Lecture 8: Estimation of Univariate ARIMA Models
Bus 41910, Time Series Analysis, Mr. R. Tsay

In time series analysis, parameter estimation is commonly carried out by the maximum likelihood method. In the case of AR models, the least squares method is also used. The likelihood function for n observations $Z_1, \cdots, Z_n$ of a stationary Gaussian time series is

$$f(Z_1, \cdots, Z_n) = (\frac{1}{2\pi})^{n/2}|V|^{-1/2}\exp[-\boldsymbol{Z}'V^{-1}\boldsymbol{Z}/2]$$

where $\boldsymbol{Z} = (Z_1, Z_2, \cdots, Z_n)'$ and V is the covariance matrix of $\boldsymbol{Z}$. Here we ignore the mean of Z_t which is always estimated by the sample mean.
The covariance matrix V is an $n \times n$ matrix. A direct evaluation of the above likelihood function is often difficult. Some simplification must be sought. For the ARMA class of models, V is a function of the parameters ϕ_i's, θ_j's and σ_a^2. Our objective is therefore to express the likelihood function in terms of the unknown parameters.

1 Predictive Error Decomposition

:
We can factor the joint density of the observations using the fact that

$$\begin{aligned} f(Z_1, \cdots, Z_n) &= f(Z_n|Z_{n-1}, \cdots, Z_1)f(Z_{n-1}, \cdots, Z_1) \\ &= \cdots \\ &= f(Z_n|Z_{n-1}, \cdots, Z_1)f(Z_{n-1}|Z_{n-2}, \cdots, Z_1)\cdots f(Z_2|Z_1)f(Z_1). \end{aligned}$$

The log-likelihood function $\ell = \ln f(Z_1, \cdots, Z_n)$ is then

$$\ell = \sum_{t=2}^{n} \ln f(Z_t|Z_{t-1}, \cdots, Z_1) + \ln f(Z_1).$$

For Guassian time series, each of the conditional densities is normal. Let $\mu_t = E(Z_t|Z_{t-1}, \cdots, Z_1)$ and $\sigma_t^2 = \text{Var}(Z_t|Z_{t-1}, \cdots, Z_1)$. Then, the log-likelihood function is

$$\ell = -\frac{n}{2}\ln(2\pi) - \sum_{t=2}^{n}[\frac{1}{2}\ln\sigma_t^2 + \frac{(Z_t-\mu_t)^2}{2\sigma_t^2}] - \frac{1}{2}\ln\sigma_z^2 - \frac{Z_1^2}{2\sigma_z^2}$$

where σ_z^2 is the variance of Z_t.
Let $e_t = Z_t - \mu_t$. For stationary ARMA processes, $\sigma_t^2 = \sigma_a^2$ (at least for most of t) and the log-likelihood function becomes

$$\ell \propto -\frac{n-1}{2}\ln\sigma_a^2 - \sum_{t=2}^{n}\frac{e_t^2}{2\sigma_a^2} - \frac{1}{2}\ln\sigma_z^2 - \frac{Z_1^2}{2\sigma_z^2}. \quad (1)$$

By using the predictive error decomposition, we have written the quadratic from $\boldsymbol{Z}'V^{-1}\boldsymbol{Z}$ as $\sum e_t^2$. It looks as though we don't have to invert the covariance matrix V! Actually, we are basically doing a Cholesky decomposition of V^{-1}:

$$V^{-1} = UDU'$$

where U is upper triangular with ones on the diagonal, D is a diagonal matrix. This factorization of V^{-1} is unique and can be computed for any positive definite symmetric matrix. The elements of D are the conditional variances σ_t^2 and if we write $\boldsymbol{e} = U'\boldsymbol{Z}$, then $\boldsymbol{Z}'V^{-1}\boldsymbol{Z} = \boldsymbol{e}'D\boldsymbol{e}$.
The Eq. (1) is only an approximation because σ_t^2 is not equal to σ_a^2 for all t,especially when t is small. To gain further insight, we consider some simple examples.

