PREDICTIONS FROM ARMAX MODELS*

Richard T. BAILLIE
Wayne State University, Detroit, MI 48202, USA
University of Aston, Birmingham B4 6TE, UK

Received April 1977, final version received July 1979

This paper gives an expression for the minimum mean squared error predictor of the single equation ARMAX model when all the parameters are known. A formula is then derived for the asymptotic prediction mean squared error when the parameters are replaced by their maximum likelihood estimates. These general results are then specialised to the regression model with ARMA errors; for this case we also consider the properties of an alternative predictor which can sometimes be marginally more efficient than the conventional predictor.

1. Introduction

The second section of this paper considers the single equation ARMAX model and given known parameters, presents a form for expressing the minimum mean squared error multistep predictor.

However, as noted by several authors, for example Pierce (1975, p. 371), the usual formulae for prediction mean squared error are inadequate in the sense that they assume known parameter values. We specifically remedy this situation by giving a parametric expression for asymptotic prediction mean squared error (amse) from the ARMAX model when the parameters are replaced by their maximum likelihood estimates. These results are then specialised in section 3 to the regression model with ARMA errors.

Section 4 considers the properties of an alternative predictor for the regression model with ARMA errors that is based on efficient estimates of the regression parameters but ignores autocorrelation of the errors. It transpires that in certain circumstances this predictor can be more efficient than the conventional predictor which takes into account error term autocorrelation. This rather surprising result conflicts with the findings of Goldberger (1962) who showed that with known error process parameters the most efficient predictor always allowed for autocorrelation. Our results

*This article is part of a chapter in the author's Ph.D. thesis at the University of London. The author is extremely grateful to his thesis supervisors J. Durbin and K.F. Wallis; and to D.F. Hendry, G.E. Mizon and J.D. Sargan for several helpful discussions. The comments of two anonymous referees are also gratefully acknowledged; although the usual caveats apply. A much earlier version of this paper was presented under a different title to the Econometric Society's European Meeting at Vienna in 1977.
essentially occur because the gain in efficiency in allowing for autocorrelation can be offset by having to estimate the parameters of the ARMA process. This situation is illustrated with a numerical example.

2. Prediction from the general ARMAX model

The single equation ARMAX model with endogenous variable $y_t$, exogenous variables $x_{1t}, \ldots, x_{kt}$, and moving average error process is

$$a(L)y_t = \sum_{i=1}^{k} \eta_i(L)x_{it} + \theta(L)e_t,$$  \hspace{1cm} (1)

where $L$ is the lag operator, $e_t$ is a white noise process with constant variance $\sigma_e^2$,

$$a(L) = 1 - \sum_{j=1}^{p} \alpha_j L^j,$$
$$\theta(L) = 1 - \sum_{j=1}^{q} \theta_j L^j,$$
$$\eta_i(L) = \eta_{i0} - \sum_{j=1}^{s_i} \eta_{ij} L^j.$$

It is also assumed that all the roots of $a(L)$ and $\theta(L)$ lie outside the unit circle. Astrom (1970, p. 167) has derived a partial moving average expansion for an ARMA model. By slightly modifying his approach, (1) can be expressed as

$$y_{n+1} = \sum_{j=0}^{l-1} \psi_j e_{n+1-j} + \sum_{i=1}^{k} \mu(L)\eta_i(L)\theta(L)^{-1} x_{in+1} + \{1 - \mu(L)a(L)\theta(L)^{-1}\} L^{-1} y_n,$$  \hspace{1cm} (2)

where

$$\theta(L)a(L)^{-1} = \sum_{j=0}^{\infty} \psi_j L^j \quad \text{and} \quad \mu(L) = \sum_{j=0}^{l-1} \psi_j L^j.$$

Eq. (2) can also be written as

$$y_{n+1} = \sum_{j=0}^{l-1} \psi_j e_{n+1-j} + \sum_{i=1}^{k} v_i(l)' X_{in+1} + \zeta(l)' Y_n,$$  \hspace{1cm} (3)

where
\[ v_{i(l)}' = [v_{i0}(l) v_{i1}(l) v_{i2}(l) \ldots], \]
\[ \xi(l)' = [\xi_{0}(l) \xi_{1}(l) \xi_{2}(l) \ldots], \]
\[ X_{in+l} = [x_{in+l} x_{in+l-1} \ldots x_{in} x_{in-1} \ldots], \]
\[ Y_{n} = [y_{n} y_{n-1} y_{n-2} \ldots]. \]

