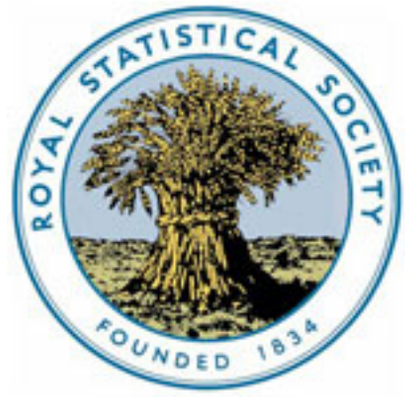

Estimation of Parameters in Time-series Regression Models

By J. DURBIN

Research Techniques Division, London School of Economics

[Received August, 1959]

SUMMARY

We consider the estimation of the coefficients in a general linear regression model in which some of the explanatory variables are lagged values of the dependent variable. For discussing optimum properties the concept of best unbiased linear estimating equations is developed. It is shown that when the errors are normally distributed the method of least squares leads to optimum estimates. The properties of the least-squares estimates are shown to be the same asymptotically as those of the least-squares coefficients of ordinary regression models containing no lagged variables, whether or not the errors are normally distributed. Finally, a method of estimation is proposed for a different model which has no lagged dependent variables but in which the errors have an autoregressive structure. The method is shown to be efficient in large samples.

1. INTRODUCTION

In many fields of study the phenomena under investigation can be represented by a regression model of the form

$$y_t+\alpha_1 y_{t-1}+\ldots+\alpha_p y_{t-p}=\beta_1 x_{1t}+\ldots+\beta_q x_{qt}+\epsilon_t \quad (t=1,\ldots,n), \tag{1}$$

where $\epsilon_1, \epsilon_2, \ldots$ is a series of independently and identically distributed random variables with zero mean and variance σ^2, and $x_{11}, x_{12}, \ldots, x_{q1}, x_{q2}, \ldots$ are either given series of constants or are random variables whose joint distribution is independent of $\epsilon_1, \epsilon_2, \ldots$. This is a generalization both of the ordinary regression model,

$$y_t=\beta_1 x_{1t}+\ldots+\beta_q x_{qt}+\epsilon_t,$$

and of the autoregression model,

$$y_t+\alpha_1 y_{t-1}+\ldots+\alpha_p y_{t-p}+\alpha_0=\epsilon_t.$$

In this paper we shall regard $y_0, y_{-1}, \ldots, y_{-p+1}$ as given numbers; alternatively, if $y_0, \ldots, y_{-p+1}$ are random variables, we shall consider only inferences conditionally on $y_0, \ldots, y_{-p+1}$ held fixed. In many applications this restriction is entirely appropriate. In others we lose the small amount of information about the parameters that can be derived from the distribution of $y_0, \ldots, y_{-p+1}$. Asymptotically there is no difference between the two cases.

The autoregression model $y_t+\alpha_1 y_{t-1}+\ldots+\alpha_p y_{t-p}+\alpha_0=\epsilon_t$ was studied by Mann and Wald (1943) who showed that ordinary least-squares theory is valid asymptotically. Simultaneous-equation models containing lagged y's and fixed x's were considered by Koopmans et al. (1950) and by Anderson and Rubin (1950) and results analogous to those of Mann and Wald were obtained; however, the properties of the single-equation model (1) cannot be deduced from the treatment given by these

authors. Moreover, their discussion of asymptotic behaviour depends on work by Rubin which was never published. It should be noted that the only explicit statement of Rubin's theorem (Koopmans et al., 1950, p. 135, Theorem 3·3, Equation 4) is incorrect unless a further condition concerning the limiting positive-definiteness of M_{xx} is met. Anderson and Rubin (1950) include such a condition but do not state Rubin's theorem.

The difficulty in handling the sampling distributions of the least-squares estimators of the parameters in (1) arises from the fact that the coefficients of the unknowns in the normal equations are random variables, whereas in ordinary least-squares theory they are constants. As a result the ordinary theory of linear estimation cannot be applied. However, the equations themselves can be regarded as unbiased in a sense to be explained later. Consequently, it seems reasonable to develop the idea of unbiased estimating equations with minimum variance. This is done in sections 2 and 3 of the paper where lower bounds similar to Cramér–Rao lower bounds for the variance of unbiased estimators are derived. It is shown in section 4 that when the errors are normally distributed these lower bounds are attained by the least-squares equations. In sections 5 and 6 the distribution of the estimates for general error distributions is considered and it is shown that for many cases ordinary least-squares regression theory applies asymptotically.

Finally, in section 7 a different type of model is considered in which, while lagged values of the dependent variable do not appear as explanatory variables, the errors in the model are of autoregressive form. A two-stage estimation procedure is proposed which leads to efficient estimates.

2. Best Unbiased Estimating Equations for a Single Parameter

The difference between ordinary regression models and models containing lagged y's can be illustrated by considering the simplest cases of both types, namely

$$y_t = \beta x_t + \epsilon_t \quad (t = 1, \ldots, n), \tag{2}$$

where $x_1, \ldots, x_n$ are constants, and

$$y_t + \alpha y_{t-1} = \epsilon_t \quad (t = 1, \ldots, n), \tag{3}$$

where y_0 is constant. In both cases $\epsilon_1, \ldots, \epsilon_n$ are taken to be independently and identically distributed with zero mean and variance σ^2.

The least-squares estimators of β and α are

$$b = \sum_{t=1}^{n} x_t y_t \Big/ \sum_{t=1}^{n} x_t^2 \quad \text{and} \quad a = -\sum_{t=1}^{n} y_t y_{t-1} \Big/ \sum_{t=1}^{n} y_{t-1}^2$$

respectively. Although they are similar in form the sampling behaviour of these estimates is very different, at least in small samples. By the Gauss–Markoff theorem, b is the unbiased linear estimator with minimum variance. In contrast a is not even unbiased and little is known about its small-sample properties. The difference arises from the fact that whereas b is a linear function of the y's and hence easy to deal with, a is a ratio of quadratic forms which is much less tractable. The "best linear unbiased" type of optimality criterion, which is so suitable to b, is clearly inappropriate to a. Nevertheless, there is nothing sacrosanct about such criteria and one has the intuitive feeling that there ought to exist a reasonable optimality criterion for the estimation of α which is satisfied by a or by something closely related to it.

Such a criterion may be developed by considering the estimating equation

$$a\sum_{t=1}^{n} y_{t-1}^2+\sum_{t=1}^{n} y_t y_{t-1}=0 \tag{4}$$

from which a is derived. On replacing a by α, the left-hand side becomes

$$\alpha\sum_{t=1}^{n} y_{t-1}^2+\sum_{t=1}^{n} y_t y_{t-1}$$

which has expected value zero. It therefore seems reasonable to call (4) an unbiased estimating equation, a concept introduced by Kendall (1951). Noting that (4) is linear in a, we are led to the following general definition.