2 Conditional and exact likelihood functions for AR(1) models

For a zero-mean stationary AR(1) series,

$$Z_t = \phi Z_{t-1} + a_t$$

we have $Z_t - E(Z_t|Z_{t-1}, \cdots, Z_1) = Z_t - \phi Z_{t-1} = a_t$ for $t \geq 2$. Also, Z_1 is normal with mean 0 and variance $\sigma_a^2/(1-\phi^2)$. Therefore, the log-likelihood function is

$$\ell(\phi, \sigma_a^2) \propto -\frac{n}{2}\ln\sigma_a^2 + \frac{1}{2}\ln(1-\phi^2) - \frac{1-\phi^2}{2\sigma_a^2}Z_1^2 - \frac{1}{2\sigma_a^2}\sum_{t=2}^{n}(Z_t - \phi Z_{t-1})^2.$$

To find the MLE of ϕ and σ_a^2, we do the following: (1) concentrate out σ_a^2; (2) differentiate the resulting function with respect to ϕ. Such estimates are the exact maximum likelihood estimates.

If we condition on Z_1, the conditional log-likelihood function is

$$\ell(\phi, \sigma_a^2|Z_1) \propto -\frac{n-1}{2}\ln\sigma_a^2 - \frac{1}{2\sigma_a^2}\sum_{t=2}^{n}(Z_t - \phi Z_{t-1})^2.$$

Maximizing this conditional log-likelihood function is equivalent to minimizing the sum of squares of residuals $\sum_{t=2}^{n}(Z_t - \phi Z_{t-1})^2$, which gives the least squares estimate of ϕ for an AR(1) model. We can justify this conditional likelihood as an approximation to the exact likelihood. We realize that the terms of the likelihood function pertaining to the first observation Z_1 are of order 1 and the other terms are of order n so that the normalized likelihood function could be well approximated by the conditional likelihood function.
There are some minor differences between the exact and conditional MLE in this AR(1) case. For instance, the term $\ln(1-\phi^2)$ keeps the exact MLE in the stationary region. There is no such a built-in constriant for the conditional MLE.
<u>Exercise</u>: Generalize the above results to AR(p) processes.

3 Conditional likelihood function for MA(1) processes

The exact likelihood functions of MA and ARMA processes are relatively complicated. For general information, you may consult Hillmer and Tiao (1979, JASA) and Ansley (1979, BKA). These two papers present two different approaches to obtain the exact likelihood function of an ARMA model. Here I shall start with the conditional likelihood function before introducing the exact likelihood function of some simple models.

For the MA(1) series

$$Z_t = a_t - \theta a_{t-1},$$

we have $a_t = Z_t + \theta a_{t-1}$. Therefore,

$$\begin{aligned}
a_1 &= Z_1 + \theta a_0 \\
a_2 &= Z_2 + \theta Z_1 + \theta^2 a_0 \\
a_3 &= Z_3 + \theta Z_2 + \theta^2 Z_1 + \theta^2 a_0 \\
&\vdots \\
a_n &= Z_n + \theta Z_{n-1} + \cdots + \theta^{n-1} Z_1 + \theta^n a_0
\end{aligned}$$

Or equivalently, in matrix form, we have

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & & & \\ \theta & 1 & 0 & \cdots & & & \\ \theta^2 & \theta & 1 & 0 & \cdots & & \\ \vdots & \vdots & & & & & \\ \theta^{n-1} & \theta^{n-2} & \cdots & & & \theta & 1 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ \vdots \\ Z_n \end{bmatrix} + \begin{bmatrix} \theta \\ \theta^2 \\ \theta^3 \\ \vdots \\ \theta^n \end{bmatrix} a_0. \qquad (2)$$

