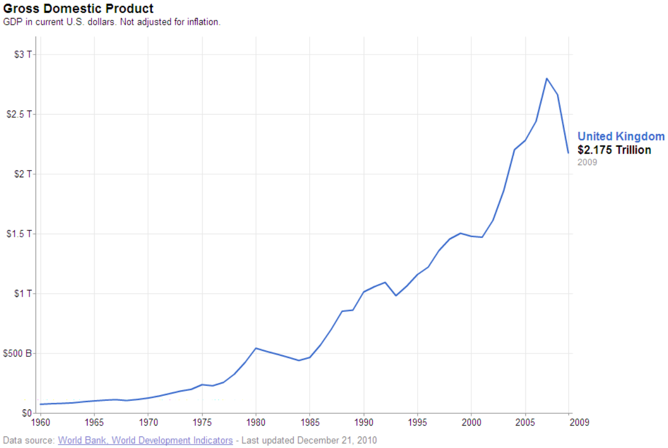
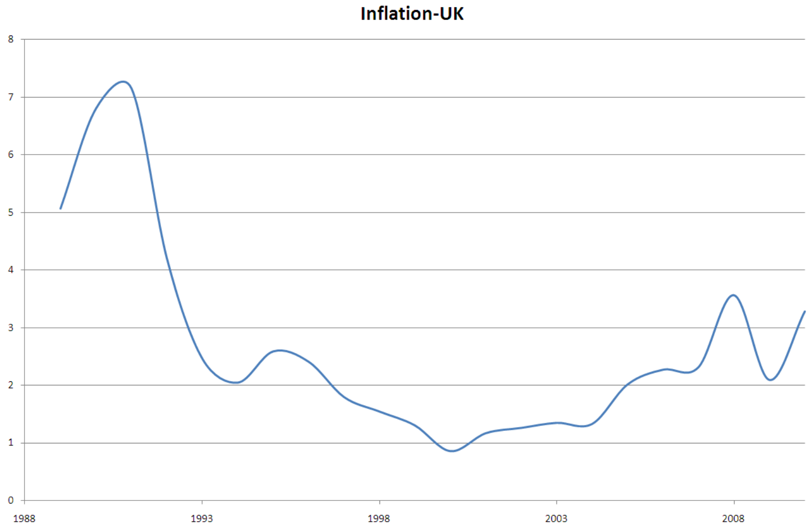
Univariate time series modelling

- Stochastic Process
- Stationary Processes
- Wold's Decomposition Theorem
- ARMA processes
- Box Jenkins Methodology
- Forecasts


Gross Domestic Product
GDP in current U.S. dollars. Not adjusted for inflation.
$3 T
$2.5 T
$2 T
$1.5 T
$1 T
$500 B
$0
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2009
United Kingdom
$2.175 Trillion
2009
Data source: World Bank, World Development Indicators - Last updated December 21, 2010


Inflation-UK
8
7
6
5
4
3
2
1
0
1988
1993
1998
2003
2008

Definition

Stochastic Process: A stochastic process $\{X_t\}_{t=-\infty}^{+\infty}$ is a sequence of random variables ordered by time.

- A sequence $\{x_t\}_{t=-\infty}^{+\infty} = \{..., x_{-2}, x_{-1}, x_0, x_1, x_2, ...\}$ is regarded as a *realization* of a *stochastic process* i.e. for each value of t (each point in time) x_t is drawn from a distribution (or population) of X_t's.
- Let $f_{X_t}(x_t)$ denote the *probability density function* (pdf) of X_t – note that it depends on t so that each element in the realization may be drawn from a different distribution.

Definition

The *expectation* (or *mean*) of X_t is given by

$$\begin{aligned} E(X_t) &= \mu_t \\ &= \int_{-\infty}^{+\infty} x_t f_{X_t}(x_t) dx_t. \end{aligned}$$

Definition

The *variance* of X_t is given by

$$\begin{aligned} var(X_t) &= E[(X_t - \mu_t)^2] = \gamma_{0t} \\ &= \int_{-\infty}^{+\infty} (x_t - \mu_t)^2 f_{X_t}(x_t) dx_t. \end{aligned}$$

Definition

The *autocovariance* of X_t are given by (for $j = 0, \pm 1, \pm 2, ...$)

$$\gamma_{jt} = cov(X_t, X_{t-j}) = E[(X_t - \mu_t)(X_{t-j} - \mu_{t-j})]$$

$$= \int_{\mathbb{R}^{j+1}} (x_t - \mu_t)(x_{t-j} - \mu_{t-j}) f_{X_t, X_{t-1}, ..., X_{t-j}}(x_t, x_{t-1}, ..., x_{t-j}) dx_t dx_{t-1} ... dx_{t-j}.$$

where $f_{X_t, X_{t-1}, ..., X_{t-j}}(x_t, x_{t-1}, ..., x_{t-j})$ denotes the joint pdf of $(X_t, X_{t-1}, ..., X_{t-j})$.

Note that all quantities are indexed with t

- In order to be able to estimate such quantities it would be necessary to obtain a sample of observations on X for each t, which is simply not possible.
- In practice we are faced with the task of making inferences about the statistical properties of the variable X from a single finite realization or set of (T) observations: $\{x_t\}_{t=1}^{T} = \{x_1, x_2, ..., x_T\}$.
- In order to do this we need to impose some structure e.g. stationarity.

Definition

A Strictly Stationary Process: A stochastic process is strictly (or strongly) stationary if for every collection of time indices $1 \leq t_1 < \ldots < t_m$ the joint distribution of $(X_{t_1}, \ldots, X_{t_m})$ is the same as that of $(X_{t_1+h}, \ldots, X_{t_m+h})$ for $h \geq 1$

Implications:

- X_1, X_2, X_3 have the same distribution
- (X_1, X_2) and (X_t, X_{t+1}) have the same joint distribution for $t \geq 1$,
- etc.

Stationary Processes

In some cases a weaker for of stationary suffices

Definition

A stochastic process $\{X_t\}_{t=-\infty}^{+\infty}$ is *covariance (or weakly or wide-sense) stationary* if

- $E(X_t) = \mu$ (does not vary with t)
- $var(X_t) = \gamma_0$ is constant,
- for any $j \geq 1$, $cov(X_t, X_{t-j}) = cov(X_t, X_{t+j}) = \gamma_j$ depends only on j and not on t.