Definition 1. *Suppose that the estimator a of a parameter α is given by the linear equation*

$$T_1 a+T_2=0, \tag{5}$$

where T_1 and T_2 are functions of the observations such that T_2/T_1 is independent of unknown parameters, and where

$$E(T_1\alpha+T_2)=0. \tag{6}$$

Then (5) *is called an unbiased linear estimating equation.*

The word linear here means linear in a, not linear in the observations. Note that this definition includes the ordinary notion of an unbiased estimator since if $T_1\equiv 1$ and (6) is satisfied, a is an unbiased estimator of α. The quantity T_1 is assumed to be non-zero with probability one. In applications, T_1 and T_2 will normally be statistics, i.e. functions of the observations only; however, it is inconvenient, for reasons that will soon be apparent, to lay this down as a requirement in the definition.

It may be remarked that a number of arguments frequently advocated in favour of the use of unbiased estimates apply, with suitable modification, to unbiased estimating equations. For example, if we wish to combine several sets of data we can do so by adding the values of T_1 and T_2 given by the different sets and solving for a at the end, rather than by taking the mean of a number of values of a. By proceeding in this way, the bias of the estimate goes down as the number of sets of observations increases.

We now require the analogue of the minimum variance of an unbiased estimator. It is clearly not enough to lay down that the variance of $T_1\alpha+T_2$ should be a minimum since (5) may be multiplied through by an arbitrary constant without affecting the value of a. It seems reasonable to standardize by dividing through by $E(T_1)$. We are therefore led to the following definition.

Definition 2. *Suppose that $t_1 a+t_2=0$ is an unbiased linear estimating equation where $E(t_1)=1$ and*

$$V(t_1\alpha+t_2)\leqslant V(t_1'\alpha+t_2'), \tag{7}$$

for all other unbiased linear estimating equations $t_1' a+t_2'=0$ having $E(t_1')=1$. Then $t_1 a+t_2=0$ is called a best unbiased linear estimating equation.

This definition includes the notion of a minimum-variance unbiased estimator since if $t_1\equiv t_1'\equiv 1$ and (7) is satisfied then a is a minimum-variance unbiased estimator of α. In other cases a consideration of $V(t_1\alpha+t_2)$ often provides a convenient approach to the study of the asymptotic efficiency of a.

We now derive a lower bound to the variance of $t_1\alpha+t_2$ by a slight variant of the derivation of the Cramér–Rao lower bound to the variance of an unbiased estimator given by Rao (1952, p. 130).

Suppose that $T_1 a+T_2=0$ is an unbiased estimating equation, where T_1 and T_2 depend only on the observations, and that the sample density is $\phi(y_1,\ldots,y_n;\alpha)$. From (6) we have

$$\int(T_1\alpha+T_2)\phi\,dy=0, \tag{8}$$

where dy stands for $dy_1\ldots dy_n$ and $\int$ denotes the multiple integral. If the conditions for differentiating under the integral sign are satisfied we have, on differentiating with respect to α and putting $t_1=T_1/E(T_1)$, $t_2=T_2/E(T_1)$,

$$\int t_1\phi\,dy+\int(t_1\alpha+t_2)\frac{\partial\phi}{\partial\alpha}dy=0.$$

Since $E(t_1)=\int t_1\phi\,dy=1$, we have

$$\int(t_1\alpha+t_2)\frac{\partial\log\phi}{\partial\alpha}\phi\,dy=-1.$$

The application of Schwarz's inequality gives

$$E(t_1\alpha+t_2)^2E\left(\frac{\partial\log\phi}{\partial\alpha}\right)^2\geqslant\left[\int(t_1\alpha+t_2)\frac{\partial\log\phi}{\partial\alpha}\phi\,dy\right]^2=1,$$

so that

$$V(t_1\alpha+t_2)\geqslant\frac{1}{E(\partial\log\phi/\partial\alpha)^2}=-\frac{1}{E(\partial^2\log\phi/\partial\alpha^2)}. \tag{9}$$

Note that the derivation can easily be extended to cover discrete distributions by replacing the integrals by sums.

If $t_1\equiv 1$ then $-t_2$ is an unbiased estimator and (9) gives the Cramér–Rao inequality

$$E(t_2+\alpha)^2\geqslant\frac{1}{E(\partial\log\phi/\partial\alpha)^2}.$$

Suppose that t_1 is not identically equal to one but converges stochastically to one as $n\to\infty$. Suppose also that δ_n is a function of n such that

$$E\left(\frac{\partial\log\phi}{\partial\alpha}\right)^2=O(\delta_n^2).$$

Then since $t_1(a-\alpha)=-(t_1\alpha+t_2)$, if the asymptotic distribution of $\delta_n(a-\alpha)$ exists it has mean zero and limiting minimum variance given by

$$\lim_{n\to\infty}\frac{\delta_n^2}{E(\partial\log\phi/\partial\alpha)^2}.$$

This follows from a theorem given by Cramér (1946a, p. 254). It will usually be the case that $\delta_n=\sqrt{n}$, though in multi-parameter problems of the kind considered later other types of functions will be found necessary.

To illustrate, let us return to the model (3) and let us make the further assumption that $\epsilon_1, \ldots, \epsilon_n$ are normal with unit variance. The density is

$$\phi = \frac{1}{(2\pi)^{\frac{1}{2}n}} \exp\left[-\tfrac{1}{2}\sum_{t=1}^{n}(y_t+\alpha y_{t-1})^2\right].$$

Maximum likelihood gives the same estimating equation as least squares, namely

$$a\sum_{t=1}^{n} y_{t-1}^2 + \sum_{t=1}^{n} y_t y_{t-1} = 0,$$

which is linear and unbiased. We find

$$-\frac{\partial^2 \log \phi}{\partial \alpha^2} = \sum_{t=1}^{n} y_{t-1}^2.$$

Putting $t_1 = \Sigma y_{t-1}^2 / E(\Sigma y_{t-1}^2)$ and $t_2 = \Sigma y_t y_{t-1} / E\Sigma y_{t-1}^2$, we have from (9),

$$V(t_1\alpha + t_2) \geqslant \frac{1}{E(\Sigma y_{t-1}^2)}.$$

But
$$t_1\alpha + t_2 = -\frac{\partial \log \phi}{\partial \alpha} \Big/ E(\Sigma y_{t-1}^2).$$

Thus
$$V(t_1\alpha + t_2) = \frac{E(\partial \log \phi/\partial\alpha)^2}{[E(\Sigma y_{t-1}^2)]^2} = \frac{1}{E(\Sigma y_{t-1}^2)}.$$

Consequently the lower bound is actually attained in this case.

The above results are exact for samples of any size regardless of whether or not $|\alpha| < 1$. In order to consider the behaviour of the estimator a itself we resort to asymptotic theory for which the assumption $|\alpha| < 1$ is convenient, though not perhaps essential. We then have

$$\frac{1}{n} E\left(\sum_{t=1}^{n} y_{t-1}^2\right) \to \frac{1}{1-\alpha^2},$$

so that in the limit $(a-\alpha)\sqrt{n}$ has zero mean and variance $1-\alpha^2$. The value $1-\alpha^2$ is the minimum possible, i.e. a is an asymptotically efficient estimator.

