COWLES COMMISSION DISCUSSION PAPERS: STATISTICS, No. 311

Statistical Problems in Forecasting from Econometric Models

by

Kenneth J. Arrow

(The problems treated in this paper were suggested by S. Cohn, E. E. Hagen, and A. Smithies, of the Bureau of the Budget, and S. Siegel, of the Board of Governors of the Federal Reserve System.)

1. The Problem of Forecasting

The theory of statistical forecasting has never been systematically explored, especially in conjunction with the now-recognized distinction between structural and regression relations. (Mention must be made of papers by Hotelling, particularly one on the standard error of a forecast, and a recent paper by Hoel on the choice between two forecasting formulas.)

The usual formulation of the forecasting problem runs as follows: We have a sample X of observations made without error at times $1, \ldots, T$, (referred to as the base period) the sample being drawn from a probability distribution generated by a structure S. S is known to belong to a model M. The structure at time $T^* > T$ is $S^* = f(S)$, where f is a known structural change (e.g., f may specify one of the parameters unspecified by the model M, or it may say that a given parameter, such as the marginal propensity to consume, is a specified proportion of its value in the base period or is a specified amount higher). Assume finally that the values of the variables predetermined at time T^* under the structure S^*, z^* are known exactly. The problem of forecasting is then to express an optimum forecast of the jointly determined variables at T^* under S^* as a function $y^* = y^*(z^*, X)$. Assuming that M identifies

S, the procedure has been to estimate S by its maximum likelihood estimate $\hat{S}(X)$. Then estimate S* by $\hat{S}^* = f(\hat{S})$, and finally use $\hat{S}^*$ to express y^* in terms of z^* by setting the disturbances in $\hat{S}^*$ equal to zero. Under the previous assumptions plus the assumption of large T, this method of forecasting will have well-known optimum properties.

The actual problem confronting the forecaster differs in several ways from the above scheme. The observations on base period are not necessarily made without error; considerations of this type lead to shock-error models, which will not be considered here. Variations on the knowledge available at the time of forecast are possible. For one thing due to surveys and lags in publications, estimates, perfect or imperfect, may be available for one or more jointly determined variables; these estimates should be combined with the forecasts from the fitted structure $\hat{S}^*$. In effect, these additional estimates permit a better forecast of y^* to be obtained by forecasting the disturbances at a level different from zero. This problem is discussed in greater detail in Section 2.

On the other hand, the assumption of perfect knowledge of the predetermined variables may be stronger than reality permits. If we bar errors of observations, it would seem as though the lagged endogenous variables should be perfectly known; but the time required for analysis may require the forecaster to work with preliminary estimates. As for the exogenous variables, assuming them to be perfectly known at time T* would require that the variances of the disturbances in the part of the structure determining the distribution of the exogenous variables be zero (this might be assumed of exogenous variables determined by government policy), unless, indeed, the "forecast" were not actually made until after

the values of the exogenous variables are observed. I conjecture that for large samples the optimum forecasting method, in the case of imperfectly known predetermined variables, would be to use $\hat{S}^*$, based on the maximum likelihood estimate $\hat{S}$, must substitute for z^* the best estimates available.

Another variation of the knowledge available at the time of forecast is to assume that several observations on the new structure S^* are available at times $T', \ldots, T^{*-1}$. If γ is a known function of S, this case can be reduced to the previous one by introducing an exogenous variable taking on the values 0 for $t = 1, \ldots, T$, and 1 for $t = T', \ldots, T^{*-1}$, and considering all observations together for maximum likelihood estimation of S, S^*. The case where the structural change f is not completely known is formulated in Section 3.

Among other points, it is suggested in Section 3, that structural change in many cases is not genuine but apparent, arising out of the incorrect specification of M. This leads naturally to the question of specification bias (error due to incorrect specification of M), which is discussed in Section 4 in a particular example.

A brief discussion of a general criterion for optimum estimates is discussed in Section 5.

