STAT 153 & 248 - Time Series
Lecture Seventeen

Fall 2025, UC Berkeley

Aditya Guntuboyina

October 28, 2025

1 Parameter Estimation in AutoRegressive Models

We shall discuss parameter estimation in $AR(p)$ models. Recall that the $AR(p)$ is given by:

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

for $t = p+1, \dots, n$. In matrix notation,

$$Y = X\beta + \epsilon$$

where

$$Y = \begin{pmatrix} y_{p+1} \\ y_{p+2} \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & y_p & y_{p-1} & \cdot & \cdot & \cdot & y_1 \\ 1 & y_{p+1} & y_p & \cdot & \cdot & \cdot & y_2 \\ 1 & y_{p+2} & y_{p+1} & \cdot & \cdot & \cdot & y_3 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & y_{n-1} & y_{n-2} & \cdot & \cdot & \cdot & y_{n-p} \end{pmatrix} \quad \beta = \begin{pmatrix} \phi_0 \\ \phi_1 \\ \phi_2 \\ \cdot \\ \cdot \\ \cdot \\ \phi_p \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_{p+1} \\ \epsilon_{p+2} \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_n \end{pmatrix}$$

The $AR(p)$ can be seen as a spcial of the usual linear regression model where the covariate matrix X as well as the response vector y are both formed from a single data set $y_1, \dots, y_n$.

We shall discuss estimation of the parameters $\phi_0, \dots, \phi_p$ and σ in $AR(p)$. For simplicity, let us first assume $p = 1$ (we shall revert to the more general case later). The AR(1) model is given by:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t \qquad \text{for } t = 2, \dots, n. \tag{2}$$

Because of the close relation between $AR(p)$ models and linear regression, let us first revisit parameter estimation in usual linear regression.

2 Detour: usual linear regression

The model (2) looks just like a usual regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{3}$$

except we are using ϕ instead of β for the coefficients, and the index is now t as opposed to i. Let the data be denoted by $(x_i, y_i), i = 1, \dots, m$ (m is the number of data points).

Under the assumption $\epsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$, the likelihood for (3) is:

$$\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right). \tag{4}$$

Let us take a closer look as to how the likelihood (4) is actually derived. We shall examine closely the assumptions that are made in deriving (4). Our main interest is to see whether the same assumptions are still true for AutoRegression.

The data is $(x_i, y_i), i = 1, \dots, m$ (m is the number of data points). The likelihood is the probability density function of the data treated as a function of the parameters $\theta = (\beta_0, \beta_1, \sigma)$ (σ is the standard deviation of the errors):

$$\text{Likelihood for model (3)} = f_{x_1, y_1, x_2, y_2, \dots, x_m, y_m | \theta}(x_1, y_1, \dots, x_m, y_m).$$

Given that the model (3) writes each y_i in terms of x_i, it makes sense to first condition on $x_1, \dots, x_m$:

$$\text{Likelihood for model (3)} = f_{y_1, \dots, y_m | x_1, \dots, x_m, \theta}(y_1, \dots, y_m) f_{x_1, \dots, x_m | \theta}(x_1, \dots, x_m).$$

Now we write $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ for each i to get

$$\begin{aligned}
&\text{Likelihood for model (3)} \\
&= f_{y_1, \dots, y_m | x_1, \dots, x_m, \theta}(y_1, \dots, y_n) f_{x_1, \dots, x_m | \theta}(x_1, \dots, x_m) \\
&= f_{\beta_0 + \beta_1 x_1 + \epsilon_1, \dots, \beta_0 + \beta_1 x_m + \epsilon_m | x_1, \dots, x_m, \theta}(y_1, \dots, y_n) f_{x_1, \dots, x_m | \theta}(x_1, \dots, x_m) \\
&= f_{\epsilon_1, \dots, \epsilon_m | x_1, \dots, x_m, \theta}(y_1 - \beta_0 - \beta_1 x_1, \dots, y_m - \beta_0 - \beta_1 x_m) f_{x_1, \dots, x_m | \theta}(x_1, \dots, x_m).
\end{aligned}$$

