Fitting a Linear Autoregressive Model for Long-Range Forecasting

C. S. YAO

Department of Meteorology, Nanjing University, People's Republic of China

(Manuscript received 10 February 1982, in final form 4 January 1983)

ABSTRACT

Methods of fitting a linear autoregressive model to a stationary time series are summarized. Parameters of the linear autoregressive model were estimated by the Durbin stepwise procedure and the order of this model was chosen by means of a t-test or F-test. An illustrative example used to forecast the monthly rainfall is also presented.

1. Introduction

An observed time series can be thought of as one particular realization of a stochastic process. A stochastic model is a formalized expression of the stochastic process. Stochastic models, such as autoregressive models and Markov chain models, are useful in meteorology.

A Markov process whose state space is discrete is called a Markov chain. Gabriel and Neumann (1962) first suggested a first-order Markov chain model to study Tel Aviv daily rainfall occurrences. The order of the Markov chain to be fitted to the Tel Aviv daily rainfall data is debatable (Gates and Tong, 1976; Katz, 1979). To test the goodness of fit of the Markov chain model to daily rainfall records, Lowry and Guthrie (1968), following the theory of Billingsley (1961) and the work of Hoel (1954), introduced a χ^2 test for the order of a Markov chain. Based on Akaike's Information Criterion (Akaike, 1974), Gates and Tong (1976) proposed a model-building procedure for estimating the order of the Markov chain model after the work of Tong (1975). Using several procedures Katz (1979) applied Akaike's Information Criterion and the Bayesian Information Criterion (Schwarz, 1978) to the optimum selection of the order of a Markov chain model and compared these two procedures. However, in order to improve the Markov property of minimal feedback in studying the probabilities of changes of wet days and dry days, Yao (1966) treated stochastic changes of weather as trials and took the effect of any past trials into consideration in the analyses of the current states. This technique made it possible to form a regular Markov chain with n-step transition without considering the problem of how to determine the order of the Markov chain.

Here the problem of fitting the linear autoregressive model to the observed series will be considered. An autoregressive model of order p will be abbreviated to an AR(p) model (Box and Jenkins, 1970). Yule (1927) developed the classical idea of periodicity and suggested that an AR process could be used to represent a harmonic oscillation disturbed by a stream of random external shocks. The realization of such a process is said to be stochastic in that the future value is partly determined by past values.

Among the methods of fitting the linear AR model to the stationary time series, the Yule-Walker equations are well known. In addition, the relevant parameters are usually estimated by a stepwise procedure which is due to Durbin (1960).

Box and Jenkins (1970) suggested a systematic method to estimate these parameters. They first find the initial estimates by means of the Yule-Walker equations and then adjust the initial estimates by an iterative procedure in order to get the least-squares estimates for the AR model.

Mann and Wald (1943) proved that the sampling properties of least-squares estimates in the AR model are asymptotically the same as those of least-squares regression estimators in multivariate normal situations. On the basis of this useful theorem, we can treat an estimation problem in AR models as one in ordinary regression. Thus, Chatfield (1980) showed that higher order AR models may be fitted by least squares in a manner similar to that involved in the fitting of ordinary regressive models. The well-known Yule-Walker equations can also be obtained by the use of the least-squares method.

A different approach to the autoregressive fitting was used by Jones (1964) who estimated the relevant parameters by means of the spectrum estimated by the periodogram. The periodogram is the finite Fourier transform of the autocovariance function.

Various tests are available for determining the order of the linear AR model. Quenouille (1947) used the partial autocorrelation coefficients to provide an

ingenious test of fit (see Kendall and Stuart, 1976b, pp. 502–505).

According to the Mann-Wald theorem, an F-test in the linear multiple regression analysis may be performed to choose the order of the AR model (Yevjevich, 1972). Zurndorfer and Glahn (1977), on the basis of previous work and further in consideration of the problem of correlations among predictors, performed a Monte Carlo technique for generating an F-statistic for testing the significance of regression models. The Monte Carlo F-value should be considered as a new approach to determining an appropriate order (i.e., without overfitting) of the linear AR model.

Another approach is to calculate the residual sum of squares on fitting extra terms in the AR model, and it may be possible to find the proper order of the model when there is little improvement in fit (Kendall and Stuart, 1976b; Chatfield, 1980). Yevjevich (1972) emphatically suggested that the method of whitening was likely to be most attractive for determining the order of the linear AR model, if the sequence of random errors did not differ significantly from a pure random sequence (white noise sequence) in the sense used by Yevjevich. Although we may assume that the sequence of random errors is a white noise sequence in the model [see Eq. (1) below], such an assumption is sometimes not appropriate and may lead to an inappropriate model, since the white noise process is a physically unrealizable phenomenon. As has been noted by Chatfield (1980), this conclusion has been given further foundation by building models III and IV described below, and these models are indeed inappropriate, though the sequences of random errors of these models more nearly approximate white noise.

The order of a linear AR model is often chosen by the method known as Akaike's Final Prediction Error (FPE), suggested by Akaike (1969). Another objective approach to identifying the order of this model is to minimize a quantity called Akaike's Information Criterion (AIC) proposed by Akaike (1974) or a quantity called the Bayesian Information Criterion (BIC) (Schwarz, 1978).

Shibata (1980) proposed an asymptotically efficient selection technique for choosing the order of AR models in which six methods of selection, such as AIC, BIC and others, were compared.

Recently, Carr (1980) developed new criteria L_1 and L_2 for determining the number of predictors included in a linear multiple regression as well as in a linear AR model.

These methods for either experimental approximations or quantitative criteria as applied to the selection of the appropriate order of a model are generally all similar in that they attempt to minimize the error and to satisfy the requirement of economy in the model building. It seems that the weak points of these usual methods of choosing the order of a model have in common the lack of a statement of the statistical significance of the parameters and the rather laborious calculations required.

