9. Estimation and selection of ARIMA models

Outline:

- Introduction
- Likelihood of ARMA processes
  - ◇ AR processes
  - ◇ MA and ARMA processes
- Kalman filter
- Properties of the estimators
- Model selection criteria

Recommended readings:

▷ Chapter 10 of D. Peña (2008).

▷ Chapters 5 and 8 of P.J. Brockwell and R.A. Davis (1996).

▷ Chapters 5 and 13 of J.D. Hamilton (1994).

Introduction

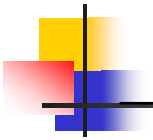
▷ This section is divided into two parts:

- First, we will study the estimation of parameters of an ARMA model.
- Second, we wil study the selection of an ARMA model from among various estimated models.

▷ We will assume that we have a stationary series, $\{\omega_t\}$, and we wish to estimate the parameters of a specified ARMA model.

The notation ω_t is used because frequently the series is a transformation of the original series, z_t. For example, with monthly economic data we often have $\omega_t = \nabla\nabla_{12}\ln z_t$.

▷ The study of the estimation starts with the simplest case: the conditional estimation of AR processes, which is similar to least squares estimation in a regression model.



Introduction

▷ Next we will look at the exact estimation of AR processes, which leads to a problem of non-linear estimation in the parameters, thus requiring the use of optimization algorithms for non-linear problems.

▷ Later, we will study the estimation of MA and ARMA models, which is always non-linear and requires two stages:

(i) To calculate the value of the likelihood function given the value of the parameters.

(ii) to find a new value of the parameters that increases the value of the function.

▷ We see how to evaluate the likelihood function by means of an efficient algorithm, the Kalman filter.

▷ The estimation consists in iterating between these two phases until the maximum of the function is obtained.

Introduction

▷ The second part of the section looks at the case in which we have several estimated ARMA models for a series and are faced with the problem of deciding between them and selecting the most suitable.

▷ The main ideas of model selection are important and are be widely used in the rest of the course:

- Adjustment criteria are not useful for model selection, because if we increase the number of parameters the fit of the model will increase.
- Then, we must turn to criteria that balance the adjustment with the number of estimated parameters.

▷ Here, we will study the Akaike information criterion and the Bayesian information criterion (also known as Schwarz IC).

Likelihood of ARMA processes

▷ Let us assume that we have an ARMA process and wish to estimate the parameters by maximum likelihood.

▷ To do this, we must write the joint density function and maximize it with respect to the parameters, considering the data as fixed.

▷ To write the joint density of the T observations $\omega_T = (\omega_1, ..., \omega_T)$, we are going to use the following relation:

$$f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) f(\mathbf{y}|\mathbf{x}). \tag{171}$$

▷ This expression remains true if all the density functions are conditional on another variable $\mathbf{z}$, such that:

$$f(\mathbf{x}, \mathbf{y}|\mathbf{z}) = f(\mathbf{x}|\mathbf{z}) f(\mathbf{y}|\mathbf{x}, \mathbf{z}). \tag{172}$$

Likelihood of ARMA processes

▷ We consider the joint density function of the T observations ω_T. Taking $\mathbf{x} = \omega_1$ and $\mathbf{y} = \omega_2, ..., \omega_T$ in (171) we can write

$$f(\omega_T) = f(\omega_1) f(\omega_2, ..., \omega_T | \omega_1)$$

and decomposing the second term, with (172) making $\mathbf{z} = \omega_1$, $\mathbf{x} = \omega_2$ and $\mathbf{y} = \omega_3, ..., \omega_T$, results in

$$f(\omega_T) = f(\omega_1) f(\omega_2|\omega_1) f(\omega_3, ..., \omega_T | \omega_1, \omega_2)$$

and repeating this process, we finally obtain

$$f(\omega_T) = f(\omega_1) f(\omega_2|\omega_1) f(\omega_3|\omega_2, \omega_1) ... f(\omega_T|\omega_{T-1}, ..., \omega_1). \quad (173)$$

▷ This expression allows us to write the joint density function of the T variables as a product of T univariate distributions.

Likelihood of ARMA processes

▷ The difference between this representation and that which is obtained using independent data points is that instead of having the product of the marginal distributions of each data point we have the marginal distribution of the first and the product of the conditionals of each data point, given the previous one.

▷ The decomposition (173) lets us write the likelihood of an ARMA model, since if we assume normality, all the conditional distributions will be normal. Its expectation is the one step ahead prediction which minimizes the quadratic prediction error, and we write:

$$E\left(\omega_t|\omega_{t-1},...,\omega_1\right) = \widehat{\omega}_{t-1}(1) = \omega_{t|t-1}.$$

▷ We let e_t denote the prediction error of ω_t made using the information from $\omega_{t-1},...,\omega_1$ and knowing the parameters of the process. Thus:

$$e_t = e_{t-1}(1) = \omega_t - \omega_{t|t-1}.$$

Likelihood of ARMA processes

▷ These prediction errors are highly related to the innovations of the process. Nevertheless, they are not identical due to a problem of initial values.

▷ To illustrate the difference, let us assume a series of zero mean and size T, generated by an AR(1) process with known parameter ϕ. Thus, since $\omega_{t|t-1} = \phi\omega_{t-1}$ for $t = 2, ..., T$ but $E(\omega_1) = \phi E(\omega_0) = 0$, we can calculate the one step ahead prediction errors by means of:

$$\begin{aligned} a_1 &= \omega_1 - \phi E(\omega_0) = \omega_1 \\ a_2 &= \omega_2 - \phi\omega_1 \\ \vdots \quad &\vdots \qquad\quad \vdots \\ a_T &= \omega_T - \phi\omega_{T-1}. \end{aligned}$$

▷ We see that the prediction errors, $e_2, ..., e_T$, coincide with the innovations of the model, $a_2, ..., a_T$, where $\omega_t = \phi\omega_{t-1} + a_t$, for $t = 2, ..., T$.

Likelihood of ARMA processes

▷ The difference appears in the first one step ahead prediction error, e_1, which is not equal to the innovation in this point, $a_1 = \omega_1 - \phi\omega_0$.

▷ This makes it so that for $t = 2, ..., T$ the variance of the one step ahead prediction errors is σ^2, that of the innovation, whereas for $t = 1$ it is different.

