This PDF is a selection from an out-of-print volume from the National Bureau of Economic Research

Volume Title: Annals of Economic and Social Measurement, Volume 3, number 4

Volume Author/Editor: Sanford V. Berg, editor

Volume Publisher: NBER

Volume URL: http://www.nber.org/books/aesm74-4

Publication Date: October 1974

Chapter Title: Estimation of Systems of Simultaneous Equations and Computational Specifications of GREMLIN

Chapter Author: David A. Belsley

Chapter URL: http://www.nber.org/chapters/c10203

Chapter pages in book: (p. 551 - 614)

Annals of Economic and Social Measurement, 3/4. 1974

ESTIMATION OF SYSTEMS OF SIMULTANEOUS EQUATIONS AND COMPUTATIONAL SPECIFICATIONS OF GREMLIN

BY DAVID A. BELSLEY*

CONTENTS

PART 0. INTRODUCTION

Several purposes are served by this paper. First, it describes the technical underpinnings of a comprehensive system of single- and multiequation econometric estimators--including the general *k*-class, three stage least squares (3SLS), instrumental variables (IV), limited and full information efficient instrumental variables (LIVE) and (FIVE), and as a byproduct of the latter, linear full-information maximum likelihood (FIML).[1] Design specifications for such estimators are, of course, not new; but the presentation given here is comprehensive and consistent, and introduces computational techniques of numerical analysis that will indeed be new and interesting to many econometricians.

* The author wishes to express gratitude to the following people for their aid, comments, discussion, and thoughts: Gregory Chow, John Dennis, Mark Eisner, Gene Golub, Jerry Hausman, Paul Holland, Dale Jorgenson, Edwin Kuh, Virginia Klema, Alexander Sarris. This research was supported under NSF Grant GJ-1154X3 to the NBER.

[1] The *k*-class and IV estimators are given in both linear and nonlinear forms. This paper only presents linear estimation for 3SLS and FIML. See Jorgenson and Laffont (elsewhere in this issue) on nonlinear 3SLS. The basis for the nonlinear FIML facility will be Gregory Chow's work (1972, 1973). Hausman (elsewhere in this issue) shows the relationship of iterated FIVE to linear FIML.

The estimation techniques described here are currently being implemented as a software system called GREMLIN (Generalized Research Environment and Modeling Language for the Integrated Network); this work is being done at the NBER Computer Research Center for Economics and Management Science. Hence, a second purpose of this paper is to give users of GREMLIN more detailed computational specifications than can be provided by the usual software documentation. In this regard it should be emphasized that the system is still being programmed and may differ in some details from the specifications given here; but this paper describes the basic design of the final product.

Third, this paper may introduce to econometricians several useful computational techniques of modern numerical analysis—in particular, the *QR decomposition* of a matrix (effected stably and efficiently by the Householder transformation) and the *singular value decomposition* of a matrix. These concepts and their properties, which are discussed in some detail here, will hardly be new to those familiar with the literature of numerical analysis; but they will be new to most econometricians, who until recently have not taken advantage of much relevant work done in that field. Both of these matrix decompositions produce efficient and stable computational schemes—efficient in the sense that the operation counts of many large econometric calculations can be reduced; and stable in the sense that the calculations are significantly less sensitive to the ill-conditioned (nearly singular) data matrices that are frequently encountered in econometric practice. In the work that follows, both the QR decomposition and the singular value decomposition are employed in widely differing situations, attesting to their power in practical computational contexts. It is also to be conjectured that the simplification of complex matrix expressions that frequently accompanies the application of these decompositions will show them to be powerful analytic tools.

0.1. Scope of this Paper

In Section 0.2, motivation will be offered for the development of the system described here. Then Part 1 treats the theory and calculations of the general k-class estimator. This discussion begins with preliminary lemmas on the QR decomposition and its application to ordinary least squares computations. This decomposition (effected by the Householder transformation) not only simplifies calculations but also yields expressions devoid of moment matrices and the need for matrix inverses—both major sources of computational problems to be avoided where possible.[2] The decomposition is then applied to the linear k-class estimator, which is in turn adapted for nonlinear (in the parameters) estimation.

Part 2 treats another important matrix decomposition, the singular value decomposition. This concept and its relation to pseudoinverses are developed and applied in the context of a general discussion of multicollinearity. Indeed, the singular value decomposition presents a means of calculation that remains stable even in the presence of perfect multicollinearity, and it also offers a promising

[2] It is advantageous to retain normal equations in moment-matrix form for the k-class estimator, although the QR decomposition still plays a central role. A linear form is possible, but for $k > 1$, it involves the need for storing matrices of complex numbers and is not readily adaptable for the iterative nonlinear estimation techniques of Section 1.5 and Appendix A.

means of detecting multicollinearity and determining if any estimates can be salvaged in spite of it.

Part 3 deals with the calculations of linear 3SLS;[3] here again, the *QR* decomposition simplifies the calculations. Part 4 examines estimation subject to linear constraints and presents a method employing the *QR* decomposition that may be applied directly to the moment matrices. This means of dealing with linear restrictions, which differs from the usual Lagrange technique or the method of substitution, is employed to allow efficient iteration for nonlinear estimation. Part 5 develops the computational procedures for several instrumental variables estimators. A method employing the *QR* decomposition is presented for the standard IV estimator, and its computational advantage is assessed. Further, several devices for constructing instruments through the use of principal components and/or preliminary regressions are developed (this draws heavily on the work of Kloek and Mennes (1960)). Finally, the resulting IV estimator is utilized to implement the Brundy–Jorgenson (1971) estimators LIVE and FIVE.[4]

GREMLIN will also include a general procedure for nonlinear full-information maximum likelihood estimates. The basis for the calculations to be employed are those developed by Gregory Chow (1972, 1973).

0.2. Background and Perspective

The last two decades have witnessed extraordinary growth not only in the theory of econometrics but also in its practice and its recognition as an essential part of virtually every phase of economics. This growth has not ceased, yet as in most rapidly growing fields, as many questions have been created as answered. The onslaught of econometric creativity has left pockets of "rubble" that must be tidied up and put into their proper place. A principal portion of this rubble in econometric theory is ignorance of the small-sample properties of the single- and multiequation estimators that have been accepted to varying degrees over the years, based primarily upon large-sample considerations or other assumed properties that have little to do with the reality from which economic data derive. Similarly, in the area of econometric practice, such examples of rubble are easily given; indeed, in considering the degree to which economic theory lacks *hard* empirical verification, one readily realizes that rubble is more the rule than the exception.

While there are many important reasons for our ignorance of small-sample properties and our incomplete empirical knowledge of economic systems, there is one ingredient, so far absent, that would help advance the profession in both areas—namely, a widely available estimation system that includes all important econometric estimators and is consistent, flexible, and efficient. The need for such a system motivates this work.

[3] A procedure for nonlinear 3SLS is given by Jorgenson and Laffont elsewhere in this issue.

[4] Jerry Hausman, elsewhere in this issue, shows the relation of iterated FIVE to linear full-information maximum likelihood.

Small-Sample Properties

The most direct source of ignorance of the small-sample properties of many econometric estimators is, of course, the intractable quality of the mathematics describing them— a difficulty that often disappears as sample sizes become indefinitely large. In order to gain the needed small-sample information, work has been in two general directions: exact, or nearly exact, results are sought in those few cases that admit such analysis; and Monte Carlo studies.

Recent theoretical results show that some exact or nearly exact answers may be possible. Light on exact small-sample properties has been shed in papers by Basmann (1961, 1963), Richardson (1968), Sawa (1969), Marino and Sawa (1971), and Kadane (1971); but these results deal with special cases and do not admit of obvious generalization to more complex and more realistic cases. Additional information has been obtained on nearly exact properties of small-sample estimators by using approximate results that take second- and even higher-order terms into account in "returning" from the asymptotic to the finite world. This promising research is exemplified by Anderson (1972), Anderson and Sawa (1970, 1973a, 1973b), and Nagar (1959).

In contrast to the theoretical work just mentioned, much effort also has been devoted to the small-sample properties and comparative efficiencies of the various estimators through Monte Carlo studies. This computation-intensive approach is well exemplified in studies by Summers (1965), Cragg (1966, 1967), Griliches and Rao (1969), Quandt (1962, 1965), Nagar (1960) and Wagner (1958); and the basic results are well summarized in Johnston (1972).

There is strong agreement in the general conclusion so far derived from both the theoretical and the Monte Carlo studies: namely, it all depends—just about anything can happen depending upon the circumstances.

Such an agnostic conclusion sounds, perhaps, more pessimistic than it is in fact; for in it there is at least the indication that in any given set of circumstances (at some specified point in the parameter and the data space), it may indeed be possible to derive meaningful small-sample conclusions for, and comparisons among, the various estimators. Since in the real world, not all circumstances are possible, and since informed limitations can be put on both the parameter and data space, theoretical analysis of important select regions of the parameter and data space may result in a less sterile conclusion than "anything can happen". This optimistic hope applies both to additional theoretical conclusions and to additional Monte Carlo results, for both tools seem most meaningfully applied when the model specification is narrowed and particularized.

One likes to think that the efficacy of theoretical studies has been limited by inadequate mathematical tools in combination with a shortage of genius, and that someday something will happen to change all that. Unfortunately, such a solution is outside our control. On the other hand, some of the main limiting factors for Monte Carlo studies can be controlled, namely, 1) the high cost of conducting studies of sufficiently varying parameter and sample conditions to gain any real overall picture;[5] 2) the lack of software estimation systems sufficiently compre-

[5] The initial version of GREMLIN may not fully exploit all computer capabilities required for truly efficient repetitive experiments; however, later versions will be made expressly with this in mind.

hensive to allow an *individual* investigator to make consistent comparisons of many different estimators, and 3) the unavailability of such software to the econometrics profession in general. The estimation facility planned here for the GREMLIN system will go a long way to relaxing these limitations.

Model Estimation

The role of large econometric models in furthering economic research and in aiding governmental and managerial policy decisions is perhaps best evidenced by the continuing use of many existing models, each highlighting some important area of theoretical or practical concern. Notable examples include the models of Klein–Goldberger (1955), the Brookings Institution (Duesenberry et al., 1965), the Wharton School, and MIT-FRB as well as the Michigan Model. The facility for building, estimating, and manipulating these models, however, is not widely available for econometric and managerial research. Whereas "regression packages" are universally available, systems which can execute all important full-system estimators upon large numbers of equations are available to but a few.

In order, then, to advance knowledge of the small-sample properties of econometric estimators and to facilitate applied econometric research in general, it seems useful to provide a comprehensive, consistent system of the important single- and multiequation estimators. Such a system should be implemented in a general research environment that includes facilties for data editing, model editing, and full-system simulation. The system should be generally available to the profession, should provide the power and flexibility needed to advance frontiers in all areas of applied econometric research, and should also provide the scope and efficiency needed for meaningful experimentation into the small-sample properties of the estimators.

Part 1. Double-k Class Calculation

1.0. Introduction

This part focuses on the calculations of the double-k class estimators of a single equation containing both endogenous and exogenous regressors. This general class of estimators includes such well-known estimators as ordinary least squares (OLS), two-stage least squares (2SLS) and limited information maximum likelihood (LIML).

First, in Section 1.0, the basic problem is defined, and notation that will be employed throughout the paper is developed. Section 1.1 presents the preliminary theoretical results that underlie the first-stage calculations given in Section 1.2. The basis of these preliminary results is the QR decomposition of a matrix, an operation that reduces the solution of the OLS problem to one whose calculations are devoid of moment matrices and inverses. The simplifications afforded by this decomposition will be frequently exploited in this paper. An outline of the final k-class computational procedure is given in Sections 1.3 and 1.4. Section 1.5 deals with estimation of equations that are nonlinear in the parameters, and Section 1.6 summarizes the computational steps.

Consider the multivariate equation

$$y = Y\gamma + X_1\beta + \varepsilon \tag{1.1}$$

where y is $T \times 1$, a vector of T observations on the normalized "dependent" variable;

Y is $T \times G$, a matrix of T observations on G endogenous variables included as regressors;

X_1 is $T \times K_1$, a matrix of T observations on K_1 included exogenous variables;

ε is $T \times 1$, a vector of stochastic disturbance terms;

γ is $G \times 1$, a vector of G unknown parameters to be estimated; and

β is $K_1 \times 1$, a vector of K_1 unknown parameters to be estimated.

In addition to these, define

X_2 to be $T \times K_2$, a matrix of T observations on K_2 additional exogenous variables (the excluded exogenous variables);

and define

$$X \equiv [X_1\ X_2], \text{ a } T \times K \text{ matrix with } K = K_1 + K_2.$$

The double-k class estimator of γ and β is a function of the data y, Y, X, and two parameters k_1 and k_2 that are determined in ways to be discussed later on. The basic form of the double-k class estimator (though not the form in which we shall calculate it) is

$$\begin{bmatrix} \hat{\gamma}_k \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} Y'Y - k_1(Y'Y)_{\perp X} & Y'X_1 \\ X_1'Y & X_1'X_1 \end{bmatrix}^{-1} \begin{bmatrix} Y'y - k_2(Y'y)_{\perp X} \\ X_1'y \end{bmatrix}. \tag{1.2}$$[6]

Were it not for the inclusion of the matrices $(Y'Y)_{\perp X}$ and $(Y'y)_{\perp X}$, (1.2) would simply be a $(G + K_1)$ square system of linear equations based on the moment matrices of y, Y, and X_1. $(Y'Y)_{\perp X}$, however, depicts the inner product of those components of Y with themselves insofar as they are orthogonal to the space spanned by the columns of X. Quite simply, $(Y'Y)_{\perp X}$ is the matrix of residual second moments resulting from regressing Y on X, and $(Y'y)_{\perp X}$ is analogously defined.[7] Thus, in calculating (1.2), the equivalent of a "first-stage" regression of Y on X is required to determine $(Y'Y)_{\perp X}$ and $(Y'y)_{\perp X}$.

[6] The notation $(Y'Y)_{\perp X}$ and $(Y'y)_{\perp X}$, which is explained immediately below, is Ruble's (1968), and will prove useful at a later stage.

[7] In projective terminology, any T vector Y can be decomposed into its orthogonal *projection* lying in the space spanned by the K columns of X, denoted $Y_{\parallel X}$ (Y parallel with the space spanned by X), and its orthogonal *projector*, denoted $Y_{\perp X}$, so that $Y = Y_{\parallel X} + Y_{\perp X}$. Since $Y'_{\parallel X}Y_{\perp X} = 0$, then $Y'Y = (Y'Y)_{\parallel X} + (Y'Y)_{\perp X}$, the standard decomposition of the second moment of Y into the "explained" and "unexplained (residual)" components.

It is a standard result of regression analysis that, when X is of full rank (i.e., $\rho(X) = K$),

$$(1.3) \qquad \begin{aligned} (Y'Y)_{\perp X} &= Y'Y - Y'X(X'X)^{-1}X'Y, \text{ and} \\ (Y'y)_{\perp X} &= Y'y - Y'X(X'X)^{-1}X'y. \end{aligned}$$

These calculations will not, however, be directly required. Indeed all of the submatrices in (1.2) may be obtained from a single QR decomposition of an appropriately expanded data matrix. This procedure has the following advantages:

1. It reduces significantly the sizes of the matrices for subsequent operations.
2. The Householder transformations that produce the QR decomposition are somewhat faster than ordinary regression calculations and are very stable.[8]
3. The calculation of $(Y'Y)_{\perp X}$ and $(Y'y)_{\perp X}$ can take place even when X is singular.[9]
4. The relevant matrices for determining the LIML value of k are given almost gratis.

We turn now, in Section 1.1, to some preliminary theoretical results that form the basis of the calculation procedure given in the Section 1.2.

1.1. Preliminary Results

The principal results for the method of calculation given here depend upon the QR decomposition of a matrix A, namely

Lemma 1.1a

For every $m \times n$ matrix A $(m \geq n)$ there exists an $m \times m$ orthogonal matrix $\tilde{Q}$ such that

$$\tilde{Q}A = \begin{bmatrix} R \\ 0 \end{bmatrix}$$

where R is $n \times n$ and upper triangular and 0 is $(m - n) \times n$.

Lemma 1.1a may be restated in another form that gives name to the QR decomposition. Let $\tilde{Q} \equiv \begin{bmatrix} Q' \\ S' \end{bmatrix}$ with Q' $n \times m$. Then since $\tilde{Q}'\tilde{Q}A = A = QR$, and conversely (since Q may always be augmented with orthogonal basis for the null space of A), we have

Lemma 1.1b

Every $m \times n$ matrix A $(m \geq n)$ can be decomposed as

$$A = QR$$

[8] On the Householder transformation see Golub (1969), Businger and Golub (1965), and Hanson and Lawson (1969).

[9] A true advantage during "first-stage" regressions where statistical tests of hypotheses are not being made, and hence no major problem arises from multicollinearity.

where Q is $m \times n$ (the same size as A) and $Q'Q = I_n$ and R is $n \times n$ and upper triangular.

Clearly the rank of R equals that of A, and hence R is a nonsingular triangular matrix if A has full rank. This makes inverting R particularly simple.

Such a decomposition may be effected either by a sequence of Householder transformations or by using classical or modified Gram–Schmidt orthogonalization. The modified Gram–Schmidt dominates classical Gram–Schmidt when A is ill-conditioned (nearly singular), as so frequently occurs in economic problems. The Householder transformations appear to be a speedy compromise, as shown in Businger and Golub (1965).

Simple regression is easily accomplished using the QR decomposition. Indeed

Lemma 1.2

In the linear equation $y = X\beta + \varepsilon$, the OLS estimator of β is $b = R^{-1}Q'y$, where $X \equiv QR$. Further $V(b) = \sigma^2 R^{-1}R'^{-1}$.

Proof

This follows from simply substituting for X in

$$\begin{aligned} b = (X'X)^{-1}X'y &= (R'Q'QR)^{-1}R'Q'y \\ &= (R'R)^{-1}R'Q'y \\ &= R^{-1}R'^{-1}R'Q'y = R^{-1}Q'y \end{aligned}$$

where the orthogonality of Q is used. Further, $V(b) = \sigma^2(X'X)^{-1} = \sigma^2 R^{-1}R'^{-1}$.

Q.E.D.

Due to the upper triangularity of R, an equation system of the form $Rb = Q'y$ is quickly solved by backsolving, and the need for a formal inversion routine is avoided. Further, moment matrices of the form $X'X$ are not required and the additional precision often necessitated by such accumulated sums of squares can be dispensed with.[10]

Somewhat more generally we have

Lemma 1.3

Let X and Y be two sets of variates of size K and M, respectively (T observations each). Then, from the QR decomposition of

$$Z \equiv [XY] = [Q_1 Q_2]\begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix},$$

[10] Unfortunately, this especially nice property of the QR decomposition in the context of OLS cannot always be exploited in more complicated estimators, particularly a method for linear equations that can also be used iteratively for solutions of nonlinear equations (Section 1.5).

(a) the moment matrix of residuals of Y regressed on X is

$$(Y'Y)_{\perp X} = R'_{22}R_{22}$$

and (b) the moment matrix of predicted values is

$$(Y'Y)_{\| X} = R'_{12}R_{12}.$$

Proof

$$Z'Z = \begin{bmatrix} X'X & X'Y \\ Y'X & Y'Y \end{bmatrix} = R'Q'QR = R'R$$

$$= \begin{bmatrix} R'_{11} & 0 \\ R'_{12} & R'_{22} \end{bmatrix}\begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} = \begin{bmatrix} R'_{11}R_{11} & R'_{11}R_{12} \\ R'_{12}R_{11} & R'_{12}R_{12} + R'_{22}R_{22} \end{bmatrix}$$

Now $(Y'Y)_{\perp X} = Y'Y - Y'X(X'X)^{-1}X'Y$; and by substitution of the appropriate moments from above

$$\begin{aligned} Y'Y - Y'X(X'X)^{-1}X'Y &= R'_{12}R_{12} + R'_{22}R_{22} - R'_{12}R_{11}(R'_{11}R_{11})^{-1}R'_{11}R_{12} \\ &= R'_{12}R_{12} + R'_{22}R_{22} - R'_{12}R_{12} \\ &= R'_{22}R_{22}. \end{aligned}$$

Thus (a) is shown.

Now (b) follows immediately from the fact that

$$Y'Y = (Y'Y)_{\perp X} + (Y'Y)_{\| X}.$$

Hence

$$\begin{aligned} (Y'Y)_{\| X} = Y'Y - (Y'Y)_{\perp X} &= R'_{12}R_{12} + R'_{22}R_{22} - R'_{22}R_{22} \\ &= R'_{12}R_{12}. \end{aligned}$$

Q.E.D.