Consequently, if we condition on $a_0 = 0$, then there is a one-to-one transformation between $\boldsymbol{Z}$ and $\boldsymbol{a} = (a_1, a_2, \cdots, a_n)'$, and the Jacobian of the transformation is unity. Therefore, the conditional likelihood function is

$$f(Z_1, \cdots, Z_n | a_0 = 0) = f(a_1, \cdots, a_n) = \prod_{t=1}^{n} f(a_t)$$

and the conditional log-likelihood function is

$$\ell(\theta, \sigma_a^2 | a_0 = 0) \propto -\frac{n}{2} \ln \sigma_a^2 - \frac{1}{2\sigma_a^2} \sum_{t=1}^{n} a_t^2$$

where $a_t = Z_t + \theta Z_{t-1} + \cdots + \theta^{t-1} Z_1$. Again, we can concentrate out σ_a^2. Maximizing the resulting concentrated log-likelihood function amounts to minimizing the sum of squares $\sum_{t=1}^{n} a_t^2$, which is a non-linear function of θ so that non-linear optimization is required. The usual procedure of non-linear optimization is to use iterative methods. You may consult any non-linear optimization routine for further information.

Turn to the exact MLE. Here a_0 is treated as an unknown initial value of the innovational noise and must be dealt with precisely. Mathematically, the joint density of $\boldsymbol{Z}$ can be obtained by integrating out a_0 as

$$f(Z_1, \cdots, Z_n) = \int f(Z_1, \cdots, Z_n, a_0) da_0.$$

Thus, we need to consider the joint density of $f(Z_1, \cdots, Z_n, a_0)$. First, to simplify the notation, we rewrite equation (2) as

$$\boldsymbol{a} = \boldsymbol{L_1 Z} + \boldsymbol{X}_1 a_0.$$

Next, augmenting the indentity $a_0 = a_0$ on top of the above equation, we obtain

$$\boldsymbol{A} = \boldsymbol{LZ} + \boldsymbol{X} a_0 \tag{3}$$

where

$$\boldsymbol{A} = \begin{bmatrix} a_0 \\ \boldsymbol{a} \end{bmatrix}_{(n+1)\times 1}, \quad \boldsymbol{L} = \begin{bmatrix} \boldsymbol{O} \\ \boldsymbol{L}_1 \end{bmatrix}_{(n+1)\times n}, \quad \boldsymbol{X} = \begin{bmatrix} 1 \\ \boldsymbol{X}_1 \end{bmatrix}_{(n+1)\times 1}$$

and $\boldsymbol{O} = (0, \cdots, 0)$ is a row-vector of n zeros. This equation can be used in two ways. First of all, it says that there is a one-to-one transformation between $\boldsymbol{A}$ and $(a_0, \boldsymbol{Z}')'$ with unit Jacobian of transformation. Consequently, we have

$$f(Z_1, \cdots, Z_n, a_0) = f(a_0, a_1, \cdots, a_n) = (\frac{1}{\sqrt{2\pi\sigma_a^2}})^n \exp[\frac{1}{2\sigma_a^2} \sum_{t=0}^{n} a_t^2].$$

Again, by (3),

$$\sum_{t=0}^{n} a_t^2 = \boldsymbol{A}'\boldsymbol{A} = (\boldsymbol{LZ} + \boldsymbol{X}a_0)'(\boldsymbol{LZ} + \boldsymbol{X}a_0).$$

Let

$$\hat{a}_0 = -(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{LZ}) = -\frac{(1-\theta^2)\sum_{t=1}^{n} \theta^t Y_t}{1-\theta^{2(n+1)}},$$

where $Y_t = Z_t + \theta Z_{t-1} + \cdots + \theta^{t-1} Z_1$, which is a function of $\boldsymbol{Z}$ and θ, not a_0. Then, the sum of squares becomes

$$\begin{aligned} \boldsymbol{A}'\boldsymbol{A} &= (\boldsymbol{LZ} + \boldsymbol{X}\hat{a}_0 - \boldsymbol{X}\hat{a}_0 + \boldsymbol{X}a_0)'(\boldsymbol{LZ} + \boldsymbol{X}\hat{a}_0 - \boldsymbol{X}\hat{a}_0 + \boldsymbol{X}a_0) \\ &= (\boldsymbol{LZ} + \boldsymbol{X}\hat{a}_0)'(\boldsymbol{LZ} + \boldsymbol{X}\hat{a}_0) + (a_0 - \hat{a}_0)'\boldsymbol{X}'\boldsymbol{X}(a_0 - \hat{a}_0). \end{aligned}$$