2. The Use of Additional Information on the Jointly Determined Variables.

Suppose a linear econometric model to have been fitted and expressed in reduced forms. Let $y_1, \ldots, y_H$ be the endogenous variables and $z_1, \ldots, z_K$ the predetermined variables.

$$(2.1) \quad y_h = \sum_{k=1}^{K} \pi_{hk} z_k + v_h ,$$

where v_h is the random disturbance in the reduced form and $E(v_h)=0$. In the absence of additional information, the optimum forecast $\hat{y}_h$ of y_h is obtained by setting $v_h = 0$ in (2.1). Suppose, however, that, by means of a survey or otherwise, another forecast y', of y, is available, and that the errors in this forecast are independent of the disturbances in the reduced forms. This may arise in a survey, where the error is due to sampling fluctuations rather than random disturbances in relations.

$$(2.2) \quad y'_1 = y_1 + w,$$

where w is independent of $v_1, \ldots, v_H$ and $c(w) = 0$

Case I: $E(w^2) = 0$

In this case, $y'_1 = y_1$, so that the optimum forecast of y_1 is simply y'_1, which is a perfect forecast under the assumption.

$$(2.3) \quad y_1 = y'_1.$$

For the other y_h's, the knowledge of y_1 yields additional information, since $v_1 = y_1 - \sum_{k=1}^{K} \pi_{1k} z_k$ is now known. As $v_2, \ldots, v_H$ are, in general, correlated with v_1, better estimates of y_h can be obtained by resubstituting for v_h its regression on v_1 instead of its expected value, 0.

$$(2.4) \quad \hat{y}_h = \sum_{k=1}^{K} \pi_{hk} z_k + E(v_h|v_1) \qquad (h = 2, \ldots, H)$$

As $E(v_h) = E(v_1) = 0$

$$E(v_h|v_1) = \rho_{h1} v_1$$

ρ_{hl} can be estimated at the same time as the π_{hk}'s. After having estimated the π_{hk}'s, compute the residuals v_h for the period in which the fit was made by the formula

$$v_h = y_h - \sum_{k=1}^{K} \pi_{hk} z_k$$

and then, by the usual regression formula,

$$\rho_{hl} = \frac{\sum_{t=1}^{T} v_{ht} v_{lt}}{\sum_{t=1}^{T} v_{lt}^2}$$

from (2.4) the best forecast of y_h is given by

$$(2.5)\quad \hat{y}_h = \sum_{k=1}^{K} \pi_{hk} z_k + \rho_{hl} \left(y'_l - \sum_{k=1}^{K} \pi_{lk} z_k \right) = \sum_{k=1}^{K} \left(\pi_{hk} - \rho_{hl} \pi_{lk} \right) z_k + \rho_{hl} y'_l$$

$$(h = 2, \ldots, H)$$

Case II. General Case

We now assume the survey information is not perfect but affected with an error w. We have two forecasts of y_1: y'_1 and $\sum_{k=1}^{K} \pi_{1k} z_k$. The first has a variance σ_w^2, the second $\sigma_{v_1}^2$. To combine these, assuming w independent of u_1, let the best forecast of y_1 be

$$y_1 = \alpha y'_1 + (1-\alpha) \sum_{k=1}^{K} \pi_{1k} z_k),$$

where α is chosen so as to minimize the variance of $\hat{y}_1 - y_1$, which is

$$(2.6)\quad \alpha^2 \sigma_w^2 + (1-\alpha)^2 \sigma_{v_1}^2$$

The minimizing value of α is

$$(2.7)\quad \frac{\sigma_{v_1}^2}{\sigma_w^2 + \sigma_{v_1}^2} .$$

i.e., our two forecasts of y_1 are weighted inversely to their variances. Hence

$$(2.8)\quad \hat{y}_1 = \frac{\sigma_{v_1}^2}{\sigma_{v_1}^2 + \sigma_w^2}\, y'_1 + \frac{\sigma_w^2}{\sigma_{v_1}^2 + \sigma_w^2} \sum_{k=1}^{K} \pi_{1k} z_k$$

Then v_1 can be estimated from the formula,

$$(2.9)\quad \hat{v}_1 = \hat{y}_1 - \sum_{k=1}^{K} \pi_{1k} z_h$$

v_h ($h = 2, 1 \ldots, H$) can be estimated from the regression relation,

$$\hat{v}_h = \rho_{h1} \hat{v}_1,$$

so that the best forecasts of y_h ($h = 2, \ldots, H$) are

$$(2.10)\quad \hat{y}_h = \sum_{k=1}^{K} \pi_{hk} z_k + \rho_{h1} \hat{v}_1 = \sum_{k=1}^{K} (\pi_{hk} - \rho_{h1} \pi_{1k} z_k) + \rho_{h1} \hat{y}_1.$$