Remark:

- Strong stationary does not imply weak stationary, though strong stationary + $E(X_t^2)$ finite implies weak stationary.
- Multivariate Normality+weak stationary⇒strong stationary

Remarks: In this case

- γ_j is denoted as the j *th* lag autocovariance
- $\rho_j = \gamma_j / \gamma_0 =$ is the *jth* lag autocorrelation.
- Henceforth stationary process will mean weakly stationary

- **Example:** The process $\{\varepsilon_t, t = 1, ...\}$ such as $E(\varepsilon_t) = 0$, $var(\varepsilon_t) = \sigma_\varepsilon^2$ and $cov(\varepsilon_t, \varepsilon_{t-j}) = 0, j \neq 0$, is known as a *white noise process* (it will be denoted as $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$. It is covariance stationary.)

Examples of nonstationary variables

1. $X_t = \beta t + \varepsilon_t, \varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$
   - In this case $E(X_t) = \beta t$ (hence nonstationary)
2. *The random walk*: $X_t = X_{t-1} + \varepsilon_t, \varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$, X_0 constant

Solving recursively we obtain $X_t = \sum_{j=1}^{t} \varepsilon_j + X_0$.

Thus $E(X_t) = X_0, var(X_t) = t\sigma_\varepsilon^2$ (hence nonstationary)

Wold's Decomposition Theorem

- The white noise process is the building block of the time series models that we are going to study.

Theorem

(Wold's Decomposition Theorem) Any covariance stationary process with mean zero can be represented as

$$X_t = \sum_{j=0}^{+\infty} \theta_j \varepsilon_{t-j} + v_t,$$

where $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$, $\theta_0 = 1$ *and* $\sum_{j=0}^{+\infty} \theta_j^2 < \infty$, $E(v_t \varepsilon_{t-j}) = 0$ *for all* j *and there exists constants* $\alpha_0, \alpha_1, \ldots$ *such that* $var(\sum_{j=0}^{\infty} \alpha_j v_{t-j}) = 0$.

- v_t is called deterministic component of X_t: It means as it can be predicted arbitrarily well from a linear function of past values of v_t.
- The term $\sum_{j=0}^{+\infty} \theta_j \varepsilon_{t-j}$ is called the indeterministic component of X_t.

Wold's Decomposition Theorem

- In practice it is usually assumed that we have a purely indeterministic process, i.e. $v_t = 0$ and try to approximate $\sum_{j=0}^{+\infty} \theta_j \varepsilon_{t-j}$.
- Obviously it is impossible to estimate $\sum_{j=0}^{+\infty} \theta_j \varepsilon_{t-j}$ because it requires the estimation of an infinite number of parameters $(\theta_1, \theta_2, ...)$.
- The traditional approach here is to approximate $\sum_{j=0}^{+\infty} \theta_j \varepsilon_{t-j}$, such that $\sum_{j=0}^{+\infty} \theta_j^2 < \infty$, by a parsimonious model that is a model with a small number of parameters.
- The most famous models are known as *Autoregressive Moving Average Models* (ARMA) (Box-Jenkins 1976).
- These models have as special cases the *Moving Average* (MA) and the *Autoregressive model* (AR).

The Lag operator L (or backshift operator) operates on an element of a time series to produce the previous element, that is

$$LX_t = X_{t-1}.$$

The lag operator can be raised to arbitrary integer powers so that if raised to the q power, we obtain

$$L^q X_t = X_{t-q}$$

Also If raised to the $-q$ power, we obtain

$$L^{-q} X_t = X_{t+q}.$$

Using this operator the first difference of x_t can be written as

$$\Delta X_t = X_t - X_{t-1} = (1 - L) X_t$$

The second difference is

$$\Delta^2 X_t = \Delta(\Delta X_t) = \Delta((1 - L) X_t) = (1 - L)^2 X_t$$

We can define a (finite or infinite order) polynomial in L or a filter according to:

$$a(L) = a_0 + a_1 L + a_2 L^2 + \dots$$

Thus

$$\begin{aligned} a(L)X_t &= a_0 X_t + a_1 L X_t + a_2 L^2 X_t + \dots \\ &= a_0 X_t + a_1 X_{t-1} + a_2 X_{t-2} + \dots \end{aligned}$$

Let $a(L)$ be a finite order polynomial in L. $a(L) = 1 - \sum_{i=1}^{p} a_i L^i$.We define $a(L)^{-1}$ to be the polynomial in L that satisfies

$$a(L)^{-1}a(L) = 1$$

That is,

$$a(L)^{-1}a(L)X_t = X_t$$

- $a(L)^{-1}$ will correspond to a series of the form $\sum_{i=0}^{\infty} b_i L^i$.

Example: Suppose

$$a(L) = 1 - \rho L.$$

Note that

$$(1 + \rho L + \rho^2 L^2 + \dots)(1 - \rho L) = 1$$

so $a(L)^{-1} = \sum_{i=0}^{+\infty} \rho^i L^i$.

- The coefficients of this infinite-order polynomial are absolutely summable if $\sum_{i=0}^{\infty} |b_i| < \infty$.
- Note that $\sum_{i=0}^{\infty} |b_i| < \infty \Rightarrow \sum_{i=0}^{\infty} b_i^2 < \infty$, that is absolute summability implies square summability.
- We will often be interested in inverses whose coefficients are absolutely summable:
  - The conditions that ensure that an inverse has absolutely summable coefficients (and therefore squared summable) play a crucial role in establishing necessary conditions for a time series to be stationary.

- A necessary and sufficient condition for an inverse to meet the absolute summability condition:
  - The characteristic roots of $a(z)$ lie outside the unit circle, where z is a complex variable.
  - That is, we have to find the zeros of the function $a(z)$. Denote one of them as z^*, for it to be outside the unit circle we must have $|z^*| > 1$.

Example: Suppose

$$a(L) = 1 - \rho L,$$

$a(L)^{-1} = \sum_{i=0}^{+\infty} \rho^i L^i$. To see if

$$\sum_{i=0}^{+\infty} \left|\rho^i\right|$$

it is convergent we have to compute the zeros of $a(z) = 1 - \rho z$. In this case it is $z^* = 1/\rho$. Thus we require $|z^*| > 1$ or $|\rho| < 1$.

