COWLES COMMISSION DISCUSSION PAPER: STATISTICS: NO. 310A

Supplement to "Statistical Methods of Measuring Economic Relationships," based on lectures given at the University of Chicago, Autumn 1948, by

Tjalling C. Koopmans

Notes taken and elaborated by
Stephen G. Allen
(Occasional substitute: George H. Borts.)

These supplementary notes embody a number of minor improvements and one substantial modification of previous class notes (Discussion Paper 310). This modification is the explicit derivation of the likelihood function corresponding to a linear system of structural equations, and its successive maximization with respect to suitably chosen subsets of its parameters. In this way we have derived the limited information method of estimating the parameters of a subset of one or more out of a complete system of equations, originally developed in a somewhat different manner by Anderson and Rubin.

Errata to these supplementary notes have been given effect to and can be disregarded. Errata to the original notes are given on pages 3, 6, and 49. The latter errata have not been given effect to.

Supplementary Notes for Economics 313

Economics 313, October 8, 1948
Lecture 3, Supplement

PAGE 15:
Insert at the conclusion of the proof of Theorem (52) the

Corollary: If the matrix B has a right inverse D (i.e. a solution of $BD=I$), then D is also a left inverse (i.e. a solution of $DB=I$); and conversely.

Proof: Multiplying

(S-1) $BD=I$

through on the left by D, we have

(S-2) $DBD=D$.

Theorem (52) implies that D has a right inverse, say D^{-1}. Multiplying (S-2) through on the right by D^{-1}, we have

(S-3) $DBDD^{-1} = DD^{-1}$, or $DB=I$,

which completes the proof of the first part of the theorem. The proof of the converse is similar.

PAGE 17:

Theorem

(66A) If B is non-singular in the system $B y' = 0$, then y' can only be the (0) vector.

PAGE 18:

Proof of Theorem (71): We write

(S-4) $\rho(A) = \rho$

and assume for the sake of argument that

(S-5) $\rho(A \quad u') > \rho$,

then there exists a non-singular square sub-matrix of $(A \quad u')$ of order $\rho+1$, which by permutation of rows and columns we can make to be

(S-6)
$$\begin{pmatrix} a_{11} & \cdots & a_{1\rho} & u_1 \\ \cdot & \cdots & \cdot & \cdot \\ a_{\rho 1} & \cdots & a_{\rho\rho} & u_\rho \\ a_{\rho+1,1} & \cdots & a_{\rho+1,\rho} & u_{\rho+1} \end{pmatrix} \equiv (A^* \; u'), \text{ say.}$$

This sub-matrix necessarily contains the u-column, because if it contained $\rho+1$ columns from A, (-4) could not be true. Furthermore

(S-7) $\rho(A^*) = \rho$

because otherwise the development of the determinant value of (S-6) according to the elements of the last column would show that (S-5) is not satisfied.

It follows that we can use the existence of solutions $x^{(j)}$ to the equations (87) and (89) occuring in the proof of Theorem (85), which we rewrite as

$$(S\text{-}8) \qquad a_{rj} = -\sum_{s=1}^{\rho} a_{rs} x_s^{(j)}, \qquad \begin{matrix} r=1,\ldots,m \\ j=\rho+1,\ldots,n \end{matrix}$$

where m is the number of rows of A, and n the number of columns. Writing out the equations (70) as

$$(S\text{-}9) \qquad \sum_{j=1}^{n} a_{rj}x_j = \sum_{s=1}^{\rho} a_{rs}x_s + \sum_{j=\rho+1}^{n} a_{rj}x_j = u_r, \quad r=1,\ldots,m$$

We substitute the right hand members from (S-8) for the a_{rj}, $j=\rho+1,\ldots,n$ in the second member of (S-9) to obtain

$$(S\text{-}10) \qquad \sum_{s=1}^{\rho} a_{rs}\xi_s = u_r, \qquad r=1,\ldots,m$$

where

$$(S\text{-}11) \qquad \xi_s = x_s - \sum_{j=\rho+1}^{n} x_s^{(j)} x_j, \qquad s=1,\ldots,\rho$$

Now (S-10) implies that the matrix (S-6) postmultiplied by the transpose of the non-vanishing vector

$$(\xi_1 \ldots \xi_\rho, \ -1)$$

becomes zero, which, because of Theorem (62), contradicts (S-5) It follows that (S-5) cannot be true.

Since the rank of a matrix can obviously not be decreased by the addition of a column, Theorem (71) follows.

Geometrically Theorem (71) means that if we consider the hyperplane of lowest dimension (=r) containing all vectors of A (represented by columns), evidently u', a linear combination of the vectors of A, is also contained in this hyperplane.

ERRATA pages 1-23:

page 3	5th line from bottom	replace "structure (5) of our" with "structure (5) within our".
page 10	line 1	delete all line 1, beginning "u =..." and replace with "$E[u^* v^*] = \frac{1}{1+\lambda} E[(u+\lambda v)v] = \frac{\lambda}{1+\lambda} E[v^2] \neq 0$ unless $\lambda = 0$"
page 10	4th & 5th lines from bottom	delete "each multiplied by an appropriate depreciation coefficient σ"
page 10	equation (30)	strike out "σ_1" and "σ_2"
page 11	equation (33)	add "+u" as the last term of the left hand member
page 11	equation (34)	add "+v" as the last term of the left hand member
page 11	equation (35)	add "+w" as the last term of the left hand member
page 23	line 1	change "rows" to"columns"
page 23	line 8	change "equation" to "equations"
page 23	line 14	should read "...it is of order $\varphi+1$, which is larger than φ, the..."
page 23	line 22	should read "...of the equations $A x' = 0$ is..."
page 23	equation (91)	should read "$X=(X_1^{(j)},\ldots,X^{(j)}, 0_{\varphi+1}, \ldots, 0_{j-1}, 1_j, 0_{j+1}, \ldots, 0_n)$"

Economics 313, October 21, 1948
Lecture 6, Supplement

PAGE 32:

Let us assume that the order conditions are satisfied, and that we are uncertain as to the rank condition because our restrictions on the other equations of the system are compatible with rank G-1 as well as with a lower rank of the criterion matrix in (137).

We may make the general statement that whenever the identifiability of any equation or any parameter depends on the values of unknown parameters, such identifiability is in principle subject to statistical test. The conclusiveness of the test is of course limited by the number of observations available.

A general argument which may be advanced to show that identifiability is subject to test is as follows:

In defining identifiability we took as given the distribution of observations. There are then two cases,

a) The distribution of observations will be such that all structures which could have generated it are alike in the parameters of the equation under investigation,

i.e. $f(y|z)$ is such that $S \rightarrow f$, $S^* \rightarrow f$,

with S and S^* inside our model, and

$$(\alpha_{11}, \ldots, \alpha_{1k}) = (\alpha^*_{11}, \ldots, \alpha^*_{1k}).$$

The first equation is then identifiable.

b) The distribution of observations is such that all the structures are not alike in the parameters of this equation,

i.e. $f(y|z)$ is such that there exists a pair in the model S and S^* such that $S \rightarrow f$, $S^* \rightarrow f$, whereas

$$(\alpha_{11}, \ldots, \alpha_{1k}) \neq (\alpha^*_{11}, \ldots, \alpha^*_{1k}).$$

Then the first equation is not identifiable.

We see from this that whether or not the equation is identifiable is a property of the distribution of the observations. We therefore have a dichotomy of all distribution functions $f(y|z)$ -- i.e. of all such functions generated by structures in the model-- and we may set up the null hypothesis that f is in one class to be tested against the alternative that f is in the other class.

The same conclusion can be arrived at by a less general argument, as follows:

Let us add a row of zeros to the matrix whose rank is in question,

$$\begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ \alpha_{2k_1} & \cdot & \cdot & \cdot & \alpha_{2k_r} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \alpha_{Gk_1} & \cdot & \cdot & \cdot & \alpha_{Gk_r} \end{pmatrix}$$

We will refer to this matrix as the criterion matrix $\alpha^{(1)}$ of the first equation.

We know that whenever S and S^* are equivalent, there exists Υ such that $\alpha^* = \Upsilon\alpha$, that is, the only changes in the coefficient matrix that preserve the function f are those obtained by non-singular linear transformations. In forming the j^{th} column of α^*, only the j^{th} column enters from the α matrix.

Therefore $\alpha^* = \Upsilon\alpha$ implies $\alpha^{(1)*} = \Upsilon\alpha^{(1)}$, for the columns of $\alpha^{(1)*}$ are related to the columns of α in the manner just specified. We know that the rank of $\alpha^{(1)}$ is not changed by the addition of a row of zeros, and we state without proof that the rank of a matrix is not changed when it is pre-multiplied by a non-singular matrix,

i.e. $\rho(\alpha^{(1)}) = \rho(\alpha^{(1)*})$,

and since the structure S^* represented by α^* is required to be in the model, then the elements of the first row of $\alpha^{(1)*}$ will also be zeros.

We may therefore conclude that between equivalent structures the rank of the criterion matrix for the identifiability of a given equation is always the same.

This rank, we conclude, depends only on the distribution of the observations; we may therefore regard the rank of the criterion matrix to be itself an identifiable parameter. It is therefore a matter of statistical test whether f(y|z) is such that the rank of the criterion matrix is $\geq G-1$.

Conclusion: Though we may be in doubt about the identifiability of a given parameter or equation because of the smallness of the sample, it remains true that a sufficiently large sample does convey some statistical information regarding the identifiability of that parameter or equation.

ERRATA:

page 18	theorem (71)	strike out "non-vanishing"
page 21	line 6	should read "...column tacked onto β), then according to Theorem 85 there"
page 31	line following equation (138)	should read "but where $\alpha_{ik} \neq \alpha^*_{ik}$ for at least one value of k"
page 31	3rd line from bottom	change "should" to "may"

Economics 313, October 29, 1948
Lecture 6, Supplement

General Observations on the Identification Problem.

In factor analysis in modern psychology, limitations on experimentation give rise to a problem of identification also. Theory suggests that the score of an individual on a given test depends on the extent to which that individual posseses certain abilities or mental factors required in the performance of that test. A model is then postulated in which the score of an individual on a test (i.e. the dependent variable) is approximately equal to a linear combination of the individual's abilities (regarded as a number for each mental factor) which are required for the performance of that test and of a random influence in that individual's score on the test (i.e. the random variable). The model thus specifies the system:

$$s_{kn} = \sum_{l=1}^{L} \alpha_{kl} \gamma_{ln} + u_{kn} \quad (k=1,\ldots,K;\ n=1,\ldots,N)$$

or in matrix notation,

$$\begin{pmatrix} s_{11} & \cdots & s_{1N} \\ \vdots & & \vdots \\ s_{K1} & \cdots & s_{KN} \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1L} \\ \vdots & & \vdots \\ \alpha_{K1} & \cdots & \alpha_{KL} \end{pmatrix} \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1N} \\ \vdots & & \vdots \\ \gamma_{L1} & \cdots & \gamma_{LN} \end{pmatrix} + \begin{pmatrix} u_{11} & \cdots & u_{1N} \\ \vdots & & \vdots \\ u_{K1} & \cdots & u_{KN} \end{pmatrix}$$

where s_{kn} is the score of the k^{th} individual on the n^{th} test, α_{kl} is the "amount" of ability l present in the k^{th} individual, γ_{ln} is a weight representing the "importance" of the l^{th} ability in the performance of the n^{th} test, and u_{kn} is a random influence in the score of the k^{th} individual on the n^{th} test.

The s_{kn} are observables; the α_{kl}, γ_{ln}, and u_{kn} are not. The α_{kl} & γ_{ln} are unknown structural parameters. Knowledge of the structure

would permit prediction of the scores. The problems of identifying the structure arise from the difficulty of devising tests requiring as few abilities as possible and of making different tests depend on different sets of abilities, i.e. of arranging for a suitable sprinkling of zeros in the matrix A to render its elements identifiable. It is thus the familiar problem of finding the unique structure that generates the observed distribution of dependent variables, namely $f(s_{kn})$.

(For a discussion of the identification problem in factor analysis, see the unpublished study by Reiersøl, Cowles Commission Discussion Paper 303).

Lecture 8, Supplement

The Situation of Changed Structure.

(Reading assignment: Tinbergen, The Statistical Testing of Business Cycle Theories, Volume II, Sections 6.8 and 6.9)

In Section 6.9, Tinbergen cites as examples of changes of structure: 1) Situations in which a change of policy replaces an entire equation by the statement that the variable previously "explained" by that model is henceforth a constant, e.g. government fixes total investment by varying government expenditures; and 2) Situations in which the value of a parameter is changed, e.g. a change in the marginal propensity to consume as a result of social security legislation. The problem requiring investigation in such cases is whether or not any other equations are simultaneously affected by such a policy.

Economics 313, November 4, 1948
Lectures 9 and 10, Supplement

Validity of the Least Squares Method: Case 1, Continued.

Suppose, instead of choosing to estimate parameters of the reduced form (191)-(192), we pick q as the "dependent" variable and p, I as the "independent" variables in the attempt to estimate structural parameters of the demand equation (189), say, by least squares. Clearly least-squares estimates in this case will be suspect, for the consistency of these estimates, in the well-known theory of least-squares regression, is based on the condition that the variables chosen as the "independent" ones can be assumed to retain a fixed value in repeated samples. But in (192) p is a linear function of v_2, which is a linear function of u_1 and u_2. This establishes a stochastic relationship between the "independent" variable p, and the random disturbance u_1, contrary to the assumptions underlying least squares theory.

Basically, least squares estimation assumes stochastic independence between the random variable and the "independent" (in the meaning of the preceding paragraph) variables, i.e. exogenous variables. It can be proved explicitly that, except for special values of α or γ, least-squares estimation of any of the structural equations (189)-(190), for any arbitrary choice of "Dependent" variable, leads to estimates which are even asymptotically subject to a finite bias. We shall not prove this statement in general, or with respect to the model (189)-(190), but with respect to a still simpler model discussed by Haavelmo.

Haavelmo's First Example.