3. Best Unbiased Estimating Equations for Several Parameters

Suppose we wish to estimate the vector $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_k\}$ of k parameters and that the vector of estimates $\mathbf{b}$ is obtained as the solution of the linear equations

$$\mathbf{T}_1\mathbf{b} + \mathbf{T}_2 = \mathbf{0}, \tag{10}$$

where $\mathbf{T}_1$ is a $k \times k$ matrix and $\mathbf{T}_2$ is a $k \times 1$ vector depending on the observations, and where $\mathbf{0}$ is a vector of k zeros. We assume that $\mathbf{T}_1$ is non-singular with probability one and that $\mathbf{T}_1^{-1}\mathbf{T}_2$ is independent of unknown parameters.

Definition 3. *If* $E(\mathbf{T}_1\boldsymbol{\beta} + \mathbf{T}_2) = \mathbf{0}$, *the set of equations* (10) *is called a set of unbiased linear estimating equations.*

Since (10) may be multiplied on the left by an arbitrary $k \times k$ matrix and still be satisfied, let us standardize by multiplying through by $[E(\mathbf{T}_1)]^{-1}$. Putting $\mathbf{t}_1 = [E(\mathbf{T}_1)]^{-1}\mathbf{T}_1$ and $\mathbf{t}_2 = [E(\mathbf{T}_1)]^{-1}\mathbf{T}_2$, then $E(\mathbf{t}_1) = \mathbf{I}$, the unit $k \times k$ matrix. Let the

variance matrix of a vector $\mathbf{z}$ be denoted by $V(\mathbf{z})$. The concept of best unbiased linear estimating equations is then defined as follows.

Definition 4. *Suppose* $\mathbf{t}_1\mathbf{b}+\mathbf{t}_2=\mathbf{0}$ *is a set of unbiased linear estimating equations with* $E(\mathbf{t}_1)=\mathbf{I}$ *such that the matrix*

$$V(\mathbf{t}_1\boldsymbol{\beta}+\mathbf{t}_2)-V(\mathbf{t}_1^*\boldsymbol{\beta}+\mathbf{t}_2^*) \tag{11}$$

is negative definite or semi-definite, $\mathbf{t}_1^*\mathbf{b}+\mathbf{t}_2^*=0$ *being any other set of linear estimating equations having* $E(\mathbf{t}_1^*)=\mathbf{I}$. *Then the equations* $\mathbf{t}_1\mathbf{b}+\mathbf{t}_2=\mathbf{0}$ *are called a set of best unbiased linear estimating equations.*

This is a natural extension of Definition 2. The implication of (11) is that the variance of any given linear function of the elements of $\mathbf{t}_1\boldsymbol{\beta}+\mathbf{t}_2$ given by a best set of equations is never greater than the variance of the same linear function of the elements given by any other unbiased set $\mathbf{t}_1^*\mathbf{b}+\mathbf{t}_2^*=\mathbf{0}$ having $E(\mathbf{t}_1^*)=\mathbf{I}$.

We now obtain a matrix lower bound to $V(\mathbf{t}_1\boldsymbol{\beta}+\mathbf{t}_2)$ analogous to the matrix form of the Cramér–Rao lower bound (see, e.g., Rao, 1952, p. 144). This will be called the minimal variance matrix.

Suppose that $\mathbf{T}_1$ and $\mathbf{T}_2$ depend only on the observations and that the sample density is $\phi(y_1,\ldots,y_n;\beta_1,\ldots,\beta_k)$. Let $\partial\log\phi/\partial\boldsymbol{\beta}'$ denote the row vector

$$\left[\frac{\partial\log\phi}{\partial\beta_1},\ldots,\frac{\partial\log\phi}{\partial\beta_k}\right],$$

$\partial\log\phi/\partial\boldsymbol{\beta}$ the transpose of $\partial\log\phi/\partial\boldsymbol{\beta}'$, and $\partial^2\log\phi/\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'$ the matrix $[\partial^2\log\phi/\partial\beta_i\,\partial\beta_j]$. Since $E(\mathbf{T}_1\boldsymbol{\beta}+\mathbf{T}_2)=\mathbf{0}$, we have

$$\int(\mathbf{T}_1\boldsymbol{\beta}+\mathbf{T}_2)\,\phi\,dy=\mathbf{0}.$$

Assuming the conditions for differentiating under the integral sign to be satisfied, we have, on differentiating with respect to $\boldsymbol{\alpha}$ and multiplying through by $[E(\mathbf{T}_1)]^{-1}$,

$$\int\mathbf{t}_1\phi\,dy+\int(\mathbf{t}_1\boldsymbol{\beta}+\mathbf{t}_2)\frac{\partial\log\phi}{\partial\boldsymbol{\beta}'}\phi\,dy=\mathbf{0},$$

i.e.

$$\int(\mathbf{t}_1\boldsymbol{\beta}+\mathbf{t}_2)\frac{\partial\log\phi}{\partial\boldsymbol{\beta}'}\phi\,dy=-\mathbf{I}.$$

On using a generalized form of Schwarz's inequality due to Cramér (1946b, p. 89) we have

$$\int(\mathbf{t}_1\boldsymbol{\beta}+\mathbf{t}_2)(\mathbf{t}_1\boldsymbol{\beta}+\mathbf{t}_2)'\,\phi\,dy-\left[\int\frac{\partial\log\phi}{\partial\boldsymbol{\beta}}\frac{\partial\log\phi}{\partial\boldsymbol{\beta}'}\phi\,dy\right]^{-1}$$

is positive definite or semi-definite, i.e.

$$V(\mathbf{t}_1\boldsymbol{\beta}+\mathbf{t}_2)-\mathscr{I}^{-1}$$

is positive definite or semi-definite, where $\mathscr{I}$ is the information matrix

$$E\left(\frac{\partial\log\phi}{\partial\boldsymbol{\beta}}\frac{\partial\log\phi}{\partial\boldsymbol{\beta}'}\right).$$

Thus $\mathscr{I}^{-1}$ is the minimal variance matrix of the equations $\mathbf{t}_1\mathbf{b}+\mathbf{t}_2=0$. On differentiating the relation $E(\partial\log\phi/\partial\boldsymbol{\beta})=0$ we have the alternative form $\mathscr{I}=-E(\partial^2\log\phi/\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}')$.

