Financial Econometrics I

Lecture 2

Fu Ouyang

November 23, 2018

Outline

Linear Time Series Models: Stationarity

Moving Average Processes

Autoregressive Processes

Stationary ARMA Processes

Fitting ARMA Models

Linear Time Series Models

Data obtained from observations collected sequentially over time are called *time series*. The purpose of analyzing time series data:

1. Recover the *data generating process* (DGP) that generates the data.
2. Forecast the future values of a time series using historical data.

In the following couple of lectures, we will study a class of models which depict the linear features (the first two moments and linear dependence) of time series.

In what follows, we use $\{X_t, t = 0, \pm 1, \pm 2, ...\}$ (or for notation simplicity $\{X_t\}$) to represent a generic stochastic process (i.e., a sequence of random variables). But sometimes it is convenient to refer to $\{X_t\}$ itself as a set of observed time series data.

Weak Stationarity

The assumption of stationarity plays a central role in forecasting, which in general refers to certain time invariance properties of the underlying DGP.

Weak Stationarity

$\{X_t\}$ is *weakly stationary* (or *second order stationary* or *covariance stationary*) if $E(X_t^2) < \infty$ and both $E(X_t)$ and $Cov(X_t, X_{t+k})$, for any integer k, do not depend on t.

- $E(X_t)$ is a constant, i.e., $E(X_t) = \mu$.
- $Cov(X_t, X_{t+k})$ is independent of t for all $k = 0, \pm 1, \pm 2, \cdots$.
- $|Cov(X_t, X_{t+k})| < \infty$ by $|E(X_t X_{t+k})|^2 \leq E(X_t^2)E(X_{t+k}^2)$ (recall the Cauchy-Schwarz inequality) and $E(X_t^2) < \infty$.
- $\{X_t\}$ is weakly stationary $\Leftrightarrow$ $\{X_t\}$ has finite and time-invariant first two moments.

Autocovariance Function

The *autocovariance function* (ACVF) is defined as

$$\gamma(k) = Cov(X_t, X_{t+k}) = E[(X_t - \mu)(X_{t+k} - \mu)]$$

for $k = 0, \pm 1, \pm 2, \cdots$. Note that $\gamma(0) = Var(X_t)$ and $\gamma(k) = \gamma(-k)$.

The variance-covariance matrix of the vector $(X_t, ..., X_{t+k})$ is

$$Var(X_t, ..., X_{t+k}) = \begin{pmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \cdots & \gamma(k) \\ \gamma(1) & \gamma(0) & \gamma(1) & \cdots & \gamma(k-1) \\ \gamma(2) & \gamma(1) & \gamma(0) & \cdots & \gamma(k-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma(k-1) & \gamma(k-2) & \gamma(k-3) & \cdots & \gamma(1) \\ \gamma(k) & \gamma(k-1) & \gamma(k-2) & \cdots & \gamma(0) \end{pmatrix}$$

Autocorrelation Function

The *autocorrelation function* (ACF) is defined as

$$\rho(k) = Corr(X_t, X_{t+k}) = \gamma(k)/\gamma(0)$$

for $k = 0, \pm 1, \pm 2, \cdots$. Note that $\rho(0) = 1$ and $\rho(k) = \rho(-k)$.

Sample ACVF and Sample ACF

How to use an observed sample $X_1, ...X_T$ to estimate ACVF and ACF?

$$\widehat{\gamma}(k) = \frac{1}{T} \sum_{t=k+1}^{T} (X_t - \bar{X})(X_{t-k} - \bar{X}), \widehat{\rho}(k) = \widehat{\gamma}(k)/\widehat{\gamma}(0)$$

where $\bar{X} = \sum_{t=1}^{T} X_t/T$. $\widehat{\gamma}(k)$ and $\widehat{\rho}(k)$ are called sample ACVF and sample ACF, respectively.

Note that the estimator $\widehat{\gamma}(k)$ use divisor T instead of $T - k$!

