Annals of Economic and Social Measurement, 6/1, 1977

COVARIANCE OF ESTIMATED PARAMETERS IN ARMA REGRESSION MODELS

BY RICHARD HILL

In this paper we derive the asymptotic covariance matrix of the maximum likelihood estimator for regression models with ARMA errors, we discuss some alternative sample estimates of this covariance matrix, and we extend some of these results to forecasting.

I. INTRODUCTION

We begin by defining a general class of regression models, having gaussian errors with unknown covariance structure. We derive the likelihood function and its derivatives, and specialize these to the case where the covariance structure is that specified by an autoregressive moving average process. Next we derive the asymptotic covariance matrix for the maximum likelihood estimator, and we discuss some alternative sample estimates of this covariance matrix. Finally we extend some of these results to forecasting.

II. THE MODEL

Let β be a $k \times 1$ vector of parameters, m a twice differentiable function m: $R^k \rightarrow R^n$, so that $m(\beta)$ is an $n \times 1$ vector. $V(\theta)$ is an $n \times n$ symmetric positive definite matrix, whose elements are a function of the $p \times 1$ vector θ.

Our model is

$$Y = m(\beta) + \epsilon, \tag{1-1}$$

where

$$\epsilon \sim N_n(0, \sigma^2 V(\theta)),$$

so that if

$$V(\theta) = [V^{1/2}(\theta)][V^{1/2}(\theta)]^T,$$

$$V^{1/2}(\theta)\,\epsilon \sim N_n(0, \sigma^2 I_n). \tag{1-2}$$

For example if $V(\theta) = I_n$, we have the usual nonlinear regression model, and if

$$m(\beta) = X\beta$$

then we have $Y - X\beta \sim N(0, \sigma^2 I_n)$ which is the usual linear regression model. For convenience, we put $f(\beta) = Y - m(\beta)$, so that $f(\beta)$ is the $n \times 1$ vector of residuals. We let $\gamma = \binom{\beta}{\theta}$, the combined parameter vector.

In our applications we will find that p, the dimension of θ, is much smaller than n, so that $V(\theta)$ is unknown only up to few parameter values, which we wish to estimate. For example, if Y were a zero mean time series, we could take $f(\beta) = Y$, and perhaps assume

$$V^{-1/2}(\theta) = \begin{pmatrix} 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ \theta & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \theta & 1 & \cdot & \cdot & \cdot & 0 \\ & \cdot & & & & & \\ & \cdot & & & & & \\ & \cdot & & & & & \\ 0 & 0 & 0 & \cdots & 0 & \theta & 1 \end{pmatrix}$$

This is a one parameter model, in which we are trying to estimate the correlation between Y_i and Y_{i+1}, assuming that Y_i and Y_{i+t} are uncorrelated for $t \geq 2$. The ARMA models described in Box and Jenkins (1970) are special cases of (1-1). In fact, they can be written as

$$Y_i - \rho_1 Y_{i-1} - \rho_2 Y_{i-2} - \cdots - \rho_a Y_{i-a} = \epsilon_i - \phi_1 \epsilon_{i-1} - \cdots - \phi_b \epsilon_{i-b},$$

(1-3)

where

$$\epsilon_1, \ldots, \epsilon_n \text{ are i.i.d. } N(0, \sigma^2) \text{ variables.}$$

In our notation

(1-4) $$P(\rho) Y = T(\phi)\epsilon,$$

where

$$\rho = (\rho_1, \ldots, \rho_a)^T, \phi = (\phi_1, \ldots, \phi_b)^T,$$

(1-5) $$P(\rho) = \begin{pmatrix} 1 & 0 & 0 & & 0 \\ -\rho_1 & 1 & 0 & & 0 \\ -\rho_2 & -\rho_1 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ & & & \cdots & \\ -\rho_a & -\rho_{a-1} & -\rho_{a-2} & & 0 \\ 0 & -\rho_a & -\rho_{a-1} & & 0 \\ 0 & 0 & -\rho_a & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & 1 \end{pmatrix}$$

and

$$(1\text{-}6) \qquad T(\phi) = \begin{pmatrix} 1 & 0 & 0 & & 0 \\ -\phi_1 & 1 & 0 & & 0 \\ -\phi_2 & -\phi_1 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ & & & \cdots & \\ -\phi_a & -\phi_{a-1} & -\phi_{a-2} & & 0 \\ 0 & -\phi_a & -\phi_{a-1} & & 0 \\ 0 & 0 & -\phi_a & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & 1 \end{pmatrix} .$$

