Modeling Cycles: MA, AR and ARMA Models

Ka-fu WONG
University of Hong Kong

January 2, 2020

Contents

1. A recap of the unobserved components model

- According to the unobserved components model of a time series, the series y_t, is made up of the sum of three independent components
  - a time trend component
  - a seasonal component
  - an irregular or cyclical component.

$$y_t = \text{time trend} + \text{seasonal} + \text{cyclical} = T_t + S_t + C_t$$

2. The Starting Point

Let y_t denote the cyclical component of the time series.

We will assume, unless noted otherwise, that y_t is a

zero-mean covariance stationary process.

Recall that part of this assumption is that the time series originated infinitely far back into the past and will continue infinitely far into the future, with the same mean, variance, and autocovariance structure.

The starting point for introducing the various kinds of econometric models that are available to describe stationary processes is the *Wold* Representation Theorem (or, simply, Wold's theorem).

3. Wold's theorem

According to Wold's theorem, any zero-mean covariance stationary process y_t can be written as:

$$y_t = \sum_{i=0}^{\infty} b_i \epsilon_{t-i} = b_0 \epsilon_t + b_1 \epsilon_{t-1} + b_2 \epsilon_{t-2} + \dots$$

where (i) the ϵ's are $WN(0, \sigma^2)$, (ii) $b_0 = 1$, and (iii) $\sum_{i=0}^{\infty} b_i^2 < \infty$.

In other words, each y_t can be expressed in terms of a single *linear* function of current and (possibly an infinite number of) *past* drawings of the *white noise* process, ϵ_t.

Condition (i) the ϵ's are $WN(0, \sigma^2)$
implies that $E(\epsilon_t) = 0$ and $Cov(\epsilon_t, \epsilon_{t-j}) = 0$ for all $j \neq 0$.

Condition (ii) $b_0 = 1$
is better understood as a *normalization*. To see this, suppose $b_0 \neq 1$, we can always define $\eta_t = b_0\epsilon_t$, $b'_i = b_i/b_0$, and we have

$$y_t = \sum_{i=0}^{\infty} b'_i \eta_{t-i}$$

where η's are $WN(0, b_0^2\sigma^2)$ and $b'_0 = 1$.
Example:

$$y_t = 5\epsilon_t + 2\epsilon_{t-1}, \qquad \epsilon_t \sim WN(0, \sigma^2)$$

$$y_t = 5\epsilon_t + \frac{2}{5} \times 5\epsilon_{t-1}, \qquad \epsilon_t \sim WN(0, \sigma^2)$$

$$y_t = \eta_t + \frac{2}{5} \times \eta_{t-1}, \qquad \eta_t = 5\epsilon_t \sim WN(0, 25\sigma^2)$$

Condition (iii) $\sum_{i=0}^{\infty} b_i^2 < \infty$
guarantees that the variance of y_t to be finite, i.e.,

$$Var(y_t) = \sum_{i=0}^{\infty} b_i^2 Var(\epsilon_{t-i}) = \left(\sum_{i=0}^{\infty} b_i^2\right) \sigma^2 < \infty,$$

a condition for stationary process.

The implication is that if y_t depends on an infinite number of past ϵ's, the absolute weights on these ϵ's, i.e., the b_i's, must be *going to zero as i gets large*. In fact, the convergence of b_i sequence to zero has to be fast enough to satisfy the condition of $\sum_{i=0}^{\infty} b_i^2 < \infty$.

Reason:

If infinitely many b_i's are larger than zero in absolute values, we must be able to find a strictly positive constant c such that $b_i^2 > c$ for all i, and

$$\sum_{i=0}^{\infty} b_i^2 > \sum_{i=0}^{\infty} c = \infty$$

If the b_i's are going to zero as i gets large, we would not be able to find such a strictly positive constant "c".

4. Innovation

ϵ_t is called the innovation in y_t because ϵ_t is the part of y_t that is not predictable from the past history of y_t, i.e., $E(\epsilon_t \mid y_{t-1}, y_{t-2}, ...) = E(\epsilon_t) = 0$. Note that because of the Wold representation, we can use the history of y_t to uncover ϵ_t.

knowing the past history of y_t

$(y_{t-1}, y_{t-2}, ...)$

$$\Updownarrow$$

knowing the past history of ϵ_t.

$(\epsilon_{t-1}, \epsilon_{t-2}, ...)$

Wold's Theorem says

$$y_t = \sum_{i=0}^{\infty} b_i \epsilon_{t-i} = b_0 \epsilon_t + b_1 \epsilon_{t-1} + b_2 \epsilon_{t-2} + \ldots$$

and thus it is obvious that

knowing the past history of ϵ_t.
$(\epsilon_{t-1}, \epsilon_{t-2}, \ldots)$
$\Downarrow$
knowing the past history of y_t
$(y_{t-1}, y_{t-2}, \ldots)$

If we can write

$$\epsilon_t = \sum_{i=0}^{\infty} d_i y_{t-i} = d_0 y_t + d_1 y_{t-1} + d_2 y_{t-2} + \ldots$$

knowing the past history of y_t

$(y_{t-1}, y_{t-2}, \ldots)$

$\Downarrow$

knowing the past history of ϵ_t:

$(\epsilon_{t-1}, \epsilon_{t-2}, \ldots)$

If a process

$$y_t = \sum_{i=0}^{\infty} b_i \epsilon_{t-i} = b_0 \epsilon_t + b_1 \epsilon_{t-1} + b_2 \epsilon_{t-2} + \ldots$$

can be inverted to

$$\epsilon_t = \sum_{i=0}^{\infty} d_i y_{t-i} = d_0 y_t + d_1 y_{t-1} + d_2 y_{t-2} + \ldots$$

the process is called *invertible*. (More on "invertibility" later.)

For those processes that are invertible

knowing the past history of y_t
$(y_{t-1}, y_{t-2}, \ldots)$

$$\Updownarrow$$

knowing the past history of ϵ_t.
$(\epsilon_{t-1}, \epsilon_{t-2}, \ldots)$

Remark: Most of the time series analysis assumes invertibility. When such invertibility assumption is invalid, we will need to make adjustment to our analysis, and hence results.

It follows that the forecast (conditional expectation) of y_t given the infinite history of y_t (i.e., y_{t-1}, y_{t-2},...)

$$\begin{aligned}
E(y_t \mid y_{t-1}, y_{t-2}, ...) &= E(y_t \mid \epsilon_{t-1}, \epsilon_{t-2}, ...) \\
&= E(\epsilon_t + b_1\epsilon_{t-1} + b_2\epsilon_{t-2} + ... \mid \epsilon_{t-1}, \epsilon_{t-2}, ...) \\
&= E(\epsilon_t \mid \epsilon_{t-1}, \epsilon_{t-2}, ...) \\
&\quad + E(b_1\epsilon_{t-1} + b_2\epsilon_{t-2} + ... \mid \epsilon_{t-1}, \epsilon_{t-2}, ...) \\
&= 0 + b_1\epsilon_{t-1} + b_2\epsilon_{t-2} + ... \\
&= b_1\epsilon_{t-1} + b_2\epsilon_{t-2} + ...
\end{aligned}$$

and the *one*-step ahead forecast error is

$$y_t - E(y_t \mid y_{t-1}, y_{t-2}, ...) = (\epsilon_t + b_1\epsilon_{t-1} + b_2\epsilon_{t-2} + ...) - (b_1\epsilon_{t-1} + b_2\epsilon_{t-2} + ...) = \epsilon_t$$

Thus, according to the Wold theorem, each covariance stationary y_t can be expressed as a weighted average of current and past innovations (or, 1-step ahead forecast errors), i.e., an infinite-order *moving average* process.

5. Problem: infinite parameters, finite observations

It is impossible to estimate the *infinite number of parameters* in Wold representation using finite observations. Estimation is possible only if the number of parameters can be reduced substantially. Two possibilities:

- only a *small* number of the parameters in Wold representation are *non-zero*.
- the infinite number of the parameters in Wold representation are related via some *simple function* that depends on small number of parameters.

It turns out that the Wold representation can usually be well-approximated by a variety of models that can be expressed in terms of a very small number of parameters.

- the moving-average (MA) models,
- the autoregressive (AR) models, and
- the autoregressive moving-average (ARMA) models.

The procedure we will follow is to describe each of these three types of models and, especially, the shapes of the autocorrelation and partial autocorrelations that they imply.

Then, the game will be to use the sample autocorrelation/partial autocorrelation functions of the data to "*guess*" *which kind of model likely generated the data*. We estimate that model and see if it provides a good fit to the data. If yes, we proceed to the forecasting step using this estimated model of the cyclical component. If not, we guess again . . .

6. Digression – The Lag Operator

The lag operator, L, is a simple but powerful device that is routinely used in applied and theoretical time series analysis, including forecasting. The operation L applied to y_t returns y_{t-1}, which is y_t "lagged" one period. We write

$$Ly_t = y_{t-1}$$

Similarly, we have

$$L^2 y_t = L(Ly_t) = y_{t-2}$$

i.e., the operation L applied twice to y_t returns y_{t-2}, y_t lagged two periods. More generally,

$$L^s y_t = y_{t-s}$$

for any integer s.