We shall show more explicitly the derivation of the Plim in (H2.19):

$$(S\text{-}12) \qquad a = \frac{m_{cy}}{m_{yy}} = 1 - \frac{m_{zy}}{m_{yy}} \qquad \text{from (H2.1)}$$

$$(S\text{-}13) \qquad y - \bar{y} = \frac{1}{1-\alpha}(z - \bar{z}) + \frac{1}{1-\alpha}(u - \bar{u}) \qquad \text{from (H2.7)}$$

Multiplying (S-2) by $(z - \bar{z})$, and summing over t, and dividing by N, we have

$$(S\text{-}14) \qquad m_{zy} = \frac{1}{1-\alpha}(m_{zz} + m_{uz})$$

Squaring each side of (S-2) and summing, we have

$$(S\text{-}15) \qquad m_{yy} = \frac{1}{(1-\alpha)^2}\left[m_{zz} + 2m_{uz} + m_{uu}\right]$$

Substituting (S-3) and (S-4) in (S-1) we obtain

$$(S\text{-}16) \qquad a = 1 - \frac{(1-\alpha)(m_{zz} + m_{uz})}{m_{zz} + 2m_{uz} + m_{uu}} = \frac{\alpha m_{zz} + (1+\alpha)m_{uz} + m_{uu}}{m_{zz} + 2m_{uz} + m_{uu}}$$

which is (H2.18).

Consider the expression

$$(S\text{-}17) \qquad m_{zu} \equiv \frac{1}{N} \sum_{t=1}^{N} \left[z(t) - \bar{z} \right] u(t) \equiv \sum_{t=1}^{N} \lambda(t)\, u(t),$$

$$\text{i.e.} \qquad \lambda(t) \equiv \frac{z(t) - \bar{z}}{T}$$

with expected value zero and variance

$$(S\text{-}18) \qquad E\left\{ \left[\sum_t \lambda(t)\, u(t) \right]^2 \right\} = E\left\{ \sum_t \sum_{t'} \lambda(t)\lambda(t')\, u(t)\, u(t') \right\}$$

$$= \sigma_u^2 \sum_t \lambda^2(t)$$

$$= \sigma_u^2 \frac{m_{zz}}{N}$$

Since $\lim_{N \to \infty} m_{zz} = \bar{m}_{zz}$ by assumption,

$$(S\text{-}19) \qquad \lim_{N \to \infty} \sigma_u^2 \frac{m_{zz}}{N} = 0 .$$

Therefore,

$$(S\text{-}20) \qquad \operatorname*{Plim}_{N \to \infty} m_{uz} = \operatorname*{Plim}_{N \to \infty} \left\{ \frac{1}{N} \sum_t \left[z(t) - \bar{z} \right] u(t) \right\} = 0 .$$

And finally

$$(S\text{-}21) \qquad \operatorname*{Plim}_{N \to \infty} a = \frac{\alpha\, \bar{m}_{zz} + \sigma_u^2}{\bar{m}_{zz} + \sigma_u^2} = \frac{\alpha + \dfrac{\sigma_u^2}{\bar{m}_{zz}}}{1 + \dfrac{\sigma_u^2}{\bar{m}_{zz}}} > \alpha$$

Economics 313, November 9, 1948
Lecture 11, Supplement

We have recognized that (208) is a necessary and sufficient condition for the identifiability of the equation (198) we are concerned with. This in itself is proof of the equivalence of (208) with the earlier rank conditions for identifiability (137), in terms of co-efficients of the structural equations, as applied to equation (198). A direct proof is as follows.

The rank of a matrix A equals the number of columns less the number of linearly independent solutions x' of

(S-22) $$A\,x' = 0$$

(This follows from the proof of theorem (85)). Since solutions of (s-22) are solutions of

(S-23) $$\Upsilon A\,x' = 0$$

and conversely provided Υ is non-singular, the rank of a matrix is not affected by pre-multiplication by a non-singular matrix. A similar argument applies to post-multiplication by a non-singular matrix.

Let $\tilde{\alpha}$ be a matrix consisting of those columns of α for which (198) prescribes zeros. Then $\tilde{\alpha}$ is obtained from the rank criterion matrix by adding a row of zeros, which does not affect rank.

We supplement (199) with the remaining equations of the reduced form, denoting by y^{Δ} the vector of dependent variables occuring in (198), and by $y^{\Delta\Delta}$ the vector of the remaining dependent variables, as follows,

(S-24)
$$\begin{aligned} y' \quad + \quad 0 \quad - {}^{\Delta}\Pi^{*}\, z'^{*} \quad - {}^{\Delta}\Pi^{**}\, z'^{**} &= v'^{\Delta}\ (198) \\ 0 \quad + \quad y'^{\Delta\Delta} - {}^{\Delta\Delta}\Pi^{**}\, z'^{*} \quad - {}^{\Delta\Delta}\Pi^{**}\, z'^{**} &= v'^{\Delta\Delta} \end{aligned}$$

From the definition of the reduced form

(S-25) $$(I - \Pi) = \beta^{-1}\tilde{\alpha}$$

and, selecting the columns from excluded from (198),

(S-26) $$\begin{pmatrix} {}^{\Delta}0^{\Delta\Delta} & -{}^{\Delta}\Pi^{**} \\ {}^{\Delta\Delta}I^{\Delta\Delta} & -{}^{\Delta\Delta}\Pi^{**} \end{pmatrix} = \beta^{-1}\tilde{\alpha}$$

where the superscripts to the zero and unit matrices serve to indicate the number of rows and columns.

Postmultiplication with the non-singular matrix

$$\begin{pmatrix} {}^{\Delta\Delta}I^{\Delta\Delta} & -{}^{\Delta\Delta}\Pi^{**} \\ {}^{\Delta\Delta}0^{**} & {}^{**}I^{**} \end{pmatrix} \tag{S-27}$$

turns the lefthand member of (S-26) into

$$\begin{pmatrix} {}^{\Delta}0^{\Delta\Delta} & -{}^{\Delta}\Pi^{**} \\ {}^{\Delta\Delta}I^{\Delta\Delta} & {}^{\Delta\Delta}0^{**} \end{pmatrix} \tag{S-28}$$

It follows that

(S-29) (S-28) has the same rank as $\tilde{\alpha}$.

the order of ${}^{\Delta\Delta}I^{\Delta\Delta}$ is G-H. We shall prove

$$\rho({}^{\Delta}\Pi^{**}) = \rho(\tilde{\alpha}) - (G-H). \tag{S-30}$$

Suppose first that in (S-30) a $\leq$ sign applies. Then we could select from ${}^{\Delta}\Pi^{**}$ a non-singular submatrix $\tilde{\Pi}$ of order $\rho({}^{\Delta}\Pi^{**})$ and combine it with ${}^{\Delta\Delta}I^{\Delta\Delta}$ to the following non-singular submatrix

$$\begin{pmatrix} 0 & \tilde{\Pi} \\ {}^{\Delta\Delta}I^{\Delta\Delta} & 0 \end{pmatrix} \tag{S-31}$$

of (S-28) with order $\rho({}^{\Delta}\Pi^{**}) + (G-H) > \rho(\tilde{\alpha})$, which contradicts (S-29). Suppose next that in (S-30) a $<$ sign applies. Then select from (S-28) a nonsingular submatrix of order $\rho(\tilde{\alpha})$, which necessarily takes the form

$$\begin{pmatrix} 0 & \Pi^{\dagger} \\ I^{\dagger} & 0 \end{pmatrix}$$

with $I^{\dagger}$ of order at most G-H, and hence with the nonsingular $\Pi^{\dagger}$ of order at least $\rho(\tilde{\alpha}) - G-H$, which contradicts the assumed inequalities negating (S-30). This completes the proof of (S-30). Since the "old" identifiability criterion was

$$\rho(\tilde{\alpha}) = G - 1$$

the "new" criterion

$$\rho({}^{\Delta}\Pi^{**}) = H - 1$$

is through (S-30) equivalent to it.

Economics 313, November 9, 1948
Lectures 11-15, Supplement

Remarks on transformation of variables in a density function.

References: Arrow, Lecture Notes for Economics 312M, pages 65-74.
Wilkes, Mathematical Statistics, pages 23-29.

Suppose we have a random variable x with a probability density function $\phi(x)$ such that (up to first order terms in ΔX)

$$P(X \leq x \leq X+\Delta X) = \phi(X)\Delta X, \tag{S-32}$$

and a random variable y such that

$$x = h(y), \tag{S-33}$$

where x is a continuous, monotonically increasing function of y. We have similarly

$$P(Y \leq y \leq Y+\Delta Y) = \psi(Y)\Delta Y \quad . \tag{S-34}$$

Then $\phi(x)$ and $\psi(y)$ can be related through

$$P(X \leq x \leq X+\Delta X) = P(Y \leq y \leq Y+\Delta Y). \tag{S-35}$$

From (S-32), (S-34) and (S-35), we have

$$\phi(X)\Delta X = \psi(Y)\Delta Y. \tag{S-36}$$

We observe that in the limit $\frac{\Delta X}{\Delta Y}$ becomes $\frac{dX}{dY}$; and thus, returning to small letters,

$$\psi(y) = \phi(x)\frac{dx}{dy} \quad . \tag{S-37}$$

If x is a monotonically decreasing function of y, we need only observe the precaution of using the absolute value of $\frac{dx}{dy}$ to preserve the equality in (S-37).

In a two dimensional space where we define

$$\begin{aligned} x_1 &= h_1(y_1, y_2) \\ x_2 &= h_2(y_1, y_2), \end{aligned} \tag{S-38}$$

we have

$$P(X_1, X_2 \leq x_1, x_2 \leq X_1+\Delta X_1, X_2+\Delta X_2) = \phi(X_1, X_2)dX_1 dX_2 \tag{S-39}$$

where again to express ψ in terms of ϕ we must say that for the small area $\Delta y_1 \Delta y_2$ in the y-space there corresponds an (in the small) proportional but not necessarily equal area in the x-space.

The factor of proportionality (in the limit for infinitely small areas is given by

$$(S\text{-}40) \qquad |J| = \left| \det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} \end{pmatrix} \right| .$$

This determinant is known as the Jacobian, abbreviated by J. Hence,

$$(S\text{-}41) \qquad \Psi(y_1, y_2) = \phi(x_1, x_2)|J| .$$

In the general case where

$$(S\text{-}42) \qquad \begin{aligned} x_1 &= h_1(y_1; \cdots, y_N) \\ &\vdots \quad \vdots \\ x_N &= h_N(y_1, \cdots, y_N) \end{aligned} ,$$

we have $\Psi(y_1, \cdots, y_N) = \phi(x_1, \cdots, x_N)|J|$, where

$$(S\text{-}43) \qquad |J| = \left| \det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \cdots\cdots & \frac{\partial x_N}{\partial y_1} \\ \vdots & & \vdots \\ \frac{\partial x_1}{\partial y_N} & \cdots\cdots & \frac{\partial x_N}{\partial y_N} \end{pmatrix} \right| ,$$

and provided J does not vanish.

If $h_1, \cdots, h_N$ are linear functions

$$(S\text{-}44) \qquad x_n = \sum_{m=1}^{N} \lambda_{nm} y_m$$

the Jacobian is simply

$$(S\text{-}45) \qquad J = \det \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{N1} \\ \vdots & & \\ \lambda_{1N} & \cdots & \lambda_{NN} \end{pmatrix} = \det \mathcal{L} .$$

Since we require J to be non-vanishing, the transformation inverse to (S-42) exists. If we introduce the notation

$$(S\text{-}46) \qquad J = J_{xy} \equiv \frac{\partial(x_1, \cdots, x_N)}{\partial(y_1, \cdots, y_N)} , \quad J_{yx} = \frac{\partial(y_1, \cdots, y_N)}{\partial(x_1, \cdots, x_N)}$$

we necessarily have $J_{xy}J_{yx}=1$, as is seen by successively applying (S-42) and its inverse. The net effect is no change in $\phi(x_1, \ldots, x_k)$. Hence

$$(S\text{-}47) \qquad J_{yx}=\frac{1}{J_{xy}} \quad .$$

Derivation of the Multi-variate Normal Distribution.

Suppose that we have G independent variables w_i with distribution N(0,1). Then

$$(S\text{-}48) \qquad E(w_i)=0, \quad E(w_iw_j) \begin{cases} = 0 & \text{for } i\neq j \\ = 1 & \text{for } i=j \end{cases} .$$

The joint probability density function of the G variables is

$$(S\text{-}49) \qquad \Omega(w_1,\ldots,w_G) = \left(\frac{1}{\sqrt{2\pi}}\right)^G e^{-\frac{1}{2}\sum_{i=1}^{G} w_i^2}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^G e^{-\frac{1}{2}ww'} ,$$

writing $w = (w_1, \ldots, w_G)$.

If we define G variables u_i by

$$(S\text{-}50) \qquad u_g = \sum_{h=1}^{G} \lambda_{gh}w_h \quad , \text{ or simply } u' = \Lambda w'$$

where Λ is non-singular, we obtain the joint probability density of the u's from that of the w's by the transformation $w' = \Lambda^{-1} u'$,

$$(S\text{-}51) \qquad \phi(u_1,\ldots,u_G) = \left(\frac{1}{\sqrt{2\pi}}\right)^G e^{-\frac{1}{2}u\Lambda^{-1}\Lambda^{-1}u'} \cdot \left|\det \Lambda^{-1}\right| .$$

The last factor is the right-hand member ofis the Jacobian with of the transformation. We shall show that the co-efficient matrix of the quadratic form in the exponent of e and the Jacobian are simple functions of the moment matrix of the u's.