Remarks:
Let

$$P(z) = a_n z^n + a_{n-1} z^{n-1} + ... + a_1 z + a_0$$

be a polynomial of degree n where a_0, a_1,a_2,..., a_n are constant coefficients.

- *Fundamental Theorem of Algebra:* every non-zero single-variable polynomial $P(z)$ of degree n has n values z_i for which $P(z_i) = 0$ (some of them possibly complex) .
- If $z = a + bi$, where a and b are real numbers and $i = \sqrt{-1}$, then $|z| = \sqrt{a^2 + b^2}$.
- the roots of a real-valued polynomial can occur in complex conjugate pairs, in which case we require their modulus to be greater than one. If $z = a + bi$ its complex conjugate is $\bar{z} = a - bi$

Example: Consider the operator

$$P(L) = (1 + L^2)$$

Does $P(L)$ have a absolutely summable inverse?

$$X_t = \sum_{j=0}^{q} \theta_j \varepsilon_{t-j}, \varepsilon_t \sim WN(0, \sigma_\varepsilon^2),\ \theta_0 = 1,$$

Or

$$X_t = \Theta(L)\varepsilon_t$$

where $\Theta(L) = \sum_{j=0}^{q} \theta_j L^j, \theta_0 = 1.$

- Notice that by the Wold decomposition theorem the true model of the data is a $MA(\infty)$.
- Here we approximate a $MA(\infty)$ process by a a $MA(q)$ process with q finite.

The $MA(q)$ is always stationary as

- $E(X_t) = 0$
- $\gamma_0 = \sum_{j=0}^{q} \theta_j^2 \sigma_\varepsilon^2$
- $\gamma_j = [\theta_j + \theta_{j+1}\theta_1 + \theta_{j+2}\theta_2 + ... + \theta_q\theta_{q-j}]\sigma_\varepsilon^2, j = 1, ..., q$
- $\gamma_j = 0, j > q$

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t, \varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$$

Or

$$\Phi(L)X_t = \varepsilon_t$$

where $\Phi(L) = 1 - \sum_{j=1}^{p} \phi_j L^j$.

- It can be shown that the $AR(p)$ is stationary if the roots of $\Phi(z)$ are outside the unit circle. Therefore $\Phi(L)$ has a a absolutely summable inverse.

This model corresponds to a $MA(\infty)$.

- **Example:** Consider the case $p = 1$

$$X_t = \phi_1 X_{t-1} + \varepsilon_t$$

- Notice that the model is equivalent to

$$\Phi(L)X_t = \varepsilon_t$$

where $\Phi(L) = 1 - \phi_1 L$. We know that $\Phi(L)$ has a absolutely summable inverse if $|\phi_1| < 1$ and it is equal to $(1 + \phi_1 L + \phi_1^2 L^2 + \ldots)$ thus multiplying both sides by $\Phi(L)^{-1}$ we have

$$\Phi(L)^{-1}\Phi(L)X_t = \Phi(L)^{-1}\varepsilon_t$$

or

$$\begin{aligned} X_t &= (1 + \phi_1 L + \phi_1^2 L^2 + \ldots)\varepsilon_t \\ &= \sum_{j=0}^{+\infty} \phi_1^j \varepsilon_{t-j} \end{aligned}$$

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=0}^{q} \theta_j \varepsilon_{t-j}, \varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$$

with $\theta_0 = 1$ or

$$\Phi(L) X_t = \Theta(L) \varepsilon_t$$

where $\Phi(L) = 1 - \sum_{j=1}^{p} \phi_j L^j$ and $\Theta(L) = \sum_{j=0}^{q} \theta_j L^j$.

- It can be shown that the $ARMA(p,q)$ is stationary if the roots of $\Phi(z)$ are outside the unit circle. Therefore $\Phi(L)$ has an absolutely summable inverse.

Under these conditions it equivalent to a $MA(\infty)$ model.
Example: For instance consider the model $ARMA(1,1)$

$$X_t = \phi_1 X_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

There are two ways to show that it is equivalent to a $MA(\infty)$ process.

- Using the method of undetermined coefficients.
- Using the operator L

Assume that the process is stationary thus

$$X_t = \sum_{i=0}^{\infty} \alpha_i \varepsilon_{t-i}$$

which implies that

$$X_{t-1} = \sum_{i=0}^{\infty} \alpha_i \varepsilon_{t-1-i}$$

The objective is to find the values of the coefficients $\alpha_0, \alpha_1, \alpha_2, ...$

- Replacing this in the equation above we have

$$\sum_{i=0}^{\infty} \alpha_i \varepsilon_{t-i} = \phi_1 \left(\sum_{i=0}^{\infty} \alpha_i \varepsilon_{t-1-i}\right) + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

- Now we match the coefficients of the terms containing $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, ...$ and obtaining
  - $\alpha_0 = 1$
  - $\alpha_1 = \phi_1 \alpha_0 + \theta_1 \Rightarrow \alpha_1 = \phi_1 + \theta_1$
  - $\alpha_i = \phi_1 \alpha_{i-1} \Rightarrow \alpha_i = \phi_1^{i-1} \alpha_1, i \geq 2$

We have

$$\Phi(L)X_t = \Theta(L)\varepsilon_t$$

with $\Phi(L) = 1 - \phi_1 L$ and $\Theta(L) = 1 + \theta_1 L$, thus assuming again that $|\phi_1| < 1$ we have

$$\begin{aligned} X_t &= \Phi^{-1}(L)\Theta(L)\varepsilon_t \\ &= (1 + \phi_1 L + \phi_1^2 L^2 + \ldots)(1 + \theta_1 L)\varepsilon_t \\ &= (1 + (\phi_1 + \theta_1)L + \phi_1(\phi_1 + \theta_1)L^2 + \ldots)\varepsilon_t \\ &= \textstyle\sum_{i=0}^{\infty} \alpha_i \varepsilon_{t-i} \end{aligned}$$

with

- $\alpha_0 = 1$
- $\alpha_1 = \phi_1 \alpha_0 + \theta_1 \Rightarrow \alpha_1 = \phi_1 + \theta_1$
- $\alpha_i = \phi_1 \alpha_{i-1} \Rightarrow \alpha_i = \phi_1^{i-1} \alpha_1, i \geq 2$

The $ARMA(p,q)$ process is said to be **invertible** if the roots of $\Theta(z) = 0$ lie outside the unit circle. Therefore $\Theta(L)$ has a a absolutely summable inverse.
If $\Theta(L)$ is invertible we can pre-multiply both sides by $\Theta(L)^{-1}$ to obtain

$$\begin{aligned} \Theta(L)^{-1}\Phi(L)X_t &= \varepsilon_t \\ \left(1 - \sum_{j=1}^{\infty} \alpha_j L^j\right) X_t &= \varepsilon_t \end{aligned}$$

for some coefficients $\alpha_1, \alpha_2, ...,$ which corresponds to a $AR(\infty)$ process.