Note that L is like *a mapping or function*. In an ordinary function such as $f(x) = x^2$, if you give me $x = 1$, I give you $f(1) = 1$; if you give me $x = 2$, I give you $f(2) = 4$, etc. The L is function that returns the lag of y_t. That is, if you give me y_{10}, I will give $Ly_{10} = y_9$, etc.

$$
\begin{array}{ccccc}
Ly_t & Ly_{t-1} & Ly_{t-2} & Ly_{t-3} & Ly_{t-4} \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
y_{t-1} & y_{t-2} & y_{t-3} & y_{t-4} & y_{t-5}
\end{array}
$$

$$
\begin{array}{ccccc}
Ly_t & L^2y_t & L^3y_t & L^4y_t & L^5y_t \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
y_{t-1} & y_{t-2} & y_{t-3} & y_{t-4} & y_{t-5}
\end{array}
$$

Consider the application of the following *polynomial in the lag operator* to y_t:

$$(b_0 + b_1 L + b_2 L^2 + ... + b_s L^s)y_t = b_0 y_t + b_1 y_{t-1} + b_2 y_{t-2} + ... + b_s y_{t-s}$$

where $b_0, b_1, ..., b_s$ are real numbers.

$$\begin{array}{ccccccccccc} b_0 y_t & + & b_1 L y_t & + & b_2 L^2 y_t & + & b_3 L^3 y_t & + & ... & + & b_s L^s y_t \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\ b_0 y_t & + & b_1 y_{t-1} & + & b_2 y_{t-2} & + & b_3 y_{t-3} & + & ... & + & b_s y_{t-s} \end{array}$$

We sometimes write this in *shorthands* as $B(L)y_t$, where

$$B(L) = b_0 + b_1 L + b_2 L^2 + ... + b_s L^s$$

Thus, we can write the Wold representation of y_t as $B(L)\epsilon_t$ where $B(L)$ is the infinite order polynomial in L: $B(L) = 1 + b_1 L + b_2 L^2 + ...$

$$\begin{aligned}
y_t &= B(L)\epsilon_t \\
&= (1 + b_1 L + b_2 L^2 + b_3 L^3 + ...)\epsilon_t \\
&= \epsilon_t + b_1 L \epsilon_t + b_2 L^2 \epsilon_t + b_3 L^3 \epsilon_t + ... \\
&= \epsilon_t + b_1 \epsilon_{t-1} + b_2 \epsilon_{t-2} + b_3 \epsilon_{t-3} + ...
\end{aligned}$$

Example:
Suppose $y_t = by_{t-1} + \epsilon_t$, we can write

$$
\begin{aligned}
y_t &= by_{t-1} + \epsilon_t \\
y_t - by_{t-1} &= \epsilon_t \\
y_t - bLy_t &= \epsilon_t \\
(1 - bL)y_t &= \epsilon_t \\
B(L)y_t &= \epsilon_t
\end{aligned}
$$

where $B(L) = 1 - bL$.

Under suitable conditions, a polynomial of lag operators is *invertible*. For instance, when $\mid b \mid < 1$, we have

$$(1 - bL)^{-1} = \frac{1}{1 - bL} = 1 + bL + b^2L^2 + b^3L^3 + \dots$$

Recall the infinite geometric sum

$$(1 - 0.5)^{-1} = \frac{1}{1 - 0.5} = 0.5^0 + 0.5^1 + 0.5^2 + 0.5^3 + \dots = 2$$

and

$$(1 - 0.8)^{-1} = \frac{1}{1 - 0.8} = 0.8^0 + 0.8^1 + 0.8^2 + 0.8^3 + \dots = 5$$

When $\mid b_1 \mid < 1$ and $\mid b_2 \mid < 1$, we have

$$
\begin{aligned}
(1 - b_1 L)^{-1}(1 - b_2 L)^{-1} &= \left(\frac{1}{1 - b_1 L}\right)\left(\frac{1}{1 - b_2 L}\right) \\
&= (1 + b_1 L + b_1^2 L^2 + b_1^3 L^3 + \ldots)\times \\
&\quad\ (1 + b_2 L + b_2^2 L^2 + b_2^3 L^3 + \ldots) \\
&= 1 + (b_1 + b_2)L + (b_1^2 + b_2^2 + b_1 b_2)L^2 + \ldots.
\end{aligned}
$$

Recall the product of two infinite geometric sums

$$
\begin{aligned}
(1 - 0.5)^{-1}(1 - 0.8)^{-1} &= \left(\frac{1}{1 - 0.5}\right)\left(\frac{1}{1 - 0.8}\right) \\
&= \left(0.5^0 + 0.5^1 + 0.5^2 + \ldots\right)\left(0.8^0 + 0.8^1 + 0.8^2 + \ldots\right)
\end{aligned}
$$

7. Moving average models

Moving average models assume y_t can be represented by *a weighted sum of recent innovations only*. That is,

$$y_t = \epsilon_t + b_1\epsilon_{t-1} + b_2\epsilon_{t-2} + \ldots + b_q\epsilon_{t-q}$$

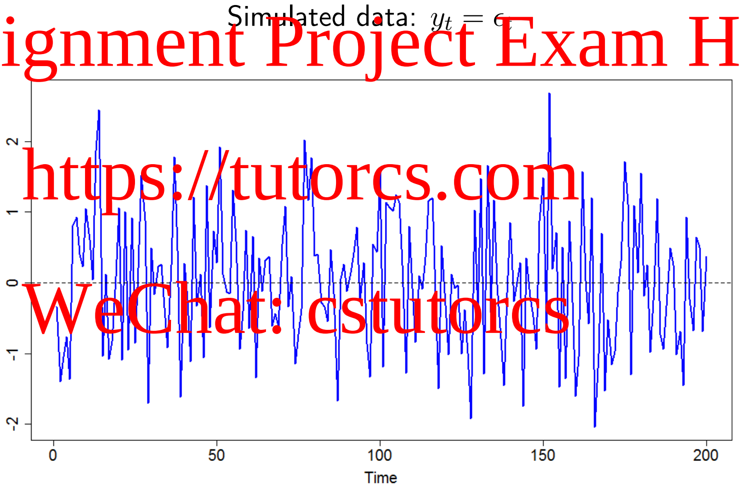
When y_t is assumed to be a weighted sum of of recent q innovations as shown above, y_t is said to follow a MA(q) process. "*q*" is known as the *order* of the moving average process.

MA(0): $y_t = \epsilon_t$

MA(1): $y_t = \epsilon_t + b_1\epsilon_{t-1}$

MA(2): $y_t = \epsilon_t + b_1\epsilon_{t-1} + b_2\epsilon_{t-2}$

7.1. **MA(0),** $y_t = \epsilon_t$.



The simplest case is the case without MA terms, i.e., when y_t is a white noise process.

$$y_t = \epsilon_t$$

where the ϵ's are $WN(0, \sigma^2)$. We compute several quantities:

(1) Unconditional expectation of y_t, i.e., $E(y_t)$. Since $y_t = \epsilon_t$,

$$E(y_t) = E(\epsilon_t) = 0$$

Note that shifting the time forward or backward does not change the unconditional expectation. Using the same reasoning,

$$E(y_{t+j}) = E(\epsilon_{t+j}) = 0$$

for all j.

(2) Expectation of y_{t+1} conditional on the history of y_t, i.e., $E(y_{t+1} \mid y_t, y_{t-1}, \ldots)$.
Since $y_t = \epsilon_t$,

$$E(y_{t+1} \mid y_t, y_{t-1}, \ldots) = E(y_{t+1} \mid \epsilon_t, \epsilon_{t-1}, \ldots)$$

and

$$E(y_{t+1} \mid \epsilon_t, \epsilon_{t-1}, \ldots) = E(\epsilon_{t+1} \mid \epsilon_t, \epsilon_{t-1}, \ldots)$$

Since the ϵ's are $WN(0, \sigma^2)$,

$$E(\epsilon_{t+1} \mid \epsilon_t, \epsilon_{t-1}, \ldots) = E(\epsilon_{t+1}) = 0$$

Note that shifting the time forward or backward does not change the conditional expectation. That is

$$\begin{aligned} E(y_{t+j} \mid y_{t+j-1}, y_{t+j-2}, \ldots) &= E(\epsilon_{t+j} \mid \epsilon_{t+j-1}, \epsilon_{t+j-2}, \ldots) \\ &= E(\epsilon_{t+j}) = 0 \end{aligned}$$

for all j.

Using the same reasoning,

$$\begin{aligned} E(y_{t+h} \mid y_t, y_{t-1}, \ldots) &= E(\epsilon_{t+h} \mid \epsilon_t, \epsilon_{t-1}, \ldots) \\ &= E(\epsilon_{t+h}) \\ &= 0 \end{aligned}$$

for all h.

(3) Unconditional variance of y_t, i.e., $Var(y_t) \equiv E[y_t - E(y_t)]^2$. It should be obvious that

$$Var(y_t) = Var(\epsilon_t) = \sigma^2$$

Note that shifting the time forward or backward does not change the unconditional variance.

$$Var(y_{t+j}) = Var(\epsilon_{t+j}) = \sigma^2$$

(4) Covariance of y_t and y_{t-j}, i.e.

$$Cov(y_t, y_{t-j}) \equiv E[y_t - E(y_t)][y_{t-j} - E(y_{t-j})].$$

Note that the *white noise property* of ϵ_t implies

$$Cov(y_t, y_{t-j}) = Cov(\epsilon_t, \epsilon_{t-j}) = 0$$

for all $j \neq 0$.

(5) Autocorrelation coefficient of y_t, i.e., $\rho(j) \equiv Cov(y_t, y_{t-j})/Var(y_t)$. It is obvious that

$$\rho(j) = \begin{cases} 1 & \text{for } j = 0 \\ 0 & \text{for } j \neq 0 \end{cases}$$

(6) Partial autocorrelation coefficients of y_t at displacement j, i.e., $\beta_j^{(j)}$ in

$$y_t = \beta_0^{(j)} + \beta_1^{(j)} y_{t-1} + \beta_2^{(j)} y_{t-2} + \ldots + \beta_j^{(j)} y_{t-j} + v_t$$

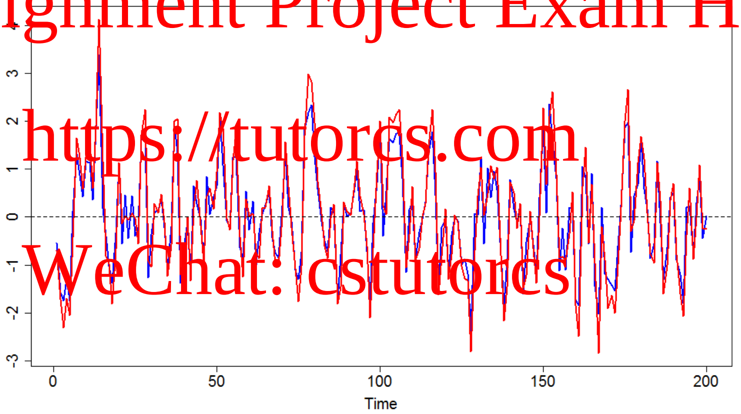
The coefficients can be interpreted as impact of y_{t-j} on y_t when its impact on y_t through $y_{t-1}, \ldots, y_{t-j+1}$ have been properly controlled for. In general, estimation of $\beta_j^{(j)}$ is slightly more involved. However, for white noise, the partial autocorrelation coefficients

$$p(j) = 0$$

for all $j > 0$.

