made by abdelmomen ben nasr

# MATHEMATICAL CONCEPTS BEHIND

# SEASONAL AND NON-SEASONAL ARIMA MODELS

# CONTENT

# TIME SERIES

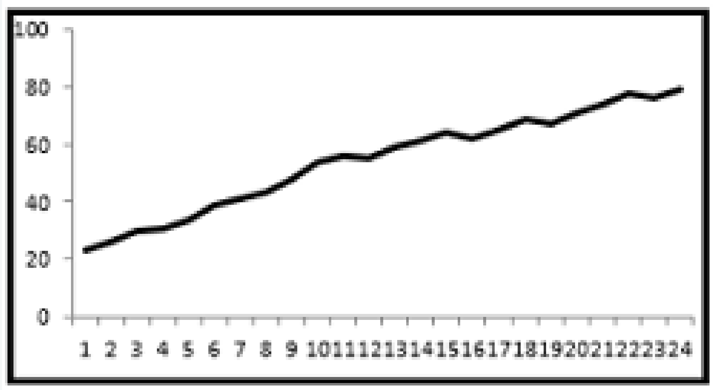
Time series refers to a sequence of data points collected and recorded over a period of time. It involves observations or measurements taken at regular intervals, such as hourly, daily, weekly, monthly, or yearly.

These data points are typically arranged in chronological order, allowing analysts to analyze and identify patterns, trends, and other characteristics over time.

# COMPONENTS OF A TIME SERIES

## Trend

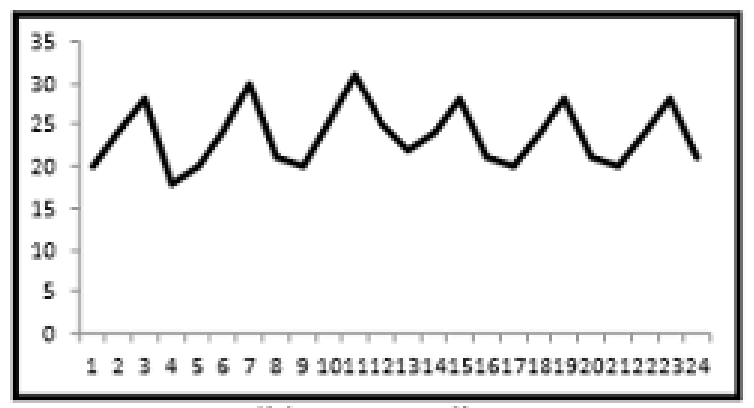
Overall direction of a data overtime (whether it's increasing or decreasing or staying the same)
a line on a graph that's either going up or down or staying flat



(a) Trend

## Seasonality

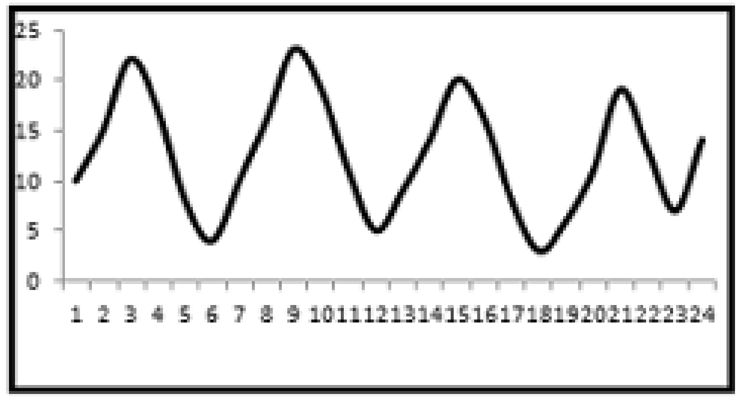
A repeating pattern of data over a set of period of time, like the way that retail sales spike during the holiday seasons



(b) Seasonality

## Cyclicality

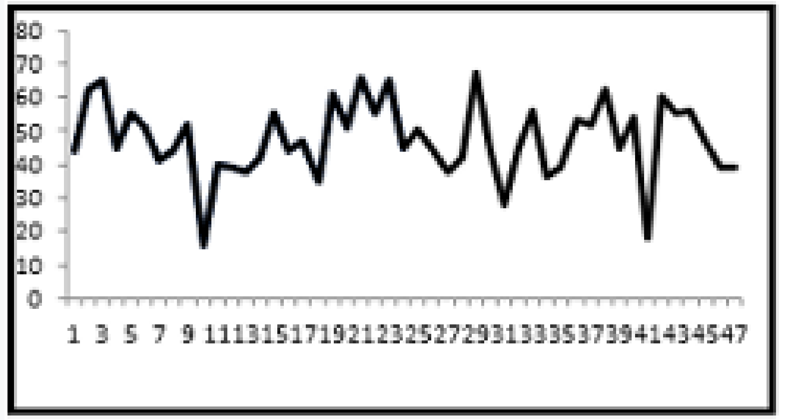
Repeating but non seasonal patterns in the data (smooth curve)