Letting

$$(1\text{-}7) \qquad \theta = (\rho_1, \ldots, \rho_a, \phi_1, \ldots, \phi_b)^T$$

and

$$(1\text{-}8) \qquad V^{-1/2}(\theta) = T^{-1}(\phi) P(\rho),$$

we have

$$V^{-1/2}(\theta) y \sim N_n(0, \sigma^2 I_n),$$

so that the Box-Jenkins models are indeed special cases of (1-1), with $m(\beta) \equiv 0$ and $V(\theta)$ given by (1-8). Throughout, we will let $P(\rho)$ and $T(\phi)$ be defined by the above matrices.

II. The Likelihood Function

We propose to estimate the parameter γ by using the method of maximum likelihood. We can only observe the $n \times 1$ vector Y, so we need the likelihood in terms of Y:

$$(2\text{-}1) \qquad L(f, \beta, \theta, \sigma) = \frac{C \det(V^{-1/2}(\theta))}{\sigma^n} \exp\left[\frac{-f^T(\beta) V^{-1}(\theta) f(\beta)}{2\sigma^2}\right]$$

where C is a constant (see Rao (1969) Section 8a.4).

For all out applications we will have $\det(V(\theta)) = 1$, so we immediately simplify things by assuming that

$$(2\text{-}2) \qquad \det(V^{1/2}(\theta)) = 1$$

for all values of θ. Hence

$$\text{(2-3)}\quad \log L(f, \beta, \theta, \sigma) = -\frac{1}{2\sigma^2} f^T(\beta)\, V^{-1}(\theta) f(\beta) - n \log \sigma + C.$$

To maximize this we differentiate and set the derivatives to 0. (Recall that $f(\beta) = Y - m(\beta)$, so for each β, $f(\beta)$ is observable.)

$$\frac{\partial \log L}{\partial \sigma} = \frac{1}{\sigma^3} f^T(\beta)\, V^{-1}(\theta) f(\beta) - \frac{n}{\sigma} = 0$$

or

$$f^T(\beta)\, V^{-1}(\theta) f(\beta) = n\sigma^2.$$

Hence

$$\text{(2-4)}\qquad \hat{\sigma}^2 = \frac{f^T(\hat{\beta})\, V^{-1}(\hat{\theta}) f(\hat{\beta})}{n},$$

and we can treat σ^2 as a constant throughout the rest of the discussion. Note that we are now trying to minimize

$$f^T(\beta)\, V^{-1}(\theta) f(\beta).$$

We write $f(\beta) = (f_1, \ldots, f_n)^T$; $V^{-1}(\theta) = (V^{ij})$ for convenience. Then

$$\begin{aligned}
\frac{\partial \log L}{\partial \beta_j} &= \frac{-1}{2\sigma^2} \frac{\partial}{\partial \beta_j}\left(\sum_{ij} f_i V^{ij} f_j\right) \\
&= \frac{-1}{2\sigma^2}\left(\sum_{ij}\left[\frac{\partial f_i}{\partial \beta_j} V^{ij} f_j + f_i V^{ij} \frac{\partial f_j}{\partial \beta_i}\right]\right) \\
&= \frac{-1}{\sigma^2}\left(\sum_{ij} \frac{\partial f_i}{\partial \beta_j} V^{ij} f_j\right) \\
&= \frac{-1}{\sigma^2}\left(\frac{\partial f(\beta)}{\partial \beta_j}\right)^T V^{-1}(\theta) f(\beta).
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \log L}{\partial \theta_m} &= \frac{-1}{2\sigma^2} \frac{\partial}{\partial \theta_m}\left(\sum_{ij} f_i V^{ij} f_j\right) \\
&= \frac{-1}{2\sigma^2}\left(\sum_{ij} f_i \frac{\partial V^{ij}}{\partial \theta_m} f_j\right) \\
&= \frac{-1}{2\sigma^2} f^T(\beta) \frac{\partial V^{-1}(\beta)}{\partial \theta_m} f(\beta).
\end{aligned}$$

