COWLES COMMISSION DISCUSSION PAPER: STATISTICS: 333

July 18, 1949

PRELIMINARY CONSIDERATIONS REGARDING TIME SERIES AND/OR CROSS-SECTION STUDIES

by

Clifford Hildreth

I. I find it convenient to have a complete (though highly simplified) model specified as a background for the discussion. Such a model follows:

Entrepreneur's Behavior

(1.1) $y_1^{it} + \beta_{14}y_4^{*t} + \gamma_{15}z_5^{*t} + \gamma_{17}z_7^{it} = U_1^{it}$

(1.2) $y_2^{it} + \beta_{25}y_5^{*t} + \beta_{26}y_6^{it} + \gamma_{22}z_2^{it} + \gamma_{27}z_7^{it} + \gamma_{28}z_8^{it} = U_2^{it}$

(1.3) $\beta_{31}y_1^{it} + y_3^{it} + \gamma_{39}z_9^{it} = U_3^{it}$

(1.4) $y_6^{it} = y_3^{it} \cdot y_5^{*t} - y_1^{it} \cdot y_4^{*t}$

(1.5) $y_7^{it} = z_7^{it} + y_6^{it} - y_2^{it} \cdot y_5^{*t}$

$i = 1, 2, \ldots, I$
$t = 1, 2, \ldots, T$

Worker's Behavior

(1.6) $y_1^{jt} + \phi_{14}y_4^{*t} + \theta_{15}z_5^{*t} + \theta_{17}z_7^{jt} + \theta_{18}z_8^{jt} = U_1^{jt}$

(1.7) $y_2^{jt} + \phi_{25}y_5^{*t} + \phi_{26}y_6^{jt} + \theta_{22}z_2^{jt} + \theta_{27}z_7^{jt} + \theta_{28}z_8^{jt} = U_2^{jt}$

(1.8) $y_6^{jt} = y_1^{jt} \cdot y_4^{*t}$

(1.9) $y_7^{jt} = z_7^{jt} + y_6^{jt} - y_2^{jt} \cdot y_5^{*t}$

$j = (I + 1), (I + 2), \ldots, (I + J)$

Market Equilibrium Conditions

(1.10) $y_1^{It} = y_1^{Jt}$

(1.11) $y_3^{It} - y_2^{It} = y_2^{Jt}$

where $y_g^{It} = \sum_{i=1}^{I} y_g^{it}$, $y_g^{Jt} = \sum_{j=I+1}^{I+J} y_g^{jt}$

The theoretical economy being considered consists of I entrepreneurs and J workers, both groups are also regarded as consumers.

y_1^{it} - labor used by i^{th} entrepreneur in time period t

y_1^{jt} - labor furnished by j^{th} worker in time period t

y_2^{it} - commodities consumed by entrepreneur i in time t

y_2^{jt} - commodities consumed by worker j in time t

$$z_2^{it} = y_2^{i,(t-1)} \; ; \; z_2^{jt} = y_2^{j,(t-1)}$$

y_3^{it} - commodities produced by entrepreneur i in time t

y_4^{*t} - wage rate in time t. The * superscript means that this variable is constant over all i and j.

y_5^{*t} - price of commodities in time t

$$z_5^{*t} = y_5^{*,(t-1)}$$

y_6^{it} - income of entrepreneur i in time t

y_6^{jt} - income of worker j in time t

y_7^{it}, y_7^{jt} - cash held at end of period t

$z_7^{it} = y_7^{i,(t-1)}$, $z_7^{jt} = y_7^{j(t-1)}$

z_8^{it} - composition of i^{th} entrepreneur's family at time t

z_8^{jt} - composition of j^{th} worker's family at time t

z_9^{it} - observed weather variables experienced by i^{th} entrepreneur in period t

II. It is now possible to consider with reference to the above model some of the identification and estimation problems that would be faced by a statistician who desired to estimate one or more structural relations and had observations on some of the variables. Thus far I have considered only a few special cases and would be glad for suggestions of other interesting cases.

