number of degrees of freedom in the first stage.

The principal components transformation produces a new set of variables which are orthogonal linear combinations of the original variables. These new variables are ordered so that each variable explains as much of the remaining variance of the original variables as possible. As a result, it is often possible to use a much smaller set of variables while still accounting for the major fraction of the variance explained by the original variables.

Given a matrix of the original variables W which is $n \times s$ where $n =$ number of observations and $s =$ number of predetermined variables, to create the principal components form the $(s \times s)$ correlation matrix $r = (r^{ij})$, where

$$r_{ij} = \frac{(1/n)\sum_t (W_{it} - \overline{W}_i)(W_{jt} - \overline{W}_j)}{\sqrt{(1/n)\sum_t (W_{it} - \overline{W}_i)^2}\sqrt{(1/n)\sum_t (W_{jt} - \overline{W}_j)^2} \tag{44}$$

Find the characteristic roots, and the characteristic vectors of the matrix R.

$$\lambda_i, i = 1, \dots, s \qquad e_i, i = 1, \dots, s$$

Then order these roots and vectors such that

$$\lambda_1 > \lambda_2 > \dots > \lambda_m$$

The factor loadings for the first principal component are then found by normalizing the eigenvectors

$$\alpha_i = e_i/\lambda_i \tag{45}$$

for the desired set of principal components.

Form the $s \times r$ matrix A, where

$$A = [\alpha_1, \alpha_2, \dots, \alpha_r]$$

$$s = \text{number of original instruments}$$

$$r = \text{number of desired instruments}$$

The new set of instruments is created from

$$\tilde{W} = WA \tag{46}$$

Note that this new set of instruments can be significantly smaller than the original set of instruments W.

VI. Generalized Least Squares for a Single Equation

A basic assumption of least-squares regression is that each implicit error term comes from a population with a constant variance, and that each error is independent of any other error.

Thus the variance-covariance matrix of the errors must be

(47) $$E[\varepsilon\varepsilon'] = \sigma^2 I$$

If this is not the case, ordinary least squares will result in estimates that are unbiased and consistent, but that are not efficient. It is not unusual, however, for this assumption to be violated. If one is performing regressions on cross-sectional data related to a set of firms, the error variance for each observation may be related to the size of the firm, and the diagonal elements of the error covariance matrix will not be constant. Often in regressions on time-series data it is reasonable to assume that errors occurring in previous time periods will be correlated with errors in the current time period, since the arbitrary time divisions used in the analysis do not correspond to the actual continuous process being analyzed. This relation between the current and previous error terms will result in non-zero off-diagonal terms in the variance-covariance matrix.

(48) $$E[\varepsilon\varepsilon'] = \sigma^2 V$$

A. Generalized Least Squares Correction

In order to provide an efficient estimate using least squares, it is necessary to transform the error term ε so that its error covariance matrix is of the correct form.

A transformation matrix *A* can be constructed such that

(49) $$\tilde{\varepsilon} = A\varepsilon$$

and

(50) $$E[\tilde{\varepsilon}\tilde{\varepsilon}'] = \sigma^2 I$$

If the covariance of the error process is known, the matrix *A* is determined as follows:

(51) $$\begin{aligned} E[\tilde{\varepsilon}\tilde{\varepsilon}'] &= E(A\varepsilon\varepsilon' A') = \sigma^2 I \\ &= AE(\varepsilon\varepsilon')A' = \sigma^2 I \\ &= A\sigma^2 VA' = \sigma^2 I \\ &= AVA' = I \end{aligned}$$

premultiply by A^{-1} and postmultiply by $(A')^{-1}$

(52) $$\begin{aligned} V &= A^{-1}A'^{-1} \\ V &= (A'A)^{-1} \end{aligned}$$

and

$$(A'A) = V^{-1}$$

Therefore the correct transformation matrix is constructed by factoring[4] the inverse of the variance-covariance matrix of the error term ε.

[4] A Choleski triangular factoring procedure is used. See Faddeev [7], p. 144.