ARMA Processes are well characterized by two functions

- The *autocorrelation function* (ACF):

$$\rho_j = \frac{\gamma_j}{\gamma_0}, j = 1, 2, \ldots$$

- The *partial autocorrelation function* (PACF).

- The partial autocorrelation function is more difficult to define. Informally it is defined as a measure of the association between X_t and X_{t-j} whilst taking away the effects of the variables $X_{t-1}, ..., X_{t-j+1}$ on this relationship (for $j \geq 1$).
- Formally the partial autocorrelation between X_t and X_{t-j} is defined as the coefficient of the variable X_{t-j} in the linear projection of X_t on $X_{t-1}, ..., X_{t-j}$.
- **Remark:** The linear projection of X_t on $X_{t-1}, ..., X_{t-j}$ is given by

$$P(X_t|X_{t-1}, ..., X_{t-j}) = \alpha_j^* + \beta_{j,1}^* X_{t-1} + ... + \beta_{j,j}^* X_{t-j},$$

where $\alpha_j^*, \beta_{j,1}^*, ..., \beta_{j,j}^*$ the values of $\alpha_j, \beta_{j,1}, ..., \beta_{j,j}$ that minimize

$$E[(X_t - \alpha_j - \beta_{j,1} X_{t-1} - ... - \beta_{j,j} X_{t-j})^2].$$

So the partial autocorrelation of order j is β_{jj}^*.

- Fortunately, there are formulas that allow us to compute the PACF in a easy way for any stochastic process which are given by the *Yule Walker* Equations:

$$
\begin{aligned}
\beta^*_{1,1} &= \rho_1 \\
\beta^*_{2,2} &= (\rho_2 - \rho_1^2)/(1 - \rho_1^2) \\
\beta^*_{j,j} &= \frac{\rho_j - \sum_{i=1}^{j-1} \beta^*_{j-1,1}\rho_{j-i}}{1 - \sum_{i=1}^{j-1} \beta^*_{j-1,1}\rho_j}, j = 3, 4, 5, ... \\
\beta^*_{j,i} &= \beta^*_{j-1,i} - \beta^*_{j,j}\beta^*_{j-i,j-i}, i = 1, 2, ..., j-1
\end{aligned}
$$

$$X_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$$

We would like to compute $\mu = E(X_t)$, $\gamma_0 = var(X_t)$, $\gamma_j = cov(X_t, X_{t-j})$ for $j \geq 1$.
Notice that $\mu = c$. and $\gamma_0 = (1 + \theta_1^2)\sigma_\varepsilon^2$. and

$$\begin{aligned} \gamma_1 &= E((X_t - c)(X_{t-1} - c)) \\ &= \theta_1 \sigma_\varepsilon^2 \end{aligned}$$

$$\begin{aligned} \gamma_j &= E((X_t - c)(X_{t-j} - c)) \\ &= 0, j > 1. \end{aligned}$$

thus the *ACF* is given by

$$\begin{aligned} \rho_1 &= \frac{\theta_1}{(1 + \theta_1^2)} \\ \rho_j &= 0, j > 1. \end{aligned}$$

PACF

Notice that $X_t = c + \Theta(L)\varepsilon_t$, assuming that $\Theta(L) = (1 + \theta_1 L)$ is invertible we have if $|\theta_1| < 1$

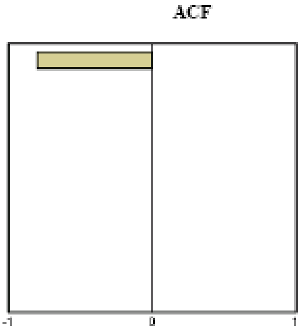
$$\Theta(L)^{-1} X_t = \Theta(L)^{-1} c + \varepsilon_t$$

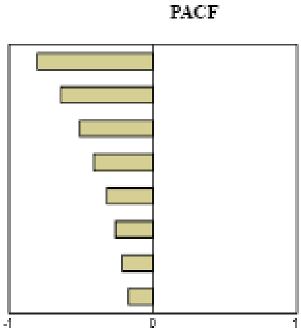
where $\Theta(L)^{-1} = (1 - \theta_1 L + \theta_1^2 L^2 - \theta_1^3 L^3 - ....)$.Thus

$$X_t = \frac{c}{1 + \theta_1} + \theta_1 X_{t-1} - \theta_1^2 X_{t-2} + \theta_1^3 X_{t-3} + ... + \varepsilon_t$$

Therefore X_t is correlated with all its lags. The PACF with exhibit a geometrically decaying pattern. If $\theta_1 < 0$ its decay is direct. If $\theta_1 > 0$ the PACF coefficients oscillate.