▷ In general, we can write:

$$Var\left(\omega_t|\omega_{t-1}, ..., \omega_1\right) = \sigma^2 v_{t|t-1}$$

where for an AR(1):

$$\begin{aligned} v_{t|t-1} &= 1 && \text{for } t = 2, ..., T \\ &= (1-\phi^2)^{-1} && \text{for } t = 1. \end{aligned}$$

Likelihood of ARMA processes

▷ With this notation, the joint density function of the sample for a general ARMA process can be written as:

$$f\left(\omega_T\right) = \prod_{t=1}^{T} \sigma^{-1} v_{t|t-1}^{-1/2} \left(2\pi\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{t=1}^{T} \frac{\left(\omega_t - \omega_{t|t-1}\right)^2}{v_{t|t-1}}\right\}$$

and taking logarithms letting $\boldsymbol{\beta} = (\mu, \phi_1, ..., \theta_q, \sigma^2)$ be the parameter vector, the support function is:

$$L(\boldsymbol{\beta}) = -\frac{T}{2} \ln \sigma^2 - \frac{1}{2} \sum_{t=1}^{T} \ln v_{t|t-1} - \frac{1}{2\sigma^2} \sum_{t=1}^{T} \frac{e_t^2}{v_{t|t-1}} \tag{174}$$

where both the conditional variances $v_{t|t-1}$ as well as the one step ahead prediction errors, e_t, depend on the parameters.

▷ Therefore, evaluating the likelihood function is reduced to the problem of calculating the one step ahead prediction errors and their variances.

▷ The maximization of the exact likelihood function is carried out using a nonlinear optimization algorithm.

Likelihood of ARMA processes

AR processes

▷ First, let us consider the AR(1) process of mean μ, $\omega_t = c + \phi\omega_{t-1} + a_t$, where $c = \mu(1-\phi)$. In this case, we have:

$$E(\omega_1) = \mu \tag{175}$$

and

$$Var(\omega_1) = E(\omega_1 - \mu)^2 = \frac{\sigma^2}{1-\phi^2}. \tag{176}$$

▷ We assume that, using the above notation, $\omega_{1|0} = \mu$ and $v_{1|0} = (1-\phi^2)^{-1}$. For ω_2 we have that, by conditioning ω_1, the moments of the conditional distribution are:

$$\omega_{2|1} = E(\omega_2|\omega_1) = c + \phi\omega_1$$

and

$$Var(\omega_2|\omega_1) = E\left[(\omega_2 - c - \phi\omega_1)^2\right] = E(a_2^2) = \sigma^2$$

resulting in $v_{2|1} = 1$.

Likelihood of AR(1) processes

▷ In the same way, we check that

$$\omega_{t|t-1} = E(\omega_t|\omega_{t-1}) = c + \phi\omega_{t-1}, \qquad t = 2, ..., T$$

and

$$Var(\omega_t|\omega_{t-1}) = \sigma^2 v_{t|t-1} = \sigma^2, \qquad t = 2, ..., T.$$

▷ As a result, the **likelihood function** is:

$$f(\omega_T) = f(\omega_1) \prod_{t=2}^{T} \sigma^{-1} (2\pi)^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{t=2}^{T} (\omega_t - c - \phi\omega_{t-1})^2 \right\}. \quad (177)$$

Likelihood of AR(1) processes

▷ Taking logarithms and using $f(\omega_1)$ as normal, with parameters given by (176) and (175), gives us the **support function**:

$$L\left(\phi,\sigma^2|\omega_T\right) = \frac{-T}{2}\ln\sigma^2 + \frac{1}{2}\ln\left(1-\phi^2\right) - \frac{\left(1-\phi^2\right)\left(\omega_1-\mu\right)^2}{2\sigma^2} \quad (178)$$
$$-\frac{1}{2\sigma^2}\sum_{t=2}^{T}\left(\omega_t - c - \phi\omega_{t-1}\right)^2.$$

▷ To obtain the estimator of ϕ we have to take the derivative of the parameter and set the result to zero.

▷ A cubic equation is obtained which has three roots, and the one that maximizes the likelihood function is the ML estimator.

Likelihood of AR(1) processes

▷ The expression (178) shows that, if we don't consider the first term, the support function has the usual expression of a linear model. If we condition the first observation we have:

$$f\left(\omega_2,...,\omega_T|\omega_1\right)=\prod_{t=2}^{T}\sigma^{-1}\left(2\pi\right)^{-1/2}\times\exp\left\{-\frac{1}{2\sigma^2}\sum_{t=2}^{T}\left(\omega_t-c-\phi\omega_{t-1}\right)^2\right\}.$$

▷ We define the **conditional likelihood** as the one associated with this joint density function:

$$L_C\left(\phi,\sigma^2|\omega_1\right)=\frac{-(T-1)}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{t=2}^{T}\left(\omega_t-c-\phi\omega_{t-1}\right)^2$$

and the estimator of the parameter ϕ that maximizes this conditional likelihood is obtained by minimizing the sum of squares

$$\sum_{t=2}^{T}\left(\omega_t-c-\phi\omega_{t-1}\right)^2=\sum_{t=2}^{T}\left(\widetilde{\omega}_t-\phi\widetilde{\omega}_{t-1}\right)^2$$

where $\widetilde{\omega}_t=\omega_t-\mu$.

Likelihood of AR(1) processes

▷ Taking the derivative and setting it to zero, given the estimator $\overline{\omega} = \widehat{\mu} = \sum_{t=1}^{T} \omega_t / T$ for the mean, the **conditional ML estimator of** ϕ is:

$$\widehat{\phi} = \frac{\sum_{t=2}^{T} (\omega_t - \overline{\omega})(\omega_{t-1} - \overline{\omega})}{\sum_{t=2}^{T} (\omega_{t-1} - \overline{\omega})^2},$$

which is that of the slope in a regression model of ω_t with respect to ω_{t-1}.

▷ The **conditional ML estimator** of the variance is:

$$\widehat{\sigma}^2 = \frac{\sum_{t=2}^{T} (\omega_t - \widehat{c} - \widehat{\phi}\omega_{t-1})^2}{T-1},$$

where $\widehat{c} = \overline{\omega}(1 - \widehat{\phi})$.

▷ In conclusion, if we condition the first term and write the likelihood of the observations from 2 to T, we have a linear model in the parameters.

▷ The difference between the estimator obtained with the conditional likelihood and the exact likelihood will generally be small, and negligible in large samples.