For fitting an observed time series, Box and Jenkins (1970) suggested the partial autocorrelation function as a measure to choose the order of a linear AR process (model). According to Box and Jenkins, the AR model should be of order p, if the partial autocorrelation function is nonzero for $k \leq p$ and zero for $k > p$. Chatfield (1980) gave a 95% confidence interval of $\pm 2/\sqrt{n}$ for testing the partial autocorrelation. But $\mathrm{var} r_k \approx 1/n$ used by Chatfield is only an approximate result, assuming that all the population autocorrelation coefficients are zero except at lag zero.

In this paper we build on the work of Box and Jenkins (1970) and suggest that the order of the linear AR model can be chosen exactly by a t-test or F-test at a preselected significance level, the calculation of which is very straightforward. But it is important to recall that the t-test or F-test is only valid for normal independent random variables.

The monthly rainfalls in the rainy season used in this paper are normally distributed (Brooks and Carruthers, 1953; Yao, 1958, 1963). According to the Central Limit Theory, the normality condition required by the t-test or F-test may be shown to be satisfied for a large sample of independent random variables having the same distribution. Furthermore, the distribution of rainfall records varies with the length of the observation period. For instance, hourly rainfall has a highly skewed distribution, while the distribution of annual rainfall is only slightly skewed (quasi-normal). In general, the distribution of rainfalls approaches the normal distribution more closely as the period of observation increases (Brooks and Carruthers, 1953). A standard statistical technique to remove skewness is to apply a transformation (such as the logarithm, square root or cube root) to transform original observations of the skewed distribution into new ones having a quasi-normal distribution (Katz and Skaggs, 1981). In this connection, in addition to the required independence of variables the normality condition required by the t-test or F-test may also be satisfied by the technique of transformation.

2. Building of linear autoregressive models

Let y_t represent the deviation of the stationary stochastic process from the mean. The linear AR model representing the stationary time series (stochastic process) may then be written

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \epsilon_t. \quad (1)$$

The model (1) is regarded as a regression of the current value y_t on the past values $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$ of the process, and $\{\epsilon_t\}$ represents a sequence of normal random variable each with mean zero and variance σ_ϵ^2 (white noise). This model with parameters $\alpha_1, \alpha_2, \ldots, \alpha_p$ is called the linear AR(p) model.

First, the linear autoregressive parameters may be estimated by solving the equations

$$\left.\begin{aligned}\rho_1 &= \alpha_1 + \alpha_2\rho_1 + \cdots + \alpha_p\rho_{p-1}\\ \rho_2 &= \alpha_1\rho_1 + \alpha_2 + \cdots + \alpha_p\rho_{p-2}\\ &\cdots\cdots\cdots\cdots\cdots\cdots\\ \rho_p &= \alpha_1\rho_{p-1} + \alpha_2\rho_{p-2} + \cdots + \alpha_p\end{aligned}\right\}. \quad (2)$$

These are the Yule-Walker equations, which are a set of linear equations with the autoregressive parameters $\alpha_1, \alpha_2, \ldots, \alpha_p$ related to each other in terms of the autocorrelation coefficients $\rho_1, \rho_2, \ldots, \rho_p$.

When the Durbin stepwise procedure (Durbin, 1960) is used as a second method to estimate the parameters of the linear AR model, the recurrence formula gives

$$\left.\begin{aligned}\alpha_k^{(k)} &= \frac{\rho_k - (\alpha_1^{(k-1)}\rho_{k-1} + \alpha_2^{(k-1)}\rho_{k-2} + \cdots + \alpha_{k-1}^{(k-1)}\rho_1)}{1 - (\alpha_1^{(k-1)}\rho_1 + \alpha_2^{(k-1)}\rho_2 + \cdots + \alpha_{k-1}^{(k-1)}\rho_{k-1})}\\ \alpha_j^{(k)} &= \alpha_j^{(k-1)} - \alpha_k^{(k)}\alpha_{k-j}^{(k-1)}\\ j &= 1, 2, \ldots, k-1; \quad k = 1, 2, \ldots, p\end{aligned}\right\}, \quad (3)$$

where the upper subscript (k) denotes the order of this model and the lower subscript k the kth parameter. Thus $\alpha_k^{(k)}$ is the last parameter in a linear AR model of order k.

According to the Mann-Wald theorem, the parameters of a linear AR model may be estimated by virtue of partial autocorrelations, the results of which are just the same as those obtained from (2) or (3). Now, as a third method to estimate the parameters of the linear AR model, we let $y_t, y_{t-1}, y_{t-2}, \ldots, y_{t-p}$ be replaced by $y, y_1, y_2, \ldots, y_p$ respectively. Then we may write (1) in the form

$$y = \alpha_1 y_1 + \alpha_2 y_2 + \cdots + \alpha_p y_p + \epsilon, \quad (4)$$

where the subscripts $1, 2, \ldots, p$ signify the past values. The parameters of a linear AR model [(4)] may then be calculated by means of the following relations between partial autocorrelations and residual variances (Cramér, 1946):

$$\left.\begin{aligned}\alpha_1^{(k)} &= \rho_{1y\cdot 23\cdots k}\frac{\sigma_{y\cdot 23\cdots k}}{\sigma_{1\cdot 23\cdots k}}\\ \alpha_2^{(k)} &= \rho_{2y\cdot 13\cdots k}\frac{\sigma_{y\cdot 13\cdots k}}{\sigma_{2\cdot 13\cdots k}}\\ &\cdots\cdots\cdots\cdots\cdots\cdots\\ \alpha_k^{(k)} &= \rho_{ky\cdot 12\cdots k-1}\frac{\sigma_{y\cdot 12\cdots k-1}}{\sigma_{k\cdot 12\cdots k-1}},\\ &(k = 1, 2, \ldots, p),\end{aligned}\right\}, \quad (5)$$

where $\rho_{1y\cdot 23\cdots k}$ will be called the partial autocorrelation coefficient of y and y_1 with respect to $y_2, \ldots, y_k$, and $\sigma^2_{y\cdot 23\cdots k}$ the residual variance of y with respect to $y_2, \ldots, y_k$. The other ρ's and σ's are similarly defined.