Lemma 1.3a

In the event that $(Y'Y)_{\| X}$ is required but $(Y'Y)_{\perp X}$ is not, the QR decomposition of Lemma 1.3 need progress only through its first K steps (a fact we call Lemma 1.3a) since the sequence of Householder transformations works one row at a time, and additional changes do not affect the rows above the row being worked on. After K Steps, therefore, R will be of the form $\begin{bmatrix} K_{11} & R_{12} \\ 0 & S \end{bmatrix}$ where S is some rectangular (not upper triangular) matrix. In subsequent steps S will change but R_{12} will not, and hence R_{12} is available after the K-th step for calculating $(Y'Y)_{\| X} = R'_{12}R_{12}$.

$R'_{12}R_{12}$ can be calculated even if X is not of full rank. Its meaning will be correct, i.e., the sum of squares and cross products of the predicted values of

Y regressed on X, a unique value in spite of the fact that there may be an infinity of representations of these predicted values in terms of linear combinations of X.

The results of Lemma 1.3 are readily extended for the case of linear regression in

Lemma 1.4

In the least squares block regression of Y ($T \times M$) on X ($T \times K$), i.e. $Y = Xb + e$ where b is $K \times M$ and e is $T \times M$, we have

(a) $b = R_{11}^{-1}R_{12}$ (notation from Lemma 1.3)

(b) $e = Q_2R_{22}$

and in the case where $M = 1$,

(c) $V(b) = \sigma^2 R_{11}^{-1}R_{11}'^{-1}$

(d) $s^2 = \dfrac{1}{T-K}R_{22}^2$.

Proof

Following the notation developed in Lemma 1.3, from the QR decomposition of Z

$$X = Q_1R_{11}$$
$$Y = Q_1R_{12} + Q_2R_{22}.$$

Hence

$$b = (X'X)^{-1}X'Y = R_{11}^{-1}Q_1'Y \qquad (b \text{ is } K \times M)$$
$$= R_{11}^{-1}Q_1'(Q_1R_{12} + Q_2R_{22})$$
$$= R_{11}^{-1}R_{12} \text{ since } Q_1'Q_1 = I \text{ and } Q_1'Q_2 = 0.$$

Now

$$e = Y - Xb = (Q_1R_{12} + Q_2R_{22}) - (Q_1R_{11})(R_{11}^{-1}R_{12})$$
$$= Q_2R_{22}.$$

Rather generally $e'e = R_{22}'R_{22}$ (result (a) of Lemma 1.3), an $M \times M$ matrix, and for $M = 1$

$$e'e = R_{22}^2.$$

Hence

$$s^2 = \frac{1}{T-K}R_{22}^2.$$

(c) is already shown in Lemma 1.2. **Q.E.D.**

1.2. The k-Class (Double-k Class) Decomposition

The preceding results are now applied to the determination of the double-k class moment matrices in equation (1.2). Returning to the notation of Section 1.0, form the augmented matrix

$$Z = [X_1 X_2 Y y]$$

a $T \times (G + K + 1)$ matrix where X_1, Y, and y are from the linear equation (1.1), and X_2 contains T observations on K_2 additional predetermined variables. If $X = [X_1 X_2]$ contains all the predetermined variables in a full system of equations (of which (1.1) is a single equation to be estimated), we are dealing with a *proper* k-class estimator. If X contains X_1, and if X_2 is a subset of the remaining predetermined variables, we are dealing with a *truncated* k-class estimator. But, rather generally, there is no reason X_2 cannot contain any additional instrumental variables (asymptotically uncorrelated with ε, correlated with X_1).

Decomposing Z into a QR gives

$$(1.4) \qquad Z = [X_1 X_2 Y y] \equiv QR = [Q_1 Q_2 Q_3 Q_4] \begin{bmatrix} R_{11} & R_{12} & R_{13} & R_{14} \\ & R_{22} & R_{23} & R_{24} \\ & & R_{33} & R_{44} \\ 0 & & & R_{44} \end{bmatrix},$$

where the Q's are the same sizes as the corresponding partitions of Z, i.e.,

$$\begin{array}{lcccc} & T \times K_1 & T \times K_2 & T \times G & T \times 1 \\ Z = [& X_1 & X_2 & Y & y] \\ Q = [& Q_1 & Q_2 & Q_3 & Q_4], \end{array}$$

and the R's are sized as

$$R = \begin{array}{c} \begin{matrix} K_1 & K_2 & G & 1 \end{matrix} \\ \begin{bmatrix} R_{11} & R_{12} & R_{13} & R_{14} \\ & R_{22} & R_{23} & R_{24} \\ & & R_{33} & R_{34} \\ 0 & & & R_{44} \end{bmatrix} \end{array} \begin{matrix} K_1 \\ K_2 \\ G \\ 1 \end{matrix}$$

Each of the diagonal blocks is square and upper triangular.

Write the basic moments of Z in terms of R as follows:

$$(1.5) \qquad Z'Z = \begin{array}{c} \begin{matrix} K_1 & K_2 & G & 1 \end{matrix} \\ \begin{bmatrix} X_1'X_1 & X_1'X_2 & X_1'Y & X_1'y \\ X_2'X_1 & X_2'X_2 & X_2'Y & X_2'y \\ Y'X_1 & Y'X_2 & Y'Y & Y'y \\ y'X_1 & y'X_2 & y'Y & y'y \end{bmatrix} \end{array} \begin{matrix} K_1 \\ K_2 \\ G \\ 1 \end{matrix}$$

$$= R'Q'QR = R'R$$

$$= \begin{bmatrix} R'_{11} & & & 0 \\ R'_{12} & R'_{22} & & \\ R'_{13} & R'_{23} & R'_{33} & \\ R'_{14} & R'_{24} & R'_{34} & R'_{44} \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} & R_{13} & R_{14} \\ & R_{22} & R_{23} & R_{24} \\ & & R_{33} & R_{34} \\ 0 & & & R_{44} \end{bmatrix}$$

$$(1.6) \quad = \begin{bmatrix} R'_{11}R_{11} & R'_{11}R_{12} & R'_{11}R_{13} & R'_{11}R_{14} \\ R'_{12}R_{11} & R'_{12}R_{12} + R'_{22}R_{22} & R'_{12}R_{13} + R'_{22}R_{23} & R'_{12}R_{14} + R'_{22}R_{24} \\ R'_{13}R_{11} & R'_{13}R_{12} + R'_{23}R_{22} & R'_{13}R_{13} + R'_{23}R_{23} + R'_{33}R_{33} & R'_{13}R_{14} + R'_{23}R_{24} + R'_{33}R_{34} \\ R'_{14}R_{11} & R'_{14}R_{12} + R'_{24}R_{22} & R'_{14}R_{13} + R'_{24}R_{23} + R'_{34}R_{33} & R'_{14}R_{14} + R'_{24}R_{24} + R'_{34}R_{34} + R'_{44}R_{44} \end{bmatrix}$$

Now partition

$$Z \equiv [X_1 X_2 Y y]$$

as

$$[X_1 X_2 | Y y] \equiv [X|W],$$

where

$X = [X_1 X_2]$, is $T \times K$ and $W = [Y y]$ is $T \times (G + 1)$.

Lemma 1.3 is applied to $[XW]$ to obtain

$$(1.7) \qquad (W'W)_{\perp X} = ([Y y]'[Y y])_{\perp X} = \begin{bmatrix} (Y'Y)_{\perp X} & (Y'y)_{\perp X} \\ (y'Y)_{\perp X} & (y'y)_{\perp X} \end{bmatrix}$$

$$(1.8) \qquad = \begin{bmatrix} R'_{33} & 0 \\ R'_{34} & R'_{44} \end{bmatrix} \begin{bmatrix} R_{33} & R_{34} \\ 0 & R_{44} \end{bmatrix}$$

$$= \begin{bmatrix} R'_{33}R_{33} & R'_{33}R_{34} \\ R'_{34}R_{33} & R'_{34}R_{34} + R'_{44}R_{44} \end{bmatrix}$$

The relevant submatrices from (1.6) and (1.8) are paired with those in (1.5) and (1.7) to obtain

(1.9)

Raw Moment Matrix		R Decomposition	Size
$X_1'X_1$	=	$R_{11}'R_{11}$	$K_1 \times K_1$
$Y'X_1$	=	$R_{13}'R_{11}$	$M \times K_1$
$(Y'Y)_{\perp X}$	=	$R_{33}'R_{33}$	$M \times M$
$Y'Y$	=	$R_{13}'R_{13} + R_{23}'R_{23} + R_{33}'R_{33}$	$M \times M$
$(Y'y)_{\perp X}$	=	$R_{33}'R_{34}$	$M \times 1$
$Y'y$	=	$R_{13}'R_{14} + R_{23}'R_{24} + R_{33}'R_{34}$	$M \times 1$
$X_1'y$	=	$R_{11}'R_{14}$	$K_1 \times 1$

These can be substituted into (1.2) to give the double-k class estimator only in terms of the R_{ij} (all of the large Q matrices are unnecessary at this stage):

$$\begin{bmatrix} \hat{\gamma}_{k_1k_2} \\ \hat{\beta}_{k_1k_2} \end{bmatrix} = \begin{bmatrix} R_{13}'R_{13} + R_{23}'R_{23} + (1-k_1)R_{33}'R_{33} & R_{13}'R_{11} \\ R_{11}'R_{13} & R_{11}'R_{11} \end{bmatrix}^{-1} \begin{bmatrix} R_{13}R_{14} + R_{23}R_{24} + (1-k_2)R_{33}R_{34} \\ R_{11}R_{14} \end{bmatrix}. \tag{1.10}$$

The system of linear equations (1.10), which is summarized as

$$c \equiv M^{-1}d \quad \text{or} \quad Mc = d, \tag{1.11}$$

can be solved by a general linear equation-solving routine like MINFIT or by some similar routine that is more directly suited to dealing with a real symmetric system of equations. (MINFIT and other such procedures will be discussed more fully in Section 2.)

It is to be noted that both R and M require storage only of the upper triangle—R because it is upper triangular and M because it is symmetric.

1.3. The Values of k and Two Special Cases (2SLS and LIML)

Calculation of M in (1.11) requires knowledge of k_1 and k_2. In the k-class estimator, as distinguished from the double-k class, $k_1 = k_2$. Various well-known estimators result from special values of k. Indeed, in the cases of $k = k_1 = k_2$ the

following estimators result:

Value of k	Estimator
$k = 0$	Ordinary least squares[11]
$k = 1$	Two-stage least squares
$k = \mu$	Limited information maximum likelihood (the determination of μ will be discussed below)
$k = 1 + \dfrac{K - K_1 - G - 1}{T}$	Nagar's[12] Unbiased to order T^{-1}.

In GREMLIN, the value of K is to be specified by the user and he may specify different values for k_1 and k_2. It is envisioned, therefore, that the entire k-class package can be invoked by a single name; or any of the specific values given above can be invoked by a special name, such as 2SLS, LIML, OBK, which automatically causes the appropriate k to be used in the calculations.

Two special cases of k—2SLS ($k = 1$) and LIML ($k = \mu$)—deserve special attention because they have specific computational implications.

2SLS ($k = 1$)

This case deserves special attention for two reasons. First, as is clear from (1.10), with $k = 1$ the terms $(1 - k)R'_{33}R_{33}$ and $(1 - k)R'_{33}R_{34}$ do not appear and therefore need not be calculated. Second, as is also clear from (1.10), with $k = 1$, the only submatrices of R that are needed are R_{11}, R_{13}, R_{14}, R_{23}, and R_{24}—all from only the first two block rows. Applying Lemma 1.3a, therefore, it is required that the QR decomposition of Z proceed for only $K = K_1 + K_2$ steps to obtain the needed submatrices. In general $K + G$ steps will be required.

Both points can be exploited to make computation of this special case less burdensome.

LIML ($k = \mu$)

The LIML estimator is calculated as a k-class estimator with k equal to the minimum eigenvalue of the eigensystem

$$|H_1 - \mu H| = 0, \tag{1.12}$$

where

$$H_1 = Y'Y - Y'X_1(X'_1X_1)^{-1}X'_1Y = (Y'Y)_{\perp X_1}$$

and

$$H = Y'Y - Y'X(X'X)^{-1}X'Y = (Y'Y)_{\perp X}.$$

[11] This is an unnecessarily cumbersome means of calculating OLS, but it offers a good means of checking the program by comparison with the OLS estimator in TROLL (Eisner and Pindyck, 1973; National Bureau of Economic Research, 1974).

[12] Where X is $T \times K$, X_1 is $T \times K_1$, Y is $T \times G$.

From (1.7) and (1.8) it follows that

$$H = (Y'Y)_{\perp X} = R'_{33}R_{33}, \tag{1.13}$$

and from (1.5) and (1.6) it follows that

$$\begin{aligned} H_1 &= (Y'Y)_{\perp X_1} = Y'Y - Y'X_1(X'_1X_1)^{-1}X'_1Y \\ &= R'_{13}R_{13} + R'_{23}R_{23} + R'_{33}R_{33} - R'_{13}R_{11}(R'_{11}R_{11})^{-1}R'_{11}R_{13} \\ &= R'_{23}R_{23} + R'_{33}R_{33}. \end{aligned} \tag{1.14}$$

The determinantal equation (1.12) thus becomes

$$|R'_{23}R_{23} - (\mu - 1)R'_{33}R_{33}| = 0 \quad \text{or} \tag{1.15}$$

$$|(R'_{33}R_{33})^{-1}R'_{23}R_{23} - (\mu - 1)I| = 0.$$

The LIML μ, then, can be calculated as either of the following:

(1.16) (a) The minimum eigenvalue, $\sigma_{\min}$, of $(R'_{33}R_{33})^{-1}R'_{23}R_{23}$, in which case $\mu = \sigma_{\min} + 1$.

(b) The maximum eigenvalue, $\sigma_{\max}$, of $(R'_{23}R_{23})^{-1}R'_{33}R_{33}$, in which case $\mu = 1/\sigma_{\max} + 1$.

Depending upon the eigenvalue finder, method (a) would have an advantage, since R_{33} is upper triangular and its inverse is more readily found to produce $(R'_{33}R_{33})^{-1} = R_{33}^{-1}R_{33}^{-1'}$.

R_{33} is required for the LIML computations, and hence the QR decomposition of Z must proceed through the first $K + G$ operations. R_{44}, however, need not be directly computed—although, since it is 1×1, no substantial saving is accomplished here.[13]

Special facility for determining the minimal or maximal eigenvalue of (1.16) will therefore be required when the LIML option has been selected by the user, but no other special considerations arise in this case.

General k-Class

The user should be able to specify any value of k or k_1 and k_2. Equation (1.10) shows that R_{33} is required for all k-class estimators except 2SLS ($k = 1$). Hence it is necessary to effect the QR decomposition of Z through its first $G + K$ steps. It is never necessary to go through all $G + K + 1$ steps.

1.4. The k-Class Calculations

The preceding calculations result in the square, symmetric linear system (1.11), repeated here,

$$Mc = d \tag{1.17}$$

from which c, the k-class estimator, can be determined.

[13] The calculations for LIML given here have an advantage over those suggested by Dent and Golub (1973) in that they avoid the need to store the large Q matrix.

There should be at least two means of solving this linear system, and the user should have the option of picking the one he wants. The first is a routine like MINFIT (briefly explained below) that can calculate the singular values of M. Such a routine would be highly useful in analysis of problems due to multicollinearity, albeit at the cost of added computation time.

Second, there should be facility to solve (1.17) using a computationally efficient and speedy procedure such as the Cholesky decomposition, described below. The increased speed will be of great value in Monte Carlo studies and repetitive sampling experiments where the added information afforded by the singular values is not as important.

MINFIT

Both the nature of a matrix's singular values and the routine MINFIT will be described in Part 2. Here it need only be noted that MINFIT produces a diagonal matrix Σ of singular values and an orthogonal matrix V such that the real symmetric matrix M in (1.17) can be decomposed as

$$M = V\Sigma V'. \tag{1.18}$$

c is then calculated as $M^+d \equiv V\Sigma^+V'd$, where M^+ and Σ^+ are the pseudoinverses of M and Σ^2 respectively. (Pseudoinverses will also be discussed in Section 2.)

The residual vector

$$e = y - Y\hat{\gamma}_k - X_1\hat{\beta}_k \tag{1.19}$$

is best formed by using the $c = \begin{bmatrix} \hat{\gamma}_k \\ \hat{\beta}_k \end{bmatrix}$ calculated above directly with the raw data y, Y, and X_1 as in (1.19).

The estimator of σ^2, namely

$$s^2 = \frac{e'e}{T - K_1 - G}, \tag{1.20}$$

is to be calculated in exactly this way.

Finally, the estimated variance–covariance matrix of c is simply

$$s^2M^{-1} = s^2V\Sigma^{-1}V'. \tag{1.21}$$

Because most applications require only the diagonal elements of (1.21) to be produced, it seems reasonable to calculate only these values in the absence of additional optional specification by the user. If $V = (v_{ij})$ and $\Sigma = \text{diag}(\sigma_1 \ldots$

σ_{G+K}), $i, j = 1 \ldots G + K$, the k-th diagonal element of M^{-1} is simply

$$m^{kk} = \sum_{j=1}^{G+K} v_{kj}^2 \cdot \frac{1}{\sigma_j}. \tag{1.22}$$

Cholesky Decomposition

It is always possible to decompose a real, symmetric, positive-definite matrix,[14] such as M, into

$$M = D'D \tag{1.23}$$

where D is upper triangular. With this decomposition, (1.17) is solved as two backsolves

$$D'f = d \quad \text{and} \quad Dc = f,$$

stable calculations that avoid matrix inversion.

The calculations for s^2, e are as in (1.19) and (1.20), but M^{-1} must now be calculated as

$$M^{-1} = D^{-1}D'^{-1}, \tag{1.24}$$

which requires a routine for inverting an upper triangular matrix.

Note the relation between the Cholesky and QR decompositions relative to a positive-definite matrix of the form $X'X$. There is an infinity of upper triangular Cholesky matrices D such that $X'X = D'D$; but only one of these, namely $D = R$, is also associated with an orthogonal Q such that $X'X = R'R$ and $X = QR$.

1.5. Nonlinear Estimation

The procedure applied here to the estimation of an equation that is nonlinear in its parameters is a generalization of the preceding calculations, since it is akin to iteration on a linearized version of the given equation.[15]

Consider a general nonlinear equation

$$-f(Z, \delta) = \varepsilon \tag{1.25}$$

where f is a random vector of size T,

$$-f = \begin{bmatrix} -f^1(Z_1, \delta) \\ \vdots \\ -f^T(Z_T, \delta) \end{bmatrix}$$

and where $Z = [X_1, Y]$;

Z_t is the t-th row of Z;

X_1 is a $T \times K_1$ matrix of exogenous variables (identified as such);

Y is a $T \times G$ matrix of endogenous variables (identified as such);

[14] See, for example, Golub (1969), who also describes several computational procedures for effecting the decomposition.

[15] A more detailed explanation of the notation employed here and the calculations involved is given in Appendix A.

X_2 is a $T \times K_2$ matrix of additional exogenous variables (identified as such);

$\delta = (\delta_1 \ldots \delta_M)'$ is a vector of M unknown parameters to be estimated; and

ε is a $T \times 1$ vector of stochastic disturbances.

Linearizing (1.25) by expanding about δ_0 (and submerging the inessential argument Z) gives

$$\varepsilon = -f(\delta) = -f(\delta_0) - f_\delta(\delta_0)(\delta - \delta_0), \tag{1.26}$$

where

$$f_\delta \equiv \frac{\partial f}{\partial \delta} = \begin{bmatrix} f_1^1 \ldots f_M^1 \\ \vdots \quad \vdots \\ f_1^T \ldots f_M^T \end{bmatrix}, \tag{1.27}$$

the Jacobian of f with respect to δ, and called the matrix of *coterms*.

