CHAPTER 5: Box-Jenkins (ARIMA) Forecasting

Prof. Alan Wan

Table of contents

Overview

- The Box-Jenkins methodology refers to a set of procedures for identifying and estimating time series models within the class of autoregressive integrated moving average (ARIMA) models.

Overview

- The Box-Jenkins methodology refers to a set of procedures for identifying and estimating time series models within the class of autoregressive integrated moving average (ARIMA) models.
- We speak also of AR models, MA models and ARMA models which are special cases of this general class.

Overview

- The Box-Jenkins methodology refers to a set of procedures for identifying and estimating time series models within the class of autoregressive integrated moving average (ARIMA) models.
- We speak also of AR models, MA models and ARMA models which are special cases of this general class.
- Models generalise regression but "explanatory" variables are past values of the series itself and unobservable random disturbances.

Overview

- The Box-Jenkins methodology refers to a set of procedures for identifying and estimating time series models within the class of autoregressive integrated moving average (ARIMA) models.
- We speak also of AR models, MA models and ARMA models which are special cases of this general class.
- Models generalise regression but "explanatory" variables are past values of the series itself and unobservable random disturbances.
- ARIMA models exploit information embedded in the autocorrelation pattern of the data.

Overview

- The Box-Jenkins methodology refers to a set of procedures for identifying and estimating time series models within the class of autoregressive integrated moving average (ARIMA) models.
- We speak also of AR models, MA models and ARMA models which are special cases of this general class.
- Models generalise regression but "explanatory" variables are past values of the series itself and unobservable random disturbances.
- ARIMA models exploit information embedded in the autocorrelation pattern of the data.
- Estimation is based on maximum likelihood; not least squares.

Overview

- The Box-Jenkins methodology refers to a set of procedures for identifying and estimating time series models within the class of autoregressive integrated moving average (ARIMA) models.
- We speak also of AR models, MA models and ARMA models which are special cases of this general class.
- Models generalise regression but "explanatory" variables are past values of the series itself and unobservable random disturbances.
- ARIMA models exploit information embedded in the autocorrelation pattern of the data.
- Estimation is based on maximum likelihood; not least squares.
- This method applies to both non-seasonal and seasonal data. In this chapter, we will only deal with non-seasonal data.

Overview

- Let $Y_1, Y_2, \cdots, Y_n$ denote a series of values for a time series of interest. These values are observable.

Overview

- Let $Y_1, Y_2, \cdots, Y_n$ denote a series of values for a time series of interest. These values are observable.
- Let $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ denote a series of random disturbances. These values are unobservable.

Overview

- Let $Y_1, Y_2, \cdots, Y_n$ denote a series of values for a time series of interest. These values are observable.
- Let $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ denote a series of random disturbances. These values are unobservable.
- The random disturbances (or noises) may be thought of as a series of random shocks that are completely uncorrelated with one another.

Overview

- Let $Y_1, Y_2, \cdots, Y_n$ denote a series of values for a time series of interest. These values are observable.
- Let $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ denote a series of random disturbances. These values are unobservable.
- The random disturbances (or noises) may be thought of as a series of random shocks that are completely uncorrelated with one another.
- Usually the noises are assumed to be generated from a distribution with identical mean 0 and identical variance σ^2 across all periods, and are uncorrelated with one another. They are called "white noises" (more on this later).

Overview

The three basic Box-Jenkins models for Y_t are:

1. Autoregressive model of order p ($AR(p)$):
   $Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \epsilon_t$
   (i.e., Y_t depends on its p previous values)
2. Moving average model of order q ($MA(q)$):
   $Y_t = \delta + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - ..... - \theta_q \epsilon_{t-q}$
   (i.e., Y_t depends on its q previous random error terms)
3. Autoregressive moving average model of orders p and q ($ARMA(p, q)$):
   $$\begin{aligned} Y_t = \delta &+ \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} \\ &+ \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - ..... - \theta_q \epsilon_{t-q} \end{aligned}$$
   (i.e., Y_t depends on its p previous values and q previous random error terms)

Overview

- Obviously, AR and MA models are special cases of ARMA.

Overview

- Obviously, AR and MA models are special cases of ARMA.
- ϵ_t is typically assumed to be "white noise"; i.e., it is identically and independently distributed with a common mean 0 and common variance σ^2 across all observations.

Overview

- Obviously, AR and MA models are special cases of ARMA.
- ϵ_t is typically assumed to be "white noise"; i.e., it is identically and independently distributed with a common mean 0 and common variance σ^2 across all observations.
- We write $\epsilon_t \sim i.i.d.(0, \sigma^2)$.

Overview

- Obviously, AR and MA models are special cases of ARMA.
- ϵ_t is typically assumed to be "white noise"; i.e., it is identically and independently distributed with a common mean 0 and common variance σ^2 across all observations.
- We write $\epsilon_t \sim i.i.d.(0, \sigma^2)$.
- The white noise assumption rules out possibilities of serial autocorrelation and heteroscedasticity in the disturbances.

Stationarity

- "Stationarity" is a fundamental property underlying almost all time series statistical models.

Stationarity

- "Stationarity" is a fundamental property underlying almost all time series statistical models.
- A time series Y_t is said to be strictly stationary if the joint distribution of $\{Y_1, Y_2, \cdots, Y_n\}$ is the same as that of $\{Y_{1+k}, Y_{2+k}, \cdots, Y_{n+k}\}$. That is, when we shift through time the behaviour of the random variables as characterised by the density function stays the same.

Stationarity

- "Stationarity" is a fundamental property underlying almost all time series statistical models.
- A time series Y_t is said to be strictly stationary if the joint distribution of $\{Y_1, Y_2, \cdots, Y_n\}$ is the same as that of $\{Y_{1+k}, Y_{2+k}, \cdots, Y_{n+k}\}$. That is, when we shift through time the behaviour of the random variables as characterised by the density function stays the same.
- Strict stationarity is difficult to fulfill or be tested in practice. Usually, when we speak of stationarity we refer to a weaker definition.

Stationarity

- A time series Y_t is said to be weakly stationary if it satisfies all of the following conditions:
    1. $E(Y_t) = \mu_y$ for all t
    2. $var(Y_t) = E[(Y_t - \mu_y)^2] = \sigma_y^2$ for all t
    3. $cov(Y_t, Y_{t-k}) = \gamma_k$ for all t

Stationarity

- A time series Y_t is said to be weakly stationary if it satisfies all of the following conditions:
  1. $E(Y_t) = \mu_y$ for all t
  2. $var(Y_t) = E[(Y_t - \mu_y)^2] = \sigma_y^2$ for all t
  3. $cov(Y_t, Y_{t-k}) = \gamma_k$ for all t
- A time series is thus weakly stationary if

  - its mean is the same at every period,

  - its variance is the same at every period, and

  - its autocovariance with respect to a particular lag is the same at every period.

Stationarity

- A time series Y_t is said to be weakly stationary if it satisfies all of the following conditions:
  1. $E(Y_t) = \mu_y$ for all t
  2. $var(Y_t) = E[(Y_t - \mu_y)^2] = \sigma_y^2$ for all t
  3. $cov(Y_t, Y_{t-k}) = \gamma_k$ for all t
- A time series is thus weakly stationary if
  - its mean is the same at every period,
  - its variance is the same at every period, and
  - its autocovariance with respect to a particular lag is the same at every period.
- A series of outcomes from independent identical trials is stationary, while a series with a trend cannot be stationary.

Stationarity

- With time series models, the sequence of observations is assumed to obey some sorts of dependence structure. Note that observations can be independent only in one way but they can be dependent in many different ways. Stationarity is one way of modeling the dependence structure.

Stationarity

- With time series models, the sequence of observations is assumed to obey some sorts of dependence structure. Note that observations can be independent only in one way but they can be dependent in many different ways. Stationarity is one way of modeling the dependence structure.
- It turns out that many useful results that hold under independence (e.g., the law of large numbers, Central Limit Theorem) also hold under the stationary dependence structure.