Likelihood of ARMA processes

AR processes

▷ Let us consider a general AR(p) process. The conditional expectation of ω_t, for $t = p+1, ..., T$ given the previous data points, $\omega_{t-1}, ..., \omega_1$ is:

$$E\left[\omega_t | \omega_{t-1}, ..., \omega_1\right] = \mu + \phi_1(\omega_{t-1} - \mu) + ... + \phi_p(\omega_{t-p} - \mu)$$

and its conditional variance is:

$$Var\left(\omega_t | \omega_{t-1}, ..., \omega_1\right) = Var(a_t) = \sigma^2.$$

▷ Hence, all the conditional distributions for $t = p+1, ..., T$ are normal, with a mean equal to the one step ahead prediction and variance σ^2.

▷ The conditional support function is obtained from the joint density of the observations $(\omega_{p+1,}, ..., \omega_T)$ conditional on the first p:

$$L_C\left(\mu, \phi, \sigma^2 | \omega_1, ..., \omega_p\right) = -\frac{(T-p)}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=p+1}^{T} \left(\omega_t - \mu - \sum_{i=1}^{p} \phi_i(\omega_{t-i} - \mu) \right)^2. \quad (179)$$

Likelihood of AR(p) processes

▷ Maximizing this function with respect to μ and ϕ is equivalent to minimizing the sum of squares of the one step ahead prediction errors and we can write:

$$S = \sum_{t=p+1}^{T} a_t^2 = \sum_{t=p+1}^{T} \left(\omega_t - \mu - \sum_{i=1}^{p} \phi_i (\omega_{t-i} - \mu) \right)^2 \tag{180}$$

where $a_t = (\omega_t - \mu - \sum_{i=1}^{p} \phi_i (\omega_{t-i} - \mu))$.

▷ Therefore, maximizing the conditional support is equivalent to least squares.

▷ The estimator of μ is obtained:

$$\sum_{t=p+1}^{T} \left(\omega_t - \mu - \sum_{i=1}^{p} \phi_i (\omega_{t-i} - \mu) \right) = 0$$

and assuming that $\sum_{t=p+1}^{T} \omega_t \approx \sum_{t=p+1}^{T} \omega_{t-i}$, which is approximately true if T is large, we find that the estimator of the mean is the sample mean of the observations considered: $\widehat{\mu} = (T-p)^{-1} \sum_{t=p+1}^{T} \omega_t$.

Likelihood of AR(p) processes

▷ A better estimator of μ is $\overline{\omega} = \sum_{t=1}^{T} \omega_t / T$, the sample mean of all the observations. Both estimators are unbiased, but the one calculated using the whole sample has less variance and is the one we will use.

▷ This is equivalent to initially estimating the mean with all the data points and then writing the likelihood for the variables in deviations to the mean.

▷ To obtain the estimator of ϕ, replacing μ with $\overline{\omega}$ in (180) and letting $\mathbf{x}_t' = (\omega_{t-1} - \overline{\omega}, ..., \omega_{t-p} - \overline{\omega})$, we get the usual least squares estimator in regression models:

$$\widehat{\phi} = \left(\sum_{t=p+1}^{T} \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \left(\sum_{t=p+1}^{T} \mathbf{x}_t (\omega_t - \overline{\omega}) \right). \tag{181}$$

Likelihood of AR(p) processes

▷ For large samples this expression is approximately,

$$\widehat{\phi} = \widehat{\Gamma}_p^{-1}\widehat{\gamma}_p, \tag{182}$$

where

$$\widehat{\Gamma}_p = \begin{vmatrix} \widehat{\gamma}_0 & \cdots & \widehat{\gamma}_{p-1} \\ \vdots & \ddots & \vdots \\ \widehat{\gamma}_{p-1} & \cdots & \widehat{\gamma}_0 \end{vmatrix}, \qquad \widehat{\gamma}_p = \begin{vmatrix} \widehat{\gamma}_1 \\ \vdots \\ \widehat{\gamma}_p \end{vmatrix},$$

which are the **Yule-Walker equations**.

▷ Nevertheless, in small samples both estimators are different and this difference is greater when the order of the process is higher.

▷ It can be proved that the least squares estimators are more accurate than those of Yule-Walker.

Likelihood of ARMA processes

MA and ARMA processes

▷ The estimation of models with MA and mixed components is more complicated than that of AR for two reasons.

- First, the likelihood function, both the conditional as well as the exact, is always non-linear in the parameters.
- Second, the procedure for conditioning certain initial values, which leads to simple results in the ARs, is more complicated for MA and ARMA processes, making the calculation of the expectations and conditional variances more difficult

▷ To illustrate these problems, we take the case of an MA(1):

$$\omega_t = a_t - \theta a_{t-1}$$

with zero marginal expectation.

Likelihood of MA(1) processes

▷ The expectation of ω_t conditional on its previous values is no longer straightforward, as in the AR, and to obtain it we must express ω_t as a function of the previous values.

▷ Starting with $t = 2$, since $\omega_2 = a_2 - \theta a_1$, and $a_1 = \omega_1 + \theta a_0$, we have that:

$$\omega_2 = -\theta\omega_1 + a_2 - \theta^2 a_0$$

and taking expectations in this expression and assuming $E(a_0|\omega_1) = 0$, we deduce that the expectation of the conditional distribution is:

$$E(\omega_2|\omega_1) = -\theta\omega_1,$$

and the variance

$$var(\omega_2|\omega_1) = \sigma^2(1 + \theta^4).$$

Likelihood of AR(p) processes

▷ Following this form for $t = 3, 4, ..,$ we obtain

$$\omega_t = -\theta\omega_{t-1} - \theta^2\omega_{t-2} - ... - \theta^{t-1}\omega_1 + a_t - \theta^t a_0$$

which leads to

$$E(\omega_t|\omega_{t-1}, ..., \omega_1) = -\theta\omega_{t-1} - \theta^2\omega_{t-2} - ... - \theta^{t-1}\omega_1$$

and

$$var(\omega_t|\omega_{t-1}, ..., \omega_1) = \sigma^2(1 + \theta^{2t}).$$

▷ These expressions are non-linear in the parameters and they are difficult to obtain in MA(q) processes.

Likelihood of MA(1) processes

▷ An alternative approach is to condition the first unobserved innovations as well. We observe that for each value of the parameters θ, the expression:

$$a_t = \omega_t + \theta a_{t-1} \tag{183}$$

permits recursive calculations of the disturbances a_t, conditional on an initial value a_0.