In general some of the columns of f_δ are functions of the endogenous Y's (as well, perhaps, as of the X's), and some are functions of the exogenous X's alone. Group the first set of coterms together in ϕ_δ, a $T \times M_1$ matrix of *endogenous coterms*; and group the second set together in χ_δ, a $T \times M_2$ matrix of *exogenous coterms*. Hence:

$$f_\delta \equiv [\phi_\delta \chi_\delta]. \tag{1.28}$$

The vector δ will be commensurately reordered and so partitioned as

$$\delta = \begin{bmatrix} \gamma \\ \beta \end{bmatrix}.$$

Equation (1.26) can be written in a form analogous to (1.1) as

$$f_\delta(\delta_0)\delta_0 - f(\delta_0) = f_\delta(\delta_0)\delta + \varepsilon. \tag{1.29}$$

In a manner described in detail in Appendix A, (1.29) leads to a Newton–Raphson iteration of the form

$$\hat{\delta}_{r+1} = \hat{\delta}_r - [\tilde{G} + f'_\delta f_\delta - k_1(f'_\delta f_\delta)_{\perp X_I}]^{-1} \cdot [f'_\delta f - k_2(f'_\delta f)_{\perp X_I}] \tag{1.30}$$

where X_I is a matrix of preliminary regressors and $\tilde{G}$ is a matrix formed of second-partials of f as

$$\tilde{G} = (g_{gk}) \equiv \Sigma \mathscr{F}^t \hat{f}^t \tag{1.31}$$

where $\mathscr{F}^t$ is the $G \times G$ matrix

$$(f^t_{gk}) \equiv \left[\frac{\partial^2 f^t}{\partial \beta_k \partial \beta_g}\right] \quad g, k = 1 \ldots G,$$

and

$$\hat{f}^t = W_I f \equiv X_I(X_I'X_I)^{-1}X_I' f.$$

The elements g_{gk} can also be computed as

$$g_{gk} = f'_{gk} W_I f = f'_{gk} \hat{f}, \tag{1.32}$$

but $\tilde{G}$ is probably best calculated as $\Sigma \mathcal{F}^t \hat{f}^t$.

Since $f_\delta = [\phi_\delta \chi_\delta]$, equation (1.30) becomes

$$\delta_{r+1} = \delta_r - \left[\tilde{G} + \begin{Bmatrix} \phi'_\delta \phi_\delta - k_1 (\phi'_\delta \phi_\delta)_{\perp X_I} & \phi'_\delta \chi_\delta \\ \chi'_\delta \phi_\delta & \chi'_\delta \chi_\delta \end{Bmatrix}\right]^{-1} \begin{bmatrix} \phi'_\delta f - k_2 (\phi'_\delta f)_{\perp X_I} \\ \chi'_\delta f \end{bmatrix}. \tag{1.33}$$

What matrix of preliminary regressors X_I should be used in (1.33)? In an analogy to the linearized equation (1.29), the included exogenous variates are χ_δ while the excluded variates are X_2. This would argue for the use of

$$X_I = [\chi_\delta X_2]. \tag{1.34}$$

The advantage of (1.34) is that the matrices needed in (1.32), except $\tilde{G}$, can be computed exactly, as in the linear case, through the QR decomposition of $Z = [\chi_\delta X_2 \phi_\delta f]$. The relevant blocks of this decomposition may be combined as in (1.10). The disadvantage is that the projection into the X_I space afforded by this decomposition in obtaining $(\phi'_\delta \phi_\delta)_{\perp X_I}$ and $(\phi'_\delta f)_{\perp X_I}$ must be recomputed at each iteration since the coterms χ_δ will change with each iteration.

An alternative technique would be to use

$$X_I = X \equiv [X_1 X_2]. \tag{1.35}$$

X is unchanging; and as has been demonstrated by Amemiya (1973), the resulting estimator retains consistency—although the comparative small-sample properties of different instruments remain an open question.

The use of (1.35) does not, however, allow full exploitation of the decomposition leading to (1.10), since X_1 and not χ_δ is employed. Rather $(\phi'_\delta \phi_\delta)_{\perp X}$ and $(\phi'_\delta f)_{\perp X}$ would be determined from a QR decomposition of $Z = [X_1 X_2 \phi_\delta f]$, with the first K steps computed only once at the first iteration and stored for repeated use in subsequent iterations. The remaining moments with χ_δ in (1.33) must be recomputed at each iteration.

1.6. Summary of Computational Steps

Linear Estimation

1. Form $Z = [X_1 X_2 Y y]$.
2. Determine k_1, k_2 or type of class.
3. Form QR decomposition of Z:
   (a) K steps only for 2SLS ($k = 1$);
   (b) $K + G$ steps otherwise.
4. Determine μ as in (1.16) if k is LIML.
5. Form (1.10) and solve for c.

6. Determine
   (a) e as from (1.19);
   (b) s^2 as from (1.20);
   (c) relevant elements of M^{-1} as from (1.21) or (1.24).
7. Output, minimally, c, s^2, M^{-1} (relevant elements), some housekeeping information on roles of variates.

Nonlinear Estimation

1. Form $f_\delta(\delta_0)$ and determine $[\phi_\delta \chi_\delta]$.
2. Form $[\chi_\delta X_2 \phi_\delta f] \equiv Z$.
3. Determine k_1, k_2 or type of class.
4. Form QR decomposition of Z:
   (a) $M_1 + K_2$ steps only for $k = 1$ (2SLS);
   (b) M steps otherwise.
5. Determine μ as in (1.16) if LIML.
6. Form $\tilde{G}$ and relevant matrices as in (1.10) for (1.31); solve for c.
7. Iterate to convergence.
8. Form final estimates and output as for linear case.

Part 2. Singular Value Decomposition, Pseudoinverses, and Multicollinearity

2.0. Introduction

This part focuses on a specific matrix decomposition, the singular value decomposition (SVD), that relates directly to the solution of the general least squares problem, including the case where X has less than full rank. The SVD is discussed in Section 2.1. The relation of the SVD to pseudoinverses is examined in Section 2.2. The two are brought together in Section 2.3 to provide a general solution to the least squares problem both when X is rank deficient and when X has full rank (the conventional OLS estimator). Section 2.4 explores the relevance of a procedure that can deal with the problem of multicollinearity even in the presence of rank deficiency. It is shown that the information given by the SVD may provide useful diagnostics for the presence and whereabouts of multicollinearity. Finally, a computational procedure that effects the SVD in the solution of the least squares problem is described. This procedure is called MINFIT.

2.1. The Singular Value Decomposition[16]

Lemma 2.1

Any $m \times n$ matrix A can be decomposed as

$$A = U\Sigma V' \tag{2.1}$$

where U and V are orthogonal matrices of sizes to be discussed below, and Σ is a diagonal matrix not necessarily square, whose nonzero diagonal elements are always positive and are called the *singular values* of A.

[16] See further Golub (1969, 1970) and Hanson and Lawson (1969).

See Lanczus (1961) or Osborne (1961) for a proof of Lemma 2.1.

U, Σ, and V can be sized in several different ways, each of which has appropriate applications. A is $m \times n$, $m \geq n$, and equation (2.1) can take the following forms:

$$\underset{m \times n}{A} = \underset{m \times n}{U} \quad \underset{n \times n}{\Sigma} \quad \underset{n \times n}{V'} \tag{2.1a}$$

$$\underset{m \times n}{A} = \underset{m \times m}{U} \quad \underset{m \times n}{\Sigma} \quad \underset{n \times n}{V'} . \tag{2.1b}$$

In addition, if A has rank $r \leq n$, then equation (2.1) can take the form

$$\underset{m \times n}{A} = \underset{m \times r}{U} \quad \underset{r \times r}{\Sigma} \quad \underset{r \times n}{V'}. \tag{2.1c}$$

In each case $U'U = V'V = I_n$. The nonzero elements of Σ are always positive and lie only on the first diagonal. In (2.1c) Σ is always square and has full rank with all its diagonal elements being strictly positive.

It is clear that

$$A'A = V\Sigma^2V' \tag{2.2}$$

and

$$AA' = U\Sigma^2U'.$$

Hence V and U are orthogonal matrices that diagonalize $A'A$ and AA', respectively. It follows that the diagonal elements of Σ are the positive square roots of the eigenvalues of $A'A$ and AA', and V and U are the matrices of eigenvectors of $A'A$ and AA', respectively.[17] U and V are necessarily of full rank. The rank of Σ, however, is equal to r, that of A; and Σ has r nonzero positive elements along its diagonal and zeros elsewhere.

2.2. Pseudoinverses

An immediate application of the SVD is in calculating the *pseudoinverse*[18] of the matrix A. The pseudoinverse of any $m \times n$ matrix A is the *unique* $n \times m$ matrix A^+ satisfying all of the following:

$$(AA^+)' = AA^+ \tag{2.3a}$$

$$(A^+A)' = A^+A \tag{2.3b}$$

$$AA^+A = A \tag{2.3c}$$

$$A^+AA^+ = A^+. \tag{2.3d}$$

For proof of the uniqueness of A^+, see Greville (1959) or Rao (1965, p. 25). It is readily verified that the pseudoinverse A^+ can be derived from the SVD of

[17] See, for example, Graybill (1969), Theorem 3.4.4.

[18] The term *pseudoinverse* is not universal. Rao (1965) refers to A^+ as the Moore Inverse, and Graybill (1969) and Theil (1971) call it the generalized inverse. This latter term, however, is more commonly reserved for any $n \times m$ matrix A^- such that for any vector Y for which $AX = Y$ is a consistent equation, $X = A^-Y$ is a solution (Rao, 1965, p. 24). In general there is an infinity of such A^-, of which A^+ is a unique special case.

$A = U\Sigma V'$ as

$$A^+ = V\Sigma^+ U' \tag{2.4}$$

where Σ^+ is the pseudoinverse of Σ. As again may be readily verified, Σ^+ is determined from Σ simply by replacing the nonzero diagonal elements of Σ by their reciprocals, leaving all other zeros, including any on the diagonal, unchanged.

2.3. SVD and Least Squares

This section begins with a review of the role of the pseudoinverse in the solution of the general least squares problem;[19] this establishes the relevance of the SVD to the least squares problem, since the SVD is a means of calculating the pseudoinverse. The analysis is then extended to the case where the data matrix of "independent" variates X is of less than full rank.

X Has Full Rank

In the linear model $y = Xb + e$, the normal equations that characterize $e'e$, the minimum sum of squared errors, are

$$X'Xb = X'y. \tag{2.5}$$

When the $T \times K$ matrix X has full rank, i.e. $\rho(X) = K \leq T$, the *unique* least squares solution is

$$b^* = (X'X)^{-1}X'y. \tag{2.6}$$

Application of the SVD to X gives

$$\underset{X}{\underset{T \times K}{}} = \underset{U}{\underset{T \times K}{}} \quad \underset{\Sigma}{\underset{K \times K}{}} \quad \underset{V'}{\underset{K \times K}{}} \tag{2.7}$$

where $U'U = V'V = I_K$ and Σ is diagonal and nonsingular. Hence (2.6) reduces to

$$\begin{aligned} b &= (V\Sigma U'U\Sigma V')^{-1}V\Sigma U'y \\ &= V\Sigma^{-1}U'y = X^+y \end{aligned} \tag{2.8}$$

where (2.4) is used and it is recognized that $\Sigma^+ = \Sigma^{-1}$ when Σ is nonsingular.

Equation (2.8) shows that knowledge of X^+ allows solution of the least squares problem without the costly and often unstable calculations of the moment matrix $X'X$ and its inverse $(X'X)^{-1}$. These calculations are required in the conventional formation of (2.6)—at least if X has full rank.

X Has Less Than Full Rank

The solution in (2.8) is general, for pseudoinverses exist even when the data matrix X has less than full rank.

[19] These basic calculations are not new, and Theil's new textbook (1971) makes them generally available. Another good exposition of the pseudoinverse in the least squares context is found in Peters and Wilkinson (1970).

Suppose now that $\rho(X) = r \leq K$. The normal equations (2.5) remain valid, but now they determine a $K - r$ dimensional space S of solutions for b, all giving the same minimized squared error length $e'e$. It will now be shown that the specific solution in (2.8) for the full-rank case $b^* = X^+y$ remains a solution in the rank-deficient case $b^* \in S$, and has the additional property that among all $b \in S$, b^* has minimum length.

X has $\rho(X) = r \leq K$. Application of the SVD to X in the form of (2.1c) gives

$$\begin{matrix} & & T \times r & r \times r & r \times K \\ (2.9) & X = & U & \Sigma & V' \end{matrix}$$

where $U'U = V'V = I_r$; and Σ is a square, diagonal, nonsingular matrix of size r. The normal equations (2.5) therefore become $V\Sigma U'U\Sigma V'b = V\Sigma U'y$ or,

$$V'b = \Sigma^{-1}U'y. \tag{2.10}$$

Premultiplying by V gives the equivalent normal equations

$$VV'b = V\Sigma^{-1}U'y \equiv X^+y. \tag{2.11}$$

Now two lemmas show:

Theorem 2.1

$b^* = X^+y$ is the *unique* vector of minimal length satisfying the normal equations (2.11) and, hence, minimizing the sum of squared residuals $e = y - Xb$, where $\rho(X) = r \leq K$.

Proof

Lemma 2.2

$b^* = X^+y$ satisfies (2.11).

Proof

$$\begin{aligned} VV'b^* &= VV'X^+y \\ &= VV'V\Sigma^+U'y = V\Sigma^+U'y = X^+y. \end{aligned}$$

Lemma 2.3

Let b^0 be any solution to $VV'b = X^+y$, and define d by $b^0 \equiv b^* + d$. Then $VV'd = 0$ and $d'b^* = 0$.

Proof

$VV'd = VV'(b^0 - b^*) = X^+y - X^+y = 0$. Hence, $b^* = VV'b^*$, and $d'b^* = d'VV'b^* = 0$.

Thus, to complete the proof to Theorem 2.1:

$$\begin{aligned} b^{0\prime}b^0 &= b^{*\prime}b^* + 2b^{*\prime}d + d'd \\ &= b^{*\prime}b^* + d'd, \end{aligned}$$

and hence $\|b^*\| \leq \|b^0\|$. The uniqueness of b^* follows from the uniqueness of the SVD and the pseudoinverse.

2.4. Multicollinearity and MINFIT

The preceding has shown that within the context of the linear regression model $y = \chi\beta + \varepsilon$, the solution of the least squares problem can always be made unique (if not economically interpretable), even when X has less than full rank, by extending the problem to that of finding the b^* of minimum length that also minimizes the sum of squared residuals. If X has full rank, this expanded problem produces the least squares estimator (2.6) that is familiar to econometricians. Thus, the use of pseudoinverses is a means of calculating least squares solutions (and predictions) even in the face of perfectly collinear data.

MINFIT is a computer routine that performs these calculations with computational stability. At the same time, MINFIT holds out the promise of being able to create diagnostics for the presence of multicollinearity. We will return to a description of MINFIT below, but before we do so, a word or two on collinear data seems in order.

Multicollinearity

As a general rule, estimation in the presence of perfectly collinear data is problematic for the econometrician. An exception is Marschak's (1953) now famous "prediction only" case, but this case is not of practical significance (except as noted below). In the prediction-only case, the collinear conditions upon which the estimation is based are expected to continue into the prediction period. Clearly such a case is, as a mechanical matter, handled effectively by simply dropping one of the collinear variates.[20]

However, one special instance of Marschak's case does occur as a practical matter: the calculation of multistage least squares estimators. In 2SLS, for example, the prediction of the $\hat{y}$'s is the sole object of the first-stage calculations; this is the special case where the observation period (upon which the estimates are based) and the prediction period are identical. It is of practical advantage, therefore, to have first-stage computational devices that proceed stably even when the first-stage regressors are linearly dependent (as they may happen to be—either through poor planning or because of their large numbers in models with many equations). Such a procedure will produce correct second-stage estimates even in those cases where standard regression packages (which require inverting $X'X$) would "blow up".

The real interest in a routine like MINFIT, however, occurs not when X is singular (of less than full rank), but when X is nearly singular (ill-conditioned). In this case, which is of extreme practical importance to the econometrician, standard programs, requiring the computation of $(X'X)^{-1}$, become computationally unstable. Clearly a routine that produces stable calculations when $X'X$ is

[20] This solution first requires that the offending variates be identified if calculations are to proceed in the conventional manner of (2.6). This requirement, and indeed the need altogether to drop offending variates, is avoided by a computational routine like MINFIT that works even in the presence of pure multicollinearity.

singular will have no computational trouble when $X'X$ is nearly singular. Equally clearly, however, such a routine does not solve the basic problem of near collinearity—the inability to separate structurally distinct but statistically confounded effects. It merely prevents this logical estimation problem from being compounded by an additional mechanical problem of unstable calculations.

There is, however, an obvious danger in using a method of calculation that always produces "unique" estimates, since perfect collinearity could make them economically meaningless. Integral to such a procedure, then, there should also be a means of diagnosing multicollinearity and alerting the user to its presence. The singular values computed by MINFIT as part of its basic calculations may well serve this purpose.

The Computations of MINFIT

MINFIT is a computational program[21] that solves the general ($\rho(X) = r \leq K$) least squares problem of Theorem 2.1. It determines the b^* of minimum length that minimizes $e'e$, namely $b^* = X^+y$. The basis of its computations is the determination of the pseudoinverse X^+ through the SVD of X, that is $X^+ = V\Sigma^+U'$ as in (2.4). The basic output of MINFIT includes $b^* = X^+y$, V, U and the singular values of X—the positive diagonal elements of Σ. It is these latter elements that help in diagnosing multicollinearity.

Conditioning of Matrices and Singular Values

The condition number[22] of an $n \times m$ matrix A, denoted $\kappa(A)$, is defined to be the ratio of its maximum to minimum nonzero singular values, $\sigma_{max}/\sigma_{min}$. In the SVD of $A = U\Sigma V'$, $\rho(A) = \rho(\Sigma)$. Hence, as A becomes "nearly singular" its minimum singular value approaches zero and $\kappa(A)$ becomes large. It is also clear that $\kappa(A) = \kappa(\lambda A)$ for any scalar λ, and hence the condition number (unlike the determinant) is a measure of near singularity or ill-conditioning that is invariant to the scale of the given matrix.

Since MINFIT, on its way to computing $b^* = X^+y$, also calculates the singular values of X, the user can be informed of $\kappa(X)$ and can thereby be alerted to the presence of multicollinearity.

SVD and the Decomposition of the Estimated Variance

The singular values and the SVD have great promise in diagnosing the source of multicollinearity and in assessing the extent of the troubles it may cause. As is well known, collinear data can cause some or all regression coefficients to be known only with very poor precision. However, not all the regression coefficients need be rendered useless by ill-conditioned data, and the extent to which this is true can be examined through a decomposition of the estimated variance into components associated with each singular value of X.

[21] MINFIT was developed by Gene Golub, Computer Sciences Department, Stanford University, and is published in Golub and Reinsch (1970). A version of MINFIT in use at the NBER Computer Research Center is published in Becker, Kaden, and Klema (1974).

[22] Also called the *spectral condition number*. See further Hanson and Lawson (1969).

Let $b^* = X^+ y$ be the OLS estimate of β in the standard linear model $y = X\beta + \varepsilon$, in which ε is appropriately distributed with zero mean, and $V(\varepsilon) = \sigma^2 I$; and X, however ill-conditioned, has full rank K. Then

$$b^* - \beta = X^+ \varepsilon \tag{2.12}$$

and, using (2.4),

$$\begin{aligned} V(b^*) &= \sigma^2 X^+ X'^+ \\ &= \sigma^2 V \Sigma^{-2} V'. \end{aligned} \tag{2.13}$$

Let b_k^* be the k-th element of b^*, and $V \equiv (v_{ij})$, $i, j = 1 \ldots K$; hence it follows from (2.13) that

$$\operatorname{var}(b_k^*) = \sigma^2 \sum_{j=1}^{K} \frac{v_{kj}^2}{\sigma_j^2} \tag{2.14}$$

The variance of b_k^* is thus seen to be a sum of components of the form v_{kj}^2/σ_j^2 each associated with one of the singular values σ_j. *Ceteris paribus*, the more nearly singular (the more ill-conditioned) the X, the smaller the certain σ_j; and hence, the larger the impact of those components on $\operatorname{var}(b_k^*)$. However, the ill effects of a very small σ_j can be mitigated, or even nullified, if the associated v_{kj}^2 in the numerator is correspondingly small. Indeed, letting X_i denote the i-th column of X, it is conjectured that if X_j is orthogonal to X_k and is nonorthogonal only to columns of X which are themselves orthogonal to X_k, then $v_{kj} = 0$. This result, which appears true in practice (an example is given below), requires formal proof. If true, however, it indicates that near singularity, resulting in very small σ_j for such X_j, would have little detrimental influence in determining the precision with which β_k can be estimated by least squares. Such a result is in accord with theory, for it is well known that in ordinary least squares, the addition of a new variate that is orthogonal to all preceding variates will not affect the preceding regression estimates. Indeed, then, adding two perfectly correlated variates, each of which is orthogonal to all preceding variates, should leave the preceding regression estimates, and the precision with which they are known, unchanged even though the augmented X matrix is singular. This result is seen in the following example.