A regression on this transformed error term will minimize the following objective function

$$\Phi = \tilde{e}'\tilde{e} = e'A'Ae = e'V^{-1}e \tag{53}$$

This approach also maximizes the likelihood function, i.e., given an error term ε with an assumed multivariate normal distribution and a known covariance matrix V the likelihood function

$$L(B) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} e^{-(1/2)e'V^{-1}e} \tag{54}$$

is maximized when the negative exponent is minimized, or when the objective function

$$\Phi = e'V^{-1}e \tag{55}$$

is minimized.

B. The Process

Given an equation form

$$-f(x, B) = \varepsilon \tag{56}$$

in which the covariance matrix of ε is known

$$E(\varepsilon\varepsilon') = \sigma^2 V \tag{57}$$

A transformation matrix A is constructed by factoring V^{-1}, i.e., A is formed such that

$$A'A = V^{-1} \tag{58}$$

The equation is linearized, and the linearized equation is then premultiplied by the transformation matrix A:

$$A\left[\sum_i B^0 \frac{\partial f}{\partial B_i} - f(B^0)\right] = \sum_i B\left(A\frac{\partial f}{\partial B_i}\right) + A\varepsilon \tag{59}$$

The ordinary least squares procedure is then invoked on this transformed equation.

C. Generalized Least Squares with a Symbolic Error Covariance Matrix

The preceding discussion assumed that the error covariance matrix was known, i.e., it could be numerically specified. However, when dealing with time-series analysis, the error process is often expressed as a function of a set of parameters. For example, the user may hypothesize that the error term follows a second-order autoregressive process of the form

$$\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \varepsilon_t^* \tag{60}$$

or that it follows a first-order moving average process of the form:

$$\varepsilon_t = \rho_1\varepsilon_{t-1}^* + \varepsilon_t^* \tag{61}$$

In these cases the error covariance matrix and its inverse will be a function of the parameters ρ_i.

To perform generalized least squares we must determine the values of the parameters ρ_i which minimize the least squares objective function

$$\Phi = e'V_\rho^{-1} e \tag{62}$$

$$\Phi = (A_\rho f(x, B))'(A_\rho f(x, B)) \tag{63}$$

The standard procedure for minimizing this function would be to derive the first order conditions, and solve the resulting set of simultaneous equations. These conditions are:

$$\frac{\partial \Phi}{\partial B_i} = 2A_\rho f(x, B)\frac{\partial f}{\partial B_i} = 0; \qquad i = 1, \ldots, n \tag{64}$$

and

$$\frac{\partial \Phi}{\partial \rho_j} = 2A_\rho f(x, B)\frac{\partial A(\rho)}{\partial \rho_j} = 0; \qquad j = 1, \ldots, r \tag{65}$$

It is normally difficult to solve these first order conditions directly, but, by partitioning the system, the problem can be greatly simplified. If an initial guess is made for ρ the first set of equations (64) reduces to a standard generalized least squares regression with a given transformation matrix A. This regression will produce a set of B's which can be used in the second set of equations to produce new values for ρ.

Given a first guess for ρ, which we will call ρ^0, the standard generalized least squares procedure will produce an estimate of B and as a result a set of observed residuals $\hat{e}$.

Given this set of residuals $\hat{e}$, the objective function

$$\Phi = (A_\rho \hat{e})'(A_\rho \hat{e}) \tag{66}$$

can be minimized via a simple direct search procedure.[5] This results in a new set of parameters $\hat{\rho}$ which are then used in the generalized least squares procedure to produce a new set of $\hat{B}$. The process is continued until no significant change can be made in either the ρ's or the B's.

The process described above is similar in some ways to the familiar Hildreth–Lu [10] procedure for correcting for first-order Markov serial correlation. In this case, a single ρ needs to be estimated such that:

$$-1 \leq \rho \leq 1$$

The solution process is again partitioned into two steps but instead of evaluating a new ρ from a previous guess the domain of ρ is divided into an equally partitioned grid. The grid point which produces the smallest sum of squared residuals is considered the solution point. A grid search has the advantage of more certainly finding the neighborhood of a global minimum as compared to the direct search which may result in a local minimum. However, when the number of ρ's to be estimated is greater than one, the process can be prohibitively expensive.