▷ Taking $a_0 = 0$ we can calculate all the remaining disturbances starting from ω_t. Thus:

$$E(\omega_t|\omega_{t-1}, ..., \omega_1, a_0) = -\theta a_{t-1}$$

and

$$var(\omega_t|\omega_{t-1}, ..., \omega_1, a_0) = E\left[(\omega_t + \theta a_{t-1})^2\right] = E\left[{a_t}^2\right] = \sigma^2$$

which leads to the conditional support:

$$L_C\left(\theta|\omega_1, a_0\right) = \frac{-\left(T-1\right)}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=2}^{T} a_t^2.$$

▷ The maximization of this function is carried out by means of a non-linear algorithm.

Likelihood of ARMA(p,q) processes

▷ The conditional estimation of ARMA(p,q) models is carried out following the same principles. Letting $r = \max(p, q)$ and $\boldsymbol{\beta} = (\mu, \phi_1, ..., \theta_q, \sigma^2)$ be the parameter vector, the conditional support function is:

$$L_C(\boldsymbol{\beta}|\mathbf{a}_0, \omega_p) = \frac{-(T-r)}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{T=r+1}^{T} a_t^2 \qquad (184)$$

where we have $a_t^2 = a_t^2(\beta|\mathbf{a}_0, \omega_p)$, in order to stress that the innovations are calculated from vectors $\mathbf{a}_0$ and ω_p from initial values.

▷ These estimated or residual innovations are calculated recursively by means of:

$$\widehat{a}_t = \omega_t - c - \phi_1\omega_{t-1} - ... - \phi_p\omega_{t-p} + \theta_1\widehat{a}_{t-1} + ... + \theta_q\widehat{a}_{t-q} \quad t = r+1, ..., T \quad (185)$$

where $c = \mu(1 - \phi_1 - ... - \phi_p)$ and it is assumed that the first r residuals are zero.

▷ The maximization of (184) requires an initial value of the parameters that can be obtained using the Hannan-Rissanen algorithm.

Hannan-Rissanen algorithm

▷ This algorithm provides initial estimators for an ARMA(p, q) process and it has two steps.

① We obtain an initial estimation of residuals of the model by adjusting a long AR of order $k > p + q$. Let $\widehat{\pi}_i$ be the coefficients estimated using (181). The residuals are calculated by means of

$$\widehat{a}_t = \omega_t - \widehat{c} - \sum_{i=1}^{k} \widehat{\pi}_i \omega_{t-i}$$

② With the estimated residuals from step 1, we estimate the regression

$$\omega_t = c + \phi_1 \omega_{t-1} + ... + \phi_p \omega_{t-p} - \theta_1 \widehat{a}_{t-1} - ... - \theta_q \widehat{a}_{t-q} + u_t. \qquad (186)$$

The estimation of this regression provides the initial estimators.

▷ This algorithm can be used to obtain estimators of ARMA models by iterating the above steps, which only require regressions. Indeed, with the parameters estimated in step 2 we can calculate new residuals and repeat the estimation of (186) until convergence is reached.

▷ In these conditions we obtain estimators close to the ML.

The Kalman filter

▷ The Kalman filter is a recursive procedure which is very fast computationally and has many applications in time series. In particular, it lets us quickly evaluate the likelihood of any ARMA model by calculating the one step ahead prediction errors and their variances.

▷ Let us assume that we observe a system that can be represented by means of an observation equation:

$$\mathbf{z}_t = \mathbf{H}_t \alpha_t + \epsilon_t \tag{187}$$

where $\mathbf{z}_t$ is a $k \times 1$ vector of observations, $\mathbf{H}_t$ is a known $k \times p$ matrix, α_t is an unobserved $p \times 1$ state vector and ϵ_t is a WN with distribution $\mathcal{N}(\mathbf{0}, \mathbf{V}_t)$.

▷ Moreover, the description of the system includes an equation that describes the dynamic evolution of the state variables, α_t, called the state equation:

$$\alpha_t = \mathbf{\Omega}_t \alpha_{t-1} + \mathbf{u}_t \tag{188}$$

where $\mathbf{\Omega}_t$ is known $p \times p$ matrix and $\mathbf{u}_t$ another WN, independent of ϵ_t, with distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{R}_t)$.

The Kalman filter

▷ The representation of a system by means of equations (187) and (188) is not unique. There is always the possibility of increasing the dimension of the state vector by putting zeros in the matrices which multiply it and we say that the state vector has a minimum dimension when it is not possible to represent the system with fewer than p state variables.

▷ Once the dimension of the state vector is fixed it is not unique either. Given a state vector α_t the system can be represented equally using the state vector $\alpha_t^* = \mathbf{A}\alpha_t$, and we can write the observation equation as

$$\mathbf{z}_t = \mathbf{H}_t\mathbf{A}^{-1}\mathbf{A}\alpha_t + \epsilon_t = \mathbf{H}_t^*\alpha_t^* + \epsilon_t$$

and that of the evolution of the state as

$$\alpha_t^* = \mathbf{\Omega}_t^*\alpha_{t-1}^* + \mathbf{u}_t$$

where now $\mathbf{\Omega}_t^* = \mathbf{A}\mathbf{\Omega}_t\mathbf{A}^{-1}$.

▷ From here on we will assume that the system has a minimum dimension.

The Kalman filter

▷ Any ARMA(p, q) model can be written in this formulation as follows. We define $m = \max(p, q + 1)$ and let $\alpha_t = (\alpha_{1,t}, \alpha_{2,t}, ..., \alpha_{m,t})'$ denote the state vector variables, which follow the state equation:

$$\begin{bmatrix} \alpha_{1,t} \\ \alpha_{2,t} \\ \vdots \\ \alpha_{m,t} \end{bmatrix} = \begin{bmatrix} \phi_1 & 1 & ... & 0 \\ \phi_2 & 0 & \ddots & 0 \\ \vdots & \vdots & ... & 1 \\ \phi_m & 0 & ... & 0 \end{bmatrix} \begin{bmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \\ \vdots \\ \alpha_{m,t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ -\theta_1 \\ \vdots \\ -\theta_m \end{bmatrix} a_t. \qquad (189)$$

▷ We observe that in this equation the state matrix, the form of $\mathbf{\Omega}_t$ is:

$$\mathbf{\Omega}_t = \begin{bmatrix} \phi_{m-1} & \mathbf{I} \\ \phi_m & \mathbf{0}' \end{bmatrix}$$

where ϕ_{m-1} is an $m - 1$ column vector, $\mathbf{I}$ is the identity matrix and $\mathbf{0}'$ is a vector of zeros.

The Kalman filter

▷ The vector of innovations in this equation is

$$\mathbf{u}_t = \theta a_t$$

where $\theta' = (1, -\theta_1, ..., -\theta_m)$. The covariance matrix of $\mathbf{u}$ is

$$\mathbf{R}_t = \theta\theta'\sigma^2.$$

▷ We are going to check whether by substituting successively in the state variables we obtain the representation of the ARMA process.