Let us now specialize to the case in which $\log\phi$ is a quadratic function of $\beta_1,\ldots,\beta_k$. Then $\partial\log\phi/\partial\boldsymbol{\beta}$ is linear in $\boldsymbol{\beta}$ and has the form $\mathbf{T}_1\boldsymbol{\beta}+\mathbf{T}_2$, where $\mathbf{T}_1$ is a matrix and $\mathbf{T}_2$ a vector depending only on the observations. The maximum likelihood equations are

$$\mathbf{T}_1\boldsymbol{\beta}+\mathbf{T}_2\Big|_{\boldsymbol{\beta}=\mathbf{b}}=\mathbf{0}. \tag{12}$$

Since for any ϕ, $E(\partial\log\phi/\partial\boldsymbol{\beta})=0$, it follows immediately that the maximum likelihood equations are unbiased linear estimating equations. Furthermore, $\partial^2\log\phi/\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'=\mathbf{T}_1$ so that $E(\mathbf{T}_1)=\mathscr{I}$. Let $\mathbf{t}_1=\mathscr{I}^{-1}\mathbf{T}_1$ and $\mathbf{t}_2=\mathscr{I}^{-1}\mathbf{T}_2$. Then $E(\mathbf{t}_1)=\mathbf{I}$. Also, $\mathbf{t}_1\boldsymbol{\beta}+\mathbf{t}_2=\mathscr{I}^{-1}(\partial\log\phi/\partial\boldsymbol{\beta})$, so that

$$V(\mathbf{t}_1\boldsymbol{\beta}+\mathbf{t}_2)=\mathscr{I}^{-1}E\left(\frac{\partial\log\phi}{\partial\boldsymbol{\beta}}\frac{\partial\log\phi}{\partial\boldsymbol{\beta}'}\right)\mathscr{I}^{-1}=\mathscr{I}^{-1}.$$

Consequently, the maximum likelihood equations are best unbiased linear estimating equations. As a corollary we deduce that if $\mathbf{T}_1$ is a matrix of constants the maximum likelihood estimates are best unbiased linear estimates; this result follows also, of course, from the sufficiency of the estimates.

G. A. Barnard has suggested to the author the following extension to non-linear estimating equations. Suppose $F(x,\hat{\theta})=0$ is an unbiased estimating equation for a single parameter θ, i.e. $E[F(x,\theta)]=0$. Let

$$f(x,\theta)=F(x,\theta)\Big/E\left(\frac{dF}{d\theta}\right).$$

By the methods given in the paper, we can show that

$$V[f(x,\theta)]\geqslant 1\Big/E\left(\frac{d\log\phi}{d\theta}\right)^2,$$

the equality being attained when $F(x,\theta)$ is a multiple of $d\log\phi/d\theta$. Thus the maximum likelihood equation

$$\frac{d\log\phi}{d\theta}\Big|_{\theta=\hat{\theta}}=0$$

is a best unbiased estimating equation for any sample size. The generalization to more than one parameter follows the treatment of the linear case given above.

The quadratic log-likelihood function is not the only one for which the minimal variance matrix is attainable by unbiased linear estimating equations, though it is the most important. The general form can be found by writing down the condition for equality in the generalized Schwarz inequality. This is

$$\frac{\partial\log\phi}{\partial\boldsymbol{\beta}}=\boldsymbol{\Lambda}(\mathbf{t}_1\boldsymbol{\beta}+\mathbf{t}_2),$$

where $\boldsymbol{\Lambda}$ is a $k\times k$ matrix depending on $\beta_1,\ldots,\beta_k$, but not on the observations.

This leads to the result

$$\phi=\exp\left(\sum_{i=1}^{k}B_iT_i+\sum_{i=1}^{k}\sum_{j=1}^{k}C_{ij}T_{ij}+T_0\right), \tag{13}$$

where $T_0, T_1, \ldots, T_k, T_{11}, \ldots, T_{kk}$ are functions of the observations only and $B_1, \ldots, B_k$, $C_{11}, \ldots, C_{kk}$ are functions of $\beta_1, \ldots, \beta_k$ only such that $(\partial B_i/\partial\beta_r)\beta_j = \partial C_{ij}/\partial\beta_r$. For this distribution

$$\frac{\partial \log\phi}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial B_1}{\partial \beta_1} & \cdots & \frac{\partial B_k}{\partial \beta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial B_1}{\partial \beta_k} & \cdots & \frac{\partial B_k}{\partial \beta_k} \end{bmatrix} [\mathbf{T}\boldsymbol{\beta}+\mathbf{t}],$$

where $\mathbf{T} = [T_{ij}]$ and $\mathbf{t} = \{T_i\}$. It may be verified that the equations $\mathbf{T}\mathbf{b}+\mathbf{t} = \mathbf{0}$ are unbiased linear estimating equations which on standardizing attain the minimal variance matrix. For $k = 1$, (13) reduces to

$$\phi = \exp(BT_1 + CT_{11} + T_0), \tag{14}$$

where $\beta(\partial B/\partial\beta) = \partial C/\partial\beta$. Equations (13) and (14) are slightly more general than the Koopman–Darmois forms for distributions admitting sufficient statistics.

So far the results of this section are exact. When we come to the study of asymptotic behaviour we encounter difficulties owing to the fact that the asymptotic properties of $\mathbf{t}_1$ and $\mathscr{I}$ will vary according to the nature of the series $\{x_{1t}\}, \ldots, \{x_{qt}\}$. Returning to the case of general ϕ, suppose that there is a diagonal matrix $\boldsymbol{\delta}_n$ depending on n such that $\boldsymbol{\delta}_n \mathscr{I}^{-1} \boldsymbol{\delta}_n$ converges to a finite positive definite matrix, $\mathbf{U}$ say. In many cases, though by no means in all, $\boldsymbol{\delta}_n$ will take the form $\mathbf{I}\sqrt{n}$. Then for any n, $V[\boldsymbol{\delta}_n(\mathbf{t}_1\boldsymbol{\beta}+\mathbf{t}_2)] - \boldsymbol{\delta}_n \mathscr{I}^{-1} \boldsymbol{\delta}_n$ will be positive definite or semi-definite. Thus

$$\lim_{n\to\infty} V[\boldsymbol{\delta}_n(\mathbf{t}_1\boldsymbol{\beta}+\mathbf{t}_2)] - \mathbf{U}$$

will be positive definite or semi-definite. We may therefore say that asymptotically $\mathbf{U}$ is the minimal variance matrix of $\boldsymbol{\delta}_n(\mathbf{t}_1\boldsymbol{\beta}+\mathbf{t}_2)$. Now $\mathbf{t}_1\boldsymbol{\beta}+\mathbf{t}_2 = -\mathbf{t}_1(\mathbf{b}-\boldsymbol{\beta})$. Also, if $\mathbf{t}_1$ converges stochastically to $\mathbf{I}$ as $n\to\infty$ and $\boldsymbol{\delta}_n\mathbf{t}_1(\mathbf{b}-\boldsymbol{\beta})$ has a limiting distribution, this limiting distribution will be the same as that of $\boldsymbol{\delta}_n(\mathbf{b}-\boldsymbol{\beta})$. Thus the limiting distribution of $\boldsymbol{\delta}_n(\mathbf{b}-\boldsymbol{\beta})$ as $n\to\infty$ has mean zero and minimal variance matrix $\mathbf{U}$. More simply, we may say that $\mathbf{b}$ is an asymptotically unbiased estimator of $\boldsymbol{\beta}$ with minimal asymptotic variance matrix $\mathscr{I}^{-1}$.