An Example

Consider the case where $T = 6$, $K = 5$ and

$$X = \begin{bmatrix} -74 & 80 & 18 & -56 & -112 \\ 14 & -69 & 21 & 52 & 104 \\ 66 & -72 & -5 & 764 & 1528 \\ -12 & 66 & -30 & 4096 & 8192 \\ 3 & 8 & -7 & -13276 & -26552 \\ 4 & -12 & 4 & 8421 & 16842 \end{bmatrix}$$

This matrix, due to Bauer (1971), has the property that X_5 is exactly twice X_4, and both X_5 and X_4 are orthogonal to X_1, X_2 and X_3. The V matrix resulting from the SVD of $X = U\Sigma V'$ is

$$\left[\begin{array}{cccccc|cc}
0.547864\text{D} & 00 & -0.625347\text{D} & 00 & 0.555685\text{D} & 00 & & \\
-0.835930\text{D} & 00 & -0.383313\text{D} & 00 & 0.392800\text{D} & 00 & & \\
0.326342\text{D} & -01 & 0.679715\text{D} & 00 & 0.732750\text{D} & 00 & & \\
\hline
-0.642653\text{D} & -15 & -0.216297\text{D} & -15 & 0.913326\text{D} & -14 & & \\
0.321423\text{D} & -15 & 0.108174\text{D} & -15 & -0.456672\text{D} & -14 & & \\
\hline
 & & & & & & 0.148362\text{D} & -18 \quad -0.543183\text{D} \quad -14 \\
 & & & & & & 0.215618\text{D} & -19 \quad -0.470435\text{D} \quad -14 \\
 & & & & & & 0.158113\text{D} & -18 \quad -0.729449\text{D} \quad -14 \\
\hline
 & & & & & & -0.447214\text{D} & 00 \quad 0.894427\text{D} \quad 00 \\
 & & & & & & -0.894427\text{D} & 00 \quad -0.447214\text{D} \quad 00
\end{array}\right].$$

The resulting singular values, the diagonal elements of Σ, are

$$\begin{aligned}
\sigma_1 &= 0.170701\text{D} \quad 03 \\
\sigma_2 &= 0.605332\text{D} \quad 02 \\
\sigma_3 &= 0.760190\text{D} \quad 01 \\
\sigma_4 &= 0.363684\text{D} \quad 05 \\
\sigma_5 &= 0.131159\text{D} \quad -11.
\end{aligned}$$

A glance at V shows that the v_{ij} corresponding to the cross terms between group X_4 and X_5 on the one hand and group X_1, X_2 and X_3 on the other are all of the magnitude of 10^{-14} or smaller and are well within the effective zero of the computational precision.

Further, one singular value, σ_5 is much smaller than the other four, indicating (within the zero tolerances of the machine) the rank deficiency of X.[23] However, σ_5, small as it is, is several orders of magnitude larger than its corresponding v_{i5} for $i = 1 - 3$; and hence the contributions of the v_{i5}^2/σ_5^2 components to calculations of var(b_1^*), var(b_2^*), and var(b_3^*) in (2.14) will be small. That is, the presence of pure multicollinearity will not significantly upset the precision with which the coefficients of other variates can be estimated, provided these other variates are reasonably isolated from the offending collinear variables through near orthogonality.

[23] Indeed σ_5 relative to σ_4, the largest σ, is of the order of 10^{-16} and, according to bounds given by the numerical analysts, is within the zero of the machine. Professor Golub claims that any σ_k having the property that $\sigma_k/\sigma_{\max} \leq \sqrt{\varepsilon}$, where ε is the machine zero, is evidence of rank deficiency.

To demonstrate this point, calculate the relative components of $\text{var}(b_1^*)$ by means of (2.14):

$$\text{(2.15)} \qquad \begin{aligned} \text{var}(b_1^*) &= \sigma^2 \sum_{j=1}^{5} \frac{v_{1j}^2}{\sigma_j^2} \\ &= \sigma^2(0.0010 + 0.0107 + 0.5343 + 0.0 + 0.0017)10^{-2} \\ &= \sigma^2(0.5488 \times 10^{-2}). \end{aligned}$$

This shows that the component of $\text{var}(b_1^*)$ affected adversely by the collinearity, namely v_{15}^2/σ_5^2, is small (0.0017×10^{-2}) relative to the total (0.5488×10^{-2}). Indeed, this term has definition only through the finite arithmetic of the machine; in theory, it is an undetermined ratio of zeros. In practice there is reason to cast out this component in actual calculations of $\text{var}(b_1^*)$.

The preceding is in stark contrast to the calculation of $\text{var}(b_4^*)$ or $\text{var}(b_5^*)$, for these are the variances of coefficients that correspond to variables involved in the singularity of X. Indeed,

$$\text{(2.16)}^{24} \qquad \begin{aligned} \text{var}(b_5^*) &= \sigma^2 \sum_{j=1}^{5} \frac{v_{5j}^2}{\sigma_j^2} \\ &= \sigma^2(0.0 + 0.0 + 0.0 + 0.0000 + 1.1626 \times 10^{23}). \end{aligned}$$

This variance is obviously huge and completely dominated by the last term and its role in causing the singularity of X.

This example strongly suggests that there are situations in which near (or even perfect) collinearity need not prevent meaningful estimations of some regression coefficients—and these situations can be diagnosed and analyzed with data from the Σ and V matrices produced by the SVD of X. The situation in which such partial salvaging seems possible is when the offending multicollinear variates are adequately isolated from the others (perfect isolation being orthogonality). Clearly the problem of multicollinearity is a continuum: it increases as the strictness of the orthogonality is violated and as the X matrix becomes more nearly singular—as evidenced by one or more very small singular values.

There is no hope of salvaging estimates among the offending variates. In spite of much current research into the recovery of all estimates even with collinear data (research strangely reminiscent of the alchemists), one cannot retrieve that which was never there in the first place. The use of the SVD does, however, deserve investigation both as a diagnostic tool and as a means of retrieving all that is available when multicollinearity is present.

[24] The use of 0.0 and 0.0000 is intended to distinguish a number within the machine's zero (0.0) from a nonzero number with small exponent. The 0.0's in (2.16) are of the order of 10^{-30}, while the 0.0000 is of the order 10^{-10}.

PART 3. THREE-STAGE LEAST SQUARES

3.0. INTRODUCTION

This part presents the basic calculations for linear 3SLS estimates of a full system of G linear equations, or of a subsystem of such equations. The procedure given here uses the same efficient and stable computational schemes for the first-stage calculations as those developed in Part 1. The result is an efficient means of calculating linear 3SLS estimates, but unfortunately, this efficiency cannot be extended to nonlinear (in the parameters) estimation. The latter requires a different approach, as discussed by Jorgenson and Laffont elsewhere in this issue.

In the single-equation calculations for the k-class estimations of Part 1, the variates in the equation were ordered first into the included exogenous variates X_1, second into the excluded exogenous variates X_2, and finally into the included endogenous variates Y. This ordering was exploited in the subsequent QR decomposition, e.g., in (2.4). When there are several equations, however, the included exogenous variates of one equation are the excluded variates of another, and no such straightforward ordering is possible. A more general approach is, therefore, indicated if many operations are not to be duplicated. Here, then, a general set of calculations will be determined (effectively the first two stages), and a means will be determined for selecting appropriate subsets to build up the third-stage calculations.

Section 3.1 develops notation and determines the 3SLS estimator to be calculated. In Section 3.2 the basic 3SLS calculations are derived. The QR decomposition is once again exploited to produce the information from the "first two stages". An indexing scheme is determined to build up the final estimates from the moments of R.

Nothing has so far been said about estimation subject to linear constraints. This is the subject of Part 4, which treats the effect of linear restraints on 3SLS as well as on the K-class.

3.1. THE BASIC 3SLS MODEL

Consider the system of G equations

$$Y\Gamma + XB + U = 0 \tag{3.1}$$

where Y is a $T \times G$ matrix of G endogenous variables (specified as such);
X is a $T \times K$ matrix of K predetermined variables (specified as such);
Γ is a $G \times G$ matrix of unknown parameters to be estimated (some of which are specified initially to be zero);
B is a $K \times G$ matrix of unknown parameters to be estimated (some of which are specified to be zero); and
U is a $T \times G$ matrix of stochastic disturbance terms.[25]

For purposes of calculation it is better to rewrite (3.1) in a way that more directly deals with the individual equations. In particular, consider the g-th

[25] It is assumed that U is the result of a G-variate stationary stoachastic process with mean 0 and variance-covariance matrix $\sum_{GG}$.

equation in (3.1):

$$Y\tilde{\gamma}_g + X\tilde{\beta}_g + u_g = 0 \tag{3.2}$$

where $\tilde{\gamma}_g$ is the g-th column of Γ,
$\tilde{\beta}_g$ is the g-th column of B, and
u_g is the g-th column of U.

Since, in general, not all G of the Y's and not all K of the X's enter this equation, the variates are assumed to be ordered so that all zero coefficients in $\tilde{\gamma}_g$ and $\tilde{\beta}_g$ come last, i.e.,

$$\tilde{\gamma}_g = \begin{bmatrix} \bar{\gamma}_g \\ 0 \end{bmatrix} \quad \text{and} \quad \tilde{\beta}_g \equiv \begin{bmatrix} \beta_g \\ 0 \end{bmatrix} \tag{3.3}$$

where[26] $\bar{\gamma}_g$ is $G_g \times 1$, and β_g is $K_g \times 1$.

G_g, then, is the number of endogenous variables included in equation g (clearly $G - G_g$ are excluded), and K_g is the number of predetermined variables included in equation g.

Partitioning Y and X in accordance with the above gives

$$\begin{aligned} Y\tilde{\gamma}_g + X\tilde{\beta}_g + u_g &= [\bar{Y}_g Y_g^*]\begin{bmatrix} \bar{\gamma}_g \\ 0 \end{bmatrix} + [X_g X_g^*]\begin{bmatrix} \beta_g \\ 0 \end{bmatrix} + u_g \\ &= \bar{Y}_g\bar{\gamma}_g + X_g\beta_g + u_g = 0 \end{aligned} \tag{3.4}$$

where $\bar{Y}_g$ is the $T \times G$ matrix of included endogenous variables,
Y_g^* is the $T \times (G - G_g)$ matrix of excluded endogenous variables,
X_g is the $T \times K_g$ matrix of included exogenous variables, and
X_g^* is the $T \times (K - K_g)$ matrix of excluded exogenous variables.

Finally, the equation is normalized (since the variance of U is assumed to be known only up to a scalar) so that one of the coefficients (usually one of the $\bar{\gamma}_g$'s) equals minus unity. This coefficient and its variate are assumed to be placed first. Thus (3.4) becomes:

$$\begin{aligned} [y_g Y_g]\begin{bmatrix} -1 \\ \gamma_g \end{bmatrix} + X_g\beta_g + u_g = 0 \qquad &\text{or} \\ y_g = Y_g\gamma_g + X_g\beta_g + u_g, & \end{aligned} \tag{3.5}$$

where $\bar{Y}_g \equiv [y_g Y_g]$;
y_g is $T \times 1$, the normalized variate;
Y_g is $T \times (G_g - 1)$, the remaining included endogenous variates;
X_g is $T \times K_g$;
γ_g is $(G_g - 1) \times 1$; and
β_g is $K_g \times 1$.

[26] The reason for which $\bar{\gamma}_g$ is given a bar but β_g is not, will become apparent below.

Equation (3.5) is usually summarized as

(3.6) $$y_g = Z_g\delta_g + u_g$$

where

$$Z_g \equiv [Y_g X_g] \qquad T \times (G_g + K_g - 1),$$

and

$$\delta_g \equiv \begin{bmatrix} \gamma_g \\ \beta_g \end{bmatrix} \qquad (G_g + K_g - 1) \times 1.$$

In this notation (which includes all zero restrictions on the elements of Γ and B), the full system of equations (3.1) can be summarized as

(3.7) $$y = Z\delta + u$$

where

$$y \equiv \begin{bmatrix} y_1 \\ \vdots \\ y_g \\ \vdots \\ y_G \end{bmatrix} \qquad GT \times 1$$

$$Z \equiv \begin{bmatrix} Z_1 & & & 0 \\ & \ddots & & \\ & & Z_g & \\ & & & \ddots \\ 0 & & & Z_G \end{bmatrix} \qquad GT \times \Sigma_g(K_g + G_g - 1)$$

$$\delta \equiv \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_g \\ \vdots \\ \delta_G \end{bmatrix} \qquad \Sigma_g(K_g + G_g - 1) \times 1$$

$$u \equiv \begin{bmatrix} u_1 \\ \vdots \\ u_g \\ \vdots \\ u_G \end{bmatrix} \qquad TG \times 1.$$

The 3SLS estimator of δ in (3.7)—which can be derived as a generalized least squares estimator—takes the form[27]

$$\delta_{3SLS} = \begin{bmatrix} s^{11}(Z_1'Z_1)_{\|X} & \dots & s^{1G}(Z_1'Z_G)_{\|X} \\ \vdots & & \vdots \\ s^{G1}(Z_G'Z_1)_{\|X} & \dots & s^{GG}(Z_G'Z_G)_{\|X} \end{bmatrix}^{-1} \begin{bmatrix} \sum_g s^{1g}(Z_1'y_g)_{\|X} \\ \vdots \\ \sum_g s^{Gg}(Z_G'y_g)_{\|X} \end{bmatrix} \equiv N^{-1}d \tag{3.8}$$

where

$$(Z_h'Z_g)_{\|X} = \begin{bmatrix} (Y_h'Y_g)_{\|X} & Y_h'X_g \\ X_h'Y_g & X_h'X_g \end{bmatrix} \quad h, g = 1 \dots G \tag{3.9}$$

is the inner product of the columns of Z_h and Z_g insofar as they lie in the space spanned by the columns of X. $(Z_h'y_g)_{\|X}$ is analogously defined.[28]

When X has full rank, it is well known that

$$(Z_h'Z_g)_{\|X} = Z_h'X(X'X)^{-1}X'Z_g \quad \text{and} \quad (Z_h'y_g)_{\|X} = Z_h'X(X'X)^{-1}X'y_g. \tag{3.10}$$

The s^{ij} in (3.8) are the elements of S^{-1} where S is the estimator of the variance-covariance matrix Σ, based on 2SLS. The calculations for S will be discussed more fully later.

It is the elements of (3.8), then, that must be calculated to determine the δ_{3SLS}. These calculations are discussed in the next section.

3.2. The Basic 3SLS Calculations

All blocks in (3.10) can be determined by a single QR decomposition of the matrix $Z = [XY]$. Notice that $X = \cup_g X_g$ and $Y = \cup_g [y_g Y_g]$, where the symbol $\cup$ indicates set union.[29] We would then have

$$Z = [XY] = QR = [Q_1 Q_2] \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} \tag{3.11}$$

where

$$Q'Q = I_{K+G},$$

and the relevant matrix sizes are

$$\begin{array}{cc} K & G \\ T[X & Y] \\ T[Q_1 & Q_2] \end{array} \qquad \begin{array}{cc} K & G \\ K[R_{11} & R_{12}] \\ G[0 & R_{22}] \end{array}.$$

R_{11} and R_{22} are upper triangular.

[27] This result is available in any standard econometrics text, e.g., Johnston (1972, p. 397).

[28] See footnote 7 above.

[29] In practice it may be useful to have the machine determine Y and X from specifications for individual equations rather than have the user additionally specify them.

Application of Lemma 1.3 gives

$$(X'Y) = R'_{11}R_{12}, \tag{3.12}$$

$$(Y'Y)_{\|X} = R'_{12}R_{12},$$

and

$$(X'X) = R'_{11}R_{11}.$$

Hence a basic matrix of size $(K + G)^2$ can be constructed

$$M \equiv \begin{bmatrix} X'X & X'Y \\ Y'X & (Y'Y)_{\|X} \end{bmatrix} \tag{3.13}$$

$$= \begin{bmatrix} R'_{11}R_{11} & R'_{11}R_{12} \\ R'_{12}R_{11} & R'_{12}R_{12} \end{bmatrix} \equiv R'_1R_1$$

where $R_1 = [R_{11}R_{12}]$, the first block-row of R. Since M is based only on the first K rows of R, Lemma 1.3a can be applied to show that only the first K steps of the QR decomposition of Z are required—thereby determining R_1. Both this fact and the fact that M is symmetric and hence requires only its upper triangle to be stored, should be exploited.

Forming the $(Z'Z)_{\|X}$

Consider $(Z'_kZ_g)_{\|X}$. All elements of this general block of (3.8) are also elements of M, and hence can be derived from M. To do this will require some straightforward indexing.

Assume that each Y and each X are numbered:

$$\begin{matrix} 1 & \dots K & K+1 \dots K+G \end{matrix}$$

$$Z = [X_1 \dots X_K \quad Y_1 \quad \dots Y_G] \tag{3.14}$$

(X_K = k-th column of X, Y_g = g-th column of Y).

These numbers will be used to identify those variates included in a specific equation. In equation g, for example, the included variates can be summarized as:

$$\tilde{Z}_g = [X_g \quad Y_g \quad y_g], \qquad T \times (G_g + K_g + 1). \tag{3.15}$$

The columns of $\tilde{Z}_g$ can be labeled by their names from Z:

$$\begin{matrix} r_1 & \dots r_{K_g} & r_{K_g+1} \dots r_{K_g+G_g} & r_{K_g+G_g+1} \end{matrix}$$

$$\tilde{Z}_g = [X_{g1} \dots X_{gK_g} \quad Y_{g1} \quad \dots Y_{gG_g} \quad y_g] \tag{3.16}$$

where

X_{gi} is the i-th column of X_g

Y_{gj} is the j-th column of Y_g

and the r_i above the columns of $\tilde{Z}_g$ are the corresponding index names in (3.14).

Hence each $\bar{Z}_g$ can be identified by its list of r's. Take $\bar{Z}_h$ and $\bar{Z}_g$,

$$\begin{aligned}\bar{Z}_h &= \{r_1 \dots r_{K_h} \quad r_{K_h+1} \dots r_{K_h+G_h} \quad r_{K_h+G_h+1}\} \\ \bar{Z}_g &= \{s_1 \dots s_{K_g} \quad s_{K_g+1} \dots s_{K_g+G_g} \quad s_{K_g+G_g+1}\}.\end{aligned} \tag{3.17}$$

Now the (1, 1) element of $(Z_h'Z_g)_{\| X}$ is simply the (r_1, s_1) element of M, and, in general, the (m, n) element of $(Z_h'Z_g)_{\| X}$ is the (r_m, s_n) element of M. These blocks will be of size $(K_h + G_h) \times (K_g + G_g)$.

Similarly, in determining the $(Z_h'y_g)_{\| X}$ vectors, which will be $(K_h + G_h) \times 1$, the n-th component will be the $(r_n, s_{K_g+G_g+1})$ element of M.

Determining the s^{ij}

As each $(Z_h'Z_g)_{\| X}$ is formed, it should be stored in its appropriate block of (3.8); note of course that if (3.8) is written as $N\delta_{3SLS} = d$, N is symmetric and only its upper triangle need be stored. At this stage, d may consist only of

$$d = \begin{bmatrix} (Z_1'y_1)_{\| X} \\ \vdots \\ (Z_G'y_G)_{\| X} \end{bmatrix}. \tag{3.18}$$

The s^{ij} are determined from 2SLS estimates on each equation separately, and these can be obtained from the data blocks already computed as a solution to:

$$(Z_g'Z_g)_{\| X}\delta_{g2SLS} = (Z_g'y_g)_{\| X} \qquad g = 1, \dots, G \tag{3.19}$$

This is a square symmetric system to be solved through backsolving by some computationally speedy procedure such as the Cholesky decomposition. The additional output of the more costly MINFIT is not required in this use.

Having δ_{g2SLS} for $g = 1, \dots, G$, the 2SLS residuals can be formed as:

$$e_g = y_g - Z_g\delta_{g2SLS}, \qquad g = 1, \dots, G, \tag{3.20}$$

a $T \times 1$ vector to be stored in

$$E = [e_1 \dots e_G] \qquad T \times G. \tag{3.21}$$

S is then a $G \times G$ matrix determined as

$$S = \frac{1}{T}E'E \qquad G \times G, \tag{3.22}$$

a matrix whose inverse gives the s^{ij} required in (3.8),

$$S^{-1} = (s^{ij}). \tag{3.23}$$

Now it is possible to finish forming (3.8) by weighting the blocks of N with the appropriate s^{ij} and by forming the sums for each component of d. This latter operation will require additional submatrices of the form $(Z_h'y_g)_{\| X}$ to be picked from M in the manner described above.