For example, suppose only aggregate data were available and that consumption of entrepreneurs and workers had not been recorded separately but only in total - i.e. y_2^{It}, y_2^{Jt} are not observable but $y_2^{I+J,t} = y_2^{It} + y_2^{Jt}$ is observable. An appropriate model would then be:

$$(2.1) \quad y_1^{It} + I \cdot \beta_{14} y_4^{*t} + I \cdot \gamma_{15} z_5^{*t} + \gamma_{17} z_7^{It} = U_1^{It}$$

$$(2.2) \quad y_2^{I+J,t} + (I \cdot \beta_{25} + J \cdot \phi_{25}) y_5^{*t} + \beta_{26} y_6^{It} + \phi_{25} y_6^{Jt}$$

$$+ \gamma_{22} z_2^{It} + \theta_{22} z_2^{Jt} + \gamma_{27} z_7^{It} + \theta_{27} z_7^{Jt}$$

$$+ \gamma_{28} z_8^{It} + \theta_{28} z_8^{Jt} = U_2^{I+J,t}$$

(2.3) $\beta_{31} y_1^{It} + y_2^{I+J,t} + \gamma_{39} z_9^{It} = u_3^{It}$

(2.6) $y_1^{It} + J \cdot \phi_{14} y_4^{*t} + J \cdot \phi_{15} z_5^{*t} + \theta_{17} z_7^{Jt} + \theta_{18} z_8^{Jt} = u_1^{Jt}$

(2.4) $y_6^{It} = y_2^{I+J,t} \cdot y_5^{*t} - y_1^{It} \cdot y_4^{*t}$

(2.8) $y_6^{Jt} = y_1^{It} \cdot y_4^{*t}$

$t = 1 \ldots T$

(2.1) has been obtained by aggregating (1.1) over all i, (2.3) is the aggregate of (1.3), etc. (2.2') is the aggregate of the sum of (1.2) and (1.7). y_1^{It} was substituted for y_1^{Jt} in (2.6) and (2.8) using equilibrium condition (1.10) and $y_2^{I+J,t}$ has been substituted for y_3^{It} in (2.3) and (2.4) using condition (1.11).

Equations (2.1), (2.3), and (2.6) are identifiable; (2.2') is not. If it were assumed that the consumption functions for workers and entrepreneurs were the same, (2.2') could be replaced by:

(2.2") $$y_2^{I+J,t} + (I \cdot \beta_{25} + J \cdot \phi_{25}) y_5^{*t} + \beta_{26} y_6^{I+J,t} + \gamma_{22} z_2^{I+J,t} + \gamma_{27} z_7^{I+J,t} + \gamma_{28} z_8^{I+J,t} = u_2^{I+J,t}$$

(2.4) and (2.8) would be replaced by:

(2.4') $y_6^{I+J,t} = y_2^{I+J,t} \cdot y_5^{*t}$

This would not change the identifiability of the equations.

Aggregation was simple in the above example because the initial relations were assumed linear. Aggregation of non-linear equations generally presents formidable problems but satisfactory procedures may exist in special cases. The special case of initial quadratic equations is discussed in the next section. Rubin has recently (Cowles Commission Papers: Statistics: 329) discussed limited information estimation for non-linear systems.

III. Assume that production (y_3^{it}) of a given entrepreneur is a quadratic function of labor (y_1^{it}) and weather (z_9^{it}). Instead of equation (1.3) above we would then have:

$$(3.1) \quad y_3^{it} + \beta_{31} y_1^{it} + \gamma_{39} z_9^{it} + \beta_{32} (y_1^{it})^2 + \beta_{33} y_1^{it} \cdot z_9^{it} + \gamma_{3,10} (z_9^{it})^2 = u_3^{it}$$

The corresponding aggregate relationship would be:

$$(3.2) \quad y_3^{It} + \beta_{31} y_1^{It} + \gamma_{39} z_9^{It} + \beta_{32} \sum_{i=1}^{I} (y_1^{it})^2 + \beta_{33} \sum_{i=1}^{I} y_1^{it} \cdot z_9^{it} + \gamma_{3,10} \sum_{i=1}^{I} (z_9^{it})^2 = u_3^{It}$$