Sample ACVF and Sample ACF

Let $Z_t \equiv X_t - \bar{X}$.

$$\mathbf{Z} = \begin{pmatrix} 0 & 0 & \cdots & 0 & Z_1 & Z_2 & \cdots & Z_{T-1} & Z_T \\ 0 & 0 & \cdots & Z_1 & Z_2 & Z_3 & \cdots & Z_T & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ Z_1 & Z_2 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 0 \end{pmatrix}_{(k+1)\times(k+T)}$$

$$\widehat{Var}(X_t, ..., X_{t+k}) = \frac{1}{T}\mathbf{Z}\mathbf{Z}'$$

Using divisor T ensures that $\widehat{Var}(X_t, ..., X_{t+k})$ is *semi-positive definite*, i.e., for any $(k+1)$-vector v, $v'\widehat{Var}(X_t, ..., X_{t+k})v \geq 0$.

Strong Stationarity

Strong Stationarity

$\{X_t\}$ is said to be *strongly stationary* or *strictly stationary* if the joint distribution of $(X_1, ...X_k)$ is the same as that of $(X_{t+1}, ..., X_{t+k})$ for any $k \geq 1$ and t.

Note that

- Provided $E(X_t^2) < \infty$, strong stationarity $\Rightarrow$ weak stationarity.
- The strong stationarity of $\{X_t\} \Rightarrow$ the strong stationarity of $\{g(X_t)\}$ for any function g.
- The assumption of strong stationarity will be needed in the context of nonlinear prediction.

Outline

Linear Time Series Models: Stationarity

Moving Average Processes

Autoregressive Processes

Stationary ARMA Processes

Fitting ARMA Models

Moving Average (MA) Processes: Definition

Let $\epsilon_t \sim WN(0, \sigma^2)$. For a fixed integer $q \geq 1$, we say $X_t \sim \text{MA}(q)$ if X_t is defined as a moving average of q successive ϵ_t as follows

$$X_t = \mu + \epsilon_t + \sum_{k=1}^{q} a_k \epsilon_{t-k}$$

where $\mu, a_1, ..., a_q$ are constant coefficients.

- μ is the stationary expectation of X_t, $E(X_t) = \mu$.
- $\{\epsilon_t\}$ stands for a sequence of innovations (shocks) to the market in each period.
- $\{a_k\}$ can be thought of as "discount" factors associated with lagged innovations $\{\epsilon_{t-k}\}$.
- All MA(q) processes are (weakly) stationary. (why?)

MA(q) Processes: ACVF and ACF

Recall $\rho(k) = Cov(X_{t+k}, X_t)/Var(X_t) = \gamma(k)/\gamma(0)$. Letting $a_0 \equiv 1$,

$$\gamma(0) = Var(X_t) = E\left[\left(\sum_{l=0}^{q} a_l \epsilon_{t-l}\right)^2\right]$$

$$\gamma(k) = Cov(X_{t+k}, X_t) = E\left[\left(\sum_{l=0}^{q} a_l \epsilon_{t-l}\right)\left(\sum_{l=0}^{q} a_l \epsilon_{t+k-l}\right)\right]$$

By $\epsilon_t \sim WN(0, \sigma^2)$, $E(\epsilon_t \epsilon_s) \neq 0$ if and only if $t = s$. Hence,

$$\gamma(0) = \sigma^2 \sum_{l=0}^{q} a_l^2$$

and $\forall k > q$, $Cov(X_{t+k}, X_t) = 0$, i.e., the ACF of MA(q) process cuts off at q.

MA(q) Processes: ACVF and ACF

For $1 \leq k \leq q$, common WN terms are $\epsilon_{t+k-q}, ..., \epsilon_{t-1}, \epsilon_t$, and so

$$\gamma(k) = \sigma^2(a_q a_{q-k} + ... + a_{k+1}a_1 + a_k a_0)$$

To sum up, we have

$$\rho(k) = \frac{a_q a_{q-|k|} + ... + a_{|k|+1}a_1 + a_{|k|}a_0}{a_0^2 + a_1^2 + ... + a_q^2} \cdot \mathbf{1}[1 \leq |k| \leq q]$$

where $\mathbf{1}[\cdot]$ is an indicator function and the $|\cdot|$ is used because of the symmetry of $\rho(k)$, i.e., $Cov(X_{t+k}, X_t) = Cov(X_{t-k}, X_t)$.