(c) Cyclicality

## Irregularity

Unpredictable ups and downs that occure in the data and cannot be explained by the other components also known as irregularity or noise



(d) Irregular

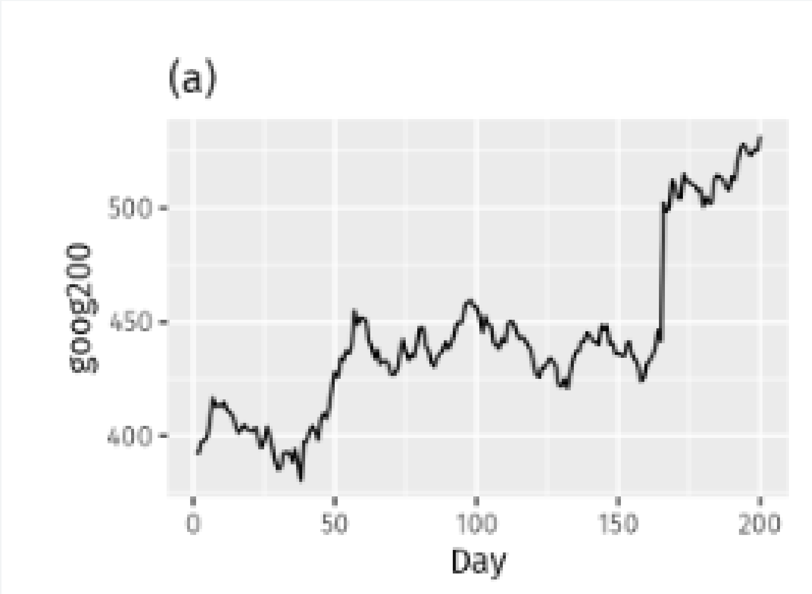
# Stationarity and differencing

A stationary time series is where its properties do not depend on the time at which the series is observed.

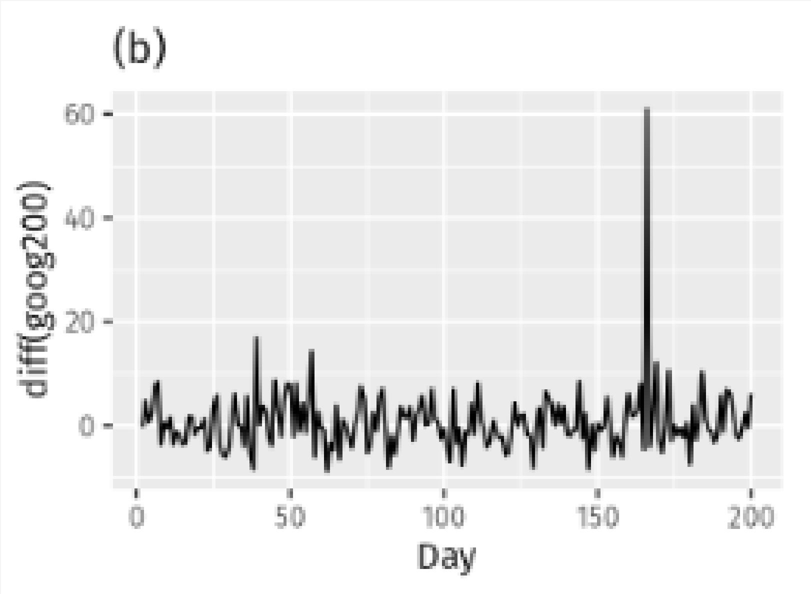
Thus, time series with trends, or with seasonality are not stationary (the trend and seasonality will affect the value of the time series at different times) .

On the other hand a time series with a cyclic behaviour (but without trend or seasonality) is stationary.

In general, a stationary time series will have no predictable patterns in the long-term. Time plots will show the series to be roughly horizontal (although some cyclic behaviour is possible), with constant variance.
(The variance is a way to the calculate how individual values in the series deviate from the mean or average value)




