Financial Econometrics
Lecture 2: ARIMA models

Richard G. Pierse

1 Introduction

The Box and Jenkins methodology of *ARIMA* modelling (Box *et al.* 1994) is a time series method for explaining variables in terms of their own past. It is a purely statistical analysis since no attempt is made to use any *a priori* economic theory. We concentrate here on univariate analysis although, as shown in the final section, it can just as easily be used to model several time series jointly. The methodology has proved popular as a method of producing short-term forecasts, particularly in the area of finance.

2 AR(I)MA models

2.1 The Wold Representation Theorem

Any *weakly stationary* stochastic process y_t with mean μ and variance σ^2 can be written

$$y_t - \mu = \psi_0\varepsilon_t + \psi_1\varepsilon_{t-1} + \psi_2\varepsilon_{t-2} + \cdots = \sum_{j=0}^{\infty}\psi_j\varepsilon_{t-j} \tag{2.1}$$

where ε_t is a sequence of uncorrelated random variables with mean 0 and constant variance σ^2. This is called the *infinite moving average* representation of y_t, or the *Wold* representation. The moving average coefficients are subject to the condition that they are *absolutely summable*

$$\sum_{j=0}^{\infty}|\psi_j| < \infty\,.$$

Using the *lag operator* L, equation (2.1) can be rewritten as

$$y_t - \mu = (1 + \psi_1 L + \psi_2 L^2 + \cdots)\varepsilon_t = \psi(L)\varepsilon_t$$

where $\psi(L)$ is a polynomial function in the lag operator. Without loss of generality, we have imposed the normalisation restriction that $\psi_0 = 1$.

The polynomial $\psi(L)$ can be factorised as the product of its roots

$$\psi(L) = \prod_{j=1}^{\infty}(1 + \beta_j L) = (1 + \beta_1 L)(1 + \beta_2 L) \cdots \tag{2.2}$$

with roots given by

$$-\frac{1}{\beta_1}, -\frac{1}{\beta_2}, \quad \text{etc.}$$

The moving average roots must satisfy the condition of *identifiability* that

$$\|\beta_j\| \leq 1 \quad , \quad \forall j \, .$$

Note that this condition of identifiability does not rule out the possibility of unit moving average roots where $\|\beta_j\| = 1$.

2.2 Invertibility and the autoregressive representation

When there are no unit roots so that *all* the roots satisfy the stronger condition that $\|\beta_j\| < 1$, then the process is said to be *invertible* and y_t can be written in the *autoregressive* representation

$$\psi(L)^{-1} y_t = \varepsilon_t \, .$$

More generally, if *some* of the roots satisfy $\|\beta_i\| < 1$, then $\psi(L)$ can be factorised into two polynomials

$$y_t = \psi(L)\varepsilon_t = \phi(L)^{-1}\theta(L)\varepsilon_t$$

or

$$\phi(L) y_t = \theta(L)\varepsilon_t \, . \tag{2.3}$$

(2.3) is a *mixed ARMA(p,q)* model where

$$\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p$$

and

$$\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q \, .$$

Finally, consider the case where y_t is not stationary but instead is *integrated* of order d, ($y_t \sim I(d)$). Then, by definition, the *dth* order difference of y_t,

$$\Delta^d y_t$$

is stationary, and can be expressed as an *ARMA(p,q)* process. Therefore, it follows that

$$\phi(L)\Delta^d y_t = \theta(L)\varepsilon_t \tag{2.4}$$

Such a process is said to be an integrated *ARMA* process or an *ARIMA(p,d,q)* process.

2.3 Mixed Processes

2.3.1 Advantages of mixed processes

Why is it necessary to consider mixed processes? Box and Jenkins (1976) stress *parsimony*. They argue that an *ARMA(p,q)* model with small values of p and q will do as well at explaining a process y_t as a high order $AR(p^*)$ or $MA(q^*)$ process. Allowing an MA component may give evidence of *over-differencing*. Suppose that $y_t = \varepsilon_t$, then $\Delta y_t = \Delta\varepsilon_t = \varepsilon_t + \theta\varepsilon_{t-1}$, where $\theta = -1$. If we find an estimated parameter $\widehat{\theta}$ close to -1, then this is evidence for over-differencing. This could not be picked up in a *pure AR* model.

2.3.2 Problems with mixed processes

One problem with estimating mixed processes is that of *common factors*. Suppose the 'true' model is *ARMA(p,q)* but the investigator mistakenly fits *ARMA(p+1,q+1)*. If the true model is given by $\phi(L)x_t = \theta(L)\varepsilon_t$, then the estimated model can be written

$$(1 - \alpha_{p+1}L)\phi(L)x_t = (1 + \beta_{q+1}L)\theta(L)\varepsilon_t\,.$$

This model is *not identified* since $\alpha_{p+1} = -\beta_{q+1} = \gamma$, reduces the model to *ARMA(p,q)* for *any* value of the root γ. This means that a *general-to-simple modelling strategy* will not work with mixed *ARMA* models.