At time \( n \) the \( l \) step predictor is given by

\[ y_{n,l} = \sum_{i=1}^{k} v_{i(l)}' X_{in+l} + \xi(l)' Y_{n}. \]  \hspace{1cm} (4)

For subsequent analysis it is convenient to obtain parametric expressions for the prediction weights \( v_{i(l)} \) and \( \xi(l) \). We first express the model (1) in companion form as

\[ (I - AL) Y_{i}^* = \sum_{i=1}^{k} (C_{i} - D_{i}L) X_{it}^* + (I - BL) \epsilon_{i}^*, \]  \hspace{1cm} (5)

where

\[
A = \begin{bmatrix}
\alpha_1 & \alpha_2 & \ldots & \alpha_m \\
1 & 0 & \ldots & 0 \\
0 & 1 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & 1 & 0
\end{bmatrix}, \quad C = \begin{bmatrix}
\eta_{10} & 1 & 0 \\
0 & 1 & \ldots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 0
\end{bmatrix}
\]

\( B \) and \( D \) are of the same form as \( A \), except their first rows consist of \((\theta_1 \ldots \theta_m)\) and \((\eta_{11} \ldots \eta_{lm})\) respectively,

\[ Y_{i}^* = [y_{i} y_{i-1} \ldots y_{i-m+1}], \]
\[ X_{it}^* = [x_{it} x_{it-1} \ldots x_{it-m+1}], \]
\[ \epsilon_{i}^* = [\epsilon_{i} 0 \ldots 0]. \]

\( I \) is the unit matrix of order \( m \) and \( m = \max(p, q, s_1, \ldots, s_n) \). Zero restrictions can be inserted on the appropriate coefficients of any polynomial of order less than \( m \).

Thus the first row of (5) realises (1) and every other row is an identity. The advantage with this representation is that by following the approach of Yamamoto (1975) for the pure ARMA process, it can be shown that
\[ v_{ij}(l) = \eta_{i0}, \quad j = 0, \]
\[ = e'(AC_i - D_i)A^{l-1}e, \quad j = 1, 2, \ldots, l - 1, \quad (6) \]
\[ = e'(D_i - BC_i)A^{l-1}B^{l-1}e, \quad j = l, l + 1, \ldots, \]

and

\[ \xi_j(l) = e'(A - B)A^{l-1}B^{l-1}e, \quad j = 0, 1, \ldots, \quad (7) \]

where \( e' = [1 \ 0 \ \ldots \ 0] \) is of dimension \( 1 \times m \).

We now want to consider the situation where \( (1) \) has been correctly identified and maximum likelihood estimates obtained for the set of parameters

\[ \gamma' = (\alpha_1 \ldots \alpha_p \ \theta_1 \ldots \theta_q \ \eta_{10} \ldots \eta_{kS_k}), \]

which can be used to estimate the predictor weights \( (6) \) and \( (7) \), which can in turn be used to construct the practical predictor

\[ \hat{y}_{n,l} = \sum_{i=1}^{k} \hat{\delta}_i(l)'X_{in+l} + \hat{\xi}(l)'Y_n. \quad (8) \]

The \( l \) step prediction error is

\[ e_{n,l} = y_{n+l} - \hat{y}_{n,l}, \]

and on taking Taylor series expansions on \( \hat{\delta}_i(l) \) and \( \hat{\xi}(l) \) around their true values, we obtain

\[ e_{n,l} = \sum_{j=0}^{l-1} \psi_j e_{n+l-j} - (\hat{\gamma} - \gamma)' \{ G(l)X_{n+l} + D(l)Y_n \}, \]

where

\[ X_{n+l} = (X_{1n+l} \ldots X_{kn+l}), \]
\[ G(l) = \{ G_1(l) \ldots G_k(l) \}, \]

and

\[ G_i(l) = [\delta v_i(l)/\delta \gamma] \quad \text{and} \quad D(l) = [\delta \xi(l)/\delta \gamma]. \]