▷ The first equation is

$$\alpha_{1,t} = \phi_1\alpha_{1,t-1} + \alpha_{2,t-1} + a_t \tag{190}$$

and the second

$$\alpha_{2,t} = \phi_2\alpha_{1,t-1} + \alpha_{3,t-1} - \theta_1 a_t \tag{191}$$

▷ Substituting $\alpha_{2,t-1}$ in (190) according to the expression (191), gives us

$$\alpha_{1,t} = \phi_1\alpha_{1,t-1} + \phi_2\alpha_{1,t-2} + \alpha_{3,t-2} + a_t - \theta_1 a_{t-1} \tag{192}$$

The Kalman filter

▷ The third equation is

$$\alpha_{3,t} = \phi_3 \alpha_{1,t-1} + \alpha_{4,t-1} - \theta_2 a_t$$

and replacing $\alpha_{3,t-2}$ now in (192) with its above expression, we begin recovering the ARMA process in the variable $\alpha_{1,t}$.

▷ The observation equation simply serves to make the observed variable, z_t, which is scalar, equal to the first component of the state vector:

$$z_t = (1, 0, ...0)\alpha_t. \tag{193}$$

▷ Equations (189) and (193) are a representation of the ARMA model in the state space. We observe that they are a particular case of (187) and (188):

- In the observation equation (187) the data vector is now scalar, the state is a $m = \max(p, q+1) \times 1$ vector, the matrix $\mathbf{H}_t = (1, 0, ..., 0)$ and there is no measurement error or noise in the observation matrix.

- In the state equation matrix $\mathbf{\Omega}_t$ is invariant in time, and the covariance matrix of $\mathbf{u}_t$ is singular of rank one.

The Kalman filter

▷ The Kalman filter is a recursive algorithm for obtaining predictions of future observations and quickly provides the one step ahead prediction errors and their variances.

▷We are going to show the general formulation of the algorithm and then indicate its particularization in order to calculate the likelihood function of an ARMA process.

▷The algorithm operates in three steps:

- First, we predict the future state from information about the current state.
- Second, we predict new observations.
- In the third step, which is carried out when a new observation enters the system, the state estimation is revised at that moment in light of the new information.

The Kalman filter - First step

▷ The first step is the prediction of the future state starting from an estimation of the current state.

▷ Let us assume that we have the data points $Z_{t-1} = \{z_1, ..., z_{t-1}\}$ and an estimator of the state vector, $\widehat{\alpha}_{t-1}$, and we wish to predict $\widehat{\alpha}_{t|t-1}$, the future state estimation using the observed data points, Z_{t-1}.

▷ This estimation is calculated taking expectations in (188) conditional on Z_{t-1} and we have:

$$\widehat{\alpha}_{t|t-1} = \mathbf{\Omega}_t \widehat{\alpha}_{t-1} \tag{194}$$

where we have used the notation $\widehat{\alpha}_{t-1|t-1} = \widehat{\alpha}_{t-1}$.

▷ We let $\mathbf{S}_{t|t-1}$ denote the covariance matrix of this estimation:

$$\mathbf{S}_{t|t-1} = E\left[(\alpha_t - \widehat{\alpha}_{t|t-1})(\alpha_t - \widehat{\alpha}_{t|t-1})'|Z_{t-1}\right]$$

The Kalman filter - First step

▷ To calculate $\mathbf{S}_{t|t-1}$ we subtract equation (194) from (188) such that:

$$\alpha_t - \widehat{\alpha}_{t|t-1} = \mathbf{\Omega}_t \left(\alpha_{t-1} - \widehat{\alpha}_{t-1}\right) + \mathbf{u}_t,$$

and plugging this expression into the definition of $\mathbf{S}_{t|t-1}$ and letting $\mathbf{S}_{t-1} = \mathbf{S}_{t-1|t-1}$, results in

$$\mathbf{S}_{t|t-1} = \mathbf{\Omega}_t \mathbf{S}_{t-1} \mathbf{\Omega}_t' + \mathbf{R}_t. \tag{195}$$

▷ This equation has a clear intuitive interpretation: the uncertainty when predicting a new state with information up to $t-1$ is the sum of the uncertainty that we had with respect to the previous state using this information, measured by $\mathbf{S}_{t-1}$, and the uncertainty of the noise in the state equation, $\mathbf{R}_t$.

▷ For example, in an AR(1) the state vector is scalar, and $\Omega_t = \phi < 1$. The variance of the estimation follows the process

$$s_{t|t-1} = \phi^2 s_{t-1} + \sigma^2$$

and only a part of the uncertainty at $t-1$ is transferred to time t.

The Kalman filter - Second step

▷ The second step is the prediction of the new observation $\mathbf{z}_t$ given information up to $t-1$:

$$\widehat{\mathbf{z}}_{t|t-1} = E(\mathbf{z}_t|Z_{t-1}) = \mathbf{H}_t\widehat{\alpha}_{t|t-1}. \tag{196}$$

▷ This prediction will have an uncertainty that is measured by the covariance matrix of the prediction errors:

$$\mathbf{e}_t = \mathbf{z}_t - \widehat{\mathbf{z}}_{t|t-1}$$

defined by:

$$\mathbf{P}_{t|t-1} = E\left[\mathbf{e}_t\mathbf{e}_t'\right].$$

▷ To calculate this matrix, subtracting prediction (196) from the observation equation (187), we have:

$$\mathbf{e}_t = \mathbf{z}_t - \widehat{\mathbf{z}}_{t|t-1} = \mathbf{H}_t(\alpha_t - \widehat{\alpha}_{t|t-1}) + \epsilon_t \tag{197}$$

The Kalman filter - Second step

▷ Plugging expression (197) into the definition of $\mathbf{P}_{t|t-1}$, we obtain

$$\mathbf{P}_{t|t-1} = \mathbf{H}_t \mathbf{S}_{\mathbf{t}|\mathbf{t}-\mathbf{1}} \mathbf{H}_t' + \mathbf{V}_t. \tag{198}$$

▷ This equation indicates that the uncertainty of the prediction accumulates the uncertainty in the state and that of the measurement error of the observation equation.

▷ The prediction error that comes from the state estimation is modulated depending on matrix $\mathbf{H}_t$.

▷ If this is the identity matrix, which means that the observations $\mathbf{z}_t$ are measurements of the state variables plus a random error, a measurement error of the observations is added to the error of the state variables.