MA(1), $\theta_1 < 0$





Let us now consider the AR(1) with a constant.

$$X_t = c + \phi_1 X_{t-1} + \varepsilon_t, \varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$$

where $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$. We would like to compute $\mu = E(X_t)$, $\gamma_0 = var(X_t)$, $\gamma_j = cov(X_t, X_{t-j})$ for $j \geq 1$.
Assuming stationary we have

$$E(X_t) = \frac{c}{1 - \phi_1}.$$

and

$$var(X_t) = \frac{\sigma_\varepsilon^2}{1 - \phi_1^2} = \gamma_0$$

The auto-covariances are given by

$$\gamma_j = \phi_1 \gamma_{j-1} = \phi_1^j \gamma_0.$$

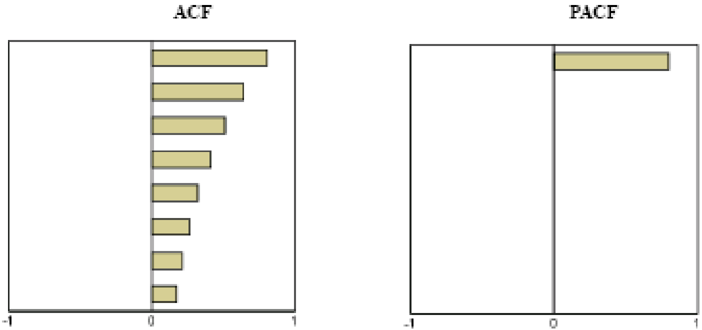
Thus the *ACF* is given by

$$\rho_j = \phi_1^j, j \geq 1$$

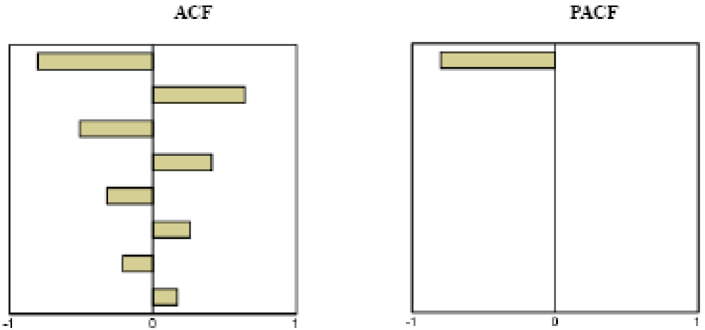
The *PACF* is given by

$$\begin{aligned} \beta_{1,1}^* &= \phi_1 \\ \beta_{j,j}^* &= 0, j > 1 \end{aligned}$$

AR(1) $\phi_1 > 0$



AR(1) $\phi_1 < 0$

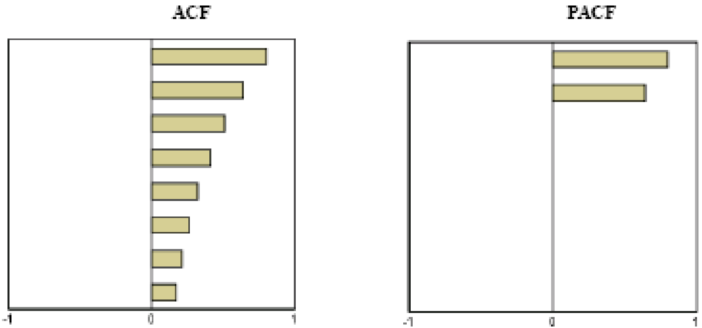


Let us now consider the AR(2) with a constant.

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t$$

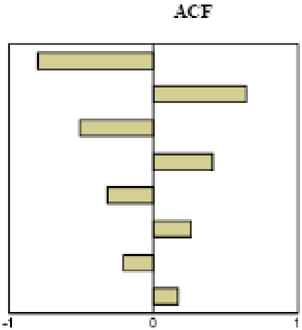
Properties

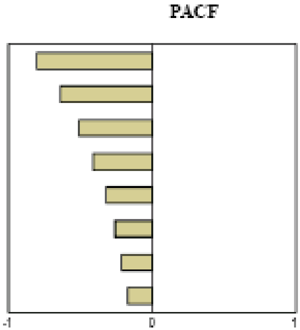
- The ACF in this case will be exponentially declining.
- AR(2) processes spike in the first two lags of the PACF and it will be equal to zero for lags bigger than two.



$$X_t = c + \phi_1 X_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$$

The ARMA(1,1) process shows exponential declines in both the ACF and the PACF.





- Autoregressive processes have an exponentially declining ACF and spikes in the first one or more lags of the PACF. The number of spikes in the PACF indicates the order of the autoregression.
- Moving average processes have spikes in the first one or more lags of the ACF and an exponentially declining PACF. The number of spikes in the ACF indicates the order of the moving average.
- Mixed (ARMA) processes typically show exponential declines in both the ACF and the PACF

Common factors

- Consider the $MA\ (\infty)$ representation of the ARMA(1,1) model (without a constant for simplicity.):

$$X_t = \sum_{j=0}^{\infty} \phi_1^j (\varepsilon_{t-j} + \theta_1 \varepsilon_{t-j-1})$$

which implies ACF

$$\rho_l = \phi_1^{l-1} \frac{(1+\phi_1\theta_1)(\phi_1+\theta_1)}{1+\theta_1^2+2\phi_1\theta_1}, l \geq 1.$$

- If $\theta_1 = -\phi_1 \Rightarrow \rho_l = 0$, $ARMA(1,1)$ reduces to white noise: AR and MA polynomials cancel in $(1-\phi_1 L)X_t = (1+\theta_1 L)\varepsilon_t$. Implies that ϕ_1 and θ_1 are not identified.
- Same occurs in $ARMA(p,q)$ models, if z^* is a zero of $\Phi(z)$ and $-z^*$ is a zero of $\Theta(z)$
- To avoid identification problems reduce the model to $ARMA(p-1, q-1)$.

Integrated Processes

Many time-series are non-stationary, but may have stationary first differences

Definition

If a stochastic process X_t must be differenced exactly d times to achieve stationarity, then the series is $I(d)$ and we write $X_t \sim I(d)$ (in words the process X_t is said to be integrated of order d.)

Remarks:

- A stationary series is $I(0)$
- $X_t \sim I(3) \Leftrightarrow \Delta^3 X_t = \Delta[\Delta(\Delta X_t)] \sim I(0)$.
- The random walk is $I(1)$: $X_t = X_{t-1} + \varepsilon_t$, $\varepsilon_t \sim WN(0, \sigma^2)$, thus $\Delta X_t = \varepsilon_t$.

