

Cowles Foundation Paper 17

Reprinted from
ECONOMETRICA, Journal of the Econometric Society, Vol. 14, No. 2, April, 1946
The University of Chicago, Chicago 37, Illinois, U.S.A.

ESTIMATING RELATIONS FROM NONEXPERIMENTAL OBSERVATIONS

Abstracts of papers presented by members of the staff of the Cowles Commission for Research in Economics at joint sessions of the American Statistical Association, The Econometric Society, and the Institute of Mathematical Statistics, at Cleveland, Ohio, Friday, January 25, 1946.

The sessions on Friday, January 25, were held jointly with the American Statistical Association and the Institute of Mathematical Statistics, and were devoted to Estimating Relations from Nonexperimental Observations. The chairman of the morning session was MORDECAI EZEKIEL, Department of Agriculture, and the following papers were presented:

JACOB MARSCHAK, Cowles Commission for Research in Economics, *The Economist's Problem of Statistical Inference*.*—In its team work, the Cowles Commission tries to adapt statistical tools to two peculiarities of economic observations: (1) the observations often have the form of time series $y_1(t), y_2(t), \dots, t=1, \dots, T$, the variables being "autoregressive"; (2) the observations are (as shown by Haavelmo) solutions of a system of stochastic simultaneous equations:

$$(1) \quad \phi_g[y_1(t), \dots, y_G(t), z_1(t), \dots, z_K(t)] = u_g(t), \quad \begin{array}{l} g = 1, \dots, G, \\ t = 1, \dots, T; \end{array}$$

where the y and z are observable variables and the u are unobservable random disturbances having an unknown joint distribution $H(u; \epsilon)$. The speaker treated the second problem only and confined himself to

* This abstract and the two following ones, by Leonid Hurwicz and by Tjalling Koopmans and Roy Bergh Leipnik, will be reprinted as Cowles Commission Papers, New Series, No. 17.

its population aspects by assuming the sample to be sufficiently large so that the joint distribution of any set of observable variables is determinable from the observations. We want to estimate the set of "structural parameters" $\theta = (\alpha, \epsilon)$, where α stands for the parameters of the functions ϕ_σ , and ϵ for the parameters of the joint distribution function $H(u)$ of the vector of "random disturbances in equations," u . The variables y, z are assumed to be observable without errors. If errors—or any other "random disturbances in variables"—are introduced, $y_\sigma(t)$ is replaced by $[y_\sigma(t) - w_\sigma(t)]$, say; this complication has not yet been studied.

It is assumed that no equations independent of (1) exist that would involve the G "jointly dependent variables" $y(t)$. The K variables $z(t)$, on the contrary, are, in Koopmans' terminology, "predetermined" in additional equations, outside of (1). The random disturbances (if any) in those equations are distributed independently of $H(u; \epsilon)$. Predetermined variables include both exogenous variables and lagged endogenous ones,¹ so that $z_i(t)$ may be $= y_\sigma(t-1)$, say.

For a given t , an experimenter would replace all but one equation in (1) by

$$y_\sigma(t) = c_\sigma(t) + v_\sigma(t), \quad \sigma = 2, \dots, G,$$

where $c_\sigma(t)$ is a controlled constant and $v_\sigma(t)$ random; this makes all variables except $y_1(t)$ predetermined. In economics experiments are not feasible and the system will in general involve two or more jointly dependent variables.

As an example—admittedly oversimplified—consider the jointly dependent variables x (national product, identical with income, or supply) and y (demand for all goods) in the system

$$\begin{aligned} y_t - \alpha x_t - \beta &= u_t && \text{(behavior of buyers),} \\ y_t - x_t &= v_t && \text{(behavior of producers);} \end{aligned}$$

here time is indicated by subscripts; u_t and v_t ("random shift of demand," "failure to adjust production") are like drawings from urns; each pair of values drawn determines $x_t = (u_t - v_t + \beta)/(1 - \alpha)$, $y_t = (u_t - \alpha v_t + \beta)/(1 - \alpha)$. If the joint distribution $H(u, v)$ is normal and has moments

$$\begin{aligned} Eu_t = Ev_t &= 0, & Eu_t u_t &= \sigma_{uu} \delta_{tt}, & Ev_t v_t &= \sigma_{vv} \delta_{tt}, \\ Eu_t v_t &= Ev_t u_t &= 0, \end{aligned}$$

then the joint distribution $F(x, y)$ is normal and has moments

¹ See T. Koopmans, "When is an Equation System Complete for Statistical Purposes?" in *Statistical Inference in Dynamic Economic Systems*, to be published as Cowles Commission Monograph No. 10.