The results of this section can be summarized as follows:

Result 1. If $\mathbf{t}_1\mathbf{b}+\mathbf{t}_2 = 0$ is a set of unbiased linear estimating equations with $E(\mathbf{t}_1) = \mathbf{I}$, the vector $\mathbf{t}_1\boldsymbol{\beta}+\mathbf{t}_2$ has minimal variance matrix $\mathscr{I}^{-1}$ under wide conditions.

Result 2. If $\mathbf{t}_1 \equiv \mathbf{I}$, then $\mathbf{b}$ is an unbiased vector estimate of $\boldsymbol{\beta}$ with minimal variance matrix $\mathscr{I}^{-1}$.

Result 3. If $\boldsymbol{\delta}_n$ is a diagonal matrix such that $\boldsymbol{\delta}_n \mathscr{I}^{-1} \boldsymbol{\delta}_n$ converges to a finite positive definite matrix $\mathbf{U}$ and if $\mathbf{t}_1$ converges stochastically to $\mathbf{I}$, then $\mathbf{b}$ is asymptotically unbiased and the minimal variance of $\boldsymbol{\delta}_n(\mathbf{b}-\boldsymbol{\beta})$ is asymptotically equal to $\mathbf{U}$.

Result 4. If $\log\phi$ is a quadratic function of $\beta_1, \ldots, \beta_k$, then the maximum likelihood equations

$$\left.\frac{\partial \log\phi}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\mathbf{b}} = \mathbf{0}$$

are best unbiased linear estimating equations.

Result 5. If $\log\phi$ is a quadratic function of $\beta_1, \ldots, \beta_k$ and $\partial^2 \log\phi/\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'$ is a constant matrix, then the maximum-likelihood estimators of $\beta_1, \ldots, \beta_k$ are best unbiased estimators.

Result 6. If $\log\phi$ is a quadratic function of $\beta_1, \ldots, \beta_k$, if $\mathscr{I}^{-1}(\partial^2 \log\phi/\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}')$ converges stochastically to $\mathbf{I}$, and if $\boldsymbol{\delta}_n \mathscr{I}^{-1} \boldsymbol{\delta}_n$ converges to a finite positive definite matrix $\mathbf{U}$, then $\boldsymbol{\delta}_n(\mathbf{b}-\boldsymbol{\beta})$ has asymptotically zero mean and variance matrix $\mathbf{U}$, i.e. asymptotically $\mathbf{b}$ is a best unbiased vector estimator.

4. Estimation of Parameters in the Regression Model (1)

We now return to the question of estimating $\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q$ in the general regression model (1), namely,

$$y_t + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} = \beta_1 x_{1t} + \ldots + \beta_q x_{qt} + \epsilon_t \quad (t = 1, \ldots, n).$$

From now on we assume that each series $x_{i1}, x_{i2}, \ldots$ $(i = 1, \ldots, q)$ is a given series of constants. In the contrary case, in which they are realizations of given stochastic processes, we consider only inferences conditional on these realizations being held fixed. Let us write $y_{t-r} = x_{q+r\,t}$, $\alpha_r = -\beta_{q+r}$, $k = p+q$. Then (1) can be written

$$y_t = \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + \epsilon_t \quad (t = 1, \ldots, n), \tag{15}$$

or in matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{16}$$

Let us assume $\epsilon_1, \ldots, \epsilon_n$ to be independently normally distributed with zero mean and variance σ^2. The density conditional on $y_0, y_{-1}, \ldots, y_{-p+1}$ fixed is

$$\phi = \frac{1}{(2\pi)^{\frac{1}{2}n}\sigma^n} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right].$$

Thus the maximum-likelihood estimates are the same as the least-squares estimates and are obtained as the solution of the equations

$$\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{X}'\mathbf{y} = \mathbf{0}. \tag{17}$$

It follows from Result 4 of the previous section that (17) is a set of best unbiased linear estimating equations.

The information matrix is $\mathscr{I} = E(\mathbf{X}'\mathbf{X})/\sigma^2$. If $\mathscr{I}^{-1}\mathbf{X}'\mathbf{X}/\sigma^2$ converges stochastically to $\mathbf{I}$ and if $\boldsymbol{\delta}_n$ can be found so that $\boldsymbol{\delta}_n \mathscr{I}^{-1} \boldsymbol{\delta}_n$ converges to $\mathbf{U}$, then from Result 6 asymptotically $\boldsymbol{\delta}_n(\mathbf{b}-\boldsymbol{\beta})$ has mean zero and variance matrix $\mathbf{U}$, i.e. $\mathbf{b}$ is asymptotically a best unbiased vector estimator.

There is no doubt that in some cases the stochastic convergence of $\mathscr{I}^{-1}\mathbf{X}'\mathbf{X}$ will be difficult to verify. An idea of the difficulties involved may be gained from a study of section 3 of the paper by Mann and Wald (1943). It appears likely that a necessary condition is that the roots of the equation $x^p + \alpha_1 x^{p-1} + \ldots + \alpha_p = 0$ each have modulus less than one; the question of sufficiency remains open, however. Fortunately, in many cases it will be reasonable to regard the above results as furnishing us with an approximation. For instance, we will often be justified in concluding that

$E(\mathbf{b}) = \boldsymbol{\beta} + O(n^{-r})$ and $V(\mathbf{b}) = \mathscr{I}^{-1}[1 + O(n^{-s})]$ where $r, s \geqslant 1$. The values of r and s will depend on the nature of x series in the model; often they will be equal to unity.

The estimation of σ^2 will be considered in the next section.

5. Non-normal Errors

If $\epsilon_1, \ldots, \epsilon_n$ are not normally distributed the least-squares estimates will not necessarily possess the optimum properties described in the last section. Nevertheless, they will usually be employed in practical work. Let us therefore investigate their sampling properties assuming only that $\epsilon_1, \ldots, \epsilon_n$ are independently and identically distributed with zero mean and variance σ^2.

The least-squares equations (17) are still a set of unbiased linear estimating equations. To investigate the variance matrix of the estimates we note that $\mathbf{X}'\mathbf{X}(\mathbf{b}-\boldsymbol{\beta}) = \mathbf{X}'\boldsymbol{\epsilon}$. The ijth element of $E(\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{X})$ is

$$E\left[\sum_{t=1}^{n} x_{it}\epsilon_t \sum_{\tau=1}^{n} x_{j\tau}\epsilon_\tau\right].$$

For $i \leqslant q$, x_{it} is constant. For $i > q$, $x_{it} = y_{t+q-i}$ and depends on $\epsilon_1, \ldots, \epsilon_{t+q-i}$. In either case ϵ_t is independent of x_{it}. Similarly, ϵ_τ is independent of $x_{j\tau}$. Thus

$$E\left[\sum_{t=1}^{n} x_{it}\epsilon_t \sum_{\tau=1}^{n} x_{j\tau}\epsilon_\tau\right] = \sigma^2 E\left[\sum_{t=1}^{n} x_{it}x_{jt}\right]$$

since we only get a contribution when $t = \tau$. We therefore have

$$E(\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{X}) = \sigma^2 E(\mathbf{X}'\mathbf{X}).$$

Consequently the variance matrix of $\mathbf{X}'\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})$ is $\sigma^2 E(\mathbf{X}'\mathbf{X})$. This is an exact result for any sample size.