7.2. **MA(1),** $y_t = \epsilon_t - \theta_1 \epsilon_{t-1}$.

Simulated data: (blue: $\theta = -0.5$; red: $\theta = -0.95$)



We compute several quantities:

(1) Unconditional expectation:

$$E(y_t) = E(\epsilon_t - \theta_1 \epsilon_{t-1}) = E(\epsilon_t) - \theta_1 E(\epsilon_{t-1}) = 0$$

(2) Unconditional variance:

$$\begin{aligned} Var(y_t) &= E[(y_t - E(y_t))^2] \\ &= E(y_t^2) = E[(\epsilon_t - \theta_1 \epsilon_{t-1})^2] \\ &= E(\epsilon_t^2) + \theta_1^2 E(\epsilon_{t-1}^2) - 2\theta_1 E(\epsilon_t \epsilon_{t-1}) \end{aligned}$$

By law of iterated expectation,

$$E(\epsilon_t \epsilon_{t-1}) = E\left[E(\epsilon_t \epsilon_{t-1} \mid \epsilon_{t-1})\right] = E\left[E(\epsilon_t \mid \epsilon_{t-1})\epsilon_{t-1}\right] = 0$$

Hence we must have

$$Var(y_t) = \sigma^2 + \theta_1^2 \sigma^2 - 2\theta_1 \times 0 = (1 + \theta_1^2)\sigma^2$$

(3) Covariances

$$
\begin{aligned}
\gamma(1) &= Cov(y_t, y_{t-1}) \\
&= E[(y_t - E(y_t))(y_{t-1} - E(y_{t-1}))] \\
&= E(y_t y_{t-1}) \\
&= E[(\epsilon_t - \theta_1 \epsilon_{t-1})(\epsilon_{t-1} - \theta_1 \epsilon_{t-2})] \\
&= E[\epsilon_t \epsilon_{t-1} - \theta_1 \epsilon_{t-1}^2 - \theta_1 \epsilon_t \epsilon_{t-2} + \theta_1^2 \epsilon_{t-1} \epsilon_{t-2}] \\
&= E(\epsilon_t \epsilon_{t-1}) - E(\theta_1 \epsilon_{t-1}^2) - E(\theta_1 \epsilon_t \epsilon_{t-2}) + E(\theta_1^2 \epsilon_{t-1} \epsilon_{t-2}) \\
&= 0 - \theta_1 \sigma^2 + 0 + 0 \\
&= -\theta_1 \sigma^2
\end{aligned}
$$

$$
\begin{aligned}
\gamma(2) &= Cov(y_t, y_{t-2}) \\
&= E[(y_t - E(y_t))(y_{t-2} - E(y_{t-2}))] \\
&= E(y_t y_{t-2}) \\
&= E[(\epsilon_t - \theta_1 \epsilon_{t-1})(\epsilon_{t-2} - \theta_1 \epsilon_{t-3})] \\
&= E[\epsilon_t \epsilon_{t-2} - \theta_1 \epsilon_{t-1}\epsilon_{t-2} - \theta_1 \epsilon_t \epsilon_{t-3} + \theta_1^2 \epsilon_{t-1}\epsilon_{t-3}] \\
&= E(\epsilon_t \epsilon_{t-2}) - E(\theta_1 \epsilon_{t-1}\epsilon_{t-2}) - E(\theta_1 \epsilon_t \epsilon_{t-3}) + E(\theta_1^2 \epsilon_{t-1}\epsilon_{t-3}) \\
&= 0 - 0 - 0 + 0 \\
&= 0
\end{aligned}
$$

Similarly, $\gamma(j) = 0$ for all $j \geq 2$.

(4) Autocorrelations

$$\begin{aligned}\rho(1) &= \gamma(1)/\gamma(0)\\ &= -\theta_1\sigma^2/[(1+\theta_1^2)\sigma^2]\\ &= -\theta_1/(1+\theta_1^2)\end{aligned}$$

Thus, $\rho(1) > 0$ if $\theta_1 < 0$ and $\rho(1) < 0$ if $\theta_1 > 0$.

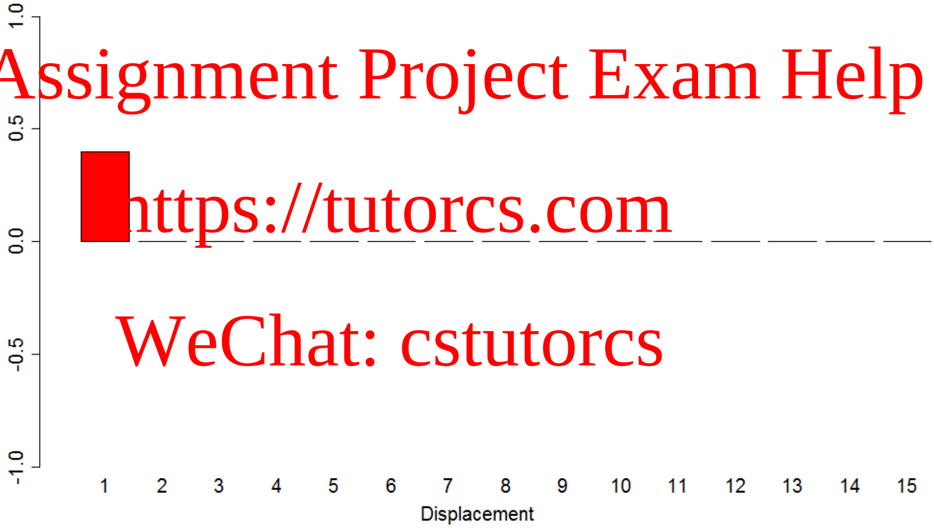
In addition, for an *invertible* MA(1) process (i.e., the MA(1) process can be written as an infinite AR process), $|\theta_1| < 1$, and hence $|\rho(1)| < 0.5$.

θ_1	$\rho(1) = -\theta_1/(1+\theta_1^2)$	θ_1	$\rho(1) = -\theta_1/(1+\theta_1^2)$
1	-0.500	-1	0.500
0.9	-0.497	-0.9	0.497
0.8	-0.488	-0.8	0.488
0.7	-0.470	-0.7	0.470
0.6	-0.441	-0.6	0.441
0.5	-0.400	-0.5	0.400

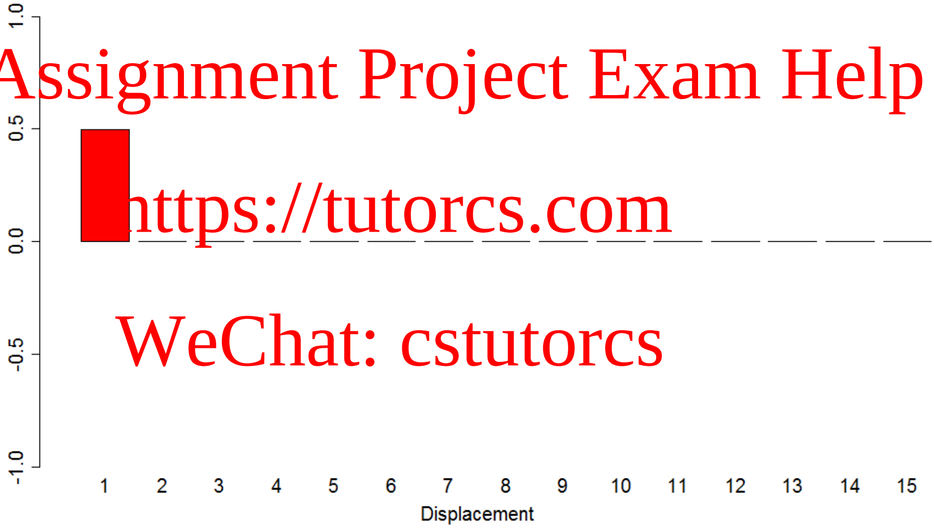
$$\begin{aligned}\rho(2) &= \gamma(2)/\gamma(0) \\ &= 0/[(1+\theta_1^2)\sigma^2] \\ &= 0\end{aligned}$$

Similarly, $\rho(j) = 0$ for all $j \geq 2$.

Population ACF of $y_t = \epsilon_t + 0.5\epsilon_{t-1}$



Population ACF of $y_t = \epsilon_t + 0.95\epsilon_{t-1}$



(5) Partial autocorrelation function (PACF): The partial autocorrelation function for the MA(1) process is a bit more tedious to derive and is related to the autocorrelation function in a complicated way. However, with the help of lag operators, one can easily verify that for an MA(1) process, the PACF is *likely nonzero for all j*. Using the property of lag operator, we have

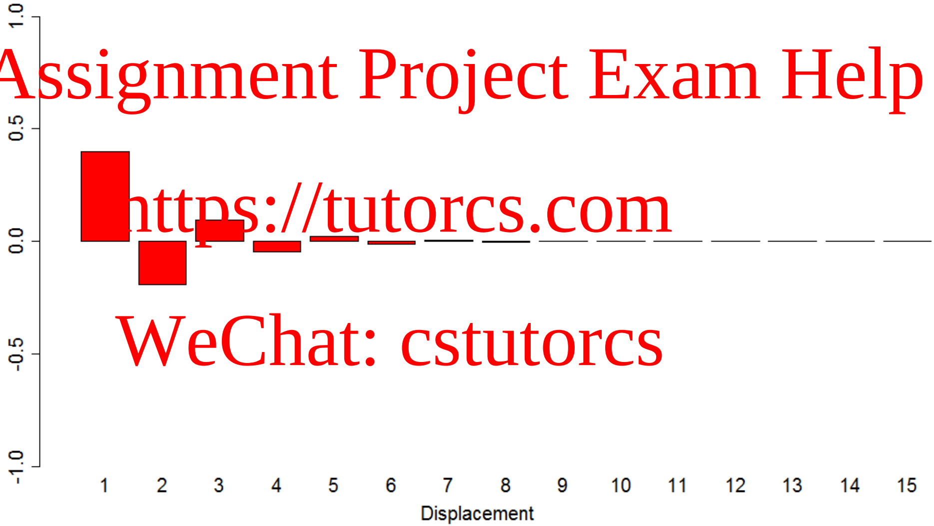
$$
\begin{aligned}
\frac{1}{1-\theta_1 L} y_t &= \epsilon_t \\
(1+\theta_1 L + \theta_1^2 L^2 + \theta_1^3 L^3 + \ldots) y_t &= \epsilon_t \\
y_t + \theta_1 y_{t-1} + \theta_1^2 y_{t-2} + \theta_1^3 y_{t-3} + \ldots &= \epsilon_t \\
y_t &= -\theta_1 y_{t-1} - \theta_1^2 y_{t-2} - \theta_1^3 y_{t-3} + \ldots + \epsilon_t
\end{aligned}
$$