In the above, we have used

$$(a_0 - \hat{a}_0)\boldsymbol{X}'(\boldsymbol{LZ} + \boldsymbol{X}\hat{a}_0) = (a_0 - \hat{a}_0)[\boldsymbol{X}'\boldsymbol{LZ} - \boldsymbol{X}'\boldsymbol{LZ}] = 0.$$

The above derivation gives that

$$f(Z_1, \cdots, Z_n, a_0) = (\frac{1}{2\pi\sigma_a^2})^{(n+1)/2} \exp[\frac{-1}{2\sigma_a^2}\{(\boldsymbol{LZ}+\boldsymbol{X}\hat{a}_0)'(\boldsymbol{LZ}+\boldsymbol{X}\hat{a}_0)+(a_0-\hat{a}_0)\boldsymbol{X}'\boldsymbol{X}(a_0-\hat{a}_0)\}].$$

Finally, using properties of normal density to integrate out a_0, we obtain the joint density function of $\boldsymbol{Z}$

$$f(Z_1, \cdots, Z_n) = (\frac{1}{2\pi\sigma_a^2})^{n/2} |\boldsymbol{X}'\boldsymbol{X}|^{-1/2} \exp[-\frac{(\boldsymbol{LZ} + \boldsymbol{X}\hat{a}_0)'(\boldsymbol{LZ} + \boldsymbol{X}\hat{a}_0)}{2\sigma_a^2}]$$

where $|\boldsymbol{X}'\boldsymbol{X}| = \frac{1-\theta^{2(n+1)}}{1-\theta^2}$.
By equation (3) with a_0 in the right hand side replaced by $\hat{a}_0$, we have that

$$(\boldsymbol{LZ} + \boldsymbol{X}\hat{a}_0)'(\boldsymbol{LZ} + \boldsymbol{X}\hat{a}_0) = \boldsymbol{A}'\boldsymbol{A}|_{a_0=\hat{a}_0} = \sum_{t=0}^{n} a_{0,t}^2$$

where $a_{0,t}$ is an estimate of a_t obtained by the recursion in (2) with a_0 replaced by $\hat{a}_0$. Consequently, the exact log-likelihood function of an MA(1) process is

$$\ell(\theta, \sigma_a^2) \propto -\frac{n}{2} \ln \sigma_a^2 - \frac{1}{2} \ln(\frac{1-\theta^{2(n+1)}}{1-\theta^2}) - \frac{1}{2\sigma_a^2} \sum_{t=0}^{n} a_{0,t}^2.$$

In practice, exact MLEs require heavier computation than the conditional MLEs. This can easily be seen from the above log-likelihood function. Essentially, for a given value of θ, one needs to go through the data twice; one to estimate $\hat{a}_0$ and the second to compute $a_{0,t}$. However, the exact method provides more accurate estimates than the conditional one does, especially when θ is close to the unit circle for which the effect of initial noise a_0 is more persistent. As a rule of thumb, one can start with conditional estimates, then uses these estimates as initial values and performs an exact estimation. Of course, for large sample sizes, the differences between exact and conditional MLE are small.

The above approach in effect applies to ARMA models in general. The algebra is, of course, more involved. See Hillmer and Tiao (1979) for details. HT's paper is for multivariate time series, but univariate time series is just a special case.

Exercise: Derive the exact likelihood function of a stationary ARMA(1,1) model. (Hint: the inital values required are Z_0 and a_0.)