3 Choosing the order of the *ARIMA* model

3.1 Identifying the correct order of differencing

The first step in choosing an appropriate *ARIMA* model is to identify the correct order of differencing. This is a question of testing for unit roots and a natural test to use is the (augmented) Dickey-Fuller (*ADF*) test. The appropriate procedure would be to test down from an initial order of integration d^* that is at least as large as the (unknown) true value d. Then a sequence of Dickey-Fuller tests are computed testing the null hypothesis $\Delta^{d*}y_t \sim I(1)$ against the alternative hypothesis that $\Delta^{d*}y_t \sim I(0)$, reducing d^* each time, until the null hypothesis fails to be rejected. The final d^* then determines d.

3.2 Choosing between AR and MA representations

The theoretical *ACF* is given by

$$\rho_s = \frac{\mathrm{cov}(y_t, y_{t-s})}{\mathrm{var}(y_t)}\,.$$

In practice, we do not observe this. However, a consistent estimate is given by the sample autocorrelation function or *correlogram*

$$\widehat{\rho}_s = \frac{\sum_{t=s+1}^{T}(y_t - \overline{y})(y_{t-s} - \overline{y})}{\sum_{t=1}^{T}(y_t - \overline{y})^2} \tag{3.1}$$

where $\overline{y}$ is the sample mean of y_t. In large samples the correlogram should mirror the shape of the theoretical autocorrelation function. Similarly, we can estimate the partial autocorrelation function by estimating the coefficients $\widehat{p}_s$ from the regression

$$\widehat{y}_t = \widehat{p}_1 y_{t-1} + \widehat{p}_2 y_{t-2} + \cdots + \widehat{p}_r y_{t-r} + \widehat{\varepsilon}_t. \tag{3.2}$$

This is called the partial correlogram. Examining the correlogram and partial correlogram may help distinguish between pure AR and pure MA processes. In pure MA processes, we know that the autocorrelations should cut off after a certain point whereas in pure AR processes, the autocorrelations will never disappear completely but only die away gradually. Conversely, in pure AR models, the partial autocorrelations should cut off after a certain point whereas in pure MA models they will only die away gradually.

In practice, noise may blur these distinctions and make it difficult to decide on the correct process. In choosing between *pure AR* or *pure MA* processes a *general-to-simple* methodology can be used. However, because of the problem of common factors, the same methodogy cannot be used to choose between mixed $ARMA$ processes and over-parameterisation must be avoided. As a consequence, different researchers can often disagree about the best $ARMA$ model to fit a particular series. For example, Box and Jenkins themselves identify two different processes for some of their test series.

3.3 Using information criteria to select models

Another way of choosing between models is to compare how well they fit. The classic measure of goodness of fit is R^2 so it might be thought that choosing the model with the largest R^2 or equivalently with the smallest mean square error, would be a sensible procedure. Note that, by definition, the OLS estimator *minimises* the sum of squared residuals

$$\sum_{t=1}^{T} \widehat{\varepsilon}_t^2$$

where

$$\widehat{\varepsilon}_t \equiv y_t - \widehat{y}_t$$

are the *OLS* residuals and so *OLS* also minimises the *Mean Square Error (MSE)* defined by

$$MSE = \frac{1}{T}\sum_{t=1}^{T}\widehat{\varepsilon}_t^2. \tag{3.3}$$

However, the *MSE* is not a good criterion for selecting a model because it will *never* favour a parsimonious model with fewer parameters. The reason is that the *MSE cannot increase* when an extra coefficient is added to a regression and can only stay the same or fall. This means that, in choosing between a *AR(2)* and an *AR(4)* model the *MSE* criterion will always favour the *AR(4)*, simply because it has two extra parameters.

Box and Jenkins stress the importance of *parsimony* in selecting a model. This suggests that, when comparing competing models, we should minimise a criterion based on the *MSE* but which attaches penalties to models with more parameters. Such criteria are known as *information criteria* and two important information criteria have been proposed in the literature. Both involve adjusting the *MSE* by a multiplicative penalty related to the number of model parameters, k. In choosing between models, the model with the smallest value of the criterion is selected.

The *Akaike Information Criterion (AIC)* of Akaike (1973) is defined by

$$AIC = \exp(\frac{2k}{T})\frac{1}{T}\sum_{t=1t}^{T}\widehat{\varepsilon}_t^2 \tag{3.4}$$

and the *Schwarz Information Criterion (SIC)* (Schwarz (1978), sometimes also known as the *Bayesian Information Criterion (BIC)*), is defined by

$$SIC = T^{\frac{k}{T}}\frac{1}{T}\sum_{t=1}^{T}\widehat{\varepsilon}_t^2. \tag{3.5}$$

Both information criteria penalise a model with more parameters (larger k) but the *SIC* involves a higher penalty than the *AIC*. This means that the *SIC* is more likely to choose the more parsimonious model than the *AIC*. Which criterion is better? It can be shown that the *SIC* is *consistent* in the sense that, when the 'true' model is among the models considered, then the probability of selecting it approaches 1 as the sample size increases. On the other hand, the *AIC* is *asymptotically efficient* in the sense that, as the sample size increases, it will select a sequence of models aproaching the 'true' model at least as fast as any other criterion. Neither criterion is clearly better than the other and so, in practice, both criteria are used and they often but not always will lead to selection of the same model. Where there is a conflict, the *parsimony principle* would suggest using the stricter *SIC* criterion.