We assume that y_3^{It}, y_1^{It}, and z_9^{It} are observable for a number of time periods and that $\sum_{i=1}^{I} (y_1^{it})^2$, $\sum_{i=1}^{I} y_1^{it} \cdot z_9^{it}$, $\sum_{i=1}^{I} (z_9^{it})^2$ are not directly observable. Consider the following relations:

(3.3) $$\sum_{i=1}^{I} (y_1^{it})^2 = I \cdot \sigma^t_{y_1^i y_1^i} + \frac{1}{I} (y_1^{It})^2$$

$$\sum_{i=1}^{I} y_1^{it} \cdot z_9^{it} = I \cdot \sigma^t_{y_1^i z_9^i} + \frac{1}{I} y_1^{It} \cdot z_9^{It}$$

$$\sum_{i=1}^{I} (z_9^{it})^2 = I \cdot \sigma^t_{z_9^i z_9} + \frac{1}{I} (z_9^{It})^2$$

where σ^t_{ab} represents the covariance of a and b at time t

In either of two cases, the investigator could use (3.2) and (3.3) to obtain an aggregate equation from which he could estimate coefficients of (3.2). First, if he had estimates of the two variances and the covariance for each time interval he could use (3.3) to get estimates of the three sums not directly observed in (3.2). Second, if the variances and the co-variance were believed to be closely approximated by linear functions of time, then (3.2) could first be rewritten

(3.4) $$y_3^{It} + \beta_{31} y_1^{It} + \gamma_{39} z_9^{It} + \frac{\beta_{32}}{I} (y_1^{It})^2 + \frac{\beta_{33}}{I} y_1^{It} \cdot z_9^{It} + \frac{\gamma_{3,10}}{I} (z_9^{It})^2 + \beta_{32} \cdot I \cdot \sigma^t_{y_1^i y_1^i} + \beta_{33} \cdot I \cdot \sigma^t_{y_1^i z_9^i} + \gamma_{3,10} \cdot I \cdot \sigma^t_{z_9 z_9} = U_3^{It}$$

and then simplified to

(3.5) $$y_3^{It} + \beta_{31} y_1^{It} + \gamma_{39} z_9^{It} + \frac{\beta_{32}}{I} (y_1^{It})^2 + \frac{\beta_{33}}{I} y_1^{It} \cdot z_9^{It} + \frac{\gamma_{3,10}}{I} (z_9^{It})^2 + \gamma_{3,11} \cdot t = U_3^{It}$$

as one member of a system of aggregated equations.

If the latter formulation were used, it would be helpful to have cross-section data available for some of the time periods to furnish some check on the assumption that variances and the covariance are linear functions of time. Such data might exist fairly frequently as the result of censuses and sample surveys.

IV. Before using some equations from the model of Section I to illustrate problems of using cross-section data, we might briefly note the nature of some types of variables that have not been included in the model but which might appear in more elaborate models. It is clear that an investigator using only cross-section data - i.e. observations on activities of a number of individuals during a single time period - could not estimate coefficients of variables with superscripts *t. These variables would be constant in his sample. We might also conceive of models that would contain variables with superscripts i* or j*. These would be factors that varied over individuals but were fixed over time for a given individual. Variables with superscripts *t might be called "chronological" variables and those with superscripts i* might be called "personal" variables to distinguish them from "general" variables with superscripts (i or j)t In some models personal variables might be used to represent such things as formal education of the individual or permanent natural resources peculiar to a particular entrepreneur. While none have been included in the present model, it can readily be seen that coefficients of personal variables could not be estimated from time-series data.

It is also possible that variables, X^{it}, might sometimes be encountered with properties that X^{it} varies over both individuals and time, that X^{it_o} is independent of U^{it_o} where to represents a fixed time period, but that $X^{i_o t}$ is not independent of $U^{i_o t}$. Such a variable could be treated as predetermined in a cross-section study but not in a time-series study. Similarly there might be variables independent of the disturbances over time for a given individual but not independent of disturbances over individuals for a given time period. Such a variable could be properly treated as predetermined in a time-series study but not in a cross section study. Neither of these two types of variables appears in the model of Section I. Speculation as to situations in which they might arise and further discussion of their statistical implications are deferred.