$$(2) \quad \begin{aligned} Ex = Ey = 0, \quad \sigma_{xx} &= (\sigma_{uu} + \sigma_{vv})/(1 - \alpha)^2, \\ \sigma_{yy} &= (\sigma_{uu} + \alpha^2\sigma_{vv})/(1 - \alpha)^2, \quad \sigma_{xy} = (\sigma_{uu} + \alpha\sigma_{vv})/(1 - \alpha)^2. \end{aligned}$$

In this case, the full set of structural parameters $\theta = (\alpha, \beta, \sigma_{uu}, \sigma_{vv})$ can be determined from the distribution $F(x, y)$. If, however, the assumption $Eu_i v_i = 0$ were not justified on a priori grounds, and no other a priori knowledge of θ were available (e.g., the knowledge of β or of σ_{uu}/σ_{vv}), the system would be "nonidentifiable": more than one value θ would correspond to the given distribution $F(x, y)$ of the variables.

Estimation of θ is necessary because policy and environment may be subject not only to known "nonstructural changes," i.e., those in the observable noneconomic variables—some of the z 's in (1)—but also to "structural changes," i.e., those in the parameters θ . Let $\theta = \theta^0$ for $t = 1, \dots, T$ (period of observation) and $\theta = \theta^1$ for $t > T$. It was suggested by Hurwicz to use an operator P , to denote a given change in policy or environment so that $\theta^1 = P(\theta^0)$. In the example above, P may mean a certain percentage change in α through introduction of income tax, or a certain absolute change in β through embarking upon government spending. Denote the old and the new distribution of x, y by $F^0 = F(x, y | \theta^0)$, and $F^1 = F(x, y | \theta^1)$ respectively; consider also the corresponding regressions of y on x :

$$E(y | x, \theta^0) = \gamma^0 x + \delta^0, \quad E(y | x, \theta^1) = \gamma^1 x + \delta^1.$$

We have $\gamma^0 = \sigma_{xy}^0 / \sigma_{xx}^0 = (\sigma_{uu}^0 + \alpha^0 \sigma_{vv}^0) / (\sigma_{uu}^0 + \sigma_{vv}^0)$, and a similar expression for γ^1 . It is seen that (1) $\gamma^0 \rightarrow \alpha^0$ or $\rightarrow 1$ as $\sigma_{uu}^0 / \sigma_{vv}^0 \rightarrow 0$ or $\rightarrow \infty$ respectively, and that (2) γ^1 cannot be obtained from γ^0 ; but γ^1 can be obtained from the knowledge of θ^0 and of the operator P . Only if no change in the structure needs to be presumed (i.e., if P is an identical transformation) can the ordinary regression coefficient γ^0 , or any other parameters of F^0 , computed from the observations on the past, be used to predict the future. If, on the other hand, the structure is supposed to undergo a given change P (as is the case when the effects of a proposed policy are discussed), the observations, and the old distribution F^0 based on them must be used to determine the structure θ^0 , if identifiable. From the latter and the known operator P the new distribution F^1 and any of its parameters (including γ^1 if so desired) can be derived. If the structure θ^0 is not identifiable, there is no way to predict effects of known structural changes.