The Kalman filter - Third step

▷ The third step is to revise the state estimation in light of the new information.

▷ Let us assume that $\mathbf{z}_t$ has been observed thus the information available becomes $Z_t = (Z_{t-1}, \mathbf{z}_t)$.

▷ The new state estimation, $\widehat{\alpha}_t = \widehat{\alpha}_{t|t} = E(\alpha_t|Z_t)$, is calculated by regression using:

$$E(\alpha_t|Z_{t-1}, \mathbf{z}_t) = E(\alpha_t|Z_{t-1}) + \\ + cov(\alpha_t, \mathbf{z}_t|Z_{t-1}) var(\mathbf{z}_t|Z_{t-1})^{-1}(\mathbf{z}_t - E(\mathbf{z}_t|Z_{t-1})). \quad (199)$$

▷ In this equation the expectations $E(\alpha_t|Z_{t-1}) = \widehat{\alpha}_{t|t-1}$ and $E(\mathbf{z}_t|Z_{t-1}) = \widehat{\mathbf{z}}_{t|t-1}$ and the matrix $var(\mathbf{z}_t|Z_{t-1}) = \mathbf{P}_{t|t-1}$ are known.

The Kalman filter - Third step

▷ All that remains to be calculated is the covariance between the state and the new observation, which is given by

$$cov(\alpha_t, \mathbf{z}_t|Z_{t-1}) = E\left[(\alpha_t - \widehat{\alpha}_{t|t-1})(\mathbf{z}_t - \widehat{\mathbf{z}}_{t|t-1})'\right] = E\left[(\alpha_t - \widehat{\alpha}_{t|t-1})\mathbf{e}_t'\right]$$

and substituting (197),

$$cov(\alpha_t, \mathbf{z}_t|Z_{t-1}) = E\left[(\alpha_t - \widehat{\alpha}_{t|t-1})((\alpha_t - \widehat{\alpha}_{t|t-1})'\mathbf{H}_t' + \epsilon_t')\right] = \mathbf{S_{t|t-1}}\mathbf{H}_t', \tag{200}$$

since the observation error ϵ_t' is white noise and independent of $\alpha_t - \widehat{\alpha}_{t|t-1}$.

▷ Plugging this covariance into (199), we can write:

$$\widehat{\alpha}_t = \widehat{\alpha}_{t|t-1} + \mathbf{K}_t(\mathbf{z}_t - \widehat{\mathbf{z}}_{t|t-1}) \tag{201}$$

where $\mathbf{K}_t$ is the matrix of regression coefficients which is called the gain of the filter, and is given by:

$$\mathbf{K}_t = \mathbf{S_{t|t-1}}\mathbf{H}_t'\mathbf{P}_{t|t-1}^{-1}.$$

The Kalman filter - Third step

▷ Equation (201) indicates that the revision we make of the prior estimation to the state depends on the prediction error, $\mathbf{e}_t = \mathbf{z}_t - \widehat{\mathbf{z}}_{t|t-1}$.

▷ If this error is zero, we do not modify the estimation, otherwise, we make a modification in the state estimation that depends on the quotient of the error in the state estimation, $\mathbf{S}_{t|t-1}$, and the prediction error $\mathbf{P}_{t|t-1}^{-1}$.

▷ An equivalent way of writing equation (201) is

$$\widehat{\alpha}_t = (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\widehat{\alpha}_{t|t-1} + \mathbf{K}_t\mathbf{z}_t$$

which indicates that the state estimation is a linear combination of the two sources of information that are available to us:

On the one hand, the prior estimation, $\widehat{\alpha}_{t|t-1}$, and on the other, observation $\mathbf{z}_t$ that also provides information about the state.

The Kalman filter - Third step

▷ The covariance matrix of this estimation is

$$\mathbf{S}_t = E\left[(\alpha_t - \widehat{\alpha}_t)(\alpha_t - \widehat{\alpha}_t)'|Z_t\right]$$

and replacing $\widehat{\alpha}_t$ with its expression in equation (201), we have

$$\mathbf{S}_t = E\left[(\alpha_t - \widehat{\alpha}_{t|t-1} - \mathbf{K}_t\mathbf{e}_t)(\alpha_t - \widehat{\alpha}_{t|t-1} - \mathbf{K}_t\mathbf{e}_t)'|Z_t\right]$$

and utilizing (200), we finally obtain

$$\mathbf{S}_t = \mathbf{S_{t|t-1}} - \mathbf{S_{t|t-1}}\mathbf{H}_t'\mathbf{P}_{\mathbf{t|t-1}}^{-1}\mathbf{H}_t\mathbf{S_{t|t-1}}. \tag{202}$$

The Kalman filter equations

▷ Equations (194), (195), (196), (198), (201) and (202) comprise the Kalman filter:

$$\widehat{\alpha}_{t|t-1} = \mathbf{\Omega}_t \widehat{\alpha}_{t-1}$$

$$\mathbf{S}_{t|t-1} = \mathbf{\Omega}_t \mathbf{S}_{t-1} \mathbf{\Omega}_t' + \mathbf{R}_t.$$

$$\widehat{\mathbf{z}}_{t|t-1} = E(\mathbf{z}_t | Z_{t-1}) = \mathbf{H}_t \widehat{\alpha}_{t|t-1}.$$

$$\mathbf{P}_{t|t-1} = \mathbf{H}_t \mathbf{S}_{\mathbf{t}|\mathbf{t-1}} \mathbf{H}_t' + \mathbf{V}_t.$$

$$\widehat{\alpha}_t = \widehat{\alpha}_{t|t-1} + \mathbf{K}_t (\mathbf{z}_t - \widehat{\mathbf{z}}_{t|t-1})$$

$$\mathbf{S}_t = \mathbf{S}_{\mathbf{t}|\mathbf{t-1}} - \mathbf{S}_{\mathbf{t}|\mathbf{t-1}} \mathbf{H}_t' \mathbf{P}_{\mathbf{t}|\mathbf{t-1}}^{-\mathbf{1}} \mathbf{H}_t \mathbf{S}_{\mathbf{t}|\mathbf{t-1}}.$$

▷ Under the hypothesis of normality the filter provides optimal estimations and predictions.

Properties of the estimators

▷ It can be proved that the asymptotic properties of the maximum likelihood method are valid, under some regularity conditions, for ML estimators of ARMA models.

▷ These conditions require the process to be stationary and that the ARMA model we are estimating not contain common factors in its AR and MA part.