Stationarity

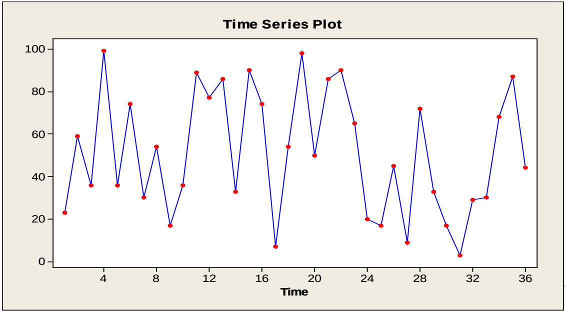
- With time series models, the sequence of observations is assumed to obey some sorts of dependence structure. Note that observations can be independent only in one way but they can be dependent in many different ways. Stationarity is one way of modeling the dependence structure.
- It turns out that many useful results that hold under independence (e.g., the law of large numbers, Central Limit Theorem) also hold under the stationary dependence structure.
- Without stationarity, the results can be spurious (e.g., the maximum likelihood estimators of the unknowns are inconsistent).

Stationarity

- The white noise series ϵ_t is stationary because
  1. $E(\epsilon_t) = 0$ for all t
  2. $var(\epsilon_t) = \sigma^2$ for all t
  3. $cov(\epsilon_t, \epsilon_{t-k}) = 0$ for all t and $k \neq 0$.

Stationarity

- The white noise series ϵ_t is stationary because
  1. $E(\epsilon_t) = 0$ for all t
  2. $var(\epsilon_t) = \sigma^2$ for all t
  3. $cov(\epsilon_t, \epsilon_{t-k}) = 0$ for all t and $k \neq 0$.
- Here is a time series plot of a white noise process:

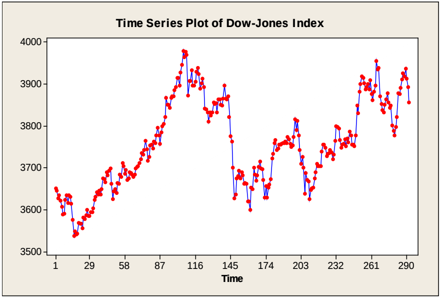


Stationarity

- However, not all time series are stationary. In fact, economic and financial time series are typically non-stationary.

Stationarity

- However, not all time series are stationary. In fact, economic and financial time series are typically non-stationary.
- Here is a time series plot of the Dow Jones Industrial Average Index. The series cannot be stationary as the trend rules out any possibility of a constant mean over time.



Stationarity

- Suppose that Y_t follows an AR(1) process without drift (an intercept). Is Y_t stationary?

Stationarity

- Suppose that Y_t follows an AR(1) process without drift (an intercept). Is Y_t stationary?
- Note that

$$
\begin{aligned}
Y_t &= \phi_1 Y_{t-1} + \epsilon_t \\
&= \phi_1(\phi_1 Y_{t-2} + \epsilon_{t-1}) + \epsilon_t \\
&= \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_1^2 \epsilon_{t-2} + \phi_1^3 \epsilon_{t-3} + \cdots + \phi_1^t Y_0
\end{aligned}
$$

- Without loss of generality, assume that $Y_0 = 0$.

Stationarity

- Hence $E(Y_t) = 0$.

Stationarity

- Hence $E(Y_t) = 0$.
- Assuming that the process started a long time ago (i.e., t is large) and $|\phi_1| < 1$, then it can be shown, for $t > k \geq 0$, that

$$\begin{aligned} var(Y_t) &= \frac{\sigma^2}{1-\phi_1^2} \\ cov(Y_t, Y_{t-k}) &= \frac{\phi_1^k \sigma^2}{1-\phi_1^2} = \phi_1^k var(Y_t) \end{aligned}$$

- That is, the mean, variance and covariances are all independent of t, provided that t is large and $|\phi_1| < 1$.
- When t is large, the necessary and sufficient condition of stationarity for an AR(1) process is $|\phi_1| < 1$.

Stationarity

- Consider the special case of $\phi_1 = 1$, i.e.,
$$Y_t = Y_{t-1} + \epsilon_t$$

Stationarity

- Consider the special case of $\phi_1 = 1$, i.e.,
$$Y_t = Y_{t-1} + \epsilon_t$$
- This process is known as a "random walk".

Stationarity

- Consider the special case of $\phi_1 = 1$, i.e.,
$$Y_t = Y_{t-1} + \epsilon_t$$
- This process is known as a "random walk".
- Now, assuming that $Y_0 = 0$, Y_t may be expressed equivalently as
$$Y_t = \sum_{j=0}^{t-1} \epsilon_{t-j}$$

Stationarity

- Consider the special case of $\phi_1 = 1$, i.e.,
$$Y_t = Y_{t-1} + \epsilon_t$$
- This process is known as a "random walk".
- Now, assuming that $Y_0 = 0$, Y_t may be expressed equivalently as
$$Y_t = \sum_{j=0}^{t-1} \epsilon_{t-j}$$
- Thus,
  1. $E(Y_t) = 0$ for all t
  2. $var(Y_t) = t\sigma^2$ for all t
  3. $cov(Y_t, Y_{t-k}) = (t-k)\sigma^2$ for all $t > k \geq 0$.

Stationarity

- Both the variance and covariance are dependent on t; the time series is thus non-stationary.

Stationarity

- Both the variance and covariance are dependent on t; the time series is thus non-stationary.
- Y_t's also vary more as t increases. That means when the data follows a random walk, the best prediction of the future is the present (a naive forecast) and the prediction becomes less accurate the further into the future we forecast.

Stationarity

- Suppose that the model is an AR(2) without drift:
  $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t$

Stationarity

- Suppose that the model is an AR(2) without drift:
  $$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t$$
- The necessary and sufficient conditions of stationarity for AR(2) are:
  $$\phi_1 + \phi_2 < 1, \;\; \phi_2 - \phi_1 < 1 \text{ and } |\phi_2| < 1$$

Stationarity

- Suppose that the model is an AR(2) without drift:
  $$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t$$
- The necessary and sufficient conditions of stationarity for AR(2) are:
  $$\phi_1 + \phi_2 < 1, \;\; \phi_2 - \phi_1 < 1 \text{ and } |\phi_2| < 1$$
- The algebraic complexity of the conditions increases with the order of the process. While these conditions can be generalised and do obey some kind of pattern, it is not necessary to learn the derivation of the conditions. The key point to note is that AR processes are not stationary unless appropriate conditions are imposed on the coefficients.

Stationarity

- Now, suppose that the model of interest is an MA(1) without drift:
  $$Y_t = \epsilon_t - \theta_1 \epsilon_{t-1}$$

Stationarity

- Now, suppose that the model of interest is an MA(1) without drift:
  $$Y_t = \epsilon_t - \theta_1 \epsilon_{t-1}$$
- It can be shown, regardless of the value of θ_1, that
  1. $E(Y_t) = 0 \quad \text{for all } t$
  2. $var(Y_t) = \sigma^2(1 + \theta_1^2) \quad \text{for all } t$
  3. $cov(Y_t, Y_{t-k}) = \begin{cases} -\theta_1 \sigma^2 & \text{if } k = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } t > k > 0$

Stationarity

- Now, suppose that the model of interest is an MA(1) without drift:

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

- It can be shown, regardless of the value of θ_1, that
  1. $E(Y_t) = 0$ for all t
  2. $var(Y_t) = \sigma^2(1+\theta_1^2)$ for all t
  3. $cov(Y_t, Y_{t-k}) = \begin{cases} -\theta_1\sigma^2 & \text{if } k = 1 \\ 0 & \text{otherwise} \end{cases}$ for all $t > k > 0$
- An MA(1) process is thus always stationary without the need to impose any condition on the unknown coefficient.