Once the final N and d are finished, δ_{3SLS} is solved from the linear system

$$N\delta_{3SLS} = d. \tag{3.24}$$

This will usually be a large system, for N has dimensions

$$[\Sigma_g(G_g + K_g)] \times [\Sigma_g(G_g + K_g)].$$

As was true for the k-class estimators, the user should have the option of solving (3.24) either by a MINFIT-like routine that produces singular values, or by a faster routine like the Cholesky decomposition.

Part 4. Linear Restrictions in OLS, k-Class, and 3SLS

4.0. Introduction

On account of the nonlinear facility of the k-class estimation system described in Part 1, linear restrictions within a single equation can be built directly into the formulation of the model. For example, in the equation

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \varepsilon \tag{4.1}$$

with the linear restriction

$$\alpha_2 = 1 - \alpha_1, \tag{4.2}$$

correct constrained estimation will result from estimating the nonlinear equation

$$y = \alpha_0 + \alpha_1 \lambda_1 + (1 - \alpha_1) x_2 + \varepsilon. \tag{4.3}$$

This procedure has the advantage that it is easy for the user to include the restrictions; further, the procedure is not limited to linear constraints among the parameters. The disadvantages are that this procedure is computationally inefficient and is not directly applicable to constraints among coefficients in different equations of simultaneous systems. The first disadvantage is, perhaps, minor. The second makes it appropriate to consider a facility for estimating 3SLS and the like subject to linear constraints.

Section 4.1 briefly reviews and compares the two most commonly employed methods of including linear restriction in OLS—the method of Lagrangean constrained maximization and the method of substitution. A third method, more useful for the current purpose, is also explained; in this method the constraints are used directly to modify the moment matrix of the normal equations being solved. This has the following advantages:

1. The routines for k-class and 3SLS estimation developed in Parts 1 and 3 can be readily adapted to estimation subject to linear constraints.
2. The size of the final system of equations that must be solved is reduced rather than increased.

Section 4.2 extends the modified moment-matrix method to introduce linear restrictions in k-class estimation, and Section 4.3 further extends it to 3SLS.

4.1. Linear Restrictions in OLS

Consider the problem of estimating

$$Y = X\beta + \varepsilon \tag{4.4}$$

where

$$X \text{ is } T \times K$$

by OLS subject to the r independent linear constraints

$$A\beta = a \tag{4.5}$$

where

$$A \text{ is } r \times K$$

$$\rho(A) = r < K.$$

Method of Lagrange

An obvious way of treating this problem is to minimize $e'e = (Y - Xb)'(Y - Xb)$ subject to (4.5), by Lagrange's method:

$$\mathscr{L}(b, \lambda) = Y'Y - 2b'X'Y + b'X'Xb - \lambda'[a - Ab] \tag{4.6}$$

$$\frac{\partial \mathscr{L}}{\partial b} = -2X'Y + 2X'Xb + A'\lambda = 0 \tag{4.6a}$$

$$\frac{\partial \mathscr{L}}{\partial \lambda} = -a + Ab = 0. \tag{4.6b}$$

Equations (4.6a) and (4.6b) give

$$\begin{aligned} b &= (X'X)^{-1}X'Y + (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}[a - A(X'X)^{-1}X'Y] \\ &= \hat{b} + (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}[a - A\hat{b}] \end{aligned} \tag{4.7}$$

where $\hat{b}$ is the OLS estimator

$$\hat{b} = (X'X)^{-1}X'Y. \tag{4.8}$$

Substitution of (4.4) into (4.7), with reference to (4.5) gives

$$b = \beta + [1 - (X'X)^{-1}A'F^{-1}A](X'X)^{-1}X'\varepsilon \tag{4.9}$$

where

$$F \equiv A(X'X)^{-1}A',$$

and hence

$$\begin{aligned} V(b) &= E(b - \beta)(b - \beta)' \\ &= \sigma^2[(X'X)^{-1} - (X'X)^{-1}A'F^{-1}A(X'X)^{-1}]. \end{aligned} \tag{4.10}$$

Estimation via (4.7) clearly involves a regression of order K and much additional computation. The method of substitution reduces the order of the regression and thus seems to warrant consideration.

Method of Substitution

Beginning with

$$A\beta = a, \tag{4.11}$$

order the β's (and also the X's) so that (4.11) can be partitioned as

$$[A_1 A_2]\begin{bmatrix}\beta_1 \\ \beta_2\end{bmatrix} = A_1\beta_1 + A_2\beta_2 = a \tag{4.12}$$

where

$$A_1 \text{ is } r \times r, \quad \rho(A_1) = r,$$

and

$$A_2 \text{ is } r \times (K - r).$$

This gives

$$\beta_1 = A_1^{-1}[a - A_2\beta_2]. \tag{4.13}$$

Substitution of (4.13) into (4.4), commensurately partitioned, gives

$$\begin{aligned} Y &= X_1\beta_1 + X_2\beta_2 + \varepsilon \\ &= X_1A_1^{-1}a + [X_2 - X_1A_1^{-1}A_2]\beta_2 + \varepsilon. \end{aligned} \tag{4.14}$$

Equation (4.14) becomes

$$\begin{aligned} [Y - X_1A_1^{-1}a] &= [X_2 - X_1A_1^{-1}A_2]\beta_2 + \varepsilon \\ V &= W\beta_2 + \varepsilon \end{aligned} \tag{4.15}$$

where

$$\begin{aligned} \beta_1 &= A_1^{-1}[a - A_2]\beta_2 \\ V &= Y - X_1A_1^{-1}a \quad T \times 1 \\ W &= X_2 - X_1A_1^{-1}A_2 \quad T \times (K - r). \end{aligned}$$

Equation (4.15) is directly amenable to OLS, and computationally is a regression of order $(K - r)$ with a preliminary decomposition of A.

The decomposition of A can be done effectively by a QR decomposition of the augmented matrix $[A\ a]$ $r \times (K + 1)$. This results in

$$Q[A\ a] = [R_1R_2R_3] \tag{4.16}$$

where

R_1 is $r \times r$, upper triangular;

R_2 is $r \times (K - r)$; and

R_3 is $r \times 1$.

Premultiplying (4.11) by Q gives

(4.17) $$QA\beta = Qa \qquad \text{or}$$

$$[R_1 R_2]\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} = [R_3].$$

Hence

(4.18) $$R_1\beta_1 = R_3 - R_2\beta_2 \qquad \text{or}$$

$$\beta_1 = R_1^{-1}[R_3 - R_2\beta_2].$$

Since R is upper triangular, its inverse—or indirectly, its back-solution—is easily accomplished. Thus, the procedure for calculating the OLS estimates of (4.4) subject to the linear constraints (4.5) is

1. QR decomposition of $[A\ a] \to [R_1 R_2 R_3]$.
2. Form back-solution to

$$R_1[c_1 c_2] = [R_2 R_3]$$

so that $c_1 = R_1^{-1}R_2$ and $c_2 = R_1^{-1}R_3$.

3. Form $V = Y - X_1 c_2$

$$W = X_2 - X_1 c_1.$$

4. Apply OLS to V, W.

The variance-covariance matrix of β can now be derived from

(4.20) $$V(b_2) = \sigma^2(W'W)^{-1}.$$

Since b_1 is estimated from (4.18) as

(4.21) $$b_\cdot = R_1^{-1}R_3 - R_1^{-1}R_2 b_2$$

we have

(4.22) $$Eb_1 = c_2 - c_1, \qquad Eb_2 = c_2 - c_1\beta_2 = \beta_1,$$

and hence

(4.23) $$b_1 - \beta_1 = -c_1(b_2 - \beta_2).$$

Thus

(4.24) $$V(b_1) = E(b_1 - \beta_1)(b_1 - \beta_1)' = c_1 V(b_2)c_1' = \sigma^2 c_1(W'W)^{-1}c_1$$

$$\text{Cov}(b_1 b_2) = E(b_1 - \beta_1)(b_2 - \beta_2)' = -c_1 V(b_2) = -\sigma^2 c_1(W'W)^{-1}.$$

Combining these gives

(4.25) $$V(b) = \sigma^2\begin{pmatrix} c_1(W'W)^{-1}c_1' & -c_1(W'W)^{-1}\\ -(W'W)^{-1}c_1' & (W'W)^{-1}\end{pmatrix}$$

$$= \sigma^2 d(W'W)^{-1}d'$$

where

$$d = [-c_1\ I].$$

Whereas this method requires a QR decomposition of $[A\ a]$, a matrix of the size $r \times (K+1)$, the additional backsolvings are very fast, and the size of the ultimate OLS computations is reduced from K to $K - r$.

Modification in Moment-Matrix Form

The substitution method can be modified for application to the normal equations (4.8) based on the unconstrained estimation—rather than being used to reduce the system before calculation as in the procedure given in the previous section. The advantage of such a modification is that the k-class and 3SLS routines developed in Parts 1 and 3 can easily be adapted for estimation subject to linear constraints. At the same time, computational advantage of the method of substitution—namely, reducing the size of the system of equations to be solved—is retained.

Define

(4.26)
$$R_1^{-1}R_3 = f$$
$$-R_1^{-1}R_2 = \tilde{F}.$$

Then (4.21) becomes

(4.27)
$$b_1 = f + \tilde{F}b_2.$$

Define

(4.28)
$$F = \begin{bmatrix} \tilde{F} \\ I \end{bmatrix}$$

so that

$$Fb = f$$

and (4.14) becomes

(4.29)
$$Y - X_1 f = [X_1\tilde{F} + X_2]\beta_2 + \varepsilon = XF\beta_2 + \varepsilon.$$

OLS applied to (4.29) gives

(4.30)
$$\hat{b}_2 = (F'X'XF)^{-1}F'X'(Y - X_1 f).$$

Equation (4.30) can be calculated by either of the following methods:

1. OLS of $Y - X_1 f$ on XF; or
2. Formation of normal equations $X'Xb = X'Y$, adapted by
   (a) forming $F'(X'X)F$, and
   (b) forming $X'X_1 f$ (from appropriate columns of $X'X$) and then $F'(X'Y - X'X_1 f)$.

In method 2 constraints can be taken into account after an unconstrained moment matrix has been formed—a procedure that will be useful for k-class estimation

and for 3SLS. Specified in slightly greater detail, Method 2 is: Given $X'X$ and $X'Y$ (or its R equivalent),

1. form $Ab = a$ from F and f as described above,
2. form $F'X'XF \equiv M$,
3. form $X'Y - X'X_1 f = c$,
4. form $F'(X'Y - X'X_1 f) = F'c$,
5. solve $\hat{b}_2$ from $M\hat{b}_2 = F'c$,
6. calculate $\hat{b}_1 = f + \tilde{F}\hat{b}_2$ where $F = \begin{bmatrix} \tilde{F} \\ I \end{bmatrix}$.

The variance-covariance matrix of b can be calculated by noting

$$v(\hat{b}_1) = \sigma^2(F'X'XF)^{-1} = \sigma^2(W'W)^{-1} \tag{4.31}$$

for W as in (4.25), and hence

$$v(\hat{b}) = \sigma^2 F(F'X'XF)^{-1}F'. \tag{4.32}$$

4.2. Linear Restrictions in k-Class Estimation

As shown in Section 1.2, the k-class estimator results in the system of equations

$$\begin{bmatrix} \hat{\gamma}_{k_1k_2} \\ \hat{\beta}_{k_1k_2} \end{bmatrix} = \begin{bmatrix} R'_{13}R_{13} + R'_{23}R_{23} + (1-k_1)R'_{33}R_{33} & R'_{13}R_{11} \\ R'_{11}R_{13} & R'_{11}R_{11} \end{bmatrix}^{-1} \cdot \begin{bmatrix} R'_{13}R_{14} + R'_{23}R_{24} + (1-k_2)R'_{33}R_{34} \\ R'_{11}R_{14} \end{bmatrix} \tag{4.33}$$

which can be shortened as

$$Mc = d. \tag{4.34}$$

For $k = k_1 = k_2$ it is straightforward to verify that (4.34) is the set of normal equations for OLS applied to

$$H'Y = H'Z\delta + H'\varepsilon \tag{4.35}$$

where

$$H = [(1-k)^{1/2} I \; k^{1/2} Q],$$

and where Q results from the QK decomposition in (1.4). That is, we have

$$M = Z'HH'Z \quad \text{and} \quad d = Z'HH'Y.$$

Hence the k-class estimator $\hat{\delta}_k$ can be obtained simply by applying OLS to

$$\tilde{Y} = \tilde{Z}'\hat{\delta}_k + \tilde{\varepsilon} \tag{4.36}$$

where the tilde denotes the given matrix premultiplied by H'.

It is clear that estimation of $\hat{\delta}_k$ subject to linear constraints can proceed exactly as for the case of OLS in the previous section.

If $A\delta = a$, then form $F\delta = f$ and determine

$$\hat{\delta}_2 = [F'(\tilde{Z}'\tilde{Z})F]^{-1}F'\tilde{Z}'(\tilde{Y} - \tilde{Z}_1 f) \tag{4.37}$$

which can be calculated in moment form (as described above) as

$$(4.38)\quad (F'\tilde{Z}'\tilde{Z}F)\hat{\delta}_2 = F'(\tilde{Z}'\tilde{y} - \tilde{Z}'\tilde{Z}_1 f) \quad \text{or} \quad (F'MF)\hat{\delta}_2 = F'(c - M_1 f)$$

where $M_1 \equiv \tilde{Z}'\tilde{Z}_1$, taken from the relevant columns of M. Clearly, as in (4.32), $\hat{\delta}_1 = f + \tilde{F}\hat{\delta}_2$ and

$$(4.39)\quad v(\hat{\delta}_k) = \sigma^2 F(F'MF)^{-1}F'.$$

4.3. Linear Restrictions in 3SLS

The 3SLS estimates come from a solution to the linear equations (3.24), repeated here,

$$(4.40)\quad N\hat{\delta}_{3SLS} = d.$$

Additional linear constraints

$$A\delta = a$$

can be taken into account exactly as for the k-class estimator. Form F and f as described above under the method of modification of the moment matrix and determine

$$(4.41)\quad (F'NF)\hat{\delta}_2 = F'(d - N_1 f)$$

where N_1 is the columns of N corresponding to δ_1. Then

$$(4.42)\quad \hat{\delta}_1 = f + \tilde{F}\hat{\delta}_2$$

and

$$(4.43)\quad v(\hat{\delta}_{3SLS}) = \sigma^2 F(F'NF)^{-1}F'.$$

Part 5. Instrumental Variables Computations

5.0. Introduction

The instrumental variables (IV) estimator is among the most general consistent estimators of linear equations since it subsumes 2SLS, LIML, and 3SLS as special cases. The usefulness of IV estimation has been further enhanced by recent work of Brundy and Jorgenson and of Hausman. Brundy and Jorgenson (1971, 1973) introduced two-stage IV-type estimators called LIVE (Limited Information Instrumental Variables Efficient) and FIVE (Full Information Instrumental Variables Efficient). LIVE and FIVE have, respectively, the same Cramer–Rao best asymptotic efficiency as 2SLS and LIML, on the one hand, and as 3SLS and FIML, on the other. This asymptotic efficiency is gained without requiring a set of preliminary regressions on *all* exogenous variables in the systems of equations—a requirement in 2SLS and 3SLS that often cannot be met for large systems with few observations. Hausman (1973) showed that the FIVE estimator[30] when iterated, converges to the FIML estimate (if it converges at all). Thus a single well-integrated IV package can afford the user a wide choice of single- and multi-

[30] See further Hausman's paper in this issue.

equation estimators that possess both consistency, a basic property of all IV estimators, and asymptotic efficiency, a property only of LIVE and FIVE estimators (which include 2SLS and 3SLS).[31]

In Section 5.1 the basic IV estimator is determined. In Section 5.2 methods for constructing and computing the more interesting and widely employed instruments are discussed. Section 5.3 presents a means of calculating IV estimators, and a computationally efficient method employing the QR decomposition is proposed. In Section 5.4 the LIVE and FIVE two-stage estimators are dealt with.

5.1. The Basic IV Estimator

Consider with the linear equation

$$y = Y\gamma + X_1\beta + \varepsilon \equiv Z\delta + \varepsilon \tag{5.1}$$

where

$$y \text{ is } T \times 1 \qquad Z = [X_1 Y] \text{ is } T \times (K_1 + G)$$

$$Y \text{ is } T \times G \qquad \delta = \begin{bmatrix} \beta \\ \gamma \end{bmatrix} \text{ is } (K_1 + G) \times 1$$

$$X_1 \text{ is } T \times K_1$$

$$\varepsilon \text{ is } T \times 1.$$

A set of $G + K_1$ linearly independent instruments, W, is picked where W is $T \times (K_1 + G)$, with $\rho(W) = K_1 + G$.

In general, the instruments should be correlated with the variates X_1, but uncorrelated (at least asymptotically) with ε. Interest centers on picking and computing these instruments, a problem to be dealt with at length in the next section. Once the instruments have been picked, form

$$W'y = W'Z\delta + W'\varepsilon, \tag{5.2}$$

which implies the IV estimator

$$\hat{\delta}_{IV} = (W'Z)^{-1}W'y \qquad \text{or} \qquad (W'Z)\hat{\delta}_{IV} = W'y, \tag{5.3}$$

a square, nonsymmetric system of equations that can be solved directly through the use of a general routine like MINFIT (Section 2.4). In Section 5.3, however, these basic normal equations for $\hat{\delta}_{IV}$ will be transformed by a QR decomposition to produce a system of equations capable of more efficient solution—even counting the cost of the QR decomposition. The variance-covariance matrix of $\hat{\delta}_{IV}$ is readily derived (Johnston, 1972, p. 283):

$$V(\hat{\delta}_{IV}) = \sigma^2(W'Z)^{-1}(W'W)(Z'W)^{-1}. \tag{5.4}$$

[31] LIVE is a bit of a misnomer, for it is not "limited information" in the sense of LIML or 2SLS where specification need be made only for the single equation being estimated. LIVE is really a "full information" estimator that ignores cross-equation corrections but essentially requires the full set of equations to be specified.

5.2. Picking the Instruments

If an IV routine is to be truly useful in an interactive system like GREMLIN, it should have a capability for nearly automatic generation of widely used classes of instruments. This section specifies these instruments and their computation.

The task is to fill the $G + K_1$ columns of W with variates that are (i) correlated with X_1 but (ii) asymptotically uncorrelated with ε. Since the columns of X_1 fit these requirements ideally, it is assumed that X_1 is always used as K_1 of the instruments. Hence it remains only to pick the additional G instruments corresponding to the G-included endogenous variates Y. W is therefore of the form

$$W = [X_1 F] \tag{5.5}$$

where F is $T \times G$, a set of G instruments to be determined.

As a practical matter, the user has at his immediate disposal a set of variates $\mathcal{F}$ that satisfies (i) and (ii). $\mathcal{F}$ usually includes the following subset:

1. X_1, the predetermined variates included in the given equation.
2. X_2 (or some subset of X_2), the set of all other predetermined (cotemporaneously uncorrelated) variates in the system of equations. ($X \equiv [X_1 X_2]$.)
3. X_{-1}, additional lagged values of the X's.
4. D, dummy variables constructed by the user.

In addition to the basic elements of $\mathcal{F}$, a facility should be available by which the user can readily augment these variates by various principal components of the elements of $\mathcal{F}$ or of elements derived from those in $\mathcal{F}$. The use of principal components in this context has been formalized by Kloeck and Mennes (1960), whose work is incorporated here. Being linear combinations of the elements of $\mathcal{F}$, these principal components also satisfy conditions (i) and (ii) and hence are legitimate possibilities. Thus, routines will be required to generate the following:

5. P_1, the principal components (or first principal components) of any subset of $\mathcal{F}$.
6. P_2, the principal components (or first principal components) of the residuals of the block regression of any subset of $\mathcal{F}$ regressed on any other subset of $\mathcal{F}$.[32]

Denote by $\mathcal{H}$ the set $\mathcal{F}$ augmented as in (5) and (6). Two methods[33] of determining F can now be usefully distinguished:

Method I, Substitution: Determine F as any G columns (presumably linearly independent) picked from G elements of $\mathcal{H}$.

Method II, Regression: Determine F as $\hat{Y}$, the G-predicted values resulting from a regression of Y on any subset of $\mathcal{H}$ of order G or greater.

