Estimating an ARMA Process

Overview

1. Main ideas
2. Fitting autoregressions
3. Fitting with moving average components
4. Standard errors
5. Examples
6. Appendix: Simple estimators for autoregressions

Main ideas

Efficiency Maximum likelihood is nice, if you know the right distribution. For time series, its more motivation for least squares. If the data are Gaussian, then ML is efficient; if the data are not Gaussian, then ML amounts to complicated weighted least squares.

Gaussian log-likelihood for the stationary process $\{X_t\}$ that generates $\boldsymbol{X} = (X_1, \ldots, X_n)'$ is (minus twice the log of the likelihood)

$$-2\ell(\mu, \phi, \theta, \sigma^2) = n \log 2\pi + \log |\Gamma_n| + (\boldsymbol{X} - \mu)' \Gamma_n^{-1} (\boldsymbol{X} - \mu) \; . \quad (1)$$

Think of the covariances $\gamma(k)$ as functions of the parameters of the process,

$$\gamma(k) = \sigma^2 g(k; \phi, \theta) \; . \quad (2)$$

To find the maximum likelihood estimates of μ, ϕ, and θ for an ARMA(p, q) process is "simply a numerical minimization" of the negative log likelihood.

"All you need to do" is express the covariances in (1) as functions of the unknown parameters. For example, for the AR(1) process $X_t = \phi_1 X_{t-1} + w_t$ with $\mu = 0$ (given), $\gamma(0) = \sigma^2/(1 - \phi_1^2)$, and $\gamma(h) = \phi^{|h|}\gamma(0)$.

Recursion The models we consider are causal, with time "flowing" in one direction. Hence, it is useful to decompose the joint distribution of $\boldsymbol{X}$ in the log-likelihood (1) as a sequence of one-sided conditional distributions:

$$f(x_1,\ldots,x_n) = f(x_1)f(x_2|x_1)\; f(x_3|x_1,x_2)\cdots f(x_n|x_1,\ldots,x_{n-1})\;.$$

MLEs for AR(1) It's useful to solve for the MLE in closed form for the simplest of models. The log-likelihood (4) simplifies for a Gaussian AR(1) process:

$$\begin{aligned}
\ell(\phi,\sigma^2) &= \log f(X_1)f(X_2|X_1)\cdots f(X_n|X_{n-1}) \\
&= \log N(0,\frac{\sigma^2}{1-\phi^2})f(X_2|X_1)\cdots f(X_n|X_{n-1}) \\
&= -\frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2}\log(1-\phi^2) - \underbrace{\left(X_1^2(1-\phi^2) + \sum_{t=2}^{n}(X_t-\phi X_{t-1})^2\right)}_{SS}/2\sigma^2\;.
\end{aligned}$$

The derivatives that give the MLE's for σ^2 and ϕ are:

$$\frac{\partial\ell}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{SS}{2\sigma^4}$$

and

$$\frac{\partial\ell}{\partial\phi} = \frac{X_1^2\phi + \sum(X_t-\phi X_{t-1})X_{t-1}}{\sigma^2} - \frac{\phi}{1-\phi^2}$$

where SS denotes the sum of squares from the exponential term in the likelihood. Setting these to zero and solving gives:

$$\hat{\sigma}^2 = SS/n$$

and

$$\hat{\phi}\left(\frac{\hat{\sigma}^2}{1-\hat{\phi}^2} - X_1^2\right) = \sum_{t=2}^{n}(X_t-\hat{\phi}X_{t-1})X_{t-1}\;.$$

Since the l.h.s. of this expression has approximate expected value zero (note the expression for the variance of X_t, the MLE can be seen to be quite similar to the usual LS estimator.

Initial values Again thinking recursively, the likelihood looks a lot like that in the normal linear model *if* we only knew enough to get started. If we condition on X_1, X_2, $\ldots, X_p$ and w_1, w_2, $\ldots, w_q$ (*i.e.*, assume we know these), then

$$-2\log f(x_{p+1}^n | x_1^p, w_1^q) = c + (n-p)\log\sigma^2$$

$$+ \sum_{t=p+1}^{n} (w_t = x_t - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p} - \theta_1 w_{t-1} - \cdots - \theta_q w_{t-q})^2$$

by the change of variables from x_t to w_t, as in the normal linear model. This expression becomes amenable to least squares. This approach is called "conditional least squares" in **R**.