So the $k + p$ normal equations are

$$\text{(2-5)}\qquad \begin{cases} \dfrac{\partial f^T(\beta)}{\partial \beta_j} V^{-1}(\theta) f(\beta) = 0 \\[2ex] f^T(\beta) \dfrac{\partial V^{-1}(\theta)}{\partial \theta_m} f(\beta) = 0. \end{cases}$$

Next, we compute the matrix of second derivatives. Omitting the details, we have:

$$-\sigma^2 \frac{\partial^2 \log L}{\partial \beta_i \beta_j} = \frac{\partial f^T(\beta)}{\partial b_i} V^{-1}(\theta) \frac{\partial f(\beta)}{\partial \beta_j} + \frac{\partial^2 f^T(\beta)}{\partial h_i \beta_j} V^{-1}(\theta) f(\beta)$$

$$-\sigma^2 \frac{\partial^2 \log L}{\partial \beta_i \theta_m} = \frac{\partial f^T(\beta)}{\partial \beta_i} \frac{\partial V^{-1}(\theta)}{\partial \theta_m} f(\beta) \tag{2-6}$$

$$-2\sigma^2 \frac{\partial^2 \log L}{\partial \theta_l \theta_m} = f^T(\beta) \frac{2 V^{-1}(\theta)}{\partial \theta_l \theta_m} f(\beta),$$

We summarize these results as follows:

$$H = \begin{pmatrix} [f']^T V^{-1} f' + f'' V^{-1} f & [f']^T [V^{-1}]' f \\ [f']^T [V^{-1}]' f & f^T [V^{-1}]'' f \end{pmatrix}$$

$$G = \begin{pmatrix} [f']^T V^{-1} f \\ f^T [V^{-1}]' f \end{pmatrix} \tag{2-7}$$

The primes denoting the appropriate derivatives.

The asymptotic information matrix $I(\gamma)$ is then given by

$$\mathrm{I}(\gamma) = E\left(\frac{1}{\sigma^2}\,\mathrm{H}\right) \tag{2-8}$$

Holland (1973) described a method for carrying out the expectation in 2-9. Since σ^2 is considered fixed, we treat it as a constant. Then

$$E\left[\frac{1}{\sigma^2}\,[f']^T V^{-1} f' + [f'']^T V^{-1} f\right] =$$

$$= \frac{1}{\sigma^2}\,[f']^T V^{-1} f' + [f'']^T V^{-1}\,[Ef] = \frac{1}{\sigma^2}\,[f']^T V^{-1} f',$$

since $f'(\beta) = m'(\beta)$ was assumed fixed;

$$E\left[\frac{1}{\sigma^2}\,[f']^T [V^{-1}]' f\right] = \frac{1}{\sigma^2}\,[f']^T [V^{-1}]' [Ef] = 0,$$

since $f(\beta) = Y - m(\beta) \sim N(0, \sigma^2 V(\theta))$, by 1-1.

$$E\left[\frac{1}{2\sigma^2}\,f^T [V^{-1}]'' f\right] = \frac{1}{2\sigma^2}\,\mathrm{trace}\,[E[f^T [V^{-1}]'' f]]$$

$$= \frac{1}{2\sigma^2}\,E\,\mathrm{trace}\,[f^T [V^{-1}]'' f] = \frac{1}{2\sigma^2}\,E\,\mathrm{trace}\,[[V^{-1}]'' f f^T]$$

$$= \frac{1}{2\sigma^2}\ \text{trace}\ \{E[[V^{-1}]'' ff^T]\} = \frac{1}{2\sigma^2}\ \text{trace}\,[[V^{-1}]''\, E[ff^T]]$$

$$= \frac{1}{2\sigma^2}\ \text{trace}\ [[V^{-1}]''\sigma^2 V] = \frac{1}{2}\ \text{trace}\,[V[V^{-1}]'']$$

So we have

$$(2\text{-}9) \qquad I(\gamma) = \begin{pmatrix} \frac{1}{\sigma^2}\,[f']^T V^{-1} f' & 0 \\ 0 & \frac{1}{2}\ \text{trace}\ V\left(\frac{\partial^2 V^{-1}}{\partial\theta_l\theta_m}\right) \end{pmatrix}.$$

We see that the ARMA coefficient estimates are asymptotically uncorrelated with the regression parameter estimates, and consequently the design of the regression experiment does not affect the precision of the estimate of the ARMA parameters.