By following the approach of Schmidt (1974) and Yamamoto (1976) a parametric expression can be obtained for \( G(l) \), and is given in the appendix. A parametric expression for the first \( p + q \) rows of \( D(l) \) has been previously obtained by Yamamoto (1975); the remaining rows are entirely full of zeros.
By denoting the expectation operator as $E$ and writing the parameter estimates asymptotic covariance matrix as

$$E\{\sqrt{n}(\hat{\gamma} - \gamma)\sqrt{n}(\hat{\gamma}' - \gamma)\}' = \frac{\sigma^2}{n} \Omega,$$

where $n$ is the number of observations in the sample, and by noting that parameter estimates and observations used in the predictor are asymptotically independent to order $o(n^{-1})$ we obtain the following expression for the prediction amse:

$$\text{amse}(\hat{y}_{n+1}) = \sigma^2 \sum_{j=0}^{l-1} \psi_j^2 + \frac{\sigma^2}{n} \text{tr}\{G(l)E(X_{n+1}X'_{n+1})G(l)' + D(l)E(Y_nY_n')D(l)' + 2D(l)E(Y_nX_{n+1}')G(l)\} \Omega. \quad (9)$$

It should be noted that both the predictor (8) and its amse (9) are formulated in terms of an infinite number of observations whereas in practice they have to be computed from a sample of $n$ observations. The validity of this procedure and in particular the convergence of individual terms in (9) depends upon $n$ being large enough for asymptotic theory to be invoked and for $\alpha(L)$ and $\theta(L)$ to possess no root lying close to the unit circle. Provided these conditions are satisfied, individual terms within (9) typically converge for a number of observations considerably less than $n$. Further details of this, based on some numerical examples are given by Baillie (1978).

3. The regression model with ARMA errors

We now consider the model

$$y_t = x_t'\beta + u_t, \quad \phi(L)u_t = \theta(L)e_t, \quad (10)$$

where $x_t$ is a $1 \times k$ vector containing $k$ exogenous variables at time $t$, $\beta$ is a $k \times 1$ vector of parameters and $u_t$ follows an ARMA $(p, q)$ process. This model can readily be seen to be a particular case of (1) with

$$\alpha(L) = \phi(L), \quad n_{10} = \beta_1 \quad \text{and} \quad n_i(L) = \phi(L), \quad i = 1, \ldots, k.$$

The practical predictor (8), based on maximum likelihood estimates becomes

$$\hat{y}_{n+1} = x_{n+1}'\hat{\beta} + \xi' (Y_n - X_n\hat{\beta}), \quad (11)$$
and the \( l \) step prediction error is then

\[
e_{n+l} = \sum_{j=0}^{l-1} \psi_j \hat{e}_{n+j} - x'_{n+l}(\hat{\beta} - \beta) - (\hat{\rho} - \rho)H(l)(Y_n - X_n\beta)
\]

\[+ \xi(l)'X_n(\hat{\beta} - \beta),\]

where \( \rho' = (\phi' \theta)' \) and \( H(l) \) is the \( p+q \) submatrix in the top left-hand corner of \( D(l) \). Pierce (1971) has shown that the maximum likelihood estimates of the regression parameters and the error process parameters are asymptotically independent. By denoting the asymptotic covariance matrices of \( \hat{\beta} \) and \( (\hat{\phi}' \hat{\theta}) \) as \( (\sigma^2/n)B^{-1} \) and \( (\sigma^2/n)V^{-1} \), respectively; it can be shown that

\[
\text{amse} \left( \hat{\phi}_{n,l} \right) = \sigma^2 \sum_{j=0}^{l-1} \psi_j^2 + \sigma^2 \left\{ x'_{n+l}B^{-1}x_{n+l} - 2x'_{n+l}B^{-1}X_n\xi(l) \right\}
\]

\[+ \xi(l)'X_nB^{-1}X_n'\xi(l) + \text{tr}H(l)\Gamma H(l)'V^{-1}, \quad (12)\]

where

\[\Gamma = \text{E}(U_n' U_n') \quad \text{and} \quad U_n' = [u_n, u_{n-1}, \ldots].\]

The first term in (12) is due to the random innovations in the forecast period, the next three terms are due to the estimation of the regression parameters and the last term is due to estimating \( \rho \) and is identical to the parameter estimation term for the univariate ARMA model obtained by Yamamoto (1975). When \( q=0 \) and just AR\((p)\) errors are present the matrices \( X_n, \xi(l), H(l) \) and \( \Gamma \) reduce in dimension to \( p \times k, p \times 1, p \times p \) and \( p \times p \) respectively; and \( V = \Gamma \) and (12) becomes the result given by Baillie (1979).