What does this leave out? Conditioning on the initial values does not leave out too much, really. We still

- Need values for w_1, $\ldots, w_q$, and
- Would like to gain further information from $X_1, \ldots, X_p$.

Autoregressions

Why start with autoregressions? Several reasons:

- These often fit quite well (don't need the MA terms) because we know that we can approximate $X_t = \sum \pi_j X_{t-j}$. This 'Markovian' approach has become more popular because of its speed.
- Estimation is fast (MLEs require some iteration).
- Avoid estimation of initial w_ts.
- Sampling properties are well-known (essentially those of normal linear model with stochastic explanatory variables).

Backwards Even if we don't want the AR model itself, these are often used to estimate the initial errors, w_1, w_2,, $\ldots, w_q$. By fitting an autoregression backwards in time, we can use the fit to estimate say, $\hat{w}_t^{(m)} = X_t - \sum_{j=1}^{m} \hat{\pi}_j X_{t+j}$ (if we assume normality, the process is reversible).

MLE for autoregression In the AR(p) case,

$$\begin{aligned} f(x_1,\ldots,x_n) &= \underbrace{f(x_1,x_2,\ldots,x_p)}_{\text{messy}}\underbrace{f(x_{p+1}|x_1^p)\cdots f(x_n|x_{n-p}^{n-1})}_{\text{simple}} \\ &= f(x_1,x_2,\ldots,x_p)\,\frac{e^{-\frac{1}{2}\sum_{p+1}^n w_i^2/\sigma^2}}{(2\pi\sigma^2)^{-(n-p)/2}}\ . \end{aligned}$$

where $w_t = (x_t-\mu) - \sum_j \phi_j(x_{t-j}-\mu)$ for $t = p+1,\ldots,n$. But for the initial terms (the "messy" part), we have the same likelihood as in the normal linear model, and the MLEs are those that we would get from the least squares regression of x_t on $x_{t-1},\ldots,x_{t-p}$ and a constant. (If we call the constant in that regression β_0, then $\hat{\mu} = \hat{\beta}_0/(1-\hat{\phi}_1-\cdots-\hat{\phi}_p)$. But for ignoring the contribution from $x_1,\ldots,x_p$, least squares matches maximum likelihood in the AR(p) case. Hence, maximum likelihood cannot improve the estimates much unless p is large relative to n.

Recursion = triangular factorization A recursion captures the full likelihood. For an AR(p) model with coefficients $\boldsymbol{\phi}_p = (\phi_{p1}, \phi_{p2}, \ldots \phi_{pp})$ express the lower-order coefficients as functions of $\boldsymbol{\phi}_p$ (*e.g.*, find $\gamma(0)$ and $\phi_{11} = \mathrm{Corr}(X_t, X_{t-1})$ in terms of $\boldsymbol{\phi}_p$). If we can do that, it is simple to model

$$f(x_1,\ldots,x_p) = f(x_1)f(x_2|x_1)\cdots f(x_p|x_1^{p-1})\ .$$

The prior AR(1) example shows this.

In general, use the *Levinson recursion* to obtain a triangular decomposition of the covariance matrix Γ_n. This is done by converting the correlated variables $X_1,\ldots,X_n$ into a collection, say U_1, U_2, …, U_n of uncorrelated variables. One has many ways of doing this, such as the Gram-Schmidt or Cholesky factorization. In the following, let P_j denote the projection onto the random variables in X_j (as in fitting a regression).