We now specialize to a subset of (1-1) for which the expressions (2-9) are easy to compute.

III. Specialization To ARMA Error Processes

We restrict ourselves to the subset of (1-1) for which

$$(3\text{-}1) \qquad V(\theta) = P^{-1}(\rho)\,T(\phi)\,[P^{-1}(\rho)\,T(\phi)]^T$$

so that

$$(3\text{-}2) \qquad V^{-1/2}(\theta) = T^{-1}(\phi)\,P(\rho),$$

where T, P are given by 1-5 and 1-6.

The error process is now an ARMA error process. Using the fact that both P and T are Toeplitz matrices, it is possible to considerably simplify the expressions 2-7 and 2-9. These computations are straightforward but tedious, and they will not be given here. They are carried out in full in Hill (1975). In particular, it can be shown that

$$\frac{1}{2}\ \text{trace}\left(V\,\frac{\partial^2 V^{-1}}{\partial\theta_l\theta_m}\right) = \text{trace}\left[(T^{-1}P)\partial\,\frac{(P^{-1}T)}{\partial\theta_l}\left((T^{-1}P)\partial\,\frac{(P^{-1}T)}{\partial\theta_m}\right)T\right]$$

(3-3)

and

$$(3\text{-}4) \qquad \begin{cases} (T^{-1}P)\,\dfrac{\partial}{\partial\rho_l}\,(P^{-1}T) = -\,\dfrac{\partial P}{\partial\rho_l}\,P^{-1} \\[2ex] (T^{-1}P)\,\dfrac{\partial}{\partial\rho_l}\,(P^{-1}T) = \dfrac{\partial T}{\partial\rho_l}\,T^{-1} \end{cases}$$

Since the matrices $\frac{\partial P}{\partial \rho_l}$, $\frac{\partial T}{\partial \phi_l}$, P^{-1} and T^{-1} are readily computed in closed form, these expressions simplify the computation of the information matrix (2-9).

i) If $T(\phi) = I$ and $\rho = 0$, then

$$I(\theta) = \begin{pmatrix} n-1 & 0 & 0 & \cdots & 0 \\ 0 & n-2 & 0 & \cdots & 0 \\ 0 & 0 & n-3 & \cdots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \cdots & n-p \end{pmatrix}$$

This is a perfectly sensible answer, since it is well known that the estimate for ρ_j is essentially based on $n - j$ observations. In particular, for $p = 1$

$$\hat{\rho} = \frac{\sum_{i=2}^{n} Y_i Y_{i-1}}{\sum_{i=2}^{n} Y_i^2}.$$

ii) Furthermore, from formulas (3-3) and (3-4) we see that if either $P(\rho) \equiv I$ or $T(\rho) \equiv I$, so that we have only ϕ's or only ρ's to estimate, the value of $I(\gamma)$ will depend only on the value of the ϕ or ρ vector, and not on whether or not it is a ϕ vector or a ρ vector. That is, $I(\phi) = I(\rho)$ whenever $\phi = \rho$ and, respectively, $P(\rho) = I$ or $T(\phi) = I$.

This result is rather surprising: it says that the asymptotic variance for the ρ's is the same as that for the ϕ's if only ρ's or ϕ's are present, even though they represent quite different models: One is

$$Y_i - \rho_1 Y_{2-1} \cdots - \rho_p Y_{i-p} \sim N(0, \sigma^2)$$

The other is

$$Y_i \sim \epsilon_i - \phi_1 \epsilon_{i-1} \cdots - \phi_p \epsilon_{i-p}$$

where

$$\epsilon_1, \ldots, \epsilon_n \sim \text{i.i.d. } N(0, \sigma_2)$$

iii) If $\rho = \phi$, then $I(\phi)$ is singular, since it has the form $\begin{pmatrix} A & -A \\ -A & A \end{pmatrix}$. This means that the parameters are not estimable, and this is reasonable since our model is now

$$Y \sim N(0, \sigma^2 I_n)$$

and many choices of ρ and ϕ will give us this model.