Note that partial autocorrelation coefficients of y_t at displacement j is $\beta_j^{(j)}$ in the regression of

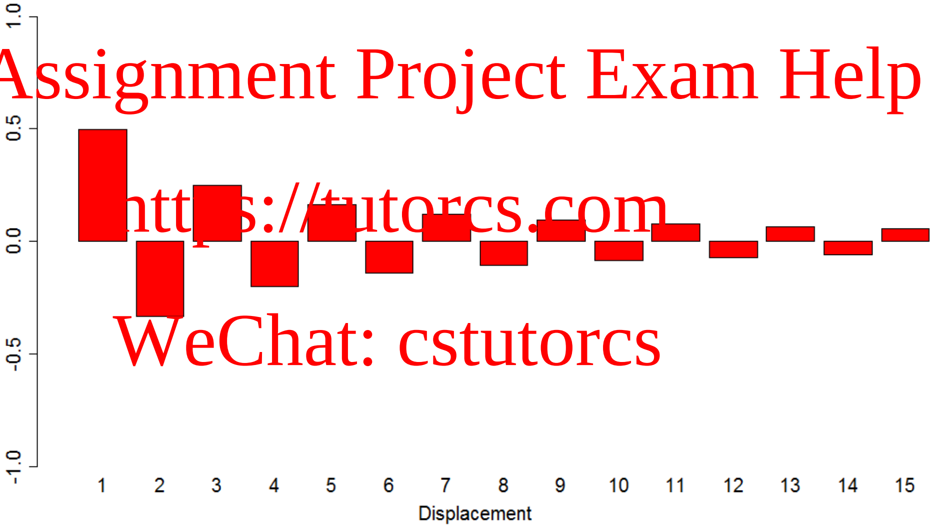
$$y_t = \beta_0^{(j)} + \beta_1^{(j)} y_{t-1} + \beta_2^{(j)} y_{t-2} + \ldots + \beta_j^{(j)} y_{t-j} + v_t$$

Rewriting the MA(1) into an infinite autoregressive process makes it clear that the PACF, $p(j)$, will likely be nonzero for all j, and likely converging to zero in absolute value as j increases.

Population PACF of $y_t = \epsilon_t + 0.5\epsilon_{t-1}$



Population PACF of $y_t = \epsilon_t + 0.95\epsilon_{t-1}$



(6) Conditional expectations:

$$
\begin{aligned}
E(y_{t+1} \mid y_t, y_{t-1}, \ldots) &= E(y_{t+1} \mid \epsilon_t, \epsilon_{t-1}, \ldots) \\
&= E(\epsilon_{t+1} - \theta_1 \epsilon_t \mid \epsilon_t, \epsilon_{t-1}, \ldots) \\
&= E(\epsilon_{t+1} \mid \epsilon_t, \epsilon_{t-1}, \ldots) - E(\theta_1 \epsilon_t \mid \epsilon_t, \epsilon_{t-1}, \ldots) \\
&= -\theta_1 \epsilon_t
\end{aligned}
$$

$$
\begin{aligned}
E(y_{t+2} \mid y_t, y_{t-1}, \ldots) &= E(y_{t+2} \mid \epsilon_t, \epsilon_{t-1}, \ldots) \\
&= E(\epsilon_{t+2} - \theta_1 \epsilon_{t+1} \mid \epsilon_t, \epsilon_{t-1}, \ldots) \\
&= E(\epsilon_{t+2} \mid \epsilon_t, \epsilon_{t-1}, \ldots) - E(\theta_1 \epsilon_{t+1} \mid \epsilon_t, \epsilon_{t-1}, \ldots) \\
&= 0
\end{aligned}
$$

Similarly, $E(y_{t+j}|y_t, y_{t-1}, ...) = 0$ for all $j \geq 2$. Thus, if the true process is MA(1), *knowledge of history helps forecast only up to 1 period ahead*.

In other words, a change in innovation at time t has impact on the conditional expectation of future y_t only for the next one period. Or, y_t will respond to change in innovation at time t only in the current period and the next one period.

7.3. **Inverting MA processes to observables.** The innovations, ϵ_t, are not observable. However, they are related to the observables. Consider MA(1).

$$y_t = \epsilon_t - \theta_1 \epsilon_{t-1}$$

The MA(1) process can be written as

$$\begin{aligned}
\epsilon_t &= y_t + \theta_1 \epsilon_{t-1} \\
&= y_t + \theta_1 [y_{t-1} + \theta_1 \epsilon_{t-2}] \\
&= y_t + \theta_1 y_{t-1} + \theta_1^2 \epsilon_{t-2} \\
&= y_t + \theta_1 y_{t-1} + \theta_1^2 [y_{t-2} + \theta_1 \epsilon_{t-3}] \\
&= y_t + \theta_1 y_{t-1} + \theta_1^2 y_{t-2} + \theta_1^3 \epsilon_{t-3} \\
&= y_t + \theta_1 y_{t-1} + \ldots + \theta_1^n y_{t-n} + \theta_1^{n+1} \epsilon_{t-(n+1)} \\
&= \lim_{n \to \infty} \sum_{j=0}^{n} \theta_1^j y_{t-j} + \lim_{n \to \infty} \theta_1^{(n+1)} \epsilon_{t-(n+1)}
\end{aligned}$$

where $\lim_{n \to \infty} \theta_1^{(n+1)} \epsilon_{t-(n+1)} = 0$ *if* $|\theta_1| < 1$.

If this condition is fulfilled (i.e., $\lim_{n\to\infty}(-\theta_1)^{(n+1)}\epsilon_{t-(n+1)} = 0$), knowing y_{t-j}, $j = 0, ..., \infty$, will allow us to uncover ϵ_t, and hence

$$E(y_{t+h} \mid y_t, y_{t-1}, ...) = E(y_{t+h} \mid \epsilon_t, \epsilon_{t-1}, ...)$$

Alternatively, using the property of lag operator, we have

$$\begin{aligned}
\frac{1}{1-\theta_1 L} y_t &= \epsilon_t \\
(1 + \theta_1 L + \theta_1^2 L^2 + \theta_1^3 L^3 + ...) y_t &= \epsilon_t \\
y_t + \theta_1 y_{t-1} + \theta_1^2 y_{t-2} + \theta_1^3 y_{t-3} + ... &= \epsilon_t \\
y_t &= -\theta_1 y_{t-1} - \theta_1^2 y_{t-2} - \theta_1^3 y_{t-3} + ... + \epsilon_t
\end{aligned}$$

8. Estimation of MA processes

There are two approaches to estimate an MA(1) model: maximum likelihood and nonlinear least squares (NLS).
Note that the MA(1) process can be rewritten as

$$y_t = \epsilon_t - \theta_1 \epsilon_{t-1} \qquad \Longrightarrow \qquad \epsilon_t = y_t + \theta_1 \epsilon_{t-1}$$

$$\begin{aligned}
\epsilon_1 &= y_1 + \theta_1 \epsilon_0 \\
\epsilon_2 &= y_2 + \theta_1 \epsilon_1 \\
\epsilon_3 &= y_3 + \theta_1 \epsilon_2 \\
\vdots &\quad \vdots \quad \vdots \\
\epsilon_T &= y_T + \theta_1 \epsilon_{T-1}
\end{aligned}$$

NLS estimator minimizes

$$\sum_{t=1}^{T} e_t(b)^2$$

where $e_t(b)$ is the residuals with an assumed value of parameter estimate b for θ_1.

By assuming a normal density, the approach of ML will maximize the joint likelihood of

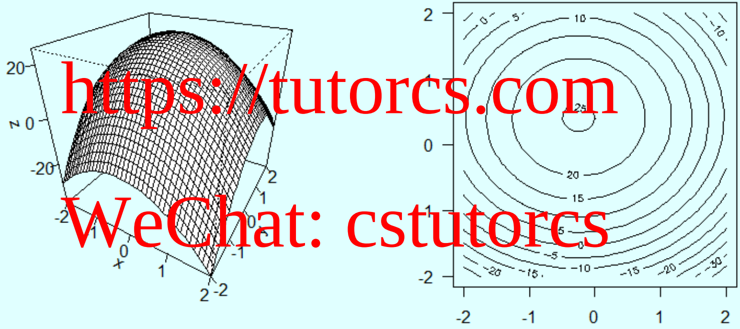
$$f(e_1, e_2, e_3, ..., e_T \mid b) = f(e_1 \mid b) \times f(e_2 \mid e_1, b) \times f(e_3 \mid e_2, e_1, b) \times$$
$$... \times f(e_T \mid e_{T-1}, ..., e_2, e_1, b)$$

where $f(. \mid .)$ is a conditional density.

Both approaches (NLS and ML) estimate the parameters *numerically*. Both approaches set e_0 and earlier e's to their unconditional expectations, i.e., zero.

The idea is to evaluate the objective functions at different values of b and choose the b that minimizes the sum of squared residuals or maximize the joint likelihood. When an initial set of parameters are supplied, most statistical packages will do the numerical optimization using Ralph-Newton-like algorithms.

Ralph-Newton-like algorithms are essentially variations of efficient hill-climing algorithms to find the maximum or minimum. The major assumption is that the hill of a relatively "smooth".