Following the Cholesky factorization, construct

$$\begin{aligned} U_1 &= X_1 \\ U_2 &= X_2 - P_1X_2 = X_2 - \phi_{1,1}X_1 \end{aligned}$$

$$\begin{aligned} U_3 &= X_3 - P_{12}X_3 = X_3 - \phi_{2,2}X_1 - \phi_{2,1}X_2 \\ U_4 &= X_4 - P_{123}X_4 = X_4 - \phi_{3,3}X_1 - \phi_{3,2}X_2 - \phi_{3,1}X_3 \\ U_j &= X_j - \sum_{k=1}^{j-1} \phi_{j-1,k}X_{j-k} \end{aligned}$$

(This sequence of projections differs from those used in the numerically superior modified Gram-Schmidt method. GS sweeps X_1 from all of the others first, filling the first column of L rather than recursively one row at a time.) Let L denote a lower triangular matrix that begins

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \dots \\ -\phi_{11} & 1 & 0 & 0 & 0 & \dots \\ -\phi_{22} & -\phi_{21} & 1 & 0 & 0 & \dots \\ -\phi_{33} & -\phi_{32} & -\phi_{31} & 1 & 0 & \dots \\ \vdots & & & & \ddots & \dots \end{bmatrix} \tag{3}$$

with diagonal elements $L_{kk} = 1$ and off-diagonal elements $L_{kj} = -\phi_{k-1,k-j}, j = 1, \dots, k-1$. Since the U_j are uncorrelated, we have

$$D_n = \text{Var}(U = LX) = L\Gamma_n L' \quad \Rightarrow \quad \Gamma_n^{-1} = L'^{-1} D_n^{-1} L^{-1}\,, \tag{4}$$

where D_n is the diagonal matrix with the conditional variances

$$\sigma_k^2 = \text{Var}\left(X_t - \sum_{j=1}^{k} \phi_{kj} X_{t-j}\right)$$

along the diagonal.

Comments

- It follows that $\sigma^2 = \lim_n |\Gamma_{n+1}|/|\Gamma_n|$. (Where have we seen this ratio of determinants before? Toeplitz matrices, Szegö's theorem, and Levinson's recursion itself.)
- If the process is indeed AR(p), the lower triangular matrix L is banded, with p subdiagonal stripes. The element $L_{kj} = 0$ for $j < k - p$.

If not an AR model? Since $\tilde{w}_t$ is a linear combination of X_s, $s = 1, \ldots, t$, the approximation is a simple way of computing/representing

$$\tilde{w}_t = X_t - \sum_j \pi_j X_{t-j} \; .$$

The terms become more accurate ($\tilde{w}_t$ approaches w_t) as t increases since

$$\|w_t - \tilde{w}_t\| = \| \textstyle\sum_{j=t}^{\infty} \pi_j X_{t-j}\| \leq \sqrt{\gamma(0)} c^t$$

for some $|c| < 1$ (since the π_j are eventually a sum of geometric series). We'll do an example of this in **R**; see the accompanying script.

ARMA via regression Given estimates of w_t, fit the "regression" model using the estimates of the error process as covariates. That is, regress X_t on its p lags and q lags of the estimated errors,

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \theta_1 \tilde{w}_{t-1} + \cdots + \theta_q \tilde{w}_{t-q} + w_t^* \; .$$

This method works rather well, plus it opens up the use of all of the tools of *regression diagnostics* for use in time series analysis. The method dates back to J. Durbin (1960) "The fitting of time series models" (*Intl Statis. Review*, 233-244) and was exploited in a model selection context by Hannan and Kavaleris (1984) (*Biometrika* 71, 273-280).

Standard errors Another insight from this approach is that it suggests the properties of estimators of ϕ and θ. Estimates of ϕ (ignoring the MA components) resemble those from usual regression, only with lags. Hence, for an AR(p) model, it follows that $\mathrm{Var}(\hat{\phi}) \approx \sigma^2 \, \hat{\Gamma}_p^{-1}$

Maximum likelihood

ARMA(p, q) models As noted earlier, the MLE is "easy" to compute if we condition on prior observations $X_0, \ldots, X_{-p+1}$ and prior errors $w_0, \ldots, w_{-q+1}$, for then the likelihood can be written in terms of $w_1, \ldots, w_n$, which are *iid* Gaussian. Estimation then reduces to minimizing $\sum_{t=1}^n w_t^2$.

Without this conditioning, the resulting estimates of $w_1, w_2, \ldots, w_n$ are *functions* of the parameters ϕ and θ, we can chose parameters that minimize this sum, giving a *nonlinear least squares* problem.

