BEST LINEAR UNBIASED ESTIMATION OF MISSING OBSERVATIONS IN AN ECONOMIC TIME SERIES

Gregory C. Chow

and

An-loh Lin

Econometric Research Program
Research Memorandum No. 173
February 1975

The research described in this paper was supported by NSF Grant GS43747X.

Econometric Research Program
PRINCETON UNIVERSITY
207 Dickinson Hall
Princeton, New Jersey

BEST LINEAR UNBIASED ESTIMATION OF MISSING OBSERVATIONS IN AN ECONOMIC TIME SERIES

Gregory C. Chow

and

An-loh Lin*

Abstract: The best linear unbiased estimator which we proposed previously for interpolating, distributing and extrapolating a time series by related series is applied to the problem of estimating missing observations. Under special assumptions, the problem reduces to the one treated by H. E. Doran. Our estimator is compared with his and is shown to be more efficient.

*Research by Gregory Chow is supported by the National Science Foundation Grant GS43747X. An-loh Lin is a member of the research staff at the National Bureau of Economic Research. Helpful discussion with Donald T. Sant is gratefully acknowledged.

1. INTRODUCTION

The purpose of this paper is two-fold. First, it is to show that the best linear unbiased estimator which we proposed previously [1] for interpolating, distributing and extrapolating a time series by the use of a regression on a set of related series has wider applicability than the examples explicitly cited in the paper. In particular, our method is applicable to the problem of estimating missing observations as treated by H. E. Doran [2]. Second, our method will be compared with Doran's method, and shown to be more efficient. This can explain why the latter method performed so poorly in the simulation experiments reported in Doran's paper.

The main problems cited in our paper [1] are the following. Given quarterly observations on a stock variable and monthly observations on some related time series, we wish to estimate the monthly observations of the first variable. This is the problem of interpolation. Given quarterly observations on a flow variable and monthly observations on some related time series, we wish to estimate the monthly observations of the first variable. This problem has been called distribution because the estimated monthly observations, on the Gross National Product for example, are supposed to sum to the observed quarterly figures. Each quarterly figure is supposed to be distributed into three monthly figures. If even quarterly observations on the variable concerned are unavailable for a certain time interval such as certain future time periods, the problem is one of extrapolation. Both a stock variable and a flow variable can be extrapolated. Our estimator applies to all these problems, and others as well.

The problem posed by Doran [2] is as follows. Given quarterly observations

on a (stock or flow) variable in an early period, and given its monthly observations in a later period, the problem is to estimate the missing monthly observations. Doran referred to our work [1], but stated that his problem is different because no related series are used. It turns out that our estimator not only applies to Doran's problem, with or without the use of a set of related series, but that it is more efficient under Doran's assumptions. We will point out the difference between these two estimators and show the relative efficiency of ours.

In section 2 we restate our method and the assumptions under which it can be applied. Section 3 applies our estimator to the problem of missing observations. Section 4 compares our estimator with Doran's and shows the efficiency of the former.

2. A METHOD OF BEST LINEAR UNBIASED ESTIMATION

It is assumed that during a sample period the monthly figures (not necessarily available) for a variable is governed by the linear model

$$y = X\beta + u \tag{2.1}$$

where y is column vector of T observations on the dependent variable to be estimated, X is a $T \times p$ matrix of observations on a set of p related series, and u is generated by a covariance-stationary stochastic process uncorrelated with the observations X and has mean zero and covariance matrix V.

The observed series $y_{\cdot}$ is assumed to be Cy, with C denoting a given linear transformation. For the problem of interpolating a stock variable, the matrix C is

$$C_I = \begin{bmatrix} 1\ 0\ 0\ 0 & \cdots \\ 0\ 0\ 0\ 1\ 0 & \cdots \\ \cdots & \cdots \\ 0\ \cdots & 1\ 0\ 0 \end{bmatrix} \tag{2.2}$$

For the problem of distributing a flow variable, it is

$$C_D = 1/3 \begin{bmatrix} 1\ 1\ 1\ 0\ .\ .\ . & & & & & & 0 \\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ .\ .\ . & & & & & & 0 \\ & . & . & . & & & \\ 0 & . & . & . & & 1\ 1\ 1 & \end{bmatrix} \tag{2.3}$$

where the factor $1/3$ is used if both the monthly and the quarterly series are expressed at annual rates. The matrix C is certainly not restricted to the form (2.2) or (2.3). Our estimator is applicable to other situations in which we observe some known linear transformation $y_* = Cy$ of the time series y. Premultiplying (2.1) by C, we find the observed series to satisfy

$$y_* = CX\beta + Cu = X_*\beta + u_* \tag{2.4}$$

where $X_* = CX$ and $u_* = Cu$.