Stationarity

- For an MA(2) process without drift:
  $$Y_t = \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2}$$

Stationarity

- For an MA(2) process without drift:
  $$Y_t = \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2}$$
- 1. $E(Y_t) = 0 \quad \text{for all } t$
  2. $var(Y_t) = \sigma^2(1 + \theta_1^2 + \theta_2^2) \quad \text{for all } t$
  3. $$cov(Y_t, Y_{t-k}) = \begin{cases} -\theta_1(1-\theta_2)\sigma^2 & \text{if } \; k = 1 \\ -\theta_2\sigma^2 & \text{if } \; k = 2 \\ 0 & \text{otherwise} \end{cases} \quad \text{for all } t > k > 0$$

Stationarity

- For an MA(2) process without drift:
  $$Y_t = \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2}$$
- 1. $E(Y_t) = 0 \quad \text{for all } t$
  2. $var(Y_t) = \sigma^2(1 + \theta_1^2 + \theta_2^2) \quad \text{for all } t$
  3. $$cov(Y_t, Y_{t-k}) = \begin{cases} -\theta_1(1-\theta_2)\sigma^2 & \text{if } k = 1 \\ -\theta_2\sigma^2 & \text{if } k = 2 \\ 0 & \text{otherwise} \end{cases} \quad \text{for all } t > k > 0$$
- Again, an MA(2) process is stationary irrespective of the values of the unknown coefficients.

Invertibility

- MA processes are always stationary regardless of the values of the coefficients, but they are not necessarily invertible.

Invertibility

- MA processes are always stationary regardless of the values of the coefficients, but they are not necessarily invertible.
- A finite order MA process is said to be invertible if it can be converted into a stationary AR process of infinite order.

Invertibility

- MA processes are always stationary regardless of the values of the coefficients, but they are not necessarily invertible.
- A finite order MA process is said to be invertible if it can be converted into a stationary AR process of infinite order.
- As an example, consider an MA(1) process:

$$
\begin{aligned}
Y_t &= \epsilon_t - \theta_1 \epsilon_{t-1} \\
&= \epsilon_t - \theta_1 (Y_{t-1} + \theta_1 \epsilon_{t-2}) \\
&= \cdots \\
&= \epsilon_t - \theta_1 Y_{t-1} - \theta_1^2 Y_{t-2} - \theta_1^3 Y_{t-3} - \cdots - \theta_1^t \epsilon_0
\end{aligned}
$$

Invertibility

- MA processes are always stationary regardless of the values of the coefficients, but they are not necessarily invertible.
- A finite order MA process is said to be invertible if it can be converted into a stationary AR process of infinite order.
- As an example, consider an MA(1) process:

$$\begin{aligned}
Y_t &= \epsilon_t - \theta_1 \epsilon_{t-1} \\
&= \epsilon_t - \theta_1 (Y_{t-1} + \theta_1 \epsilon_{t-2}) \\
&= \cdots \\
&= \epsilon_t - \theta_1 Y_{t-1} - \theta_1^2 Y_{t-2} - \theta_1^3 Y_{t-3} - \cdots - \theta_1^t \epsilon_0
\end{aligned}$$

- In order for the MA(1) to be equivalent to an AR(∞), the last term on the r.h.s. of the above equation has to be zero.

Invertibility

- Assume that the process started a long time ago. With $|\theta_1| < 1$, we have $lim_{t\to\infty}\theta_1^t\epsilon_0 = 0$. So the condition $|\theta_1| < 1$ enables us to convert an MA(1) to an AR(∞).

Invertibility

- Assume that the process started a long time ago. With $|\theta_1| < 1$, we have $lim_{t \to \infty} \theta_1^t \epsilon_0 = 0$. So the condition $|\theta_1| < 1$ enables us to convert an MA(1) to an AR(∞).
- The conditions of invertibility for an MA(q) process are analogous to those of stationary for an AR(q) process.

Invertibility

- Assume that the process started a long time ago. With $|\theta_1| < 1$, we have $lim_{t \to \infty} \theta_1^t \epsilon_0 = 0$. So the condition $|\theta_1| < 1$ enables us to convert an MA(1) to an AR(∞).
- The conditions of invertibility for an MA(q) process are analogous to those of stationary for an AR(q) process.
- Now consider the following two MA(1) processes:
  1. $Y_t = \epsilon_t + 2\epsilon_{t-1}; \quad \epsilon_t \sim i.i.d.(0, 1)$
  2. $Y_t = \epsilon_t + (1/2)\epsilon_{t-1}; \quad \epsilon_t \sim i.i.d.(0, 4)$

Invertibility

- Assume that the process started a long time ago. With $|\theta_1| < 1$, we have $lim_{t \to \infty} \theta_1^t \epsilon_0 = 0$. So the condition $|\theta_1| < 1$ enables us to convert an MA(1) to an AR(∞).
- The conditions of invertibility for an MA(q) process are analogous to those of stationary for an AR(q) process.
- Now consider the following two MA(1) processes:
    1. $Y_t = \epsilon_t + 2\epsilon_{t-1}; \quad \epsilon_t \sim i.i.d.(0,1)$
    2. $Y_t = \epsilon_t + (1/2)\epsilon_{t-1}; \quad \epsilon_t \sim i.i.d.(0,4)$
- The second process is invertible but the first process is non-invertible. However, both processes generate the same mean, variance and covariances of Y_t's.

Invertibility

- For any non-invertible MA, there is always an equivalent invertible representation up to the second moment. The converse is also true. We prefer the invertible representation because if we can convert an MA process to an AR process, we can find the unobservable ϵ_t based on the past values of observable Y_t. If the process is non-invertible, then, in order to find the value of ϵ_t, we have to know all future values of Y_t that are unobservable at time t.
- To explain, note that
  $\epsilon_t = Y_t + \theta_1 Y_{t-1} + \theta_1^2 Y_{t-2} + \theta_1^3 Y_{t-3} + \cdots + \theta_1^t \epsilon_0$
  or
  $\epsilon_t = -1/\theta_1 Y_{t+1} - 1/\theta_1^2 Y_{t+2} - 1/\theta_1^3 Y_{t+3} + \cdots + 1/\theta_1^k \epsilon_{t+k}$,
  where $k > 0$.(write $Y_{t+1} = \epsilon_{t+1} - \theta_1 \epsilon_t$)

Invertibility

- Also, when expressing the most recent error as a combination of current and past observations by the AR(∞) representation, for an invertible process, $|\theta_1| < 1$, and so the most recent observation has a higher weight than any observation from the past. But when $|\theta_1| > 1$, the weights increase as the lags increase, so the more distant the observations the greater their influence on the current error. When $|\theta_1| = 1$, the weights are constant in size, and the distant observations have the same influence as the current observation. As neither of these make much sense, we prefer the invertible process.

Invertibility

- Also, when expressing the most recent error as a combination of current and past observations by the AR(∞) representation, for an invertible process, $|\theta_1| < 1$, and so the most recent observation has a higher weight than any observation from the past. But when $|\theta_1| > 1$, the weights increase as the lags increase, so the more distant the observations the greater their influence on the current error. When $|\theta_1| = 1$, the weights are constant in size, and the distant observations have the same influence as the current observation. As neither of these make much sense, we prefer the invertible process.
- Invertibility is a restriction programmed into time series software for estimating MA coefficients. It is not something that we check for data analysis.

Differencing

- Often non-stationary series can be made stationary through differencing.

Differencing

- Often non-stationary series can be made stationary through differencing.
- For example,
  1. $Y_t = Y_{t-1} + \epsilon_t$ is non-stationary, but
     $W_t = Y_t - Y_{t-1} = \epsilon_t$ is stationary
  2. $Y_t = 1.7Y_{t-1} - 0.7Y_{t-2} + \epsilon_t$ is non-stationary, but
     $W_t = Y_t - Y_{t-1} = 0.7W_{t-1} + \epsilon_t$ is stationary

Differencing

- Often non-stationary series can be made stationary through differencing.
- For example,
  1. $Y_t = Y_{t-1} + \epsilon_t$ is non-stationary, but
     $W_t = Y_t - Y_{t-1} = \epsilon_t$ is stationary
  2. $Y_t = 1.7Y_{t-1} - 0.7Y_{t-2} + \epsilon_t$ is non-stationary, but
     $W_t = Y_t - Y_{t-1} = 0.7W_{t-1} + \epsilon_t$ is stationary
- Let Δ be the difference operator and B the backward shift (or lag) operator such that $\Delta Y_t = Y_t - Y_{t-1}$ and $BY_t = Y_{t-1}$.