Since $E(u_i) = 0$ for all i, we have from (S-38)

$$\sigma_{u_i u_j} = E(u_i u_j) = E\left[\left(\sum \lambda_{ig} w_g\right)\left(\sum \lambda_{jh} w_j\right)\right]$$

$$= \sum_{g,h} \lambda_{ig}\, \lambda_{jh}\, E(w_g w_h)$$

$$= \sum_{g} \lambda_{ig}\, \lambda_{jg} \qquad \begin{matrix}(i = 1,\ldots,G)\\(j = 1,\ldots,G)\end{matrix} .$$

Let us write for the moment matrix of the u's $\Sigma = (\sigma_{u_i u_j})$. Then we have found that

$$\Sigma = \Lambda \Lambda' \qquad (S\text{-}52)$$

which is obviously symmetric ($\Sigma = \Sigma'$). We then have

$$\det \Sigma = \det \Lambda \det \Lambda' = \det^2 \Lambda$$

$$\det \Lambda^{-1} = \det^{-1} \Lambda = \det^{-\frac{1}{2}} \Sigma$$

and $$\Sigma^{-1} = \Lambda'^{-1} \Lambda^{-1}$$

Thus we may rewrite (S-51) as

$$\phi(u_1,\ldots,u_G) = (2\pi)^{-\frac{G}{2}} \cdot e^{-\frac{1}{2} u \Sigma^{-1} u'} \cdot \det^{-\frac{1}{2}} \Sigma \qquad (S\text{-}53)$$

(There is no page 17. Proceed to page 18.)

For further discussion on the multi-variate normal distribution, see:

Arrow, Lecture Notes for Economics 312M, page 51-55

Wilks, Mathematical Statistics, pages 63-69

Derivation of the Likelihood Function of the Sample.

In the preceding supplement we defined the moment matrix of the u's as

$$\Sigma = (\sigma_{u_i u_j}) \equiv E(u'u)$$

and found

$$\Sigma = \Lambda\Lambda' \quad .$$

Σ is the expectation of a matrix of rank 1, but if we require that the u's are linearly independent functions of the w's, i.e. that Λ is nonsingular, then the rank[1] of Σ is G.

Recalling that $u(t) \equiv \{u_1(t),\ldots,u_G(t)\}$ has the joint density function in (S-53) and that $\phi[u(t)|t] = \phi(u)$, we have for the joint density function of $\underline{u} \equiv \{u(1)\ldots u(t)\ldots u(T)\}$

$$(S\text{-}54)\quad \underline{\Phi}(\underline{u}) = \prod_{t=1}^{T} (2\pi)^{-\frac{1}{2}G} \cdot \det^{-\frac{1}{2}}\Sigma \cdot \exp\left\{-\tfrac{1}{2}u(t)\Sigma^{-1}u'(t)\right\}$$

$$= (2\pi)^{-\frac{1}{2}GT} \cdot \det^{-\frac{T}{2}}\Sigma \cdot \exp\left\{-\frac{1}{2}\sum_{t=1}^{T} u(t)\,\Sigma^{-1}u'(t)\right\}$$

for which the corresponding volume element is $d\underline{u} \equiv du_1(1)\ldots du_G(T)$.

We shall use the transformations defined (76) to derive $F(\underline{y}|\underline{z},\underline{y},\underline{z})$ from $\underline{\Phi}(\underline{u})$

where $\underline{y} \equiv \{y(1)\ldots y(T)\} \equiv \{y_1(1)\ldots y_G(T)\}$

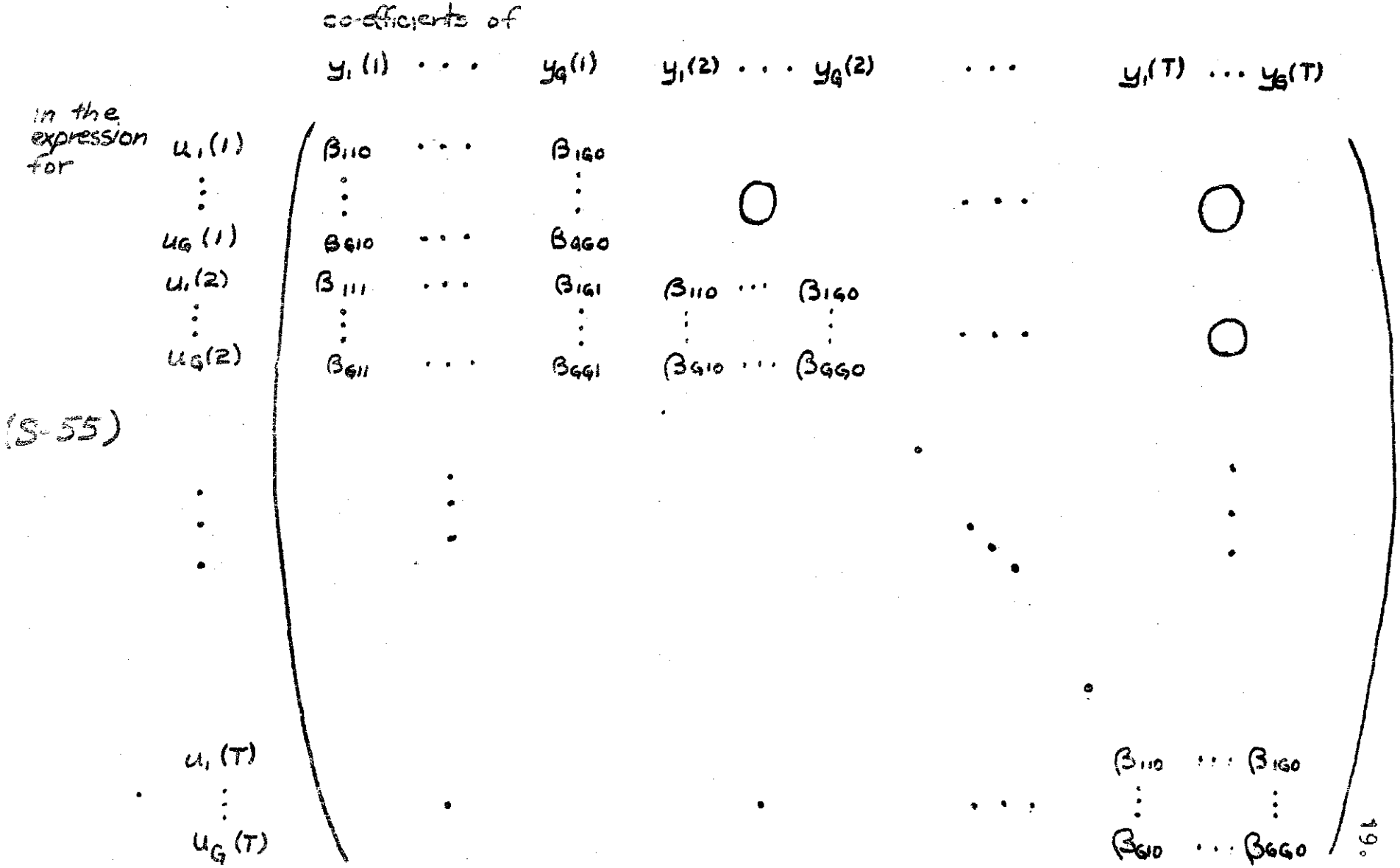
$\underline{z} \equiv \{z(1)\ldots z(T)\} \equiv \{z_1(1)\ldots z_K(T)\}$

$\underline{\underline{y}} \equiv y(1-\tau^{\alpha})\ldots y(0) \equiv$ etc.

$\underline{\underline{z}} \equiv z(1-\tau^{\beta})\ldots z(0) \equiv$ etc.

First we will indicate more explicitly the coefficients of the variables $y_G(t)$ in the system (76). These can be arranged in the following matrix put together from the matrices $\beta_0, \beta_1, \ldots \beta_{\tau}$ noted in (82): (see following page)

[1] H. Cramer: Mathematical Methods of Statistics §22.5


co-efficients of
$y_1(1) \cdots y_G(1)$ $y_1(2) \cdots y_G(2)$ $\cdots$ $y_1(T) \cdots y_G(T)$
in the expression for
$u_1(1)$
$u_G(1)$
$u_1(2)$
$u_G(2)$
$u_1(T)$
$u_G(T)$
$B_{110} \cdots B_{1G0}$
$B_{G10} \cdots B_{GG0}$
$B_{111} \cdots B_{1G1}$
$B_{G11} \cdots B_{GG1}$
$B_{110} \cdots B_{1G0}$
$B_{G10} \cdots B_{GG0}$
$B_{110} \cdots B_{1G0}$
$B_{G10} \cdots B_{GG0}$

(S-55)

Or more succinctly:

$$
\begin{array}{c} \\ u(1) \\ u(2) \\ u(3) \\ \vdots \\ u(\tau^0+1) \\ \vdots \\ u(T) \end{array}
\begin{array}{c}
\begin{array}{cccccc} y(1) & y(2) & y(3) & \cdots & y(T-\tau^0) & \cdots & y(T) \end{array} \\
\begin{pmatrix}
\beta_0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\
\beta_1 & \beta_0 & 0 & \cdots & 0 & \cdots & 0 \\
\beta_2 & \beta_1 & \beta_0 & \cdots & 0 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & & \vdots \\
\beta_{\tau^0} & \ddots & \ddots & \ddots & \ddots & & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & & \vdots \\
0 & 0 & 0 & \cdots & \beta_{\tau^0} & \cdots & \beta_0
\end{pmatrix}
\end{array}
\qquad (S\text{-}55A)
$$

We do not need the coefficient matrices of the $z(t)$, which are all exogenous, nor for the $y(t-\tau^0)$ $(t=0,1,\ldots,\tau^0)$ in order to determine the Jacobian of the transformations (76). Our reasoning in regard to the $y(t)$ for time points prior to $t=1,\ldots,T$ is similar to that with regard to the $z(t)$: We will only consider repeated samples arising from other drawings $u(t)$ over the time-points $t=1,\ldots,T$, and thus the values of the endogenous variables $y(t)$ for $t=0$ should be held constant in these repeated samples. This in no way alters or restricts the joint distribution of the disturbances $u(t)$ $t=1,\ldots,T$; for our assumption of independence between the $u(t)$ and $u(t-\tau)$ and between the $u(t)$ and the $z(t)$ implies

$$\Phi(\underline{u} \mid \underline{z}, \underline{y}, \underline{z}) = \Phi(\underline{u}) .$$

The Jacobian of the transformation from u to y is thus the determinant of the matrix in (S-55), which we may denote $\underline{\beta}(T)$. Using the Laplace expansion by blocks on the first G rows (the first row of (S-55A)), we obtain $\det \underline{\beta}(1) = \det \beta_0 \cdot \det \underline{\beta}(T-1)$. Applying the same procedure to $\underline{\beta}(T-1)$, etc., and writing $\beta_0 = \beta$ as in (83), we have

$$J = \det \underline{\beta}(T) = \det^T \beta_0$$

and hence

$$(S\text{-}56) \quad F(\underline{y} \mid \underline{z}, \underline{y}, \underline{z}) = (2\pi)^{-\frac{1}{2}GT} \cdot \left|\det^T \beta\right| \cdot \det^{-\frac{1}{2}T} \Sigma \cdot$$

$$\exp\left\{-\frac{T}{2} \sum_{t=1}^{T} x(t) A' \Sigma^{-1} A x'(t)\right\}$$

where

$$\alpha \equiv (\beta_0, \beta_1, \ldots \beta_{\tau^0}, \Gamma_0 \ldots \Gamma_{\tau^0}) \equiv (\beta\Gamma),$$

a G by $(G + K)(\tau^0 + 1)$ matrix

and where

$$x(t) \equiv y(t)\, y(t-1) \ldots y(t-\tau^0)\, z(t) \ldots z(t-\tau^0)$$
$$\equiv y_1(t) \ldots y_G(t) \ldots y_1(t-\tau^0) \ldots y_G(t-\tau^0)\, z_1(t) \ldots z_K(t-\tau^0)$$

a row vector with $(G + K)(\tau^0 + 1)$ elements.

The volume element corresponding to F is

$$dy_1(1) \ldots dy_G(T).$$

Observe that the exponent in F is the quadratic form

$$(S\text{-}57) \quad -\frac{1}{2}\sum_{t=1}^{T}\sum_{\substack{i=1\\j=1}}^{G} x_i(t)\,\Theta_{ij}\,x_j(t)$$

where $\Theta \equiv (\Theta_{ij}) \equiv \alpha'\Sigma^{-1}\alpha$

Writing $\frac{1}{T}\sum_t x_i(t)x_j(t) = m_{ij}$ and $(m_{ij}) = M$, a symmetric matrix of order G, (S-57) becomes

$$(S\text{-}58) \quad -\frac{1}{2}T\sum_{i,j}\Theta_{ij}m_{ji} = -\frac{1}{2}T\sum_i\left[\sum_j \Theta_{ij}m_{ji}\right] = -\frac{1}{2}T\sum_i (\Theta M)_{ii}$$

where $(\Theta M)_{ii}$ is the i^{th} element of the main diagonal in the matrix ΘM. The sum of the diagonal elements of a matrix is known as the trace (abbreviated tr). We may rewrite (S-56) as

$$(S\text{-}59) \quad F = (2\pi)^{-\frac{1}{2}GT}\left|\det{}^{T}\beta\right|\det{}^{-\frac{T}{2}}\Sigma \exp -\frac{T}{2}\mathrm{tr}(\alpha'\Sigma^{-1}\alpha M)$$

Taking the log of F and multiplying by $\frac{1}{T}$, we have

$$(S\text{-}60) \quad \frac{1}{T}\log F \equiv L = -\frac{1}{2}G\log 2\pi + \log\left|\det\beta\right| - \frac{1}{2}\log\det\Sigma - \frac{1}{2}\mathrm{tr}(\alpha'\Sigma^{-1}\alpha M)$$

which we shall call the logarithmic likelihood function of the parameters given the sample.

It is seen that the moments m_{ij} are the only functions of the observations that enter into L. Inserting these sample moments in (S-60), we may determine maximum likelihood estimates of Σ and α as functions of M by maximizing L with respect to the former; such values of the parameters if true would maximize the probability of "Drawing" our sample.

Invariance of L for Linear Transformations.