MA(∞) Processes

If we permit the order q of an MA(q) process to increase to infinity, i.e.,

$$X_t = \mu + \sum_{j=0}^{\infty} a_j \epsilon_{t-j}$$

with $\epsilon_t \sim WN(0, \sigma^2)$, we obtain a MA($\infty$) process. MA($\infty$) is well-defined (i.e., $\sum_{j=0}^{\infty} a_j \epsilon_{t-j}$ converges in mean-square) if $\sum_{j=1}^{\infty} a_j^2 < \infty$ as

$$E\left[|\sum_{j=0}^{n} a_j \epsilon_{t-j} - \sum_{j=0}^{m} a_j \epsilon_{t-j}|^2\right] = \sum_{j=m}^{n} \sigma^2 a_j^2 \to 0$$

Using the same derivation for MA(q), we obtain that for a MA(∞) process,

$$\gamma(0) = \sigma^2 \sum_{j=0}^{\infty} a_j^2 < \infty, \gamma(k) = \sigma^2 \sum_{j=0}^{\infty} a_j a_{j+|k|}$$

Outline

Linear Time Series Models: Stationarity

Moving Average Processes

Autoregressive Processes

Stationary ARMA Processes

Fitting ARMA Models

Autoregressive Processes: Definition

For a time series $\{X_t\}$, it is more intuitive to predict X_t using its history,

$$X_t = c + b_1 X_{t-1} + \cdots b_p X_{t-p} + \epsilon_t$$

where $\epsilon_t \sim WN(0, \sigma^2)$ and $c, b_1, \ldots, b_p$ are unknown parameters. We refer to this model as an *autoregressive* (AR) *process* of order p, $X_t \sim \text{AR}(p)$.

Suppose $X_t \sim \text{AR}(p)$ is stationary. Then,

$$\mu \equiv E(X_t) = c + \mu(b_1 + \cdots + b_p) \Rightarrow \mu = \frac{c}{1 - (b_1 + \cdots + b_p)}$$

and so

$$X_t - \mu = b_1(X_{t-1} - \mu) + \cdots + b_p(X_{t-p} - \mu) + \epsilon_t$$

In what follows, we assume X_t is "centralized", i.e., $E(X_t) = 0$ and $c = 0$.

AR(1) Processes

Example: AR(1) Model

$$X_t = bX_{t-1} + \epsilon_t$$

Assuming that $Cov(\epsilon_t, X_{t-k}) = 0, \forall k \geq 1$, $E(X_t^2) = b^2 E(X_{t-1}^2) + \sigma^2$.

"$\Rightarrow$": For weakly stationary AR(1), $E(X_t^2) = E(X_{t-1}^2)$, which then implies $|b| < 1$ and $E(X_t^2) = \sigma^2/(1-b^2)$. Therefore, $|b| < 1$ is a necessary condition for the stationarity of AR(1).

"$\Leftarrow$": By recursive substitution,

$$\begin{aligned} X_t &= bX_{t-1} + \epsilon_t \\ &= b^2 X_{t-2} + \epsilon_t + b\epsilon_{t-1} \\ &= \epsilon_t + b\epsilon_{t-1} + \cdots + b^k \epsilon_{t-k} + b^{k+1} X_{t-k-1} \end{aligned}$$

If $|b| < 1$, $X_t = \sum_{j=0}^{\infty} b^j \epsilon_{t-j}$ (in a mean squared error sense). To see this...

AR(1) Processes

Example: AR(1) Model

$$
\begin{aligned}
E[(X_t - \sum_{j=0}^{\infty} b^j \epsilon_{t-j})^2] &= \lim_{k\to\infty} E[(X_t - \sum_{j=0}^{k} b^j \epsilon_{t-j})^2] \\
&= \lim_{k\to\infty} |b|^{2(k+1)} E(X_{t-k-1}^2)
\end{aligned}
$$

Hence, if $|b| < 1$, $|b|^{2(k+1)} \to 0$ as $k \to \infty$, and $E[(X_t - \sum_{j=0}^{\infty} b^j \epsilon_{t-j})^2] = 0$.

AR(1) process is effectively a MA(∞) process and so weakly stationary.

To sum up, an AR(1) process is weakly stationary if and only if $|b| < 1$.

AR(1) Processes

Example: AR(1) Model

Stationary AR(1) models exhibit the *mean-reversion* property. To see this, consider

$$X_t = c + bX_{t-1} + \epsilon_t$$

with $E(X_t) = \mu$. Then, $|b| < 1$, $c = (1-b)\mu$, and

$$\begin{aligned}\Delta X_t \equiv X_t - X_{t-1} &= c + (b-1)X_{t-1} + \epsilon_t \\ &= \kappa(X_{t-1} - \mu) + \epsilon_t\end{aligned}$$

where $\kappa = b - 1 < 0$. This implies that $E[\Delta X_t | X_{t-1}] < 0$ when $X_{t-1} > \mu$, while $E[\Delta X_t | X_{t-1}] > 0$ when $X_{t-1} < \mu$.

