STAT 153 & 248 - Time Series
Lecture Eighteen

Spring 2025, UC Berkeley

Aditya Guntuboyina

April 01, 2025

1 AR models: estimation, inference and prediction

The AR(p) model is given by:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t \tag{1}$$

The unknown parameters are $\phi_0, \phi_1, \ldots, \phi_p$ as well as σ (σ is the standard deviation of ϵ_t). These need to be estimated from the observed data $y_1, \ldots, y_n$.

The likelihood is (below θ denotes the vector consisting of all the parameters $\phi_0, \ldots, \phi_p$ and σ):

$$f_{y_1,\ldots,y_n|\theta}(y_1, \ldots, y_n) = f_{y_{p+1},\ldots,y_n|y_1,\ldots,y_p,\theta}(y_{p+1}, \ldots, y_n) f_{y_1,\ldots,y_p|\theta}(y_1, \ldots, y_p).$$

The conditional likelihood is calculated as

$$\begin{aligned}
&f_{y_{p+1},\ldots,y_n|y_1,\ldots,y_p,\theta}(y_{p+1}, \ldots, y_n) \\
&= \prod_{t=p+1}^{n} f_{y_t|y_{t-1},\ldots,y_1,\theta}(y_t) \\
&= \prod_{t=p+1}^{n} f_{\phi_0+\phi_1 y_{t-1}+\cdots+\phi_p y_{t-p}+\epsilon_t|y_{t-1},\ldots,y_1,\theta}(y_t) \\
&= \prod_{t=p+1}^{n} f_{\epsilon_t|y_{t-1},\ldots,y_1,\theta}(y_t - \phi_0 - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p}).
\end{aligned}$$

In order to proceed further, we shall make the following assumption:

$$\epsilon_t \mid y_{t-1}, \ldots, y_1 \sim N(0, \sigma^2) \qquad \text{for each } t = p+1, \ldots, n. \tag{2}$$

This is equivalent to assuming that $\epsilon_t \sim N(0, \sigma^2)$ **and that** ϵ_t **is independent of** $y_1, \ldots, y_{t-1}$. With (2), we get

$$\begin{aligned}
&f_{y_{p+1},\ldots,y_n|y_1,\ldots,y_p,\theta}(y_{p+1}, \ldots, y_n) \\
&= \prod_{t=p+1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_t - \phi_0 - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p})^2}{2\sigma^2}\right) \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n-p} \exp\left(-\frac{1}{2\sigma^2} \sum_{t=p+1}^{n} (y_t - \phi_0 - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p})^2\right).
\end{aligned}$$

Observe that, in order to write the above formula, we only used the model equation (1) for $t = p+1, \ldots, n$.

The conditional joint density $f_{y_{p+1},\ldots,y_n|y_1,\ldots,y_p,\theta}(y_{p+1},\ldots,y_n)$ is called the **conditional likelihood** of the AR(p) model. The full likelihood is

$$\begin{aligned}
&f_{y_1,\ldots,y_n|\theta}(y_1,\ldots,y_n) \\
&= f_{y_{p+1},\ldots,y_n|y_1,\ldots,y_p,\theta}(y_{p+1},\ldots,y_n) f_{y_1,\ldots,y_p|\theta}(y_1,\ldots,y_p) \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n-p} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=p+1}^{n}(y_t - \phi_0 - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p})^2\right) f_{y_1,\ldots,y_p|\theta}(y_1,\ldots,y_p).
\end{aligned}$$

If we assume that $f_{y_1,\ldots,y_p|\theta}(y_1,\ldots,y_p)$ does not depend on θ, then maximizing the full likelihood is equivalent to maximizing the conditional likelihood.

If we want to derive $f_{y_1,\ldots,y_p|\theta}(y_1,\ldots,y_p)$ in a more principled way, then we have to use the model equation (1) for smaller values of t (i.e., $t = p, p-1, p-2, \ldots, 0, -1, \ldots$). This makes the analysis complicated and is not really worth it. It also only works under some "stationarity" assumptions on $\phi_0, \ldots, \phi_p$. It is much simpler working with the conditional likelihood.