As an exercise we shall show explicitly that the form of the likelihood function is preserved by linear transformation. For then we have

$$\alpha^* = \Upsilon\alpha, \text{ in particular } \beta^* = \Upsilon\beta,$$

and $$u'^* = \Upsilon u' ,$$

(where Υ non-singular). That this should be so follows from theorem (117), but it is instructive to see a direct proof.

Recalling that for one time-point t, we have

$$\Sigma \equiv E(u'u)$$

and thus

$$\Sigma^* \equiv E(u^{*\prime}u^*) = E(\Upsilon u'u\Upsilon')$$
$$= \Upsilon[E(u'u)]\Upsilon' = \Upsilon\Sigma\Upsilon'$$

Inserting these expressions in a function L^* formed analagously to (S-60) we have

$$(S\text{-}61)\quad L^* = -\tfrac{1}{2}G\log 2\pi + \log|\det\beta^*| - \tfrac{1}{2}\log\det\Sigma^* - \tfrac{1}{2}\mathrm{tr}(\alpha^{*\prime}\Sigma^{*-1}\alpha^*_M)$$
$$= -\tfrac{1}{2}G\log 2\pi + \log|\det\Upsilon\beta| - \tfrac{1}{2}\log\det(\Upsilon\Sigma\Upsilon') - \tfrac{1}{2}\mathrm{tr}\{\alpha'\Upsilon'(\Upsilon\Sigma\Upsilon')^{-1}\Upsilon\alpha_M\}$$
$$= -\tfrac{1}{2}G\log 2\pi + \log|\det\Upsilon| + \log|\det\beta| - \tfrac{1}{2}\log(\det\Upsilon\det\Sigma\det\Upsilon') - \tfrac{1}{2}\mathrm{tr}(\alpha'\Upsilon'\Upsilon^{-1}\Sigma^{-1}\Upsilon^{-1}\Upsilon\alpha_M)$$

Since $-\frac{1}{2}\log(\det\Upsilon\det\Sigma\det\Upsilon') = -\frac{1}{2}\log(\det^2\Upsilon\det\Sigma)$

$$= -\tfrac{1}{2}(2\log|\det\Upsilon| + \log\det\Sigma)$$
$$= -\log|\det\Upsilon| - \tfrac{1}{2}\log\det\Sigma ,$$

we have

$$L^* = -\tfrac{1}{2}G\log 2\pi + \log|\det\beta| - \tfrac{1}{2}\log\det\Sigma - \tfrac{1}{2}\mathrm{tr}(\alpha'\Sigma^{-1}\alpha_M) = L$$

q.e.d.

Economics 313, November 13, 1948
Lectures 11-15, Supplement

(Note: The Supplements to Lectures 11-15, of which this is the second, replace the material of pages 65-89 .)

Further Remarks on Identification.

The function L of (S-60) is seen to be a function of observations, i.e. M, and parameters, i.e. $\Theta = (\alpha, \Sigma)$. When we consider L as a likelihood function, L appears with observed values of F inserted; thus L becomes a function of parameters only, once the observations are given. If then the parameters of L are unique, i.e. there exists no $\Theta^* = (\alpha^*, \Sigma^*)$ such that

(S-62) $\quad L(\Theta, M) = L^*(\Theta^*, M)$, for all M,

we shall say the parameters are identifiable.

On the other hand, what happens if we attempt estimation by maximum likelihood of parameters which are actually not identifiable, i.e. are such that $T \neq I$ exists for which (S-62) is true? We need to distinguish two cases.

First assume that a $T \neq I$ making (S-62) true exists for any value of Θ satisfying the given restrictions. Then the attempt to maximize $L(\Theta, M)$ with respect to Θ cannot lead to a unique "estimate" $\hat{\Theta}(M)$, and lack of identifiability of Θ is detected in this way if it was not noticed earlier.

Secondly, however, there is the possibility that the restrictions on Θ are sufficient in number and variety to produce identifiability for almost all values of Θ, the exceptional values being those in which the appropriate rank criteria of identification are not met. In that case, if the true Θ happens to be one of these exceptional values, this need not show up in an M derived from a finite sample, because of sampling variation. This is therefore a deceptive case, which will only reveal itself through large estimated (as against infinitely large true) sampling variances of the seemingly unique "estimates" $\hat{\Theta}(M)$.

Computation of maximum Likelihood Estimates.

Obviously for a large system of equations the computation of maximum likelihood estimates will be long and tedious. We may in some cases sacrifice some of the desirable properties of such estimates in favor of estimates whose computation is less involved.

Before pursuing this further, let us consider special assumptions for a model which reduce substantially the computation necessary without sacrifice of desirable properties.

CASE I: The equation system may be split into two subsets.

The model is in this case assumed to specify that the joint distribution of disturbances occuring in the first subset is independent of that of the disturbances in the second subset, i.e.

$$(S\text{-}63)\qquad \Sigma = \begin{pmatrix} \sigma_{11} \cdots \sigma_{1G_I} & 0 \\ \vdots & \\ \sigma_{G_I,1} \cdots \sigma_{G_I,G_I} & \sigma_{G_I+1,G_I+1} \cdots \sigma_{G_I+1,G} \\ 0 & \vdots \\ & \sigma_{G,G_I+1} \cdots \sigma_{GG} \end{pmatrix} = \begin{pmatrix} \Sigma_{II} & 0 \\ 0 & \Sigma_{II\,II} \end{pmatrix}$$

and that the second subset has as many equations as dependent variables, i.e.

$$(S\text{-}64)\qquad \beta = \begin{pmatrix} \beta_{11} \cdots \beta_{1} & \cdots & \beta_{1G} \\ \vdots & & \vdots \\ \beta_{G_I,1} \cdots & \cdots & \beta_{G_IG} \\ 0 \cdots 0\ \beta_{G_I+1,G_I+1} & \cdots & \beta_{G_I+1,G} \\ \vdots \quad \vdots & & \vdots \\ 0 \cdots 0\ \beta_{G,G_I+1} & \cdots & \beta_{GG} \end{pmatrix} = \begin{pmatrix} \beta_{II} & \beta_{I\,II} \\ 0 & \beta_{II\,II} \end{pmatrix}$$

where $G = G_I + G_{II}$.

Our system thus looks as follows:

1st set: $\beta_{II}\, y_I' + \beta_{I\,II} y_{II}' + \Gamma_I z' = u_I'$

2nd set: $\beta_{II\,II} y_{II}' \quad \Gamma_{II} z_{II}' = u_{II}'$

where u_I' is independent of u_{II}' .

Since the second set is sufficient to determine y_{II}' , then in the first set y_{II}' may be regarded as "predetermined" variables "with respect to the first set" (not in a temporal sense, however).

From (S-63) we have

(S-65) $$\log \det \Sigma = \log \det \Sigma_{II} + \log \det \Sigma_{II\,II}$$

and

(S-66) $$\Sigma^{-1} = \begin{pmatrix} \Sigma_{II}^{-1} & O \\ O & \Sigma_{II\,II}^{-1} \end{pmatrix}$$

Writing

$$\alpha = \begin{pmatrix} \alpha_I \\ \alpha_{II} \end{pmatrix}$$

we have

(S-67) $$\alpha'\Sigma^{-1}\alpha = (\alpha_I' \alpha_{II}') \begin{pmatrix} \Sigma_{II}^{-1} & O \\ O & \Sigma_{II\,II}^{-1} \end{pmatrix} \begin{pmatrix} \alpha_I \\ \alpha_{II} \end{pmatrix}$$
$$= \alpha_I' \Sigma_{II}^{-1} \alpha_I + \alpha_{II}' \Sigma_{II\,II}^{-1} \alpha_{II}$$

From (S-64) we have

(S-68) $$\log \left|\det \beta\right| = \log \left|\det \beta_{II}\right| + \log \left|\det \beta_{II\,II}\right|$$

Then using (S-65)-(S-68), we obtain from (S-60)

(S-69) $$L = L_I + L_{II}$$

where

$$L_I = \text{constant} + \log \left|\det \beta_{II}\right| - \tfrac{1}{2}\log \det \Sigma_{II} - \tfrac{1}{2}\operatorname{tr}(\alpha_I' \Sigma_{II}^{-1} \alpha_I M)$$
$$L_{II} = \text{constant} + \log \left|\det \beta_{II\,II}\right| - \tfrac{1}{2}\log \det \Sigma_{II\,II} - \tfrac{1}{2}\operatorname{tr}(\alpha_{II}' \Sigma_{II\,II}^{-1} \alpha_{II} M) .$$

It will be noticed that the parameters occuring in L_I are different from those in L_{II}. Unless the two sets of parameters are linked by restrictions, therefore, we may maximize L by independently maximizing L_I and L_{II} with respect to the parameters occuring in them, and thus obtain maximum likelihood estimates of those parameters.

CASE II: The Recursive Form.

Our assumptions are now that the model specifies

$$(S\text{-}70) \qquad \mathfrak{B} = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1G} \\ & \ddots & \vdots \\ 0 & & \beta_{GG} \end{pmatrix}, \ \mathfrak{B} \text{ non-singular}$$

$$(S\text{-}71) \qquad \Sigma = \begin{pmatrix} \sigma_{11} & & 0 \\ & \sigma_{22} & \\ & \ddots & \\ 0 & & \sigma_{GG} \end{pmatrix}$$

It may be left as an exercise to prove that these restrictions are sufficient for identification of all equations. If no further restrictions connect parameters of different equations, our system may be split into as many subsets as there are equations. Applying repeatedly the argument of Case I we find

$$(S\text{-}72) \qquad L = L_1 + L_2 + \cdots + L_G \ .$$

It follows that maximum likelihood estimation is in this case equivalent to the method of least squares if we choose as the dependent variable in each equation that variable whose coefficient appears on the main diagonal of $\mathfrak{B}$. The estimates obtained, being maximum likelihood estimates, are consistent and asymptotically efficient, but not necessarily unbiased in finite samples. For we may be violating the assumption that (in least squares terminology) "independent" variables can be held constant in repeated samples. While the variables $y_{g+1}, \ldots, y_G$ may be called predetermined with respect to the gth equation, they may depend on earlier values of the y's including y_g, and thus are not truly independent as required by least squares theory.

As an example of the recursive form we shall select a model that may have particular relevance to agricultural markets where supply for a period may be dependent only on prices of a preceding period.

Demand: $\alpha p + q + \xi = u$

Supply: $q + \gamma p_{-1} + \eta = v$

where $\Sigma = \begin{pmatrix} \sigma_{uu} & 0 \\ 0 & \sigma_{vv} \end{pmatrix}$ and $\beta = \begin{pmatrix} \alpha & 1 \\ 0 & 1 \end{pmatrix}$.

We must choose q as the "dependent" variable in the second equation and p in the first. Clearly we may not assume constancy of p_{-1} in repeated samples "over the period", and our least squares estimate, δ will consequently have a bias. This bias is of order $\frac{2}{T}$, which is not an insignificant magnitude for finite samples. See Hurwicz: "Least Squares Bias in Time Series" in Cowles Commission Monograph 10, or the finite sample bias of certain consistent and asymptotically efficient estimates.

Lecture 12, Supplement

Properties of Maximum Likelihood Estimates with Respect to Efficiency. (Page 64)

Wald has now published (Annals of Mathematical Statistics, March, 1948) a proof of the asymptotic efficiency of maximum likelihood estimates for the case of general stochastic processes depending on only one parameter. H. Rubin has given a proof for the case of stable linear stochastic difference equation systems with many parameters (Cowles Commission Discussion Paper 301).

Economics 313, November 23, 1948
Lectures 11-15, Supplement

ERRATA to Supplemental Notes:

page 11	2nd line of (S-24)	change "z" to "z'^*"
page 11	Equations (S-25) thru (S-28)	The minus signs between sub-matrices in these equations should be moved to the right so that they will be read as referring to the second sub-matrix, not as instructions to form the difference between two matrices
page 21	3rd line	Read "G by $(G+K)(\tau^n+1)$"
page 21	6th line	"$z_k(t-\tau^n)$" should read "$z_K(t-\tau^n)$"
page 21	7th line	"$(t-1)$" should read "(τ^n+1)"
page 21	7th line from bottom	Before "sample" insert "parameters given the"
page 21	2nd line from bottom	Change "estimates" to "values of the parameters if true would"
page 22	3rd line	Strike out the word "the"
page 22	13th line	Should read: $\Sigma^* = E(u^{*\prime}u^*) = E(\Gamma u'u\Gamma')$

CASE II: The Recursive Form (cont'd)

We will state without proof that any model by a suitable transformation of variables can be put into the Recursive Form in as many ways as we can order equations and dependent variables. This form, as we have observed, always has identifiable parameters. Now if the parameters of any equation of the Recursive Form are estimated by the method of maximum likelihood, the estimates obtained are at the same time least squares estimates if we choose for the dependent variable in each equation that variable whose coefficient appears on the main diagonal of Γ --- subject to the qualification that we either have no or use no restrictions by the model "in excess" of those specified by the Recursive Form itself.

Now the last equation of the Recursive Form (i.e. that equation having only one dependent variable in it) is identical with the last equation of the reduced form (that arranges dependent variables in the same order). This follows from the fact that these equations exclude the same set of G-1 dependent variables (all but the last) and that such exclusion is sufficient for its identification (check on the rank condition!). Thus maximum likelihood estimation of the last equation of the reduced form is equivalent to least squares estimation. Obviously, this conlcusion is not specific to the last equation, since the variables can be placed in any desired order.

We thus have obtained the important additional result that least squares estimates of the parameters of the reduced form are consistent and asymptotically efficient even when lagged endogenous variables appear among the predetermined variables in the reduced form (for we have made no restrictions against their appearance in the corresponding Recursive Form).