Draper and Smith (1967) have used partial correlation coefficients to give a description of the forward selection procedure for selecting the best multiple regression model, but F-tests are still used for the building of the regression model. The author (Yao, 1977a,b) suggested a method for building the multiple regression model with the help of the t-test of the partial correlation coefficient throughout the whole procedure, the result being the same as that obtained from stepwise regression. Similarly, the parameters of a linear AR model may be estimated by (5) and the order of the model may be chosen by the t-test of the partial autocorrelation coefficient:

$$t = \frac{r_{ky\cdot 12\cdots k-1}}{\sqrt{1 - r^2_{ky\cdot 12\cdots k-1}}}\nu^{1/2}, \quad (6)$$

where the degrees of freedom are $\nu = n - 1 - k$, and $r_{ky\cdot 12\cdots k-1}$ denotes the estimated value of the coefficient of partial autocorrelation. Eq. (6) is of the same form as that for the t-test of the simple correlation, except that the degrees of freedom here are $n - 2 - (k + 1 - 2) = n - 1 - k$ (von Mises, 1964, p. 608).

For increasingly higher orders for an AR model, stepwi se calculations in (5) are indeed laborious. Fortunately, the following relations hold for the linear AR model:

$$\left.\begin{aligned}\sigma^2_{y\cdot 12\cdots k-1} &= \sigma^2_{1\cdot 23\cdots k} = \sigma^2_{k\cdot 12\cdots k-1}\\ \sigma^2_{y\cdot 23\cdots k} &= \sigma^2_{y\cdot 12\cdots k-1} = \sigma^2_{1\cdot 23\cdots k}\\ \sigma^2_{y\cdot 13\cdots k} &= \sigma^2_{1\cdot y2\cdots k-1} = \sigma^2_{2\cdot 13\cdots k}\\ &\cdots\cdots\cdots\cdots\cdots\cdots\end{aligned}\right\}. \quad (7)$$

But

Hence, it follows from (7) that the last equation of (5) may be expressed as

$$\alpha_k^{(k)} = \rho_{ky\cdot 12\cdots k-1}, \quad k = 1, 2, \ldots, p. \quad (8)$$

This relation states the fact that the last parameter $\alpha_k^{(k)}$ of a linear AR model of order k is just the partial autocorrelation coefficient at lag k (Box and Jenkins, 1970).

For example, if $k = 1$,

$$\alpha_1^{(1)} = \rho_{1y} = \rho_1.$$

If $k = 2$,

$$\alpha_2^{(2)} = \rho_{2y\cdot 1} = \frac{\rho_{2y} - \rho_{12}\rho_{1y}}{[(1 - \rho_{12}^2)(1 - \rho_{1y}^2)]^{1/2}} = \frac{\rho_2 - \alpha_1^{(1)}\rho_1}{1 - \alpha_1^{(1)}\rho_1}.$$

If $k = 3$,

$$\alpha_3^{(3)} = \rho_{3y\cdot 12} = \frac{\rho_{3y\cdot 1} - \rho_{23\cdot 1}\rho_{2y\cdot 1}}{[(1 - \rho^2_{23\cdot 1})(1 - \rho^2_{2y\cdot 1})]^{1/2}}$$

$$= \frac{(\rho_3 - \rho_1\rho_2) - \rho_1(\rho_1\rho_3 - \rho_2^2) + \rho_1(\rho_1^2 - \rho_2)}{(1 - \rho_1^2) - \rho_1(\rho_1 - \rho_1\rho_2) + \rho_2(\rho_1^2 - \rho_2)}$$

$$= \frac{\rho_3 - [\alpha_1^{(2)}\rho_2 + \alpha_2^{(2)}\rho_1]}{1 - [\alpha_1^{(2)}\rho_1 + \alpha_2^{(2)}\rho_2]}.$$

The above results are just the ones obtained from (3).

In view of (8), the sample coefficient of the partial autocorrelation can be tested by means of (6) with the purpose of choosing the order of a linear AR model. Substituting the sample coefficient of the partial autocorrelation of (8) into (6) yields

$$t = \frac{\hat{\alpha}_k^{(k)}}{[1 - (\hat{\alpha}_k^{(k)})^2]^{1/2}} \nu^{1/2}, \quad k = 1, 2, \ldots, p, \quad (9)$$

where the degrees of freedom are $\nu = n - 1 - k$, n being the number of pairs of observational data used to compute the partial autocorrelation coefficient at lag k. Thus, the order of a linear AR model can easily be chosen by the t-test according to a preselected significance level, if the last parameter $\alpha_k^{(k)}$ has been estimated for each step in the recurrence procedure. A detailed justification why we can use the t-test to find the order of a linear AR process with interrelationships will be discussed at the end of this paper.

It is known that the F value

$$F(1, n - k - 1) = \frac{r_{y\cdot 12\ldots k}^2 - r_{y\cdot 12\ldots k-1}^2}{1 - r_{y\cdot 12\ldots k}^2}(n - k - 1), \quad (10)$$

where $r_{y\cdot 12\ldots k}$ is the sample multiple correlation coefficient, can be used as a test statistic in a procedure of selecting predictors in the linear multiple regression model (Klein *et al.*, 1959; Burr, 1974, p. 385) as well as in the linear AR model. If F is above the critical value $F_{1-\alpha}$ at a certain preselected significance level, then the additional variable y_k is to be selected as the kth predictor, a linear AR model of order k being built.