[32] P_1 allows for instruments corresponding to Kloeck and Mennes (1960) methods 1 and 4, while P_2 allows for their methods 2 and 3.

[33] Clearly Method II is but another means of augmenting the set $\mathcal{H}$ to include additional instruments. But it seems useful to separate this case so that its relation to multistage least squares techniques can be kept in mind.

Options for Method I, Direct Substitution

In general, the user should be able to choose F as any subset of G elements of $\mathscr{H}$. He should have options for the following special cases:

(a) F taken to be any subset of $\mathscr{F}$ of order G, not including those elements in X_1.

(b) F taken to be the G largest principal components of any subset of $\mathscr{F}$ of order G or greater.
   1) $F = G$ largest principal components of $\mathscr{F}$.
   2) $F = G$ largest principal components of $\mathscr{F}$ excluding X_1.

(c) F taken to be the G largest principal components of the residuals of any subset of $\mathscr{F}$ (exclusive of X_1) regressed on X_1; i.e., let P be the matrix whose columns are members of $\mathscr{F}$ not also included in X_1, and then form F as the G largest principal components of the residual matrix $P - X_1(X_1'X_1)^{-1}X_1'P$.
   1) $P = X_2$.
   2) $P = [X_2X_{-1}D]$, i.e., $\mathscr{F}$ exclusive of X_1.

(d) As in (b) except that the ordering is not by descending eigenvalues σ_k^2, but by descending values of $\sigma_k^2(1 - r_k^2)$ where r_k^2 is the multiple correlation coefficient of the k-th variate in $\mathscr{F}$ on X_1. This ordering can be applied to either 1) or 2) in (b).[34]

These options require that the IV routine have access to a principal components finder and an OLS package to find multiple correlation coefficients in (d).

Options for Method II, Preliminary Regression

In general, the user should be able to choose any subset of G or more elements of $\mathscr{F}$ to act as preliminary regressors in determining $\hat{Y}$ as F. Denote the matrix of such regressors by L.

(a) L = any subset of G or more elements of $\mathscr{F}$.

(b) L = the $G + n$ ($n \geq 0$) largest principal components of any subset of $\mathscr{F}$ of order $G + n$ or greater.
   1) $L = G + n$ largest principal components of $\mathscr{F}$.
   2) $L = G + n$ largest principal components of $\mathscr{F}$ excluding X_1.

(c) As in Method I(c) except that $G + n$ principal components can be taken.

(d) As in Method I(d) except that $G + n$ principal components can be taken.

Calculation of Instruments

Let B be a $T \times M$ matrix whose M columns are composed of the basic set of instruments from the set $\mathscr{F}$. These variables, supplied by the operator, can serve

[34] The numbering of methods here corresponds to numbering of methods in Kloeck and Mennes (1960) as follows:

Kloeck and Mennes	This Paper
1	I(b) 2)
2	I(c)
3	I(d)
4	I(b) 1)

as instruments by themselves, or they can be transformed into other instruments, as, for example, by taking various of their principal components. B can be defined for a whole system of equations, but for single-equation IV estimation it will be particularized for that equation. For a given equation, B will always contain X_1, the set of included exogenous variables. Hence write

$$B \equiv [X_1 B_1]. \tag{5.6}$$

As described earlier, B_1 can contain a subset of X_2 (the excluded predetermined variates), a subset of X_{-1} (lagged values of any of the predetermined variates), and D, a matrix of appropriate dummy variates.

The discussions of Methods I and II indicate the need for generating various types of principal components of B and its submatrices. In particular, the following computational routines are needed:

PC(k: *LIST*). This routine produces the k largest principal components of the variates given in LIST—all columns of B. The user specifies k subject to certain restrictions that should be automatically checked and flagged if violated. The restrictions are:

1) If Method I is used, $k = G$ and LIST must have $G = k$ or more elements.

2) If Method II is used, $k \geq G$ and LIST must have k or more elements.

A default option should be provided that assumes LIST indicates all of B if no list is given. Further, a symbol should be available which causes LIST to include only the elements of B_1 (B exclusive of X_1), such as *PC*(k:B_1).

This routine implements Methods I(b) and II(b).

*PC*1(k: *LIST*). This routine produces the k largest principal components of the residual matrix of the variables in LIST regressed on X_1. In this case no variables composing X_1 should be permitted in LIST, for this will guarantee perfect collinearity in the ultimate IV equations. A check for such consistency is desirable.

These calculations can be accomplished as follows. Let C be the matrix whose columns are in LIST. Applying Lemma 1.3, decompose $A = [X_1 C]$ by the QR routine to obtain

$$A = [Q_1 Q_2] \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}.$$

The matrix of residuals is $Q_2 R_{22} \equiv U$ (Lemma 1.4); the k largest principal components of this matrix are sought. If the principal components of $U \equiv Q_2 R_{22}$ are calculated by forming the eigenvectors of $U'U$, Q is orthogonal and $U'U$ is simply $R'_{22} R_{22}$. However, Q_2 must be preserved in this instance so that the principal components of U can be calculated. If V is the matrix of eigenvectors of $U'U$, then $P \equiv Q_2 R_{22} V$ is sought as the principal components of $Q_2 R_{22}$.

The same checks on the relation of k to G described above for the two Methods should be made. This routine implements Methods I(c) and II(c).

*PC*2(k: *LIST*). This routine, used in conjunction with *PC*(k: LIST), modifies the ordering of that routine and takes the k largest principal components according to the new ordering. In particular this routine does the following:

1. Forms the principal components of LIST (always exclusive of X_1)—call these by the matrix P—along with their corresponding eigenvalues μ_j^2.

2. Determines the multiple correlation of the elements of P regressed on X_1 (let the j-th such multiple correlation be denoted r_j^2).
3. Orders the principal components according to the ranking

$$\lambda_j = \mu_j^2(1 - r_j^2). \tag{5.7}$$

The μ_j^2 result from the determination of the principal components in Step 1 and can be obtained from *PC*(k: LIST) where k is made the same size as the order of LIST.

The r_j^2 are formed as follows. Decompose $A = [X_1 P]$ to obtain

$$QA = \begin{bmatrix} R_{11} & R_{12} \\ 0 & S \end{bmatrix}$$

stopping after K_1 steps since only R_{12} is required. Now $(P'P)_{\|X_1} = R'_{12}R_{12}$ by Lemma 1.3. Hence the diagonal elements[35] are $(P'_jP_j)_{\|X_1} = \hat{P}'_j\hat{P}_j \equiv \pi_j$. Further $P'_jP_j = \mu_j^2$ since these are principal components; hence

$$r_j^2 = \left[\frac{\hat{P}'_j\hat{P}_j}{P'_jP_j}\right] = \frac{\pi_j}{\mu_j^2} \tag{5.8}$$

or

$$\mu_j^2 r_j^2 = \hat{P}'_j\hat{P}_j = \pi_j. \tag{5.9}$$

Hence

$$\lambda_j \equiv \mu_j^2(1 - r_j^2) = \mu_j^2 - \mu_j^2 r_j^2 = \mu_j^2 - \pi_j. \tag{5.10}$$

PC2(k: LIST) now determines those principal components in P corresponding to the k largest values of the λ_j in (5.10).

Relation of IV to Multistage Least Squares

Methods I and II above offer many ways of choosing instruments to form W. Those given in Method II involve preliminary regressions to determine the instruments, and as is well known, these IV estimators bear a relation to multistage (truncated or augmented two-stage) least squares estimators.

The multistage 2SLS ($k = 1$) estimator and the corresponding IV estimator are identical when (and only when) the list of preliminary regressors includes a basis for X_1, the set of exogenous variates included in the particular equation being estimated (Brundy and Jorgenson, 1973). Only when this is true will the multistage least squares estimator be consistent—although the IV estimator is consistent regardless of whether the instruments or the preliminary regressors contain a basis for X_1.

The method of k-class estimation given in Part 1 guarantees that the included predetermined variates X_1 are utilized in the first-stage regressions along with the matrix X_2—which can be any of the other predetermined variates used here as instruments. Therefore, the k-class package should have access to the instrument-generating routines discussed here. When an X_2 is employed that does not contain

[35] Since only the diagonal elements of $R'_{12}R_{12}$ are needed, only these inner products need be computed from R_{12}.

all the "other predetermined" variates in the system of equations, the resulting multistage least squares estimator will lack the efficiency of 2SLS, but such estimators are popular and their easy access is desirable.

Although the truncated 2SLS estimator described above is inefficient, so also, rather generally, is IV estimation. But either of these estimators can be used separately or together to form a set of consistent—not necessarily efficient—estimates of the full system of equations; and these estimates can be employed in a multistage IV procedure such as LIVE or FIVE, to produce asymptotically efficient estimators. This will be discussed more fully in Section 5.4.

5.3. The IV Computational Procedure

In the notation of Section 5.1, the task is to calculate the $\hat{\delta}_{IV}$ solving

$$(W'Z)\hat{\delta}_{IV} = W'y \tag{5.11}$$

where

$$W = [X_1 F] \text{ is } T \times (G \times K_1)$$

$$Z = [X_1 Y] \text{ is } T \times (G + K_1)$$

and F has been determined as a $T \times G$ matrix of instruments (Section 5.2).

The variance-covariance matrix of $\hat{\delta}_{IV}$ is

$$V(\hat{\delta}_{IV}) = \sigma^2(W'Z)^{-1}(W'W)(Z'W)^{-1}. \tag{5.12}$$

(5.11) is a square, nonsymmetric system of equations that can be solved with MINFIT or a similar routine after the relevant moment matrices $W'Z$ and $W'y$ have been formed, and it may be useful to have facility for carrying out these direct calculations. However, an alternative procedure is given here that, in terms of operations counts, is faster and more efficient.

The Calculations

Form the QR decomposition of

$$A \equiv [X_1 F Y y] \tag{5.13}$$

to get

$$QA = Q[X_1 F Y y] = \begin{bmatrix} R_{11} & R_{12} & R_{13} & R_{14} \\ & R_{22} & R_{23} & R_{24} \\ 0 & & S & S \\ & & S & S \end{bmatrix} \tag{5.14}$$

where only $K_1 + G$ steps in the decompositions are taken, and the S's represent the remaining parts of A after the first $K_1 + G$ rows are formed using Householder transformations. The S elements are essentially discarded for subsequent calculations. Q is orthogonal and R_{11} and R_{22} are upper triangular.

Now,

$$W'Z = W'Q'QZ = (QW)'QZ$$

and

$$QW = Q[X_1 F] = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \\ & 0 \end{bmatrix} \equiv \begin{bmatrix} M \\ 0 \end{bmatrix} \tag{5.15}$$

$$QZ = Q[X_1 Y] = \begin{bmatrix} R_{11} & R_{13} \\ & R_{23} \\ & S \\ 0 & S \end{bmatrix} \equiv \begin{bmatrix} \overline{M} \\ F \end{bmatrix}$$

with

$$M = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$$

and

$$\overline{M} = \begin{bmatrix} R_{11} & R_{13} \\ 0 & R_{23} \end{bmatrix}$$

and

$$Qy = \begin{bmatrix} R_{14} \\ R_{24} \\ S \\ S \end{bmatrix} \equiv \begin{bmatrix} m \\ f \end{bmatrix}$$

$$m = \begin{bmatrix} R_{14} \\ R_{24} \end{bmatrix}.$$

Further,

$$W'W = W'Q'QW = [M'0'] \begin{bmatrix} M \\ 0 \end{bmatrix} = M'M \tag{5.16}$$

$$W'Z = W'Q'QZ = [M'0'] \begin{bmatrix} \overline{M}' \\ F \end{bmatrix} = M'\overline{M}$$

$$W'y = W'Q'Qy = [M'0'] \begin{bmatrix} m \\ f \end{bmatrix} = M'm.$$

By substitution of (5.15) and (5.16) into (5.11), $\hat{\delta}_{IV}$ becomes

$$\hat{\delta}_{IV} = (W'Z)^{-1} W'y = (M'\overline{M})^{-1} M'm = \overline{M}^{-1} m. \tag{5.17}$$

$V(\hat{\delta}_{IV})$ in (5.12) becomes

$$V(\hat{\delta}_{IV}) = \sigma^2(W'Z)^{-1}W'W(Z'W)^{-1} \tag{5.18}$$
$$= \sigma^2(M'\overline{M})^{-1}M'M(\overline{M}'M)^{-1}$$
$$= \sigma^2\overline{M}^{-1}\overline{M}'^{-1}.$$

Now

$$\overline{M}^{-1} = \begin{bmatrix} R_{11} & R_{13} \\ 0 & R_{23} \end{bmatrix}^{-1} = \begin{bmatrix} R_{11}^{-1} & -R_{11}^{-1}R_{13}R_{23}^{-1} \\ 0 & R_{23}^{-1} \end{bmatrix} \tag{5.19}$$

and (5.17) becomes

$$\hat{\delta}_{IV} \equiv \begin{bmatrix} b \\ c \end{bmatrix} = \begin{bmatrix} R_{11}^{-1} & -R_{11}^{-1}R_{13}R_{23}^{-1} \\ 0 & R_{22}^{-1} \end{bmatrix}\begin{bmatrix} R_{14} \\ R_{24} \end{bmatrix} \tag{5.20}$$
$$= \begin{bmatrix} R_{11}^{-1}R_{14} & -R_{11}^{-1}R_{13}R_{23}^{-1}R_{24} \\ & R_{23}^{11}R_{24} \end{bmatrix}.$$

Thus, the following computational steps result in the IV estimator:

1. Form $A = [X_1FYy]$ (order is important).
2. Take $K_1 + G$ steps in the QR decomposition of A to get

$$\begin{array}{c} \\ K \\ G \end{array}\begin{array}{c} \begin{array}{cccc} K & G & G & 1 \end{array} \\ \begin{bmatrix} R_{11} & R_{12} & R_{13} & R_{14} \\ 0 & R_{22} & R_{23} & R_{24} \end{bmatrix} \end{array}.$$

3. Solve $R_{23}c = R_{24}$, a rectangular system.
4. Solve $R_{11}b = R_{14} - R_{13}c$, a triangular system.
5. Obtain $e \equiv y - Yc - X_1b$ and form $s^2 = e'e/(T - K_1)$.
6. Form R_{11}^{-1}, R_{23}^{-1} and the $\overline{M}^{-1}\overline{M}'^{-1}$.

Operation Counts

The computational scheme just proposed for the IV estimator and its variance-covariance matrix has two advantages over direct computation of the moment matrices in the normal equations (5.3) and in (5.4): first, the proposed scheme employs the computationally stable QR decomposition and hence has advantages in dealing with collinear data; second, in most cases the proposed scheme is computationally more efficient in a direct comparison of operations counts. The exception occurs if $G \gg K_1$ (not a likely occurrence), and even here the disadvantage occurs in the computation only of $\hat{\delta}_{IV}$ but not of $V(\hat{\delta}_{IV})$.

Operation counts were made first for computing $\hat{\delta}_{IV}$ directly as in (5.3) and then for computing it as in Steps 1 through 4 above. The relative counts were

based on the following evaluation of numbers of operations:

Operation	Count
1. Solution to general square linear system ($n \times n$)	$\frac{1}{3}n^3 + n^2$
2. Solution to triangular linear system ($n \times n$)	$\frac{1}{2}n^2$
3. QR decomposition on $m \times n$ matrix (proportionately less if cut off early)	$\frac{2}{3}mn^2$
4. Inner product of $m \times n$ and $n \times p$ matrices	mnp
5. Inversion $n \times n$ matrix	$\frac{4}{3}n^3$

The method of calculation suggested above has in its favor $\frac{1}{3}T(K_1^2 - G^2) + 2TGK_1 + \frac{4}{3}TG^2 + \frac{2}{3}T(G + K_1) + \frac{1}{3}G^3 + G^2 + \frac{1}{2}K_1^2 + K_1G$ counts through the calculation of $\hat{\delta}_{IV}$ in (5.17) in comparison with the direct calculation of $\hat{\delta}_{IV}$ through (5.3). Only if $G \gg K_1$ will (5.3) prove more efficient. Comparison of the calculation of $V(\hat{\delta}_{IV})$ by (5.18) with the direct calculation of (5.4) offers clear additional evidence that the QR decomposition has a computational edge in all cases. Indeed, the calculation of (5.18) instead of (5.4) has these advantages: the entire $W'W$ matrix newly required by (5.12) need not be formed; only one matrix product need be taken instead of two; and inversion in (5.18) is of a $K_1 \times K_1$ upper triangular matrix and a $G \times G$ general matrix instead of the $(K_1 + G) \times (K_1 + G)$ general matrix $(W'Z)^{-1}$.

5.4. LIVE AND FIVE

The advantages of estimation by instrumental variables have been extended by the work of Brundy and Jorgenson (1971). Instrumental variables estimators, by their very structure, are consistent; but only in special cases do they also possess relative efficiency. Through a two-stage instrumental variables procedure, however, Brundy and Jorgenson (1971, 1973) have determined two efficient IV estimators, LIVE (Limited Information Instrumental Variables Efficient) and FIVE (Full Information Instrumental Variables Efficient). Whereas LIVE is called a "limited" information estimator, in fact both LIVE and FIVE are based on estimation of the full system of G equations. LIVE is "limited information" in the sense that it does not take into account any information on across-equation covariation. As a result, LIVE has the same asymptotic efficiency (Cramer–Rao lower bound) as LIML and 2SLS; while FIVE, which does employ information on across-equation covariation, has the same asymptotic efficiency as FIML. Indeed Hausman (1974) has shown that FIVE iterates to FIML.

In what follows, the calculations leading to LIVE and FIVE are examined in turn. The set of G equations to be estimated is

$$y_g = Y_g\gamma_g + X_g\beta_g + \varepsilon_g \qquad g = 1 \ldots G \tag{5.21}$$
$$= Z_g\delta_g + \varepsilon_g$$
$$\equiv [X_g Y_g]\begin{bmatrix}\beta_g \\ \gamma_g\end{bmatrix} + \varepsilon_g$$

where y_g is $T \times 1$, a vector of T observations on the normalized endogenous variable of equation g;
Y_g is $T \times G_g$, G_g endogenous variables included in equation g;
X_g is $T \times K_g$, K_g predetermined variables included in equation g;
γ_g is $G_g \times 1$, a vector of G_g nonzero parameters to be determined;
β_g is $K_g \times 1$, a vector of K_g nonzero parameters to be determined; and
ε_g is $T \times 1$, a vector of disturbance terms.
Further, define

$$U = [\varepsilon_1 \dots \varepsilon_G] \tag{5.22}$$

$$\Sigma_{\varepsilon\varepsilon} = \operatorname{plim} \frac{1}{T} U'U.$$

The First-Stage Estimates

Both LIVE and FIVE are two-stage estimators and assume that consistent (perhaps inefficient) estimates of the β_g and γ_g $g = 1 \dots G$ have been obtained in the first stage. In GREMLIN the user should be able to specify that any available consistent single-equation technique be used on any equation in the first stage. The k-class estimators or any IV estimator discussed above is a legitimate estimator for this purpose. The role of the LIVE and FIVE routine in the first stage is principally bookkeeping: specifying each equation in the system; generating data for the first-stage estimator for each equation; calling the relevant single-equation estimation package to carry out the estimation; and, finally, summarizing the first-stage results for use by the second-stage LIVE or FIVE estimator. This routine, therefore, draws upon all completed packages discussed above. The user should also be able simply to enter first-stage consistent estimators obtained from any other source.

Let $Y = \bigcup_g Y_g$, and $X = \bigcup_g X_g$, where $\bigcup$ denotes set union; and rewrite the system (5.21) as

$$Y\Gamma + X\beta + U = 0 \tag{5.23}$$

where Γ is a $G \times G$ square, nonsingular matrix whose g-th column contains (a) the associated elements of γ_g for each slot corresponding to a column of Y_g, (b) the value -1 corresponding to y_g, and (c) the value 0 elsewhere.
β is a $K \times G$ rectangular matrix whose g-th column is composed of the associated element of β_g for each slot corresponding to a column of X_g, and zeros elsewhere.
U is as in (5.22).

The stage-one estimation (assumed already accomplished) results in estimated vectors $\hat{\gamma}_g$ and $\hat{\beta}_g$, $g = 1 \dots G$ which together compose consistent estimates of Γ and β, denoted by $\hat{\Gamma}$ and $\hat{\beta}$.