To proceed further, we assume that $\epsilon_1, \dots, \epsilon_n$ are independent of $x_1, \dots, x_n$ (given the parameters). This allows us to remove the conditioning on $x_1, \dots, x_n$ in the first term above, leading to:

$$\begin{aligned}
&\text{Likelihood for model (3)} \\
&= f_{\epsilon_1, \dots, \epsilon_m | \theta}(y_1 - \beta_0 - \beta_1 x_1, \dots, y_m - \beta_0 - \beta_1 x_m) f_{x_1, \dots, x_m | \theta}(x_1, \dots, x_m).
\end{aligned}$$

We now use the assumption that $\epsilon_1, \dots, \epsilon_n \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$ to write:

$$\begin{aligned}
&\text{Likelihood for model (3)} \\
&= \left[\prod_{i=1}^{m} f_{\epsilon_i | \theta}(y_i - \beta_0 - \beta_1 x_i)\right] f_{x_1, \dots, x_m | \theta}(x_1, \dots, x_n) \\
&= \left[\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right)\right] f_{x_1, \dots, x_m | \theta}(x_1, \dots, x_m).
\end{aligned}$$

How do we deal with the last term $f_{x_1, \dots, x_m | \theta}(x_1, \dots, x_m)$? We simply assume that this term does not depend on θ so it only becomes a constant (in terms of θ) multiplicative factor in the likelihood that can be omitted leading to:

$$\text{Likelihood for model (3)} \propto \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^m \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - \beta_0 - \beta_1 x_i)^2\right),$$

which coincides with (4).

To summarize, we used the following assumptions to derive the likelihood (4):

1. The model equation (3)
2. Independence of the errors $\epsilon_1, \dots, \epsilon_m$ with the covariates $x_1, \dots, x_m$
3. $\epsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$
4. The density of $x_1, \dots, x_m$ does not depend on $\theta = (\beta_0, \beta_1, \sigma)$.

Using the likelihood, frequentist inference first computes the Maximum Likelihood Estimates by maximizing the likelihood or the log-likelihood. This gives:

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad \text{and} \quad \hat{\sigma}_{\text{MLE}} := \sqrt{\frac{RSS}{n}}$$

where X is the $m \times 2$ matrix with the first column consisting of all ones, and the second column consists of $x_1, \dots, x_m$. Then the goal is to derive the distribution of the MLEs $\hat{\beta}$ and $\hat{\sigma}_{\text{MLE}}$ (given the parameters θ). For this, one again needs to use the above assumptions.

Bayesian inference proceeds by combining the prior $\beta_0, \beta_1, \log \sigma \overset{\text{i.i.d}}{\sim} \text{unif}(-C, C)$ with the likelihood to compute the posterior. We have seen previously that the posterior is given by:

$$\beta \mid \text{data} \sim t_{m-2,2}\left(\hat{\beta}, \hat{\sigma}^2 (X^T X)^{-1}\right) \qquad \text{where } \hat{\sigma} := \sqrt{\frac{RSS}{m-2}}.$$

A posterior for σ can also be derived (we did this in Homework one). One important point about Bayesian inference is that once the likelihood is written, we no longer care about the assumptions that were needed for writing the likelihood. Once the likelihood is written, the subsequent inference (via the posterior distribution) only uses the likelihood. In contrast, frequentist inference uses the assumptions twice: first to write the likelihood in order to compute the MLEs, and then to derive the distribution of the MLEs.

3 Likelihood for AR(1)

Let us now write the likelihood for the AR(1) model (2). Superficially, (2) looks the same as (3) with $i = t$, $x_i = y_{t-1}$ and $\beta_0 = \phi_0$ and $\beta_1 = \phi_1$. However, some of the other regression assumptions listed above do not hold for (2):

1. The density of $x_t = y_{t-1}$ will depend on θ (because $y_{t-1} = \phi_0 + \phi_1 y_{t-2} + \epsilon_{t-1}$ so ϕ_0, ϕ_1 and σ certainly affect y_{t-1}).
2. It is also unclear why $\epsilon_2, \dots, \epsilon_n$ should be independent of $x_2 = y_1, \dots, x_n = y_{n-1}$.