The monthly series to be estimated is denoted by z, which satisfies

$$z = X_z\beta + u_z \tag{2.5}$$

with X_z and u_z respectively denoting the related series and the random residuals corresponding to z. The new symbol z is introduced because we may be dealing with the problem of extrapolation outside the sample period. For the problems of interpolation and distribution using the transformations (2.2) and (2.3), z is identical to y. For the problem of missing observations to be studied in section 3, z is a subvector of y.

In our previous paper, the linear estimator $\hat{z} = Ay$. which is unbiased,

with $E(\hat{z}-z) = 0$, and which minimizes the trace of the covariance matrix $Cov(\hat{z}-z)$ is found to be

$$\hat{z} = X_z\hat{\beta} + [V_{z.}V_{..}^{-1}]\hat{u}_. \tag{2.6}$$

where

$$V_{..} = Eu_.u_.' = CVC' \ , \tag{2.7}$$

$$V_{z.} = Eu_zu_.' \ , \tag{2.8}$$

$$\hat{\beta} = (X_.'V_{..}^{-1}X_.)^{-1}X_.V_{..}^{-1}y_. = \beta + (X_.'V_{..}^{-1}X_.)^{-1}X_.'V_{..}^{-1}u_. \ , \tag{2.9}$$

and

$$\hat{u}_. = y_. - X_.\hat{\beta} = [I - X_.(X_.'V_{..}^{-1}X_.)^{-1}X_.'V_{..}^{-1}]u_. \ . \tag{2.10}$$

The estimator (2.6) consists of two parts. The first results from applying the generalized least squares estimator $\hat{\beta}$ obtained from the regression model (2.4) to the related series X_z associated with the vector z to be estimated. The second is an estimate of the residual u_z associated with z. It amounts to applying the coefficients $[V_{z.}V_{..}^{-1}] = (Eu_zu_.')(Eu_.u_.')^{-1}$ in the multivariate regressions of u_z on $u_.$ to the estimated residuals $\hat{u}_.$ in the regression model (2.4). The estimator (2.6) requires using the covariance matrix of the regression residuals. We suggested imposing some autoregressive structure to the process generating u and estimating the

structure by using the observed residuals $\hat{u}_\cdot$. The estimate of the covariance matrix $V_{\cdot\cdot}$ can be used in (2.10) to obtain a new set of residuals $\hat{u}_\cdot$, and the process can be iterated.

Using (2.5), (2.6) and the second equalities of (2.9) and (2.10), one finds the vector of estimation errors to be

$$\hat{z} - z = [(X_z - V_{z\cdot}V_{\cdot\cdot}^{-1}X_\cdot)(X_\cdot'V_{\cdot\cdot}^{-1}X_\cdot)^{-1}X_\cdot' + V_{z\cdot}]V_{\cdot\cdot}^{-1}u_\cdot - u_z \qquad (2.11)$$

From (2.11), the covariance matrix of estimation errors follows.

$$\begin{aligned} \mathrm{Cov}(\hat{z}-z) = {} & (X_z - V_{z\cdot}V_{\cdot\cdot}^{-1}X_\cdot)(X_\cdot'V_{\cdot\cdot}^{-1}X_\cdot)^{-1}(X_z' - X_\cdot'V_{\cdot\cdot}^{-1}V_{\cdot z}) \\ & + (V_{zz} - V_{z\cdot}V_{\cdot\cdot}^{-1}V_{\cdot z}) \end{aligned} \qquad (2.12)$$

The first component of the error covariance matrix is due to the error in estimating $X_z\beta$ by $X_z\hat{\beta}$. The second component is due to the error in estimating u_z by regression on $u_\cdot$, with V_{zz} denoting Eu_zu_z' .