From the relation (Kendall and Stuart, 1976a, p. 355)

$$r_{y\cdot 12\ldots k}^2 = 1 - (1 - r_{1y}^2)(1 - r_{2y\cdot 1}^2) \times (1 - r_{3y\cdot 12})\cdots(1 - r_{ky\cdot 12\ldots k-1}^2), \quad (11)$$

it follows that

$$F(1, n - k - 1) = \frac{r_{y\cdot 12\ldots k}^2 - r_{y\cdot 12\ldots k-1}^2}{1 - r_{y\cdot 12\ldots k}^2}(n - k - 1)$$

$$= \frac{(1 - r_{y1}^2)(1 - r_{y2\cdot 1}^2)\cdots(1 - r_{y(k-1)\cdot 12\ldots k-2}^2) - (1 - r_{y1}^2)(1 - r_{y2\cdot 1}^2)\cdots(1 - r_{yk\cdot 12\ldots k-1}^2)}{(1 - r_{y1}^2)(1 - r_{y2\cdot 1}^2)\cdots(1 - r_{yk\cdot 12\ldots k-1}^2)}(n - k - 1)$$

$$= \frac{r_{yk\cdot 12\ldots k-1}^2}{1 - r_{yk\cdot 12\ldots k-1}^2}(n - k - 1).$$

Therefore, from (8)

$$F(1, n - k - 1) = \frac{[\hat{\alpha}_k^{(k)}]^2}{1 - [\hat{\alpha}_k^{(k)}]^2}(n - k - 1) \quad (12)$$

has the F distribution with 1 and $n - k - 1$ degrees of freedom. The order of a linear AR model may also be chosen by the F-test. Thus, from (9) and (12), we can write

$$F(1, n - k - 1) = t^2(n - k - 1), \quad (13)$$

with $\nu_1 = 1$ and $\nu_2 = n - k - 1$ degrees of freedom. Thus, as shown in (9) and (12), it is more convenient to combine (3) with (8), and choose the order of a linear AR model by means of either a t-test or an F-test, provided that the last parameter has been estimated for each step in the recurrence procedure (3).

3. Practical fitting

To illustrate the above fitting method for a linear AR model, we make the following calculation which should serve adequately to demonstrate its practicality. A time series of June rainfall records over 40 years (1921–1960) for Shanghai is given in Table 1. The first 30 values of this time series were used for fitting the linear AR model and the last 10 for discussing forecasting effectiveness of models of different orders.

For the purpose of showing that the 30-year time series (1921–50) for Shanghai is stationary, the trend equation

$$Y = \alpha + \beta t' + \epsilon_{t'}$$

is fitted to the data. Here, α and β are parameters, $\epsilon_{t'}$ the random variable, and t' the observation number, such as -5, -3, -1, 1, 3, 5, etc., if an artificial time origin is used. It has been found that

$$\left.\begin{aligned} \hat{\alpha} &= \bar{Y} = 180.06, \\ \hat{\beta} &= 0.117, \qquad \mathrm{var}\hat{\beta} = 0.939 \\ \hat{\sigma}^2 &= 8444, \qquad t = 0.12 \end{aligned}\right\}.$$

It can be said that the trend is highly insignificant

TABLE 1. June rainfalls (mm) in Shanghai.

Years	0	1	2	3	4	5	6	7	8	9
192		256.9	230.8	165.5	234.7	42.0	251.2	205.5	215.3	70.3
193	178.5	139.9	181.9	110.4	42.1	217.1	111.9	112.6	468.9	103.8
194	93.6	292.3	198.0	152.3	140.4	327.7	89.0	233.7	142.2	153.2
195	240.2	134.5	176.0	142.6	288.2	167.6	244.5	212.3	91.8	105.2
196	210.0									

[$t \ll t_{0.95}(28)$] and this time series, being stationary, can be used for fitting linear AR models.

Using the stationary time series of June rainfall records of 30 years (1921–50) in Shanghai, the sample values of autocorrelation coefficients have been calculated as follows:

$$r_1 = -0.339 \quad r_2 = -0.137 \quad r_3 = 0.248$$

$$r_4 = -0.0393 \quad r_5 = -0.116 \quad r_6 = 0.0705.$$

As a first step, from (3), (9) and (12), we found that

$$\left.\begin{aligned} &\hat{\alpha}_1^{(1)} = -0.34, \\ &t = -1.87, \quad t < -t_{0.95}(27); \\ &F = 3.51, \quad F > F_{0.90}(1, 27) \end{aligned}\right\}.$$

A first-order linear AR model is of the form

$$Y_t = 241.11 - 0.34Y_{t-1}. \tag{I}$$

The calculation of (3) should be continued, if a 0.10 level of significance for the one-tailed t-test was selected. As a second step, from (3), (9) and (12), we found that

$$\left.\begin{aligned} &\hat{\alpha}_1^{(2)} = -0.44, \quad \hat{\alpha}_2^{(2)} = -0.28 \\ &t = -1.48, \quad t < -t_{0.90}(25) \\ &F = 2.20, \quad F > F_{0.80}(1, 25) \end{aligned}\right\}.$$

Therefore, based on the significance test of the parameter, a linear AR model of order 2 can be fitted as

$$Y_t = 309.70 - 0.44Y_{t-1} - 0.29Y_{t-2}, \tag{II}$$

where Y_t denotes the current June rainfall used as the predictand and Y_{t-1} and Y_{t-2} denote the past two years' June rainfall used as the predictors for the June rainfall prediction in Shanghai.