Let $E(\mathbf{X}'\mathbf{X}) = \mathbf{M}$ and suppose $\boldsymbol{\delta}_n$ is a diagonal matrix such that $\boldsymbol{\delta}_n \mathbf{M}^{-1} \boldsymbol{\delta}_n$ converges to a finite positive-definite matrix $\mathbf{W}$ as $n \to \infty$. Suppose also that $\mathbf{M}^{-1}\mathbf{X}'\mathbf{X}$ converges stochastically to the unit matrix. Then asymptotically $\boldsymbol{\delta}_n(\mathbf{b}-\boldsymbol{\beta})$ has zero mean and variance matrix $\sigma^2\mathbf{W}$.

The estimate of σ^2 is obtained by considering the sum of squares

$$\begin{aligned}\sum_{t=1}^{n} \epsilon_t^2 &= (\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b}) + (\mathbf{b}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\mathbf{b}-\boldsymbol{\beta}) \\ &= (\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b}) + \operatorname{trace} \boldsymbol{\delta}_n^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\delta}_n^{-1}\boldsymbol{\delta}_n(\mathbf{b}-\boldsymbol{\beta})(\mathbf{b}-\boldsymbol{\beta})'\boldsymbol{\delta}_n.\end{aligned}$$

For large n,

$$E[\operatorname{trace} \boldsymbol{\delta}_n^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\delta}_n^{-1}\boldsymbol{\delta}_n(\mathbf{b}-\boldsymbol{\beta})(\mathbf{b}-\boldsymbol{\beta})'\boldsymbol{\delta}_n] = \operatorname{trace} \mathbf{W}^{-1}\sigma^2\mathbf{W} + O(n^{-1}) = k\sigma^2 + O(n^{-1}).$$

Thus $E(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b}) = (n-k)\sigma^2 + O(n^{-1})$. Thus

$$s^2 = \frac{1}{n-k}\sum_{t=1}^{n}(y_t - b_1 x_{1t} - \ldots - b_k x_{kt})^2$$

is an unbiased estimator of σ^2 to order n^{-1}.

We may summarize the results of this section by saying that approximately in large samples $\mathbf{b}$ is an unbiased estimator of $\boldsymbol{\beta}$ with estimated variance matrix $s^2(\mathbf{X}'\mathbf{X})^{-1}$ where $(n-k)s^2$ = residual sum of squares. In other words, ordinary least-squares theory is approximately correct even if the model includes lagged dependent variables.

6. Asymptotic Distribution of $b_1, \ldots, b_k$

If $\mathbf{M}^{-1}\mathbf{X}'\mathbf{X}$ converges to the unit matrix the asymptotic distribution of $\boldsymbol{\delta}_n(\mathbf{b}-\boldsymbol{\beta})$ is the same as that of $\mathbf{W}\boldsymbol{\delta}_n^{-1}\mathbf{X}'\boldsymbol{\epsilon}$. The asymptotic properties of $b_1, \ldots, b_k$ may therefore be investigated by considering the asymptotic properties of the vector $\mathbf{X}'\boldsymbol{\epsilon} = \boldsymbol{\xi}$ say. For simplicity we confine ourselves to the case in which ϵ_t has finite moments of all orders and all values of x_{it} are bounded.

Suppose that in the original model

$$y_t + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} = \beta_1 x_{1t} + \ldots + \beta_q x_{qt} + \epsilon_t \quad (t = 1, \ldots, n), \tag{1}$$

the roots of $x^p + \alpha_1 x^{p-1} + \ldots + \alpha_p = 0$ have modulus less than one. Solving (1) as a difference equation we have that

$$y_t = w_t + z_t,$$

where w_t is the solution of the non-stochastic difference equation

$$w_t + \alpha_1 w_{t-1} + \ldots + \alpha_p w_{t-p} = \beta_1 x_{1t} + \ldots + \beta_q x_{qt} \quad (t = 1, \ldots, n),$$

with $w_0 = w_{-1} = \ldots = w_{-p+1} = 0$, while z_t is the solution of the stochastic difference equation

$$z_t + \alpha_1 z_{t-1} + \ldots + \alpha_p z_{t-p} = \epsilon_t \quad (t = 1, \ldots, n),$$

where $z_r = y_r$ $(r = 0, -1, \ldots, -p+1)$.

Writing $\xi = \{\xi_1, \ldots, \xi_k\}$, put $\xi_i = \eta_i + \zeta_i$ where

$$\eta_i = \sum_{t=1}^{n} x_{it}\epsilon_t \quad \text{and} \quad \zeta_i = 0 \quad (i = 1, \ldots, q),$$

$$\eta_i = \sum_{t=1}^{n} w_{t+q-i}\epsilon_t \quad \text{and} \quad \zeta_i = \sum_{t=1}^{n} z_{t+q-i}\epsilon_t \quad (i = q+1, \ldots, k).$$

It will be noted that each η_i is a linear function of the ϵ_t's $(i = 1, \ldots, k)$ and each ζ_i is a quadratic function of the ϵ_t's $(i = q+1, \ldots, k)$.

In what follows we shall confine ourselves exclusively to the case in which all x_{it} are bounded $(i = 1, \ldots, k;\ t = 1, 2, \ldots)$. This includes periodic regressions, but excludes polynomial regressions. Since the roots of $x^p + \alpha_1 x^{p-1} + \ldots + \alpha_p = 0$ have modulus less than one all w_{t+q-i} are also bounded. Thus each η_i is $O(\sqrt{n})$. Also each ζ_i is $O(\sqrt{n})$. Consequently, we may take $\boldsymbol{\delta}_n$ to be the matrix $\mathbf{I}\sqrt{n}$.

Suppose that m_0, s_0 can be found such that for all $m \geqslant m_0$, $s \geqslant s_0$ and $i = 1, \ldots, k$, the variance of $\sum_{t=s+1}^{s+m} x_{it}\epsilon_t$ has a positive lower bound.