Approximate noise terms Assume for this section that $\mu = 0$ so that we can focus on the estimates of ϕ and θ. Let $\beta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$ denote the vector of unknown coefficients. Consider setting the prior unknown X_j and w_j for $j \le 0$ to their marginal expected values (0) and define

$$\begin{aligned}
\tilde{w}_1(\beta) &= X_1, \\
\tilde{w}_2(\beta) &= X_2 - \phi_1 X_1 - \theta_1 \tilde{w}_1, \\
\tilde{w}_3(\beta) &= X_3 - \phi_1 X_2 - \phi_2 X_1 - \theta_1 \tilde{w}_2 - \theta_2 \tilde{w}_1, \\
&\cdots \\
\tilde{w}_n(\beta) &= X_n - \sum_j \phi_j X_{n-j} - \sum_j \theta_j \tilde{w}_{n-j}
\end{aligned}$$

Backcasting In general, algorithms *start* like this, and then find estimates of x_t and w_t for $t \le 0$ (the method is known as "backcasting"). We will avoid those methods and consider (coming lecture) methods based on the *Kalman filter* (used in **R**). We will continue this discussion, though, to explore how the *iterative* estimation proceeds and discover the form of the asymptotic variance matrix of the estimates.

Nonlinear least squares The nonlinear SS is approximately a least squares problem

$$\sum_{t=1}^{n} \tilde{w}_t(\beta)^2 \approx \sum_{t=1}^{n} (\tilde{w}_t(\beta_0) - D_t'(\beta - \beta_0))^2, \tag{5}$$

where the derivative term is

$$D_{tj} = -\frac{\partial \tilde{w}_t}{\partial \beta_j} \, .$$

The negative sign is chosen to make the sum (5) look like a regression sum-of-squares $\sum_i (y_i - x_i'\beta)^2$ with D_t playing the role of x_i and $\tilde{w}_t(\beta_0)$ as y_i. The estimate of β is thus

$$\hat{\beta} = \beta_0 + (D'D)^{-1} D' \tilde{w}(\beta_0) \, .$$

Continuing the analogy to least squares, the variance matrix of $\hat{\beta}$ is then (approximating $\tilde{w}_t \approx w_t$),

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (D'D)^{-1}.$$

Derivatives simplify once the estimates are beyond the direct presence of the initial conditions where we filled in zeros. For $t > \max(p, q)$,

$$D_{tj} = -\frac{\partial \tilde{w}_t}{\partial \phi_j} = X_{t-j} - \sum_{k=1}^{q} \theta_k D_{t-k,j}, \quad j = 1, \ldots, p,$$

and, for the moving average terms, (use the product rule for derivatives)

$$D_{t,p+j} = \tilde{w}_{t-j} - \sum_{k=1}^{q} \theta_k D_{t-k,p+j}, \quad j = 1, \ldots, q.$$

That is, $\theta(B) D_{t,j} = X_{t-j}$ for the AR terms and $\theta(B) D_{t,p+j} = \tilde{w}_{t-j}$ for the MA terms. The derivatives thus behave as two AR processes for which (see Property 3.9)

$$D_{tj} \approx B^j u_t, \quad u_t = \frac{1}{\theta(B)} X_{t-j} = \frac{1}{\phi(B)} w_t$$

and

$$D_{t,p+j} \approx B^j v_t, \quad v_t = \frac{1}{\theta(B)} w_t \; .$$

Hence, the matrix $D'D/n$ is approximately the variance covariance of two distinct AR processes, both with errors w_t.

Example: ARMA(1,1) process The variance/covariance matrix of the estimates is (page 135 in 2nd edition, 134 in 3rd)

$$n \; \mathrm{Var}(\tilde{\phi}, \tilde{\theta}) \approx \sigma^2 \begin{bmatrix} \mathbb{E}\, u_t^2 & \mathbb{E}\, u_t \, v_t \\ \mathbb{E}\, u_t \, v_t & \mathbb{E}\, v_t^2 \end{bmatrix}^{-1} = \begin{bmatrix} 1/(1-\phi^2) & 1/(1+\phi\theta) \\ 1/(1+\phi\theta) & 1/(1-\theta^2) \end{bmatrix}^{-1}$$

A related issue at this point is whether it is better to use the expression here for the *expected Fisher information*, or to use the observed matrix of second derivatives (which is available within the actual estimation routine as $D'D$). There is a long literature that compares the use of expected versus observed Fisher information.