It is also instructive to see if a linear AR model of order 3 could be fitted to data of the June rainfall in Shanghai. Thus the following calculations were carried out:

$$\left.\begin{aligned} &\hat{\alpha}_1^{(3)} = -0.40, \quad \hat{\alpha}_2^{(3)} = -0.24, \quad \hat{\alpha}_3^{(3)} = 0.11 \\ &t = 0.55, \quad t > t_{0.70}(23) \\ &F = 0.30, \quad F < F_{0.50}(1, 23) \end{aligned}\right\}.$$

The linear AR model of order 3,

$$Y_t = 274.57 - 0.40Y_{t-1} - 0.24Y_{t-2} + 0.11Y_{t-3}, \tag{III}$$

could not be chosen on the basis of a significance level of 0.10 for the one-tailed t-test. Furthermore, an attempt might be made to identify a model of order 4, specifically,

$$Y_t = 251.64 - 0.41Y_{t-1} - 0.22Y_{t-2} + 0.15Y_{t-3} + 0.08Y_{t-4}, \tag{IV}$$

if we only pay attention to the problem of the goodness of fit.

It is well known that the improvement of predicability depends on the significance of autocorrelation in the time series of June rainfalls, provided that the whole series, composed of both data for fitting and data for forecasting, is stationary and ergodic also in autocorrelations. It can be concluded that model IV identified by considering the goodness of fit without a significance test may be less efficient for prediction. As a matter of fact, the t-test of model IV gave

$$t = 0.38, \quad t < t_{0.70}(21);$$

and accordingly model IV should not be adopted.

The probability of rejecting an alternative hypothesis is usually large, if the significance level α is small. Therefore, when the power of the test is not considered in meteorological problems, the value of $\alpha = 0.10$ is often selected, in order that a most powerful test may be constructed (compare Thom and Thom, 1972). Consequently, model II for purposes of long-range prediction is superior, if the significance level $\alpha = 0.10$ for the t-test or $\alpha = 0.20$ for the F-test is adopted based on the work of Thom and Thom.

In order to choose the order of a linear AR model, Akaike calculated the estimate of the Final Prediction Error at each step:

$$\text{FPE}_k = \hat{\sigma}_k^2[1 + (k + 1)/n], \quad k = 1, 2, \ldots, p. \tag{14}$$

If the estimated linear AR model is unbiased, FPE is the expected independent sample mean square error.

The recurrence formula for the residual sum of squares is

$$S_k = S_{k-1}[1 - (\hat{\alpha}_k^{(k)})^2]. \tag{15}$$

The mean square error, when fitting the linear AR

model of order k (Jones, 1975), is

$$\hat{\sigma}_k^2 = S_k(n-k-1)^{-1}, \tag{16}$$

that is to say, $S_k(n-k-1)^{-1}$ is the sample residual variance of y with respect to $y_1, y_2, \ldots, y_k$. The initial value of S_k is $S_0 = (n-1)\hat{\sigma}^2$, and $\hat{\sigma}^2$ is the unbiased sample variance of y_t.

It was Jones' (1975) experience that all orders for fitting the model to daily data were usually up to some maximum such as 25 or 50. The order for which FPE_k is a minimum is selected as the best fit.

Applying Akaike's Final Prediction Error to choose the order of a linear AR model for June rainfall in Shanghai (Table 1), it was found that Eqs. (15), (16) and (14) gave

$$\text{FPE}_1 = 7975, \quad \text{FPE}_3 = 8279,$$

$$\text{FPE}_2 = 7838, \quad \text{FPE}_4 = 8801.$$

Since the Final Prediction Error is a minimum for model II, this model must be adopted to describe the time series of the June rainfall in Shanghai. The conclusion reached from the method of Akaike is the same as that obtained here by selected significance tests of the parameters.

Using the results of Lorenz (1977), Carr (1980) derived two new estimators of the independent sample mean error which depend only upon dependent sample statistics:

$$L_1 = \frac{n-1}{(n-k-1)(n-k-2)} S_k, \tag{17}$$

$$L_2 = \frac{n(n-1)}{(n-k-1)^2(n-k-1)} S_k, \tag{18}$$

$$(k = 1, 2, \ldots, p)$$

where L_1 uses the biased residual variance, while L_2 uses the unbiased residual variance. These estimators were later applied to an AR model (Carr, 1981).

Applying Carr's criteria L_1 and L_2 to determine the order of a linear AR model for June rainfall in Shanghai (Table 1), Eqs. (17) and (18) gave

$$L_1 \begin{cases} L_{1,1} = 8030 \\ L_{1,2} = 7948 \\ L_{1,3} = 8474 \\ L_{1,4} = 9116 \end{cases}$$

$$L_2 \begin{cases} L_{2,1} = 8297 \\ L_{2,2} = 8504 \\ L_{2,3} = 9041 \\ L_{2,4} = 10501, \end{cases}$$

where the second subscripts for both L_1 and L_2 denote the order of the models. Calculations of L_1 indicate that model II should be accepted as the best forecast model, the result being the same as that given either by the FPE or by the t-test given in Eq. (9). Calculations of L_2 give a result that model I could be the best forecast model corresponding to the results of the t-test having the significance level of $\alpha = 0.05$.

For the sake of a thorough examination for the problem of the goodness of fit and effectiveness of prediction, the residual sums of squares for the four models have been calculated by means of Eq. (15) for the June rainfall records of 30 years (1921–50) in Shanghai and the error sums of squares calculated by means of (I), (II), (III) and (IV) for the records of another 10 years (1951–60), the results of which are shown in Table 2.

It is evident from Table 2 that the residual sum of squares of model IV is a minimum. The error sum of squares of model I is a minimum, a fact being consistent with Carr's conclusion (Carr, 1981) that the L_2 estimator may be able to perform better than FPE in determining the order of an AR model. However, the range of errors show of model II is a minimum, i.e., the changes of individual errors are less variation for model II than for model I. Therefore, it is of particular interest in consideration of the fact that the use of the significance level $\alpha = 0.10$ for the t-test may give a most powerful test during the process of model building for climatic prediction.