To return to situations that can be examined within the framework of the present model, suppose that an investigator wishes to estimate parameters in the relations determining workers' behavior and that the available data consist of observations of the variables in equations (1.6) and (1.7) from a random sample of workers in a single time period. An appropriate model, derived from equations (1.6), (1.7), and (1.8) would be -

$$(4.1)\quad y_1^{jt} + \theta_{17} z_7^{jt} + \theta_{18} z_8^{jt} + [\phi_{14} y_4^{*t} + \theta_{15} z_5^{*t}] = U_1^{jt}$$

$$(4.2)\quad y_2^{jt} + (\phi_{26} \cdot y_4^{*t})\, y_1^{jt} + \theta_{22} z_2^{jt} + \theta_{27} z_7^{jt} + \theta_{28} z_8^{jt} + [\phi_{25} y_5^{*t}] = U_2^{jt}$$

It is evident that (4.1) is identifiable, in fact it could be estimated by least squares. The constant term in the estimated equation would be an estimate of the quantity in brackets plus the mean (over j) of U_1^{jt}. Thus the cross-section study would not furnish estimates of ϕ_{14} or θ_{15} and the constant term would be peculiar to the time period for which data were available. (4.2) would not be identifiable as the model stands, but would be identifiable if either θ_{27} or θ_{28} were known to be zero. Even in the latter case, no estimate of ϕ_{25} would be obtained, the coefficient of y_1^{jt} could be expected to change in other time periods as y_4^{*t} changed, and the constant term could be expected to be different in different time periods.

V. A question that may occasionally arise in cross-section studies is whether an equation that is unidentifiable in a model utilizing observations of a sample of individuals in a single time period could be identified by a making use of similar observations for another or several other time periods (If enough time periods were available and the model contained a sufficient number of predetermined chronological variables - z_k^{*t}, the use of these might produce identifiability. This section treats the case where the number of time periods are too few to provide useful estimates of moments involving the z_k^{*t}.)

An investigator might face the situation contemplated here if he had survey or census data for two or more periods and if a relation in which he was interested was not identifiable in his cross-section model or if some of the predetermined variables necessary for identifiability had not been recorded by the census or survey takers.

The cross-section model for a particular time period $\underline{t}$ might be represented -

$$(5.1)\ \mathcal{B}\, y^t + \Gamma\, z^t = U^t$$

where $\mathcal{B}$ and Γ are matrices of coefficients of orders G·G and G·K, and

y^t, z^t, U^t are column vectors with respectively G, K, G elements. If the equations of (5.1) are multiplied through by the z^t in order and expected values of the individual terms are taken, the following set of equations results -

$$(5.2) \quad \beta \mu^t_{yz} + \Gamma \mu^t_{zz} = 0$$

where μ^t_{yz} is the matrix of covariances of y^{it} and z^{it} for fixed t; μ^t_{zz} is the variance - covariance matrix of the z's; and μ^t_{uz} is the covariance matrix of the u's and z's and is null since the z's are pre-determined.

Suppose the investigator is interested in the relation whose coefficients are the first rows of β and Γ. Let β_1 and Γ_1 be the first rows of β and Γ. The part of (5.2) that includes these coefficients can be written:-

$$(5.3) \quad (\beta_1 \ \Gamma_1) \begin{pmatrix} \mu^t_{yz} \\ \mu^t_{zz} \end{pmatrix} = 0$$

This can be reduced to -

$$(5.4) \quad (\beta_1^0 \ \Gamma_1^0) \begin{pmatrix} \mu^{t0}_{yz} \\ \mu^{t0}_{zz} \end{pmatrix} = 0$$

where β_1^0 and Γ_1^0 are obtained from β_1 and Γ_1 by eliminating elements known to be zero and $\begin{pmatrix} \mu^{t0}_{yz} \\ \mu^{t0}_{zz} \end{pmatrix}$ is obtained from the corresponding matrix of (5.3) by omitting the rows corresponding to zero elements of $(\beta_1 \ \Gamma_1)$.