Using the matrix notation:

$$Y_{(n-p)\times 1} = \begin{pmatrix} y_{p+1} \\ y_{p+2} \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \qquad X_{(n-p)\times(p+1)} = \begin{pmatrix} 1 & y_p & y_{p-1} & \cdots & y_1 \\ 1 & y_{p+1} & y_{p+2} & \cdots & y_2 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & y_{n-1} & y_{n-2} & \cdots & y_{n-p} \end{pmatrix} \qquad \beta_{(p+1)\times 1} = \begin{pmatrix} \phi_0 \\ \phi_1 \\ \cdot \\ \cdot \\ \cdot \\ \phi_p \end{pmatrix},$$

the conditional likelihood (which is also proportional to the full likelihood under the assumption that $f_{y_1,\ldots,y_p|\theta}(y_1,\ldots,y_p)$ does not depend on θ) becomes:

$$\text{likelihood} \propto \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n-p} \exp\left(-\frac{\|Y - X\beta\|^2}{2\sigma^2}\right). \tag{3}$$

This likelihood is the same as the likelihood in linear regression with $n-p$ observations. We can therefore infer the parameters $\phi_0, \ldots, \phi_p$ and σ as in usual linear regression with the prior:

$$\phi_0, \phi_1, \ldots, \phi_p, \log\sigma \overset{\text{i.i.d}}{\sim} \text{unif}(-C, C).$$

This will allow us to write down the joint posterior of (β, σ). Integrating over σ leads to the posterior of β alone. As in Lecture 4, this leads to

$$\beta \mid \text{data} \sim t_{n-2p-1,p+1}\left(\hat{\beta}, \hat{\sigma}^2(X^TX)^{-1}\right)$$

where

$$\hat{\beta} := (X'X)^{-1}X'Y \quad \text{and} \quad \hat{\sigma} = \sqrt{\frac{\|Y - X\hat{\beta}\|^2}{n-2p-1}}.$$

Note that the degrees of freedom of the t-distribution above is $n-2p-1$ as the number of observations equals $n-p$ and the number of components of β is $p+1$. If inference for σ is desired, one can use:

$$\frac{\|Y - X\hat{\beta}\|^2}{\sigma^2} \mid \text{data} \sim \chi^2_{n-2p-1}.$$

Note that Bayesian inference for AR models is identical to Bayesian inference for linear regression models because the likelihood (3) is the same as in the usual linear model (with $n-p$ observations). Bayesian inference only cares about the likelihood.

Frequentist inference for the AR(p) model is based on the MLE which is given by $\hat{\beta}$ and

$$\hat{\sigma}_{\text{MLE}} = \sqrt{\frac{\|Y - X\hat{\beta}\|^2}{n-p}}.$$

To obtain frequentist confidence intervals for the parameters ϕ_i, one needs to find the distribution of $\hat{\beta}$. Here the analysis is quite different from that used in linear regression (see, for example, Section 3.5 of the book by Shumway and Stoffer titled *Time Series Analysis and its applications* (Fourth Edition)). The results turn out to be quite close to those obtained by the Bayesian method.

Unlike Bayesian inference, frequentist inference for the AR model is not identical to frequentist inference for the usual linear regression model. For example, one does not use t-distributions for inferring the ϕ parameters in AR(p) models. Instead, one uses normal distributions (e.g., z-scores as opposed to t-scores) which are justified by asymptotic arguments that are different from and more complicated than those used for linear regression.

2 Predictions given by $AR(p)$ models

One important goal of time series analysis is prediction also known as forecasting: given the observed data $y_1, \ldots, y_n$, what can we say about the future observations $y_{n+1}, \ldots, y_{n+k}$ for some $k \geq 1$? In the Bayesian context, prediction is done via the joint probability distribution of

$$y_{n+1}, \ldots, y_{n+k}$$

conditional on the observed data $y_1, \ldots, y_n$. For example, point predictions can be obtained by the conditional expectations:

$$\mathbb{E}\left(y_{n+1} \mid y_1, \ldots, y_n\right), \ldots, \mathbb{E}\left(y_{n+k} \mid y_1, \ldots, y_n\right).$$

Uncertainty quantification for the predictions can be done via the conditional variances:

$$\text{var}\left(y_{n+1} \mid y_1, \ldots, y_n\right), \ldots, \text{var}\left(y_{n+k} \mid y_1, \ldots, y_n\right).$$

Let us focus on point predictions using conditional expectations for now. We shall deal with uncertainty quantification for the predictions in the next class. The conditional expectations can be written as

$$\mathbb{E}\left(y_{n+i} \mid y_1, \ldots, y_n\right) = \int \mathbb{E}\left(y_{n+i} \mid y_1, \ldots, y_n, \theta\right) f_{\theta|y_1,\ldots,y_n}(\theta) d\theta \tag{4}$$

for $i = 1, \ldots, n$.