Having fitted an AR model to a stationary time series, it is often necessary to see if the residuals are random in order to verify that the fitted model may provide an adequate description of the data. After a review of all the statistical tools for residual analysis for AR processes, such as the Durbin-Watson test, the Portmanteau lack of fit test, etc., Chatfield (1980) arrived at the conclusion that he wishes to look at the first few values of the residual r_k, particularly at lag 1, to see if any are significantly different from zero; although there would not be enough evidence to reject the model, in case only one value of r_k is just significant.

For model I, the residuals have a mean -3.35, variance 85.34^2, and the first few values of r_k are:

$$r_1 = -0.1289, \quad r_2 = -0.2133, \quad r_3 = 0.2324,$$

$$r_4 = 0.5069, \quad r_5 = -0.1434.$$

For model II the residuals have a mean -7.03, variance 81.65^2, and the first few values of r_k are:

$$r_1 = -0.0160, \quad r_2 = 0.0175, \quad r_3 = 0.153,$$

$$r_4 = 0.0422, \quad r_5 = -0.0749.$$

After performing the t-test, the sequences of residuals for both model I and model II were found to be sequences of mutually independent variables (a random series with $r_k \approx 0$). However, the following facts should be pointed out: 1) the sequence of residuals

for model II approximates a white noise sequence with $\rho_k = 0$; 2) there is a large improvement in fit for model II; 3) there is a little improvement in fit either for model III or for model IV (see Table 2); 4) other approximations and estimators of the independent sample mean square error appearing in the literature (Carr, 1980) are similar to L_1 (Carr, 1981), but L_1 is in good agreement with the t-test or F-test procedure used here. Therefore, in consideration of the facts mentioned above we should be fully convinced that model II identified by means of the t-test at a significance level $\alpha = 0.10$ or the F-test at a significance level $\alpha = 0.20$ is superior for describing the data.

There is a large improvement in fit for model II; therefore, from the above method of calculating the residual sum of squares on fitting extra terms in the AR model, it is apparent that there certainly exists a systematic component at least in the sequence of residuals calculated from model I. This is a problem of the AR model with a moving average error (Kendall and Stuart, 1976b, pp. 508–509; Katz and Skaggs, 1981). It is also equivalent to fitting an autoregressive-moving average (ARMA) model to the data so that information still contained in the data can be extracted. However, it is much harder to estimate the parameters of an ARMA model than an AR model, and thus if low-order AR processes adequately fit the time series, it may be unnecessary to consider the more general ARMA processes, as was concluded by Katz and Skaggs (1981). Indeed, this is another reason why the author prefers to adopt model II rather than model I to which the moving average terms should be added.

4. Degrees of freedom in tests of hypotheses for autoregressive models

It has been noted that the linear AR model may be identified by means of the t-test or the F-test. But the tests of hypotheses are strictly applicable to an independent normal random variate, while meteorological time series generally possess pronounced temporal interrelationships. This serial correlation might decrease the effective degrees of freedom, with the result that a test of a statistical hypothesis may underestimate the significance. Nordø (1966) discussed this problem in detail with respect to the regression equation

$$Y_{(t)} = \beta_0 + \beta_1 X_1(t) + \beta_2 X_2(t) + \cdots + \beta_k X_k(t) + \epsilon(t). \quad (19)$$

It has been verified by Nordø that the estimated degrees of freedom of residual variance is approximately equal to

$$\nu = n - n^{-1} \sum_{j=0}^{k} \sum_{u,v=1}^{n} R_{u-v}\rho_{(j)u-v}, \quad (20)$$

TABLE 2. Residual sum of squares, error sum of squares and other errors.

Model		I	II	III	IV
Residual sum of squares (1921–50)		209343	192399	189925	188600
Error sum of squares (1951–60)		35749	38312	39734	41282
Errors	max	95.4	90.7	97.2	93.5
	min	−104.8	−104.1	−110.2	−112.8
	range	200.2	194.8	207.4	206.3

where

$$R_{u-v} = \frac{E[\epsilon(u)\epsilon(v)]}{\sqrt{E[\epsilon^2(u)]E[\epsilon^2(v)]}} \quad (21)$$

is the population correlation coefficient of residuals at $(u - v)$ lag, and

$$\rho_{(j)u-v} = \frac{E[x_j(u)x_j(v)]}{\sqrt{E[x_j^2(u)]E[x_j^2(v)]}} \quad (22)$$

is the population correlation coefficient of deviation values of $x_j(t)$ separated by $(u - v)$ units of time; $x_0(t)$ is constant. The second term in (20) will represent the part of degrees of freedom due to interrelationships.

If the time series is described by the simple Markov process, then (20) becomes

$$\nu = n - \sum_{j=0}^{k} \frac{1 + \rho_{(j)}R_1}{1 - \rho_{(j)}R_1} + \frac{2}{n}\sum_{j=0}^{k} \rho_{(j)}R_1 \frac{1 - \rho_{(j)}^n R_1^n}{(1 - \rho_{(j)}R_1)^2}, \quad (23)$$

where $\rho_{(j)}$ is the autocorrelation coefficient of the first order, and R_1 the residual autocorrelation coefficient of the first order.

Now, according to the Mann-Weld theorem, we can apply the above result of Nordø to the linear AR model. Since

$$\rho_{(0)} = 1 \quad \text{and} \quad \rho_{(1)} = \rho_{(k)} = \rho_1,$$

Eq. (23) becomes

$$\nu = n - \left[\frac{1 + R_1}{1 - R_1} + k\frac{1 + \rho_1 R_1}{1 - \rho_1 R_1}\right] + \frac{2}{n}\left[\frac{R_1(1 - R_1^n)}{(1 - R_1)^2} + k\rho_1 R_1 \frac{1 - \rho_1^n R_1^n}{(1 - \rho_1 R_1)^2}\right]. \quad (24)$$

$$(k = 1, 2, \cdots)$$

The degrees of freedom ν in (9) should be replaced by ν in (24), in case the linear AR model can be

identified by a t-test or F-test of the parameters. However, the question is how to estimate the population autocorrelation coefficient of residuals. The larger the value of k, the smaller is the value of R_1. If k is large enough, the residuals approach random numbers, and R_1 approaches zero.