[5] See Hildreth–Lu [10] or Box [2], Chapter 3.

For example, assume that a grid of twenty points is used for the first order Hildreth–Lu process. To achieve the same accuracy with a second-order process, four hundred grid points would have to be evaluated. One can achieve a compromise between the two methods by performing a thin grid with a small number of distinct regressions in which the $\hat{\rho}$ is fixed. The regression with the smallest sum of squares can be considered to be at a relatively "global" minimum. Then a direct search can be undertaken to improve the estimate of this minimum point.

D. Use of Generalized Least Squares

The generalized least squares procedure described above can be applied either to linear or nonlinear equations. However, if a symbolic transformation matrix A is used in conjunction with a nonlinear equation, two iterative processes are involved and extensive computation may be required to reach a solution. The process is well defined if a symbolic form is provided for the matrices A, V, or V^{-1}. However, it is only computationally efficient if the symbolic matrix A is supplied directly. Symbolic A matrices are available for example, for autoregressive and moving average error processes.

E. Statistics for Generalized Least Squares

When applying generalized least squares, the standard procedure for deriving statistics must be slightly modified.

If we return to the original equation and evaluate it to obtain the residuals, values for the untransformed error term will be produced. However, the statistics generated for least-squares are valid only if the error terms satisfy the least-squares assumptions, i.e., is serially uncorrelated and is homoscedastic. Therefore, these residuals must be transformed to reflect the transformed errors. This is accomplished by multiplying the resulting residuals by the specified transformation matrix, i.e., forming

$$\tilde{e} = Ae \tag{67}$$

It is these transformed residuals, $\tilde{e}$, that are used in deriving all statistics for generalized least squares.

The correct matrix to use in producing the variance-covariance matrix of the coefficients when performing generalized least squares is the inverse of the cross-products matrix generated in the regression at solution.[6] This can be seen clearly by substituting the following transformed variables in the standard derivation of the covariance matrix of the coefficients:

$$\tilde{X} = AX \tag{68}$$

$$\tilde{e} = Ae \tag{69}$$

$$\tilde{Y} = AY \tag{70}$$

[6] We assume here that there are no lagged dependent variables in the equation.

then

$$(71)\qquad \begin{aligned}\hat{B} - B &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y - B\\ &= (\tilde{X}'\tilde{X})^{-1}[\tilde{X}B + \tilde{\varepsilon}] - B\\ &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\varepsilon}\end{aligned}$$

and

$$(72)\qquad \begin{aligned}E[(\hat{B} - B)(\hat{B} - B)'] &= E[(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\varepsilon}\tilde{\varepsilon}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}]\\ &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'E[\tilde{\varepsilon}\tilde{\varepsilon}']\tilde{X}(\tilde{X}'\tilde{X})^{-1}\\ &= \sigma^2(\tilde{X}'\tilde{X})^{-1}\\ &= \sigma^2((AX)'(AX))^{-1}\\ &= \sigma^2(X'V^{-1}X)^{-1}\end{aligned}$$

The calculation of the variance of the dependent term (i.e., the LHS) also requires a variable transformation, i.e., given $w = g(x, B)$ as the LHS, the transformed variable

$$\tilde{w} = Aw$$

is used in constructing the LHS covariance.

VII. Combining Statistical Procedures

It would not be unusual for an econometrician to want to combine two or more of the statistical techniques that have been described in the previous sections of this paper. For example, one might want to apply both generalized least squares and instrumental variables to a nonlinear equation with a polynomial distributed lag operator on one or more of its co-terms. One of the greatest strengths of a system such as TROLL/1 is, in fact, that it allows one to freely combine statistical techniques in this way.

Questions arise, however, as to how these techniques should be combined, and in particular, in what order they should be combined. For example, should a generalized least squares transformation be made before or after an instrumental variable sLbstitution is made? These ordering questions must be resolved so as to best insure consistency and efficiency in the estimates—directly in the case of a linear estimation, and for each iterative linearization in the case of a nonlinear estimation.