IV. Approximations To The Covariance Matrix

It is usual to assume that $I^{-1}(\hat{\gamma}) \xrightarrow{P} I^{-1}(\gamma)$, and in fact Rao (1965) shows that if F_n is the distribution function of γ and G_n is the distribution function of a random variable distributed $N(0, I^{-1}(\hat{\gamma}))$, then

$$\lim_{n \to \infty} | F_n - G_n | = 0,$$

under suitable regularity conditions.

By the strong law of large numbers, and consistency, we also have

$$\frac{1}{n} H(\hat{\gamma}) - \frac{1}{n} I(\gamma) \xrightarrow{P} 0, \quad \text{since } EH(\gamma) = EI(\gamma).$$

(Note that it is *not* true that $H(\hat{\gamma}) \xrightarrow{P} I(\gamma)$, in fact $H(\gamma)$ need not converge to $I(\gamma)$, as we will see later.)

On the basis of this result, it has been suggested that we use $H(\hat{\gamma})$ rather than $I(\hat{\gamma})$ as an estimate of $I(\gamma)$. We point out some disadvantages to this approach.

i) Suppose that $f(\beta) = Y$, and that $\hat{\rho} = \hat{\phi}$, so that $I(\hat{\gamma})$ is singular. $H(\hat{\gamma})$ is not necessarily singular: in fact, let $\hat{\rho} = \hat{\phi} = 0$, and $p = 2$. Then

$$H(\hat{\gamma}) = \begin{pmatrix} \sum_{i=2}^{n} Y_i^2 & \sum_{i=2}^{n} Y_i^2 - \sum_{i=3}^{n} Y_{i-2}Y_i \\ \sum_{i=2}^{n} Y_i^2 - \sum_{i=3}^{n} Y_{i-2}Y_i & \sum_{i=2}^{n} Y_i^2 + \sum_{i=3}^{n} Y_{i-2}Y_i \end{pmatrix}.$$

ii) Let $f(\beta) = Y - \beta$, $k = 1$, $T(\phi) = 1$, $p = 1$, $\sigma = 1$. Then

$$H(0) = \begin{pmatrix} n & Y_1 - Y_n \\ Y_1 - Y_n & \sum_{i=2}^{n} Y_{i-1}^2 \end{pmatrix}$$

Whereas

$$I(0) = \begin{pmatrix} n & 0 \\ 0 & n - 1 \end{pmatrix}$$

The form for $H(0)$ is most easily derived by observing that here

$$2 \log L(f, \beta, \rho) = (Y_1 - \beta)^2 + \sum_{i=2}^{n} [(Y_i - \beta) - \rho(Y_{i-1} - \beta)]^2$$

So

$$\frac{\partial \log L}{\partial \rho} = - \sum_{2}^{n} [(Y_i - \beta) - \rho(Y_{i-1} - \beta)](Y_{i-1} - \beta)$$

$$\frac{\partial \log L}{\partial \beta} = -(Y_i - \beta) + 2\sum_{2}^{n}[(y_i - \beta) - \rho(Y_{i-1} - \beta)][-1 + \rho]$$

$$\frac{\partial \log L}{\partial \rho^2} = \sum_{2}^{n}(Y_{i-1} - \beta)^2$$

$$\frac{\partial^2 \log L}{\partial \beta^2} = 1 + \sum_{2}^{n}[-1 + \rho][-1 + \rho] = N - 2(N - 1)\rho + (N - 1)\rho^2$$

$$\begin{aligned}\frac{\partial^2 \log L}{\partial \rho \beta} &= -\sum_{2}^{n}[-1 + \rho](Y_{i-1} - \beta)\\ &\quad + \sum_{2}^{n}[(Y_i - \beta) - \rho(Y_{i-1} - \beta)](-1)\\ &= \sum_{2}^{n}(Y_{i-1} - \beta) - \rho\sum_{2}^{n}(Y_{i-1} - \beta)\\ &\quad - \sum_{2}^{n}(Y_i - \beta) + \rho\sum_{2}^{n}(Y_{i-1} - \beta)\\ &= (Y_1 - \beta) - (Y_n - \beta) = Y_1 - Y_n.\end{aligned}$$

Clearly $H(0)$ does *not* converge in probability to $I(0)$; however, under the assumption $\rho = 0$

$$Y_1 - Y_n \sim N(0, 2).$$

and

$$\sum_{i=2}^{n} Y_{i-1}^2 \sim X_{n-1}^2,$$

so we see that

$$\frac{Y_1 - Y_n}{n} = O_p\left(\frac{1}{n}\right)$$

and

$$\frac{1}{n}\left(\sum_{i=2}^{n} Y_{i-1}^2 - n + 1\right) = O_p\left(\frac{1}{n}\right).$$

In this case, however, $I(0)$ is the correct answer, so we see that $H(0)$ is not as good.