As a result, we cannot use the same principles as in usual linear regression to write the likelihood for AR models. Instead we shall proceed as follows (using a different set of assumptions). As the data is $y_1, \dots, y_n$, the likelihood is given by (below $\theta = (\phi_0, \phi_1, \sigma)$

denotes the set of parameters)

$$
\begin{aligned}
&\text{Likelihood for Model (2)} \\
&= f_{y_1,\dots,y_n|\theta}(y_1,\dots,y_n) \\
&= f_{y_1|\theta}(y_1) f_{y_2|y_1,\theta}(y_2) f_{y_3|y_1,y_2,\theta}(y_3) \cdots f_{y_n|y_1,\dots,y_{n-1},\theta}(y_n) \\
&= f_{y_1|\theta}(y_1) \prod_{t=2}^{n} f_{y_t|y_1,\dots,y_{t-1},\theta}(y_t) \\
&= f_{y_1|\theta}(y_1) \prod_{t=2}^{n} f_{\phi_0+\phi_1 y_{t-1}+\epsilon_t|y_1,\dots,y_{t-1},\theta}(y_t) \\
&= f_{y_1|\theta}(y_1) \prod_{t=2}^{n} f_{\epsilon_t|y_1,\dots,y_{t-1},\theta}(y_t-\phi_0-\phi_1 y_{t-1}).
\end{aligned}
$$

Now we assume that ϵ_t is independent of $y_1,\dots y_{t-1}$. This gives

$$
\begin{aligned}
&\text{Likelihood for Model (2)} \\
&= f_{y_1|\theta}(y_1) \prod_{t=2}^{n} f_{\epsilon_t|y_1,\dots,y_{t-1},\theta}(y_t-\phi_0-\phi_1 y_{t-1}) = f_{y_1|\theta}(y_1) \prod_{t=2}^{n} f_{\epsilon_t}(y_t-\phi_0-\phi_1 y_{t-1}).
\end{aligned}
$$

With $\epsilon_t \sim N(0,\sigma^2)$, we get

$$\text{Likelihood for Model (2)} = f_{y_1|\theta}(y_1) \prod_{t=2}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_t-\phi_0-\phi_1 y_{t-1})^2\right)$$

which is equivalent to:

$$\text{Likelihood for (2)} = f_{y_1|\theta}(y_1) \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n-1} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=2}^{n}(y_t-\phi_0-\phi_1 y_{t-1})^2\right). \tag{5}$$

To sum up, we used the following assumptions to derive the likelihood (5):

1. The model equation (2).
2. Independence of ϵ_t and $y_1,\dots,y_{t-1}$ for each $t = 2,\dots,n-1$.
3. $\epsilon_t \sim N(0,\sigma^2)$.

The likelihood (5) has the term $f_{y_1|\theta}(y_1)$ which we should make explicit before we can compute estimators. Note that the model equation (2) is only for $t = 2,\dots,n$ which means that y_1 never appears on the left side. So it is not possible to compute $f_{y_1|\theta}(y_1)$ using (2). There are two approaches of dealing with $f_{y_1|\theta}(y_1)$.

3.1 Approach One: Assume $f_{y_1|\theta}(y_1)$ does not depend on θ

Here one simply assumes that $f_{y_1|\theta}(y_1)$ does not depend on θ. Then $f_{y_1|\theta}(y_1)$ becomes a constant factor in (5) that can be ignored in proportionality leading to

$$\text{Likelihood for (2)} \propto \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n-1} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=2}^{n}(y_t-\phi_0-\phi_1 y_{t-1})^2\right). \tag{6}$$

This likelihood is identical to the likelihood in usual regression with $x_t = y_{t-1}$ (even though the assumptions that led to the likelihood are different from the case of linear regression).

Because of this, Bayesian inference here (with the prior $\phi_0, \phi_1, \log \sigma \overset{\text{i.i.d}}{\sim} \text{uniform}(-C, C)$ will lead to identical results as in the case of linear regression. In other words, the posterior for (ϕ_0, ϕ_1) will be:

$$(\phi_0, \phi_1) \mid y_1, \dots, y_n \sim t_{n-3,2}\left((\hat{\phi_0}, \hat{\phi_1}), \hat{\sigma}^2 (X^T X)^{-1}\right)$$

where $\hat{\phi}_0$ and $\hat{\phi}_1$ represent the least squares estimates obtained by regressing y_t on y_{t-1}. Note that the residual degrees of freedom (i.e., the degrees of freedom of the t-distribution) equal $n - 3$ because the number of observations in this regression equals $n - 1$ (as $t = 2, \dots, n$) and the number of columns of X equals 2.