Differencing

- Often non-stationary series can be made stationary through differencing.
- For example,
  1. $Y_t = Y_{t-1} + \epsilon_t$ is non-stationary, but $W_t = Y_t - Y_{t-1} = \epsilon_t$ is stationary
  2. $Y_t = 1.7Y_{t-1} - 0.7Y_{t-2} + \epsilon_t$ is non-stationary, but $W_t = Y_t - Y_{t-1} = 0.7W_{t-1} + \epsilon_t$ is stationary
- Let Δ be the difference operator and B the backward shift (or lag) operator such that $\Delta Y_t = Y_t - Y_{t-1}$ and $BY_t = Y_{t-1}$.
- Thus,

  $$\Delta Y_t = Y_t - BY_t = (1 - B)Y_t$$

  and

  $$\begin{aligned}\Delta^2 Y_t = \Delta\Delta Y_t &= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\ &= Y_t - 2Y_{t-1} + Y_{t-2} = (1 - 2B + B^2)Y_t = (1 - B)^2 Y_t\end{aligned}$$

Differencing and Order of Integration

- In general, a d^{th} order difference can be written as $\Delta^d Y_t = (1 - B)^d Y_t$

Differencing and Order of Integration

- In general, a d^{th} order difference can be written as $\Delta^d Y_t = (1 - B)^d Y_t$
- The differenced series has $n - 1$ observations after taking the first difference, $n - 2$ observations after taking the second difference, and $n - d$ observations after taking d differences.

Differencing and Order of Integration

- In general, a d^{th} order difference can be written as $\Delta^d Y_t = (1 - B)^d Y_t$
- The differenced series has $n - 1$ observations after taking the first difference, $n - 2$ observations after taking the second difference, and $n - d$ observations after taking d differences.
- The number of times the original series must be differenced in order to achieve stationarity is called the order of integration, denoted by d.

Differencing and Order of Integration

- In general, a d^{th} order difference can be written as $\Delta^d Y_t = (1 - B)^d Y_t$
- The differenced series has $n - 1$ observations after taking the first difference, $n - 2$ observations after taking the second difference, and $n - d$ observations after taking d differences.
- The number of times the original series must be differenced in order to achieve stationarity is called the order of integration, denoted by d.
- In practice, it is seldom necessary to go beyond second difference, because real data generally involve only first or second level non-stationarity.

General ARIMA representation

- If Y_t is integrated of order d, we write $Y_t \sim I(d)$.

General ARIMA representation

- If Y_t is integrated of order d, we write $Y_t \sim I(d)$.
- Now, Suppose that $Y_t \sim I(d)$ and the stationary series after a d^{th} order differencing, W_t, is represented by an $ARMA(p, q)$ model.

General ARIMA representation

- If Y_t is integrated of order d, we write $Y_t \sim I(d)$.
- Now, Suppose that $Y_t \sim I(d)$ and the stationary series after a d^{th} order differencing, W_t, is represented by an $ARMA(p,q)$ model.
- Then we say that Y_t is an $ARIMA(p,d,q)$ process, that is,

$$\begin{aligned}(1-B)^d Y_t = W_t \\ &= \delta + \phi_1 W_{t-1} + \phi_2 W_{t-2} + \ldots.. + \phi_p W_{t-p} \\ &\quad + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \ldots.. - \theta_q \epsilon_{t-q}\end{aligned}$$

Autocorrelation Function

The question is, in practice, how can one tell if the data are stationary?

Autocorrelation Function

The question is, in practice, how can one tell if the data are stationary?

- Define the autocovariance at lag k as
  $\gamma_k = cov(Y_t, Y_{t-k})$.
  (Hence $\gamma_0 = var(Y_t)$).

Autocorrelation Function

The question is, in practice, how can one tell if the data are stationary?

- Define the autocovariance at lag k as
  $\gamma_k = cov(Y_t, Y_{t-k})$.
  (Hence $\gamma_0 = var(Y_t)$).
- Define the autocorrelation (AC) at lag k as
  $\rho_k = \frac{\gamma_k}{\gamma_0}$

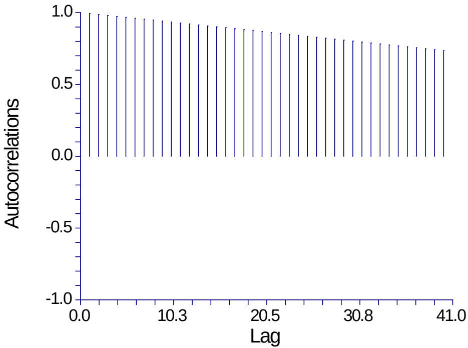
Autocorrelation Function

The question is, in practice, how can one tell if the data are stationary?

- Define the autocovariance at lag k as
  $\gamma_k = cov(Y_t, Y_{t-k})$.
  (Hence $\gamma_0 = var(Y_t)$).
- Define the autocorrelation (AC) at lag k as
  $\rho_k = \frac{\gamma_k}{\gamma_0}$
- Consequently, for a random walk process,
  $\rho_1 = (t-1)/t$,
  $\rho_2 = (t-2)/t$,
  .
  $\rho_k = (t-k)/t$

Autocorrelation Function

This produces the following autocorrelation function (ACF). The ACF dies down to zero extremely slowly as k increases.



Autocorrelation Function

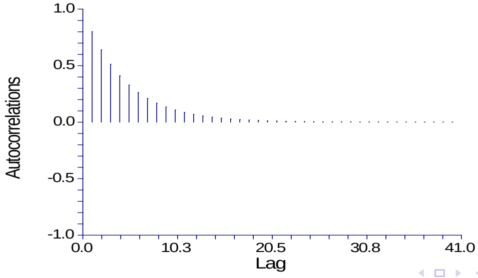
- Now, if the process is a stationary AR(1) process, i.e., $|\phi_1| < 1$. It can be easily verified that $\rho_k = \phi_1^k$.

Autocorrelation Function

- Now, if the process is a stationary AR(1) process, i.e., $|\phi_1| < 1$. It can be easily verified that $\rho_k = \phi_1^k$.
- So the ACF dies down to zero relatively quickly as k increases.

Autocorrelation Function

- Now, if the process is a stationary AR(1) process, i.e., $|\phi_1| < 1$. It can be easily verified that $\rho_k = \phi_1^k$.
- So the ACF dies down to zero relatively quickly as k increases.
- Suppose that $\phi_1 = 0.8$, then the ACF would look as follows:



Autocorrelation Function

- Now, consider an AR(2) process without drift:
  $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t$

Autocorrelation Function

- Now, consider an AR(2) process without drift:
  $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t$
- It can be shown that the AC coefficients are:
  $\rho_1 = \frac{\phi_1}{1-\phi_2}$,
  $\rho_2 = \phi_2 + \frac{\phi_1^2}{1-\phi_2}$, and
  $\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2}$ for $k > 2$.
- Hence the ACF dies down to zero according to a mixture of damped exponentials and/or damped sine waves.
- In general, the ACF of a stationary AR process dies down to zero as k increases.

Autocorrelation Function

- Consider an MA(1) process with no drift:
  $Y_t = \epsilon_t - \theta_1 \epsilon_{t-1}$

Autocorrelation Function

- Consider an MA(1) process with no drift: $Y_t = \epsilon_t - \theta_1 \epsilon_{t-1}$
- It can be easily shown that $$\rho_k = \frac{\gamma_k}{\gamma_0} = \begin{cases} \frac{-\theta_1}{1+\theta_1^2} & \text{if} \quad k = 1 \\ 0 & \text{otherwise} \end{cases}$$
- Hence the ACF cuts off at zero after lag 1.

Autocorrelation Function

- Similarly, for an MA(2) process, it can be shown that
  $\rho_1 = \frac{-\theta_1(1-\theta_2)}{1+\theta_1^2+\theta_2^2}$
  $\rho_2 = \frac{-\theta_2}{1+\theta_1^2+\theta_2^2}$
  $\rho_k = 0$ for $k > 2$
- The ACF of an MA(2) process thus cuts off at zero after 2 lags.

Autocorrelation Function

- In general, if a time series is non-stationary, its ACF dies down to zero slowly and the first autocorrelation is near 1. If a time series is stationary, the ACF dies down to zero relatively quickly in the case of AR, cuts off after certain lags in the case of MA, and dies down to zero relatively quickly in the case of ARMA (as the dying down pattern produced by the AR component would dominate the cutting off pattern produced by the MA component).