After the proper ordering of techniques is resolved, questions still remain as to the method of obtaining statistics when techniques are combined within a single estimation. How, for example, should residuals be calculated when generalized least squares is combined with instrumental variables, and what is the proper asymptotic variance-covariance of the estimated coefficients?

It may be that some of the standard statistics that econometricians are used to looking at have no meaning in estimation problems such as the one mentioned above. It is not a goal in this paper, however, to attempt to prove or even demonstrate when this is or is not the case. In the end the econometrician will have to make this decision for himself, based on considerations of his particular estimation problem.

It is our goal to set forth what we believe is a sensible way to order procedures and calculate standard statistics. Unfortunately, even large-sample statistical properties are unknown for many of the estimation problems that we will be faced with. As a result, our arguments and approach will be largely heuristic, and undoubtedly some questions will remain unresolved.

A. The Basic Ordering of Statistical Procedures

We outline below the basic ordering which is used for the procedures discussed in previous sections of this paper. For single equation estimation:

(1) Generalized least squares transformation
(2) Instrumental variable substitution (including two-stage least squares)
(3) Polynomial distributed lag operators
(4) Ordinary least squares procedure (as applied to linear or nonlinear equations)

If, for example, a set of three equations was to be estimated by using all four of the above procedures, the method would be as follows. First, each equation would be transformed by its corresponding GLS A matrix. The A matrices might be known, or they might be in symbolic form.

Next, the instrumental variable substitution will be made. Note that the first-stage regression will be run on the transformed (by A) variables of the original equation.

Finally, one or more polynomial distributed lag constraints can be imposed. This is done by transforming the data with the scrambling matrix S described in section IV. If one or more of the equations happened to be nonlinear or if the GLS A matrix was symbolic, then steps (1) to (4) would be repeated iteratively until convergence.

The following pages describe the logic behind this ordering approach, as well as the methods for calculating standard statistics.

B. Combining Generalized Least Squares with Instrumental Variables

The method described in this section is a more general alternative to the one proposed by Fair [6] for combining GLS and Two-Stage Least Squares for the case of an autoregressive error process.

We begin by considering the case of a regression on a linear equation, i.e., an equation of the form

$$y = xb + \varepsilon \tag{73}$$

Assume that x is correlated with ε, and that ε has a *known* variance-covariance matrix:

$$E[\varepsilon\varepsilon'] = \sigma^2 V \tag{74}$$

It is easy to see that an application of instrumental variables first and then generalized least squares will result in an estimate of b for which there is no guarantee of consistency.

Suppose W, for example, is a set of instruments, and we performed the first-stage regression:

$$x = WP + v \tag{75}$$

resulting in

$$x = W\hat{P} + v \tag{76}$$

$$\hat{x} = W\hat{P} \tag{77}$$

If we then substituted $\hat{x}$ into the original equation we would have:

$$Y = \hat{x}b + (\varepsilon + vb) \tag{78}$$

Now to estimate b by generalized least squares we would have to find the matrix A such that

$$A'A = V^{-1} \tag{79}$$

and then transform our equation to yield:

$$AY = A\hat{x}b + (A\varepsilon + Avb) \tag{80}$$

Note that we have no guarantee that $A\hat{x}$ will be uncorrelated with $(A\varepsilon + Avb)$ and, therefore, no guarantee of consistency. Also, consider the variance-covariance matrix of the error term in this transformed equation:

$$\begin{aligned} E[A\varepsilon + Avb)(A\varepsilon + Avb)'] = E[A\varepsilon\varepsilon'A'] + [A\varepsilon b'v'A'] + E[Avb\varepsilon'A'] \\ + E[Avbb'A'] \end{aligned} \tag{81}$$

While the probability limit of the second and third terms in the above equation is zero, the fourth term might introduce heteroscedasity, which is exactly what we had hoped to eliminate by using generalized least squares.