Frequentist inference for AutoRegression will be quite different from inference in linear regression. First note that, under the likelihood (6), the MLEs of ϕ_0, ϕ_1, σ will be identical to the MLEs in linear regression (because the likelihood (6) is the same as for linear regression). However, in order to derive the distribution of the MLEs, we now have to use the AR assumptions which are more complicated. This part is usually done by asymptotics i.e., by letting $n \to \infty$ (see for example Shumway and Stoffer [1, Chapter 4]). In this analysis, inference is based on the normal distribution and justified in the large sample limit as $n \to \infty$; in other words, t-distributions no longer arise.

This is one concrete problem where Bayesian inference and frequentist inference differ. Bayesian inference is simpler while frequentist inference is more complicated and uses asymptotic arguments (specifically, Central Limit Theorems and Laws of Large Numbers for dependent random variables).

Usual library functions for AutoRegression (such as the function `AutoReg` in the `statsmodels` library) use the frequentist formulas so they give different results from those obtained by just running the OLS function for regressing y_t on $x_t = y_{t-1}$. For example, they give standard errors and z-scores as opposed to t-scores. Most of the time in practice, the difference between the two kinds of inferences is negligible. However, strictly speaking, they are different.

3.2 Approach Two: Computing $f_{y_1|\theta}(y_1)$ by extending (2) to $t \le 1$

So far we assumed the model equation (2) only for $t = 2, \dots, n$. With this, y_1 never appears on the left hand side in (2) which means that we have not assumed anything about $f_{y_1|\theta}(y_1)$. In order to be able to compute it, a natural idea is to extend the model equation for $t = 1, 0, -1, \dots$. This allows computation of $f_{y_1|\theta}(y_1)$ in the following way. Applying (2) for $t = 1, 0, -1, -2, \dots$ recursively, we get

$$\begin{aligned}
y_1 &= \phi_0 + \phi_1 y_0 + \epsilon_1 \\
&= \phi_0 + \phi_1 \left(\phi_0 + \phi_1 y_{-1} + \epsilon_0\right) + \epsilon_1 \\
&= \phi_0 (1 + \phi_1) + \phi_1^2 y_{-1} + \phi_1 \epsilon_0 + \epsilon_1 \\
&= \phi_0 (1 + \phi_1) + \phi_1^2 \left(\phi_0 + \phi_1 y_{-2} + \epsilon_{-1}\right) + \phi_1 \epsilon_0 + \epsilon_1 \\
&= \phi_0 \left(1 + \phi_1 + \phi_1^2\right) + \phi_1^3 y_{-2} + \phi_1^2 \epsilon_{-1} + \phi \epsilon_0 + \epsilon_1 .
\end{aligned}$$

Continuing this way with using (2) for $t = -2, -3, \dots, -M$ (for some large M), we get

$$\begin{aligned}
y_1 &= \phi_0 \left(1 + \phi_1 + \phi_1^2 + \dots + \phi_1^M\right) + \phi_1^{M+1} y_{-M} + \phi_1^M \epsilon_{-M+1} + \phi_1^{M-1} \epsilon_{-M+2} + \dots + \phi \epsilon_0 + \epsilon_1 \\
&= \phi_0 \sum_{j=0}^{M} \phi_1^j + \phi_1^{M+1} y_{-M} + \sum_{j=0}^{M} \phi_1^j \epsilon_{1-j} .
\end{aligned}$$

This equation is not enough to allow us to deduce $f_{y_1|\theta}(y_1)$ because it involves the unknown quantity y_{-M}. If $|\phi_1| < 1$, then the coefficient ϕ_1^{M+1} in front of y_{-M} is very small. In this case, it might make sense to ignore the term $\phi_1^{M+1} y_{-M}$ when M is large. This allows us to write

$$y_1 \approx \phi_0 \sum_{j=0}^{M} \phi_1^j + \sum_{j=0}^{M} \phi_1^j \epsilon_{1-j} \approx \phi_0 \sum_{j=0}^{\infty} \phi_1^j + \sum_{j=0}^{\infty} \phi_1^j \epsilon_{1-j} = \frac{\phi_0}{1-\phi_1} + \sum_{j=0}^{\infty} \phi_1^j \epsilon_{1-j}.$$