Autocorrelation Function

- Question: How are the AC coefficients estimated in practice?

Autocorrelation Function

- Question: How are the AC coefficients estimated in practice?
- The sample AC at lag k is calculated as follows:

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{\sum_{t=1}^{n}(Y_t - \overline{Y})^2},$$

where $\overline{Y} = \frac{\sum_{t=1}^{n} Y_t}{n}$.

Autocorrelation Function

- Question: How are the AC coefficients estimated in practice?
- The sample AC at lag k is calculated as follows:

  $$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{\sum_{t=1}^{n}(Y_t - \overline{Y})^2},$$

  where $\overline{Y} = \frac{\sum_{t=1}^{n} Y_t}{n}$.
- Thus, r_k measures the linear association between the time series observations separated by a lag of k time units in the sample, and is an estimator of ρ_k.

Autocorrelation Function

- Question: How are the AC coefficients estimated in practice?
- The sample AC at lag k is calculated as follows:

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{\sum_{t=1}^{n}(Y_t - \overline{Y})^2},$$

  where $\overline{Y} = \frac{\sum_{t=1}^{n} Y_t}{n}$.
- Thus, r_k measures the linear association between the time series observations separated by a lag of k time units in the sample, and is an estimator of ρ_k.
- In other words, computing the sample ACs is similar to performing a series of simple regressions of Y_t on Y_{t-1}, then on Y_{t-2}, then on Y_{t-3}, and so on. The autocorrelation coefficients reflect only the relationship between the two quantities included in the regression.

Autocorrelation Function

- The standard error of r_k is $s_{r_k} = \sqrt{\frac{1+2\sum_{j=1}^{k-1} r_j^2}{n}}$.

Autocorrelation Function

- The standard error of r_k is $s_{r_k} = \sqrt{\frac{1+2\sum_{j=1}^{k-1} r_j^2}{n}}$.
- To test $H_0 : \rho_k = 0$ vs. $H_0 : \rho_k \neq 0$, we use the statistic

$$t = \frac{r_k}{s_{r_k}}$$

Autocorrelation Function

- The standard error of r_k is $s_{r_k} = \sqrt{\frac{1+2\sum_{j=1}^{k-1} r_j^2}{n}}$.
- To test $H_0 : \rho_k = 0$ vs. $H_0 : \rho_k \neq 0$, we use the statistic

  $$t = \frac{r_k}{s_{r_k}}$$

- The rule of thumb is to reject H_0 at an approximately 5% level of significance if

  $$|t| > 2,$$

  or equivalently,

  $$|r_k| > 2s_{r_k}.$$

Autocorrelation Function

- Sample ACF of a stationary AR(1) process:

```
                              The ARIMA Procedure
                                 Name of Variable = y
                            Mean of Working Series    -0.08047
                            Standard Deviation         1.123515
                            Number of Observations           99
                                     Autocorrelations
Lag   Covariance   Correlation  -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1     Std Error
 0     1.262285     1.00000     |                    |********************|            0
 1     0.643124     0.50949     |                 .  |**********          |     0.100504
 2     0.435316     0.34486     |                .   |*******             |     0.123875
 3     0.266020     0.21075     |                .   |****.               |     0.133221
 4     0.111942     0.08868     |                .   |**  .               |     0.136547
 5     0.109251     0.08655     |                .   |**  .               |     0.137127
 6     0.012504     0.00991     |               .    |     .              |     0.137678
 7    -0.040513     -.03209     |               .   *|     .              |     0.137685
 8    -0.199299     -.15789     |               . ***|     .              |     0.137761
 9    -0.253309     -.20067     |               .****|     .              |     0.139576

                        "." marks two standard errors
```

Autocorrelation Function

- Sample ACF of an invertible MA(2) process:

```
                              The ARIMA Procedure
                              Name of Variable = y
                       Mean of Working Series     0.020855
                       Standard Deviation         1.168993
                       Number of Observations           98
                                Autocorrelations
 Lag    Covariance    Correlation   -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1      Std Error
   0      1.366545        1.00000   |                    |********************|            0
   1     -0.345078        -.25252   |               *****|   .                |     0.101015
   2     -0.288095        -.21082   |                ****|   .                |     0.107263
   3     -0.064644        -.04730   |                .  *|   .                |     0.111411
   4      0.160680        0.11758   |                .   |**  .               |     0.111616
   5     0.0060944        0.00446   |               .    |    .               |     0.112873
   6     -0.117599        -.08606   |               .  **|    .               |     0.112875
   7     -0.104943        -.07679   |               .  **|    .               |     0.113542
   8      0.151050        0.11053   |               .    |**  .               |     0.114071
   9      0.122021        0.08929   |               .    |**  .               |     0.115159
                            "." marks two standard errors
```

Autocorrelation Function

- Sample ACF of a random walk process:

```
                                      The ARIMA Procedure
                                     Name of Variable = y
                              Mean of Working Series     16.79147
                              Standard Deviation          9.39551
                              Number of Observations           98
                                       Autocorrelations
Lag    Covariance    Correlation    -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1      Std Error
  0     88.275614        1.00000     |                   |********************|             0
  1     85.581769        0.96948     |               .   |******************* |      0.101015
  2     81.637135        0.92480     |            .      |******************  |      0.171423
  3     77.030769        0.87262     |          .        |*****************   |      0.216425
  4     72.573174        0.82212     |         .         |****************    |      0.249759
  5     68.419227        0.77506     |        .          |****************    |      0.275995
  6     64.688289        0.73280     |       .           |***************     |      0.297377
  7     61.119745        0.69237     |      .            |**************      |      0.315265
  8     57.932253        0.65627     |      .            |*************       |      0.330417
  9     55.302847        0.62648     |     .             |*************.      |      0.343460
                 "." marks two standard errors
```

Autocorrelation Function

- Sample ACF of a random walk process after first difference:

```
                           The ARIMA Procedure
                          Name of Variable = y
                Period(s) of Differencing                           1
                Mean of Working Series                       0.261527
                Standard Deviation                           1.160915
                Number of Observations                             97
                Observation(s) eliminated by differencing           1
                             Autocorrelations
Lag    Covariance    Correlation    -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1      Std Error
  0      1.347723        1.00000    |                    |********************|             0
  1      0.712219        0.52846    |                .   |***********         |      0.101535
  2      0.263094        0.19521    |               .    |****.               |      0.126757
  3     -0.043040        -.03194    |               .   *|    .               |      0.129820
  4     -0.151081        -.11210    |               .  **|    .               |      0.129901
  5     -0.247540        -.18367    |               .****|    .               |      0.130894
  6     -0.285363        -.21174    |               .****|    .               |      0.133525
  7     -0.274084        -.20337    |               .****|    .               |      0.136943
  8     -0.215508        -.15991    |              .  ***|     .              |      0.140022
  9     -0.077629        -.05760    |              .    *|     .              |      0.141892
                              "." marks two standard errors
```

Partial Autocorrelation Function

- The partial autocorrelation (PAC) at lag k, denoted as r_{kk}, measures the degree of association between Y_t and Y_{t-k} when the effects of other time lags (1,2,3,...,k-1) are removed. Intuitively, the PAC may be thought of as the AC of time series observations separated by k time units with the effects of the intervening observations eliminated. The PACs at lags 1, 2, 3,... make up the partial autocorrelation function (PACF).

Partial Autocorrelation Function

- The partial autocorrelation (PAC) at lag k, denoted as r_{kk}, measures the degree of association between Y_t and Y_{t-k} when the effects of other time lags (1,2,3,...,k-1) are removed. Intuitively, the PAC may be thought of as the AC of time series observations separated by k time units with the effects of the intervening observations eliminated. The PACs at lags 1, 2, 3,... make up the partial autocorrelation function (PACF).
- Computing the PACs is more in the spirit of multiple regression. The PAC removes the effects of all lower order lags before computing the autocorrelation. For example, the 2nd PAC is the effects of the observation two periods ago on the current observation, given that the effect of the observation one period ago has been removed.

Partial Autocorrelation Function

- Thus, if the model is an AR(2), in theory the first two PACs would be non-zero and all other PACs would be zero.