LEONID HURWICZ, Cowles Commission, Guggenheim Foundation Fellow, *Sampling Aspects of Structural Estimation and Prediction*.—*I. Model and definitions*. The model is given by

$$\begin{array}{l}
 (1.1) \quad \left. \begin{array}{l} \phi_g[\eta_1(t), \dots, \eta_G(t), \eta_1(t-1), \dots; \\ \zeta_1(t), \dots, \zeta_K(t)] = u_g(t), \end{array} \right\} \begin{array}{l} g = 1, 2, \dots, G, \\ k = 1, 2, \dots, K, \\ t = 1, 2, \dots, T. \end{array} \\
 (1.21) \quad y_g(t) = \eta_g(t) + v_g^{(y)}(t) \\
 (1.22) \quad z_k(t) = \zeta_k(t) + v_k^{(z)}(t)
 \end{array}$$

The z 's and the y 's are the *observed* values, the ζ 's and the η 's the "*true*" (nonobservable) values of the *exogenous* and *endogenous* variables respectively. The u 's and the v 's are respectively the disturbances "in relationships" and "in variables" (e.g., observation errors); they are nonobservable stochastic variables. If, given the cumulative distribution F of the y 's and z 's, it is possible to determine uniquely the relationship vector $\Phi = \{\phi_1, \dots, \phi_G\}$ and the cumulative distribution H of the disturbances, the model is said to be identified. The estimation of Φ and H is called *structural estimation*. The author has shown¹ that there exist identifiable systems where both u 's and v 's are present and the u 's may be autocorrelated. Most of the recent contributions, however, assume either (A) nonautocorrelated v 's and absence of u 's, or (B) nonautocorrelated u 's and absence of v 's; the rest of this note applies to (B).

Prediction (under unchanged structural conditions) is defined² as estimation of the regression coefficients of the *predictand* [say $y_1(t)$] on (some or all of) the *predictors* $y_2(t), \dots, y_1(t-1), \dots, z_1(t), \dots$. Thus in the linear case prediction consists in finding the estimates q of the χ 's in

$$\begin{array}{l}
 (2) \quad E[y_1(t) | y_2(t), \dots] \\
 = \chi_{20}^{(y)} y_2(t) + \dots + \chi_{11}^{(y)} y_1(t-1) + \dots + \chi_{1(z)}^{(y)} z_1(t) + \dots
 \end{array}$$

II. *Structural estimation*. Foundations for large-sample theory of structural estimation have been laid by Mann and Wald,³ and the consistency proof extended by Koopmans⁴ and Rubin.⁵ The efficiency of estimates has not yet been examined. Only the rudiments of small-sample theory exist. General multivariate theorems cannot be applied to *autoregressive* (i.e., difference equations) systems (where lagged values of the endogenous variables appear, and possibly also exogenous

¹ Leonid Hurwicz, "Variable Parameters in Stochastic Processes," in *Statistical Inference in Dynamic Economic Systems*, to be published as Cowles Commission Monograph No. 10.

² Leonid Hurwicz, "Prediction and Least Squares," in *ibid.*

³ H. B. Mann and A. Wald, "On the Statistical Treatment of Linear Stochastic Difference Equations," *ECONOMETRICA*, Vol. 11, October, 1943, pp. 173-220.

⁴ In an unpublished manuscript.

⁵ Herman Rubin, "Consistency of Maximum-Likelihood Estimates in the Explosive Case," in *Statistical Inference* above cited.

variables are present). Work done on small-sample properties of the autoregressive systems has been mostly confined to the first-order difference equation

$$(3) \quad y(t) = \alpha y(t-1) + u(t), \quad t = 2, 3, \dots, T,$$

where the u 's are independently normally distributed with a common variance and a common mean, and $|\alpha| < 1$. As to the initial value $y(1)$ we may assume

$$(4.1) \quad y(1) = \text{fixed variate, or}$$

$$(4.2) \quad y(1) = \text{normally distributed according to the marginal distribution of the other } y\text{'s, or}$$

$$(4.3) \quad y(1) = \alpha y(T) + u(1) \text{ where } u(1) \text{ has the same properties as the other } u\text{'s (the circular case).}$$

All the small-sample literature known to the author is either based on (4.3) (which is seldom realistic and is justified mainly as a convenient approximation) or only applies to the case $\alpha = 0$. In the author's opinion (4.2) is most realistic, although unlike (4.1) it does not imply the equivalence of the least-squares and the maximum-likelihood estimates.