There is another approximation which is clearly superior to $H(\hat{\gamma})$:

$$H_2(\hat{\gamma}) = \begin{pmatrix}[f']^T V^{-1} f' & 0\\ 0 & f^T[V^{-1}]'' f\end{pmatrix}. \tag{4-1}$$

This is obtained by eliminating those components in (2-7) whose expectation is obviously 0. For the example we have

$$H_2(0) = \begin{pmatrix} n & 0 \\ 0 & \sum_{i=2}^{n} Y_{i-1}^2 \end{pmatrix},$$

which is still not as good as $I(0)$. H_2 also still suffers from disadvantage *i*) above; in fact, the lower right corner of H_2 is identical to that of H.

We conclude that the variance of the θ's (ARMA coefficients) should not be estimated from $H(\hat{\gamma})$, but from $I(\hat{\gamma})$, since the two can differ significantly; a numerical example follows.

We generated Y by taking 100 points from a normal (0,1) distribution, so that $Y \sim N_n(0,I)$. Then we fit the model (1-1) with $m(\beta) = \beta$, where β is a scalar and $\theta = \begin{pmatrix} \rho \\ \phi \end{pmatrix}$, so that we fit a first order moving average, first order autoregressive process. (I.e., both P and T are present, but each depends only on one parameter.) We found

$$\hat{\gamma} = \begin{pmatrix} .13536 \\ -.36 \\ -.3125 \end{pmatrix},$$

$$H^{-1}(\hat{\gamma}) = \begin{pmatrix} .009328 & .000476 & .000826 \\ & .347967 & .345833 \\ & & .350404 \end{pmatrix}.$$

$$I^{-1}(\hat{\gamma}) = \begin{pmatrix} .009319 & 0 & 0 \\ & 3.1134 & 3.16488 \\ & & 3.22633 \end{pmatrix}$$

Since admissibility requires $|\rho| < 1$, $|\phi| < 1$, this last expression means that ρ and ϕ are essentially inestimable.

It is to be noted that the large observed variances for $\hat{\rho}$ and $\hat{\phi}$ are not accidental: if we had found $\hat{\rho} = -.36$, $\hat{\phi} = -.36$, then $I(\hat{\gamma})$ would have been singular, and the variances would have been infinite. In fact, if we fix $\hat{\rho}$ at $-.36$ and vary $\hat{\phi}$, we get a smooth progression from reasonable variance estimates to absurdly large ones.

$\hat{\phi}$	Estimated variance of $\hat{\phi}$
0	.0787
−.2	.33
−.3	2.06

One might conclude from this example that the estimated variances given by $H^{-1}(\hat{\gamma})$ are absurd.

In this context Wall (1973) has suggested looking at the estimated correlation matrix for ρ and ϕ, this is

$$\begin{pmatrix} i & .99041 \\ & 1 \end{pmatrix} \text{ for } H^{-1}(\hat{\gamma})$$

and

$$\begin{pmatrix} 1 & .99859 \\ & 1 \end{pmatrix} \text{ for } I^{-1}(\hat{\gamma}).$$

This indicates at once that the estimates for ρ and ϕ are unreliable, since they are so highly correlated. We could also look at the condition number for the covariance matrix of ρ and ϕ. For H^{-1} the eigenvalues are .0033505, .695021, the condition number 207: for I^{-1} .00448, 6.33525 and 1,414. The condition numbers for the correlation matrices are 208 for H^{-1} and 1,417 for I^{-1}. So we see that in fact the estimated covariance matrix is nearly singular, for H^{-1} as well as I^{-1}; this indicates that the parameters are "nearly inestimable". That is, we can reasonably conjecture that the estimated variances given by H^{-1} are much too small.