CASE III: One-Equation Systems

In such models we have

$$\Sigma = (\sigma^2), \beta = (-1), \text{ and } \Gamma' = (\gamma_1, \ldots, \gamma_k).$$

Since $\mathrm{tr}(\alpha'\Sigma^{-1}\alpha_M) = \mathrm{tr}(\Sigma^{-1}\alpha_M\alpha')$ (See (S-76) below) we have

$$(S\text{-}73) \qquad L_1 = \text{const} - \log\sigma - \frac{1}{2\sigma^2}\,\mathrm{tr}(-1, \gamma_1, \ldots, \gamma_K) M \begin{pmatrix} -1 \\ \gamma_1 \\ \vdots \\ \gamma_K \end{pmatrix}$$

$$= \text{const} - \log\sigma - \frac{1}{2\sigma^2} \sum_{\substack{k=0 \\ l=0}}^{K} \gamma_k m_{kl} \gamma_l$$

where $\gamma_0 = -1$. Raising e to the power L_1, we have

$$(S\text{-}74) \qquad f = \text{const}\ \frac{1}{\sigma} \exp\left\{ -\frac{1}{2\sigma^2} \sum \gamma_k m_{kl} \gamma_l \right\},$$

which is the likelihood function corresponding to the linear normal regression model of one variable on a set of other variables.

Note how the expression (S-73) is generalized in (S-60). The $\frac{1}{\sigma^2}$ becomes incorporated in the trace term as a factor Σ^{-1}. The $-\log\sigma$ becomes $-\frac{1}{2}\log\det\Sigma$. Finally, an additional term $\log|\det\beta|$ arises from the Jacobian of the transformation from the disturbances to the dependent variables.

Mathematical Digression.

The trace is only defined for a square matrix. Three obvious properties of the trace are as follows:

(S-75) $\mathrm{tr}X = \mathrm{tr}X'$;

if X is a p by n matrix and Y an n by p matrix,

(S-76) $\mathrm{tr}XY = \mathrm{tr}YX$

$$\text{since } \sum_{i=1}^{p}\sum_{j=1}^{n} x_{ij}y_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{p} y_{ji}x_{ij};$$

from (S-75) and (S-76), we have

(S-77) $\mathrm{tr}XY = \mathrm{tr}YX = \mathrm{tr}(YX)' = \mathrm{tr}X'Y'$.

Some less obvious properties of the trace are:

(S-78) $\mathrm{tr}X = \mathrm{tr}OXO'$ if O is orthogonal;

if X is real and symmetric then

(S-79) $\mathrm{tr}X = \sum \lambda_i$,

where the λ_i are the "characteristic roots" of X; since such matrices are always orthogonally similar to diagonal matrices with characteristic roots as elements, (S-79) follows as a corollary to (S-78). The latter property we shall not prove.

Differentiation of $\mathrm{tr}XY'$ with respect to a parameter:

Here we have X and Y two rectangular matrices with equal number of columns and equal number of rows. We shall employ various assumptions about the elements of X and Y to determine the derivative of $\mathrm{tr}XY'$ for such cases.

Case 1. Assume all elements of X but none of Y are functions of some parameter, say ξ. Then we have

(S-79) $$\frac{d}{d\xi}\mathrm{tr}XY' = \frac{d}{d\xi}\sum_{i,j} x_{ij}y_{ij} = \sum_{i,j}\left(\frac{dx_{ij}}{d\xi}\, y_{ij}\right)$$

$$= \mathrm{tr}\left(\frac{dX}{d\xi}\, Y\right)$$

where $\frac{dX}{d\xi} = \left(\frac{dx_{ij}}{d\xi}\right)$.

Case 2. Assume $\xi = x_{kl}$ and that the remaining elements of X and Y are independent of ξ.

(a) (S-79) becomes

$$(S\text{-}80) \qquad \frac{d}{dx_{kl}} \operatorname{tr} XY' = \sum_{i,j} \left(\frac{dx_{ij}}{dx_{kl}} y_{ij}\right) = y_{kl} \quad ,$$

since $\frac{dx_{ij}}{dx_{kl}} = \begin{cases} 0 \text{ for } (i,j) \neq (k,l) \\ 1 \text{ for } (i,j) = (k,l) \end{cases}$.

(b) Now assume $X = X'$. This makes $x_{lk} = x_{kl}$, but we assume that all other elements of X and all elements of Y are independent of x_{kl}. Then we have

$$(S\text{-}81) \qquad \frac{d}{dx_{kl}} \operatorname{tr} XY' = \begin{cases} y_{kl} + y_{lk}, & \text{if } k \neq l \\ y_{kk}, & \text{if } k = l \end{cases}$$

since $\frac{dx_{ij}}{dx_{kl}} = 0$ except when $(i,j) = (k,l)$ or $(j,i) = (k,l)$.

Differentiation of log det X with respect to a parameter:

Obviously we have X a square matrix, and we shall suppose $\det X > 0$. We shall employ the same assumptions of the above two cases, ignoring here the specifications concerning Y.

Case 1. We have

$$(S\text{-}82) \qquad \frac{d}{d\xi} \log \det X = \sum_{i,j} \frac{d \log \det X}{dx_{ij}} \frac{dx_{ij}}{d\xi}$$

$$= \sum_{i,j} \frac{\frac{d}{dx_{ij}} \det X}{\det X} \frac{dx_{ij}}{d\xi} \quad .$$

If we expand the i^{th} row of X by the method of LaPlace and write X_{ij} as the cofactor of x_{ij}, we get

$$(S\text{-}83) \qquad \det X = \sum_{k} x_{ik} X_{ik}$$

where all terms of the expansion except $x_{ij}X_{ij}$ are independent of x_{ij}, and, in the latter term, X_{ij} is independent of x_{ij}. Thus we have

$$(S\text{-}84) \qquad \frac{d}{dx_{ij}} \det X = X_{ij}$$

Substituting (S-84) in (S-82), we find

$$(S\text{-}85)\quad \frac{d}{d\xi}\log\det X = \sum_{i,j}\frac{X_{ij}}{\det X}\frac{dx_{ij}}{d\xi} = \sum_{i,j}x^{ji}\frac{dx_{ij}}{d\xi}$$

$$= \operatorname{tr}\left(X^{-1}\frac{dX}{d\xi}\right)$$

where x^{ij} is the i^{th} row, j^{th} column element of X^{-1} and where

$$\frac{X_{ij}}{\det X} = x^{ji} \quad .$$

Case 2. (a) Since

$$\frac{dx_{ij}}{dx_{kl}} = \begin{cases} 1 \text{ for } (i,j) = (k,l) \\ 0 \text{ for } (i,j) \neq (k,l) \end{cases}$$

the sum in (S-85) reduces to

$$(S\text{-}86)\quad \frac{d}{dx_{kl}}\log\det X = x^{lk} \quad .$$

(b) Since

$$\frac{dx_{ij}}{dx_{kl}} = \begin{cases} 1 \text{ for } (i,j) = (k,l) \text{ or } (j,i) = (k,l) \\ 0 \text{ otherwise} \end{cases}$$

the sum in (S-85) becomes

$$(S\text{-}87)\quad \frac{d}{dx_{kl}}\log\det X = x^{lk} + x^{kl} = 2x^{kl}, \text{ if } k\neq l$$
$$= x^{kk}, \text{ if } k=l \quad .$$

Computation of Maximum Likelihood Estimates.

We have mentioned previously in this connection the possibility of sacrificing some of the desirable properties of these estimates in favor of less involved computation. We might then choose between the following estimation "programs":

(1) Sacrifice nothing: estimate every equation of the system by maximizing the likelihood function with respect to every parameter.

(2) Less involved computation: estimate one equation at a time, ignoring each time the restrictions on all other equations, and pursue this procedure for as many equations as we wish to "estimate".

(3) Compromise: estimate simultaneously a subset of equations taking full acount of the restrictions on this subset but ignoring all those in the remainder of the system. If desired, this can be done again successively for different (non-overlapping) subsets.

Economics 313, November 30, 1948
Lectures 11-15, Supplement

A corollary to theorems on positive definite matrices (See page 105) :

A moment matrix Σ of linearly independent random variables u_g is positive definite.

If our variables u_i are linearly independent, then we have

(S-88) If $u\xi' = \xi u' = v$, then $E(v^2) > 0$,

where ξ is any vector but the zero-vector.

It follows that, for any $\xi \neq 0$,

(S89) $$0 < E(v^2) = E(\xi u' u \xi') = \xi E(u'u) \xi' = \xi \Sigma \xi$$

if

(S-90) $$\Sigma \equiv E(u'u)$$ Q.E.D

Partial Diagonalization of the moment matrix Σ:

Consider the positive definite moment matrix Σ (necessarily non-singular). Now every sub-matrix of Σ is non-singular, in particular $_{GG}\Sigma$ consisting of the first $G - 1$ rows and columns of Σ. Then we may find unique λ_i such that

(S-91) $$\sum_{i=1}^{G-1} \lambda_i \sigma_{ij} = \sigma_{iG} \qquad j=(1,\ldots,G) \quad \text{(by Theorem 66)}$$

and, because Σ is symmetric,

(S-92) $$\sum_{j=1}^{G-1} \lambda_j \sigma_{ji} = \sigma_{Gi} \qquad i=(1,\ldots,G) \quad .$$

Now subtract (S-91) from the G^{th} column of Σ and (S-92) from the G^{th} row. Then divide both the G^{th} row and the G^{th} column of the resulting matrix by the square root of the element appearing in the G^{th} row and G^{th} column of that matrix.

We obtain

$$(S\text{-}93) \qquad \Sigma_1 = \Lambda_1 \Sigma \Lambda_1' = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1,G-1} & 0 \\ \vdots & & \vdots & \vdots \\ \sigma_{G-1,1} & \cdots & \sigma_{G-1,G-1} & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

where pre- and post-multiplication by the matrix Λ_1, which can easily be evaluated explicitly (make this an exercise!), is equivalent to the linear operations on rows and columns described.

By reiteration of this process on

$$_{GG}\Sigma \equiv \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1,G-1} \\ \vdots & & \vdots \\ \sigma_{G-1,1} & \cdots & \sigma_{G-1,G-1} \end{pmatrix},$$

we can find G matrices Λ_i, $i=1,\ldots,G$, such that

$$(S\text{-}94) \qquad \Lambda_G \cdots \Lambda_1 \Sigma \Lambda_1' \cdots \Lambda_G' = \Lambda \Sigma \Lambda' = I .$$

However, we shall be interested in transformations Υ that reduce only a part of the rows and columns of Σ to diagonality, as follows:

$$(S\text{-}95) \qquad \Sigma^* = \Upsilon \Sigma \Upsilon' = \begin{pmatrix} \Sigma_{rr} & 0_{r\bar{r}} \\ 0_{\bar{r}r} & I_{\bar{r}\bar{r}} \end{pmatrix}$$

where Σ_{rr} is a square submatrix of Σ of order G_r and I is the identity matrix of order $G_{\bar{r}} = G - G_r$, and where every other element of Σ^* is zero.

(S-94) will be recognized as a special case of the more general theorem that any real symmetric matrix can, by real nonsingular transformation of the type considered, be reduced to diagonal form such that each diagonal element is either 1, 0 or -1. Positive definiteness rules out the possibility that the diagonal element is either 0 or -1.

Computation of Maximum Likelihood Estimates (Cont'd):

We shall now derive maximum likelihood estimates of the parameters of our likelihood function by the stepwise maximization procedure discussed on page 65.

We shall first consider the following subdivision of parameters:

parameters "η": Σ

"Θ": $\alpha \equiv (\beta \quad \Gamma)$.

By writing $\Psi \equiv \Sigma^{-1} = \Psi'$ and $U \equiv \alpha M \alpha' = U'$, we have from (S-60)

(S-96) $$L = \text{const} + \log|\det \beta| + \tfrac{1}{2}\log\det\Psi - \tfrac{1}{2}\operatorname{tr}\Psi U$$

Differentiating with respect to ψ_{ij}, we obtain

(S-97) $$\frac{\partial L}{\partial \psi_{ij}} = \tfrac{1}{2}(\psi^{ij} + \psi^{ji}) - \tfrac{1}{2}(u_{ij} + u_{ji})$$
$$= \psi^{ij} - u_{ij}$$
$$= \sigma_{ij} - u_{ij} \qquad \text{for } i \neq j$$

since $U = U'$ and $\Psi = \Psi'$, and

(S-98) $$\frac{\partial L}{\partial \psi_{ij}} = \tfrac{1}{2}(\sigma_{ii} - u_{ii}) \text{ for } i=j.$$

From $\frac{\partial L}{\partial \psi_{ij}} = 0$, $i,j = 1,\ldots,G$, we obtain the "conditional" maximum likelihood estimate $\hat{\Sigma} = \hat{\Sigma}(\alpha)$ of Σ for given values of α:

(S-99) $$\hat{\Sigma} = U \equiv \alpha M \alpha' .$$

This can be given three interpretations:

(a) If α is actually known, (S-99) gives the maximum likelihood estimate of Σ.

(b) If α is not known, (S-99) gives a conditional maximum likelihood estimate for presumed values of α. Furthermore,

(c) if α is not known, insertion of (s-99) into the likelihood function (s-60) gives a "reduced" likelihood function in terms of α alone, which can be used for further maximization with respect to α.

Writing out our results (S-99) more explicitly we have

$$\hat{\sigma}_{gh} = \sum_{k,l} \alpha_{gk} m_{kl} \alpha_{hl}$$

(S-100) $$= \frac{1}{T} \sum_t \left[\left(\sum_k \alpha_{gk} x_k(t) \right) \left(\sum_l \alpha_{hl} x_l(t) \right) \right]$$

$$= \frac{1}{T} \sum_t u_g(t)\, u_h(t) \quad ,$$

where

$$m_{kl} = \frac{1}{T}\sum_{t} x_k(t)\, x_l(t)$$

and $u_g(t) = \sum_{k} \alpha_{gh}\, x_k(t)$, etc.

Apparently (S-100) represents a simple generalization of "least "squares" results if we think of the $u_g(t)$ as the residuals obtained by inserting the observed variables $x_k(t)$ in the linear expression $\sum_{t} \alpha_{gk} x_k(t)$ with "presumed" co-efficient values α_{gk}. The generalization applies to each of the three cases (a), (b), (c) above, which can also be distinguished in least squares theory.