*GOOGLE STOCK PRICE FOR 200 CONSECUTIVE DAYS*

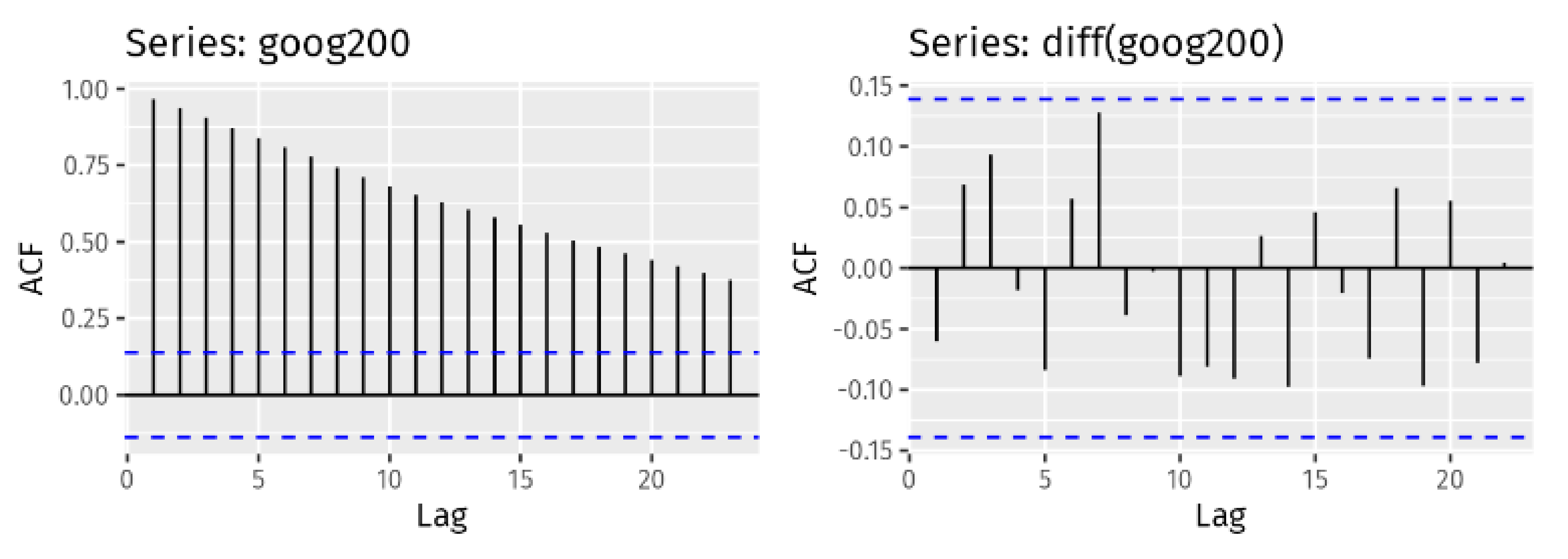



*DAILY CHANGE IN THE GOOGLE STOCK PRICE FOR 200 CONSECUTIVE DAYS*

(a) has trends and changing levels thus it can't be stationary. But in (b), it's shown an irregularity but no trends or seasonality thus it's stationary.
This shows one way to make a non-stationary time series stationary — compute the differences between consecutive observations. This is known as **differencing**.
Transformations such as logarithms can help to stabilise the variance of a time series, Differencing can help stabilise the mean of a time series by removing changes in the level of a time series, and therefore eliminating (or reducing) trend and seasonality.

# Stationarity and differencing

As well as looking at the time plot of the data, the ACF (Auto-Correlation Function) plot is also useful for identifying non-stationary time series. For a stationary time series, the ACF will drop to zero relatively quickly, while the ACF of non-stationary data decreases slowly. Also, for non-stationary data, the value of r1 ACF(1) is often large and positive.




The formula for autocorrelation, also known as the autocorrelation function (ACF), is as follows:

$$ACF(k) = \frac{1}{n-k} \sum [(x_i - \bar{x})(x_{i+k} - \bar{x})]$$

where:

- ACF(k) represents the autocorrelation at lag k.
- n is the total number of observations in the dataset.
- x_i is the value of the time series at time i.
- x̄ is the mean value of the time series.
- Σ denotes the summation symbol, summing over all observations from i = 1 to n - k.

# White noise and correlation

A time series is white noise if the variables are independent and identically distributed with a mean of zero.
All variables have the same variance (sigma^2) and each value has a zero correlation with all other values in the series.
White noise is an important concept in time series analysis and forecasting.
It is important for two main reasons:

1. Predictability: If your time series is white noise, then, by definition, it is random. You cannot reasonably model it and make predictions.
2. Model Diagnostics: The series of errors from a time series forecast model should ideally be white noise.

The formula for calculating the correlation between two time series (the Pearson correlation coefficient) is as follows ,

$$r = \frac{\sum((X_i - \bar{X})(Y_i - \bar{Y}))}{\sqrt{\sum(X_i - \bar{X})^2 \cdot \sum(Y_i - \bar{Y})^2}}$$

In this formula:

- X_i and Y_i represent the individual values of the two time series at a specific time point "i."
- X̄ and Ȳ represent the means of the X and Y series, respectively.
- Σ denotes the summation symbol, meaning you need to sum up the calculations for all time points.