The solution is simply to reverse the order in which these procedures are applied. In other words, begin by applying the GLS transformation to the original equation:

$$Ay = Axb + A\varepsilon \tag{82}$$

Next, if W is the set of instruments, regress:

$$Ax = WP + v \tag{83}$$

giving us

$$\widehat{Ax} = W\hat{P}$$

Now perform ordinary least squares on the equation:

$$Ay = \widehat{Ax}b + (A\varepsilon + vb) \tag{84}$$

If the instruments were chosen properly, $\widehat{Ax}$ will be uncorrelated with both $A\varepsilon$ and vb and a consistent estimate will result. This is essential for the use of the combined procedures.

In addition, the procedure also preserves the efficiency of the estimate by maintaining a homoscedastic variance-covariance matrix. Consider the variance covariance matrix of the new error term

$$(85)\qquad E[(A\varepsilon + vb)(A\varepsilon + vb)'] = E[A\varepsilon\varepsilon'A'] + [A\varepsilon b'v'A'] + E[Avb\varepsilon'A'] + E[vbb'v']$$

Again the probability limit of the second and third terms is zero, but the expected value of the fourth term is a scalar diagonal.

If the original equation happened to be nonlinear, the same procedure would hold, but would be repeated at each iteration. In other words, if the equation was of the general form,

$$(86)\qquad -f(x, B) = \varepsilon$$

it would be linearized, and then be premultiplied by the transformation matrix A:

$$(87)\qquad A\left[\sum_i B^0\frac{\partial f}{\partial B_i} - f(x, B^0)\right] = \sum_i B_i\left(A\frac{\partial f}{\partial B_i}\right) + A\varepsilon$$

An instrumental variable substitution could then be applied to the co-terms. If W was the set of instruments, we would obtain:

$$(88)\qquad \widehat{A\frac{\partial f}{\partial B}} = W\hat{P}$$

Ordinary least squares would then be applied to the following equation:

$$(89)\qquad A\left[\sum_i B^0\frac{\partial f}{\partial B_i} - f(x, B^0)\right] = \sum_i B_i\left(\widehat{A\frac{\partial f}{\partial B_i}}\right) + A\varepsilon$$

If the variance-covariance matrix V of the error term was in symbolic form (i.e., a function of one or more unknown parameters) the same ordering of procedures would apply. The estimation, however, would now also involve an iterative process over the unknown parameters in the V matrix, and the instrumental variable substitution would have to be repeated for each iteration.

Consider the case of a second-order autoregressive error process, i.e., error terms of the form

$$(90)\qquad \varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \varepsilon_t^*$$

We would begin by using a "first guess" for ρ_1 and ρ_2 to calculate the V^0 matrix and the resulting transformation matrix A^0. (If no "first guess" could be supplied, the initial values of ρ_1 and ρ_2 could be set to zero, making $A^0 = I$.) The equation to be estimated would then be linearized around B^0 and premultiplied by the transformation matrix A^0. The instrument variable substitution would then be applied, and the resulting equation will be of the form

$$(91)\qquad A_\rho^0\left[\sum_i B^0\frac{\partial f}{\partial B_i} - f(x, B^0)\right] = \sum_i B_i\left(\widehat{A_\rho^0\frac{\partial f}{\partial B_i}}\right) + A_\rho^0\varepsilon$$

Ordinary least squares would be applied to (91), and the resulting residuals would be used to calculate new estimates for ρ_1 and ρ_2 using the search procedure described previously. A new transformation matrix, $\hat{A}_\rho$ would result. This process would be repeated until ρ_1 and ρ_2 converged. Note, however, that at each iteration a new instrumental variable regression and substitution must be performed, and the computational cost of this procedure could therefore be considerable if the equations are nonlinear.

We remind the reader again that in the case of a nonlinear estimation, all statistics (e.g., standard error, t's, etc.) relate to the linearized regression in the last iteration of the process. We can thus discuss the calculation of statistics by considering the linearized regression:

$$Y = Xb + \varepsilon \tag{92}$$

where X is the matrix of co-terms $(\partial f/\partial B_i)$ and Y is the constant term:

$$Y = \sum_i B^0 \frac{\partial f}{\partial B_i} - f(B^0)$$

We now consider the problem of calculating the residuals. We explained in section VI that when generalized least squares is used, the only relevant residuals for use in calculating statistics are the transformed residuals:

$$\tilde{e} = A_\rho e \tag{93}$$

where A is the GLS transformation matrix and e is the residual vector from the original equation.