3.4 Estimation

Once the order of the *ARIMA(p,d,q)* process has been identified, the parameters can then be estimated. Pure *AR* models are dynamic regression models and can be estimated efficiently by *OLS*. A mixed *ARMA* process can be thought of as a dynamic regression model with autocorrelated disturbances:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + u_t$$

$$u_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}.$$

In this model, *OLS* estimates will be *inconsistent*. Efficient estimates are provided by maximum likelihood estimation methods. In general, it is rather more difficult to estimate models with autocorrelated disturbances when these take the moving average rather than the autoregressive form. However, most econometric packages now have facilities for the estimation of models with moving average disturbances.

4 Diagnostic checking

Having estimated an *ARIMA* model of chosen order, the residuals should be checked for evidence of autocorrelation.

4.1 Box-Pierce or Q statistic

This statistic, proposed by Box and Pierce (1970), is a *portmanteau* statistic for testing autocorrelation. The form of the statistic is given by

$$Q = T \sum_{j=1}^{n} \widehat{\rho}_j^2 \sim_a \chi_n^2 \tag{4.1}$$

where $\widehat{\rho}_j^2$ is the *squared* sample autocorrelation coefficient of jth order.

4.2 Box-Ljung (or Modified Box-Pierce) statistic

Ljung and Box (1978) suggested a correction to the Box-Pierce statistic that has better properties in small samples. The modified statistic is given by

$$Q^* = T(T+2) \sum_{j=1}^{n} (T-j)^{-1} \widehat{\rho}_j^2 \sim_a \chi_n^2 \tag{4.2}$$

4.3 Lagrange Multiplier tests

The Box-Pierce statistic can be shown to be a *Lagrange Multiplier* test of the hypothesis

$$H_0 : AR(0) \quad \text{or} \quad MA(0)$$

against the general alternative

$$H_1 : AR(n) \quad \text{or} \quad MA(n)\,.$$

It is well known that this test has little power against specific alternatives within the general class. Alternatively, we can construct LM tests against a more *specific* alternative hypothesis. These are likely to have higher power.

5 Multivariate models

The *ARIMA* methodology can easily be generalised to the multivariate context to model several series jointly. The model can be written

$$\mathbf{\Phi}(L)\Delta^d \mathbf{y}_t = \mathbf{\Theta}(L)\varepsilon_t \tag{5.1}$$

$$\varepsilon_t \sim iid(\mathbf{0}, \mathbf{\Sigma})$$

where $\mathbf{y}_t$ is an $n \times 1$ vector of observations on n variables at time period t and $\mathbf{\Phi}(L)$ and $\mathbf{\Theta}(L)$ are $n \times n$ polynomial matrices in the lag operator defined by

$$\mathbf{\Phi}(L) = \mathbf{\Phi}_0 - \mathbf{\Phi}_1 L - \mathbf{\Phi}_2 L^2 - \cdots - \mathbf{\Phi}_p L^p$$

and

$$\mathbf{\Theta}(L) = \mathbf{\Theta}_0 + \mathbf{\Theta}_1 L + \mathbf{\Theta}_2 L^2 + \cdots + \mathbf{\Theta}_q L^q.$$

This is called the *VARIMA(p,d,q)* model. The matrices $\mathbf{\Phi}_j$, $j = 0, \cdots, p$, and $\mathbf{\Theta}_j$, $j = 0, \cdots, q$, are all $n \times n$ matrices of coefficients with normalisation restrictions $\mathbf{\Phi}_0 = \mathbf{I}$ and $\mathbf{\Theta}_0 = \mathbf{I}$. There are pn^2 elements in the parameter matrices $\mathbf{\Phi}$ and qn^2 elements in the parameter matrices $\mathbf{\Theta}$. These parameters can be estimated using maximum likelihood (*ML*) techniques.

An important special case of this model is the *VARIMA(p,0,0)* or *Var(p)* model

$$\mathbf{\Phi}(L)\mathbf{y}_t = \varepsilon_t.$$

This is a form of *SURE* model and the pn^2 elements of the parameter matrices $\mathbf{\Phi}$ in this model can be efficiently estimated by *OLS*.

References

[1] Akaike, H. (1973), 'Information theory and an extension of the maximum likelihood principle' in B. Petrov and F. Csake *eds. Second International Symposium on Information Theory*, Akademiai Kiado, Budapest.

[2] Box, G.E.P, G.M. Jenkins and G.C. Reinsel (1994), *Time Series Analysis: Forecasting and Control*, (3rd ed.), Prentice Hall, Englewood Cliffs, NJ.

[3] Box, G.E.P. and D.A. Pierce, (1970), 'Distribution of residual autocorrelations in autoregressive integrated moving average time series models', *Journal of the American Statistical Association*, 65, 1509–1526.

[4] Ljung, G.M. and G.E.P. Box (1978), 'On a measure of lack of fit in time series models', *Biometrika*, 66, 67–72.

[5] Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics*, 6(2), 461–464.