The numerator of the formula represents the covariance between the two series, while the denominator calculates the product of their standard deviations. Dividing the covariance by the product of the standard deviations gives the correlation coefficient.
r is a value between -1 and 1 where
r=-1 is a perfect negative correlation and r =1
is a perfect positive correlation and r= 0 indicates no correlation.

# Random walk model

The differenced series is the change between consecutive observations in the original series, and can be written as $y'_t = y_t - y_{t-1}$.

The differenced series will have only T-1 values, since it is not possible to calculate a difference $y'_1$ for the first observation.

When the differenced series is white noise, the model for the original series can be written as $y_t - y_{t-1} = \varepsilon_t$,

where $\varepsilon_t$ denotes white noise. Rearranging this leads to the “random walk” model $y_t = y_{t-1} + \varepsilon_t$.

Random walk models are widely used for non-stationary data, particularly financial and economic data. Random walks typically have:

- long periods of apparent trends up or down
- sudden and unpredictable changes in direction.

The forecasts from a random walk model are equal to the last observation, as future movements are unpredictable, and are equally likely to be up or down. Thus, the random walk model underpins naïve forecasts.

A closely related model allows the differences to have a non-zero mean. Then

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

or

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

The value of c is the average of the changes between consecutive observations. If c is positive, then the average change is an increase in the value of $y_t$. Thus, $y_t$ will tend to drift upwards. However, if c/is negative, $y_t$ will tend to drift downwards.

# Second-order and seasonal differencing

Occasionally the differenced data will not appear to be stationary and it may be necessary to difference the data a second time to obtain a stationary series:

$$y''_t = y'_t - y'_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}$$

In this case, y′′t will have T-2 values. Then, we would model the “change in the changes” of the original data. In practice, it is almost never necessary to go beyond second-order differences.

A seasonal difference is the difference between an observation and the previous observation from the same season.
So

$$y'_t = y_t - y_{t-m}$$

where m= the number of seasons, These are also called “lag-m differences”, as we subtract the observation after a lag of m periods.
If seasonally differenced data appear to be white noise, then an appropriate model for the original data is yt=yt-m+εt.
Forecasts from this model are equal to the last observation from the relevant season. That is, this model gives seasonal naïve forecasts.

# Backshift notation

The backward shift operator B is a useful notational device when working with time series lags:

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

Two applications of B to yt shifts the data back two periods:

$$B(By_t) = B^2 y_t = y_{t-2}\,.$$

For monthly data, if we wish to consider "the same month last year," the notation:

$$B^{12} y_t = y_{t-12}$$

The backward shift operator is convenient for describing the process of differencing. A first difference can be written as

$$y'_t = y_t - y_{t-1} = y_t - By_t = (1-B)y_t$$

Note that a first difference is represented by (1-B). Similarly, if second-order differences have to be computed, then:

$$y''_t = y_t - 2y_{t-1} + y_{t-2} = (1 - 2B + B^2)y_t = (1-B)^2 y_t$$

In general, a dth-order difference can be written as

$$(1-B)^d y_t$$

Backshift notation is particularly useful when combining differences, as the operator can be treated using ordinary algebraic rules. In particular, terms involving B can be multiplied together.

For example, a seasonal difference followed by a first difference can be written as

$$\begin{aligned}(1-B)(1-B^m)y_t &= (1 - B - B^m + B^{m+1})y_t \\ &= y_t - y_{t-1} - y_{t-m} + y_{t-m-1}\end{aligned}$$

# AutoRegressive Models

In an autoregression model, we forecast the variable of interest using a linear combination of past values of the variable. The term autoregression indicates that it is a regression of the variable against itself.
Thus, an autoregressive model of order p can be written as