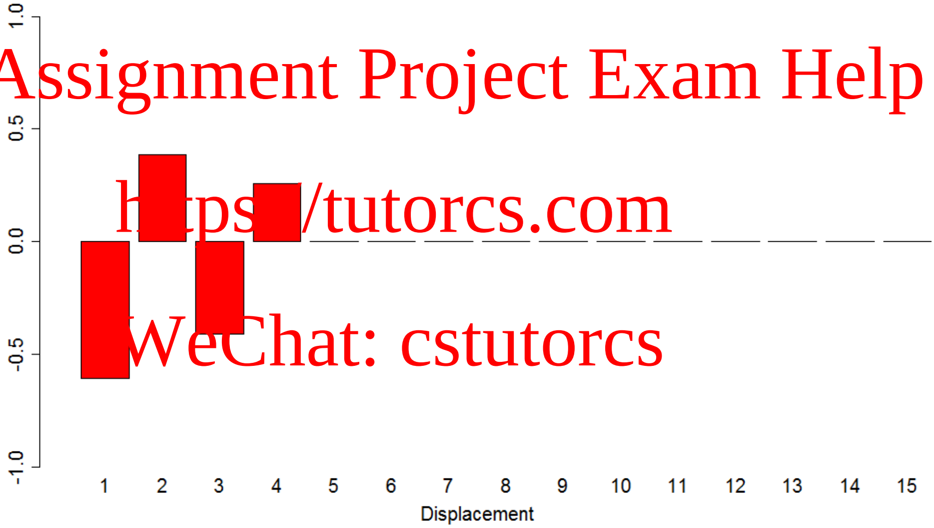


The property and estimation of the more general case of MA(q)

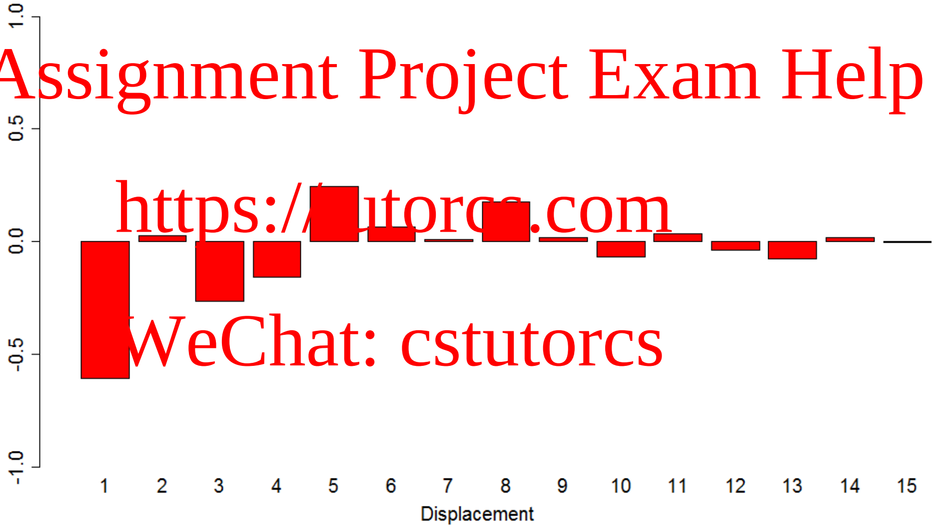
$$y_t = \epsilon_t + b_1\epsilon_{t-1} + b_2\epsilon_{t-2} + ... + b_q\epsilon_{t-q}$$

are similar but the derivation is more complicated algebraically.

Population ACF of $y_t = \epsilon_t - 0.6\epsilon_{t-1} + 0.3\epsilon_{t-2} - 0.5\epsilon_{t-3} + 0.5\epsilon_{t-4}$



Population PACF of $y_t = \epsilon_t - 0.6\epsilon_{t-1} + 0.3\epsilon_{t-2} - 0.5\epsilon_{t-3} + 0.5\epsilon_{t-4}$



9. Autoregressive models, AR(p)

Under certain circumstances, an infinite MA process (as in the general Wold representation for y_t) can be "inverted" into a finite-order autoregressive form, i.e.,

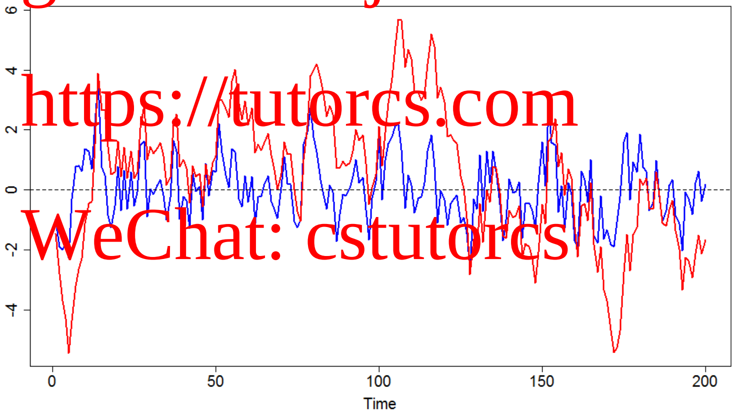
$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t$$

This is called a p-th order autoregressive process AR(p). Note that the process has p unknown coefficients only: $\phi_1, ..., \phi_p$. The approximation of the wold representation by a AR(p) process is important because a process of infinite parameters is now reduced to a process of finite parameters, making estimation feasible.

Note that the AR(p) model looks like a standard linear regression model with errors that are zero-mean, homoskedastic, and serially uncorrelated, and hence can be estimated with standard linear regression packages.

9.1. **AR(1)**, $y_t = \phi_1 y_{t-1} + \epsilon_t$.

Simulated data: (blue: $\phi = 0.5$; red: $\phi = 0.95$)



We may obtain Wold representation of AR(1) by repeated substitution.

$$
\begin{aligned}
y_t &= \phi_1 y_{t-1} + \epsilon_t \\
&= \phi_1(\phi_1 y_{t-2} + \epsilon_{t-1}) + \epsilon_t \\
&= \phi_1^2 y_{t-2} + \phi_1 \epsilon_{t-1} + \epsilon_t \\
&= \phi_1^2(\phi_1 y_{t-3} + \epsilon_{t-2}) + \phi_1 \epsilon_{t-1} + \epsilon_t \\
&= \phi_1^3 y_{t-3} + \phi_1^2 \epsilon_{t-2} + \phi_1 \epsilon_{t-1} + \epsilon_t \\
&= \ldots \\
&= \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_1^2 \epsilon_{t-2} + \phi_1^3 \epsilon_{t-3} + \ldots \phi_1^{n-1} \epsilon_{t-n-1} + \phi_1^n y_{t-n} \\
&= \sum_{i=0}^{n-1} \phi_1^i \epsilon_{t-i} + \phi_1^n y_{t-n}
\end{aligned}
$$

Taking n to infinity, i.e., infinite repeated substitutions, we have

$$y_t = \lim_{n\to\infty} \sum_{i=0}^{n-1} \phi_1^i \epsilon_{t-i} + \lim_{n\to\infty} \phi_1^n y_{t-n}$$

If $y_{t-\infty}$ is finite and $|\phi_1| < 1$, we must have $\lim_{n\to\infty} \phi_1^i y_{t-n} = 0$. Then we have

$$y_t = \sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i}$$

(1) Unconditional expectation

$$E(y_t) = E\left(\sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i}\right) = \sum_{i=0}^{\infty} \phi_1^i E(\epsilon_{t-i}) = 0$$

Alternative way to compute unconditional expectation is to use the stationary property $E(y_t) = E(y_{t-1})$.

$$y_t = \phi_1 y_{t-1} + \epsilon_t$$

$$E\left(y_t\right) = \phi_1 E\left(y_{t-1}\right) + E\left(\epsilon_t\right)$$

$$E\left(y_t\right) = \phi_1 E\left(y_t\right) + 0$$

$$E\left(y_t\right) = \frac{0}{1-\phi_1} = 0$$

(2) Unconditional variance: One way to compute it is to start from the Wold representation.

$$Var(y_t) = Var\left(\sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i}\right) = \sum_{i=0}^{\infty} \phi_1^{2i} Var(\epsilon_{t-i}) + 2\sum_{i \neq j} \phi_1^{i+j} Cov(\epsilon_{t-i}, \epsilon_{t-j})$$

$$= \sum_{i=0}^{\infty} \phi_1^{2i} Var(\epsilon_{t-i})$$

$$= \sigma^2 \sum_{i=0}^{\infty} \phi_1^{2i} = \frac{1}{1-\phi_1^2}\sigma^2$$

We can also use the property $E(y_t) = 0$ to imply $Var(y_t) = E[y_t - E(y_t)]^2 = E(y_t^2)$

$$\begin{aligned}
Var(y_t) &= E(y_t^2) = E\left(\sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i}\right)^2 \\
&= E\left(\sum_{i=0}^{\infty} \phi_1^{2i} \epsilon_{t-i}^2 + 2 \sum_{i \neq j} \phi_1^{i+j} \epsilon_{t-i} \epsilon_{t-j}\right) \\
&= E\left(\sum_{i=0}^{\infty} \phi_1^{2i} \epsilon_{t-i}^2\right) = \sum_{i=0}^{\infty} \phi_1^{2i} E(\epsilon_{t-i}^2) \\
&= \sigma^2 \sum_{i=0}^{\infty} \phi_1^{2i} \\
&= \frac{1}{1-\phi_1^2} \sigma^2
\end{aligned}$$

Alternatively, we can use the property that $Var(y_t) = Var(y_{t-s})$.

$$
\begin{aligned}
y_t &= \phi_1 y_{t-1} + \epsilon_t \\
Var\left(y_t\right) &= Var(\phi_1 y_{t-1} + \epsilon_t) \\
Var\left(y_t\right) &= \phi_1^2 Var\left(y_{t-1}\right) + Var\left(\epsilon_t\right) + 2\phi_1 Cov\left(y_{t-1}, \epsilon_t\right) \\
Var\left(y_t\right) &= \phi_1^2 Var\left(y_t\right) + \sigma^2 + 0 \\
Var\left(y_t\right) &= \frac{1}{1-\phi_1^2}\sigma^2
\end{aligned}
$$

(3) Two sets of *cross moments* are used often, and thus are introduced here for later reference. For all $j > 0$,

$$\begin{aligned} E(y_{t-j}\epsilon_t) &= E[E(y_{t-j}\epsilon_t \mid y_{t-j})] \\ &= E[y_{t-j}E(\epsilon_t \mid y_{t-j})] \\ &= E[y_{t-j} \times 0] \\ &= 0 \end{aligned}$$

$$\begin{aligned} E(y_t\epsilon_t) &= E[(\phi_1 y_{t-1} + \epsilon_t)\epsilon_t] \\ &= \phi_1 E(y_{t-1}\epsilon_t) + E(\epsilon_t^2) \\ &= \phi_1 \times 0 + \sigma^2 \\ &= \sigma^2 \end{aligned}$$