It is possible to evaluate, assuming (4.2) and zero mean for the disturbances, the bias of the least-squares estimate a of α .⁶ This bias is not negligible. For instance,

$$(5) \quad R_T \equiv \lim_{\alpha \rightarrow 0} \frac{E(a)}{\alpha} = \frac{T^2 - 2T + 3}{T^2 - 1}$$

where $a = \sum_{t=2}^T y(t)y(t-1) / \sum_{t=2}^T y^2(t-1)$ and T is the size of the sample. Thus for $T = 20$, $R = 0.910$, i.e., there is a bias of 9 per cent. This bias tends to disappear as $|\alpha| \rightarrow 1$, but not too rapidly.

III. Prediction. Prediction, as defined in *I*, is possible regardless of the identification of the model.⁷ The author has studied some of the optimal properties of the least-squares estimates \tilde{q} of the χ 's in (2). The following are some of the results:

(A) In the general *autoregressive* case the \tilde{q} 's are biased. This is implied by (5) above, since $E[y(t) | y(t-1)] = \alpha y(t-1)$ so that $\alpha \equiv \chi_{11}^{(y)}$; hence there is a case where \tilde{q} (*viz.*, a) is biased.

(B) In the *nonautoregressive* case (no lagged y 's among predictors) the \tilde{q} 's are the best (i.e., least variance) linear *absolutely* unbiased estimates of the χ 's. (An *absolutely* unbiased estimate is unbiased for

⁶ Leonid Hurwicz, "Least-Squares Bias in Time Series," in *ibid.*

⁷ Provided no structural change occurs. Cf. Jacob Marschak, in preceding paper, p. 167.

$$(4) \quad \mathbf{x}_d = \|\|y_1 \cdots y_G\|, \quad \mathbf{x}_p = \|\|z_1 \cdots z_K\|$$

are row vectors combining the *d*ependent and *p*redetermined variables respectively. The latter $[\mathbf{x}_p(t)]$ are either noneconomic (exogenous) variables causally and stochastically independent of the disturbances $\mathbf{u}(t)$ and dependent variables $\mathbf{x}_d(t)$, or economic (endogenous) variables dependent only on values of $\mathbf{u}(t-\tau)$, $\mathbf{x}_d(t-\tau)$ at earlier time points ($\tau=1, 2, \dots$).

The equations (3) are invariant under a linear transformation

$$(5) \quad \mathbf{A}_{dx}^* = \mathbf{\Pi}_{dd}\mathbf{A}_{dx}, \quad \Sigma_{dd}^* = \mathbf{\Pi}_{dd} \Sigma_{dd} \mathbf{\Pi}_{dd}',$$

with nonsingular matrix $\mathbf{\Pi}_{dd}$. In order to identify individual equations within (3) as structural (i.e., economic behavior) equations, it is necessary to impose linear a priori restrictions on \mathbf{A}_{dp} specifying which variables enter into which equations. The system (3) is called completely identified if when \mathbf{A}_{dp} satisfies these restrictions, the only transformations $\mathbf{\Pi}_{dd}$ for which \mathbf{A}_{dp}^* satisfies the same restrictions are diagonal (corresponding to a change of normalization in each equation).

On the basis of the assumptions stated, the logarithm of the distribution function of the observations $\mathbf{x}_d(t)$, $t=1, \dots, T$, is, after division by T and subtraction of a constant,

$$(6) \quad L(\mathbf{A}_{dp}, \Sigma_{dd}) = \log \det \mathbf{A}_{dd} - \frac{1}{2} \log \det \Sigma_{dd} \\ - \frac{1}{2} \text{tr} \Sigma_{dd}^{-1} \mathbf{A}_{dx} \mathbf{M}_{dd} \mathbf{A}_{dx}'.$$