Backshift Operator

The recursive substitution can be compactly represented by the *backshift operator* B, i.e., for $k = \pm 1, \pm 2, ...$

$$B^k X_t = X_{t-k}$$

Then an AR(1) model can be written as $(1 - bB)X_t = \epsilon_t$. Recall the infinite series expansion of $(1 - bx)^{-1}$, we have

$$(1 - bx)^{-1} = \sum_{j=0}^{\infty} b^j x^j$$

as $(1 - bx)(1 + bx + b^2x^2 + \cdots) = 1$. An analogous definition of $(1 - bB)^{-1}$ gives the MA(∞) representation of the AR(1) process

$$X_t = (1 - bB)^{-1}\epsilon_t = \sum_{j=0}^{\infty} b^j \epsilon_{t-j}$$

Backshift Operator

The backshift operator B is useful in handling general AR(p) process:

$$X_t = b_1 X_{t-1} + \cdots + b_p X_{t-p} + \epsilon_t$$

which can be written as $b(B)X_t = \epsilon_t$ with $b(x) \equiv 1 - b_1 x - \cdots - b_p x^p$.

Let $\alpha_1^{-1}, ..., \alpha_p^{-1}$ be roots of $b(x) = 0$, i.e.,

$$b(x) = \prod_{j=1}^{p} (1 - \alpha_j x)$$

There is a sequence $\{a_k\}$ with each a_k determined by $\alpha_1, ..., \alpha_p$ such that

$$b(x)^{-1} = \prod_{j=1}^{p} (1 - \alpha_j x)^{-1} = \prod_{j=1}^{p} \left(\sum_{l=0}^{\infty} \alpha_j^l x^l \right) = 1 + \sum_{k=1}^{\infty} a_k x^k$$

The MA(∞) Representation of AR(p) Processes

From the derivation above, we know for $k = 1, 2, ...,$

$$|a_k| = O\left(\max_{1 \le j \le p} |\alpha_j|^k\right)$$

If $|\alpha_j| < 1$ for all $1 \le j \le p$, then the AR(p) process can be written as

$$X_t = \epsilon_t + \sum_{k=1}^{\infty} a_k \epsilon_{t-k}$$

with $\sum_{k=1}^{\infty} a_k^2 < \infty$ (since $p < \infty$), i.e., $X_t \sim$ MA(∞) and stationary.

The MA(∞) representation for a stationary AR(p) process $\{X_t\}$ indicates that it is a *causal process*, i.e., X_t only depends on $\{\epsilon_t, \epsilon_{t-1}, ...\}$, and is uncorrelated with any future innovations.

The MA(∞) Representation of AR(p) Processes

Recall that for MA(∞) process

$$\gamma(0) = \sigma^2 \sum_{j=0}^{\infty} a_j^2, \gamma(k) = \sigma^2 \sum_{j=0}^{\infty} a_j a_{j+|k|}$$

and so for the AR(p) process, $\rho(k) = O(\max_{1 \le j \le p} |\alpha_j|^k) \to 0$ as $k \to \infty$, which means it only suitable for modeling *short memory* data.

AR(p) Model

1. An AR(p) process is stationary if the p roots of the characteristic equation $1 - b_1 x - \cdots - b_p x^p = 0$ are outside the unit cycle.
2. The ACF of a stationary AR(p) process decays at an exponential rate, i.e., $\rho(k) = O(\alpha^k)$ for some $\alpha \in (0, 1)$.