(4) Covariances

$$
\begin{aligned}
\gamma(1) &= Cov(y_t, y_{t-1}) \\
&= E[(y_t - E(y_t))(y_{t-1} - E(y_{t-1}))] \\
&= E[y_t y_{t-1}] \\
&= E[(\phi_1 y_{t-1} + \epsilon_t) y_{t-1}] \\
&= E(\phi_1 y_{t-1}^2) + E(\epsilon_t y_{t-1}) \\
&= \phi_1 E(y_{t-1}^2) + E(\epsilon_t y_{t-1}) \\
&= \phi_1 Var(y_t) + 0 \\
&= \phi_1 Var(y_t)
\end{aligned}
$$

$$
\begin{aligned}
\gamma(2) &= Cov(y_t, y_{t-2}) \\
&= E[(y_t - E(y_t))(y_{t-2} - E(y_{t-2}))] \\
&= E[y_t y_{t-2}] \\
&= E[(\phi_1 y_{t-1} + \epsilon_t) y_{t-2}] \\
&= E[(\phi_1(\phi_1 y_{t-2} + \epsilon_{t-1}) + \epsilon_t) y_{t-2}] \\
&= E[(\phi_1^2 y_{t-2} + \phi_1 \epsilon_{t-1} + \epsilon_t) y_{t-2}] \\
&= E[\phi_1^2 y_{t-2}^2 + \phi_1 \epsilon_{t-1} y_{t-2} + \epsilon_t y_{t-2}] \\
&= \phi_1^2 E(y_{t-2}^2) + \phi_1 E(\epsilon_{t-1} y_{t-2}) + E(\epsilon_t y_{t-2}) \\
&= \phi_1^2 Var(y_t)
\end{aligned}
$$

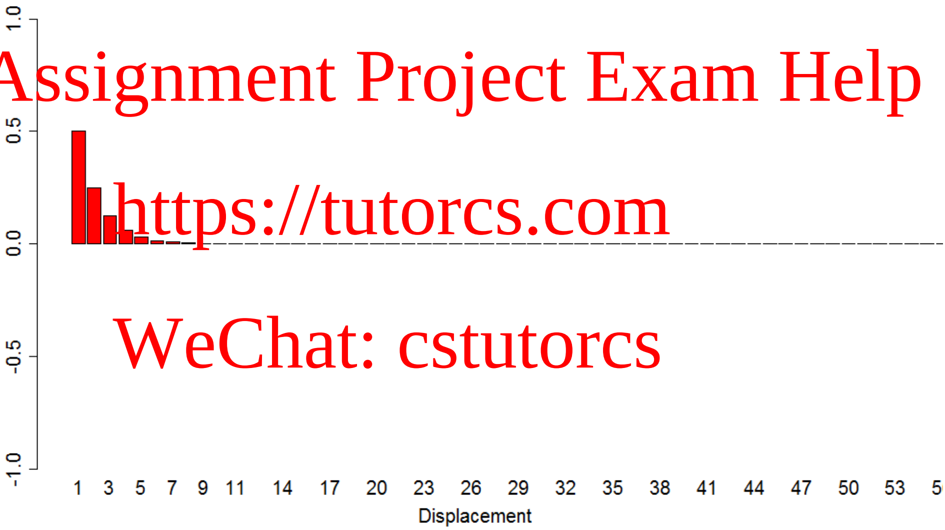
We can repeat the above steps to show that , $\gamma(j) = \phi_1^j Var(y_t)$ for all $j \geq 0$.

(5) Autocorrelations: For all $j \geq 0$,

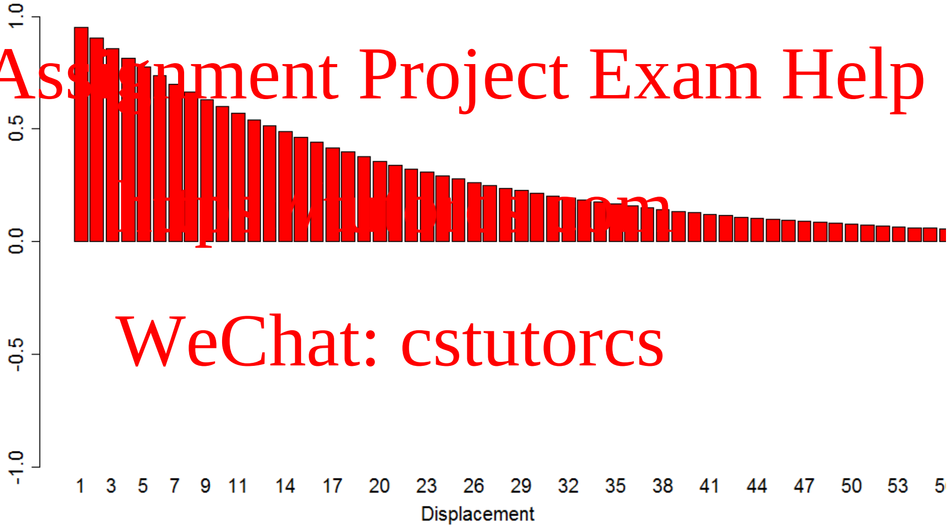
$$\begin{aligned} \rho(j) &= Corr(y_t, y_{t-j}) \\ &= \gamma(j)/\gamma(0) \\ &= \phi_1^j Var(y_t)/Var(y_t) \\ &= \phi_1^j \end{aligned}$$

(6) Partial autocorrelation function (PACF): The partial autocorrelation function for the AR(1) process is a relatively easy to derive. The PACF, $p(1)$, equals to ϕ_1. For $j \geq 2$, $p(j) = 0$.

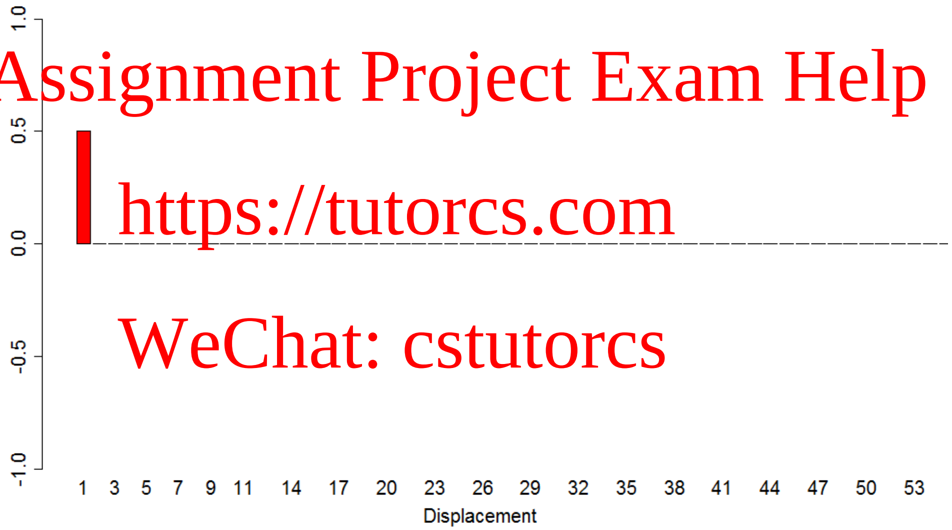
Population autocorrelation function of $y_t = 0.5y_{t-1} + \epsilon_t$



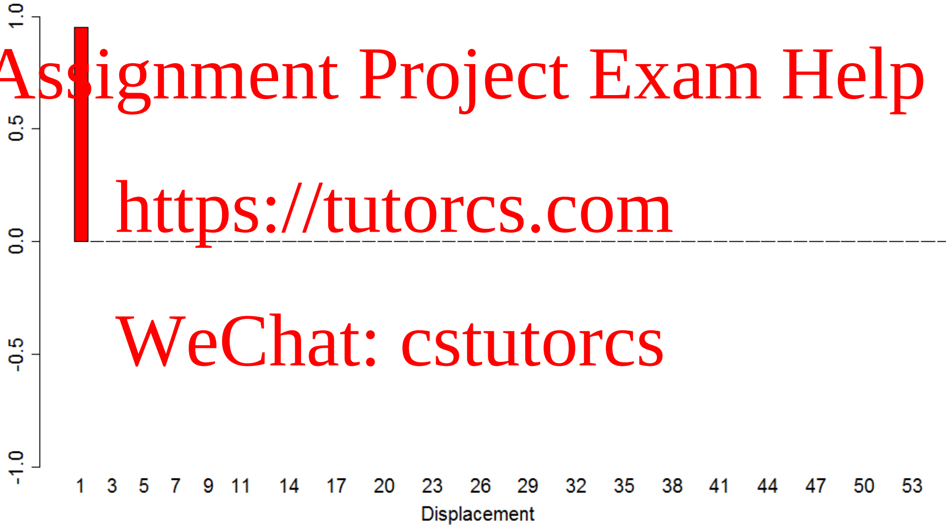
Population autocorrelation function of $y_t = 0.95y_{t-1} + \epsilon_t$



Population partial autocorrelation function of $y_t = 0.5y_{t-1} + \epsilon_t$



Population partial autocorrelation function of $y_t = 0.95y_{t-1} + \epsilon_t$



(7) Conditional expectations:

$$
\begin{aligned}
E(y_{t+1} \mid y_t, y_{t-1}, ..., \epsilon_t, \epsilon_{t-1}, ...) &= E(\phi_1 y_t + \epsilon_{t+1} \mid y_t, y_{t-1}, ..., \epsilon_t, \epsilon_{t-1}, ...) \\
&= \phi_1 y_t
\end{aligned}
$$