The estimator (2.6) is fairly general. It is applicable whenever the assumptions underlying the regression models (2.1) and (2.5) are approximately valid and there exists a set of observations $y_\cdot = Cy$ which is a given linear transformation of y . The related series X may consist of dummy variables, lagged values or future values of some related variables, and trend variables. It may merely consist of the single dummy variable 1, which serves to estimate the mean of the series.

3. APPLICATION TO THE PROBLEM OF MISSING OBSERVATIONS

In the problem of missing observations posed by Doran [2], monthly observations on the time series are available in a later period but only quarterly observations are available in an earlier period. In the framework of our estimator, the observed series can be written as

$$y_{\cdot} = Cy = \begin{bmatrix} R & O \\ O & I \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \simeq \begin{bmatrix} y_{1\cdot} \\ y_2 \end{bmatrix} \tag{3.1}$$

where the matrix R has the form (2.2) for a stock variable and the form (2.3) for a flow variable. By choosing the matrix C as given in (3.1), our estimator applies to his problem.

However, Doran [2] prefers to deal with the problem when no related series are allowed. We consider this to be a highly artificial situation because there are likely to be some related series, including dummy variables, which can usefully serve as regressors. The least that one should do is to use a single dummy variable identically equal to one; its coefficient gives the mean of the time series. Doran assumes that the mean of the series is known to be zero, so that even the regression intercept is dispensed with. Although we believe that the assumptions of having no related series and of zero mean for the series to be estimated are not very useful, we will consider our estimator under these special assumptions and show that it will have a smaller error covariance matrix than the estimator proposed by Doran.

When the only regressor is the dummy variable 1, the estimator (2.6) becomes

$$\hat{z} = \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} \hat{\beta} + [V_{z.} V_{..}^{-1}] \hat{u}_{.} \tag{3.2}$$

where the estimated mean is

$$\hat{\beta} = \left(\sum_{ij} v_{..}^{ij} \right)^{-1} \left(\sum_{ij} v_{..}^{ij} \, y_{.j} \right) \tag{3.3}$$

with $v_{..}^{ij}$ denoting the i-j element of $V_{..}^{-1}$, and the estimated vector of residuals is

$$\hat{u}_{.} = y_{.} - \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} \hat{\beta} \quad . \tag{3.4}$$

When even the mean is assumed to be zero, the vector β in (2.1) is a zero vector, y is identical with u, and $y_{.}$ and $u_{.}$ are also the same. For the problem of missing observations defined by (3.1), our estimator is reduced to

$$\hat{z} = [V_{z.} V_{..}^{-1}] y_{.} \tag{3.5}$$

or

$$\hat{y}_1 = [Ey_1y_{1.}' \quad Ey_1y_2] \begin{bmatrix} Ey_{1.}y_{1.}' & Ey_{1.}y_2' \\ Ey_2y_{1.}' & Ey_2y_2' \end{bmatrix}^{-1} \begin{bmatrix} y_{1.} \\ y_2 \end{bmatrix} \tag{3.6}$$

$$= [V_{11}R' \quad V_{12}] \begin{bmatrix} RV_{11}R' & RV_{12} \\ V_{21}R' & V_{22} \end{bmatrix}^{-1} \begin{bmatrix} Ry_1 \\ y_2 \end{bmatrix}$$

where V_{ij} denotes Ey_iy_j' $(i,j = 1,2)$.

Utilizing the well-known partitioned inverse

$$\begin{bmatrix} RV_{11}R' & RV_{12} \\ V_{12}R' & V_{22} \end{bmatrix}^{-1} = \begin{bmatrix} (RV_{1\cdot2}R')^{-1} & -(RV_{1\cdot2}R')^{-1}RV_{12}V_{22}^{-1} \\ -V_{22}^{-1}V_{21}R'(RV_{1\cdot2}R')^{-1} & V_{22}^{-1} + V_{22}^{-1}V_{21}R'(RV_{1\cdot2}R')^{-1}RV_{12}V_{22}^{-1} \end{bmatrix} . \tag{3.7}$$

where $V_{1\cdot2}$ denotes the covariance matrix of the residuals of the regressions of y_1 on y_2 , i.e.,

$$V_{1\cdot2} = V_{11} - V_{12}V_{22}^{-1}V_{21} \quad , \tag{3.8}$$

we rewrite the estimator (3.6) as

$$\hat{y}_1 = V_{1\cdot2}R'(RV_{1\cdot2}R')^{-1}y_{1.} + [I - V_{1\cdot2}R'(RV_{1\cdot2}R')^{-1}R]V_{12}V_{22}^{-1}y_2 \quad . \tag{3.9}$$

