# STAT 153 & 248 - Time Series
# Lecture Eighteen

**Fall 2025, UC Berkeley**

Aditya Guntuboyina

October 30, 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 $\sigma$ ($\sigma$ is the standard deviation of $\epsilon_t$). These need to be estimated from the observed data $y_1, \ldots, y_n$.

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

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

The first term on the right hand side above $f_{y_{p+1},\ldots,y_n|y_1,\ldots,y_p,\theta}(y_{p+1}, \ldots, y_n)$ is the conditional likelihood of $y_{p+1}, \ldots, y_n$ given $y_1, \ldots, y_p$. This 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:

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

which can be ensured by assuming that $\epsilon_t \sim N(0, \sigma^2)$ **and that $\epsilon_t$ 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 $\theta$, 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 $\theta$) 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 $\sigma$ 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 $(\beta, \sigma)$. Integrating over $\sigma$ leads to the posterior of $\beta$ 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 $\beta$ is $p+1$. If inference for $\sigma$ 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 $\phi_i$, 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 $\phi$ 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.

On the computer, estimation and inference for AR models can be done in two ways:

1. Just create $y$ and $X$ as above, and use OLS in statsmodels.
2. Use the AutoReg function in statsmodels.

Both methods give the same estimates of $\phi_0, \ldots, \phi_p$. The estimate of $\sigma$ is slightly different: OLS gives $\hat{\sigma}_{OLS} := \sqrt{RSS/(n-2p-1)}$ and AutoReg gives $\hat{\sigma}_{MLE} := \sqrt{RSS/(n-p)}$. The two methods also give slightly different standard errors corresponding to the coefficient estimates. OLS gives square roots of the diagonal entries of $\hat{\sigma}^2_{OLS}(X^TX)^{-1}$ while AutoReg considers $\hat{\sigma}^2_{MLE}(X^TX)^{-1}$. Finally OLS does coefficient inference using the $t$-distribution (with $n-2p-1$ degrees of freedom) while AutoReg recommends inference using the normal distribution.

## 2 Predictions from $AR(p)$ models

If the parameters $\phi_0, \ldots, \phi_p$ and $\sigma$ (collectively denoted by $\theta$) of the AR(p) model are fixed, then the prediction for a future value $y_{n+i}$ is given by

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

These values are calculated recursively for $i = 1, 2, \ldots$ using the following recursion:

$$\hat{y}_{n+i}(\theta) = \phi_0 + \phi_1 \hat{y}_{n+i-1}(\theta) + \phi_2 \hat{y}_{n+i-2}(\theta) + \cdots + \phi_p \hat{y}_{n+i-p}(\theta) \qquad \text{for } i = 1, 2, \ldots. \tag{4}$$

which is initialized by

$$\hat{y}_j(\theta) = y_j \qquad \text{for } j = n, n-1, \ldots, n+1-p, \tag{5}$$

Since $\theta$ is unknown, we can replace it by the conditional MLE $\hat{\theta}$.

# 3 Prediction Standard Errors

To compute prediction standard errors, we can again fix the parameter values $\theta$, and then attempt to calculate:

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

The prediction standard error corresponding to the predicted value for $y_{n+i}$ can then be taken to be $\sqrt{V_i(\hat{\theta})}$ (note that $\theta$ is replaced by the conditional MLE $\hat{\theta}$).

It turns out that it is not really possible to setup a recursion for $V_i(\theta)$. To see this, note that:

$$\begin{aligned}
V_1(\theta) &= \text{var}(y_{n+1} \mid y_1, \ldots, y_n, \theta) \\
&= \text{var}(\phi_0 + \phi_1 y_n + \phi_2 y_{n-1} + \cdots + \phi_p y_{n+1-p} + \epsilon_{n+1} \mid y_1, \ldots, y_n, \theta) \\
&= \text{var}(\epsilon_{n+1} \mid y_1, \ldots, y_n, \theta) = \sigma^2.
\end{aligned}$$