▷ For stationary processes in large samples the ML estimators have an asymptotical normal distribution and they are asymptotically unbiased and efficient.

▷ In particular, the matrix of second derivatives of the support evaluated at its maximum directly provide the variances and covariances of the estimators:

$$\mathbf{Var}\left(\widehat{\beta}_{MV}\right) = -\left[\frac{\partial^2 L\left(\widehat{\beta}_{MV}\right)}{\partial\beta\partial\beta'}\right]^{-1}$$

Properties of the estimators

▷ The condition under which there are no common factors in the AR and MA part is important.

▷ For example, if ω_t is white noise and we estimate the model

$$(1 - \phi B)\,\omega_t = (1 - \theta B)\,a_t$$

all the values of the parameters with the condition $\phi = \theta$ are compatible with the data and it can be proved that the variance of the estimators is infinite.

▷ In general, if the model is overparameterized and simultaneously has redundant AR and MA factors we will have a situation of strong multicollinearity which can give rise to multiple maximums in the likelihood function.

Estimation of ARIMA models - Example

Example 88. *We are going to estimate the models identified for the Spanish vehicle registration series. The four programs we will use are TSW, which uses the Kalman filter and exact maximum likelihood, the SCA which also uses exact ML and EViews and Minitab which use conditional ML.*

Program	*Model*	$\widehat{\sigma}_a$
TSW	$\nabla\nabla_{12}\ln M_t = (1-0.61B)\left(1-0.78B^{12}\right)a_t$	*0.123*
SCA	$\nabla\nabla_{12}\ln M_t = (1-0.61B)\left(1-0.78B^{12}\right)a_t$	*0.122*
Minitab	$\nabla\nabla_{12}\ln M_t = (1-0.62B)\left(1-0.84B^{12}\right)a_t$	*0.119*
EViews	$\nabla\nabla_{12}\ln M_t = (1-0.59B)\left(1-0.85B^{12}\right)a_t$	*0.119*
TSW	$\left(1-0.21B^{12}\right)\nabla\nabla_{12}\ln M_t = (1-0.61B)\left(1-0.89B^{12}\right)a_t$	*0.121*
SCA	$\left(1-0.14B^{12}\right)\nabla\nabla_{12}\ln M_t = (1-0.61B)\left(1-0.85B^{12}\right)a_t$	*0.117*
Minitab	$(1-0.25B)\nabla\nabla_{12}\ln M_t = (1-0.62B)\left(1-0.95B^{12}\right)a_t$	*0.116*
EViews	$\left(1-0.20B^{12}\right)\nabla\nabla_{12}\ln M_t = (1-0.59B)\left(1-0.92B^{12}\right)a_t$	*0.116*

Estimation of ARIMA models - Example

Dependent Variable: D(LREG,1,12)
Method: Least Squares
Date: 02/11/08 Time: 16:02
Sample (adjusted): 1961M02 1999M12
Included observations: 467 after adjustments
Convergence achieved after 12 iterations
Backcast: 1960M01 1961M01

Variable	Coefficient	Std. Error	t-Statistic	Prob.
MA(1)	-0.590730	0.036126	-16.35184	0.0000
SMA(12)	-0.849112	0.022492	-37.75125	0.0000

R-squared	0.506632	Mean dependent var	-0.000828
Adjusted R-squared	0.505571	S.D. dependent var	0.169220
S.E. of regression	0.118988	Akaike info criterion	-1.415315
Sum squared resid	6.583538	Schwarz criterion	-1.397557
Log likelihood	332.4760	Durbin-Watson stat	2.023869

Dependent Variable: D(LREG,1,12)
Method: Least Squares
Date: 02/11/08 Time: 16:06
Sample (adjusted): 1962M02 1999M12
Included observations: 455 after adjustments
Convergence achieved after 10 iterations
Backcast: 1961M01 1962M01

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(12)	0.200860	0.048032	4.181812	0.0000
MA(1)	-0.588512	0.037092	-15.86627	0.0000
SMA(12)	-0.919005	0.017842	-51.50692	0.0000

R-squared	0.535343	Mean dependent var	-0.000767
Adjusted R-squared	0.533287	S.D. dependent var	0.167318
S.E. of regression	0.114306	Akaike info criterion	-1.493309
Sum squared resid	5.905740	Schwarz criterion	-1.466142
Log likelihood	342.7278	Durbin-Watson stat	2.078833

▷ *We can see that the best model from the point of view of residual variance is that which has an ARMA(1,1) in the seasonal part.*

▷ *One conclusion from this exercise is that the greatest difference between the exact and conditional estimations appears when the model has moving average terms close to the unit value such that the process is close to non-invertibility.*

Model selection criteria

▷ Let us assume that we have estimated a set of models, $M_1, ..., M_m$, and we wish to select the one which best explains the observed series.

▷ Selecting the model by its fit to our given sample does not yield suitable results, since the model with the most parameters always leads to greater likelihood and a smaller sum of squares error within the sample.

▷ For example, if we compare an AR(p) with an AR($p+1$) the fit of the AR($p+1$) cannot be worse than that of the AR(p), and we will always choose the most complex model. Therefore, in order to choose between models we must turn to other principles.

▷ The problem can be looked at as one of discrimination: we have different models M_i and a stationary series, $\omega_T = (\omega_1, ..., \omega_T)$, and we wish to select the model most compatible with the observed series; this can be approached from a classical or Bayesian point of view.

Model selection criteria

▷ Beginning with the classical approach, we see that comparing the likelihood of different models is of little use because the model with the most parameters will always have greater likelihood.

▷ However, we can calculate the expected value of the likelihood for each one of the models and select the model that produces an expected value that is higher than this expected likelihood. This is the approach that leads to the Akaike information criterion.

▷ If we have the a priori probabilities for each model, $P(M_i)$, we could use the Bayesian approach and select the model that has maximum probability given the data:

$$P(M_i|\omega_T) = \frac{P(\omega_T|M_i)P(M_i)}{\sum_{i=1}^{m} P(\omega_T|M_j)P(M_j)}. \tag{203}$$

▷ If we assume that the a priori probabilities of all the models are the same, this approach leads to the Bayesian information criterion, explained below.