The term $\sum_{j=0}^{\infty} \phi_1^j \epsilon_{1-j}$ is the sum of independent normal random variables, so it is Normal with mean zero (as each ϵ_{1-j} has mean zero) and with variance:

$$\mathrm{var}\left(\sum_{j=0}^{\infty} \phi_1^j \epsilon_{1-j}\right) = \sum_{j=0}^{\infty} \mathrm{var}\left(\phi_1^j \epsilon_{1-j}\right) = \sum_{j=0}^{\infty} \phi_1^{2j} \mathrm{var}(\epsilon_{1-j}) = \sigma^2 \sum_{j=0}^{\infty} \phi_1^{2j} = \frac{\sigma^2}{1-\phi_1^2}.$$

Thus when $|\phi_1| < 1$, we can write

$$y_1 \sim N\left(\frac{\phi_0}{1-\phi_1}, \frac{\sigma^2}{1-\phi_1^2}\right).$$

which gives

$$f_{y_1|\theta}(y_1) = \frac{\sqrt{1-\phi_1^2}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1-\phi_1^2}{2\sigma^2}\left(y_1 - \frac{\phi_0}{1-\phi_1}\right)^2\right).$$

Plugging this in (5), we get

Likelihood for (2)

$$= \frac{\sqrt{1-\phi_1^2}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1-\phi_1^2}{2\sigma^2}\left(y_1 - \frac{\phi_0}{1-\phi_1}\right)^2\right) \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n-1} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=2}^{n}(y_t - \phi_0 - \phi_1 y_{t-1})^2\right). \tag{7}$$

This is a more complicated likelihood compared to (6). This is applicable only when $|\phi_1| < 1$. We shall see later the implications of this assumption.

4 Two AR(1) Models

We therefore have two different versions of AR(1) corresponding to each of the likelihoods (6) and (7).

The first version is:

$$\begin{aligned} & y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t \qquad \text{for } t = 2, \dots, n \\ & \epsilon_t \text{ independent of } y_1, \dots, y_{t-1} \qquad \text{for } t = 2, \dots, n \\ & \epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2) \text{ and } y_1 \text{ is a constant.} \end{aligned}$$

The second version assumes

$$y_1 = \frac{\phi_0}{1-\phi_1} + \sum_{j=0}^{\infty} \phi_1^j \epsilon_{1-j}.$$

Note that this, along with $y_2 = \phi_0 + \phi_1 y_1 + \epsilon_2$ implies that

$$y_2 = \phi_0 + \phi_1 \left(\frac{\phi_0}{1-\phi_1} + \sum_{j=0}^{\infty} \phi_1^j \epsilon_{1-j} \right) + \epsilon_2 = \frac{\phi_0}{1-\phi_1} + \sum_{j=0}^{\infty} \phi_1^j \epsilon_{2-j}.$$

By induction, one can show that

$$y_t = \frac{\phi_0}{1-\phi_1} + \sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j}.$$

So the second version of the AR(1) model can be simply written as:

$$y_t = \frac{\phi_0}{1-\phi_1} + \sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j} \qquad \text{for all } t = \dots, -3, -2, -1, 0, 1, 2, 3, \dots \tag{8}$$

where $\epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$. Note that this equation automatically satisfies: $y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t$ and also that ϵ_t is independent of $y_{t-1}, y_{t-2}, \dots$.

It is important to note that the second model (8) only makes sense when $|\phi_1| < 1$. So this second definition is not valid unless $|\phi_1| < 1$. When $|\phi_1| = 1$, one cannot make sense of the right hand side of (8) (this is becase $\sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j}$ fails to converge when $|\phi_1| \geq 1$).

We shall see in the next lecture that (8) is an example of a stationary model. We will explore in depth the notion of stationarity.

References

[1] Shumway, R. H. and D. S. Stoffer (2010). *Time series analysis and its applications: with R examples* (fourth ed.). Springer Science & Business Media.