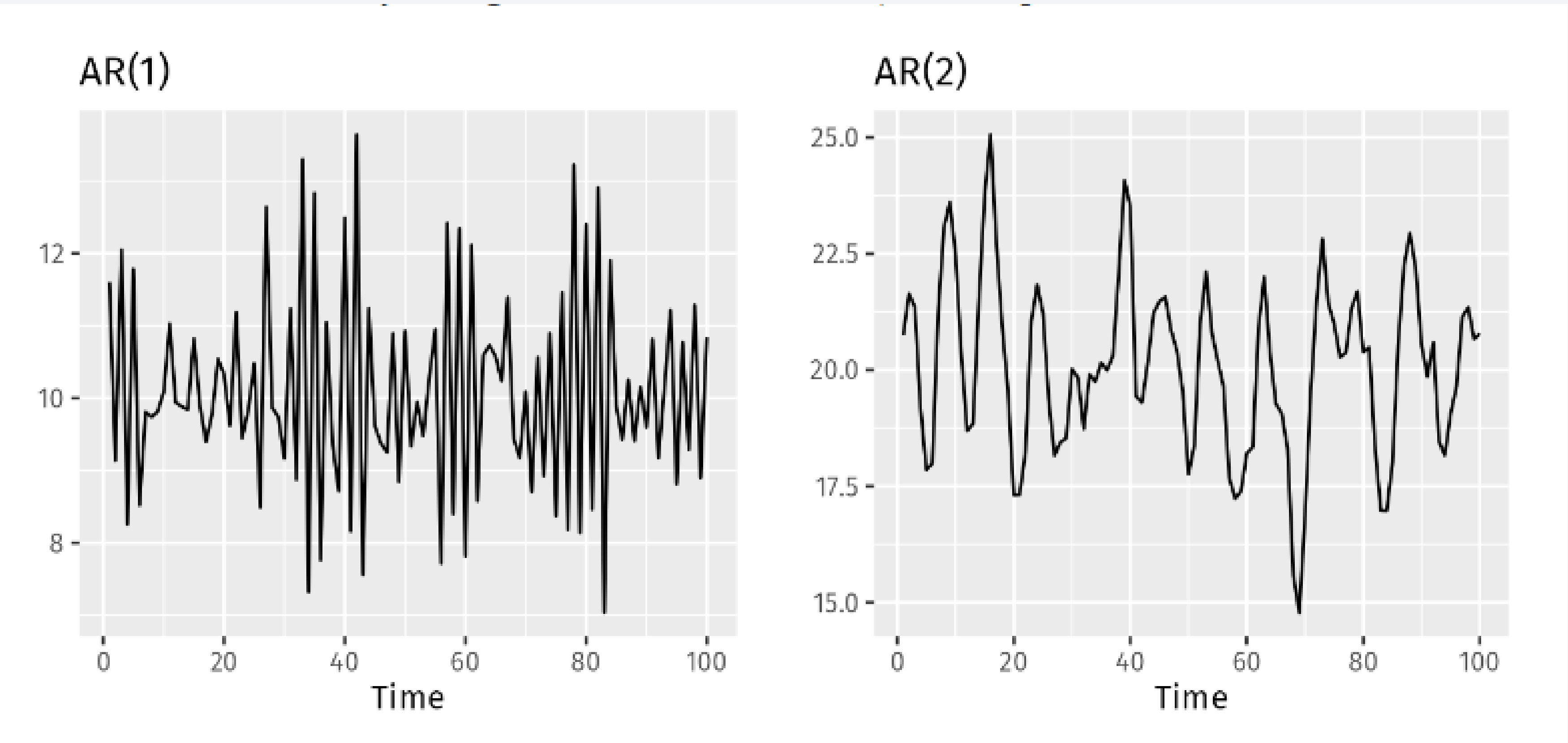
$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t$$

where

- εt is white noise.
- ϕ1, ϕ2, ..., ϕp are the autoregressive parameters that represent the relationship between the current value (yt) and the previous p values $(y_{t-1}, y_{t-2}, \dots, y_{t-p})$

This is like a multiple regression but with lagged values of yt as predictors. We refer to this as an AR(p) model, an autoregressive model of order p.

Autoregressive models are remarkably flexible at handling a wide range of different time series patterns. The two series shown below are from an AR(1) model and an AR(2) model. Changing the parameters ϕ1,…,ϕp results in different time series patterns. The variance of the error term εt will only change the scale of the series, not the patterns.




Two examples of data from autoregressive models with different parameters. Left: AR(1) with yt=18−0.8yt−1+εt. Right: AR(2) with yt=8+1.3yt−1−0.7yt−2+εt. In both cases, ε is normally distributed white noise with mean zero and variance one.

# AutoRegressive Models

For an AR(1) model:

- when $\phi_1=0$ $y_t$ is equivalent to white noise;
- when $\phi_1=1$ and $c=0$, $y_t$ is equivalent to a random walk;
- when $\phi_1=1$ and $c\neq0$, $y_t$ is equivalent to a random walk with drift;
- when $\phi_1<0$, $y_t$ tends to oscillate around the mean.

We normally restrict autoregressive models to stationary data, in which case some constraints on the values of the parameters are required.

- For an AR(1) model: $-1<\phi_1<1$
- For an AR(2) model: $-1<\phi_2<1$, $\phi_1+\phi_2<1$, $\phi_2-\phi_1<1$

When $p\geq3$ the restrictions are much more complicated. Programming languages take care of these restrictions when estimating a model.

# Moving-Average Models

A moving average model uses past forecast errors in a regression rather than using past values of the forecast variable in a regression-like model.