The approach of Ansley (1979) to evaluating exact likelihood function is slightly different. This approach is a direct generalization of the AR(1) case we discussed earlier. For a stationary ARMA(p, q) model, let $r = \max(p, q)$. Then, consider the likelihood function of $(Z_1, \cdots, Z_r)$, which involves inversion of an $r \times r$ covariance matrix. Then, partition the jointly density as

$$f(Z_1, \cdots, Z_n) = f(Z_{r+1}, \cdots, Z_n | Z_1, \cdots, Z_r) f(Z_1, \cdots, Z_r)$$

for which the first term in the right hand side can be transformed into $a_{r+1}, \cdots, a_n$.

4 Asymptotic properties of MLE

For a stationary and invertible ARMA model

$$\phi(B)Z_t = \theta(B)a_t$$

where $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$ are polynomials with no common factors and $\{a_t\}$ is a sequence of *iid* Gaussian random variables with mean zero and variance σ_a^2, the MLE (both conditional and exact) of the parameters ϕ's and θ's are (a) consistent and (b) asymptotically normal. More precisely, let $\boldsymbol{\beta} = (\phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q)'$, and $\hat{\boldsymbol{\beta}}$ the MLE of $\boldsymbol{\beta}$. Then,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N(\boldsymbol{O}, \boldsymbol{V}_\beta) \quad \text{as} \quad n \to \infty$$

where n is the sample size and $\boldsymbol{V}_\beta$ is a $(p+q) \times (p+q)$ covariance matrix defined by

$$\boldsymbol{V}_\beta = \sigma_a^2 \begin{bmatrix} E(\boldsymbol{u}_t \boldsymbol{u}_t') & E(\boldsymbol{u}_t \boldsymbol{v}_t') \\ E(\boldsymbol{v}_t \boldsymbol{u}_t') & E(\boldsymbol{v}_t \boldsymbol{v}_t') \end{bmatrix}^{-1}$$

where $\boldsymbol{u}_t = (u_t, u_{t-1}, \cdots, u_{t-p+1})'$ and $\boldsymbol{v}_t = (v_t, v_{t-1}, \cdots, v_{t-q+1})'$ with u_t and v_t satisfying

$$\phi(B)u_t = a_t, \quad \theta(B)v_t = -a_t.$$

The asymptotic normality of the MLE follows basically the usual argument as that of the *iid* case with CLT replaced by certain functional central limit theorem. You may consult standard time series textbooks for details, e.g. Box and Jenkins (1976) and Brockwell and Davis (1991, Ch. 8 & 10). Here we shall discuss the reason supporting the asymptotic variance $\boldsymbol{V}_\beta$.

Recall that the log-likelihood function of Z_t is approximately

$$\ell_n(\boldsymbol{\beta}) \propto -\frac{n}{2} \ln \sigma_a^2 - \frac{1}{2\sigma_a^2} \sum_{t=1}^{n} a_t^2$$

where $a_t = Z_t - \phi_1 Z_{t-1} - \cdots - \phi_p Z_{t-p} + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q}$. The asymptotic covariance matrix of the MLE of $\boldsymbol{\beta}$ is then the inverse of the expected Fisher information matrix

$$E[-\frac{\partial^2 \ell_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}] = E[(\frac{\partial \ell_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}})(\frac{\partial \ell_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}})'].$$

From the log-likelihood function, it is important to consider the derivatives of a_t with respect to $\boldsymbol{\beta}$ in order to compute the Fisher information matrix. It can also be shown that the MLE of $\boldsymbol{\beta}$ is asymptotically uncorrelated with that of σ_a^2 so that we can work on $\boldsymbol{\beta}$ and σ_a^2 separately.

Let

$$-\frac{\partial a_t}{\partial \phi_i} = u_{t-i}, \quad -\frac{\partial a_t}{\partial \theta_j} = v_{t-j}.$$