Determining the Second-Stage Instruments

The reduced form of (5.23) is

$$Y = XB\Gamma^{-1} - U\Gamma^{-1} \tag{5.24}$$
$$= X\Pi + V$$
$$\Pi \equiv -B\Gamma^{-1} \qquad V \equiv -U\Gamma^{-1},$$

and the corresponding consistent estimator of Π (with zero restrictions) is

$$\hat{\Pi} \equiv -\hat{B}\hat{\Gamma}^{-1}. \tag{5.25}$$

The predicted values of Y from this estimated reduced form are simply

$$\hat{Y} \equiv X\hat{\Pi}. \tag{5.26}$$

These linear functions of the predetermined variables serve as the instruments in the second stage of LIVE and FIVE.

The predicted values $\hat{Y}$ can be computed in either of two equivalent ways. First, as implied by (5.26), $\hat{Y}$ can be computed directly by determining $\hat{\Gamma}$ and $\hat{B}$, inverting $\hat{\Gamma}$, and computing $-\hat{B}\hat{\Gamma}^{-1}$. Second once each of the G equations in the system has been consistently estimated in the first stage, the system can be subjected to static simulation to determine the $\hat{Y}$'s. A simulation facility such as that in TROLL (National Bureau of Economic Research, 1974) makes this second alternative attractive.

LIVE

For each equation g ($g = 1 \ldots G$)

$$y_g = Z_g\delta_g + \varepsilon_g = [X_g Y_g]\begin{bmatrix}\beta_g \\ \gamma_g\end{bmatrix} + \varepsilon_g \tag{5.27}$$

form a matrix of instruments $\hat{Y}_g$ as the G_g-predicted values from (5.21)—or the simulation—corresponding to those variates included in Y_g, the included endogenous variables of equation g.

The IV estimation technique of Section 5.3 can now be applied to the matrix

$$A_g = [X_g \hat{Y}_g Y_g y_g] \tag{5.28}$$

to obtain $\hat{\delta}_{g_{\text{LIVE}}}$, an efficient LIVE estimate of δ_g.

σ_g^2, the variance of ε_g, can be consistently estimated through the use of the first-stage consistent estimates $\hat{\delta}_g$ by forming

$$e_g \equiv y_g - Z_g\hat{\delta}_g \tag{5.29}$$

where $\hat{\delta}_g = \begin{bmatrix}\hat{\beta}_g \\ \hat{\gamma}_g\end{bmatrix}$, and $\hat{\beta}_g$ and $\hat{\gamma}_g$ are the first-stage consistent estimates used in (5.25). $\hat{\sigma}_g^2$ can now be estimated consistently as

$$s_g^2 = \frac{e_g'e_g}{T - K_1 - G} \quad \text{or} \quad \frac{e_g'e_g}{T}. \tag{5.30}$$

The Variance-Covariance Matrix. The variance-covariance matrix of the LIVE estimator takes a simpler form than that of the usual IV estimator. The asymptotic variance-covariance matrix of $T^{1/2}(\hat{\delta}_{\text{LIVE}}^{g} - \delta_g)$ is

$$\text{plim } T(W_g'Z_g)^{-1}(W_g'\varepsilon_g\varepsilon_g'W_g)(Z_g'W_g)^{-1} \tag{5.31}$$

$$= \sigma_g^2\Sigma_{W_gZ_g}^{-1}\Sigma_{W_gW_g}\Sigma_{Z_gW_g}^{-1} = \sigma_g^2\Sigma_{W_gW_g}^{-1}$$

since plim $T^{-1}W_g'Z_g = \text{plim } T^{-1}W_g'W_g$.

Hence a good estimator of the approximate variance-covariance matrix of $\hat{\delta}_{\text{LIVE}}^{g}$ is

$$V(\hat{\delta}_{\text{LIVE}}^{g}) = s_g^2(W_g'W_g)^{-1} \tag{5.32}$$

where $W = [X_g\hat{Y}_g]$, the $T \times (G_g + K_g)$ matrix of instruments. Reference to (5.16) indicates this is easily calculated from the elements of the QR decomposition already used to calculate $\hat{\delta}_{\text{LIVE}}^{g}$ as

$$s_g^2(M'M)^{-1} = s_g^2M^{-1}M'^{-1} \tag{5.33}$$

where $M = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$ as in (5.15) and is a matrix that is easily inverted due to its upper triangularity.

Across-Equation Covariance Matrix. It is also possible to make use of the LIVE estimates to obtain estimates of the asymptotic covariance between $\hat{\delta}_{\text{LIVE}}^{g}$ and $\hat{\delta}_{\text{LIVE}}^{h}$, the estimated coefficients from two separate equations. Indeed

$$\text{plim } T(\hat{\delta}_{\text{LIVE}}^{g} - \delta_g)(\hat{\delta}_{\text{LIVE}}^{h} - \delta_h)' \tag{5.34}$$

$$= \text{plim } T(W_g'Z_g)^{-1}W_g'\varepsilon_g\varepsilon_hW_h(Z_h'W_h)^{-1}$$

$$= \sigma_{gh}\Sigma_{W_gW_g}^{-1}\Sigma_{W_gW_h}\Sigma_{W_hW_h}^{-1}$$

where σ_{gh} is the gh-th element of $\Sigma_{\varepsilon\varepsilon}$ from (5.22) and $\Sigma_{W_gW_h} = \text{plim } T^{-1}W_g'W_h$ by definition.

Hence the approximate covariance between $\hat{\delta}_{\text{LIVE}}^{g}$ and $\hat{\delta}_{\text{LIVE}}^{h}$ is estimated by

$$s_{gh}(\overline{M}_g'\overline{M}_g)^{-1}(W_g'W_h)(\overline{M}_h'\overline{M}_h)^{-1} \tag{5.35}$$

The first and last of the three matrix terms in (5.35) have already been computed in (5.33), when $V(\hat{\delta}_{\text{LIVE}}^{g})$ and $V(\hat{\delta}_{\text{LIVE}}^{h})$ were computed. The middle term, $W_g'W_h$, must be computed anew for these calculations. The estimated covariance is calculated as

$$s_{gh} = \frac{e_g'e_h}{T} \tag{5.36}$$

where e_g and e_h are determined from (5.29).

Summary of Computational Steps for LIVE. The computational steps for the LIVE estimator can be summarized as follows:

1. For each g determine first-stage consistent estimates of $\hat{\delta}_g$ and s_g^2 as in (5.29).

2. From these $\hat{\delta}_g$, form $\hat{\Gamma}$, $\hat{\mathbf{B}}$ and $\hat{\Pi} = -\hat{\mathbf{B}}\hat{\Gamma}^{-1}$.
3. Form the LIVE instruments $\hat{Y} = X\hat{\Pi}$, *or* determine $\hat{Y}$ directly by simulation from Step 1. (In the latter case skip Step 2.)
4. Form $A_g = [X_g \hat{Y}_g Y_g y_g]$ and use the IV estimator of Section 5.1 to determine $\hat{\delta}_{\text{LIVE}}$.
5. Form M^{-1} and calculate $V(\hat{\delta}_{\text{LIVE}})$ as $s_g^2 M^{-1} M'^{-1}$ from (5.33).
6. Calculate $s_{gh} W_g' W_{hg}$ and form (5.35) along with the $(M_g' M_g)^{-1}$ from Step 5.

FIVE

FIVE, like LIVE, is a full-system estimator. It is "full information" relative to LIVE not in the sense that it requires full specification of the entire system (for both LIVE and FIVE require this), but in the sense that FIVE takes into account the across-equation covariation ignored by LIVE. The asymptotic efficiency of FIVE, therefore, is the same as that of 3SLS and FIML.

The FIVE estimator uses the same building blocks as LIVE, but unfortunately the resulting equation system cannot be solved in a way that exploits the computationally efficient algorithm of Section 5.3. Instead the IV normal equations must be solved in their basic form (5.11).

The FIVE Instruments. FIVE begins exactly as does LIVE: for each equation $g = 1 \dots G$, a $\hat{\delta}_g$ and $e_g = y_g - Z_g \hat{\delta}_g$ are determined from (5.29) through some consistent (but perhaps inefficient) single-equation estimator. From these $\hat{\delta}_g$ and e_g one forms $\hat{\Gamma}$, $\hat{\mathbf{B}}$, and $\hat{\Pi} = \hat{\mathbf{B}}\hat{\Gamma}^{-1}$ from (5.25) and $S = T^{-1}U'U$, where $U = [e_1 \dots e_G]$, a $T \times G$ matrix of estimated residuals. S is clearly an estimated covariance matrix whose elements s_{gh} will be used to weight the blocks in the FIVE normal equations.

For each equation, then, a set of instruments is formed as

$$W_g = [X\hat{\Pi}_g \; X_g], \qquad \text{a } T \times (G_g + K_g) \text{ matrix}, \; g = 1 \dots G, \tag{5.37}$$

where $\hat{\Pi}_g$ is the $K \times G_g$ submatrix of $\hat{\Pi}$ formed by taking only the columns of $\hat{\Pi}$ corresponding to the G_g endogenous variables Y_g included in equation g.[36] From these a set of cross-equation blocks is formed as

$$W_{gk} = s^{gk} W_k, \qquad g, k = 1 \dots G \tag{5.38}$$

where s^{gk} is the gk element of $\bar{S}^{-1}$.

Finally, a complete instrument matrix is formed as

$$\overline{W} = \begin{bmatrix} W_{11} & W_{12} & W_{1G} \\ \vdots & \vdots & \vdots \\ W_{G1} & W_{G2} \dots & W_{GG} \end{bmatrix}, \tag{5.39}$$

a matrix of size $GT \times \Sigma_g (G_g + K_g)$.

[36] That is, if equation g includes only $Y_1 Y_5 Y_8$ and Y_9, then $\hat{\Pi}_g$ would consist of columns 1, 5, 8, and 9 of $\hat{\Pi}$—or equivalently, $\hat{Y}_g \equiv X\hat{\Pi}_g$ would consist of columns 1, 5, 8, and 9 of Y.

The FIVE Normal Equations. The g-th equation of the system is $y_g = Z_g\delta_g + \varepsilon_g$, and the full system to which the instruments $\overline{W}$ of (5.39) are applied is

$$y = Z\delta + \varepsilon \tag{5.40}$$

where

$$y \equiv \begin{bmatrix} y_1 \\ \vdots \\ y_G \end{bmatrix}, \quad Z \equiv \begin{bmatrix} Z_1 & & 0 \\ & \ddots & \\ 0 & & Z_G \end{bmatrix}, \quad \delta \equiv \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_G \end{bmatrix} \text{ and } \varepsilon \equiv \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_G \end{bmatrix}$$

The normal equations from which the FIVE estimator $\hat{\delta}_{\text{FIVE}}$ is solved are

$$(\overline{W}'Z)\hat{\delta}_{\text{FIVE}} = \overline{W}'y, \tag{5.41}$$

in general a very large system, for $\overline{W}'Z$ is square and of size $\Sigma_g(G_g + K_g)$. $\overline{W}'Z$ should be formed directly, and MINFIT or some other suitable routine should be applied directly to (5.41). However, $\overline{W}$ need not be formed and stored as a whole, for its G^2 blocks are composed only of the G matrices W_g from (5.37) and the elements from the $G \times G$ matrix S. $\overline{W}$ can be formed piecemeal, as required, from these building blocks, while $\overline{W}'Z$ and $\overline{W}'y$ in (5.41) are being formed. Likewise, the full $TG \times \Sigma(G_g + K_g)$ matrix Z need never be formed, for it is block diagonal with blocks Z_g, $g = 1 \ldots G$, from (5.40). The block multiplication which forms $\overline{W}'Z$ can therefore take advantage of the sparsity of both $\overline{W}$ and Z.

The Variance-Covariance Matrix. The estimated variance-covariance matrix of $\hat{\delta}_{\text{FIVE}}$ is easily formulated but presents computational difficulties because it is usually very large. The true asymptotic variance-covariance matrix of $T^{1/2}(\hat{\delta}_{\text{FIVE}} - \delta)$ is

$$\begin{aligned} &= \text{plim } T(\overline{W}'Z)^{-1}\overline{W}'\varepsilon\varepsilon'\overline{W}(Z'\overline{W})^{-1} \\ &= [\sigma^{gh}\Sigma_{w_g w_h}]^{-1}, \end{aligned} \tag{5.42}$$

where σ^{gh} is gh-th element of $\Sigma_{\varepsilon\varepsilon}^{-1}$ and

$$\Sigma_{w_g w_h} = \text{plim } \frac{1}{T} W_g' W_h.$$

Hence the estimated approximate variance-covariance matrix of $\hat{\delta}_{\text{FIVE}}$ is given by

$$\hat{V}(\hat{\delta}_{\text{FIVE}}) = [s^{gh}W_g'W_h]^{-1} \qquad g, h = 1 \ldots G \tag{5.43}$$

a square, symmetric matrix of size $\Sigma_g(G_g + K_g)$.

In general $V(\hat{\delta}_{\text{FIVE}})$ is large, and an inversion routine capable of such matrices is required.

Prior Restriction on $\Sigma_{\varepsilon\varepsilon}$. As in 3SLS, the calculations involved in computing $\hat{\delta}_{\text{FIVE}}$ from (5.41) and $\hat{V}(\hat{\delta}_{\text{FIVE}})$ from (5.43) can be substantially reduced if some of the s^{gh} are constrained to be zero. This would be the case if Σ were assumed to be block diagonal from the outset. So also, then, would be Σ^{-1}, and both (5.41) and (5.43) would be sparse. In this case routines exploiting the resulting block de-

composition should be utilized to reduce the calculations to several systems of smaller size.

Summary of Steps for FIVE. The computational steps for the FIVE estimator can be summarized as follows:

1. For each g, determine first-stage consistent estimates $\hat{\delta}_g$; and for each g and h, determine s_{gh} as in (5.36) and form $S = (s_{gh})$.
2. From the $\hat{\delta}_g$, form $\hat{\Gamma}$, $\hat{\mathbf{B}}$ and $\hat{\Pi} = -\hat{\mathbf{B}}\hat{\Gamma}^{-1}$ as in (5.25).
3. Form the FIVE instruments $\hat{Y} = X\hat{\Pi}$, *or* determine $\hat{Y}$ directly by simulation from Step 1. (In the latter case skip Step 2.)
4. Form $W_g = [\hat{Y}_g \; X_g] \equiv [X\hat{\Pi}_g \; X_g)$ for $g = 1 \ldots G$.
5. Calculate S^{-1} and form $W_{gh} = s^{gh}W_h, \quad g, h = 1 \ldots G$.
6. Form $(\overline{W}'Z)$ and $(\overline{W}'y)$ as

$$(\overline{W}'Z) = \{s^{hg}W_g'Z_h\}$$

$$(\overline{W}'y) \equiv \left\{\sum_{h=1}^{G} s^{hg}W_g'y_h\right\}$$

$$= \left\{W_g' \sum_{h=1}^{G} s^{hg}y_h\right\}.$$

7. Calculate $\hat{\delta}_{\text{FIVE}}$ as $(\overline{W}'Z)\hat{\delta}_{\text{FIVE}} = \overline{W}'y$.
8. Form $V = [s^{gh}W_g'W_h], g, h = 1 \ldots G$, and calculate V^{-1} as $V(\hat{\delta}_{\text{FIVE}})$, recalling that V is symmetric.

Steps 6–8 should take advantage of any zero restrictions given on Σ_{zz}.

Iterative FIVE. Hausman (1974) has shown that the FIVE estimator (5.41) iterates to the FIML estimator of δ. Iteration of FIVE proceeds as follows: an initial estimate $\hat{\delta}^{(1)}_{\text{FIVE}}$ is determined as in the previous section. The $\hat{\delta}^{(1)}_{\text{FIVE}}$ becomes the $\hat{\delta}_g$ of Step 1, and a new estimate, $\hat{\delta}^{(2)}_{\text{FIVE}}$, is produced. This in turn is used at Step 1 until an effective convergence of $\hat{\delta}^{(r)}_{\text{FIVE}} \doteq \hat{\delta}^{(r-1)}_{\text{FIVE}}$ occurs. Step 8 need be calculated only once, at the end.

The user should have the option of stopping the iterations prior to convergence. Because FIVE is a consistent and asymptotically efficient estimator for any consistent initial estimates in Step 1, each $\delta^{(r)}_{\text{FIVE}}$ is consistent and asymptotically efficient. Stopping before convergence, therefore, is costless in terms of these asymptotic properties. Only when convergence is reached, however, will the iterated FIVE estimate also be FIML.

Appendix. Iterative Procedures for Nonlinear Equations

A.0. Introduction

The purpose of this appendix is to examine estimation of a single equation that is nonlinear in its parameters and to develop in detail the notation and terminology utilized in Section 1.5.

A model that is appropriate to OLS is considered first. Then the results are extended to a model that is appropriate to estimation of one equation from a simultaneous system—i.e., one equation having endogenous regressors. A Gauss–Newton method (using first derivatives only) is developed first; this technique was employed in earlier versions of TROLL (National Bureau of Economic Research, 1974) but often failed to converge. A Newton–Raphson (second-derivative) technique was used with greater success, and this technique is presented next and adapted for use in simultaneous equations.

A.1. Procedure with Exogenous Coterms

Assume T observations on the outcome of a nonlinear random function f^t in K observed arguments $x(t)$ and having G unknown constant parameters (nonlinear) β which are the object of estimation.[37] Hence in period t assume

$$-f^t(x(t), \beta) = \varepsilon_t, \tag{A.1}$$

where ε_t is a random variable having mean zero, constant variance and independent across time.

In matrix summary we have

$$-f(X, \beta) = \varepsilon = -\begin{bmatrix} f^1(x(1), \beta) \\ \vdots \\ f^T(x(T), \beta) \end{bmatrix} \tag{A.2}$$

where f is a T-vector function

X is the $T \times K$ data matrix $X = \begin{bmatrix} x'(1) \\ x'(T) \end{bmatrix}$

β is the G-vector of parameters to be estimated

ε is distributed with mean 0, and

$V(\varepsilon) = \sigma^2 I$.

Further, define the Jacobian matrix of *coterms*[38]

$$f_\beta \equiv \frac{\partial f}{\partial \beta} \equiv \begin{bmatrix} f_1^1 & \cdots & f_G^1 \\ \vdots & & \\ f_1^T & & f_G^T \end{bmatrix} \text{ a } T \times G \text{ matrix.} \tag{A.3}$$

In this section f_β is assumed to be nonstochastic; i.e., the partial of f (which is a stochastic function) with respect to all parameters is assumed to be nonstochastic. This assumption is appropriate to a nonlinear generalization of the context of

[37] The notation f^t means not that function f is different in each period—it is in fact the same function for all t—but that it is evaluated at different $x(t)$.

[38] The meaning of "coterms" will become apparent in equation (A.5) below, where coterms are paired up with their corresponding β's in the linearized approximation. Also see Eisner and Pindyck, 1973.

OLS with all the regressors exogenous. The assumption will be relaxed in Section A.2.

Linearized OLS: A Gauss–Newton Procedure

Using the first two terms of a Taylor expansion about β_0, linearize (A.2) as

$$\varepsilon = -f(\beta) = -f(\beta_0) - f_\beta(\beta_0)[\beta - \beta_0] \tag{A.4}$$
$$= [-f(\beta_0) + f_\beta(\beta_0)\beta_0] - f_\beta(\beta_0)\beta$$

or

$$f_\beta(\beta_0)\beta_0 - f(\beta_0) = f_\beta(\beta_0)\beta + \varepsilon, \tag{A.5}$$

[39]

where all partials are evaluated at β_0. For given β_0, OLS can be applied to (A.5) to obtain the least squares estimator.

$$\hat{\beta} = [f'_\beta(\beta_0)f_\beta(\beta_0)]^{-1}f'(\beta_0)[f_\beta(\beta_0)\beta_0 - f(\beta_0)] \tag{A.6}$$
$$= \beta_0 - [f'_\beta(\beta_0)f_\beta(\beta_0)]^{-1}f'_\beta(\beta_0)f(\beta_0).$$

The form of (A.6) suggests the iterative procedure

$$b_{r+1} = b_r - [f'_\beta(b_r)f'_\beta(b_r)]^{-1}f'_\beta(b_r)f(b_r). \tag{A.7}$$

This method, employed in earlier versions of TROLL, displayed some difficulties in converging, and was replaced by the Newton–Raphson procedure described next.