ARIMA Models

- If $\Delta^d X_t$ follows a stationary and invertible ARMA(p,q) model,

$$\Phi(L)\Delta^d X_t = c + \Theta(L)\varepsilon_t$$

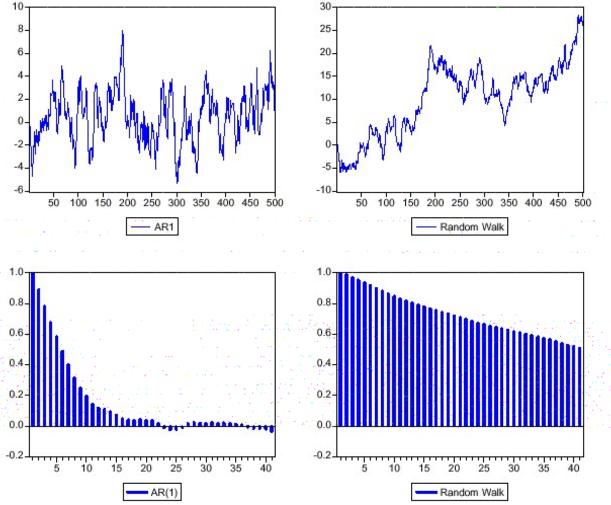
  with all roots of $\Phi(z)$ and $\Theta(z)$ outside the unit circle, then X_t follows an *autoregressive integrated moving average* model of order (p,d,q) denoted $ARIMA(p,d,q)$.
- The $ARIMA(p,d,q)$ is a non-stationary $ARMA(p+d,q)$ where the autoregressive polynomial $\Phi^*(L) = \Phi(L)(1-L)^d$ has d unit roots. Therefore testing procedures to determine d focus on the number of autoregressive unit roots.

Usually choice between $I(0)$ or $I(1)$. Main differences are:

- If $X_t \sim I(0)$, then
  - Shock ε_t has a transient decaying effect on X_{t+k} as $k \to \infty$;
  - X_t fluctuates around its mean, i.e. displays *mean-reversion*
  - ACF of X_t has either a cut-off point or decays exponentially.
- If $X_t \sim I(1)$, then
  - Shock ε_t has a *permanent or persistent* effect on X_{t+k} as $k \to \infty$;
  - X_t is not mean-reversion and *displays a (time-varying) trend*.
  - ACF of X_t is not defined, but sample ACF (defined later) stays close to one. Decays slowly (approximately linearly).

Stationary versus Integrated Processes

Examples of simulated AR(1) models with $\phi_1 = 0.9$ (left) and $\phi_1 = 1$ (right):



Box Jenkins Methodology

1. Transform the data, if necessary, so that the assumption of covariance stationarity is a reasonable one (E.g. take first differences.)
2. **Identification:** Make an initial guess for the values of p and q
3. Estimate the parameters of the proposed $ARMA(p, q)$ model
4. Perform diagnostic analysis to confirm that the proposed model adequately describes the data (e.g. examine residuals from fitted model)

Intuition: The autocorrelations and partial autocorrelations define the properties of an ARMA(p,q) model. A natural way to identify an ARMA model is to match the pattern of the observed sample autocorrelations (partial autocorrelations) with the patterns of the theoretical autocorrelations (partial autocorrelations) of a particular ARMA(p, q) model.

Sample autocovariances

$$\hat{\gamma}_j = \frac{1}{T}\sum_{t=j+1}^{T}(X_t - \bar{X})(X_{t-j} - \bar{X}), j \geq 0$$

Sample autocorrelations

$$\hat{\rho}_j = \frac{\hat{\gamma}_j}{\hat{\gamma}_0}, j = 1, 2, ...$$

Result: If X_t is **i.i.d**. for all t then $\rho_j = 0$ we have

$$\sqrt{T}\hat{\rho}_j \xrightarrow{D} N(0,1).$$

for all $j = 1, ..., k$. Thus to test $H_0 : \rho_j = 0$ vs $H_1 : \rho_j \neq 0$ we can use the statistic $\sqrt{T}\hat{\rho}_j$.
Rejection rule: Let $z_{\alpha/2}$ be the $100 \times \alpha\%$ critical value (that is the constant such that $\mathcal{P}(Z > z_{\alpha/2}) = \alpha/2$ where $Z \sim N(0,1)$) Reject H_0 in favour of H_1 if $\left|\sqrt{T}\hat{\rho}_j\right| > z_{\alpha/2}$. For instance for $\alpha = 0.05$ reject H_0 if $\left|\hat{\rho}_j\right| > 1.96/\sqrt{T}$

Remark: $\hat{\rho}_1, ..., \hat{\rho}_k$ are asymptotically independent (assuming that X_t is i.i.d.):

$$\sqrt{T}\begin{bmatrix} \hat{\rho}_1 \\ \vdots \\ \hat{\rho}_k \end{bmatrix} \xrightarrow{D} N(0, I_k)$$

- *Box-Pierce Portmonteau statistic:*

Let

$$Q_k = T\sum_{j=1}^{k} \hat{\rho}_j^2.$$

If X_t is i.i.d for all t, then $Q_k \xrightarrow{D} \chi^2(k)$.
$H_0 : \rho_j = 0, j = 1, ..., k$ vs H_0 : there is at least one $\rho_j \neq 0$.
Rejection Rule: Reject H_0 if $Q_k > c_\alpha$ where c_α is the $100 \times \alpha\%$ critical value (that is the constant such that $\mathcal{P}(X > c_\alpha) = \alpha$ where $X \sim \chi^2(k)$).

Identification of Stationary ARMA(p,q) Processes

- *Ljung and Box* showed that a simple degrees-of freedom adjustment improves the finite sample performance:

$$Q_k^* = T(T+2)\sum_{j=1}^{k}\frac{\hat{\rho}_j^2}{T-j}.$$

If X_t is i.i.d for all t, then $Q_k^* \xrightarrow{D} \chi^2(k)$ (same Rejection rule)

The jth order sample partial autocorrelation of X_t $\hat{\beta}_{jj}$ is the estimated coefficient of X_{t-j} in the regression of X_t on $X_{t-1}, X_{t-2}, ..., X_{t-j}$.for $j = 1, 2....$

- Result: If X_t is **i.i.d** for all t, then $\beta_{jj} = 0$ thus for all $j = 1, 2, ...$ we have

$$\sqrt{T}\hat{\beta}_{jj} \xrightarrow{D} N(0, 1).$$

  Thus to test $H_0 : \beta_{jj} = 0$ vs $H_1 : \beta_{jj} \neq 0$ we can use the statistic $\sqrt{T}\hat{\beta}_{jj}$.