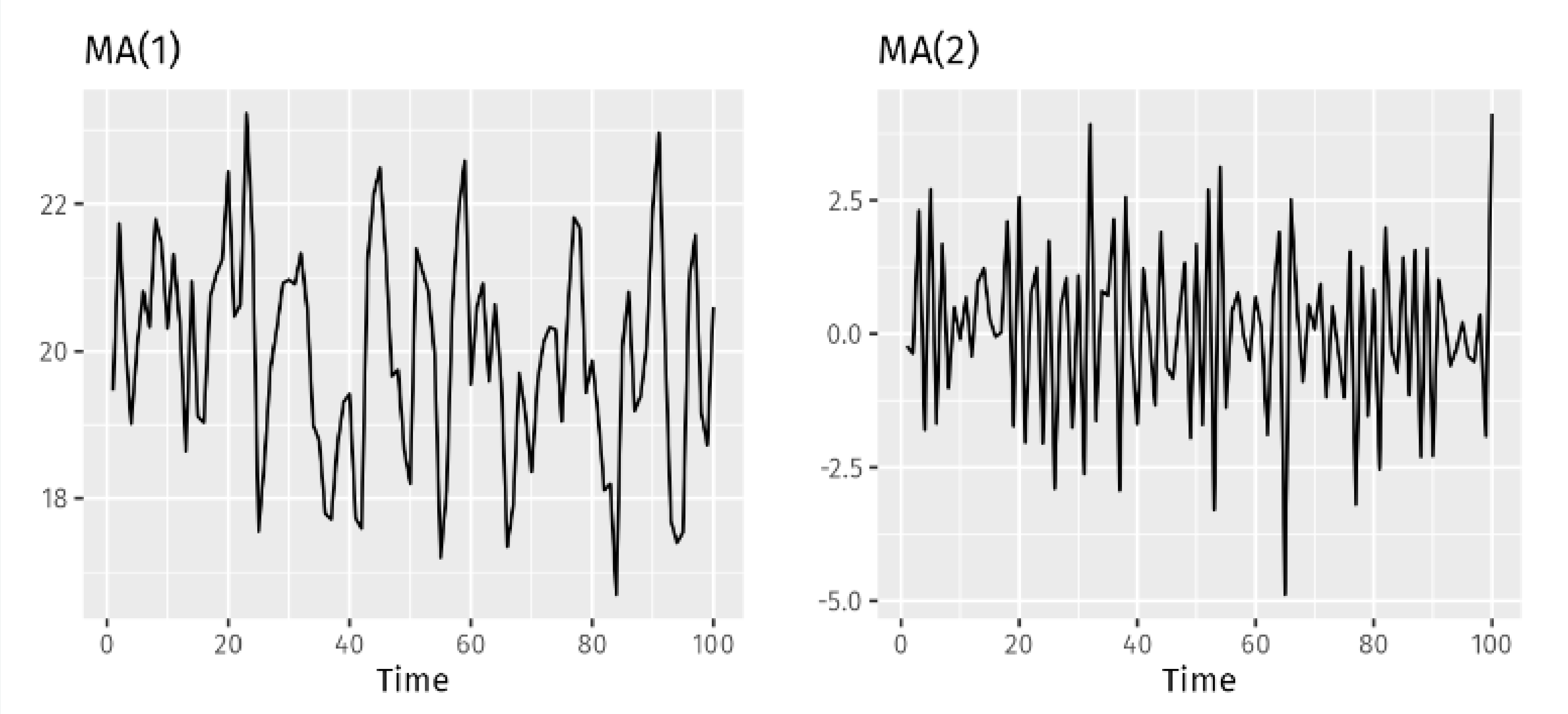
$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}$$

where

- εt is white noise.
- θ1, θ2, ..., θq are the parameters of the moving average model, representing the weights assigned to the past forecast errors.
- yt represents the value of the time series at time t.
- c is a constant term or intercept.

We refer to this as an MA(q) model, a moving average model of order q.
we do not observe the values of εt, so it is not really a regression in the usual sense.

The figures bellow show, some data from an MA(1) model and an MA(2) model. Changing the parameters θ1,…,θq results in different time series patterns. As with autoregressive models, the variance of the error term εt will only change the scale of the series, not the patterns.




Two examples of data from moving average models with different parameters. Left: MA(1) with yt=20+εt+0.8εt-1. Right: MA(2) with yt=εt-εt-1+0.8εt-2In both cases, εt is normally distributed white noise with mean zero and variance one.

# Moving-Average Models

It is possible to write any stationary AR(p) model as an MA(∞) model. For example, using repeated substitution, we can demonstrate this for an AR(1) model:

$$= \phi_1 y_{t-1} + \varepsilon_t = \phi_1(\phi_1 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \phi_1^2 y_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t = \phi_1^3 y_{t-3} + \phi_1^2 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t + \dots$$

Provided -1<ϕ1<1, the value of ϕk1 will get smaller as k gets larger. So eventually we obtain

$$y_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2^1 \varepsilon_{t-2} + \phi_3^1 \varepsilon_{t-3} + \dots$$

an MA(∞) process.