4. Properties of a structural predictor

We now consider a predictor of model (10) that is based on efficient estimates of the regression parameters and only takes into account the structural part of the model. Thus the stochastic properties of \( u_t \) are considered in estimating \( \beta \) but are omitted from the predictor. Hence the predictor (11) simplifies to

\[
\hat{y}_{n,l} = x'_{n+l}\hat{\beta}, \quad (13)
\]

which we call the structural predictor. The \( l \) step prediction error is found to be

\[
\hat{e}_{n,l} = \sum_{j=0}^{l-1} \psi_j \hat{e}_{n+l-j} - x'_{n+l}(\hat{\beta} - \beta) + \xi(l)'U_n,
\]
and after some algebra we find that

\[
\text{amse}(\hat{y}_{n,t}) = \sigma^2 \sum_{j=0}^{t-1} \psi_j^2 + \frac{\sigma^2}{n} \{x'_{n+1} B^{-1} x_{n+1} - 2 x'_{n+1} B^{-1} X'_n \zeta(l)\} + \zeta(l) \Gamma \zeta(l). \tag{14}
\]

It follows that the gain in efficiency through using the predictor (11) instead of the structural predictor (13) is given to order \(O(n^{-1})\) by subtracting (14) from (12) to obtain

\[
\xi(l) \left\{ \Gamma - \frac{\sigma^2}{n} X_n B^{-1} X'_n \right\} \xi(l) - \frac{\sigma^2}{n} \text{tr}\{H(l) \Gamma H(l) V^{-1}\}. \tag{15}
\]

The first part of the above expression corresponds to the formula given by Goldberger (1962) who noted that the predictor allowing for autocorrelation would always be more efficient than the structural predictor. However, Goldberger's results assume known error process parameters and in general when they have been estimated it is necessary to also consider the second term of (15). Our formula (15) shows that situations may arise where the gain in efficiency obtained by taking into account autocorrelation can be offset by having to estimate the \(\phi\) and \(\theta\) parameters. To illustrate we consider the following numerical example from the model

\[
y_t = 2x_t + u_t, \tag{16}
\]

\[
u_t = 0.250u_{t-1} + 0.375u_{t-2} + \varepsilon_t, \quad \sigma^2 = 1.
\]

We also assume that \(x_t\) is stationary and is generated by the AR(1) process,

\[
x_t = 0.500x_{t-1} + \alpha_t, \quad \sigma^2 = 1.
\]

We note that \(u_t\) is stationary and has variance 1.3853. In order to safeguard against a 'freak' realisation of the exogenous variable, (12) was evaluated unconditionally by substituting the stochastic properties of \(x_t\) This implied a variance for \(\beta\) of 1.75/n, and taking \(n\) to be 40 we found the results given in tables 1 and 2.

We note that the gain in efficiency through allowing for autocorrelation is considerable in the first two periods and decreases as the effect of autocorrelation dies out for longer lead times. After seven periods the estimation of the error process parameters becomes of more importance than allowing for autocorrelation.
Table 1
Prediction amse from model (16) using the true predictor (11).

<table>
<thead>
<tr>
<th>( l )</th>
<th>( \sum_{j=0}^{l-1} \hat{\psi}_j^2 )</th>
<th>Due to estimation of ( \beta )</th>
<th>Due to estimation of ( \phi )</th>
<th>Total prediction amse</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0000</td>
<td>0.0118</td>
<td>0.0500</td>
<td>1.0618</td>
</tr>
<tr>
<td>2</td>
<td>1.0625</td>
<td>0.0518</td>
<td>0.0348</td>
<td>1.1490</td>
</tr>
<tr>
<td>3</td>
<td>1.2539</td>
<td>0.0546</td>
<td>0.0452</td>
<td>1.3537</td>
</tr>
<tr>
<td>4</td>
<td>1.2952</td>
<td>0.0587</td>
<td>0.0373</td>
<td>1.3912</td>
</tr>
<tr>
<td>5</td>
<td>1.3413</td>
<td>0.0586</td>
<td>0.0316</td>
<td>1.4315</td>
</tr>
<tr>
<td>6</td>
<td>1.3582</td>
<td>0.0588</td>
<td>0.0223</td>
<td>1.4393</td>
</tr>
<tr>
<td>7</td>
<td>1.3710</td>
<td>0.0586</td>
<td>0.0185</td>
<td>1.4481</td>
</tr>
<tr>
<td>8</td>
<td>1.3769</td>
<td>0.0585</td>
<td>0.0135</td>
<td>1.4490</td>
</tr>
</tbody>
</table>

Table 2
Gain in efficiency through using (11) rather than (13), when predicting from model (16).