- **Rejection rule:** Let $z_{\alpha/2}$ be the $100 \times \alpha\%$ critical value (that is the constant such that $\mathcal{P}(Z > z_{\alpha/2}) = \alpha/2$ where $Z \sim N(0, 1)$) Reject H_0 in favour of H_1 if $\left|\sqrt{T}\hat{\beta}_{jj}\right| > z_{\alpha/2}$. For instance for $\alpha = 0.05$ reject H_0 if $\left|\hat{\beta}_{jj}\right| > 1.96/\sqrt{T}$

Maximum likelihood estimation of ARMA models.

- For *i.i.d.* date the marginal pdf $f(x_t, \gamma)$, the joint pdf for a sample $(X_1, ..., X_T)$ is

$$f(x_1, ..., x_T; \gamma) \underset{independence}{=} \prod_{t=1}^{T} f(x_t; \gamma).$$

The likelihood function is this joint density treated as a function of the parameters given the random sample

$$\mathcal{L}(\gamma | X_1, ..., X_T) = \prod_{t=1}^{T} f(X_t; \gamma).$$

- The log-likelihood is given by

$$\log(\mathcal{L}(\gamma | X_1, ..., X_T)) = \sum_{t=1}^{T} \log f(X_t; \gamma).$$

- **Problem:** in time series

$$f(x_1, ..., x_T; \gamma) \neq \prod_{t=1}^{T} f(x_t; \gamma).$$

because the random variables in sample $(X_1, ..., X_T)$ are not iid.

One possible solution: Conditional factorization of the density function.
Intuition: Suppose that X_t only depends on X_{t-1} (as in a AR(1) process).
Consider the joint density of two adjacent observations $f(x_2, x_1; \gamma)$.
The joint density can always be factored as the product of the conditional density of x_2 given x_1 and the marginal density of x_1 :

$$f(x_2, x_1; \gamma) = f(x_2|x_1; \gamma) f(x_1; \gamma)$$

For three observations, the factorization becomes.

$$f(x_3, x_2, x_1; \gamma) = f(x_3|x_2, x_1; \gamma) f(x_2|x_1; \gamma) f(x_1; \gamma)$$

Continuing with this reasoning we have

$$f(x_T, ..., x_1; \gamma) = (\prod_{t=2}^{T} f(x_t|F_{t-1}; \gamma)) f(x_1; \gamma)$$

where $F_t = (x_t, ..., x_1) =$ information available at time t.

Maximum likelihood estimation of ARMA models

The *exact log-likelihood function:*

$$\begin{aligned}
\log \mathcal{L}(\gamma|X_1,...,X_T) &= \textstyle\sum_{t=2}^{T} \ell_t(\gamma) + \ell_1(\gamma) \\
\ell_t(\gamma) &= \log f(X_t|F_{t-1};\gamma) \\
\ell_1(\gamma) &= \log f(X_1;\gamma)
\end{aligned}$$

The *conditional log-likelihood:*

$$\log \mathcal{L}^*(\gamma|X_1,...,X_T) = \textstyle\sum_{t=2}^{T} \ell_t(\gamma)$$

In a $AR(1)$ process we have

$$X_1 \sim N(\frac{c}{1-\phi_1}, \frac{\sigma^2}{1-\phi_1^2})$$

Remark: The assumption of (unconditional) normality (gaussianity) is imposed.
Thus

$$f(x_1, \gamma) = \frac{1}{\sigma\sqrt{\frac{2\pi}{1-\phi_1^2}}} \exp\left\{-\frac{(x_1 - \frac{c}{1-\phi_1})^2}{\frac{2\sigma^2}{1-\phi_1^2}}\right\}$$

where $\gamma = (c, \phi_1, \sigma^2)'$.

Maximum likelihood estimation of ARMA models

We know that $X_t = c + \phi_1 X_{t-1} + \varepsilon_t$ if we assume that $\varepsilon_t \sim N(0, \sigma^2)$ and *i.i.d.*, then $X_t | X_{t-1} = x_{t-1} \sim N(c + \phi_1 x_{t-1}, \sigma^2)$.

$$f(x_t | \underbrace{x_{t-1}, ..., x_1}_{F_{t-1}}, \gamma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(x_t - c - \phi_1 x_{t-1})^2}{2\sigma^2} \right\}$$

- The log-likelihood function for the general ARMA(p,q) model can be constructed in a similar way.

Maximum likelihood estimation of ARMA models

- Two types of maximum likelihood estimates (mles) may be computed. The first type is based on maximizing the conditional log-likelihood function $\log \mathcal{L}^*(\gamma|X_1, ..., X_T)$. This estimator is called *conditional Maximum Likelihood Estimator (CML)* $[\hat{\gamma}_{CML}]$.
- The second type is based on maximizing the exact loglikelihood function $\log \mathcal{L}(\gamma|X_1, ..., X_T)$ and is called *exact Maximum Likelihood estimator (EML)*.
- It is possible to show that both estimators are consistent and asymptotically normal under some regularity conditions:

$$\sqrt{T}(\hat{\gamma}_{CML} - \gamma_0) \xrightarrow{D} N(0, A_0^{-1})$$

  where $A_0 = E[-\frac{\partial^2 \ell_t(\gamma_0)}{\partial\gamma\partial\gamma'}]$ and γ_0 are the true parameter values. Moreover, the CML and EML estimators are asymptotically equivalent.
- They will not yield the same estimates in finite samples.
- Inferences similar to the i.i.d. case.

Testing

Diagnostic testing involves checking if the residuals $\hat{\varepsilon}_t$ have white noise properties:

- Check the Sample autocorrelation function and Sample partial autocorrelation function of the residuals [check if the absolute values are bigger than $1.96/\sqrt{T^*}$ where $T^* = T - p$ (effective sample size)].
- Use the Box-Pierce and Ljung-Box Statistics applied to the residuals.
- **Alternative test:** Test for serial correlation using the (**Breusch Godfrey**) Lagrange multiplier statistic. **Example:** Test for white noise against *rth* order autocorrelation in the residuals in a $AR(p)$ model amounts to Test $\beta_1 = ...\beta_r = 0$ in auxiliary regression

$$\hat{\varepsilon}_t = \alpha_0 + \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + \beta_1 \hat{\varepsilon}_{t-1} + ... + \beta_r \hat{\varepsilon}_{t-r} + e_t,$$

where $\hat{\varepsilon}_t$ are the residuals of the model. Test Statistic.
$LM = T \times R^2 \xrightarrow{D} \chi^2(r)$ (R^2 of the auxiliary regression.)