A Newton–Raphson Procedure

If (A.7) converges so that $b_{r+1} = b_r = b$, then it reduces to

$$-[f'_\beta(b)f_\beta(b)]^{-1}f'_\beta(b)f(b) = 0 \tag{A.8}$$

or equivalently

$$f'_\beta(b)f(b) = 0. \tag{A.9}$$

This set of normal equations, whose solution is necessarily the same as a convergent solution of (A.7), can also be derived from minimizing the sampling sum of squared errors from

$$e = -f(X, b), \tag{A.10}$$

i.e., the solution of

$$\min_b e'e = f'f. \tag{A.11}$$

[40]

Define $F(b)$ as

$$F(b) \equiv f'_\beta(b)f(b) = 0 \tag{A.12}$$

[39] The use of "coterms" should be clear from (A.5). In the linearized model, the f_β serve the same function relative to the parameters β as the X's do in the standard linear model $y = X\beta + \varepsilon$.

[40] Differentiating (A.11) produces $2f_\beta'(b)f(b) = 0$.

and expand F about b_0 to obtain

$$0 = F(b) = F(b_0) + F_\beta(b_0)[b - b_0] \tag{A.13}$$

where $F_\beta = \partial F/\partial\beta$, a $G \times G$ nonstochastic matrix.

Solving (A.13) gives

$$b = b_0 - F_\beta^{-1}(b_0)F(b_0), \tag{A.14}$$

which, rewritten in terms of f, becomes

$$b = b_0 - \left[\sum_{t=1}^{T} \mathscr{F}^t f^t + f_\beta' f_\beta\right]^{-1} f_\beta' f, \tag{A.15}$$

where $\mathscr{F}^t$ is the $G \times G$ Hessian matrix $\left[\dfrac{\partial^2 f^t}{\partial\beta_k \partial\beta_g}\right] \equiv (f^t_{gk})$.

Iteration in terms of (A.15) is like that in terms of (A.7), except that a second-derivative term $\Sigma_{t=1}^{T} \mathscr{F}^t f^t$ is included additively in the inverse.[41]

A.2. Procedure with Endogenous Coterms

In Section A.1 the coterm matrix f_β is assumed nonstochastic; this is the nonlinear analog to the OLS case. Now, however, nonlinear estimation is extended to simultaneous equations; hence it is assumed that:

1. f_β is a stochastic matrix (some of whose elements may be nonstochastic).
2. X_I (distinct from X) is a set of H preliminary regressors assumed independent of ε (i.e., of the stochastic elements of f—and hence, also of f_β).

Instrumental Variables in the Limited Information Case

Begin with the linearized equation (A.5), in which, however, f_β is no longer independent of ε. Application of OLS to (A.5) is no longer indicated; instead, a set of instruments $\hat{f}_\beta$ is introduced by regressing f_β on X_I.

$$\begin{aligned} \hat{f}_\beta &= X_I P = X_I(X_I'X_I)^{-1}X_I' f_\beta \\ &\equiv Z_I f_\beta \end{aligned} \tag{A.16}$$

where

$$Z_I \equiv X_I(X_I'X_I)^{-1}X_I'. \tag{A.17}$$

In the spirit of instrumental variables, f_β in the right-hand side of (A.5) is replaced by $\hat{f}_\beta$ from (A.16). As will be clear from (A.20), there is no need to purge the left-hand side of its stochastic terms; hence, the estimation is based on

$$f_\beta(\beta_0)\beta_0 - f(\beta_0) = \hat{f}_\beta(\beta_0)\beta + \eta \tag{A.18}$$

where

$$\begin{aligned} \eta &= \varepsilon + [f_\beta(\beta_0) - \hat{f}_\beta(\beta_0)]\beta \\ &\equiv \varepsilon + V\beta. \end{aligned}$$

[41] $F = f_\beta' f = [\sum_{t=1}^{T} f^t_g f^t] \equiv [F^g]$ a $G \times 1$ vector. Then $F^g_k \equiv \partial F^g/\partial\beta_k = \sum_t (f^t_{gk} f^t + f^t_g f^t_k)$, and hence $F_\beta = [F^g_k] = [\sum_{t=1}^{T} \mathscr{F}^t f^t_\beta + f_\beta' f_\beta]$.

Least squares applied to (A.18) gives

$$\hat{\beta} = (\hat{f}'_\beta \hat{f}'_\beta)^{-1} \hat{f}'_\beta (f_\beta \beta_0 - f) = \beta_0 - (\hat{f}'_\beta \hat{f}_\beta)^{-1} \hat{f}'_\beta f \tag{A.19}$$

which uses the fact that

$$\hat{f}'_\beta f_\beta = (f'_\beta Z_I) f_\beta = (f' Z_I)(Z_I f_\beta) = \hat{f}'_\beta \hat{f}_\beta \tag{A.20}$$

due to the idempotency of Z_I. This last fact proves that f_β need not be adjusted by Z_I on the left-hand side of (A.18).

The iterative procedure suggested by (A.19), and analogous to (A.7), is

$$b_{r+1} = b_r - [\hat{f}'_\beta(b_r) \hat{f}_\beta(b_r)]^{-1} \hat{f}'_\beta(b_r) f(b_r). \tag{A.21}$$

Newton–Raphson in the Limited Information Case

If (A.21) converges to $b_{r+1} = b_r = b$, then again the normal equations

$$\hat{f}'_\beta(b) f(b) = 0 \tag{A.22}$$

must be satisfied by b. An alternative to finding b is therefore offered by solving (A.22) for b by Newton's method.

Using (A.16), let

$$\tilde{F} \equiv \hat{f}'_\beta f = f'_\beta Z_I f = 0. \tag{A.23}$$

Expanding $\tilde{F}$ gives

$$0 = \tilde{F}(\beta) = \tilde{F}(\beta_0) + \tilde{F}_\beta(\beta_0)(\beta - \beta_0) \tag{A.24}$$

or

$$\beta = \beta_0 - \tilde{F}_\beta^{-1}(\beta_0) \tilde{F}(\beta_0). \tag{A.25}$$

Rewriting (A.25) in terms of f gives

$$\beta = \beta_0 - [G + \hat{f}'_\beta \hat{f}_\beta]^{-1} \hat{f}'_\beta f \tag{A.26}$$

where

$$G \equiv (g_{gk})$$

and

$$g_{gk} = \Sigma_t \Sigma_\tau f^t_{gk} Z_{t\tau} f^\tau \qquad g, k = 1 \ldots G,$$

$$Z \equiv (Z_{t\tau}),$$

$$f^t_{gk} = \left(\frac{\partial^2 f^t}{\partial \beta_k \partial \beta_g} \right);$$

alternatively,

$$g_{gk} = f'_{gk} Z f$$

where

$$f_{gk} = \begin{bmatrix} f^1_{gk} \\ \vdots \\ f^T_{gk} \end{bmatrix}.$$

Also

$$G = \Sigma_t \mathcal{F}^t \hat{f}^t.$$

The properties of (A.26) as an estimator need to be investigated. Clearly the estimator is consistent if the stochastic nature of the auxiliary relationship between f_β and the instruments X_I approaches that assumed behind (A.16), i.e., if $\hat{f}_\beta$ constantly estimates f_β. Otherwise the properties of the resulting estimator depend upon the true stochastic relation between f_β, X_I, V, and ε.

A.3. The Double-k Class Adaptation

The preceding adjustment procedure can be generalized to the double-k class context. Instead of regressing f_β on X_I (effectively the 2SLS option), calculate

(A.27) $$(f'_\beta f_\beta)_{\perp X_I} \quad \text{and} \quad (f'_\beta f)_{\perp X_I},$$

to use in an iterative scheme generalizing the basic double-k class estimator (1.2).

Gauss–Newton Generalization

Applying the Gauss–Newton iterative procedure analogous to (A.7) to the double-k class estimator (1.2) results in the following iterative scheme:

(A.28) $$b_{r+1} = b_r - [f'_\beta f_\beta - k_1(f'_\beta f_\beta)_{\perp X_I}]^{-1}[f'_\beta f - k_2(f'_\beta f)_{\perp X_I}].$$

The Newton–Raphson Generalization

The analogous adaptation of the Full–Newton Step would be

(A.29) $$b_{r+1} = b_r - [\bar{G} + f'_\beta f_\beta - k_1(f'_\beta f_\beta)_{\perp X_I}]^{-1}[f'_\beta f - k_2(f'_\beta f)_{\perp X_I}]$$

where

$$\bar{G} = (\bar{g}_{gk}), \quad \bar{g}_{gk} = f'_{gk} J f$$
$$J = I - k_1 H$$
$$H = I - Z$$
$$Z = X_I (X'_I X_I)^{-1} X'_I ,$$

or

$$\bar{g}_{gk} = f'_{gk} f - k_1 (f'_{gk} f)_{\perp X_I}.$$

Some f_β Nonstochastic

When not every element of f_β is stochastic, some partials can be functions of the X_I alone, and f_β can be partitioned as

(A.30) $$f_\beta = [\phi_\beta \; \chi_\beta]$$

where ϕ_β is the matrix of stochastic coterms, and χ_β is the matrix of nonstochastic coterms. Estimation can now proceed by adjusting only the ϕ_β, as, for example,

with a Newton–Raphson step of

$$(A.31)\qquad b_{r+1} = b_r - \left[\tilde{G} + \begin{bmatrix} \phi_\beta'\phi_\beta - k_1(\phi_\beta'\phi_\beta)_{\perp X_I} & \phi_\beta'\chi_\beta \\ \chi_\beta'\phi_\beta & \chi_\beta'\chi_\beta \end{bmatrix}\right]^{-1} \cdot \begin{bmatrix} \phi_\beta' f - k_2(\phi_\beta' f)_{\perp X_I} \\ \chi_\beta' f \end{bmatrix}.$$

Should the χ_β be included with the X_I as instruments? Some may already be there if, for example, χ_β has a term *linear* in the X_I. Either these linear equivalences must somehow be purged; or, as is the case with most procedures considered in this paper, the determination of $(\phi_\beta'\phi_\beta)_{\perp X_I^*}$ (where X_I^* is the set of X_I augmented by χ_β) must be able to proceed even if X_I^* is singular. At least one computational consideration is apparent: with a fixed X_I, many calculations can be saved in determining $(\phi_\beta'\phi_\beta)_{\perp X_I}$, but X_I^* will change with each iteration and cause recalculation of $Z_I = X_I^*(X_I'^*X_I^*)^{-1}X_I'^*$.

Boston College and
NBER Computer Research Center

BIBLIOGRAPHY

Amemiya, Takeshi (1966). "On The Use of Principal Components of Independent Variables in Two-Stage Least Squares Estimation," *International Economic Review*, Vol. 7, No. 3, September, pp. 283–303.

Amemiya, Takeshi (1973). "The Non-linear Two Stage Least Squares Estimator," Working Paper No. 25, The Economic Series, Institute for Mathematical Studies in The Social Sciences, Stanford University, June.

Anderson, T. W. (1972). "An Asymptotic Expansion of The Distribution of the Limited Information Maximum Likelihood Estimate of a Coefficient in a Simultaneous Equation System." Technical Report No. 73, The Economic Series Institute for Mathematical Studies in The Social Sciences, Stanford University, September.

Anderson, T. W. and Takamitsu Sawa (1973). "Distribution of Estimates of Coefficients of a Single Equation in a Simultaneous System and Their Asymptotic Expansions." *Econometrica*, Vol. 41.

Basmann, R. L. (1961). "A Note on the Exact Finite Sample Frequency Functions of Generalized Classical Linear Estimators in Two Leading Overidentified Cases," *Journal of The American Statistical Association*, 56, pp. 619–636.

Basmann, R. L. (1963). " A Note on the Exact Finite Sample Frequency Functions of Generalized Classical Linear Estimators in a Leading Three Equation Case." *Journal of the American Statistical Association*, 58, pp. 161–171.

Bauer, F. L. (1971). "Elimination with Weighted Row Combinations for Solving Linear Equations and Least Squares Problems," pp. 119–133, in *Handbook for Automatic Computation, Vol. II: Linear Algebra*, Eds. Wilkinson, J. H. and Reinsch, C. New York: Springer-Verlag.

Becker, R., N. Kaden, and V. Klema (1974). "The Singular Value Analysis in Matrix Computation," NBER Working Paper No. 46, Computer Research Center for Economics and Management Science, National Bureau of Economic Research, Inc.

Brundy, James M. and Dale W. Jorgenson (1971). "Efficient Estimation of Simultaneous Equations by Instrumental Variables," *The Review of Economics and Statistics*, Vol. 53 No. 3, August, pp. 207–224.

Brundy, James M. and Dale W. Jorgenson (1973). "Consistent Efficient Estimation of Systems of Simultaneous Equations by Means of Instrumental Variables," Technical Report No. 92, March, Institute for Mathematical Studies in the Social Sciences, Stanford University.

usinger, P. and Golub, G. H. (1965). "Linear Least Squares Solutions by Householder Transformations." *Numerische Mathematik*, Vol. 7, pp. 269–276.

Chow, Gregory C. (1972). "On The Computation of Full-Information Maximum Likelihood Estimates for Non-linear Equation Systems," Econometric Research Program Research Memorandum No. 142, Princeton University. Forthcoming *Review of Economics and Statistics*.

Chow, Gregory C. (1973). "A Family of Estimators for Simultaneous Equation Systems," Econometric Research Program, Research Memorandum No. 155, Princeton University. October.

Chow, Gregory C. and Ray C. Fair (1973). "Maximum Likelihood Estimation of Linear Equation Systems with Auto-regressive Residuals," *Annals of Economic and Social Measurement*, Vol. 2, No. 1.

Cragg, J. G. (1966). "On The Sensitivity of Simultaneous Equations Estimators to The Stochastic Assumptions of The Models," *Journal of The American Statistical Association*, Vol. 61, pp. 136–151.

Cragg, J. G. (1967). "On The Relative Small Sample Properties of Several Structural Equation Estimators," *Econometrica*, Vol. 35, pp. 89–110.

Dent, Warren and Gene H. Golub (1973). "Computation of The Limited Information Maximum Likelihood Estimator." Unpublished Manuscript, STAN-CS-73-339, Computer Science Department, Stanford University.

Duesenberry, J., G. Fromm, L. Klein, and E. Kuh (editors) (1965). *The Brookings Quarterly Econometric Model of The United States*. Amsterdam: North-Holland Publishing Company.

Eisenpress, Harry and John Greenstadt (1966). "The Estimation of Nonlinear Econometric Estimators," *Econometrica*, Vol. 34, No. 4, October, pp. 851–861.

Eisner, Mark and Robert S. Pindyck (1973). "A Generalized Approach to Estimation as Implemented in the TROLL/1 System," *Annals of Economic and Social Measurement*, Vol. 2, No. 1.

Fair, Ray C. (1972). "Efficient Estimation of Simultaneous Equations with Auto-regressive Errors by Instrumental Variables," *Review of Economics and Statistics*, Vol. LIV, No. 4, November, pp. 444–449.

Forsythe, G. and C. Moler (1967). *Computer Solution of Linear Algebraic Systems*, Englewood Cliffs, N.J.: Prentice-Hall.

Golub, Gene H. (1969). "Matrix Decompositions and Statistical Calculations." *Statistical Computation*, Academic Press, New York, pp. 365–397.

Golub, G. H. and C. Reinsch, (1970). "Singular Value Decomposition and Least Squares Solutions." *Numerische Mathematik*, Vol. 14, pp. 403–420.

Golub, Gene H. and V. Pereyra (1973). "Differential of Pseudo-inverses, Separable Nonlinear Least Squares Problems, and Other Tales." Unpublished Manuscript, Computer Science Department, Stanford University.

Graybill, F. A. (1969). *Introduction to Matrices with Applications in Statistics*. Belmont, California: Wadsworth Publishing Company, Inc.

Greville, T. N. E. (1959). "The Pseudo-inverse of a Rectangular or Singular Matrix and Its Application to the Solution of Systems of Linear Equations," *SIAM Review*, Vol. 1, pp. 38–43.

Griliches, Zvi and Potluri Rao (1969). "Small Sample Properties of Several Two-Stage Regression Methods in The Context of Auto-Correlated Errors," *Journal of The American Statistical Association*, March.

Hanson, Richard J. and Charles L. Lawson (1969). "Extensions and Applications of The Householder Algorithm for Solving Linear Least Squares Problems." *Mathematics of Computation*, Vol. 23, No. 108, October, pp. 787–812.

Hausman, J. A. (1973). "An Instrumental Variable Interpretation of FIML and Associated Estimators," Unpublished Manuscript, Dept. of Economics, Massachusetts Institute of Technology.

Johnston, J. (1972). *Econometric Methods*, 2nd Edition, McGraw-Hill, New York.

Jorgenson, Dale W. (1974). *Annals of Economics and Social Measurement*, This issue.

Kadane, J. B. (1971). "Comparison of *k*-Class Estimators When the Disturbances Are Small," *Econometrica*, 39, 723–737.

Kelejian, Harry H. (1971). "Two-Stage Least Squares and Econometric Systems, Linear in Parameters but Nonlinear in The Endogenous Variables," *Journal of The American Statistical Association*, Vol. 66, No. 334, June, pp. 373–374.

Kelejian, H. H. (1973). "Efficient Instrumental Variable Estimation of Large Scale Nonlinear Econometric Models," Unpublished Manuscript, New York University.

Klein, L. and A. S. Goldberger (1955). *An Econometric Model of the United States*, 1929–52. North-Holland Publishing Co., Amsterdam.

Kloek, T. and L. B. M. Mennes (1960). "Simultaneous Equations Estimation Based on Principal Components of Predetermined Variables," *Econometrica*, Vol. 28, No. 1, pp. 45–61.

Lanczos, C. (1961). *Linear Differential Operators*, Van Nostrand: London, Chapter 3.

Mariano, Roberto S. and Takamitser Sawa (1971). "Exact Finite Sample Distribution of the Limited-Information Maximum Likelihood Estimator in the Case of Two Included Endogenous Variables," Forthcoming, *Journal of the American Statistical Association*.

Marschak, J. (1953). "Economic Measurement for Policy and Prediction," in "Studies in Econometric Method," Hood, W. C. and T. C. Koopmans, eds., *Cowles Commission* Monograph, No. 14, John Wiley: New York.

Nagar, A. L. (1959). "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica*, XXVII, October, pp. 575–595.
Nagar, A. L. (1960). "A Monte Carlo Study of Alternative Simultaneous Equation Estimators," *Econometrica*, Vol. 28, p. 573–590.
National Bureau of Economic Research, Inc. (1971). *TROLL Reference Manual* (*Standard System*), Computer Research Center for Economics and Management Science, D0062, Chapter 11.
Osborne, E. E. (1961). "On Least Squares Solutions of Linear Equations," *Jour. Assoc. Comput. Mach.*, Vol. 8, p. 628–636.
Peters, G. and J. H. Wilkenson (1970). "The Least Squares Problem and Pseudo Inverses." *The Computer Journal*, Vol. 13, No. 3, August, pp. 309–316.
Quandt, Richard E. (1962). "Some Small Sample Properties of Certain Structural Equation Estimators," Economic Research Program, Princeton University, Research Memorandum No. 48, December.
Quandt, Richard E. (1965). "On Certain Small Sample Properties of k-Class Estimators," *International Economic Review*, Vol. 6, pp. 92–104.
Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*. New York: John Wiley and Sons, Inc.
Richardson, David H. (1968). "The Exact Distribution of a Structural Coefficient," *Journal of the American Statistical Association*, Vol. 13, pp. 1214–1226.
Ruble, W. L. (1968). *Improving the Computation of Simultaneous Stochastic Linear Equations*. Doctoral Dissertation, Michigan State University, Department of Statistics.
Sawa, T. (1969). "The Exact Sampling Distribution of Ordinary Least Squares and Two-Stage Least Squares Estimators," *Journal of the American Statistical Association*, 64, pp. 923–937.
Sawa, T. (1970). "Finite-Sample Properties of the k-Class Estimators," Technical Report No. 36, Institute of Matnematical Studies, Stanford University, forthcoming *Econometrica*.
Sawa, T. (1973a). "Almost Unbiased Estimator in Simultaneous Equations Systems." *International Economic Review*, Vol. 14, No. 1, February, pp. 97–106.
Sawa, T. (1973b). "The Mean Square Error of a Combined Estimator and Numerical Comparison with the TSLS Estimator." *Journal of Econometrics* #1, pp. 115–132.
Summers, R. (1965). "A Capital Intensive Approach to The Small Sample Properties of the Various Simultanous Equation Estimators," *Econometrica*, Vol. 33, pp. 1–41.
Theil, H. (1971). *Principles of Econometrics*, New York: John Wiley and Sons, Inc.
Wagner, H. (1958). "A Monte Carlo Study of Estimates of Simultaneous Linear Structural Equations," *Econometrica*, Vol. 26, pp. 117–133.