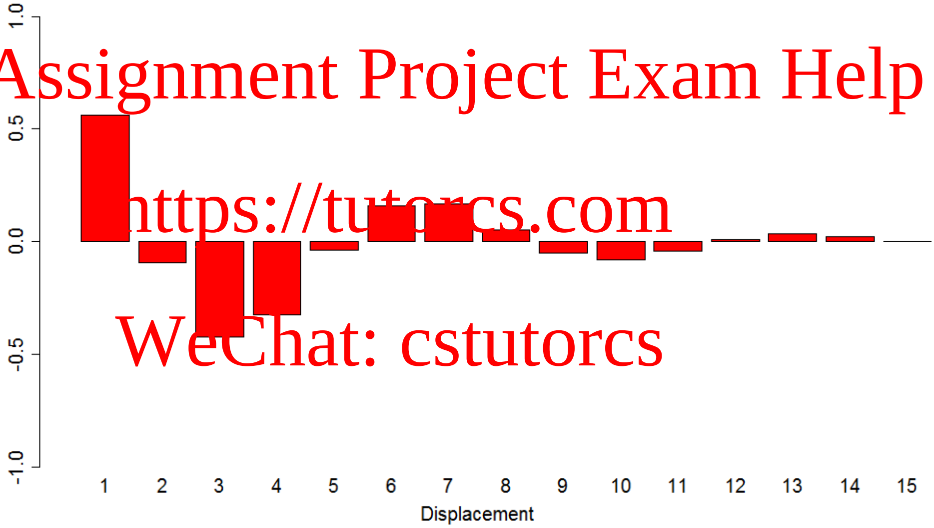
$$
\begin{aligned}
E(y_{t+2} \mid y_t, y_{t-1}, ...) &= E(\phi_1 y_{t+1} + \epsilon_{t+2} \mid y_t, y_{t-1}, ..., \epsilon_t, \epsilon_{t-1}, ...) \\
&= \phi_1 E(y_{t+1} \mid y_t, y_{t-1}, ..., \epsilon_{t1}, \epsilon_{t-1}, ...) \\
&= \phi_1^2 y_t
\end{aligned}
$$

Generally,

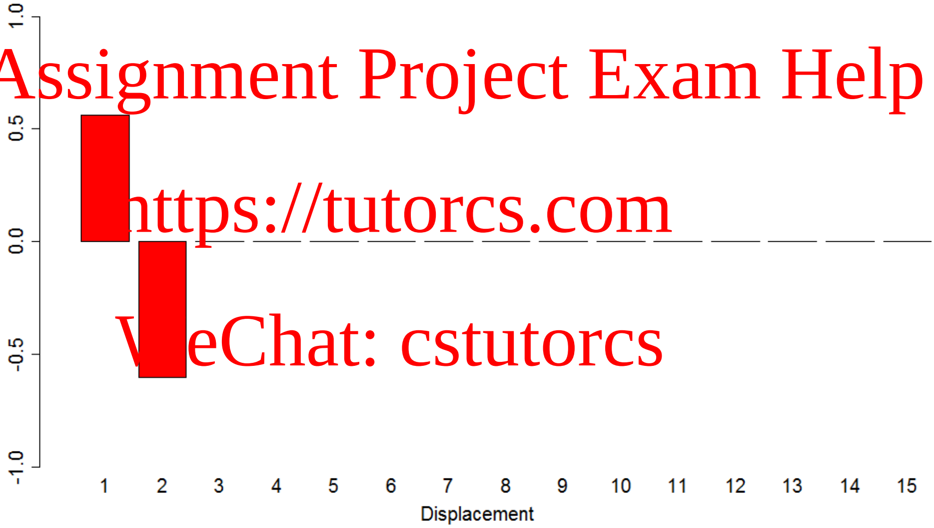
$$E(y_{t+j} \mid y_t, y_{t-1}, ..., \epsilon_t, \epsilon_{t-1}, ...) = \phi_1^j y_t$$

for all $j \geq 1$. Thus, if the true process is AR(1), knowledge of the most recent realization (i.e., y_t) helps forecast future value (i.e., y_{t+j}), even into the distant future, i.e., large j.

Population ACF of $y_t = 0.9y_{t-1} - 0.6y_{t-2} + \epsilon_t$



Population PACF of $y_t = 0.9y_{t-1} - 0.6y_{t-2} + \epsilon_t$



9.2. **Estimation.** The estimation of AR models is simpler than that of the MA models. Given the AR(1) is $y_t = \phi_1 y_{t-1} + \epsilon_t$ and ϵ_t has the same variance across observation, it is simple to apply *ordinary least squares* regression (OLS) to estimate the parameters. Suppose we have $y_1, \ldots, y_T$, we will have

$$
\begin{aligned}
e_2 &= y_2 - by_1 \\
e_3 &= y_3 - by_2 \\
\cdots & \cdots \cdots \\
e_T &= y_T - by_{T-1}
\end{aligned}
$$

Then, the parameter estimates are chosen to minimize

$$\sum_{t=2}^{T} e_t(b)^2 = \sum_{t=2}^{T} (y_t - by_{t-1})^2$$

Given the AR(2) is $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$ and ϵ_t has the same variance across observation, it is simple to apply *ordinary least squares* regression (OLS) to estimate the parameters. Suppose we have $y_1, ..., y_T$, we will have

$$\begin{aligned} e_3 &= y_3 - b_1 y_2 - b_2 y_1 \\ e_4 &= y_4 - b_1 y_3 - b_2 y_2 \\ ... &\quad ... \quad ... \\ e_T &= y_T - b_1 y_{T-1} - b_2 y_{T-2} \end{aligned}$$

Then, the parameter estimates are chosen to minimize

$$\sum_{t=3}^{T} e_t(b)^2 = \sum_{t=3}^{T} (y_t - b_1 y_{t-1} - b_2 y_{t-2})^2$$

The AR(p) process

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t$$

are similar to AR(1) and AR(2).

The parameter estimates are chosen to minimize

$$\sum_{t=p+1}^{T} e_t(b)^2 = \sum_{t=p+1}^{T} (y_t - b_1 y_{t-1} - b_2 y_{t-2} - ... - b_p y_{t-p})^2$$

9.3. **Stationarity of an AR(p) process.**

We know that "AR(0)"

$$y_t = \epsilon_t$$

is stationary.

We know that AR(1)

$$y_t = \phi_1 y_{t-1} + \epsilon_t$$

$$(1 - \phi_1 L) y_t = \epsilon_t$$

is stationary if $\mid \phi_1 \mid < 1$. If it is stationary, we can also write

$$\begin{aligned} y_t &= (1 - \phi_1 L)^{-1} \epsilon_t \\ &= \left(\phi_1^0 L^0 + \phi_1^1 L^1 + \phi_1^2 L^2 + \ldots\right) \epsilon_t \\ &= \epsilon_t + \phi_1^1 \epsilon_{t-1} + \phi_1^2 \epsilon_{t-2} + \ldots \\ &= \sum_{i=0}^{\infty} b_i \epsilon_{t-i} \end{aligned}$$

If we focus on $(1-\phi_1 L)$ and think about

$$(1-\phi_1 x)=0,$$

i.e., replacing the L in $(1-\phi_1 L)$ by x and set it to zero.

$$(1-\phi_1 L) \quad \longrightarrow \quad (1-\phi_1 x) \quad \longrightarrow \quad (1-\phi_1 x) = 0$$

We can see that

$$|\phi_1|<1 \iff |x|>1$$

That is, instead of checking the absolute value of ϕ_1, we can check the stationarity by

(1) Writing the AR(1) in the form of lag operators:

$$(1-\phi_1 L)y_t = e_t$$

(2) Replacing lag operator L by x and set the polynomial of lag operators to zero:

$$(1-\phi_1 L) \longrightarrow (1-\phi_1 x) \longrightarrow (1-\phi_1 x) = 0$$

(3) Checking whether the solution of x to the equation is larger than 1 in absolute values, i.e.,

$$|x| > 1$$

Consider AR(2) in y_t

$$y_t = (\phi_1 + \phi_2)y_{t-1} - \phi_1\phi_2 y_{t-2} + \epsilon_t$$

$$\left[1 - (\phi_1 + \phi_2)L + \phi_1\phi_2 L^2\right] y_t = \epsilon_t$$

$$(1 - \phi_1 L)(1 - \phi_2 L)y_t = \epsilon_t$$

is stationary if $\mid \phi_1 \mid < 1$ and $\mid \phi_2 \mid < 1$. If it is stationary, we can also write

$$\begin{aligned} y_t &= (1 - \phi_2 L)^{-1}(1 - \phi_1 L)^{-1}\epsilon_t \\ &= \left(\phi_1^0 L^0 + \phi_1^1 L^1 + \phi_1^2 L^2 + \ldots\right)\left(\phi_2^0 L^0 + \phi_2^1 L^1 + \phi_2^2 L^2 + \ldots\right)\epsilon_t \\ &= \left(\sum_{i=0}^{\infty} b_i L\right)\epsilon_t \\ &= \sum_{i=0}^{\infty} b_i \epsilon_{t-i} \end{aligned}$$

$$y_t = (\phi_1 + \phi_2)y_{t-1} - \phi_1\phi_2 y_{t-2} + \epsilon_t$$

$$\left[1 - (\phi_1 + \phi_2)L + \phi_1\phi_2 L^2\right] y_t = \epsilon_t$$

$$(1 - \phi_1 L)(1 - \phi_2 L)y_t = \epsilon_t$$

If we focus on $(1 - \phi_1 L)(1 - \phi_2 L)$ and think about

$$(1 - \phi_1 x)(1 - \phi_2 x) = 0$$

i.e., replacing the L in $(1 - \phi_1 L)(1 - \phi_2 L)$ by x and set it to zero.

$$(1 - \phi_1 L)(1 - \phi_2 L) \longrightarrow (1 - \phi_1 x)(1 - \phi_2 x) \longrightarrow (1 - \phi_1 x)(1 - \phi_2 x) = 0$$

or equivalently

$$1 - (\phi_1 + \phi_2)x + \phi_1\phi_2 x^2 = 0,$$

we can see that

$$|\ \phi_1\ | < 1 \text{ and } |\ \phi_2\ | < 1 \quad \Longleftrightarrow \quad |\ x_1\ | > 1 \text{ and } |\ x_2\ | > 1$$

For a general AR(p) process, we can check stationarity of y_t in two steps.