Model selection criteria - Akaike information criterion

▷ The likelihood function of an ARIMA model is given by (174). Multiplying by -2 and taking expectations in this expression we have

$$E(-2L(\beta)) = T \ln \sigma^2 + \sum_{t=1}^{T} \ln v_{t|t-1} + E\left[\sum_{t=1}^{T} \frac{e_t^2}{\sigma^2 v_{t|t-1}}\right].$$

▷ It is proved that

$$AIC = E(-2L(\beta)) = T \ln \widehat{\sigma}^2_{MV} + 2k, \tag{204}$$

where T is the sample size used to estimate the model, $\widehat{\sigma}^2_{MV}$ the ML estimator of the variance of the innovations and k the number of parameters estimated to calculate the one step ahead predictions.

▷ Therefore, selecting the model with maximum expected likelihood is equivalent to choosing that which minimizes the likelihood with a negative sign given by (204). This criterion is known as the **AIC criterion**.

Model selection criteria - Bayeasian information criterion

▷ An alternative criterion was proposed by Schwarzusing a Bayesian approach. The criterion is to maximize the a posteriori probability of the model, $P(M_i|\omega)$, assuming that the a priori probabilities are the same for all the models.

▷ Since, according to (203), $P(M_i|\omega)$ is proportional to $P(\omega|M_i)P(M_i)$, if the a priori probabilities are the same, the a posteriori probability of the model is proportional to $P(\omega|M_i)$.

▷ Selecting the model that maximizes this probability is equivalent to selecting the model that minimizes $-2\ln P(\omega|M_i)$.

▷ It can be proved that the model that asymptotically minimizes this quantity is the one that minimizes the criterion:

$$BIC = T \ln \widehat{\sigma}^2_{MV} + k \ln T, \tag{205}$$

where T is the sample size, $\widehat{\sigma}^2_{MV}$ the ML estimator of the variance and k the number of parameters.

Comparison of model selection criteria

▷ A problem with the AIC is that it tends to overestimate the number of parameters in the model and this effect can be important in small samples.

▷ If we compare the expression (205) with (204) we see that the BIC penalizes the introduction of new parameters more than the AIC does, hence it tends to choose more parsimonious models.

▷ It can be proved that the BIC criterion is consistent, in the sense that when the data have been generated by an ARIMA model the BIC selects the appropriate order of the model with a probability of one.

▷ On the other hand, the AIC criterion is efficient, in the sense that if the data are generated by a model that could be of infinite order, and we consider a sequence of estimators whose order increases with the sample size, the selected predictor is that with the lowest expected prediction error.

Selection of ARIMA models - Example

Example 89. *We are going to apply the model selection criteria in order to choose a model for the vehicle registration series. The table gives the model, the residual variance, the number of parameters and the value of the corresponding selection criterion:*

Modelo	$\widehat{\sigma}^2$	*T*	*k*	*BIC*	*AIC*
ARIMA$(0,1,1)\times(0,1,1)_{12}$	*0.119*^{2}	*466*	*2*	*-1.9799*	*-1.9716*
ARIMA$(0,1,1)\times(1,1,1)_{12}$	*0.116*^{2}	*466*	*3*	*-2.0017*	*-1.9892*

▷ *The best model using both the BIC as well as the AIC is the second one, which obtains the lowest value using both criteria. This would then be the model chosen.*

▷ Notice that EViews use an slight different definitions for AIC and BIC:

$$AIC_{EViews} = -2\ln(L/T)+2k/T \text{ and } BIC_{EViews} = -2\ln(L/T)+\ln Tk/T,$$

where L is the full likelihood function (including inessential constant terms).

Estimation and selection of ARIMA models - Example

Example 90. *We are going to compare some models estimated for the series on work related accidents:*

- ARIMA$(0,1,1) \times (0,1,1)$
- ARIMA$(2,1,0) \times (0,1,1)$
- ARIMA$(2,1,0) \times (1,1,1)$

Dependent Variable: D(LWA,1,12)
Method: Least Squares
Date: 02/11/08 Time: 16:14
Sample (adjusted): 1980M02 1998M12
Included observations: 227 after adjustments
Convergence achieved after 15 iterations
Backcast: 1979M01 1980M01

Variable	Coefficient	Std. Error	t-Statistic	Prob.
MA(1)	-0.493807	0.056078	-8.805724	0.0000
SMA(12)	-0.893467	0.022294	-40.07662	0.0000

R-squared	0.457425	Mean dependent var	0.000856
Adjusted R-squared	0.455013	S.D. dependent var	0.087703
S.E. of regression	0.064745	Akaike info criterion	-2.627936
Sum squared resid	0.943193	Schwarz criterion	-2.597760
Log likelihood	300.2707	Durbin-Watson stat	2.156429

Dependent Variable: D(LWA,1,12)
Method: Least Squares
Date: 02/11/08 Time: 16:16
Sample (adjusted): 1980M04 1998M12
Included observations: 225 after adjustments
Convergence achieved after 14 iterations
Backcast: 1979M04 1980M03

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	-0.578750	0.064125	-9.025366	0.0000
AR(2)	-0.279150	0.064026	-4.359932	0.0000
MA(12)	-0.912116	0.016276	-56.04015	0.0000

R-squared	0.482486	Mean dependent var	0.001194
Adjusted R-squared	0.477824	S.D. dependent var	0.087672
S.E. of regression	0.063353	Akaike info criterion	-2.666935
Sum squared resid	0.891029	Schwarz criterion	-2.621387
Log likelihood	303.0301	Durbin-Watson stat	2.001199

Dependent Variable: D(LWA,1,12)
Method: Least Squares
Date: 02/11/08 Time: 16:17
Sample (adjusted): 1981M04 1998M12
Included observations: 213 after adjustments
Convergence achieved after 11 iterations
Backcast: 1980M04 1981M03

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	-0.584658	0.066305	-8.817657	0.0000
AR(2)	-0.301333	0.066217	-4.550720	0.0000
SAR(12)	0.253553	0.067781	3.740758	0.0002
MA(12)	-0.928355	0.014642	-63.40277	0.0000

R-squared	0.500921	Mean dependent var	0.000865
Adjusted R-squared	0.493757	S.D. dependent var	0.086068
S.E. of regression	0.061238	Akaike info criterion	-2.729493
Sum squared resid	0.783772	Schwarz criterion	-2.666370
Log likelihood	294.6910	Durbin-Watson stat	2.009470

▷ Of the three models the best ones are the last two, according to the BIC criterion.

▷ The first is an approximation of the second, because $(1 - 0.54B)^{-1} = (1 + 0.54B + .29B^2 + .16B^3 + ...)$ and if we truncate the series and keep the first two values we have an AR(2) similar to the one estimated by the second model.

▷ The third model seems to pick up the seasonality better since the AR term is significant, and it has the smallest BIC value of the three although the differences between the second and third model are slight.