This example points out that blind acceptance of variances estimated from H^{-1}, without examination of correlation coefficients, eigenvalues or condition numbers, can be quite misleading for this class of problems.

V. Variance of Forecasts

The results of section II and III are easily extended to the forecasting case, if we take the view that the forecasts merely use additional unknown parameters to be estimated via maximum likelihood. We maintain the notation of section I, but we now assume that $y_n, y_{n-1}, \ldots, y_{n-t+1}$ are unknown, and to be estimated. Formally, the expression 2-3 still holds, hence 2-4 and 2-5 are still correct, with the understanding that $\hat{y}_n, \ldots, \hat{y}_{n-t+1}$ must be used in the computation of $f(\beta)$.

We now have the additional t normal equations

(5-1) $$\frac{\partial f^T(\beta)}{\partial y_{n-q}} V^{-1}(\theta) f(\beta) = 0, q = 0, \ldots, t-1:$$

(5-2) $$\left[\text{Note that from 1-1 } \frac{\partial f^T(\beta)}{\partial y_{n-q}} = (0, 0 \ldots 1 \underbrace{0 .. 0}_{\substack{q-1 \\ \text{entries}}}\right].$$

The additional second derivative terms are given by

$$\frac{\partial^2 \log L}{\partial y_{n-q} \partial \beta_j} = \frac{\partial f^T(\beta)}{\partial \beta_j} V^{-1}(\theta) \frac{\partial f(\beta)}{\partial y_{n-q}} \tag{5-3}$$

$$\frac{\partial^2 \log L}{\partial y_{n-q} \partial \theta_m} = \frac{\partial f^T(\beta)}{\partial y_{n-q}} \frac{\partial V^{-1}(\theta)}{\partial \theta_m} f(\beta) \tag{5-4}$$

$$\frac{\partial^2 \log L}{\partial y_{n-q} \partial y_{n-r}} = \frac{\partial f^T(\beta)}{\partial y_{n-q}} V^{-1}(\theta) \frac{\partial f(\beta)}{\partial y_{n-r}} \tag{5-5}$$

We note that (5-4) has expectation 0, and (5-3) and (5-5) are not stochastic. In particular, recalling (5-2), we see that

$$\frac{\partial^2 \log L}{\partial y_{n-q} \partial y_{n-r}} = V^{n-q,n-r}(\theta) \tag{5-6}$$

The expression (2-9) must now be modified to take into account the fact that we have only $n - t$ observations; essentially this means that the estimates for β and γ are only based on $Y_1, \ldots, y_{n-t}$, and, with this proviso, (2-9) is still correct, so we have

$$(5\text{-}7)\ I(\gamma) = \begin{bmatrix} \frac{1}{\sigma^2}[f']^T V^{-1} f' & 0 & \frac{1}{\sigma^2} [f']^T V^{-1} \frac{\partial f}{\partial y_{n-r}} \\ 0 & \frac{1}{2} \operatorname{trace} \left(V \frac{\partial V^{-1}}{\partial \theta\ \theta_m} \right) & 0 \\ \frac{1}{\sigma^2} \frac{\partial f^T}{\partial y_{n-q}} V^{-1} [f'] & 0 & \frac{1}{\sigma^2} V^{n-q,n-r} \end{bmatrix}$$

For simplicity, we now assume that $m(\beta) = 0$, so the variance of a forecast is given by inverting the appropriate segment of V^{-1}.

In particular, for the pure autoregressive case, $V^{-1} = P^T P$ so we see from 1-5 that

$$\operatorname{Var}(\hat{y}_n) = \sigma^2 1,$$

if $t = 1$. That is, the one step ahead asymptotic prediction variance is always σ^2, regardless of the order of the process. This result follows at once from (1-3), since, asymptotically, we know $\rho_1, \ldots, \rho_a$ exactly.

Similarly, the two step ahead asymptotic covariance matrix is given by

$$\sigma^2 \begin{pmatrix} 1 + \rho_1^2 & -\rho_1^{-1} \\ -\rho_1 & 1 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 + \rho_1^2 \end{pmatrix},$$

So the two step ahead variance is $\sigma^2(1 + \rho_1^2)$. This also follows at once from (1-3).