Asymptotic distributions

Information matrix MLEs in "regular" problems are asymptotically normal, unbiased, and have variance determined by the information matrix (to this order of approximation). The information matrix is the inverse of the matrix of the expected values of the second derivatives of the log-likelihood function. That is,

$$\sqrt{n}(\tilde{\theta} - \theta) \sim\rightarrow N(0, I_n)$$

where $I_n = -(E\, \frac{\partial^2 \ell}{\partial \theta^2})^{-1}$. For the AR(1) example, we have

$$\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \phi^2}\right) = \mathbb{E}\left(\frac{X_1^2 + \sum_2^n X_{t-1}^2}{\sigma^2} - \frac{1-3\phi^2}{(1-\phi^2)^2}\right) \approx \frac{n}{1-\phi^2}\,,$$

as we obtained for least squares.

Independence of $\tilde{\sigma}^2$ and $\tilde{\phi}$ The mixed partial is

$$\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \sigma^2 \partial \phi}\right) = \mathbb{E}\left(\frac{\partial SS/\partial \phi}{2\sigma^4}\right) = -\mathbb{E}\left(\frac{\phi X_1^2 + \sum w_t X_{t-1}}{\sigma^4}\right) = 0$$

so that the estimated coefficients and error variance estimate are asymptotically independent.

Invariance The variance of $\hat{\phi}$ or $\hat{\theta}$ is invariant of σ^2. The accuracy of the estimate of ϕ only depends upon the correlations of the process, not the level of the noise variance.

Examples

Examples using US unemployment are in the file `macro.R` on the text web site.

Alternative conditions. These results are utopian in the sense that they hold when all of the assumptions (large n, normality, known orders,...) hold. Simulation makes it easy to explore what happens when these assumptions fail. Interesting examples that you can explore are:

- Skewness in the distribution of the AR(1) estimator.

- Effect of outliers in the w_t process.
- Boundary effects as ϕ_1 nears 1.
- Effect of model specification errors.

Effects on prediction What are the optimal predictors once we replace ϕ and θ by estimates, particularly when the fitted model may not match p and q of the underlying data generating mechanism? (As if the underlying process were an ARMA process!)

Appendix: Estimators for Autoregressions

Introduction In applications, easily fitting autoregressions is important for obtaining initial values of parameters and in getting estimates of the error process. Least squares is a popular choice, as is the Yule Walker procedure. Unlike the YW procedure, LS can produce a non-stationary fitted model. Both the YW and LS estimators are non-iterative and consistent, so they can be used as *starting values* for iterative methods like ML.

Yule-Walker estimators are basically moment estimators, and so have the consistency and asymptotic normality associated with smooth functions of moment estimators (Slutsky's theorem). For an autoregression of order p, the Yule-Walker estimator solves the system $\hat{\Gamma}_p\hat{\phi}_p = \hat{\gamma}_p$.

The Yule-Walker estimator has two important properties not shared by other estimators for autoregressions:

1. Estimates for a sequence of models of order $p = 1, 2, \ldots, n-1$ are easily computed recursively using the Levinson algorithm applied to the estimated covariance matrix $\hat{\Gamma}_n$.
2. The associated estimates define stationary processes. That is the zeros of $\hat{\phi}(z)$ *must* must lie outside the unit circle.

The stationarity follows from observing that $\hat{\Gamma}_p$ is positive definite.

Asymptotics The result is essentially that which obtains in a corresponding regression equation. *Assuming* that the process is indeed AR(p), we have the same asymptotic distribution for all three of the estimators:

$$\sqrt{n}(\hat{\phi} - \phi) \sim AN(\phi, \Gamma_p^{-1})$$

A feature useful for *identification* of p is embedded here. If we overfit a model of order $m > p$, the fitted model is still "correct" (though not efficient) with $\phi_j = 0$ for $j = p+1, \ldots, m$. Consequently, the overfit coefficients have mean zero with variance that is $O(1/n)$.