Mann and Wald (1943, p. 187) have shown that by taking the sums of m successive terms of $\sum_{t=1}^{n} z_{t+q-i}\epsilon_t$, then for sufficiently large m, ζ_i behaves asymptotically like

$$\sum_{j=1}^{r} b_{ij} \quad (i = q+1, \ldots, k),$$

where $b_{i1}, b_{i2}, \ldots$ are independent, where $r \to \infty$ as $n \to \infty$ and where each b_{ir} is (quoting from Mann and Wald's paper) "a quadratic function of the ϵ_t's, the number of terms and the coefficients being bounded functions of j. Hence since ϵ_t has finite moments

by assumption, the third absolute moment of (b_{ij}) is a bounded function of j. Furthermore, (the variance of b_{ij}) converges to a finite positive limit as $j \to \infty$". (Brackets indicate different symbols from those used by Mann and Wald. In fact, Mann and Wald's model differs slightly from that under consideration here since their model contains a constant α_0, but this does not affect the argument essentially.)

We shall have established the asymptotic normality of $\xi_1, \ldots, \xi_k$ if we can show that $\sum_{i=1}^{k} \lambda_i \xi_i$ is asymptotically normal for all $\lambda_1, \ldots, \lambda_k$. This follows from Mann and Wald's Lemma 2 (1943, p. 188). From the above it follows that $\sum_{i=1}^{k} \lambda_i \xi_i$ behaves asymptotically like $\sum_{j=1}^{\infty} (c_j + b_j)$ where each c_j is the sum of m successive terms in the series $\sum_{i=1}^{q} \lambda_i x_{it} \epsilon_t + \sum_{i=q+1}^{k} \lambda_i w_{t+q-i} \epsilon_t \; (t = 1, 2, \ldots)$ and $b_j = \sum_{i=q+1}^{k} \lambda_i b_{ij}$. The variables $c_j + b_j \; (j = 1, 2, \ldots)$ are independent, are quadratic in the ϵ_t's with finite coefficients and have finite variances with a positive lower bound in the limit. Thus the Liapounoff conditions for the Central Limit Theorem are satisfied (see, e.g., Cramér, 1946a, p. 215), whence $\sum_{i=1}^{k} \lambda_i \xi_i$ is asymptotically normal. It follows that $\boldsymbol{\delta}_n(\mathbf{b} - \boldsymbol{\beta})$ is asymptotically multinormal with mean vector zero and variance matrix $\sigma^2 \mathbf{W}$.

These results justify the application to model (1) of the following result in large-sample regression theory, when the above conditions are satisfied. Let R_1 be the residual sum of squares after fitting k_1 coefficients in addition to the mean and let R_2 be the residual sum of squares after fitting a subset of k_2 of these in addition to the mean. Then, provided the k_2 coefficients and the mean account for all the non-zero coefficients of (1), the quantity $[(n - k_1 - 1)(R_2 - R_1)]/[(k_1 - k_2) R_1]$ is approximately distributed as χ^2 with $k_1 - k_2$ degrees of freedom. This result can be used to provide an approximate test of the hypothesis that the omitted $k_1 - k_2$ coefficients have true values zero.

7. An Alternative Model

For many situations a different model is appropriate, namely,

$$y_t = \beta_1 x_{1t} + \ldots + \beta_q x_{qt} + u_t \quad (t = 1, \ldots, n), \tag{18}$$

where $\{u_t\}$ is a stationary autoregressive series generated by

$$u_t + \alpha_1 u_{t-1} + \ldots + \alpha_p u_{t-p} = \epsilon_t \quad (t = \ldots -1, 0, 1, \ldots), \tag{19}$$

and where each series $x_{i1}, x_{i2}, \ldots \; (i = 1; \ldots, q)$ is a given series of constants. We assume the ϵ_t's to be independently and identically distributed with zero mean and constant variance. The model (18) differs from (1) in that it does not contain lagged y's and has autocorrelated error terms. The possession of the latter property makes it an appropriate model for the regression analysis of certain kinds of time series data.

On multiplying successive terms of (18) by $1, \alpha_1, \ldots, \alpha_p$ and adding, we obtain

$$\begin{aligned} y_t + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} \\ = \beta_1 x_{1t} + \ldots + \beta_q x_{qt} + \alpha_1 \beta_1 x_{1\,t-1} + \ldots + \alpha_p \beta_q x_{q\,t-p} + \epsilon_t. \end{aligned} \tag{20}$$

In principle we can apply least squares to (20) and this will lead to optimum estimates when $\epsilon_1, \ldots, \epsilon_n$ are normally distributed. Letting $S = \frac{1}{2}\sum_{t=1}^{n} \epsilon_t^2$, the least-squares equations can be written

$$\left.\frac{\partial S}{\partial \boldsymbol{\alpha}}\right|_{\substack{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}} \\ \boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}} = \mathbf{0}, \tag{21}$$

$$\left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\substack{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}} \\ \boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}} = \mathbf{0}. \tag{22}$$

It is obvious, however, that these equations are non-linear in $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ and hence are difficult to solve. In fact, some sort of iterative procedure would have to be employed. For this reason a direct attempt to solve the least-squares equations would not be very practical. Other methods based on successive approximation have been proposed by Cochrane and Orcutt (1949) and Champernowne (1948), but these also are rather onerous on account of the computational burden. A method will now be suggested which requires only two stages of computation but which leads to estimates having the same variance matrix as the least-squares estimates in large samples.

First we evaluate the large-sample variance matrix of the least-squares estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. Asymptotically (21) and (22) are equivalent to

$$\left.\frac{\partial S}{\partial \boldsymbol{\alpha}}\right|_{\substack{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}} \\ \boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}} = \mathbf{0}, \quad \left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\substack{\boldsymbol{\alpha}=\boldsymbol{\alpha} \\ \boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}} = \mathbf{0}, \tag{23}$$

which are a set of unbiased linear estimating equations. On applying the methods given earlier in the paper, the estimates constituting the solution of (23) are found to have variance matrix

$$\sigma^2 \begin{bmatrix} E\dfrac{\partial^2 S}{\partial \boldsymbol{\alpha}\,\partial \boldsymbol{\alpha}'} & E\dfrac{\partial^2 S}{\partial \boldsymbol{\alpha}\,\partial \boldsymbol{\beta}'} \\ E\dfrac{\partial^2 S}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\alpha}'} & E\dfrac{\partial^2 S}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}'} \end{bmatrix}^{-1} = \sigma^2 \begin{bmatrix} \left\{E\left(\dfrac{\partial^2 S}{\partial \boldsymbol{\alpha}\,\partial \boldsymbol{\alpha}'}\right)\right\}^{-1} & \mathbf{0} \\ \mathbf{0} & \left\{E\left(\dfrac{\partial^2 S}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}'}\right)\right\}^{-1} \end{bmatrix} \tag{24}$$

since

$$\frac{\partial^2 S}{\partial \alpha\,\partial \beta'} = \left[\sum_{t=1}^{n} x_{j\,t-1}\,\epsilon_t\right]$$

which has zero expectation. Consequently $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are asymptotically uncorrelated. Also, $\hat{\boldsymbol{\alpha}}$ has the same variance matrix as would have been given by least squares if $\boldsymbol{\beta}$ were accurately known and vice versa. If we can find estimates $\mathbf{a}, \mathbf{b}$ of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ having these properties we shall regard them as asymptotically efficient.