Similar results can be obtained for the moving average case, but the expressions quickly become more complicated.

One drawback of the asymptotic formula (5-7) is that the variance of the estimated γ parameters is not taken into account. We may use the δ method (see Rao (1965)) to derive approximations which include the $1/n$ term due to the variability of $\hat{\gamma}$ and which are conditional on $y_1, \ldots, y_{n-1}$. We illustrate the general approach with two examples. In the first order autoregressive case we have

$$\hat{\hat{y}}_i = \hat{\rho} y_{i-1},$$

and hence

$$(4\text{-}8) \qquad (\hat{y}_i - y_i) = \hat{\rho} y_{i-1} - \rho y_{i-i} - \epsilon_i = y_{i-1}(\hat{\rho} - \rho) - \epsilon_i .$$

Since y_i is not observed, it is not used in estimating $\hat{\rho}$, and so $\hat{\rho}$ and ϵ_i are independent. Thus

$$\text{Var}\,(\hat{y}_i - y_i) = \sigma^2 + y_{i-1}^2 \,\text{Var}\,(\hat{\rho}).$$

From the results of section 3, we have

$$(5\text{-}9) \qquad \text{Var}\,(\hat{y}_i - y_i) = \sigma^2 + y_{i-1}^2 \frac{1}{(n-1) + (n-2)\rho^2 + (n-3)\rho^4 + \cdots + \rho^{2(n-2)}}$$

If $\rho = 0$, this reduces to

$$(5\text{-}10) \qquad \text{Var}\,(\hat{y}_i - y_i) = \sigma^2 + \frac{y_{i-1}^2}{n-1}.$$

For the two step ahead predictor we have

$$\hat{y}_i = \hat{\rho}\hat{y}_{i-1} = \hat{\rho} y_{i-2}^2,$$

and hence

$$(5\text{-}11) \qquad (\hat{y}_i - y_i) = \hat{\rho}^2 y_{i-2} - \hat{\rho}^2 y_{i-2} - \hat{\rho}\epsilon_{i-1} - \epsilon_i$$

So

$$(5\text{-}12) \qquad \text{Var}(\hat{y}_i - y_i) = \sigma^2(1 + \rho^2) + y_{i-2}^2 \,\text{Var}\,(\hat{\rho}^2 - \rho^2).$$

Since $\hat{\rho}^2$ converge to ρ^2 in probability at rate $1/\sqrt{n}$, we may write the expansion

$$\hat{\rho}^2 = \rho^2 + (\rho - \hat{\rho})2\rho + 0_p(1/\sqrt{n}).$$

So

$$(5\text{-}13) \qquad E(\hat{\rho}^2 - \rho^2) = \text{Var}\,(\hat{\rho}^2 - \rho^2) = 4\rho^2\,\text{Var}\,(\hat{\rho}).$$

Substituting (5-13) into (5-12), and using the results of section 3, we find

$$\mathrm{Var}(\hat{y}_i - y_i) = \sigma^2(1 + \rho^2) + y_{i-2}^2 4\rho^2 \frac{1}{(n-1) + (n-2)\rho^2 + \cdots + \rho^{2(n-2)}}$$

If $\rho = 0$, this reduces to

(5-14) $$\mathrm{Var}(\hat{y}_i - y_i) = \sigma^2(1 + \rho^2).$$

and we note that the $1/n$ term does not appear. This occurs because $\hat{\rho}^2$ converge to ρ^2 at a rate greater than $1/\sqrt{n}$, if $\rho = 0$.

Similar results can be derived in more general cases, by appropriate linearization and substitutions, but the more general expressions are difficult to interpret, and are not presented here.

National Bureau of Economic Research

Submitted January 1976
Revised June 1976

REFERENCES

Box, G. E. P. and G. M. Jenkins, (1970). Time Series Analysis. Holden-Day, San Francisco, CA.

Hill, R. W. (1975). "Certain Aspects of Generalized Box-Jenkins Models, N.B.E.R. Working Paper No. 82.

Holland, P. W. (1973). Personal Communication.

Rao, C. R. (1965). "Linear Statistical Influence and its Applications". John Wiley and Sons, New York, NY.

Wall, K. (1973). Personal Communication.