Next

$$\begin{aligned}
V_2(\theta) &= \text{var}(y_{n+2} \mid y_1, \ldots, y_n, \theta) \\
&= \text{var}(\phi_0 + \phi_1 y_{n+1} + \phi_2 y_n + \cdots + \phi_p y_{n+2-p} + \epsilon_{n+2} \mid y_1, \ldots, y_n, \theta) \\
&= \text{var}(\phi_1 y_{n+1} + \epsilon_{n+2} \mid y_1, \ldots, y_n, \theta) \\
&= \phi_1^2 \text{var}(y_{n+1} \mid y_1, \ldots, y_n, \theta) + \text{var}(\epsilon_{n+2} \mid y_1, \ldots, y_n, \theta) \\
&= \phi_1^2 V_1(\theta) + \sigma^2 = \sigma^2(1 + \phi_1^2).
\end{aligned}$$

The next term is

$$\begin{aligned}
V_3(\theta) &= \text{var}(y_{n+3} \mid y_1, \ldots, y_n, \theta) \\
&= \text{var}(\phi_0 + \phi_1 y_{n+2} + \phi_2 y_{n+1} + \cdots + \phi_p y_{n+3-p} + \epsilon_{n+3} \mid y_1, \ldots, y_n, \theta) \\
&= \text{var}(\phi_1 y_{n+2} + \phi_2 y_{n+2} + \epsilon_{n+3} \mid y_1, \ldots, y_n, \theta) \\
&= \text{var}(\phi_1 y_{n+2} + \phi_2 y_{n+2} \mid y_1, \ldots, y_n, \theta) + \text{var}(\epsilon_{n+3} \mid y_1, \ldots, y_n, \theta).
\end{aligned}$$

The first term in the right hand side above cannot be written down in terms of $V_1(\theta)$ a nd $V_2(\theta)$ alone. It also involves the covariance between $y_{n+1}$ and $y_{n+2}$ (given $y_1, \ldots, y_n, \theta$).

Instead of working with the variances alone, we will get the recursion by working with the conditional **covariance matrices** of $y_{n+1}, \ldots, y_{n+k}$ (given $\theta$ and the data) for $k = 1, 2, \ldots$.

Below we review some basic facts about 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 $i$th 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 in $AR(p)$

We shall set up a recursion for the covariance matrices:

$$\Gamma_k(\theta) := \text{Cov}\left( \begin{pmatrix} y_{n+1} \\ \cdot \\ \cdot \\ \cdot \\ y_{n+k} \end{pmatrix} \mid \theta, y_1, \ldots, y_n \right)$$

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

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

The diagonal entries of $\Gamma_k(\theta)$ equal $V_1(\theta), \ldots, V_k(\theta)$.

To initialize the recursion for $\Gamma_k(\theta)$, note that

$$\begin{aligned} \Gamma_1(\theta) &= \text{var}\left(y_{n+1} \mid y_1, \ldots, y_n, \theta\right) \\ &= \text{var}\left(\phi_0 + \phi_1 y_n + \ldots \phi_p y_{n+1-p} + \epsilon_{n+1} \mid y_1, \ldots, y_n, \theta\right) \\ &= \text{var}\left(\epsilon_{n+1} \mid y_1, \ldots, y_n, \theta\right) = \sigma^2 \end{aligned}$$

We now relate $\Gamma_{k+1}(\theta)$ to $\Gamma_k(\theta)$ to establish the recursion. We can write

$$\Gamma_k(\theta) = \begin{pmatrix} \Gamma_{k-1}(\theta) & \gamma_{k1}(\theta) \\ \gamma_{k1}^T(\theta) & V_k(\theta) \end{pmatrix}$$

where

$$\gamma_{k1}(\theta) := \text{Cov}\left( \begin{pmatrix} y_{n+1} \\ \cdot \\ \cdot \\ \cdot \\ y_{n+k-1} \end{pmatrix}, y_{n+k} \mid \theta, y_1, \ldots, y_n \right)$$

and, as before,

$$V_k(\theta) = \mathrm{var}\left(y_{n+k} \mid \theta, y_1, \ldots, y_n\right).$$