Let us first calculate

$$\hat{y}_{n+i}(\theta) := \mathbb{E}\left(y_{n+i} \mid y_1, \ldots, y_n, \theta\right)$$

for fixed parameters θ. These can be calculated recursively for $i = 1, 2, \ldots$ as follows. Assuming the validity of the model equation (1) also for $t > n$, we get

$$\begin{aligned}
\hat{y}_{n+i}(\theta) &= \mathbb{E}\left(y_{n+i} \mid y_1, \ldots, y_n, \theta\right) \\
&= \mathbb{E}\left(\phi_0 + \phi_1 y_{n+i-1} + \phi_2 y_{n+i-2} + \cdots + \phi_p y_{n+i-p} \mid y_1, \ldots, y_n, \theta\right) \\
&= \phi_0 + \phi_1 \hat{y}_{n+i-1}(\theta) + \phi_2 \hat{y}_{n+i-2}(\theta) + \cdots + \phi_p \hat{y}_{n+i-p}(\theta).
\end{aligned}$$

We thus have the following recursion for the predictions $\hat{y}_{n+i}(\theta)$:

$$\hat{y}_{n+i}(\theta) = \phi_0 + \phi_1 \hat{y}_{n+i-1}(\theta) + \phi_2 \hat{y}_{n+i-2}(\theta) + \cdots + \phi_p \hat{y}_{n+i-p}(\theta) \qquad \text{for } i = 1, 2, \ldots. \tag{5}$$

If we initialize this recursion with

$$\hat{y}_j(\theta) = y_j \qquad \text{for } j = n, n-1, \ldots, n+1-p, \tag{6}$$

then (5) can be evaluated in sequence for $i = 1, 2, \ldots$ to calculate $\hat{y}_{n+i}(\theta)$ for all $i \geq 1$.

Let us get back to the conditional expectation (4):

$$\mathbb{E}(y_{n+i} \mid y_1, \ldots, y_n) = \int \hat{y}_{n+i}(\theta) f_{\theta|y_1,\ldots,y_n}(\theta) d\theta \tag{7}$$

To compute the integral above, we can do one of two things:

1. We can first generate posterior samples $\theta^{(1)}, \ldots, \theta^{(N)}$ from the posterior $f_{\theta|y_1,\ldots,y_n}(\theta)$. Then (7) is approximated as

$$\mathbb{E}(y_{n+i} \mid y_1, \ldots, y_n) \approx \frac{1}{N} \sum_{\ell=1}^{N} \hat{y}_{n+i}(\theta^{(\ell)}).$$

2. For a simpler approach, we can use the fact that the posterior density $f_{\theta|y_1,\ldots,y_n}(\theta)$ is usually highly concentrated around the point estimate $\hat{\theta} = (\hat{\beta}, \hat{\sigma})$. We can then ignore the small uncertainty of θ around $\hat{\theta}$ to write

$$\mathbb{E}(y_{n+i} \mid y_1, \ldots, y_n) = \int \hat{y}_{n+i}(\theta) f_{\theta|y_1,\ldots,y_n}(\theta) d\theta \approx \hat{y}_{n+i}(\hat{\theta}).$$

   This second method avoids posterior sampling is faster and simpler.

3 Prediction Uncertainty

We shall next discuss how to provide uncertainty quantification for predictions given by the $AR(p)$ model. This can be done via the variance of the future observations given the data. Specifically by

$$\text{var}\,(y_{n+i} \mid y_1, \ldots, y_n) \qquad \text{for } i = 1, 2, \ldots.$$

and the corresponding standard deviations. We approximate these conditional variances as (below we write "data" for $y_1, \ldots, y_n$)

$$\text{var}\,(y_{n+i} \mid \text{data}) = \mathbb{E}\,(\text{var}\,(y_{n+i} \mid \theta, \text{data}) \mid \text{data}) + \text{var}\,(\mathbb{E}\,(y_{n+i} \mid \theta, \text{data}) \mid \text{data})$$