The reverse result holds if we impose some constraints on the MA parameters. Then the MA model is called invertible. That is, we can write any invertible MA(q) process as an AR(∞) process. Invertible models are not simply introduced to enable us to convert from MA models to AR models. They also have some desirable mathematical properties.

For example, consider the MA(1) process,

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

In its AR(∞) representation, the most recent error can be written as a linear function of current and past observations:

$$\varepsilon_t = \sum_{j=0}^{\infty} (-\theta)^j y_{t-j}$$

# Moving-Average Models

When $|\theta|>1$, the weights increase as lags increase, so the more distant the observations the greater their influence on the current error. When $|\theta|=1$, the weights are constant in size, and the distant observations have the same influence as the recent observations. As neither of these situations make much sense, we require $|\theta|<1$, so the most recent observations have higher weight than observations from the more distant past. Thus, the process is invertible when $|\theta|<1$

The invertibility constraints for other models are similar to the stationarity constraints.

- For an MA(1) model: $-1<\theta_1<1$.
- For an MA(2) model: $-1<\theta_2<1$, $\theta_2+\theta_1>-1$, $\theta_1-\theta_2<1$

More complicated conditions hold for $q\geq 3$. Again, R will take care of these constraints when estimating the models.

# Non-seasonal ARIMA models

If we combine differencing with autoregression and a moving average model, we obtain a non-seasonal ARIMA model. ARIMA is an acronym for AutoRegressive Integrated Moving Average (in this context, "integration" is the reverse of differencing). The full model can be written as

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

where y′t is the differenced series (it may have been differenced more than once). The "predictors" on the right hand side include both lagged values of yt and lagged errors. We call this an ARIMA(p,d,q) model, where

- P: The autoregressive (AR) order, which represents the number of lagged observations used as predictors in the model.
- D: The differencing order, which indicates the number of times the time series needs to be differenced to achieve stationarity.
- Q: The moving average (MA) order, which denotes the number of lagged forecast errors included in the model.

The same stationarity and invertibility conditions that are used for autoregressive and moving average models also apply to an ARIMA model. Many of the models we have already discussed are special cases of the ARIMA model

| | |
|---|---|
| White noise | ARIMA(0,0,0) |
| Random walk | ARIMA(0,1,0) with no constant |
| Random walk with drift | ARIMA(0,1,0) with a constant |
| Autoregression | ARIMA($p$,0,0) |
| Moving average | ARIMA(0,0,$q$) |

Using the Backshifting notations we can rewrite the model as :

$$\underset{\substack{\uparrow \\ \text{AR}(p)}}{(1 - \phi_1 B - \cdots - \phi_p B^p)} \quad \underset{\substack{\uparrow \\ d \text{ differences}}}{(1 - B)^d y_t} \quad = \quad c + \underset{\substack{\uparrow \\ \text{MA}(q)}}{(1 + \theta_1 B + \cdots + \theta_q B^q)\varepsilon_t}$$

# Non-seasonal ARIMA models

The constant c has an important effect on the long-term forecasts obtained from these models.

- If c=0 and d=0, the long-term forecasts will go to zero.
- If c=0 and d=1, the long-term forecasts will go to a non-zero constant.
- If c=0 and d=2, the long-term forecasts will follow a straight line.
- If c≠0 and d=0, the long-term forecasts will go to the mean of the data.
- If c≠0 and d=1, the long-term forecasts will follow a straight line.
- If c≠0 and d=2, the long-term forecasts will follow a quadratic trend.

The value of d also has an effect on the prediction intervals — the higher the value of d, the more rapidly the prediction intervals increase in size. For d=0, the long-term forecast standard deviation will go to the standard deviation of the historical data, so the prediction intervals will all be essentially the same.

The value of p is important if the data show cycles. To obtain cyclic forecasts, it is necessary to have p≥2, along with some additional conditions on the parameters. For an AR(2) model, cyclic behaviour occurs if

$$\phi_1^2 + 4\phi_2 < 0$$

In that case, the average period of the cycles is

$$\frac{2\pi}{\text{arc cos}(-\phi_1(1-\phi_2)/(4\phi_2))}.$$

# ACF/PACF

Visual inspection of ACF and PACF plots: By examining the autocorrelation and partial autocorrelation plots of the time series, leads to identifying potential values for P and Q. The ACF plot helps determine the MA order (Q) by observing the significant spikes at various lags. The PACF plot helps identify the AR order (P) by identifying significant spikes at different lags.