When the values of α's used are estimates of α's, then the $\hat{\sigma}_{ij}$ will in finite samples have a negative bias due to loss of degrees of freedom in the estimation of the α's. In both cases (a) and (c), however, the maximum likelihood estimates are consistent and asymptotically efficient.

Derivation of estimates under a "compromise" program.

Having obtained (S-99) (which has an interest in intself) from the "η-θ classification" of parameters indicated, we shall not now pursue the insertion of (S-99) in the likelihood function. Instead, we shall introduce another "η-θ classification", which is particularly economical for certain purposes. We shall further make use of the fact that, in the terminology of page 65, in stepwise maximization under program (3) of (Supplement) page 32, it is sufficient, if we wish to estimate η by maximum likelihood, to maximize $g^*(\eta, x)$, where g^* is obtained from $f^*(\eta, \theta^*, x)$ by inserting $\hat{\theta}^* = \hat{\theta}^*(\eta, x)$, and where θ^* may be such a transform of θ as to reduce the labor of mathematical derivation.

The "η-θ classification" is chosen on the basis of the following considerations. We shall retain all a priori restrictions on the subset of the first G_I equations with the coefficient matrix α_I in

$$\alpha = \begin{pmatrix} \alpha_I \\ \alpha_{II} \end{pmatrix}$$

Referring to the matrix Υ of (S-95), we shall choose its form to be:

$$\Upsilon = \begin{pmatrix} I_{II} & O_{III} \\ \Upsilon_{III} & \Upsilon_{III} \end{pmatrix}$$

We then have $\alpha^* = \Upsilon\alpha$, or

$$\alpha_I^* = I_{II}\alpha_I + 0\cdot\alpha_{II} = \alpha_I$$

$$\alpha_{II}^* = \Upsilon_{II,I}\alpha_I + \Upsilon_{II,II}\alpha_{II}$$

Now the matrices $\Upsilon_{II,I}$, $\Upsilon_{II,II}$ are to be, and according to (S-93) can be, chosen such that

$$\Sigma^* = \Upsilon\Sigma\Upsilon' = \begin{pmatrix} \Sigma_{II,I} & 0 \\ 0 & I_{II,II} \end{pmatrix}$$

We thus choose:

parameters "η": α_I, Σ_{II}

"θ": α_{II}, $\Sigma_{I,II}$, $\Sigma_{II,II}$ $\quad(\Sigma_{I,II} = \Sigma_{II,I})$

transformed parameters "θ^*": α_{II}^*

The choice of θ's is made because of our earlier decision, in the interest of computational simplification, both to abstain from estimating α_I and to disregard any restrictions on α_{II} that may be known. The choice of the θ^*'s is made for simple mathematical derivation. It does not matter that there are fewer θ^*'s than θ's, because our disregard of restrictions on the θ's entails considerable lack of identifiability. The essential point is that each set of values θ is represented by at least one set of 1/ values θ^*, as proved in (S-95).

Since

$$\Sigma^{*-1} = \begin{pmatrix} \Sigma_{II,I}^{-1} & 0 \\ 0 & I_{II,II} \end{pmatrix},$$

the logarithmic likelihood function corresponding to f^* above is

$$(S\text{-}101)\quad L^* = \text{const} + \log\left|\det\begin{pmatrix} \beta_I \\ \beta_{II}^* \end{pmatrix}\right| - \tfrac{1}{2}\log\det\begin{pmatrix} \Sigma_{II,I} & 0 \\ 0 & I_{II,II} \end{pmatrix}$$

$$- \tfrac{1}{2}\operatorname{tr}\begin{pmatrix} \Sigma_{II,I}^{-1} & 0 \\ 0 & I_{II,II} \end{pmatrix}(\alpha_I \alpha_{II}^*)M\begin{pmatrix} \alpha_I' \\ \alpha_{II}'^* \end{pmatrix}$$

---

1. In fact, there is even no identifiability among θ^*'s, because $\Sigma_{II,II} = I_{II,II}$ permits of further orthogonal transformation. This does not bother us either, because the θ^*'s are going to be "maximized out" anyway.

But since the trace of a matrix product XY' is the sum of corresponding elements in each matrix X and Y, L* may be written (taking Σ for X, and $\alpha M \alpha'$ for Y)

$$(S\text{-}102) \quad L^* = \text{const} + \log\left|\det\begin{pmatrix}\beta_I \\ \beta^*_{II}\end{pmatrix}\right| - \tfrac{1}{2}\log\det \Sigma^{II}_{II}$$

$$- \tfrac{1}{2}\mathrm{tr}\,\Sigma^{-1}_{II}(\alpha_I M \alpha_I') - \tfrac{1}{2}\mathrm{tr}\,\alpha^*_{II} M \alpha^{*\prime}_{II} \quad .$$

Suppose we differentiate the second term in L* with with respect to an element $\alpha^*_{g_o k_o}$ which appears in β^*_{II} in $\alpha^*_{II} = (\beta^*_{II} \quad \Gamma^*_{II})$. The resulting partial will be 2/

$$(S\text{-}103) \quad \frac{\partial \log|\det\beta^*|}{\partial \alpha^*_{g_o k_o}} = (\beta'^{*-1})_{g_o k_o} \quad , \text{(by S-86)}$$

where β^* denotes $\begin{pmatrix}\beta_I \\ \beta^*_{II}\end{pmatrix}$. (S-103) is the element in the k_o^{th} row and g_o^{th} column of β'^{*-1}. The matrix of all such partials will be a submatrix in β'^{*-1}, call it $(\beta'^{*-1})_{II}$.

On the other hand, differentiation of the same term in L* with respect to an element $\alpha_{g_o k_o}$ which appears in Γ^*_{II} will give a value of zero. The matrix of all derivatives of $\log|\det\beta^*|$ with respect to the elements of α^*_{II} is therefore

$$\left[(\beta'^{*-1})_{II} \quad 0\right] = (\beta'^{*-1})_{II} \cdot (I \quad 0).$$

occasionally
We shall/use subscripts y, z to indicate numbers of rows and columns in matrices as follows:

y indicates the presence of G rows or columns (G = number of dependent variables).
z indicates the presence of K columns (K = number of predetermined variables).

Similarly, we have already used the notations:

I indicates the presence of G_I rows or columns (G_I = number of equations in set I)
II indicates the presence of G_{II} rows or columns (G_{II} = number of equations in set II).

---

2. Exercise: prove that this is true also when det β^* is negative.

In this notation, our matrix of derivatives of $\log|\det B|$ is

$$(B'^{*-1})_{\pi y} \cdot (I_{yy} \quad 0_{yz}).$$

The partial derivative of the second trace term appearing in L^* with respect to $\alpha^*_{g_o k_o}$ is

$$\frac{\partial \frac{1}{2}\sum_{g=G_0+1}^{G}\sum_{k,l=1}^{K}\alpha^*_{gk}m_{kl}\alpha^*_{gl}}{\partial \alpha^*_{g_o k_o}} = \frac{1}{2}\left(\sum_l m_{k_o l}\alpha^*_{g_o l} + \sum_k \alpha^*_{g_o k}m_{k k_o}\right)$$

$$= \sum_k \alpha^*_{g_o k}m_{k k_o}$$

These can be put together in the matrix

$$(\alpha^* M)_{\pi} = \alpha^*_{\pi} M$$

or, in more explicit notation, $\alpha^*_{\pi x}M_{xx}$.

Since no other terms in L^* depend on α^*_{Π} we have symbolically

$$\frac{\partial L^*}{\partial \alpha^*_{\pi}} = (B'^{*-1})_{\pi y}(I_{yy}\ 0_{yz}) - \alpha^*_{\Pi x}M_{xx} = 0. \qquad (S\text{-}104)$$

Economics 313, December 2, 1948
Lectures 11-15, Supplement

ERRATA to Supplement:

Page 36	line 3	Delete "$\frac{1}{T}$"
page 36	line 5	Delete "disturbances"
page 37	lines 2 and 3	Change "$\alpha_{\pi\pi}$" appearing in both lines to read "α_{π}"
page 37	line 9	Change "$\Sigma_{\pi I}$" to read "$\Sigma_{\pi\pi}$"
page 37	line 18	After "identifiability" insert "among the θ's"
page 38	line 4	Change "$\mathbb{L}$" to "L "
page 38	line 4	Change "$\begin{pmatrix} \Sigma_{II} & 0 \\ 0 & I_{\pi I} \end{pmatrix}$" to read "$\Sigma_{II}$"
page 38	line 11	After "column" insert "of"
page 38	line 15	After "derivatives" insert "of"
page 38	line 18	After "shall" insert "occasionally"
page 38	footnote	After "det" insert "β^*"
page 39	line 2	Change "$(\beta'^{*-1})_y$" to "$(\beta'^{*-1})_{\pi y}$"
page 39	line 12	To (S-104) add "=0"

Estimates under a "compromise" program (Cont'd):

(Continuing from page 39): as a necessary condition on α^*_{π} for a maximum of L* with respect to variation of α^*_{π}.

One might think that the next step would be to solve (S-104) for α^*_{π} and substitute the solution in (S-102). However, there is no unique solution α^*_{π} of (S-104) because, as pointed out in footnote 1 one page 37, α^*_{π} is not even identifiable. However, we can show, by eliminating α^*_{π} from (S-102) and (S-104), that in all points in the space of α^*_{π} which satisfy (S-104) for given values of α_I, Σ_{II}, the likelihood function (S-102) attains the same maximum value. The maximum value of course depends on α_I, Σ_{II}, and will be denoted by $L(\alpha_I, \Sigma_{II})$.

Then by maximizing the resulting function $L(\alpha_I, \Sigma_{II})$ with respect to the parameters α_I, Σ_{II}, of the first subset, we will obtain the needed estimates of the first subset in a computationally more "economical" method than would have been required for simultaneous estimation of all elements of α.

Post-multiplying (S-104) by $(\alpha^{*}_{\Pi x})'$ and recalling that

$$(\alpha^{*}_{\Pi x})' = (\beta^{*}_{\Pi y} \quad \Gamma^{*}_{\Pi z})'$$

we obtain

$$(S\text{-}105)\quad \alpha^{*}_{\Pi x} M_{xx} (\alpha^{*}_{\Pi x})' = (B'^{*-1})_{\Pi y} (I_{yy} \; 0_{yz}) \begin{pmatrix} (\beta^{*}_{\Pi y})' \\ (\Gamma^{*}_{\Pi z})' \end{pmatrix}$$

$$= (B'^{*-1})_{\Pi y} (\beta^{*}_{\Pi y})' = I_{\Pi\Pi},$$

i.e. the $\Pi \cdot \Pi$ submatrix of the unit matrix of order G

$$(S\text{-}105A)\qquad (B'^{*-1}) B' = I_{yy}$$

From (S-105) we then find

$$(S\text{-}106)\quad -\tfrac{1}{2}\mathrm{tr}\,\alpha^{*}_{\Pi x} M_{xx} \alpha'^{*}_{\Pi x} = -\tfrac{1}{2}\mathrm{tr} I_{\Pi\Pi} = \text{constant}.$$

To eliminate $\beta^{*}_{\Pi} \equiv \beta^{*}_{\Pi y}$ from L^{*} in (S-102), we will partition M_{xx} and α^{*}_{Π} as follows:

$$M_{xx} = \begin{pmatrix} M_{yy} & M_{yz} \\ M_{zy} & M_{zz} \end{pmatrix}, \quad \alpha^{*}_{\Pi} \equiv \alpha^{*}_{\Pi x} \equiv (\beta^{*}_{\Pi y} \quad \Gamma^{*}_{\Pi z}).$$

We may then write

$$(S\text{-}107)\quad \alpha^{*}_{\Pi x} M_{xx} = (\beta^{*}_{\Pi y} \quad \Gamma^{*}_{\Pi z}) \begin{pmatrix} M_{yy} & M_{yz} \\ M_{zy} & M_{zz} \end{pmatrix}$$

$$= (\beta^{*}_{\Pi y} M_{yy} + \Gamma^{*}_{\Pi z} M_{zy} \quad \beta^{*}_{\Pi y} M_{yz} + \Gamma^{*}_{\Pi z} M_{zz}).$$

Using the substitution from (S-107) in (S-104) we have

$$(S\text{-}108)\quad \left[(B'^{*-1})_{\Pi y} - \beta^{*}_{\Pi y} M_{yy} - \Gamma^{*}_{\Pi z} M_{zy} \quad -\beta^{*}_{\Pi y} M_{yz} - \Gamma^{*}_{\Pi z} M_{zz}\right]$$

$$= (0_{yy} \quad 0_{yz})$$

which evidently requires

$$(S\text{-}108A)\quad (B'^{*-1})_{\Pi y} - \beta^{*}_{\Pi y} M_{yy} - \Gamma^{*}_{\Pi z} M_{zy} = 0_{yy}$$

$$(S\text{-}108B)\quad -\beta^{*}_{\Pi y} M_{yz} - \Gamma^{*}_{\Pi z} M_{zz} = 0_{yz}$$

Now if our variables z(t) are linearly independent then M^{-1}_{zz} exists. Post-multipying (S-107B) by M^{-1}_{zz} we obtain

$$(S\text{-}108C)\quad \Gamma^{*}_{\Pi z} = -\beta^{*}_{\Pi y} M_{yz} M^{-1}_{zz}.$$

Using the substitution from (S-108C) in (S-108A) we obtain

$$(B'^{*-1})_{\Pi y} - \beta^{*}_{\Pi y} (M_{yy} - M_{yz} M^{-1}_{zz} M_{zy}) = 0$$

or

$$(S\text{-}109)\quad \beta^{*}_{\Pi y} W_{yy} = (B'^{*-1})_{\Pi y},$$

where we define

$$W \equiv W_{yy} \equiv (M_{yy} - M_{yz}M_{zz}^{-1}M_{zy}).$$

Note that W is observable, being a function of observable matrices. That W is a positive definite matrix we shall leave for a later proof.