From (3.9) we obtain the covariance matrix of estimation errors.

$$\mathrm{Cov}(\hat{y}_1-y_1) = [I - V_{1\cdot 2}R'(RV_{1\cdot 2}R')^{-1}R]\, V_{1\cdot 2}\, [I - R'(RV_{1\cdot 2}R')^{-1}RV_{1\cdot 2}] \qquad (3.10)$$

$$= V_{1\cdot 2} - V_{1\cdot 2}R'(RV_{1\cdot 2}R')^{-1}RV_{1\cdot 2} \quad .$$

The expression (3.10) could also be derived from the second component of (2.12) using the partitioned inverse (3.7).

We conclude this section by showing that the error covariance matrix (3.10) of our estimator (3.9) is smaller than that of any other estimator which is a linear combination of y_1. and y_2 . Let an alternative estimator be written as

$$y_1^* = [V_{1\cdot 2}R'(RV_{1\cdot 2}R')^{-1} + B_1]Ry_1 + \{[I - V_{1\cdot 2}R'(RV_{1\cdot 2}R')^{-1}R]V_{12}V_{22}^{-1} + B_2\}y_2 \qquad (3.11)$$

Its error covariance matrix is

$$E(y_1^*-y_1)(y_1^*-y_1)' \qquad (3.12)$$

$$= \{[I - V_{1\cdot 2}R'(RV_{1\cdot 2}R')^{-1}R][-I \quad V_{12}V_{22}^{-1}] + [B_1R \quad B_2]\} \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \times$$

$$\{[-I \quad V_{12}V_{22}^{-1}]'[I - R'(RV_{1\cdot 2}R')^{-1}RV_{1\cdot 2}] + [B_1R \quad B_2]'\}$$

$$= \mathrm{Cov}(\hat{y}_1-y_1) + [B_1R \quad B_2] \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \begin{bmatrix} R'B_1' \\ B_2' \end{bmatrix}$$

$$- [I - V_{1\cdot 2}R'(RV_{1\cdot 2}R')^{-1}R]V_{1\cdot 2}R'B_1' - B_1RV_{1\cdot 2}[I - R'(RV_{1\cdot 2}R')^{-1}RV_{1\cdot 2}] \quad .$$

Since the last two terms of (3.12) are zero, the error covariance matrix of the alternative estimator exceeds $\mathrm{Cov}(\hat{y}_1 - y_1)$ by a positive definite matrix. Although our estimator was derived from minimizing the trace of $\mathrm{Cov}(\hat{y}_1 - y_1)$, it has the smallest covariance matrix among all linear functions of $y_{1\cdot}$ and y_2 .

4. COMPARISON WITH DORAN'S ESTIMATOR

In this section, the estimator of section 3 will be compared with the one proposed by Doran [2] in terms of the method of derivation and of efficiency. We seek an unbiased estimator $\hat{y}_1$ which is a linear function of both $y_{1\cdot}$ and y_2 and has the smallest expected sum of squared errors. By contrast, Doran tries to find a linear function y_1^* of y_2 alone which will have the smallest expected sum of squared errors, subject to the restriction that $Ry_1^* = Ry_1 = y_{1\cdot}$. (In Doran's notations, x and y stand for our y_1 and y_2 respectively, his estimator being denoted by $\hat{x}$, with $R\hat{x} = Rx = c$.)