When generalized least squares and instrumental variables are combined in the linear regression above, the question arises as to whether the transformed residuals should be calculated with or without the instrumental variable substitution. As explained in section V, residuals should be calculated without the substitution. In other words, the residuals

$$\tilde{e} = AY - AX\hat{b} \tag{94}$$

should be used, and *not* the residuals

$$\tilde{e}_1 = AY - \widehat{AX}\hat{b} \tag{95}$$

Finally, note that the proper asymptotic variance-covariance matrix must contain the instrumental variable substitution. For a linear regression,

$$\begin{aligned} \hat{b} &= [(\widehat{AX})'(\widehat{AX})]^{-1}(\widehat{AX})'AY \\ &= [(\widehat{AX})'(\widehat{AX})]^{-1}(\widehat{AX})'[(\widehat{AX})b + A\varepsilon + vb] \end{aligned} \tag{96}$$

where v are the residuals in the instrumental variable regression. Since $\widehat{AX}$ is orthogonal to v,

$$\hat{b} = b + [(\widehat{AX})'(\widehat{AX})]^{-1}(\widehat{AX})'A\varepsilon \tag{97}$$

Then the variance-covariance matrix of the estimated coefficient $\hat{b}$ is given by:

$$E[(\hat{b} - b)(\hat{b} - b)'] = \sigma^2_{A\varepsilon}[(\widehat{AX})'(\widehat{AX})]^{-1} \tag{98}$$

Here we use the estimated value of the error variance which is calculated exactly as it is in the simple GLS case, i.e., using the transformed residuals $\tilde{e}$.

C. Combining Polynomial Distributed Lags with Generalized Least Squares and Instrumental Variables

This section discusses the application of a polynomial distributed lag (PDL) operator to a regression that also involves a generalized least squares transformation and an instrumental variable substitution. As mentioned before, the PDL operation can be viewed as simply a constrained estimation in which the included co-terms X^{INC} are postmultiplied by the constraining "scrambling" matrix S before the ordinary least squares procedure is applied. Now, if a GLS transformation and an instrumental variable substitution are also to be applied, the only question is at what point should the PDL constraint (i.e., postmultiplication by S) be imposed.

Our approach (and we offer no formal proof of its validity at this time) is to include the lagged data in the original data matrix but to apply the PDL constraint after the GLS transformation and the instrumental variable substitution has been made. In other words, we apply the constraint after the data has been "cleaned up," i.e., after problems of heteroscedasticity, serial correlation, and correlations between co-terms and error terms have been removed. Our method is outlined below. We begin with a linearization of the basic regression equation:

$$(99)\qquad \sum a_\tau^0 \frac{\partial f}{\partial a_\tau} + \sum B_i^0 \frac{\partial f}{\partial B_i} - f(B^0, a_\tau^0) = \sum a_\tau \frac{\partial f}{\partial a_\tau} + \sum B_i \frac{\partial f}{\partial B_i} + \varepsilon$$

(a) We obtain an estimate (or first-guess) of the error variance-covariance matrix V_ρ, and factor this to get the transformation matrix A_ρ. We then premultiply the linearized equation by A_ρ.

(b) We next perform the instrumental variable regression and make the substitution into the linearized equation. We then have:

$$(100)\ A_\rho\left[\sum a_\tau^0 \frac{\partial f}{\partial a} + \sum B_i^0 \frac{\partial f}{\partial B_i} - f(B^0, a_\tau^0)\right] = \sum a_\tau \widehat{\left(A_\rho \frac{\partial f}{\partial a_\tau}\right)} + \sum B_i \widehat{\left(A_\rho \frac{\partial f}{\partial B_i}\right)} + A\varepsilon$$

(c) At this point the PDL constraining transformation is applied. Those co-terms which have been expanded in a sum of lags and are to be included in the PDL operation are postmultiplied by the scrambling matrix S. Note that these are no longer the same co-terms that appeared in the regression equation—they have already undergone a GLS transformation and may also have undergone an instrumental variable substitution.