Yule-Walker Equation

Using the MA(∞) representation to compute $\gamma(k)$ and $\rho(k)$ of the AR(p) model below is cumbersome

$$X_t = b_1 X_{t-1} + \cdots + b_p X_{t-p} + \epsilon_t$$

An alternatively way is to use the *Yule-Walker equation*: For $k \geq 1$, we have

$$\gamma(k) = b_1 \gamma(k-1) + \cdots + b_p \gamma(k-p)$$

which by $\gamma(-k) = \gamma(k)$ yields

$$\begin{pmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(p) \end{pmatrix} = \begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(0) \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}$$

ACVF and ACF

For $k = 0$, we have

$$\gamma(0) = b_1\gamma(1) + \cdots + b_p\gamma(p) + \sigma^2$$

Putting all equations together, we have

$$\begin{pmatrix} \gamma(0) - \sigma^2 \\ \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(p) \end{pmatrix} = \begin{pmatrix} \gamma(1) & \gamma(2) & \cdots & \gamma(p) \\ \gamma(0) & \gamma(1) & \cdots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(0) \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}$$

i.e., $p + 1$ linear equations to solve $p + 1$ unknowns $\gamma(0), \gamma(1), ..., \gamma(p)$.

$\forall k > p$, $\gamma(k)$ can be obtained recursively using the Yule-Walker equation.

Partial Autocorrelation Function

The *partial autocorrelation function* (PACF) at lag k, denoted by $\pi(k)$, is the conditional correlation between X_1 and X_{1+k} given all the intermediate variables $X_2, ..., X_k$. More concretely, let

$$(b_{k1}, ..., b_{kk}) \equiv \arg \min_{\beta_1,...,\beta_k} E\left[(X_{1+k} - \beta_1 X_k - \cdots - \beta_k X_1)^2\right]$$

and then $\pi(k) \equiv b_{kk}$. PACF plays the same role as ACF for MA processes.

For a stationary AR(p) process, the PACF cuts off at p, i.e., $\pi(k) = 0$ for all $k > p$. To see this, recall that

$$E(X_{1+k}|X_1, ..., X_k) = \arg \min_{g \in \mathcal{G}} E[(X_{1+k} - g(X_1, ..., X_k))^2]$$

and by definition of AR(p),

Partial Autocorrelation Function

$$E(X_{1+k}|X_1, ..., X_k) = b_1 X_k + \cdots + b_p X_{k-p+1} + 0 \cdot X_{k-p} + \cdots + 0 \cdot X_1$$

Hence $b_{k1} = b_1, ..., b_{kp} = b_p, b_{k,p+1} = 0, ..., b_{kk} = 0$.

The sample PACF at lag k, denoted by $\widehat{\pi}(k)$, is the sample analogue of $\pi(k)$, i.e.,

$$(\widehat{b}_{k1}, ..., \widehat{b}_{kk}) = \arg \min_{\beta_1,...,\beta_k} \sum_{t=1+k}^{T} (X_t - \beta_1 X_{t-1} - \cdots - \beta_k X_{t-k})^2$$

and $\widehat{\pi}(k) \equiv \widehat{b}_{kk}$, one can obtain $\widehat{\pi}(k)$ by running a least square estimation.

The sample PACF of a stationary AR(p) process does *not* necessarily cuts off at p. The sample PACF will be used for model selection (next lecture).

Outline

Linear Time Series Models: Stationarity

Moving Average Processes

Autoregressive Processes

Stationary ARMA Processes

Fitting ARMA Models

ARMA Processes: Definition

A general *autoregressive and moving average* (ARMA) *model* with the order (p, q) is the combination of a AR(p) and a MA(q) process:

$$X_t = b_1 X_{t-1} + \cdots b_p X_{t-p} + \epsilon_t + a_1 \epsilon_{t-1} + \cdots + a_q \epsilon_{t-q}$$

where $\epsilon_t \sim WN(0, \sigma^2)$ and $(b_1, ..., b_p, a_1, ..., a_q)$ are unknown parameters.

Let

$$\begin{aligned} a(x) &= 1 + a_1 x + \cdots + a_q x^q \\ b(x) &= 1 - b_1 x - \cdots - b_p x^p \end{aligned}$$

Then an ARMA(p,q) model can be compactly represented as

$$b(B) X_t = a(B) \epsilon_t$$

provided that $a(x) = 0$ and $b(x) = 0$ do not have common roots.

ARMA Processes: Properties

Stationarity of ARMA(p,q)

When the p roots of $b(x) = 0$ are all outside of the unit cycle, the ARMA(p,q) process $\{X_t\}$ is stationary and has an MA(∞) representation

$$X_t = b(B)^{-1}a(B)\epsilon_t \sim \text{MA}(\infty)$$

Similar to stationary AR(p) processes, a stationary ARMA(p,q) process

1. $\{X_t\}$ is a causal process, i.e., X_t only depends on $\{\epsilon_t, \epsilon_{t-1}, ...\}$.
2. $\{X_t\}$ has short memory, i.e., $\rho(k) = O(\theta^k)$ as $k \to \infty$ for some $|\theta| < 1$.