Formally, under the assumptions of having no regressors and of zero mean for y , we seek $\hat{y}_1 = Ay_{\cdot} = A_1y_{1\cdot} + A_2y_2$ which minimizes $\mathrm{tr}\, E(\hat{y}_1 - y_1)(\hat{y}_1 - y_1)'$. The outcome of this minimization is the estimator (3.9). Doran also minimizes $\mathrm{tr}\, E(y_1^* - y_1)(y_1^* - y_1)'$ but confines his estimator y_1^* to be a linear function of y_2 alone, subject to the restriction $Ry_1^* = c$. The resulting estimator,

given by his equation (2.9), is

$$y_1^* = R'(RR')^{-1}y_{1\cdot} + [I - R'(RR')^{-1}R]V_{12}V_{22}^{-1}y_2 \quad . \tag{4.1}$$

Using (4.1) we find the covariance matrix of estimation errors to be

$$\mathrm{Cov}(y_1^*-y_1) = [I - R'(RR')^{-1}R]V_{1\cdot 2}[I - R'(RR')^{-1}R] \tag{4.2}$$

As we have shown at the end of section 3, our estimator has the smallest error covariance matrix among all estimators which are linear combinations of $y_{1\cdot}$ and y_2 . Since Doran's estimator (4.1) turns out also to be a linear combination of $y_{1\cdot}$ and y_2 , it must have a larger error covariance matrix. By (3.12), this covariance matrix exceeds ours by

$$[B_1R \quad B_2] \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \begin{bmatrix} R'B_1' \\ B_2' \end{bmatrix} \tag{4.3}$$

where, by comparison of (3.9) and (4.1),

$$B_1 = R'(RR')^{-1} - V_{1\cdot 2}R'(RV_{1\cdot 2}R')^{-1} \tag{4.4}$$

$$B_2 = [V_{1\cdot 2}R'(RV_{1\cdot 2}R')^{-1}R - R'(RR')^{-1}R]V_{12}V_{22}^{-1} \quad .$$

Therefore, the excess (4.3) of Doran's error covariance matrix is

$$\mathrm{Cov}(y_1^*-y_1) - \mathrm{Cov}(\hat{y}_1-y_1) \tag{4.5}$$

$$= [V_{1\cdot 2}R'(RV_{1\cdot 2}R')^{-1}R - R'(RR')^{-1}R]V_{1\cdot 2}[R'(RV_{1\cdot 2}R')^{-1}RV_{1\cdot 2} - R'(RR')^{-1}R]$$

$$= V_{1\cdot 2}R'(RV_{1\cdot 2}R')^{-1}RV_{1\cdot 2} + R'(RR')^{-1}RV_{1\cdot 2}R'(RR')^{-1}R$$

$$- V_{1\cdot 2}R'(RR')^{-1}R - R'(RR')^{-1}RV_{1\cdot 2}$$

The expression after the second equality sign of (4.5) could have been obtained by subtracting (3.10) from (4.2). The expression after the first equality sign of (4.5) was obtained by rewriting the positive definite matrix (4.3). (4.5) is zero if and only if R is an identity matrix (in which case the problem of missing observations does not arise), $V_{1\cdot 2}$ not being an identity matrix. (4.5) is also zero if and only if $V_{1\cdot 2} = I$, provided $R \neq I$.

The above theoretical analysis may explain why Doran's estimator performs poorly as compared with some simple estimators according to the sampling experiments reported in his paper [2]. As it was stated explicitly in our paper [1], we chose not to constrain our monthly estimates to be equal to the observed monthly data every three months in the case of interpolation, or to sum to the observed quarterly data in the case of distribution. Rather we tried to find a best linear function of all the observed data in the sense of having a small error covariance matrix. It was shown that the resulting estimator satisfies the above constraints. By premultiplying (3.9) by R, we can easily see that the constraint $R\hat{y}_1 = y_1$. is satisfied. By imposing this constraint but considering a linear function of only a subset y_2 of the observed data, one loses efficiency as we have described.

Both papers [1] and [2] contain methods for dealing with the unknown covariance matrix V, developed from treating the data in the time domain or the frequency domain. These will not be repeated here.

REFERENCES

[1] Chow, G. C. and Lin, A., "Best Linear Unbiased Interpolation, Distribution and Extrapolation of Time Series by Related Series", The Review of Economics and Statistics, 53 (November 1971), 372-375.

[2] Doran, H. E., "Prediction of Missing Observations in the Time Series of an Economic Variable", Journal of the American Statistical Association, 69 (June 1974), 546-554.