We compute $\hat{\gamma}_{k1}$ as

$$\begin{aligned}
\gamma_{k1}(\theta) &:= \mathrm{Cov}\left(\begin{pmatrix} y_{n+1} \\ \cdot \\ \cdot \\ \cdot \\ y_{n+k-1} \end{pmatrix}, y_{n+k} \mid \theta, y_1, \ldots, y_n\right) \\
&= \mathrm{Cov}\left(\begin{pmatrix} y_{n+1} \\ \cdot \\ \cdot \\ \cdot \\ y_{n+k-1} \end{pmatrix}, \phi_0 + \phi_1 y_{n+k-1} + \phi_2 y_{n+k-2} + \cdots + \phi_p y_{n+k-p} + \epsilon_{n+k} \mid \theta, \mathrm{data}\right) \\
&= \mathrm{Cov}\left(\begin{pmatrix} y_{n+1} \\ \cdot \\ \cdot \\ \cdot \\ y_{n+k-1} \end{pmatrix}, \phi_1 y_{n+k-1} + \phi_2 y_{n+k-2} + \cdots + \phi_p y_{n+k-p} \mid \theta, \mathrm{data}\right) \\
&= \mathrm{Cov}\left(\begin{pmatrix} y_{n+1} \\ \cdot \\ \cdot \\ \cdot \\ y_{n+k-1} \end{pmatrix}, \sum_{i=1}^{k-1} a_i y_{n+i} \mid \theta, \mathrm{data}\right)
\end{aligned}$$

where, for $i = 1, \ldots, k-1$,

$$a_i = \begin{cases} \phi_{k-i} & \text{provided } k-p \le i \le k-1 \\ 0 & \text{otherwise} \end{cases}$$

Thus if $a$ is the $(k-1) \times 1$ vector with entries $a_1, \ldots, a_{k-1}$, we have

$$\gamma_{k1}(\theta) = \mathrm{Cov}\left(\begin{pmatrix} y_{n+1} \\ \cdot \\ \cdot \\ \cdot \\ y_{n+k-1} \end{pmatrix}, a^T \begin{pmatrix} y_{n+1} \\ \cdot \\ \cdot \\ \cdot \\ y_{n+k-1} \end{pmatrix} \mid \theta, \mathrm{data}\right) = \Gamma_{k-1}(\theta) a.$$

Further

$$\begin{aligned}
V_k(\theta) &= \mathrm{var}\left(y_{n+k} \mid \theta, \mathrm{data}\right) \\
&= \mathrm{var}\left(\phi_0 + \phi_1 y_{n+k-1} + \phi_2 y_{n+k-2} + \cdots + \phi_p y_{n+k-p} + \epsilon_{n+k} \mid \theta, \mathrm{data}\right) \\
&= \mathrm{var}\left(\phi_0 + \phi_1 y_{n+k-1} + \phi_2 y_{n+k-2} + \cdots + \phi_p y_{n+k-p} \mid \theta, \mathrm{data}\right) + \sigma^2 \\
&= \mathrm{var}\left(\sum_{i=1}^{k-1} a_i y_{n+i} \mid \theta, \mathrm{data}\right) + \sigma^2 = a^T \Gamma_{k-1}(\theta) a + \sigma^2.
\end{aligned}$$

The equation for obtaining $\Gamma_k(\theta)$ from $\Gamma_{k-1}(\theta)$ is therefore

$$\Gamma_k(\theta) = \begin{pmatrix} \Gamma_{k-1}(\theta) & \gamma_{k1}(\theta) \\ \gamma_{k1}^T(\theta) & V_k(\theta) \end{pmatrix} = \begin{pmatrix} \Gamma_{k-1}(\theta) & \Gamma_{k-1}(\theta) a \\ a^T \Gamma_{k-1}(\theta) & a^T \Gamma_{k-1}(\theta) a + \sigma^2 \end{pmatrix}$$

The algorithm for calculating the variances $V_i(\theta)$ for $i = 1, 2, \ldots, K$ is thus given by

1. Initialize with $\Gamma_1(\theta) = V_1(\theta) = \sigma^2$.

2. For $k = 2, 3, \ldots, K$, repeat the following

   a) Form the $(k-1) \times 1$ vector $a$ whose $i^{th}$ entry is $\phi_{k-i}$ if $k - p \leq i \leq k - 1$ and 0 otherwise.

   b) Calculate $\Gamma_k(\theta)$ using $\Gamma_{k-1}(\theta)$ and $a$ by the formula given above.

3. The variances $V_i(\theta), i = 1, 2, \ldots, K$ are given by the diagonal entries of the matrix $\Gamma_K(\theta)$.

Because $\theta$ is unknown, in practice, we run this recursion with $\theta$ replaced by its conditional MLE $\hat{\theta}$. The prediction standard errors are then the square roots of $V_i(\hat{\theta})$. These prediction errors coincide with those given by the `get_prediction` function in `statsmodels`.