The above analysis has assumed the π_{hk}'s to be known exactly. If they are themselves estimates, the argument remains the same except that $\sigma_{v_1}^2$ is replaced by the variance of forecast of y_1 in relation (2.6)-(2.8). This variance is given by the formula

$$(2.11)\quad \sigma_{F_1}^2 = \sum_{j=1}^{K} \sum_{k=1}^{K} \sigma_{\pi_{1j}\pi_{1k}} z_j z_k + \sigma_{v_1}^2,$$

where $\sigma_{\pi_{1j}\pi_{1k}}$ is the covariance of the estimates of π_{1j} and π_{1k}.

A more delicate problem, and one not yet formalized, is to use survey information to arrive at a schedule relating, e.g., investment intentions to anticipated sales.

3. Structural Change.

Ignoring other complications, we find in practice that structural change may occur in an unknown way. For example, it is widely argued that the present structure of the American economy is different from the inter-war period. In other sciences, we do not like to consider the possibility of the structure changing in an unknown way; why should this possibility arise in economics?

I think the correct answer lies in the realization that our models do not really include all structures we believe possible. For reasons of simplicity, our models exclude more structures than are justified by our a priori information. Presumably, we seek to have so many structures in the model that the "distance" (as measured by the costs of choosing a false structure) between any structure which is compatible with our genuine a priori information and the nearest structure in the model is less than some preassigned level. But this distance may depend on the values of the exogenous variables. E.g., assuming relations to be linear when they are not means that the best linear approximation to the true structure may depend on the values of the exogenous variables. There will be an apparent structural shift as between the base period and T+1. Fitting of curvilinear relations will not be too satisfactory; if a linear approximation is a good fit in the base period the estimates of the higher Taylor coefficients will probably be very unreliable. An alternative is to make use of a priori knowledge as to the nature of the curvilinearity by saying that if the apparent structure (the best linear approximation in the base period) were S the apparent structure in the forecast period,

S*, belongs to a certain class or model M_S, depending on S. The previous assumption, S* = f(S), where f is a known function, is included in this assumption as the special case where M_S contains just one structure for each value of S.

If T and T*-T', the numbers of observations where structures S and S*, respectively, are both large, the optimum procedure is clearly to estimate S from the base period by the maximum likelihood estimate $\hat{S}$, assuming the model is M, and then estimate S* by maximum likelihood, assuming the model is $M_{\hat{S}}$. But in practice T*-T' is very small, 2 or 3 in the present circumstances. Assuming still that T is large, we then have to consider the problem of small-sample estimation under $M_{\hat{S}}$. In practice, we would hope $M_{\hat{S}}$ will be a simple system, so that small-sample methods can really be applied.

4. An Example of Specification Bias

Consider the following simple model:

$$(4.1) \quad C = \beta_1 Y + \sum_{k=1}^{K} \beta_{1k} z_k + u_1,$$

$$(4.2) \quad I = \beta_2 Y + \sum_{k=1}^{K} \beta_{2k} z_k + u_2,$$

$$(4.3) \quad C + I = Y$$

Here C, I, Y have their usual economic meanings. Some of the β_{1k}'s may be zero a priori. We may estimate β_1 under two different false assumptions: (1) assume that I is exogenous instead of satisfying (4.2) and that the a priori restrictions on the β_{1k}'s are not used; in that case, find the conditional regression coefficient, α, of C on I given z, and estimate β_1 by $\beta_1' = \frac{\alpha}{1+\alpha}$; (2) assume that Y is exogenous instead of satisfying (4.2) and (4.3), in which case β_1

is estimated by β_1'', the conditional regression coefficient of C on Y, given z. I will compare the errors in estimating the marginal propensity to consume by the two procedures, ignoring sampling fluctuations.