<table>
<thead>
<tr>
<th>( l )</th>
<th>( \xi(l)/\Gamma \xi(l) )</th>
<th>( \frac{\sigma^2}{n} \xi(l)'X_{e}B^{-1}X_{e}'\xi(l) )</th>
<th>( \frac{\sigma^2}{n} H(l)'H(l)V^{-1} )</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.3853</td>
<td>0.0118</td>
<td>0.0500</td>
<td>0.3234</td>
</tr>
<tr>
<td>2</td>
<td>0.3228</td>
<td>0.0117</td>
<td>0.0348</td>
<td>0.2763</td>
</tr>
<tr>
<td>3</td>
<td>0.1314</td>
<td>0.0040</td>
<td>0.0452</td>
<td>0.0822</td>
</tr>
<tr>
<td>4</td>
<td>0.0901</td>
<td>0.0030</td>
<td>0.0373</td>
<td>0.0498</td>
</tr>
<tr>
<td>5</td>
<td>0.0440</td>
<td>0.0014</td>
<td>0.0316</td>
<td>0.0110</td>
</tr>
<tr>
<td>6</td>
<td>0.0271</td>
<td>0.0009</td>
<td>0.0223</td>
<td>0.0039</td>
</tr>
<tr>
<td>7</td>
<td>0.0143</td>
<td>0.0005</td>
<td>0.0185</td>
<td>-0.0005</td>
</tr>
<tr>
<td>8</td>
<td>0.0084</td>
<td>0.0003</td>
<td>0.0135</td>
<td>-0.0054</td>
</tr>
</tbody>
</table>

5. Conclusion

In our derivation of prediction amse from various dynamic models we have assumed that the exogenous variables are known in the forecast period. In fact all the formulae we have developed can be appropriately extended at the cost of computational complexity, to cover the situation of the exogenous variables being forecast. It is also possible to modify the prediction amse formulae for the situation where (1) embodies parameter restrictions representing the fact that the estimated model is a transfer function or an autoregressive final form equation from a dynamic simultaneous equation model, as given by Zellner and Palm (1974). Further details of these situations are given by Baillie (1978).
Appendix

By writing $G_i(l) = [g_{i0}(l) \ g_{i1}(l) \ g_{i2}(l) \ ...]$ it can be shown that $g_{i0}(l)$ is a column vector composed entirely of zeros, except for a unit element in the position:

\[
(p + q + \sum_{h=1}^{l-1} s_h + i)
\]

Compactly expressing the other results, we obtain

\[
g_{ij}(l) = \begin{bmatrix}
    n_{i0} & 0' \\
    0 & M_{p-1}
\end{bmatrix} A^{j-1} e + N_p \sum_{k=0}^{j-2} A^n \otimes A^{j-2-k} (AC_i - D_i)'(e \otimes e)
\]

where

\[
N_p = (I_p \ 0 \ ... \ 0), \quad p < m,
\]

so that $M_p$ is $p \times m$, $N_p$ is $p \times m^2$ and the $(p+1)$th to $(p+q+\sum_{h=1}^{l-1} s_h + i - 1)$ elements of $g_{ij}(l)$ are composed entirely of zeros, as are the $(p+q+\sum_{h=1}^{l} s_h + i + 1)$ to $(p+q+\sum_{h=1}^{l} s_h + k)$ elements, for $j = 1, \ldots l-1$. 

\[
g_{ij}(l) = \begin{bmatrix}
    -N_p \sum_{h=0}^{l-2} A^{ih} \otimes A^{j-2-h} (D_i - BC_i)'(e \otimes e) \\
    -n_{i0} & 0' \\
    0 & M_{q-1}
\end{bmatrix} A^{i-1} B^{j-1} e - N_q \sum_{h=0}^{j-1} B^{ih} \otimes B^{j-1-h} \{A'^{i-1} (D_i - BC_i)'(e \otimes e) \}
\]

where

\[
\begin{align*}
0, e' A^{i-1} B^{j-1} e \\
-M_{s_i} A^{i-1} B^{j-1} e
\end{align*}
\]

for $j = l, l + 1, \ldots$
References

Yamamoto, T., 1975, Asymptotic mean square error of multi step prediction from mixed ARMA model, Discussion Paper no. 7521 (CORE, Université de Louvain).