We now show that such estimates of $\mathbf{a}$ and $\mathbf{b}$ can be calculated in the following way. Treat (20) as a general linear regression model of the form (1) by writing it in the form

$$y_t + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} = \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_r x_{rt} + \epsilon_t, \tag{25}$$

where the α's and β's are regarded as capable of independent variation but where $q(p+1)-r$ variables have been eliminated to leave a linearly independent set, r being the number of linearly independent vectors in the set

$$\{x_{1t}\}, \ldots, \{x_{qt}\}, \quad \{x_{1\,t-1}\}, \ldots, \{x_{q\,t-1}\}, \ldots, \{x_{q\,t-p}\}.$$

Let $a_1, \ldots, a_p$, $b'_1, \ldots, b'_r$ be the least-squares estimates obtained by minimizing

$$\sum_{t=1}^{n} (y_t + a_1 y_{t-1} + \ldots + a_p y_{t-p} - b'_1 x_{1t} - \ldots - b'_r x_{rt})^2, \tag{26}$$

i.e. the fact that $\beta_{q+1}, \ldots, \beta_r$ are expressible in terms of $\alpha_1, \ldots, \alpha_p$ and $\beta_1, \ldots, \beta_q$ is ignored when performing this fitting.

Suppose that the transformation $y_{t-i} = \beta_1 x_{1\,t-i} + \ldots + \beta_q x_{q\,t-i} + u_{t-i}$ $(i = 1, \ldots, p)$, derived from (18), had been effected before carrying through the minimization. We would then have minimized

$$\sum_{t=1}^{n} (y_t + a_1 u_{t-1} + \ldots + a_p u_{t-p} - b''_1 x_{1t} - \ldots - b''_r x_{rt})^2, \tag{27}$$

with respect to $a_1, \ldots, b''_r$, and the values $a_1, \ldots, a_p$ so obtained would have been exactly the same as those obtained by minimizing (26). Since $u_{t-1}, \ldots, u_{t-p}$ are completely independent of $x_{1t}, \ldots, x_{rt}$, it follows that $a_1, \ldots, a_p$ are asymptotically uncorrelated with $b''_1, \ldots, b''_r$, i.e. with all linear functions of the form

$$\sum_{t=1}^{n} x_{it} \epsilon_t \quad (i = 1, \ldots, r).$$

Moreover, the asymptotic variance matrix of the vector $\mathbf{a} = \{a_1, \ldots, a_p\}$ is the same as that of $a'_1, \ldots, a'_p$ obtained by minimizing

$$\sum_{t=1}^{n} (u_t + a'_1 u_{t-1} + \ldots + a'_p u_{t-p})^2.$$

But $a'_1, \ldots, a'_p$ are the estimates that would have been used if $\beta_1, \ldots, \beta_q$ had been known exactly. Hence $\mathbf{a}$ has the same asymptotic variance matrix as $\hat{\boldsymbol{\alpha}}$ given by (23).

We estimate $\boldsymbol{\beta}$ by the vector $\mathbf{b}$ obtained by substituting $\mathbf{a}$ for α in the least-squares equations for $\boldsymbol{\beta}$, i.e. $\mathbf{b}$ is the solution of

$$\left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\substack{\boldsymbol{\alpha}=\mathbf{a} \\ \boldsymbol{\beta}=\mathbf{b}}} = 0; \tag{28}$$

in other words, $\mathbf{b}$ is obtained by minimizing

$$\sum_{t=1}^{n} (y_t + a_1 y_{t-1} + \ldots + a_p y_{t-p} - b_1 x_{1t} - \ldots - a_p b_q x_{qt-p})^2 \tag{29}$$

with respect to $b_1, \ldots, b_q$. The estimate $\mathbf{b}$ has the same asymptotic distribution as $\boldsymbol{\beta}^*$ given by

$$\left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\substack{\boldsymbol{\alpha}=\boldsymbol{\alpha} \\ \boldsymbol{\beta}=\boldsymbol{\beta}^*}} = \mathbf{0}.$$

Now by ordinary regression theory the distribution of $\boldsymbol{\beta}^*$ depends only on the linear functions

$$\sum_{t=1}^{n} x_{it}\epsilon_t \quad (i = 1, \ldots, n).$$

Since **a** is asymptotically uncorrelated with these linear functions, **a** is asymptotically uncorrelated with $\boldsymbol{\beta}^*$ and hence with **b**. We conclude that **a** and **b** have asymptotically the same mean vector and variance matrix as the proper least-squares estimators defined by (21) and (22).

The estimating procedure that has been recommended can be summarized as follows. Let $-a_1, \ldots, -a_p$ be the coefficients of $y_{t-1}, \ldots, y_{t-p}$ in the fitted least-squares regression of y_t on $y_{t-1}, \ldots, y_{t-p}, x_{1t}, \ldots, x_{qt}, x_{1\,t-1}, \ldots, x_{q\,t-1}, \ldots, x_{1\,t-p}, \ldots, x_{q\,t-p}$. Let $v_t = y_t + a_1 y_{t-1} + \ldots + a_p y_{t-p}$ and let $w_{it} = x_{it} + a_1 x_{i\,t-1} + \ldots + a_p x_{i\,t-p}$ $(i = 1, \ldots, q)$. Then $b_1, \ldots, b_q$ are the coefficients of $w_{1t}, \ldots, w_{qt}$ in the fitted least-squares regression of v_t on $w_{1t}, \ldots, w_{qt}$.

REFERENCES

ANDERSON, T. W. & RUBIN, H. (1950), "The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations", *Ann. Math. Statist.*, **21**, 570–582.

CHAMPERNOWNE, D. G. (1948), "Sampling theory applied to autoregressive sequences", *J. R. Statist. Soc.*, B, **10**, 204–231.

COCHRANE, D. & ORCUTT, G. H. (1949), "Application of least-squares regression to relationships containing autocorrelated error terms", *J. Amer. Statist. Ass.*, **44**, 32–61.

CRAMÉR, H. (1946a), *Mathematical Methods of Statistics*. Princeton University Press.

—— (1946b), "A contribution to the theory of statistical estimation", *Skand. Akt.*, **29**, 85–94.

KENDALL, M. G. (1951), "Regression, structure and functional relationship—I", *Biometrika*, **38**, 11–25.

KOOPMANS, T. C., RUBIN, H. & LEIPNIK, R. B. (1950), "Measuring the equation systems of dynamic economics". Chapter 2 of *Statistical Inference in Dynamic Economic Models*, edited by T. C. Koopmans. New York: Wiley.

MANN, H. B. & WALD, A. (1943), "On the statistical treatment of linear stochastic difference equations", *Econometrica*, **11**, 173–220.

RAO, C. R. (1952), *Advanced Statistical Methods in Biometric Research*. New York: Wiley.