(1) Write the AR(p) process

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t$$

in lag polynomials

$$(1 - \phi_1 L - ... - \phi_p L^p) y_t = \epsilon_t.$$

(2) Factorize this polynomial of lag operators as

$$1 - \phi_1 L - ... - \phi_p L^p = (1 - c_1 L)(1 - c_2 L)...(1 - c_p L)$$

for some, potentially complex, numbers $c_1, c_2, ..., c_p$. Then, stationarity requires that $| c_1 | < 1$, $| c_2 | < 1$, ..., $| c_p | < 1$.

Alternatively, we can solve the root in the polynomial

$$1 - \phi_1 x - ... - \phi_p x^p = 0$$

and check if all the roots are larger than 1 in absolute values. In case some of the roots are complex numbers

$$x = a + bi,$$

check whether the modulus

$$\mid x \mid = \sqrt{a^2 + b^2}$$

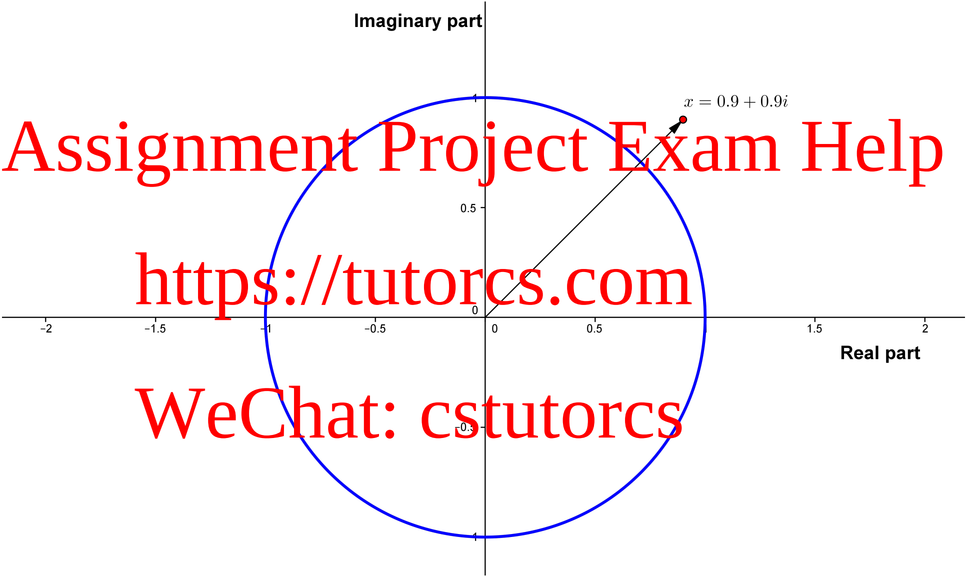
is larger than 1.

When the real and imagine parts of complex roots are represented in a coordinate plane (real on the x-axis, and imaginary on the y-axis), the check of a modulus of a root is like checking the root against a circle with unit distance.

Thus, we often say stationary requires all the roots x in

$$1 - \phi_1 x - \ldots - \phi_p x^p = 0$$

lies outside the unit circle.


Imaginary part
Real part
$x = 0.9 + 0.9i$
-2
-1.5
-1
-0.5
0
0.5
1.5
2
1
0.5
-0.5
-1

10. Model Selection

Two competing and complementary strategies:

(1) Given a set of ARMA(p,q) models, select the model with the smallest AIC (or equivalently, the smallest SIC).

		MA order 0	MA order 1	MA order 2	MA order 3
AR order	0	ARMA(0,0)	ARMA(0,1)	ARMA(0,2)	ARMA(0,3)
	1	ARMA(1,0)	ARMA(1,1)	ARMA(1,2)	ARMA(1,3)
	2	ARMA(2,0)	ARMA(2,1)	ARMA(2,2)	ARMA(2,3)
	3	ARMA(3,0)	ARMA(3,1)	ARMA(3,2)	ARMA(3,3)

(2) Start with an ARMA(p,q) model, stop if the residuals look like white noise. Otherwise, continue to add AR and/ or MA terms. If there are several models that yields white noise residuals, choose the one with the smallest number of parameters, and the one that is easier to estimate (usually those with less MA terms).

11. Model Selection – Example

A time series of 200 observations of unknown ARMA order is given. Can we identify the ARMA order correctly?

AIC

		MA order			
		0	1	2	3
AR order	0	677.1333	599.2117	565.2696	561.279
	1	551.3753	553.2422	553.8763	555.8711
	2	553.2054	552.1556	551.9593	553.9592
	3	553.4265	554.0429	552.1716	553.9638

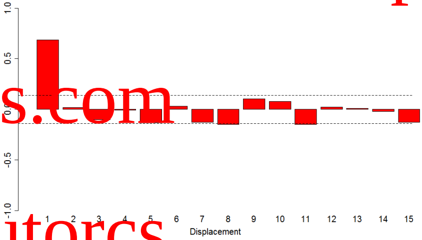
SIC

		MA order			
		0	1	2	3
AR order	0	683.7299	609.1066	578.4629	577.7706
	1	561.2702	566.4355	570.3679	575.661
	2	566.3987	568.6472	571.7492	577.0475
	3	569.918	573.8328	575.2599	580.3504

ACF and PACF of estimated ϵ_t in an assumed model of ARMA(0,0)

$$y_t = c + \epsilon_t$$

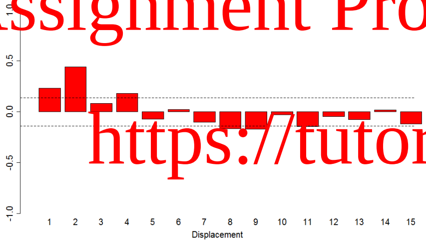
ACF



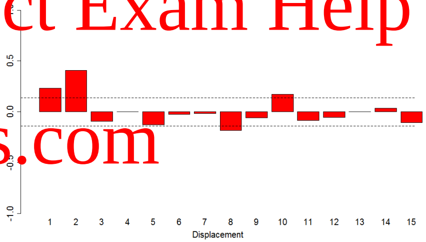
PACF

ACF and PACF of estimated ϵ_t in an assumed model of ARMA(0,1)

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1}$$



ACF

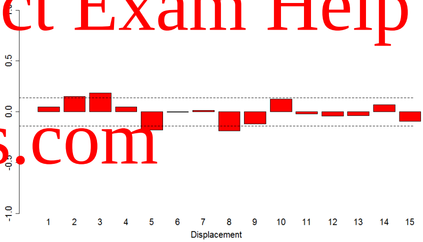


PACF

ACF and PACF of estimated ϵ_t in an assumed model of ARMA(0,2)

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}$$

ACF

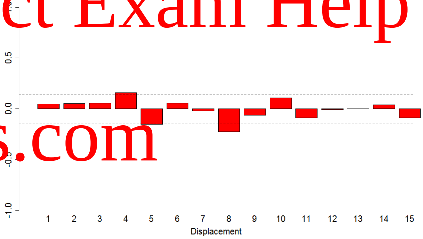


PACF

ACF and PACF of estimated ϵ_t in an assumed model of ARMA(0,3)

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \theta_3 \epsilon_{t-3}$$

ACF

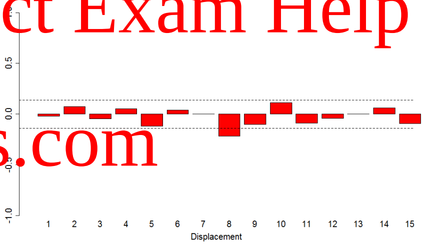


PACF

ACF and PACF of estimated ϵ_t in an assumed model of ARMA(1,0)

$$y_t = c + \phi_1 y_{t-1} + \epsilon_t$$

ACF

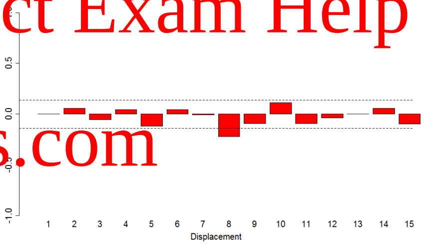


PACF

ACF and PACF of estimated ϵ_t in an assumed model of ARMA(2,0)

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$$

ACF

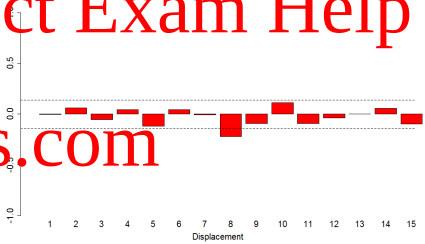


PACF

ACF and PACF of estimated ϵ_t in an assumed model of ARMA(1,1)

$$y_t = c + \phi_1 y_{t-1} + \epsilon_t + \theta_1 \epsilon_{t-1}$$

ACF



PACF

p-value of Box Test

		MA order					
		0	1	2	3	4	5
AR order	0	0	0	0.0019	0.0099	0.3452	0.3393
	1	0.1073	0.1179	0.1006	0.1031	0.3366	0.3874
	2	0.1202	0.1325	0.2218	0.2214	0.3576	0.4011
	3	0.0954	0.1628	0.2277	0.2492	0.3490	0.9302
	4	0.0974	0.1962	0.2422	0.2323	0.6820	0.7036
	5	0.3414	0.4016	0.756	0.5999	0.8795	0.9687

We cannot reject the null of white noise at 5% level with ARMA(p,q) whenever $p \geq 1$, or whenever $q \geq 4$.

The truth is:

$$y_t = 0.7y_{t-1} + \epsilon_t$$

In this simulation exercise, both model selection criteria (AIC/SIC and white noise test) suggest the same model.

In actual applications, the several model selection criteria may point to different models. In that case, researchers would have to make a *judgement*. Some researcher will choose the largest implied order when the data series is not too short (estimation uncertainty is relatively small even when the model contains a lot of parameters); some may always choose the smallest implied order.

12. Approximation

Nobody knows the true AR and MA orders. The key is approximation.

- Any MA process may be approximated by an AR(p) process, for sufficient large p. And the residuals will look like white noise.
- Any AR process may be approximated by a MA(q) process, for sufficient large q. And the residuals will look like white noise.

In fact, if an AR(p) process can be written exactly as a MA(q) process, the AR(p) process is called invertible.

Similarly, if a MA(q) process can be written exactly as an AR(p) process, the MA(q) process is called invertible.