Autocorrelation

- For a time series $y_1, \ldots, y_n$, the autocorrelation at lag L is defined as the Pearson coefficient of correlation between y_t and y_{t+L}

immediate window to predict

$$\text{Autocorrelation}(L) = \frac{\text{Covariance}_t(y_t, y_{t+L})}{\text{Variance}_t(y_t)}$$

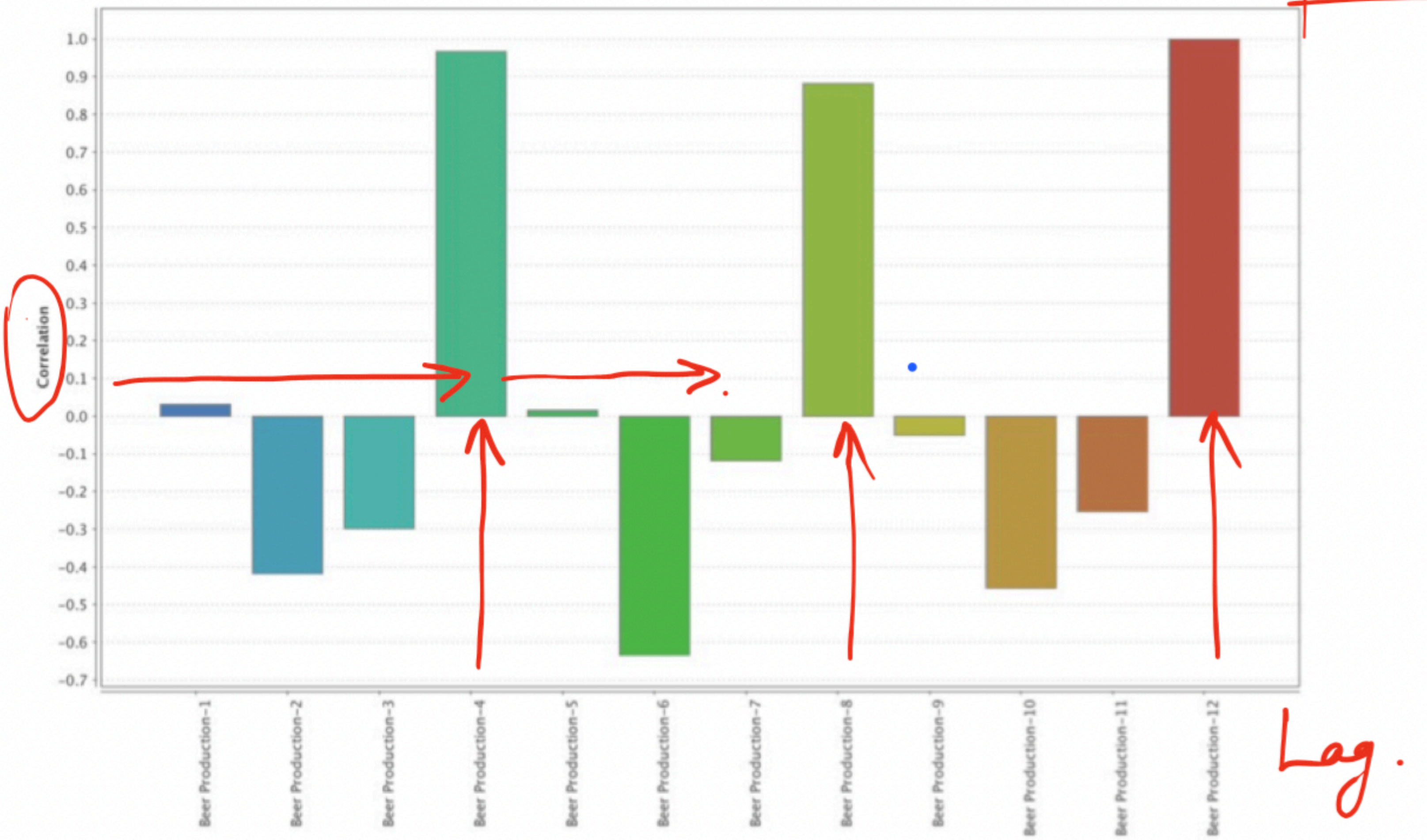
Autoregression

- The autocorrelation always lies in the range $[-1,1]$
- Adjacent values in time series are very similar, therefore, for small values of lag L, autocorrelation is positive.
- With increasing lag, the similarity drops off, and so does autocorrelation

again rises with seasonal periodicity.