$$C - E(C|z) = \frac{(1-\alpha_2)u_1 + \alpha_1 u_2}{1 - \alpha_1 - \alpha_2} ,$$

$$I - E(I|z) = \frac{(1-\alpha_1)\, u_2 + \alpha_2\, u_1}{1 - \alpha_1 - \alpha_2}$$

Let σ_1, σ_2 be the standard deviations of u_1, u_2, respectively, and ρ their correlation coefficient. Then, under procedure (1),

$$\beta_1' = \frac{\alpha_2 (1-\alpha_2)\sigma_1^2 + (1-\alpha_1-\alpha_2+2\alpha_1\alpha_2)\rho\sigma_1\sigma_2 + \alpha_1(1-\alpha_1)\sigma_2^2}{\alpha_2\sigma_1^2 + (1-\alpha_1+\alpha_2)\rho\sigma_1\sigma_2 + (1-\alpha_1)\sigma_2^2}$$

$$Y - E(Y|z) = \frac{u_1 + u_2}{1-\alpha_1-\alpha_2}$$

Then, $\beta_1'' = \dfrac{(1-\alpha_2)\sigma_1^2 + (1-\alpha_2+\alpha_1)\rho\sigma_1\sigma_2 + \alpha_1\sigma_2^2}{\sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2}$

Let the errors be $E_1' = \beta_1' - \beta_1$, $E_1'' = \beta_1'' - \beta_1$.

$$E_1' = \frac{N'}{D'} , \quad E_1'' = \frac{N''}{D''}$$

where $N' = (1-\alpha_1-\alpha_2)\left(\sigma_1^2 + \dfrac{1-\alpha_1}{\alpha_2}\rho\sigma_1\sigma_2\right)$,

$$D' = \sigma_1^2 + \left(\frac{1-\alpha_1}{\alpha_2} + 1\right)\rho\sigma_1\sigma_2 + \frac{1-\alpha_1}{\alpha_2}\sigma_2^2 ,$$

$$N'' = (1-\alpha_1-\alpha_2)(\sigma_1^2 + \rho\sigma_1\sigma_2) ,$$

$$D'' = \sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2 .$$

For $\alpha_1 + \alpha_2 < 1$, $\alpha_2 > 0$, $\rho \geq 0$, it follows that $N' > 0$, $D' > 0$, so that $\alpha_1' > \alpha_1$ (i.e., Haavelmo's method overestimates the marginal propensity to consume). Also,

$$\frac{D'}{D''} = 1 + \frac{(1-\alpha_1-\alpha_2)}{\alpha_2} \cdot \frac{\rho\sigma_1\sigma_2 + \sigma_2^2}{\sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2} ,$$

$$\frac{N'}{N''} = 1 + \frac{(1-\alpha_1-\alpha_2)}{\alpha_2} \quad \frac{\rho\sigma_1\sigma_2}{\sigma\sigma_1^2+\rho\sigma_1\sigma_2} \quad .$$

Under these assumptions, it can easily be shown that

$$\frac{\rho\sigma_1\sigma_2}{\sigma_1^2+\rho\sigma_1\sigma_2} \leq \frac{\rho\sigma_1\sigma_2 + \sigma_2^2}{\sigma_1^2+2\rho\sigma_1\sigma_2+\sigma_2^2}$$

and therefore $\frac{D'}{D''} \geq \frac{N'}{N''}$, the equality holding only if $\rho = 1$.

Hence, $E_1'' \geq E_1'$. Hence, while Haavelmo's method yields an overestimate of the marginal propensity to consume, least squares yields a still larger overestimate.

6. Forecasting as the Purpose of Fitting.

Suppose the model is constructed solely to forecast one variable, say y_1. In that case it seems reasonable that the fitting be designed to minimize the variance of forecast of y_1, as given by (2.11). This is, I believe, an application of the Wald minimax principle, where the decisions are forecasts, and the weight function is the squared error of the forecast. This criterion would supply a basis for small sample estimation. Note, however, that the form of (2.11) suggests that in general the optimum estimates of the parameters will depend on the values of the exogenous variables at the time of forecast.

The minimum forecast variance criterion appears to be a very flexible one in principle, and can be modified to take account of all the varieties of forecasting mentioned in this paper. Its actual application may be difficult, since it involves the solution of a calculus of variations problem.