If there are G^* non-zero elements of β_1 and K^* of Γ_1, then $\begin{pmatrix} \mu^{t0}_{yz} \\ \mu^{t0}_{zz} \end{pmatrix}$ is of order $(G^* + K^*)$ by K. If the rank of this matrix is $G^* + K^* - 1$, there exists a unique, except for normalizing, solution $(\beta_1^0 \; \Gamma_1^0)$ of (5.4) and the first equation of (5.1) is said to be identifiable.

Let μ^{t0} represent the moment matrix of (5.4). The case to be considered is one in which the relation is unidentifiable in this model but the investigator has observations of the relevant variables for another time period; say τ. First suppose that μ^{t0} has full column rank, rank K and that $G^* + K^* - 1 > K$ (other cases will be noted below). For time τ we have a matrix equation similar to (5.4) -

$$(5.5) \quad (\beta_1^0 \; \Gamma_1^0) \begin{pmatrix} \mu^{\tau 0}_{yz} \\ \mu^{\tau 0}_{zz} \end{pmatrix} = 0$$

which can be combined with (5.4) to give -

$$(5.6) \quad (\beta_1^0 \; \Gamma_1^0) \begin{pmatrix} \mu^{t0}_{yz} & \mu^{\tau 0}_{yz} \\ \mu^{t0}_{zz} & \mu^{\tau 0}_{zz} \end{pmatrix} = 0$$

We are interested to know under what circumstances the moment matrix of (5.6) , $(\mu^{t0} \; \mu^{\tau 0})$, can have rank greater than K and thus possibly provide identifiability of $(\beta_1^0 \; \Gamma_1^0)$.

From (5.2) we have -

$$(5.7) \quad \mu^t_{yz} = -\beta^{-1} \Gamma \mu^t_{zz}$$

givingG linear restrictions on rows of μ^t where μ^t represents the moment matrix of (5.3). If the structure of the system has not changed between time t and time τ then we also have -

(5.8) $\mu_{yz}^{\tau} = -\beta^{-1}\Gamma\mu_{zz}^{\tau}$ and we have G linear restrictions on the G + K rows of $(\mu^{t}\ \mu^{\tau}) = \begin{pmatrix} \mu_{yz}^{t} & \mu_{yz}^{\tau} \\ \mu_{zz}^{t} & \mu_{zz}^{\tau} \end{pmatrix}$. Thus $(\mu^{t}\ \mu^{\tau})$ has rank no greater than K and $(\mu^{t0}\ \mu^{\tau 0})$ derived by deleting (G + K) − (G^{*} + K^{*}) rows of $(\mu^{t}\ \mu^{\tau})$ can have rank no greater than K and the presence of observations for time τ would not help identify the first relation.

However, if the structure of the system had changed between times t and τ, with the first equation not changing, the restrictions (5.8) would differ from those of (5.7). $(\mu^{t0}\ \mu^{\tau 0})$ might then have rank higher than K and the first equation might be identifiable.

If we had a situation in which either (a) $G^{*} + K^{*} - 1 \geq K$, rank $\mu^{t0} < K$ or (b) $G^{*} + K^{*} - 1 < K$ rank $\mu^{t0} < G^{*} + K^{*} - 1$, this would imply that the variables at time t were restricted by relations other than those of (5.1). If those additional restrictions and the restrictions of (5.1) still held in period τ, then observations for the latter period would not identify a previously unidentifiable relation, but a change in the total set of restrictions could make the additional observations effective in helping to identify relations.

Situations in which the ideas of this section could be applied may not be uncommon. In Section IV, if another endogenous variable, say y_{8}^{jt} had been added to (4.1) along with its coefficient β_{18} and we supposed that the investigator did not have observations on z_{2}^{jt}, then consistent estimates of coefficients of (4.1) could not be obtained from cross-section data for only one time period. Consistent estimates for (4.1) could, however, be obtained from cross-section data for two periods for which y_{4}^{*t}, the wage rate, differed. In agriculture where cross section data are frequently available, changes in government programs might fairly often provide changes in structure which would not effect the coefficients of agricultural production functions or responses of farmers to prices received and prices paid.