Yule-Walker Equation

For all $k > q$ (why?),

$$\gamma(k) = b_1\gamma(k-1) + \cdots + b_p\gamma(k-p)$$

Invertibility: Definition

If an MA(q) process

$$X_t = a(B)\epsilon_t$$

where $a(x) = 1 + a_1 x + \cdots + a_q x^q$ can be written as an AR(∞) process, then it is *invertible*, i.e., the innovations $\epsilon_t, \epsilon_{t-1}, ...$ can be recovered from the observed $X_t, X_{t-1}, ...$

With similar derivation as for the MA(∞) representation of an AR(p) process, we can show that $X_t \sim$ AR(∞) (again in a mean-squared error sense) if the q roots of $a(x)$ are outside the unit cycle.

In practice, the above invertibility condition is imposed to the MA(q) process for the identification of $(a_1, ..., a_q)$ in terms of its ACF.

The following example shows the necessity of doing so.

Invertibility: Example

Consider two MA(1) models with $|a| < 1$,

$$X_t = \epsilon_t + a\epsilon_{t-1}, \epsilon_t \sim WN(0, \sigma^2)$$
$$Y_t = e_t + a^{-1}e_{t-1}, e_t \sim WN(0, a^2\sigma^2)$$

It is easy to show that $\{X_t\}$ and $\{Y_t\}$ share the same ACF, and so they are *not* distinguishable in terms of the ACF. However, $\{Y_t\}$ is not invertible. In fact, by recursive substitution, we have

$$\begin{aligned} Y_t &= a^{-1}e_{t-1} - \sum_{j=1}^{k} (-a)^j Y_{t+j} + (-a)^k e_{t+k} \\ &\overset{m.s.}{\to} a^{-1}e_{t-1} - \sum_{j=1}^{\infty} (-a)^j Y_{t+j} \end{aligned}$$

as $k \to \infty$, so $\{Y_t\}$ is "invertible in the future", not useful for forecasting.

Outline

Linear Time Series Models: Stationarity

Moving Average Processes

Autoregressive Processes

Stationary ARMA Processes

Fitting ARMA Models

Least Square Estimation for AR(p) Models

Consider the AR(p) model

$$X_t = b_0 + b_1 X_{t-1} + \cdots + b_p X_{t-p} + \epsilon_t, \epsilon_t \sim WN(0, \sigma^2)$$

With observations $\{X_t\}_{t=1}^T$, we can estimate parameters $b \equiv (b_0, b_1, \ldots, b_p)$ via a linear regression:

$$\widehat{b} \equiv (\widehat{b}_0, \widehat{b}_1, ..., \widehat{b}_p) = \arg \min_{b \in \mathbb{R}^{p+1}} \sum_{t=p+1}^{T} (X_t - b_0 - b_1 X_{t-1} - \cdots - b_p X_{t-p})^2$$

which is called the *least square estimator* (LSE) for b.

Note that both $\widehat{b}$ and $\widehat{Var}(\widehat{b})$ have explicit expressions. Hypothesis tests can be conducted easily.

Least Square Estimation for AR(p) Models

Once $\widehat{b}$ is obtained, we can compute the LSE for σ^2 by

$$\widehat{\sigma}^2 = \frac{1}{T-2p-1} \sum_{t=p+1}^{T} \left(X_t - \widehat{b}_0 - \widehat{b}_1 X_{t-1} - \cdots - \widehat{b}_p X_{t-p} \right)^2$$

where the divider is $T-2p-1$ because the effective sample size is $T-p$ and the number of parameters is $p+1$.

Least Square Estimation for ARMA(p, q) Model

Consider the ARMA(p, q) model

$$X_t = b_0 + b_1 X_{t-1} + \cdots + b_p X_{t-p} + \epsilon_t + a_1 \epsilon_{t-1} + \cdots + a_q \epsilon_{t-q}$$

where $\epsilon_t \sim WN(0, \sigma^2)$ and ϵ_{p+1-k} is assumed to be 0 for all $1 \leq k \leq q$.