Similar properties The methods for deriving properties of the YW and LS estimators are similar. The LS estimator is considered here, and properties of the YW estimator follow similarly. The two types of

estimators are asymptotically equivalent (though rather different in reasonably sized applications).

Estimating equations Write the normal equations for the LS estimator as

$$\sqrt{n}(\hat{\phi} - \phi) = (\boldsymbol{X}'\boldsymbol{X}/n)^{-1}(\boldsymbol{X}'\boldsymbol{w})/\sqrt{n}$$

where the $n-p \times p$ matrix $\boldsymbol{X}$ and vector $\boldsymbol{w}$ are defined as

$$\begin{aligned} \boldsymbol{X} &= (\boldsymbol{X}'_p, \ldots, \boldsymbol{X}'_n)', \quad \boldsymbol{X}_t = (X_t, \ldots, X_{t-p+1})' \\ \boldsymbol{w} &= (w_p, \ldots, w_n)' \end{aligned}$$

The cross-product matrix is nonsingular due to the a.s. convergence

$$\boldsymbol{X}'\boldsymbol{X}/n \stackrel{\text{a.s.}}{\to} \Gamma_p$$

and non-singularity of Γ_n. Asymptotically, then,

$$\sqrt{n}(\hat{\phi} - \phi) \sim \Gamma_n^{-1}(\boldsymbol{X}'\boldsymbol{w})/\sqrt{n} \; .$$

Return of the sandwich To make life easier, assume that the w_t are independent. It follows that

$$\sqrt{n}(\hat{\phi} - \phi) \sim AN(0, \Gamma_n^{-1} \operatorname{Var}(\boldsymbol{X}'\boldsymbol{w}/\sqrt{n})\Gamma_n^{-1}),$$

so we are left to deal with $\boldsymbol{X}'\boldsymbol{w}/\sqrt{n}$. Consider the jth element of this vector, (not related to terms in Levinson's method considered earlier)

$$U_j = \sum_{t=p+1}^{n} w_t X_{t-j}/\sqrt{n} \; .$$

From the assumed independence, the variance and covariance of the U_j's are:

$$\begin{aligned} n \operatorname{Cov}(U_j, U_k) &= E \sum_{s,t=p+1}^{n} w_t X_{t-j} w_s X_{s-k} \\ &= n\sigma^2\gamma(j-k) \; , \end{aligned}$$

or, in matrix notation, $\operatorname{Var}(U) = \sigma^2\Gamma_p$. Hence, the asymptotic distribution is

$$\sqrt{n}(\hat{\phi} - \phi) \sim AN(0, \sigma^2\Gamma_n^{-1}) \; ,$$

as one would anticipate from the results for regression ($\sigma^2(X'X)^{-1}$).

Weaker assumptions. You can obtain the same result without independence by assuming that the w_t's are stationary to order 4 and writing the process as an infinite moving average $X_t = \sum \psi_j w_{t-j}$, or by assuming that the w_t's form a martingale difference sequence. Taking the former approach,

$$U_j = \sum_{t=p+1}^{n} w_t \left(\sum_{m=0}^{\infty} \psi_k w_{t-j-m} \right) / \sqrt{n} \, ,$$

and the covariance is (assuming $\kappa_4 = 0$)

$$\begin{aligned} n \operatorname{Cov}(U_j, U_k) &= \sum_{s,t=p+1}^{n} \sum_{\ell,m} \psi_\ell \psi_m E(w_t w_s w_{t-j-m} w_{s-k-\ell}) \\ &= \sum_{s,t} \sum_{\ell,m} \psi_\ell \psi_m E(w_t w_s) E(w_{t-j-m} w_{s-k-\ell}) \\ &= \sum_{t} \sum_{m=k-j+\ell} \psi_\ell \psi_m E(w_t^2) E(w_{t-k-m} w_{t-k-\ell}) \\ &= (n-p)\sigma^4 \sum_{m} \psi_\ell \psi_{\ell+k-j} \\ &= (n-p)\sigma^2 \gamma(k-j) \, , \end{aligned}$$

as before.