(d) Ordinary least squares is applied

(e) The results are unscrambled, i.e., transformed back into the unconstrained data space. Residuals are calculated and a new estimate of V_ρ is obtained. We can then go back to step (a), and repeat the process until convergence is reached.

As before, in producing regression statistics we must be concerned with the method of calculating the residuals, and the method of calculating the variance-covariance matrix of the coefficients.

The residuals should be calculated exactly as they were when GLS and instrumental variables were combined without a PDL operation. In other words, simply take the residuals from the original equation (without the instrumental variable substitution), and multiply them by the transformation matrix A_ρ as before. The PDL operation is simply a constraint on the ordinary least squares regression, and does not imply a transformation of the error term (as does, for example, GLS). As a result, it does not affect our calculation of the residuals.

In calculating the variance-covariance matrix of the coefficients, we must apply an "unscrambling" process. Recall that the PDL constraining transformation took the form

$$Z = X^{\text{INC}}S \tag{101}$$

with the rank of Z less than the rank of X^{INC}. Combining those variables X^{UN} that are not included in the PDL operation, we wrote the constrained (scrambled) data matrix $\tilde{X}$ as

$$\tilde{X} = [Z \;\vdots\; X^{\text{UN}}] \tag{102}$$

Then the constrained variance-covariance matrix will be

$$C = \sigma^2(\tilde{X}'\tilde{X})^{-1} \tag{103}$$

As discussed in section IV, the unconstrained variance-covariance matrix C can be found from the transformation $C = D\tilde{C}D'$. It is important to keep in mind here that the matrices X^{INC} and X^{UN} have been transformed by the GLS transformation matrix A_ρ and contain the instrumental variable substitutions.

National Bureau of Economic Research
Massachusetts Institute of Technology

REFERENCES

[1] Almon, S., "The Distributed Lag between Capital Appropriations and Expenditures," *Econometrica*, January 1965.

[2] Box, M. S., D. Davies, and W. H. Swann, *Nonlinear Optimization Techniques*, I.C.I. Monograph No. 3, 1967.

[3] Cochrane, D., and G. H. Orcutt, "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms," *Journal of the American Statistical Association*, March 1949.

[4] Cooper, J. P., "Asymptotic Covariance Matrix of Procedures for Linear Regression in the Presence of First Order Serially Correlated Disturbances," forthcoming in *Econometrica* (University of Chicago, CMSBE Report 7047).

[5] Cooper, J. P., "Two Approaches to Polynomial Distributed Lags Estimation: An Expositional Note and Comment," forthcoming in *The American Statistician* (University of Chicago, CMSBE Report 7051).

[6] Fair, R., "The Estimation of Simultaneous Equation Models with Lagged Endogenous Variables and First Order Serially Correlated Errors", *Econometrica*, May 1970.

[7] Faddeev, D. R. and V. N. Faddeeva, *Computational Methods of Linear Algebra*, San Francisco, Freeman and Co., 1963.

[8] Fisher, F. M. "Simultaneous Equations Estimation: The State of the Art," working paper, M.I.T. Department of Economics, June 1970.

[9] Hall, R. E., "Calculating Least Squares," M.I.T. Econometrics Working Paper No. 2, July 1967.

[10] Hildreth, C. and J. Y. Lu, "Demand Relations with Autocorrelated Disturbances," Michigan State University, Agricultural Experimental Station, *Technical Bulletin* 276, November 1960.

[11] Johnston, J., *Econometric Methods*, New York, McGraw-Hill, 1963.

[12] Malinvaud, E., *Statistical Methods in Econometrics*, Chicago, 1966.

[13] Marquardt, D., "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *SIAM*, Vol. II, 1963, p. 431.

[14] Zellner, A., "An Efficient Method of Estimating Seemingly Unrelated Regressions," *Journal of the American Statistical Association*, June 1962.

[15] Zellner, A., and H. Theil, "Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations," *Econometrica*, January 1962.