We find from (1) that

$$\begin{aligned} E[\epsilon^2(t)] = \sigma^2[&(\alpha_0^2 + \alpha_1^2 + \alpha_2^2 + \cdots + \alpha_k^2) \\ &+ 2\rho_1(-\alpha_0\alpha_1 + \alpha_1\alpha_2 + \alpha_2\alpha_3 + \cdots + \alpha_{k-1}\alpha_k) \\ &+ 2\rho_2(-\alpha_0\alpha_2 + \alpha_1\alpha_3 + \alpha_2\alpha_4 + \cdots + \alpha_{k-2}\alpha_k) \\ &+ 2\rho_3(-\alpha_0\alpha_3 + \alpha_1\alpha_4 + \alpha_2\alpha_5 + \cdots + \alpha_{k-3}\alpha_k) \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &+ 2\rho_k(-\alpha_0\alpha_k)], \end{aligned} \tag{25}$$

where $\alpha_0 = 1$; and

$$\begin{aligned} E[\epsilon(t)\epsilon(t-1)] = \sigma^2[&(\alpha_0^2 + \alpha_1^2 + \alpha_2^2 + \cdots + \alpha_k^2)\rho_1 \\ &+ (\alpha_1\alpha_2 + \alpha_2\alpha_3 + \cdots + \alpha_{k-1}\alpha_k)(\rho_0 + \rho_2) \\ &+ \alpha_1\alpha_3(\rho_1 + \rho_3) + \alpha_1\alpha_4(\rho_2 + \rho_4) \\ &+ \cdots + \alpha_1\alpha_k(\rho_{k-2} + \rho_k) - \alpha_1(\rho_0 + \rho_2) \\ &- \alpha_2(\rho_1 + \rho_3) - \cdots - \alpha_k(\rho_{k-1} + \rho_{k+1})], \end{aligned} \tag{26}$$

where $\rho_0 = 1$.

If we let $k = 1$,

$$R_1 = \frac{(1 + \alpha_1^2)\rho_1 - \alpha_1(1 + \rho_2)}{(1 + \alpha_1^2) - 2\alpha_1\rho_1};$$

while for $k = 2$,

$$R_1 = \frac{(1 + \alpha_1^2 + \alpha_2^2)\rho_1 + \alpha_1\alpha_2(1 + \rho_2) - \alpha_1(1 + \rho_2) - \alpha_2(\rho_1 + \rho_3)}{1 + \alpha_1^2 + \alpha_2^2 + 2\rho_1(\alpha_1\alpha_2 - \alpha_1) - 2\alpha_2\rho_2}. \tag{27}$$

For the linear AR model II of June rainfall in Shanghai, the results of calculation from (27) and (24) in terms of sample values of α and ρ are given as follows:

$$R_1 = 0.032,$$

$$\nu = 28 - 3.024.$$

These calculations explain to a great extent the fact that for time series of monthly rainfall with certain temporal interrelationships the t-test or F-test may be used to choose the appropriate order of the linear AR model, because $\nu = 28 - 3.024 \approx 25$ for model II. This value of degrees of freedom shows the fact that model II has been adequately built on random variables.

Since this kind of interrelationship may effectively decrease the degrees of freedom according to (24), the degrees of freedom required by the critical t-value found in the tabulated form of the t distribution can be larger than that required by the observed t-value calculated from (9). Suppose, for example, $r_1 = 0.3114$ and $t = 1.7027$. But the 95% critical t-value for $\nu = 27$ is $t_{0.95}(27) = 1.703$. Since $t < t_{0.95}(27)$, we should conclude at the 5% level of significance that the null hypothesis $\rho_1 = 0$ must not be rejected. Thus, there is an insignificant autocorrelation in the time series. However, the observed t-value must be larger than 1.7027, because the degrees of freedom must be larger than 27, in case the condition of independent random variables is actually satisfied during the test procedure. At the same time, the critical t-value found from the table must still be $t_{0.95}(27) = 1.703$, for the t distribution is valid for independent normal random variables. In other words, the temporal interrelationships in the time series would raise the threshold of selection; as a result the forward selection by means of the t-test with decreased degrees of freedom should be right in judging significance, for the observed t-value is still higher than the increased selection threshold. At the same time, values of the coefficient of the partial autocorrelation will in general continue to decrease due to the decreased effects of interrelationships for increasing lags of the time interval during the process of forward selection. When an insignificant parameter is finally found at a certain step of the selection procedure, the condition of approximate independence required by the t-test should be satisfied for determining the order of the linear AR model, as is shown by the above calculation.

On the basis of this discussion, we should point out that it is by no means necessary, in practice, to perform the t-test or F-test procedure for the larger values of $\hat{\alpha}_k^{(k)}$, just like the stepwise autoregression suggested by Granger and Newbold in 1977 (cf. Chatfield, 1981), and the problem of whether or not we continue to calculate (3) may not be decided very simply. Thus the t-test or F-test is necessary only when $\hat{\alpha}_k^{(k)}$ is close to zero. This procedure is similar to that performed by Chatfield (1980, p. 69).

Acknowledgments. The author wishes to thank Dr. R. E. Livezey and Dr. Meg Brady Carr for a careful reading of the manuscript, for giving constructive suggestions and comments in all respects, and for sending the relevant references. The author is also grateful to Dr. Kevin E. Trenberth who was kind enough to give me many helpful suggestions for the further improvement of this paper.