Partial Autocorrelation Function

- Thus, if the model is an AR(2), in theory the first two PACs would be non-zero and all other PACs would be zero.
- In general, the PACF of a stationary AR(p) process would cut off at zero after lag p.

Partial Autocorrelation Function

- Thus, if the model is an AR(2), in theory the first two PACs would be non-zero and all other PACs would be zero.
- In general, the PACF of a stationary AR(p) process would cut off at zero after lag p.
- Because an invertible finite order MA process can be written as a stationary AR(∞), the PACF of a MA process would die down to zero in the same way as the ACF of the analogous AC process dies down to zero.

Partial Autocorrelation Function

- Thus, if the model is an AR(2), in theory the first two PACs would be non-zero and all other PACs would be zero.
- In general, the PACF of a stationary AR(p) process would cut off at zero after lag p.
- Because an invertible finite order MA process can be written as a stationary AR(∞), the PACF of a MA process would die down to zero in the same way as the ACF of the analogous AC process dies down to zero.
- It is also clear the AC and PAC at lag 1 of a given process are identical.

Partial Autocorrelation Function

- The sample PAC at lag k is:

$$r_{kk} = \begin{cases} r_1 & \text{if} \quad k = 1 \\ \frac{r_k - \sum_{j=1}^{k-1} r_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j} r_k}, & \text{if} \quad k > 1 \end{cases}$$

where $r_{kj} = r_{k-1,j} - r_{kk} r_{k-1,k-j}$ for $j = 1, 2, \cdots, k-1$.

Partial Autocorrelation Function

- The sample PAC at lag k is:

$$r_{kk} = \begin{cases} r_1 & \text{if} \quad k = 1 \\ \frac{r_k - \sum_{j=1}^{k-1} r_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j} r_k}, & \text{if} \quad k > 1 \end{cases}$$

  where $r_{kj} = r_{k-1,j} - r_{kk} r_{k-1,k-j}$ for $j = 1, 2, \cdots, k-1$.

- The standard error of r_{kk} is $s_{r_{kk}} = \sqrt{\frac{1}{n}}$.

Partial Autocorrelation Function

- The sample PAC at lag k is:
$$r_{kk} = \begin{cases} r_1 & \text{if} \quad k = 1 \\ \frac{r_k - \sum_{j=1}^{k-1} r_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j} r_k}, & \text{if} \quad k > 1 \end{cases}$$
where $r_{kj} = r_{k-1,j} - r_{kk} r_{k-1,k-j}$ for $j = 1, 2, \cdots, k-1$.
- The standard error of r_{kk} is $s_{r_{kk}} = \sqrt{\frac{1}{n}}$.
- The statistic for testing $H_0 : \rho_{kk} = 0$ vs. $H_0 : \rho_{kk} \neq 0$ is

$$t = \frac{r_{kk}}{s_{r_{kk}}}$$

Partial Autocorrelation Function

- The sample PAC at lag k is:
$$r_{kk} = \begin{cases} r_1 & \text{if} \quad k = 1 \\ \frac{r_k - \sum_{j=1}^{k-1} r_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j} r_k}, & \text{if} \quad k > 1 \end{cases}$$
where $r_{kj} = r_{k-1,j} - r_{kk} r_{k-1,k-j}$ for $j = 1, 2, \cdots, k-1$.
- The standard error of r_{kk} is $s_{r_{kk}} = \sqrt{\frac{1}{n}}$.
- The statistic for testing $H_0 : \rho_{kk} = 0$ vs. $H_0 : \rho_{kk} \neq 0$ is
$$t = \frac{r_{kk}}{s_{r_{kk}}}$$
- The rule of thumb is to reject H_0 at an approximately 5% level of significance if
$$|t| > 2, \text{ or equivalently, } |r_{kk}| > 2 s_{r_{kk}}.$$

Partial Autocorrelation Function

- Sample PACF of a stationary AR(2) process:

```
                              Partial Autocorrelations

Lag    Correlation    -10 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 10

1        0.23561      |                  . |*****               |
2       -0.48355      |          **********| .                  |
3       -0.04892      |                  .*| .                  |
4        0.07474      |                  . |*.                  |
5       -0.00086      |                  . | .                  |
6       -0.03688      |                  .*| .                  |
7        0.00335      |                  . | .                  |
8        0.05250      |                  . |*.                  |
9       -0.07018      |                  .*| .                  |
10      -0.06710      |                  .*| .                  |
```

Partial Autocorrelation Function

- Sample PACF of an invertible MA(1) process:

```
                        Partial Autocorrelations

 Lag    Correlation   -10 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 10

   1        0.47462    |                   . |*********           |
   2       -0.28275    |               ******|  .                 |
   3        0.22025    |                   . |****                |
   4       -0.08660    |                   **|  .                 |
   5        0.07201    |                   . |*.                  |
   6       -0.02120    |                   . |  .                 |
   7        0.06986    |                   . |*.                  |
   8       -0.06343    |                   .*|  .                 |
   9        0.00465    |                   . |  .                 |
  10       -0.03059    |                   .*|  .                 |
```

Partial Autocorrelation Function

Behaviour of AC and PAC for specific non-seasonal ARMA models:

Model	AC	PAC
AR(1)	Dies down in a damped exponential fashion	cuts off after lag 1
AR(2)	Dies down according to a mixture of damped exponentials and/or damped sine waves	cuts off after lag 2
MA(1)	cuts off after lag 1	Dies down in a damped exponential fashion
MA(2)	cuts off after lag 2	Dies down according to a mixture of damped exponentials and/or damped sine waves
ARMA(1,1)	Dies down in a damped exponential fashion	Dies down in a damped exponential fashion

Model Identification

- The ACF and PACF give insights into what models to fit to the data.

Model Identification

- The ACF and PACF give insights into what models to fit to the data.
- Refer to Class Examples 1, 2 and 3 for the sample ACF and PACF of simulated AR(2), MA(1) and ARMA(2,1) processes respectively.

Model Identification

- We work backwards in identifying the underlying ARMA model for a time series: we understand the theoretical properties of the ACF and PACF of a given ARMA process; if the sample ACF and PACF from the data have recognisable patterns then we will fit the ARMA model that would produce those ACF and PACF patterns to the data.

Model Identification

- We work backwards in identifying the underlying ARMA model for a time series: we understand the theoretical properties of the ACF and PACF of a given ARMA process; if the sample ACF and PACF from the data have recognisable patterns then we will fit the ARMA model that would produce those ACF and PACF patterns to the data.
- Note that the ARMA model is unduly restrictive - linear in coefficients, white noise error assumption, etc. In practice, no data series is generated exactly by an ARMA process. Hence we look for the best ARMA approximation to the real data only.

Parameter Estimation

- The method that is frequently used for estimating unknowns in ARMA model is maximum likelihood (M.L.).

Parameter Estimation

- The method that is frequently used for estimating unknowns in ARMA model is maximum likelihood (M.L.).
- Given n observations $Y_1, Y_2, \cdots, Y_n$, the likelihood function L is defined to be the probability of obtaining the data actually observed. It is really just the joint density function of the data.

Parameter Estimation

- The method that is frequently used for estimating unknowns in ARMA model is maximum likelihood (M.L.).
- Given n observations $Y_1, Y_2, \cdots, Y_n$, the likelihood function L is defined to be the probability of obtaining the data actually observed. It is really just the joint density function of the data.
- The maximum likelihood estimator (M.L.E.) are those values of the parameters that would lead to the highest probability of producing the data actually observed; that is, they are the values of the unknowns that maximise the likelihood function L.

Parameter Estimation

- In an ARMA model, L is a function of δ, ϕ's, θ's and σ^2 given the data $Y_1, Y_2, \cdots, Y_n$. One would also need to make an assumption of the distribution of the data.

Parameter Estimation

- In an ARMA model, L is a function of δ, ϕ's, θ's and σ^2 given the data $Y_1, Y_2, \cdots, Y_n$. One would also need to make an assumption of the distribution of the data.
- The M.L.E. of these unknowns are their values that make the observation of $Y_1, Y_2, \cdots, Y_n$ a most likely event; that is, assuming that the data come from a particular distribution (e.g., Gaussian), it is most likely that the unknown parameters take on the values of the M.L.E. in order for the data $Y_1, Y_2, \cdots, Y_n$ to be observed.