If the data are from an ARIMA(p,d,0) or ARIMA(0,d,q) model, then the ACF and PACF plots can be helpful in determining the value of p or q.
If p and q are both positive, then the plots do not help in finding suitable values of p and q.
The data may follow an ARIMA(p,d,0) model if the ACF and PACF plots of the differenced data show the following patterns:

- the ACF is exponentially decaying or sinusoidal;
- there is a significant spike at lag p in the PACF, but none beyond lag p.

The data may follow an ARIMA(0,d,q) model if the ACF and PACF plots of the differenced data show the following patterns:

- the PACF is exponentially decaying or sinusoidal;
- there is a significant spike at lag q in the ACF, but none beyond lag q.

# Maximum likelihood estimation and Information Criteria

Once the model order has been identified (i.e., the values of p, d and q), we need to estimate the parameters c, ϕ1,…,ϕp, θ1,…,θq generally using Maximum likelihood estimation for Arima models. model is based on the assumption that the residuals (the differences between the observed values and the predicted values) follow a Gaussian (normal) distribution.

Let's denote the time series data as {y_t}, where t = 1, 2, ..., T represents the time index. The ARIMA(p, d, q) model can be written as:

$$\underset{\substack{\uparrow \\ \text{AR}(p)}}{(1-\phi_1 B-\cdots-\phi_p B^p)} \quad \underset{\substack{\uparrow \\ d\text{ differences}}}{(1-B)^d y_t} \quad = \quad \underset{\substack{\uparrow \\ \text{MA}(q)}}{c+(1+\theta_1 B+\cdots+\theta_q B^q)\varepsilon_t}$$

To obtain the likelihood function, we assume that the errors ε_t are normally distributed. Therefore, the density function of ε_t is given by:

$$f(\varepsilon_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon_t^2}{2\sigma^2}\right)$$

The likelihood function is the joint probability density function of the errors, given the observed data. Assuming that the errors are independent, the likelihood function can be written as the product of the individual error density functions:

$$L(\phi_1,\ldots,\phi_p,\theta_1,\ldots,\theta_q,c,\sigma^2) = \prod f(\varepsilon_t)$$

# Maximum likelihood estimation and Information Criteria

Taking the logarithm of the likelihood function simplifies calculations, as it transforms the product into a sum:

$$\log(L) = \sum \log(f(\varepsilon_t))$$

The goal is to find the parameter values that maximize the log-likelihood function. This can be done using optimization techniques, such as
gradient descent algorithms.

Akaike’s Information Criterion (AIC), which was useful in selecting predictors for regression, is also useful for determining the order of an ARIMA model. It can be written as

$$\text{AIC} = -2\log(L) + 2(p+q+k+1).$$

where L is the likelihood of the data, k=1 if c≠0 and k=0if c=0. Note that the last term in parentheses is the number of parameters in the model (including σ2, the variance of the residuals).
For ARIMA models, the corrected AIC can be written as

$$\text{AICc} = \text{AIC} + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2},$$

# Seasonal Arima Model

ARIMA models are also capable of modelling a wide range of seasonal data. A seasonal ARIMA model is formed by including additional seasonal terms in the ARIMA models we have seen so far. It is written as follows:

$$\text{ARIMA} \quad \underbrace{(p, d, q)}_{\substack{\uparrow \\ \text{Non-seasonal part} \\ \text{of the model}}} \quad \underbrace{(P, D, Q)_m}_{\substack{\uparrow \\ \text{Seasonal part of} \\ \text{of the model}}}$$

where m= number of observations per year. We use uppercase notation for the seasonal parts of the model, and lowercase notation for the non-seasonal parts of the model

The seasonal part of the model consists of terms that are similar to the non-seasonal components of the model, but involve backshifts of the seasonal period. For example, an ARIMA(1,1,1)(1,1,1)4 model (without a constant) is for quarterly data (m=4), and can be written as

$$(1 - \phi_1 B)\ (1 - \Phi_1 B^4)(1 - B)(1 - B^4) y_t = (1 + \theta_1 B)\ (1 + \Theta_1 B^4)\varepsilon_t.$$

The additional seasonal terms are simply multiplied by the non-seasonal terms.

# ACF/PACF

The seasonal part of an AR or MA model will be seen in the seasonal lags of the PACF and ACF. For example, an ARIMA(0,0,0)(0,0,1)12 model will show:

- a spike at lag 12 in the ACF but no other significant spikes;
- exponential decay in the seasonal lags of the PACF (i.e., at lags 12, 24, 36, …).

Similarly, an ARIMA(0,0,0)(1,0,0)12 model will show:

- exponential decay in the seasonal lags of the ACF;
- a single significant spike at lag 12 in the PACF.

The modelling procedure is almost the same as for non-seasonal data, except that we need to select seasonal AR and MA terms as well as the non-seasonal components of the model.

# THANKS FOR YOUR ATTENTION !