The second term above is generally small. This is because $\mathbb{E}\,(y_{n+i} \mid \theta, \text{data})$ is a function of θ and then we take the expectation of θ with respect to the posterior distribution. Because the posterior distribution is usually quite concentrated, the variance will be small. We shall thus ignore the second term and write

$$\text{var}\,(y_{n+i} \mid \text{data}) \approx \mathbb{E}\,(\text{var}\,(y_{n+i} \mid \theta, \text{data}) \mid \text{data})$$

To calculate this, the main task is to compute

$$V_i(\theta) := \text{var}\,(y_{n+i} \mid \theta, \text{data})$$

It turns that it is difficult to directly setup a recursion for $V_i(\theta)$. Instead, we will get the recursion by working with the conditional **covariance matrices** of $Y_{n+1}, \ldots, Y_{n+k}$ (given θ and the data) for $k = 1, 2, \ldots$. Let us first review some basic formulae for covariance matrices.

3.1 Covariance Matrices

A finite number of random variables can be viewed together as a random vector. More precisely, a random vector is a vector whose entries are random variables. Let $Y = (Y_1, \ldots, Y_n)^T$ be an $n \times 1$ random vector. Its Expectation $\mathbb{E}Y$ is defined as a vector whose ith entry is the expectation of Y_i i.e., $\mathbb{E}Y = (\mathbb{E}Y_1, \mathbb{E}Y_2, \ldots, \mathbb{E}Y_n)^T$. The covariance matrix of Y, denoted by $\text{Cov}(Y)$, is an $n \times n$ matrix whose (i,j)th entry is the covariance between Y_i and Y_j. Two important but easy facts about $\text{Cov}(Y)$ are:

1. The diagonal entries of $\text{Cov}(Y)$ are the variances of $Y_1, \ldots, Y_n$. More specifically the (i,i)th entry of the matrix $\text{Cov}(Y)$ equals $var(Y_i)$.
2. $\text{Cov}(Y)$ is a symmetric matrix i.e., the (i,j)th entry of $\text{Cov}(Y)$ equals the (j,i) entry. This follows because $\text{Cov}(Y_i, Y_j) = \text{Cov}(Y_j, Y_i)$.

The following formulae are very important:

1. $\mathbb{E}(AY+c) = A\mathbb{E}(Y)+c$ for every deterministic matrix A and every deterministic vector c.
2. $\text{Cov}(AY + c) = A\text{Cov}(Y)A^T$ for every deterministic matrix A and every deterministic vector c.

As a consequence of the second formula above, we get

$$\text{var}(a^T Y) = a^T \text{Cov}(Y) a = \sum_{i,j} a_i a_j \text{Cov}(Y_i, Y_j) \qquad \text{for every } p \times 1 \text{ vector } a.$$

Given two random vectors Y $(p \times 1)$ and W $(q \times 1)$, we use $\text{Cov}(Y, W)$ to denote the $p \times q$ matrix whose $(i,j)^{th}$ entry equals the covariance $\text{Cov}(Y_i, W_j)$ between Y_i and W_j. With this definition, the previous notion of $\text{Cov}(Y)$ equals simply $\text{Cov}(Y, Y)$. It can be checked that

$$\text{Cov}(AY + c, BW + d) = A\text{Cov}(Y, W)B^T.$$

3.2 Covariance Recursion for Future Variables in $AR(p)$

We shall set up a recursion for the covariance matrices:

$$\Gamma_k(\theta) := \text{Cov}\left(\begin{pmatrix} y_{n+1} \\ \cdot \\ \cdot \\ \cdot \\ y_{n+k} \end{pmatrix} \mid \theta, \text{data} \right)$$

The $(i,j)^{th}$ entry of $\Gamma_k(\theta)$ is

$$\text{Cov}\left(y_{n+i}, y_{n+j} \mid y_1, \ldots, y_n, \theta\right).$$

We shall see how to do this in the next lecture.

3.3 Optional Additional Reading for Today

1. For more on fitting $AR(p)$ models to data, see Section 3.5 of the book by Shumway and Stoffer titled *Time Series Analysis and its applications* (Fourth Edition).
2. For more on prediction with AR models, see Section 3.4 of the Shumway-Stoffer book.