Parameter Estimation

Estimation of MA processes. Consider an MA(1) process: $Y_t = \delta + \epsilon_t - \theta_1 \epsilon_{t-1}$.

- Assume $\epsilon_0 = 0$.
- Use $\hat{\delta} = \bar{Y}$ and $r_1 = -\hat{\theta}_1/(1 + \hat{\theta}_1^{\,2})$ to obtain initial estimates of δ and θ_1.
- Obtain series of observations of ϵ_t's using the relation: $\epsilon_t = Y_t - \hat{\delta} + \hat{\theta}_1 \epsilon_{t-1}$.
- Use M.L.E. to obtain improved estimates of δ and θ_1.
- Repeat steps until differences in estimates are small.

Parameter Estimation

- Before estimating any ARMA model, one would typically test if the drift term should be included. A test of
$$H_0 : \delta = 0 \text{ vs. } H_1 : \delta \neq 0$$
may be conducted using the statistic
$$t = \frac{\bar{z}}{s_z/\sqrt{n'}},$$
where $\bar{z}$ is the mean of the working series z, s_z is the s.d. of z, and n' is the number of observations of the working series.

Parameter Estimation

- Before estimating any ARMA model, one would typically test if the drift term should be included. A test of
$$H_0 : \delta = 0 \text{ vs. } H_1 : \delta \neq 0$$
may be conducted using the statistic
$$t = \frac{\bar{z}}{s_z/\sqrt{n'}},$$
where $\bar{z}$ is the mean of the working series z, s_z is the s.d. of z, and n' is the number of observations of the working series.
- For example, if the series is $I(0)$, then $n' = n$; if the series is $I(1)$, then $n' = n - 1$.

Parameter Estimation

- Before estimating any ARMA model, one would typically test if the drift term should be included. A test of
  $$H_0 : \delta = 0 \text{ vs. } H_1 : \delta \neq 0$$
  may be conducted using the statistic
  $$t = \frac{\bar{z}}{s_z/\sqrt{n'}},$$
  where $\bar{z}$ is the mean of the working series z, s_z is the s.d. of z, and n' is the number of observations of the working series.
- For example, if the series is $I(0)$, then $n' = n$; if the series is $I(1)$, then $n' = n - 1$.
- The decision rule is to reject H_0 at an approximately 5% level of significance if $|t| > 2$.

Parameter Estimation

- Refer to the MA(2) example seen before. Here, $t = 0.020855/(1.168993/\sqrt{98}) = 0.176608 < 2$. Thus, the drift term should not be included.

```
                          The ARIMA Procedure
                          Name of Variable = y
                   Mean of Working Series    0.020855
                   Standard Deviation        1.168993
                   Number of Observations          98
                            Autocorrelations
Lag   Covariance   Correlation   -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1     Std Error
  0     1.366545     1.00000     |                     |********************|            0
  1    -0.345078     -.25252     |                *****|   .                |     0.101015
  2    -0.288095     -.21082     |                 ****|   .                |     0.107263
  3    -0.064644     -.04730     |                 .  *|   .                |     0.111411
  4     0.160680     0.11758     |                 .   |** .                |     0.111616
  5    0.0060944     0.00446     |                .    |    .               |     0.112873
  6    -0.117599     -.08606     |                .  **|    .               |     0.112875
  7    -0.104943     -.07679     |                .  **|    .               |     0.113542
  8     0.151050     0.11053     |                .    |**  .               |     0.114071
  9     0.122021     0.08929     |                .    |**  .               |     0.115159
                     "." marks two standard errors
```

Diagnostic Checking

- Often it is not straightforward to determine a single ARMA model that most adequately represents the data generating process, and it is not uncommon to identify several candidate models at the identification stage. The model that is finally selected is the one considered best based on a set of diagnostic checking criteria. These include:
  1. t-tests for coefficient significance
  2. Residual portmanteau test
  3. AIC and BIC for model selection

Diagnostic Checking

- Consider the data series of Class Example 4.
  - First, the data appear to be stationary;
  - Second, the drift term is significant;
  - Third, the sample ACF and PACF indicate that an AR(2) model probably best fits the data.

Diagnostic Checking

- Consider the data series of Class Example 4.
  - First, the data appear to be stationary;
  - Second, the drift term is significant;
  - Third, the sample ACF and PACF indicate that an AR(2) model probably best fits the data.
- For purposes of illustration, suppose that an MA(1) and an ARMA(2,1) are fitted in addition to the AR(2). A priori, we expect the AR(2) to be the preferred model of the three.

Diagnostic Checking

- The SAS program for estimation is as follows:

```
data example4;
input y;
cards;
4.0493268
7.0899911
4.7896497
.
.
2.2253477
2.439893;
proc arima data=example4;
identify var=y;
estimate p=2 method=ml printall;
estimate q=1 method=ml printall;
estimate p=2 q=1 method=ml printall;
run;
```

T-Tests for Coefficient Significance

- The AR(2) specification produces the model
$$\hat{Y}_t = 4.68115 + 0.35039\,Y_{t-1} - 0.49115\,Y_{t-2}$$

T-Tests for Coefficient Significance

- The AR(2) specification produces the model
  $$\hat{Y}_t = 4.68115 + 0.35039\,Y_{t-1} - 0.49115\,Y_{t-2}$$
- The t-statistics indicate that both ϕ_1 and ϕ_2 are significantly different from zero.

T-Tests for Coefficient Significance

- The AR(2) specification produces the model
  $\hat{Y}_t = 4.68115 + 0.35039 Y_{t-1} - 0.49115 Y_{t-2}$
- The t-statistics indicate that both ϕ_1 and ϕ_2 are significantly different from zero.
- Note that for any $ARMA(p,q)$ model, the mean value and the drift term are related through the formula
  $\mu = \frac{\delta}{1-\phi_1-\phi_2-\cdots-\phi_p}$

T-Tests for Coefficient Significance

- The AR(2) specification produces the model
  $\hat{Y}_t = 4.68115 + 0.35039\,Y_{t-1} - 0.49115\,Y_{t-2}$
- The t-statistics indicate that both ϕ_1 and ϕ_2 are significantly different from zero.
- Note that for any $ARMA(p, q)$ model, the mean value and the drift term are related through the formula
  $\mu = \frac{\delta}{1-\phi_1-\phi_2-\cdots-\phi_p}$
- Hence for the present AR(2) model,
  $\hat{\mu} = 4.10353 = 4.681115/(1 - 0.35039 + 0.49115)$

T-Tests for Coefficient Significance

- The MA(1) specification produces the model
  $$\hat{Y}_t = 4.10209 + 0.48367 e_{t-1},$$
  where $e_j = Y_j - \hat{Y}_j$ is the prediction error for the j^{th} period.

T-Tests for Coefficient Significance

- The MA(1) specification produces the model
  $$\hat{Y}_t = 4.10209 + 0.48367 e_{t-1},$$
  where $e_j = Y_j - \hat{Y}_j$ is the prediction error for the j^{th} period.
- The t-statistic indicates significance of θ_1.

T-Tests for Coefficient Significance

- The MA(1) specification produces the model
  $$\hat{Y}_t = 4.10209 + 0.48367 e_{t-1},$$
  where $e_j = Y_j - \hat{Y}_j$ is the prediction error for the j^{th} period.
- The t-statistic indicates significance of θ_1.
- $\hat{\mu} = 4.10209 = \hat{\delta}$.

T-Tests for Coefficient Significance

- The ARMA(2,1) specification produces the model
$$\hat{Y}_t = 4.49793 + 0.40904\,Y_{t-1} - 0.50516\,Y_{t-2} - 0.07693 e_{t-1}$$
- The t-statistic indicates that ϕ_1 and ϕ_2 are significantly different from zero but θ_1 is not significantly different from zero.