Postmultiplying (s-109) by [1] $\beta'^* = [(\beta_I)' \quad (\beta_{II}^*)']$ we obtain

$$(S\text{-}110) \quad \beta_{II}^* W \beta'^* = (\beta'^{*-1})_{II}\beta'^* = \{(\beta'^{*-1})_I(\beta_I)' \quad (\beta'^{*-1})_{II}(\beta_{II}^*)'\} = (0_{III} \quad I_{III}),$$

another submatrix of (S-105A). Using this result in the following product matrix we obtain

$$(S\text{-}111) \quad \beta^* W \beta'^* = \begin{pmatrix} \beta_I W \beta'^* \\ \beta_{II} W \beta'^* \end{pmatrix} = \begin{pmatrix} \beta_I W(\beta_I)' & \beta_I W(\beta_{II}^*)' \\ 0_{III} & I_{III} \end{pmatrix}$$

and therefore

$$(S\text{-}112) \quad \det \beta^* W \beta'^* = \det \beta_I W(\beta_I)' .$$

Since we may write

$$\log|\det\beta^*| = \log \frac{\sqrt{\det \beta^* W \beta'^*}}{\sqrt{\det W}}$$

we have finally

$$(S\text{-}113) \quad \log|\det\beta^*| = \tfrac{1}{2}\log\det\beta_I W \beta_I' - \tfrac{1}{2}\log\det W.$$

Using the substitutions from (S-113) and (S-106), we obtain

$$(S\text{-}114) \quad L(\alpha_I, \Sigma_{II}) = \text{const} + \tfrac{1}{2}\log\det\beta_I W \beta_I' - \tfrac{1}{2}\log\det W - \tfrac{1}{2}\log\det\Sigma_{II} - \tfrac{1}{2}\mathrm{tr}\,\Sigma_{II}^{-1}\alpha_I M \alpha_I' .$$

From this expression all parameters of the second subset of equations have disappeared. Using the same steps as we used in deriving (S-99), we find $\hat{\Sigma}_{II}(\alpha_I) = \alpha_I M \alpha_I'$ as the conditional maximizing value of Σ_{II}, and inserting $\Sigma_{II}^{-1} = (\alpha_I M \alpha_I')^{-1}$ in $L(\alpha_I, \Sigma_{II})$ of (S-114), our result is

$$(S\text{-}115) \quad L(\alpha_I) = \text{const} + \tfrac{1}{2}\log\det\beta_I W \beta_I' - \tfrac{1}{2}\log\det W - \tfrac{1}{2}\log\det\alpha_I M \alpha_I' - \tfrac{1}{2}\mathrm{tr}\{(\alpha_I M \alpha_I')^{-1}\alpha_I M \alpha_I'\}$$

$$= \text{const} + \tfrac{1}{2}\log\det\beta_I W \beta_I' - \tfrac{1}{2}\log\det W - \tfrac{1}{2}\log\det\alpha_I M \alpha_I',$$

as the trace of the identity matrix is a constant.

We are now in a position to approach maximum likelihood estimation of the parameters we are interested in, i.e. those

1. In this derivation we place the transposition sign outside parentheses in order to

of the first subset, by less laborious computation. The possible loss of efficiency in the estimates obtained under the present program will be discussed in a subsequent lecture. We may check our results in (S-114) by comparing the form of $L(\alpha_I, \Sigma_{II})$ with that of L^* in (S-102) under the assumption that the subset Π is empty, so that I coincides with the whole set of structural equations.

The relation of W to the disturbances of the reduced form.

Our task here will be to show that the maximum likelihood estimate of the moment matrix of disturbances in the reduced form is precisely W as defined in (S-109(.

We may obtain the likelihood function of the reduced form by the same steps used to reach L^* in (S-102) if we use

$$\Upsilon = B^{-1}$$

We then have

$$\alpha^* = B^{-1}\alpha = B^{-1}(B \quad \Gamma) = (I \quad -\Pi)$$

and $\Sigma^* = B^{-1}\Sigma B'^{-1} = \Omega$, where $\Omega \equiv E(v'v)$.
Our likelihood function is

$$(S\text{-}116) \qquad L(\Pi, \Omega) = \text{const} - \tfrac{1}{2}\log\det\Omega - \tfrac{1}{2}\operatorname{tr}\left\{\Omega^{-1}(I \quad -\Pi)M\begin{pmatrix} I \\ -\Pi' \end{pmatrix}\right\}.$$

In what follows we shall either assume there are no restrictions on Π such as might result from a large number of restrictions on α, or if there are such restrictions on Π, we shall disregard them in maximizing (S-116). First taking Π as given we conditionally estimate Ω by the reasoning underlying (S-99). Our result is

$$(S\text{-}117) \qquad \hat{\Omega}(\Pi) = (I \quad -\Pi)M\begin{pmatrix} I \\ -\Pi' \end{pmatrix} .$$

Substituting (S-117) in (S-116) and maximizing $L(\Pi, \hat{\Omega}(\Pi))$ with respect to Π, we will obtain

$$\hat{\Pi} = P$$

where P is the estimate obtained by the method of least squares. This follows from our previous result regarding the equivalence of least squares estimates and maximum likelihood estimates in the general case of the reduced form[1] (See Supplement page 29). We shall now establish that W satisfies

$$(S\text{-}118) \qquad \hat{\Omega} = \hat{\Omega}(P) = W .$$

(Footnote cont'd) to I, Π etc precedes transposition. We have not consistently followed this notational precaution in previous pages and will again drop it later on when misunderstanding is unlikely. The general rule, consistently applied, is that partitioning preceeds transposition unless where transposition is indicated inside parentheses, partitioning outside.

1. It can however also be proved directly from (S-117) without great difficulty.

If we minimize the mean sum of squares for the i^{th} equation of the reduced form, i.e. minimizing the sum

$$\frac{1}{T}\sum_t\left[y_i(t)-\sum_k \pi_{ik}z_k(t)\right]^2 ,$$

we obtain the so-called normal equations (See equation (275) on page 77, where $\theta_o=-1$)

$$(S\text{-}119)\qquad \begin{pmatrix} m_{z_1y_i}\\ \vdots\\ m_{z_Ky_i}\end{pmatrix} = \begin{pmatrix} m_{z_1z_1} & \cdots & m_{z_1z_K}\\ \vdots & & \\ m_{z_Kz_1} & \cdots & m_{z_Kz_K}\end{pmatrix}\begin{pmatrix} p_{i1}\\ \vdots\\ p_{iK}\end{pmatrix}$$

where $m_{z_ky_i}=\frac{1}{T}\sum_t z_k(t)y_i(t)$, $m_{z_gz_k}=\frac{1}{T}\sum_t z_g(t)z_k(t)$ and p_{ik} are the maximizing values of the π_{ik}.

Repeating this procedure for all G equations of the reduced form, we obtain

$$(S\text{-}120)\qquad \begin{pmatrix} m_{z_1y_1} & \cdots & m_{z_1y_G}\\ \vdots & & \vdots\\ m_{z_Ky_1} & \cdots & m_{z_Ky_G}\end{pmatrix} = \begin{pmatrix} m_{z_1z_1} & \cdots & m_{z_1z_K}\\ \vdots & & \vdots\\ m_{z_Kz_1} & \cdots & m_{z_Kz_K}\end{pmatrix}\begin{pmatrix} p_{11} & \cdots & p_{G1}\\ \vdots & & \\ p_{1K} & \cdots & p_{GK}\end{pmatrix}$$

or more succinctly

$$(S\text{-}120A)\qquad M_{zy}=M_{zz}P' .$$

Premultiplying (S-120A) by M_{zz}^{-1} we have

$$(S\text{-}120B)\qquad M_{zz}^{-1}M_{zy}=P'$$

Since $M_{yz}=M'_{zy}$ and $(M_{zz}^{-1})'=M_{zz}^{-1}$, we take the transpose of P' in (S-120B) to obtain

$$(S\text{-}120C)\qquad P=M_{yz}M_{zz}^{-1}$$

Now substituting these estimates P, P' for Π and Π' in (S-117), we obtain

$$(S\text{-}123)\qquad \hat{\Omega}=\hat{\Omega}(P)=(I_{yy}\ \ -M_{yz}M_{zz}^{-1})\begin{pmatrix} M_{yy} & M_{yz}\\ M_{zy} & M_{zz}\end{pmatrix}\begin{pmatrix} I\\ -M_{zz}^{-1}M_{zy}\end{pmatrix}$$

$$= M_{yy}-M_{yz}M_{zz}^{-1}M_{zy}-M_{yz}M_{zz}^{-1}M_{yz}+M_{yz}M_{zz}^{-1}M_{zz}M_{zz}^{-1}M_{zz}$$
$$= M_{yy}-M_{yz}M_{zz}^{-1}M_{zy}=W ,$$

the result required by (S-118).. We might re-emphasize that this result has been obtained under the disregard of any restrictions on Π possibly arising from restrictions on α .

Economics 313, December 4, 1948
Lectures 11-15, Supplements

ERRATA: page 11, Supplement line 26 Change "y'" to "y" "

To summarize our derivation of $L(\alpha_I)$ in (S-115) in terms of the step-wise maximization process:
We have maximized out $\theta = (\alpha_{II}, \Sigma_{III}, \Sigma_{II\,I})$, given $\eta = (\alpha_I, \Sigma_{I\,I})$. But first we have transformed θ to $\theta^* = (\alpha^*_{II}, \Sigma^*_{II\,II})$ to simplify our task. We then obtained a logarithmic likelihood depending only on η. We then maximized out $\Sigma_{I\,I}$, having obtained $\hat{\Sigma}_{I\,I}(\alpha_I) = \alpha_I M \alpha_I'$. Hence our final result a logarithmic likelihood function depending only on α_I. This result, i.e. formula (S-115) for $L(\alpha_I)$, was originally derived by H. Rubin (See C. C. Discussion Paper No. 308) by a derivation different from the one we have given.

Estimation of a Single Equation of the System.

(The reader should refer again to pages 54-60 and also to the statement of least squares formulae pages 76-81).
In this case the first subset of equations is a single equation, and we wish to estimate the parameters of this equation taking full account of the restrictions on this equation but ignoring any restriction we may know on any others. Our α_I then becomes a row vector of (G+L) elements (if L is the number of predetermined variables) which we write as follows: 1/

$$\alpha_I = \alpha = (\beta \quad \gamma) = (\beta^{\Delta} \quad \beta^{\Delta\Delta} \quad \gamma^{*} \quad \gamma^{**}),$$

where the restrictions on the equation are

$$\beta^{\Delta\Delta} = 0 \quad \text{and} \quad \gamma^{**} = 0 \ .$$

Since the determinant value of a one element matrix is that element itself, our $L(\alpha_I)$ in this case becomes

$$(S\text{-}124) \qquad L(\alpha) = \text{constant} + \tfrac{1}{2}\log\beta(M_{yy} - M_{yz}M^{-1}_{zz}M_{zy})\beta'$$

$$- \tfrac{1}{2}\log\left[(\beta \quad \gamma^{*}) \begin{pmatrix} M_{yy} & M_{yz^{*}} \\ M_{z^{*}y} & M_{z^{*}z^{*}} \end{pmatrix} \begin{pmatrix} \beta' \\ \gamma'^{*} \end{pmatrix}\right]$$

1. Note that the * now denotes partitioning, as on pages 54-60, and not transformation, as on Supplement pages 34-41.

where in writing the expression within square brackets we make explicit use of the fact that $\gamma = (\gamma^* \; 0)$. Let us call this expression within brackets $V^*(\beta, \gamma^*)$ which when the indicated matrix multiplication is carried out becomes

$$(S\text{-}125) \quad V^*(\beta,\gamma^*) = \beta M_{yy}\beta' + \beta M_{yz^*}\gamma^{*\prime} + \gamma^* M_{z^*y}\beta' + \gamma^* M_{z^*z^*}\gamma^{*\prime}$$
$$= \beta M_{yy}\beta' + 2\beta M_{yz^*}\gamma^{*\prime} + \gamma^* M_{z^*z^*}\gamma^{*\prime} \; .$$

We continue on the road toward the maximization of $L(\alpha)$ by first giving β certain arbitrarily fixed values and find the conditionally maximizing value $\gamma^*(\beta)$ of γ^*. Since γ^* occurs only in the last term of (S-124), instead of maximizing $L(\alpha)$ we may minimize $V^*(\beta, \gamma^*)$ with respect to γ^*. Before carrying out this plan, let us interpret the meaning of V^*. For our assumed given values of β write

$$(S\text{-}126) \quad \beta y'(t) = \sum_g \beta_g y_g(t) = \tilde{y}_1(t),$$

a new dependent variable which is thus a given linear function of the orginial dependent variables $y_g(t)$. For $m_{\tilde{y}_1\tilde{y}_1}$, we have

$$(S\text{-}127) \quad m_{\tilde{y}_1\tilde{y}_1} = \frac{1}{T}\sum_t \{\tilde{y}_1(t)\}^2 = \frac{1}{T}\sum_t (\beta y'(t)\; y(t)\beta')$$
$$= \beta\left[\frac{1}{T}\sum_t y'(t)\; y(t)\right]\beta'$$
$$= \beta M_{yy}\beta' \; .$$

For $m_{\tilde{y}_1 z_k}$,

$$(S\text{-}128) \quad m_{\tilde{y}_1 z_k} \equiv \frac{1}{T}\sum_t \tilde{y}_1(t)\; z_k(t) = \beta\left[\frac{1}{T}\sum_t y'(t)\; z_k(t)\right]$$
$$= \beta \begin{pmatrix} m_{y_1 z_k} \\ \vdots \\ m_{y_G z_k} \end{pmatrix} \; ,$$

and consequently

$$(S\text{-}129) \quad (m_{\tilde{y}_1 z_1} \quad m_{\tilde{y}_1 z_2} \quad \cdots \quad m_{\tilde{y}_1 z_L}) = \beta M_{yz} \; .$$