From $a_t = Z_t - \phi_1 Z_{t-1} - \cdots - \phi_p Z_{t-p} + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q}$, we have

$$u_{t-i} = Z_{t-i} + \theta_1 u_{t-1-i} + \cdots + \theta_q u_{t-q-i}$$

and

$$v_{t-j} = \theta_1 v_{t-1-j} + \cdots + \theta_q v_{t-q-j} - a_{t-j}.$$

In other words, we have

$$\theta(B) u_t = Z_t, \quad \text{and} \quad \theta(B) v_t = -a_t.$$

Next, since $Z_t = \frac{\theta(B)}{\phi(B)} a_t$, we further obtain

$$\theta(B) u_t = \frac{\theta(B)}{\phi(B)} a_t$$

so that

$$\phi(B) u_t = a_t.$$

From the stationarity and invertibility of Z_t, both u_t and v_t are stationary processes. Therefore, the Fisher information matrix is

$$\boldsymbol{V}_\beta^{-1} = \frac{n}{\sigma_a^2} E[\begin{pmatrix} \boldsymbol{u}_t \\ \boldsymbol{v}_t \end{pmatrix} (\boldsymbol{u}_t', \boldsymbol{v}_t')]$$

where $\boldsymbol{u}_t$ and $\boldsymbol{v}_t$ are defined as above.
In practice, the Fisher information matrix is evaluated by substituting $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$.

5 Some special cases

In what follows, we consider results of some simple ARMA models.
AR(p) Model: $\phi(B) Z_t = a_t$. For this model, the derived process u_t is an AR(p) process also. Therefore,

$$\boldsymbol{V}_\beta^{-1} = \frac{n}{\sigma_a^2} \boldsymbol{\Gamma}_p$$

where $\boldsymbol{\Gamma}_p$ is the covariance matrix of $(u_t, \cdots, u_{t-p+1})'$, or equivalently, the covariance matrix of $(Z_t, \cdots, Z_{t-p+1})'$. Consequently, we have

- AR(1): $\hat{\phi}_1 \sim N(\phi_1, \frac{1-\phi_1^2}{n})$.
- AR(2): $\hat{\boldsymbol{\phi}} \sim N(\boldsymbol{\phi}, \boldsymbol{V})$ where $\boldsymbol{\phi} = (\phi_1, \phi_2)'$ and

$$\boldsymbol{V} = \frac{1}{n} \begin{bmatrix} 1 - \phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1 - \phi_2^2 \end{bmatrix}.$$

MA(q) Model: $Z_t = \theta(B)a_t$. Since the derived process v_t satisfies $\theta(B)v_t = -a_t$, which is an AR(q) model, we obtain the same results as those of pure AR models.

- MA(1): $\hat{\theta}_1 \sim N(\theta_1, \frac{1-\theta_1^2}{n})$.
- MA(2): $\hat{\theta} \sim N(\boldsymbol{\theta}, V)$ where $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ and

$$\boldsymbol{V} = \frac{1}{n}\begin{bmatrix} 1-\theta_2^2 & -\theta_1(1+\theta_2) \\ -\theta_1(1+\theta_2) & 1-\theta_2^2 \end{bmatrix}.$$

Mixed ARMA(1,1) Model: $Z_t - \phi Z_{t-1} = a_t - \theta a_{t-1}$. For this process, the asymptotic covariance matrix of MLE of $\boldsymbol{\beta} = (\phi, \theta)'$ is

$$\boldsymbol{V} = \frac{1}{n}\frac{1-\phi\theta}{(\phi-\theta)^2}\begin{bmatrix} (1-\phi^2)(1-\phi\theta) & (1-\phi^2)(1-\theta^2) \\ (1-\phi^2)(1-\theta^2) & (1-\theta^2)(1-\phi\theta) \end{bmatrix}.$$

6 Illustration

R and SCA demonstration, including commands used.

Example 1. Use R to simulate and estimate the MA(1) model

$$Z_t = a_t - 0.6z_{t-1}, \quad \sigma_a^2 = 1.$$

Compare the theoretical result with the empirical result, e.g. the variance of the MA(1) estimate.

Example 2. Consider the ARMA(1,1) model

$$Z_t = 0.6Z_{t-1} + a_t + 0.4a_{t-1}, \quad \sigma_a^2 = 1.$$

Use simulation and estimation to check the asymptotic results.

Example 3. Real data analysis

Finally, Discuss model specification and parameter constriants in R and SCA.