T-Tests for Coefficient Significance

- The ARMA(2,1) specification produces the model
$$\hat{Y}_t = 4.49793 + 0.40904 Y_{t-1} - 0.50516 Y_{t-2} - 0.07693 e_{t-1}$$
- The t-statistic indicates that ϕ_1 and ϕ_2 are significantly different from zero but θ_1 is not significantly different from zero.
- $\hat{\mu} = 4.10349 = 4.49793/(1 - 0.40904 + 0.50516)$

Residual Portmanteau Test

- If an ARMA(p,q) model is an adequate representation of the data generating process of Y_t, the residuals resulting from the zero should be uncorrelated.

Residual Portmanteau Test

- If an ARMA(p,q) model is an adequate representation of the data generating process of Y_t, the residuals resulting from the zero should be uncorrelated.
- We postulate that $e_t = \eta_1 e_{t-1} + \eta_2 e_{t-2} + \eta_3 e_{t-3} + \cdots$.
- If e_t's are uncorrelated, then $\eta_1 = \eta_2 = \eta_3 = \cdots = 0$.

Residual Portmanteau Test

- Hence we test
$$H_0 : \eta_1 = \eta_2 = \eta_3 = \cdots = \eta_m = 0 \quad \text{vs.} \quad H_1\text{: otherwise}$$

Residual Portmanteau Test

- Hence we test
$$H_0 : \eta_1 = \eta_2 = \eta_3 = \cdots = \eta_m = 0 \quad \text{vs.} \quad H_1\text{: otherwise}$$
- The portmanteau test statistic for H_0 is
$$Q(m) = (n-d)(n-d+2)\sum_{k=1}^{m} \frac{\tilde{r}_k^2}{n-d-k},$$
where $\tilde{r}_k$ is the sample AC of the e_t's at lag k. Under H_0, $Q(m) \sim \chi^2_{m-p-q}$.

Residual Portmanteau Test

- Hence we test
$$H_0 : \eta_1 = \eta_2 = \eta_3 = \cdots = \eta_m = 0 \quad \text{vs.} \quad H_1\text{: otherwise}$$
- The portmanteau test statistic for H_0 is
$$Q(m) = (n-d)(n-d+2)\sum_{k=1}^{m} \frac{\tilde{r}_k^2}{n-d-k},$$
where $\tilde{r}_k$ is the sample AC of the e_t's at lag k. Under H_0, $Q(m) \sim \chi^2_{m-p-q}$.
- The decision rule is to reject H_0 at the α level of significance if $Q(m) \geq \chi^2_{\alpha,m-p-q}$.

Residual Portmanteau Test

- For example, for the AR(2) model, for $m = 6$, $Q(6) = 3.85$ with a d.o.f. of 4 and a p-value of 0.4268. Hence H_0 cannot be rejected and the residuals are uncorrelated up to lag 6. The same conclusion is reached for $m = 12, 18, 24, 30$ if one sets α to 0.05.

Residual Portmanteau Test

- For example, for the AR(2) model, for $m = 6$, $Q(6) = 3.85$ with a d.o.f. of 4 and a p-value of 0.4268. Hence H_0 cannot be rejected and the residuals are uncorrelated up to lag 6. The same conclusion is reached for $m = 12, 18, 24, 30$ if one sets α to 0.05.
- On the other hand, residuals from the MA(1) model are significantly correlated. Residuals from the ARMA(2,1) model are uncorrelated.

Model Selection Criteria

Two most commonly adopted model selection criteria are the AIC and BIC:

- Akaike Information Criterion (AIC):
  $AIC = -2ln(L) + 2g$
- Bayesian Information Criterion (BIC):
  $BIC = -2ln(L) + gln(n)$,
  where L= value of the likelihood function, g=number of coefficients being estimated, and n=number of observations.

Model Selection Criteria

Two most commonly adopted model selection criteria are the AIC and BIC:

- Akaike Information Criterion (AIC):
  $AIC = -2ln(L) + 2g$
- Bayesian Information Criterion (BIC):
  $BIC = -2ln(L) + gln(n)$,
  where $L=$ value of the likelihood function, g=number of coefficients being estimated, and n=number of observations.
- BIC is also known as Schwartz Bayesian Criterion (SBC).

Model Selection Criteria

Two most commonly adopted model selection criteria are the AIC and BIC:

- Akaike Information Criterion (AIC):
  $AIC = -2ln(L) + 2g$
- Bayesian Information Criterion (BIC):
  $BIC = -2ln(L) + gln(n)$,
  where $L=$ value of the likelihood function, $g=$number of coefficients being estimated, and $n=$number of observations.
- BIC is also known as Schwartz Bayesian Criterion (SBC).
- Both the AIC and BIC are "penalised" versions of the log likelihood.

Model Selection Criteria

- The penalty term is larger in the BIC than in the AIC; in other words, the BIC penalises additional parameters more than the AIC does, meaning that the marginal cost of adding explanatory variables is greater with the BIC than with the AIC. Hence the BIC tends to select "parsimonious" models.

Model Selection Criteria

- The penalty term is larger in the BIC than in the AIC; in other words, the BIC penalises additional parameters more than the AIC does, meaning that the marginal cost of adding explanatory variables is greater with the BIC than with the AIC. Hence the BIC tends to select "parsimonious" models.
- Ideally, both the AIC and BIC should be as small as possible.
- In our example,
  AR(2): AIC=1424.660, SBC=1437.292
  MA(1): AIC=1507.162, SBC=1515.583
  ARMA(2,1): AIC=1425.727, SBC=1442.570

Diagnostic Checking

- Summary of model diagnostics and comparisons:

	t-test	Q-test	AIC	BIC
AR(2)	$\surd$	$\surd$	1424.660	1437.292
MA(1)	$\surd$	$\times$	1507.162	1515.583
ARMA(2,1)	$\surd$(partial)	$\surd$	1425.727	1442.57

- Judging from these results, the AR(2) model is the best, corroborating our conjecture.

Forecasting

- The h-period ahead forecast based on an ARMA(p,q) model is given by:

$$\begin{aligned}\hat{Y}_{t+h} = \hat{\delta} + \hat{\phi}_1 Y_{t+h-1} + \hat{\phi}_2 Y_{t+h-2} + \cdots + \hat{\phi}_p Y_{t+h-p} \\ -\hat{\theta}_1 e_{t+h-1} - \hat{\theta}_2 e_{t+h-2} - \cdots - \hat{\theta}_q e_{t+h-q},\end{aligned}$$

where quantities on the r.h.s. of the equation may be replaced by their estimates when the actual values are unavailable.

Forecasting

- For example, for the AR(2) model,

$$\begin{aligned}\hat{Y}_{498+1} &= 4.681115 + 0.35039 \times 2.4339893 \\ &\quad -0.49115 \times 2.2253477 \\ &\quad = 4.440981 \\ \hat{Y}_{498+2} &= 4.681115 + 0.35039 \times 4.440981 \\ &\quad -0.49115 \times 2.4339893 \\ &\quad = 5.041784\end{aligned}$$

- For the MA(1) model,

$$\begin{aligned}\hat{Y}_{498+1} &= 4.10209 + 0.48367 \times -0.7263 = 3.7508 \\ \hat{Y}_{498+2} &= 4.10209 + 0.48367 \times 0 = 4.10209 \\ \hat{Y}_{498+3} &= 4.10209\end{aligned}$$

- Class exercise: Example 5 - Finland's quarterly construction building permit (in thousands), seasonally adjusted, 1977Q1-1987Q1

Summary

- ARIMA models are regression models that use past values of the series itself and unobservable random disturbances as explanatory variables.
- Stationarity of data is a fundamental requirement underlying ARIMA and most other techniques involving time series data.
- AR models are not always stationary; MA models are always stationary but not necessarily invertible; ARMA models are neither necessarily stationary nor invertible.
- Differencing transformation is required for non-stationary series. The number of times a series must be differenced in order to achieve stationarity is known as the order of integration. Differencing always results in a loss of information.

Summary

- Tentative model identification is based on recognisable patterns of ACF and PACF. We look for the best approximation to the data by a member of the ARIMA family. Usually more than one candidate model is identified at the initial stage.
- ARIMA models are estimated by maximum likelihood.
- Diagnostic checking comprises t-tests of coefficient significance, residual portmanteau test, and AIC and BIC model selection criteria.
- Sometimes the diagnostics can lead to contradictory results and no one single model is necessarily superior to all others. We can consider combining models in such situations.