Goodness of Fit

- Inspection of the SACF and SPACF to identify ARMA models is somewhat of an art rather than a science. A Less arbitrary procedure to identify an ARMA model is to use formal model selection criteria. The two most widely used criteria are the *Akaike information criterion* (AIC) and the *Bayesian (Schwarz) Information criterion (BIC or SIC)*. The usual definitions are:
  - $AIC(p,q) = \log(\hat{\sigma}) + 2\frac{(p+q)}{T^*}$, where $\hat{\sigma}$ is the estimate of σ_ε.
  - $BIC(p,q) = \log(\hat{\sigma}) + \frac{\log(T^*)(p+q)}{T^*}$. **(recommended)**
- Given several models we should choose the one having the lowest information criteria.

Interpretation:

- *Models with a good fit should have a low* $\log(\hat{\sigma})$
- $2\frac{(p+q)}{T^*}$ and $\frac{\log(T^*)(p+q)}{T^*}$ *penalize models with a large number of parameters.* Penalty of extra parameters is more severe in BIC.

Remark: Models with a large number of parameters have a poor forecast ability.

Forecasting

Preliminaries

- $F_s = \{X_s, X_{s-1}, ...\}$, information on process X_t up to s.
- **Conditional expectation** $E(X_t|F_s)$, $s < t$, is best (under squared error loss) predictor of X_t given F_s :

$$E((X_t - E(X_t|F_s))^2) \leq E((X_t - g(F_s)))^2)$$

  for *all functions* $g(F_s)$.
- **Best Linear predictor:** $P(X_t|F_s)$, $s < t$,

$$E((X_t - P(X_t|F_s))^2) \leq E((X_t - g(F_s)))^2)$$

  for *all linear functions* $g(F_s)$.

Some definitions

- A the process X_t is a **martingale** if $E(X_{t+1}|F_t) = X_t$, for all t.
- A the process Y_t is a **martingale difference sequence** if $E(Y_{t+1}|F_t) = 0$ for all t.
- **Remark:** If X_t is a martingale, $Z_t = X_t - X_{t-1}$ is a martingale difference sequence.

- Under the assumptions considered on the white noise process ε_t so far we are able to estimate $P(X_t|F_s)$ [$cov(\varepsilon_j, \varepsilon_i) = 0$ for $i \neq j$].
- If we assume that the errors are *i.i.d.* or a *martingale difference sequence,* that is

$$E(\varepsilon_{t+1}|F_t) = 0, \text{ for all t,}$$

we are able to estimate $E(X_t|F_s)$.

Useful properties of conditional expectations

- $E(E(X_t|F_s)) = E(X_t)$ (*law of iterated expectations*)
- $E(E(X_t|F_s)|F_r) = E(X_t|F_r)$ $(r < s < t)$ (*tower property*).

Let us write $E_t(X_{t+l}) = E(X_{t+l}|F_t)$ to simplify the notation.
Assuming that the errors ε_t are a *martingale difference sequence*, we can use these properties to show that the estimator for best forecast $E_t(X_{t+l})$ is given by

$$E_t(X_{t+l}) = c + \sum_{j=1}^{p} \phi_j E_t(X_{t+l-j}) + \sum_{j=1}^{q} \theta_j \varepsilon_t(l-j)$$

where $E_t(X_{t+l-j}) = X_{t+l-j}$ for $j \geq l$ and

$$\varepsilon_t(l-j) = \begin{cases} \varepsilon_{t+l-j} & j \geq l \\ 0 & j < l \end{cases}$$

Forecasting

Example: For an $ARMA(1,1)$ process

$$X_t = c + \phi_1 X_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}.$$

As $X_{t+1} = c + \phi_1 X_t + \varepsilon_{t+1} + \theta_1 \varepsilon_t$ we have

$$E_t(X_{t+1}) = c + \phi_1 X_t + \theta_1 \varepsilon_t$$

Also $X_{t+2} = c + \phi_1 X_{t+1} + \varepsilon_{t+2} + \theta_1 \varepsilon_{t+1}$ and

$$E_t(X_{t+2}) = c + \phi_1 E_t(X_{t+1})$$

- The variance of the prediction errors $e_t(l) = X_{t+l} - E_t(X_{t+l})$ is obtained from the $MA(\infty)$ representation of the $ARMA(p,q)$ process $X_t = E(X_t) + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$. Assuming that $var(\varepsilon_t) = \sigma_\varepsilon^2$ for all t, one can show that

$$var(e_t(l)) = \sigma_\varepsilon^2 \sum_{j=0}^{l-1} \psi_j^2$$

- For stationary process as $l \to \infty$ we have

$$E_t(X_{t+l}) \to E(X_t)$$

and

$$var(e_t(l)) \to var(X_t)$$

Forecasting

- In practice to make (out of sample) predictions we have to replace the unknown parameters by their estimators yielding

$$\hat{E}_T(X_{T+l}) = \hat{c} + \sum_{j=1}^{p} \hat{\phi}_j \hat{E}_T(X_{T+l-j}) + \sum_{j=1}^{q} \hat{\theta}_j \hat{\varepsilon}_T(l-j), l > 0$$

where $E_T(X_{t+l-j}) = X_{T+l-j}$ for $j \geq l$, $\hat{c}$, $\hat{\phi}_j$ and $\hat{\theta}_j$ are estimators of c, ϕ_j and θ_j and

$$\hat{\varepsilon}_t(l-j) = \begin{cases} \hat{\varepsilon}_{T+l-j} & j \geq l \\ 0 & j < l \end{cases},$$

where $\hat{\varepsilon}_t$, $t = p+1, ..., T$ are the residuals.

- For $I(1)$ processes the above methods are applied to ΔX_t, yielding $E(\Delta X_{t+l}|F_t)$. The forecasts of X_{t+l} are given by

$$E_t(X_{t+l}) = X_t + \sum_{i=1}^{l} E_t(\Delta X_{t+i}).$$

Comparison of the forecasts among different ARMA/ARIMA models:

- leave the last observations of the time series out of the estimation of the models,
- produce forecasts for these observations for each model;
- choose the model that yields the minimum value of the (sample) *mean squared prediction error* among the estimated models.