REFERENCES

Akaike, H., 1969: Fitting autoregressions for prediction. *Ann. Inst. Statist. Math.*, Tokyo, **21,** 243–247.

——, 1974: A new look at the statistical model identification. *IEEE Trans. Auto. Control,* **19,** 716–723.

Billingsley, P., 1961: Statistical methods in Markov chains. *Ann. Math. Statist.*, **32,** Institute of Mathematical Statistics, 12–40.

Brooks, C. E. P., and N. Carruthers, 1953: *Handbook of Statistical Methods in Meteorology.* HMSO, London, 412 pp.

Box, G. P., and G. M. Jenkins, 1970: *Time Series Analysis, Forecasting and Control.* Holden-Day, 553 pp.

Burr, I. W., 1974: *Applied Statistical Methods.* Academic Press, 478 pp.

Carr, M. B., 1980: On determining the number of predictors in a regression equation used for prediction. Cooperative thesis No. 59, Florida State University and National Center for Atmospheric Research, 261 pp.

——, 1981: Determining the order of an autoregressive model. *Preprints, Seventh Conf. Probability and Statistics in Atmospheric Sciences,* Monterey, Amer. Meteor. Soc., 174–176.

Chatfield, C., 1980: *The Analysis of Time Series: An Introduction.* Chapman and Hall, London, 268 pp.

Cramér, H., 1946; *Mathematical Methods of Statistics.* Princeton University Press, 575 pp.

Draper, N., and H. Smith, 1967: *Applied Regression Analysis.* Wiley, 407 pp.

Durbin, J., 1960: The fitting of time series models. *Rev. Intern. Inst. Statist.*, **28,** 233–244.

Gabriel, K. R., and J. Neumann, 1962: A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quart. J. Roy. Meteor. Soc.*, **88,** 90–95.

Gates, P., and H. Tong, 1976: On Markov chain modeling to some weather data. *J. Appl. Meteor.*, **15,** 1145–1151.

Hoel, P. G., 1954: A test of Markov chains. *Biometrika,* **41,** 430–433.

Jones, R. H., 1964: Spectral analysis and linear prediction of meteorological time series. *J. Appl. Meteor.*, **3,** 45–52.

——, 1975: Estimating the variance of time averages. *J. Appl. Meteor.*, **14,** 159–163.

Katz, R. W., 1979: Estimating the order of a Markov Chain: Another look at the Tel Aviv rainfall data. *Preprints Sixth Conf. Probability and Statistics in Atmospheric Sciences,* Banff, Amer. Meteor. Soc., 217–221.

——, and R. H. Skaggs, 1981: On the use of autoregressive-moving average processes to model meteorological time series. *Mon. Wea. Rev.*, **109,** 479–484.

Kendall, M. G., and A. Stuart, 1976a: *The Advanced Theory of Statistics,* Vol. 2. Griffin, London, 748 pp.

——, and ——, 1976b: *The Advanced Theory of Statistics,* Vol. 3. Griffin, London, 585 pp.

Klein, W., B. M. Lewis and I. Enger, 1959: Objective predication of five-day mean temperature during winter. *J. Meteor.*, **16,** 672–682.

Lorenz, E. N., 1977: An experiment in nonlinear statistical weather forecasting. *Mon. Wea. Rev.*, **105,** 590–602.

Lowry, W. P., and D. Guthrie, 1968: Markov chains of order greater than one. *Mon. Wea. Rev.*, **96,** 787–801.

Mann, H. B., and A. Wald, 1943: On the statistical treatment of linear stochastic difference equations. *Econometrica,* **11,** 173–220.

Nordø, J., 1966: Significance of statistical relations derived from geophysical data. *Tellus,* **18,** 39–53.

Quenouille, M. H., 1947: A large-sample test for the goodness of fit of autoregressive schemes. *J. Roy. Statist. Soc.*, **110,** 123–129.

Shibata, R., 1980: Selection of the number of regression parameters in small sample cases. *Statistical Climatology, Proc. First Int. Conf. Statistical Climatology,* Hachioji, Japan, S. Ikeda *et al.*, Eds., Elsevier, 137–148.

Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Statist.*, **6,** 461–464.

Thom, H. C. S., and M. D. Thom, 1972: Tests of significance for temperature and precipitation normals. *Mon. Wea. Rev.*, **100,** 503–508.

Tong, H., 1975: Determination of the order of a Markov chain by Akaike's information Criterion. *J. Appl. Prob.*, **12,** 488–497.

von Mises, R., 1964: *Mathematical Theory of Probability and Statistics.* Academic Press, 694 pp.

Yao, C. S., 1958: Guarantee probabilities of rainfalls in every month in eastern China. (in Chinese) *J. Nanking Univ.*, Nanking, 40–60.

——, 1963: *Climatic Statistics.* Science Publ., Peking, 246 pp. (in Chinese).

——, 1966: Probabilities of random changes of the wet day and dry day. (in Chinese) *Acta Meteor. Sinica,* **36,** 249–260.

——, 1977a: Orthogonal regression and partial correlation screening. (in Chinese) *J. Nanking Univ.*, Nanking, 119–140.

——, 1977b: A method of establishment of multiple regression equation by means of partial correlation screening (in Chinese). *Tech. Data Meteor.*, Central Meteor. Bureau, Peking, **5,** 6–8.

Yevjevich, V., 1972: *Stochastic Processes in Hydrology.* Water Resour. Publ., Fort Collins, CO, 276 pp.

Yule, G. U., 1927: On a method investigating periodicities in disturbed series, with special reference to Wolfer's Sunspot Numbers. *Phil. Trans. Roy. Soc. London,* **A226,** 267.

Zurndorfer, E. A., and H. R. Glahn, 1977: Significance testing of regression equations developed by screening regression. *Preprints Fifth Conf. Probability and Statistics in Atmospheric Sciences.* Amer. Meteor. Soc., 95–99.