Maximum-likelihood estimation requires finding those values \mathbf{A}_{dx} , \mathbf{S}_{dd} of \mathbf{A}_{dx} , Σ_{dd} for which this function attains its highest maximum under the restrictions stated. This problem has been studied in two cases. In one case, no restrictions are imposed on the covariance matrix Σ_{dd} of the disturbances. In the other case, to which the remainder of the paper is confined, the disturbances in different structural equations are assumed to be uncorrelated. In this case we can normalize each equation such that Σ_{dd} equals the unit matrix \mathbf{I}_{dd} , and the function (6) becomes

$$(7) \quad L(\mathbf{A}_{dp}) = \log \det \mathbf{A}_{dd} - \frac{1}{2} \text{tr} \mathbf{A}_{dx} \mathbf{M}_{xx} \mathbf{A}_{dx}'.$$

Three iterative methods to find a maximum of this function were discussed. In the first method, from an initial value $\mathbf{A}_{dx}^{(0)} = \mathbf{A}_{dx}^{(0,0)}$ satisfying the a priori restrictions, a revised value $\mathbf{A}_{dx}^{(0,1)}$, is derived in which only the first row of $\mathbf{A}_{dx}^{(0,0)}$ is replaced by new values which maximize L subject to the restrictions. To find these new values is essentially a linear problem. In the same way $\mathbf{A}_{dx}^{(0,2)}$ is derived from $\mathbf{A}_{dx}^{(0,1)}$ by revising the second row, etc., until the revision of the last row gives $\mathbf{A}_{dx}^{(0,G)} = \mathbf{A}_{dx}^{(1)} = \mathbf{A}_{dx}^{(1,0)}$ as the result of the first iteration.

In a variant of this method each $A_{dx}^{(0,\theta)}$ is derived from $A_{dx}^{(0,0)}$ instead of from $A_{dx}^{(0,\theta-1)}$, and the g th row of $A_{dx}^{(0,\theta)}$ is taken as the g th row of $A_{dx}^{(1)}$. This procedure is a special case of the second method suggested by J. von Neumann. In this method iterations are determined by

$$(8) \quad A_{dx}^{(n+1)} = A_{dx}^{(n)} + \epsilon_n D_{dx}^{(n)}$$

where

$$(9) \quad \mathcal{L}\{D_{dx}^{(n)}M_{xx}\} = \mathcal{L}\{(A_{dd}^{(n)})'^{-1}I_{dx} - A_{dx}^{(n)}M_{xx}\}.$$

Here $I_{dx} = \begin{bmatrix} I_{dd} & O_{dp} \\ 0 & I_{dp} \end{bmatrix}$, and the operator \mathcal{L} denotes a projection on the linear subspace defined by the a priori restrictions on A_{dx} , while $D_{dx}^{(n)}$ satisfies the same restrictions $D_{dx}^{(n)} = \mathcal{L}\{D_{dx}^{(n)}\}$. It can be shown that for sufficiently small positive values of ϵ_n the substitution (8) always increases L unless an extremum is already reached in $A_{dx}^{(n)}$. By taking $\epsilon_n = 1$ the special case just described is obtained. Alternatively, one may take a value of ϵ_n that maximizes the sum of the linear and quadratic terms in the Taylor expansion of $L(A_{dx}^{(n+1)}) - L(A_{dx}^{(n)})$ with respect to ϵ_n .

Faster convergence is obtained by the Newton method, in which, instead of (9), one defines $D_{dx}^{(n)}$ by

$$(10) \quad \begin{aligned} &\mathcal{L}\{(A_{dd}^{(n)})'^{-1}D_{dd}^{(n)}(A_{dd}^{(n)})'^{-1}I_{dx} + D_{dx}^{(n)}M_{xx}\} \\ &= \mathcal{L}\{(A_{dd}^{(n)})'^{-1}I_{dx} - A_{dx}^{(n)}M_{xx}\}. \end{aligned}$$

In a few examples this advantage was more than offset by the increased work per iteration, since the solution of (10) for $D_{dx}^{(n)}$ has to be repeated for each iteration, while in (9) one matrix inversion provides the solution for each value of n . The Newton method is therefore recommended only as the concluding step in the last iteration, since it yields estimates of the sampling variances and covariances of the maximum-likelihood estimates as a by-product. Otherwise, (9) with $\epsilon_n = 1$ has proved to be the most economic procedure in the cases studied.

Unsolved problems are the best choice of initial values, and the conditions under which the highest maximum of L is obtained.