Now using (S-126) and (S-129) we may write V^* in (S-125) also as follows:

$$(S\text{-}130)\quad V^*(\beta,\gamma^*) = m_{\tilde{y}_1\tilde{y}_1} + 2m_{\tilde{y}_1 z^*}\gamma^* + \gamma^* M_{z^*z^*}\gamma^*$$

$$= \frac{1}{T}\sum_t \left[\tilde{y}_1(t) + \gamma^*_{z'^*}(t)\right]\left[z^*(t)\gamma'^* + \tilde{y}_1(t)\right]$$

$$= \frac{1}{T}\sum_t \left[\tilde{y}_1(t) + \gamma^*_{z'^*}(t)\right]^2 .$$

Apparently V^* is that sum of squares of residuals which is minimised if we determine a least squares regression of $\tilde{y}_1(t)$ as defined in (S-126) on the predetermined variables occuring in the equation to be estimated (the sole equation of subset I). Thus we conclude that least squares estimates of γ^* from V^* in (S-130) are at the same time conditional maximum likelihood estimates of γ^*, for given values of β. These estimates are therefore readily obtainable as follows:

minimizing)

$$(S\text{-}131)\quad \frac{\partial V^*(\beta,\gamma^*)}{\partial \gamma^*} = \beta M_{yz^*} + \gamma^* M_{z^*z^*} = 0 .$$

Thus

$$(S\text{-}132)\quad \hat{\gamma}^*(\beta) = -\beta M_{yz^*}M^{-1}_{z^*z^*} .$$

Inserting $\hat{\gamma}^*(\beta)$ in (S-125), we have

$$(S\text{-}133)\quad V^*\{\beta,\hat{\gamma}^*(\beta)\} \equiv V^*(\beta) = \beta M_{yy}\beta' - 2\beta M_{yz^*}M^{-1}_{z^*z^*}M_{z^*y}\beta' + \beta M_{yz^*}M^{-1}_{z^*z^*}M_{z^*z^*}M^{-1}_{z^*z^*}M_{z^*y}\beta'$$

$$= \beta(M_{yy} - M_{yz^*}M^{-1}_{z^*z^*}M_{z^*y})\beta'$$

$$= \beta W^*\beta' .$$

If we define

$$(S\text{-}134)\quad W^* \equiv M_{yy} - M_{yz^*}M^{-1}_{z^*z^*}M_{z^*y}$$

Inserting $V^*(\beta)$ of (S-133) in $L(\alpha)$, we obtain

$$(S\text{-}135)\quad L(\beta) - \text{const} = \tfrac{1}{2}\log\beta W\beta' - \tfrac{1}{2}\log\beta W^*\beta'$$

$$= -\tfrac{1}{2}\log\frac{\beta W^*\beta'}{\beta W\beta'}$$

$$= -\tfrac{1}{2}\log\frac{V^*(\beta)}{V(\beta)} .$$

Here the quantity $V(\beta)$ is formed analogously to $V^*(\beta)$ from

(S-136) $$V(\beta,\gamma) \equiv \frac{1}{T}\sum_t \{\tilde{y}_1(t) + \gamma z'(t)\}^2$$

as follows: Let $\hat{\gamma}(\beta)$ be that value of γ minimizing (S-136) for given β. Comparison of (S-136) with (S-130) and (S-131) gives

(S-137) $$V(\beta) \equiv V\{\beta, \hat{\gamma}(\beta)\} = \beta W \beta' .$$

Thus $V(\beta)$ is the residual variance from a regression of $\tilde{y}_1$ on all predetermined variables z, while $V^*(\beta)$ is the residual variance from a regression of $\tilde{y}_1$ on those predetermined variables z^* allowed to occur in the equation to be estimated. The ratio of the latter to the former variance is to be minimized.

We call $\frac{V^*}{V}$ the variance ratio, and obviously maximum likelihood estimation of β is equivalent to minimizing the variance ratio. This estimation procedure can therefore be called the least variance ratio principle. Reasons why this would appear to be a satisfactory principle of estimation (even if it were not known to be a consequence of maximum likelihood estimation) are stated on pages 90-91.

It is important to remember that we can use $\beta = (\beta^\Delta \quad 0)$. The simplest way to put this into effect is to redefine the meaning of our symbols so that y stands for those dependent variables (previously denoted y^Δ) occuring in the equation of set I, and β (previously β^Δ) for its coefficient vector. The same notation was used on pages 54-60.

We shall now show that the variance ratio is always greater than or equal to one. Consider the inequality for all γ

(S-138) $$V(\beta,\gamma) \geq V\{\beta, \hat{\gamma}(\beta)\}$$ by definition of $\hat{\gamma}(\beta)$.

From (S-136) and (S-130),

(S-139) $$V^*(\beta,\gamma^*) = V\{\beta, (\gamma^* \quad 0)\},$$ and hence

(S-140) $$V^*(\beta) \equiv V^*\{\beta, \hat{\gamma}^*(\beta)\} = V[\beta, \{\hat{\gamma}^*(\beta) \quad 0\}] \geq V\{\beta, \hat{\gamma}(\beta)\} \equiv V(\beta).$$

This result expresses the well known fact that the inclusion of additional variables in the set on which a regression is formed can never increase the minimum value of the sum of squares of residuals. Therefore from (S-130) we obtain

(S-141) $$\frac{V^*(\beta)}{V(\beta)} \geq 1,$$ as desired.

The remaining problem is how to compute the value of $\hat{\beta}$ of β for which $V^*(\beta)/V(\beta)$ reaches its minimum. For the treatment of this problem we return to the middle of page 95 of last years notes.

Economics 313, December 7, 1948
Lecture 16, Supplement

ERRATA in original Notes:

page 3	2nd line preceding (5)	delete "form of the"
page 19	equation (76)	in the third subscripts to β, δ replace "t" with " "
page 26	4th line from bottom	replace "know how" with "have a method"

ERRATA in Supplementary Notes:

page 41	equation (S-106)	subscript to α'^* should be "Πx"
page 42	equation (S-110)	a closing bracket should follow "$(\beta^*_\Pi)'$"
page 42	line below equation (S-110)	should read "submatrix of (S-105A)"
page 42	equation (S-112)	delete "$\det \beta^*_\Pi W(\beta^*_\Pi)' =$ "
page 42	2nd line preceding equation (S-115)	replace "Σ" with "Σ_{II}"
page 43	2nd line preceding equation (S-116)	should read "$\Omega \equiv E(v'v)$"
page 43	equation (S-116)	the trace term should be closed with a bracket
page 43	2nd line preceding equation (S-118)	a footnote reference should follow "form"
page 43	2nd line of footnote	replace "practice" with "notational precaution"
page 43	4th line of footnote	insert "general" after "The"
page 43	bottom of page	insert footnote referred to 2nd line preceding equation (S-118) "It can however also be proved directly from (S-117) without great difficulty"
page 44	line following (S-120C)	delete "using"

ERRATA in Supplementary notes (contd):

page 46	4th line following equation (S-125)	Change "occured" to "occurs"
page 47	equation (S-130)	should read "$m_{\tilde{y}_1\tilde{y}_1}$ + "
page 47	6th line following equation (S-130)	insert "minimizing" after "from"
page 48	line 21	replace "have used" with "can use"
page 48	line 25	insert "β" after "and"
page 48	line 28	insert "for all δ " after "inequality"

The Variance Ratio.

Our numerator in the variance ratio of (S-141) is the residual variance of regression of $\tilde{y}_1 = \beta y'$ on the predetermined variables z^*. The denominator is the residual variance of a regression of y_1 on all predetermined variables $z = (z^* \quad z^{**})$. In minimizing this ratio with respect to β, we are choosing $\hat{\beta}$ in such a way that the reduction of the residual variance obtained by including z^{**} in the estimated regression is as little as possible (in a relative sense).

We may also note that our $\hat{\gamma}^{**} = \hat{\gamma}^{**}(\hat{\beta})$, being consistent estimates of $\gamma^{**} = 0$, are such that as our sample grows infinite we have

$$\frac{V^*(\beta)}{V(\beta)} = 1 \quad ,$$

and for finite samples such that

$\frac{V^*(\beta)}{V(\beta)}$ is as little above 1 as possible.

Economics 313, December 9, 1948
Lecture 18, Supplement

Recalling that $E(v'v)=\Omega$ and defining Ω^* as the matrix of covariances of deviations of $y_1,\dots y_H$, respectively, from their regressions on elements of z^* in the population, we may thus define λ_i in the population to be the roots to the equation

$$\det(\Omega^* - \lambda\Omega) = 0 \quad ,$$

numbered in increasing order: $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_H$.

Since we always have that

(A) $\beta\Pi_{yz^{**}} = 0$ with $\beta \neq 0$,

we have

(B) $\rho(\Pi_{yz^{**}}) \leq H - 1$.

It can also be shown to follow that

(C) $\lambda_1 = 1, \quad \beta(\Omega^* - \Omega) = 0$.

We shall classify all possible cases with respect to the identification and estimation of the equation to be estimated. The first criterion of classification is the number K^{**} of predetermined variables z^{**} excluded from the equation.

CASE I: $K^{**} < H - 1$; $\lambda_1 = \lambda_2 = 1$ (because β not identifiable).

In this case $\rho(P_{yz^{**}}) \leq K^{**} < H - 1$, and we will find $\hat{\lambda}_1 = \hat{\lambda}_2 = 1$; and thus $\hat{\beta}$ is not uniquely defined by (380).

CASE II: $K^{**} = H - 1$.

Here we must distinguish whether or not the equation is actually identifiable.

(a) $\rho(\Pi_{yz^{**}}) < H - 1$; $\lambda_1 = \lambda_2 = 1$ (i.e. β not identifiable).

(a.1) But in this case we have with probability one that $\rho(P_{yz^{**}}) = H - 1$. Barring peculiar samples, we have $\hat{\lambda}_2 > \hat{\lambda}_1 = 1$, and $\hat{\beta}$ is determinate but does not estimate anything identifiable. The warnings we should watch for are very large estimated [1] sampling variances

---

1/ We cannot observe the infinite "true" sampling variances.

of the $\hat{\beta}_i$, and the convergence of $\hat{\lambda}_2$ to the value one as our sample grows large.

(a.2) We would find $\hat{\beta}$ indeterminate should $\rho(P_{yz}^{**}) < H - 1$, an event occuring with a probability zero.

(b) $\rho(\pi_{yz}^{**}) = H - 1$; $\lambda_2 > \lambda_1 = 1$ (i.e. β identifiable).

(b.1) Here we have with probability one that $\rho(P_{yz}^{**}) = H - 1$. Barring degenerate samples, we have $\hat{\lambda}_2 > \hat{\lambda}_1 = 1$, and $\hat{\beta}$ is determinate. Also the sampling variances of the $\hat{\beta}_i$ will be finite.

(b.2) $\rho(P_{yz}^{**}) < H - 1$, an event occuring with probability zero. We would then have $\hat{\lambda}_1 = \hat{\lambda}_2 = 1$, and $\hat{\beta}$ indeterminate. This case will be straightened out simply if we take more observations.

CASE III: $K^{**} > H - 1$.

(a) $\rho(\pi_{yz}^{**}) < H - 1$; $\lambda_1 = \lambda_2 = 1$ (i.e. β not identifiable).

(a.1) $\rho(P_{yz}^{**}) \geq H - 1$, an event occuring with probability one. We have, barring peculiar samples, $\hat{\lambda}_2 > \hat{\lambda}_1 = 1$, and $\hat{\beta}$ is determinate but does not estimate anything identifiable. Again we must watch for large estimated sampling variances of the $\hat{\beta}_i$.

(a.2) Same as (a.2) under Case II.

(b) $\rho(\pi_{yz}^{**}) = H - 1$; $\lambda_2 > \lambda_1 = 1$ (i.e. β identifiable).

(b.1) We have with probability one that $\rho(P_{yz}^{**}) \geq H - 1$, and barring peculiar samples $\hat{\lambda}_2 > \hat{\lambda}_1 \geq 1$. Thus $\hat{\beta}$ is determinate.

(b.2) Same as (b.2) under Case II.

Test on Totality of Restrictions.

When we have hypothezied only the number of restrictions which just identify or fail to identify the equation being investigated, we always have $\lambda_1 = 1$. But suppose we assume "overidentifying" restrictions which actually are not satisfied. Then we are leaving out some of the z^{**} which actually do occur as "explanatory" variables in the true regression of $\tilde{y}_1$ on the z's. Even with an infinite sample it would be impossible to obtain a variance ratio, i.e. λ, equal to one. Thus a very natural test as to whether or not the totality of restrictions we have imposed are correct is a test of the hypothesis $\lambda_1 = 1$ against the alternative $\lambda_1 > 1$. Our criterion for the test is of course $\hat{\lambda}_1$, whose asymptotic sampling distribution has been determined by Anderson and Rubin, as being equivalent to a χ^2 test. (See "Estimation of a single equation from a complete system of stochastic difference equations" to be published in two installments in Annals of Mathematical Statistics). Note that it is only where we have imposed more than the minimum number of restrictions necessary for identification of the equation that we imply anything about reality which is subject to statistical test. Hence it is only an hypothesis embodying over-identifying restrictions which is subject to test!

Suppose that we reject our hypothesis that we have the proper restrictions as a result of the test. Then in principle we should not use our data to test the validity of a smaller number of over-identifying restrictions. For the fact that our data have led to a rejection of the original (first) hypothesis (without that we would never have tested the present (second) hypothesis!) disqualifies the assumption that we have a random sample i.e. the assumption used in evaluating the risks of error under the second test.

What is really needed is a procedure to choose, in one act, one out of several alternative hypotheses, rather than a sequence of choice each between just two alternatives, and each of which prejudiced the use of the same data as a basis for the next choice.

Test for Identifiability.

Suppose $K^{**} \geq H - 1$ but $\rho(\Pi_{yz^{**}}) < H - 1$ (i.e. β not identifiable). We may as result of our data believe we we have a determinate β (See Case III (a.1)). We might test the hypothesis that β is unidentifiable by the test of the hypothesis $\lambda_2 = 1$ (as against $\lambda_2 > 1$) using $\hat{\lambda}_2$ as the criterion of the test. The sampling distribution (even asymptotically) of the latter is still unknown.