Let $a \equiv (a_1, ..., a_q)$ and $b \equiv (b_0, b_1, \ldots, b_p)$. We can compute the LSE for (a, b) using the iterative algorithm below:

Iterative Linear Approximation

(1) Start from initial values of $\epsilon_{p+1-q} = 0, ..., \epsilon_p = 0$. For $t \geq p + 1$, Define

$$\epsilon_t(a, b) = X_t - b_0 - \sum_{j=1}^{p} b_j X_{t-j} - \sum_{l=1}^{q} a_l \epsilon_{t-l}(a, b)$$

Least Square Estimation for ARMA(p, q) Model

Iterative Linear Approximation

(2) Compute the following iterative estimator with some starting values $(a, b) = (\bar{a}, \bar{b})$:

$$\begin{aligned}
& (\widehat{a}_k, \widehat{b}_k) \\
= \quad & \arg\min_{a,b} \sum_{t=p+1}^{T} [\epsilon_t(\widehat{a}_{k-1}, \widehat{b}_{k-1})]^2 \\
= \quad & \arg\min_{a,b} \sum_{t=p+1}^{T} [X_t - b_0 - \sum_{j=1}^{p} b_j X_{t-j} - \sum_{l=1}^{q} a_l \epsilon_{t-l}(\widehat{a}_{k-1}, \widehat{b}_{k-1})]^2
\end{aligned}$$

for $k = 1, 2, ...$, where $\epsilon_t(\cdot)$ is defined in (1).

(3) Repeat (2) till $(\widehat{a}_k, \widehat{b}_k) \approx (\widehat{a}_{k-1}, \widehat{b}_{k-1})$.

Gaussian Maximum Likelihood Estimation

If we assume $\epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$, then $\theta \equiv (a, b, \sigma^2)$ can be more efficiently estimated using the Gaussian maximum likelihood estimation, and the resulting estimator is called the *maximum likelihood estimator* (MLE).

Consider a general ARMA(p, q) model:

$$X_t = b_0 + b_1 X_{t-1} + \cdots + b_p X_{t-p} + \epsilon_t + a_1 \epsilon_{t-1} + \cdots + a_q \epsilon_{t-q}$$

Let $X^t \equiv (X_1, ..., X_t)$ for all $1 \leq t \leq T$. We define the likelihood function for the model as

$$\begin{aligned} L(\theta) &\equiv f(X^T; \theta) \\ &= f(X_T; \theta | X^{T-1}) \times f(X_{T-1}; \theta | X^{T-2}) \times \cdots \times f(X_{p+1}; \theta | X^p) \times f(X^p) \end{aligned}$$

Taking log and dropping the "constant" term $\log f(X^p)$ leads to the log-likelihood function for the model.

Gaussian Maximum Likelihood Estimation

$$l(\theta) \equiv \log L(\theta) - \log f(X^p) = \sum_{t=p+1}^{T} \log f(X_t; \theta | X^{t-1})$$

Given $\epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$, $f(X_t; \theta | X^{t-1})$ has explicit expression. Then we can employ a Newton-Raphson algorithm to compute the Gaussian MLE via the following optimization procedure:

$$\widehat{\theta} = \arg\max_{\theta} l(\theta) = \arg\max_{\theta} \sum_{t=p+1}^{T} \log f(X_t; \theta | X^{t-1})$$

The covariance matrix of $\widehat{\theta}$ can be consistently estimated as

$$\widehat{Var}(\widehat{\theta}) = -\ddot{l}(\widehat{\theta})^{-1}$$

where $\ddot{l}(\cdot)$ is the Hessian matrix of the log-likelihood function $l(\cdot)$.

Gaussian Maximum Likelihood Estimation

Under mild conditions,

- Gaussian MLE is consistent and asymptotically normal.
- Gaussian MLE is often used when ϵ_t is not normal. The resulting estimator is called *quasi-MLE*.
- Gaussian MLE is more efficient than LSE as it makes (and make uses of) stronger assumption.
- Statistical inference on $g(\widehat{\theta})$ can be done using the *Delta method* when $g(\cdot)$ is differentiable.

$$\widehat{Var}(g(\widehat{\theta})) \approx -\dot{g}(\widehat{\theta})' \ddot{l}(\widehat{\theta})^{-1} \dot{g}(\widehat{\theta})$$

  where $\dot{g}(\cdot)$ is the gradient or Jacobian of $g(\cdot)$.

Now let's apply LSE and MLE to an empirical example [R markdown file].