Autocorrelation Function

Can be used to detect the periodicity



Variance and Covariance

$$\text{Covariance}_t(y_t, y_{t+L}) = \frac{\sum\ (y_t - \bar{y}_t)\ \ (y_{t+L} - \bar{y}_{t+L})}{N}$$

standard formulae

Here N is the number of pairs of $\left(y_t, y_{t+L}\right)$ considered

$$\text{Variance}_t(y_t) = \frac{\sum (y_t - \bar{y}_t)^2}{n-1}$$

where n is the number of observations of y_t

Autocorrelation

- <u>High</u> <u>absolute</u> <u>values</u> of <u>autocorrelation</u> imply that the value at a given position in the series can be <u>predicted</u> as a <u>function</u> of the values in the immediately <u>preceding</u> <u>window</u>.

Autocorrelation → Autoregressive series.

(detect periodicity)

L.

L.

Autoregressive Model

—— (reference). Average trend line

- In the autoregressive model, the value y_t at time t is defined as a linear combination of values in the immediately preceding window of length p

level.

$$\boxed{y_t} = L + \underbrace{\alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p}}_{\text{past p values are used for prediction}} + \epsilon_t$$

error (zero mean).

(L is the level, ϵ_t is the noise)

- A model that uses the preceding window of length p is referred to as $AR(p)$ model

p: window size

- The coefficients α's are unknown and learned using linear regression using training data

for auto regression.

α's remain constant for the entire series.

?

Autoregressive Model

- The larger the value of p, the greater the lag incorporated for autocorrelations.

Natural time series, temperature

speech.

- Autocorrelation often reduces with increasing value of the lag L.

20 19 21

correlation

stroke-level correlation

- Size of the regression window (p) should be adequate so as to achieve good generalisation and avoid overfitting

hyper-parameter

Types of time series

- Time series can be either stationary or non-stationary

- **Stationary** stochastic process:

  Its parameters such as mean and variance do not change with time.

- Non-stationary stochastic process:

  Its parameters change with time

Prediction is impossible using the conventional method.

Convert to a stationary series (transformation).

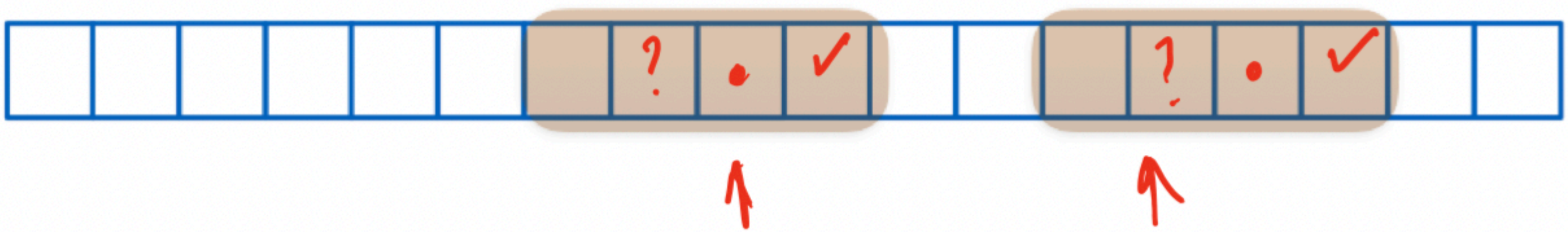
Stationary stochastic process

- White noise is the strongest form of stationarity

  It has

  - zero mean,
  - constant variance,
  - zero covariance between series values separated by a fixed lag

covariance matrix is diagonal.



Strictly stationary time series

Strict / weak.

- A strictly stationary time series is one in which the probabilistic distribution of the values in any time interval $[a, b]$ is identical to that in the shifted interval $[a + s, b + s]$ for any value of the time shift s.

any shift

- That is, all multivariate distributions of subsets of variables must match with their shifted counterparts.

1	2	3	4	5

- Since parameters do not vary over different time windows, they can be estimated reliably.

Non-stationary stochastic process

- Most series in real applications are non-stationary

- Price of an industrial commodity such as crude oil

  - average price level may increase over time

- Prediction of future values cannot be done using the current statistical parameters such as mean, variances, correlations, etc.

Time series with weak stationarity property

- Weak stationarity

1) Covariance

2) Mean

1). **Covariance Stationarity**:

The mean of the series, and the covariance between approximately adjacent time series values may be nonzero but constant over time

⇒ General statistics over any subset of variables need not remain constant.

- covariance stationarity can be assessed easily

- it is also useful for forecasting models

Time series with weak stationarity property

- Weak stationarity

Trend Stationarity:

Instead of assuming that the mean is constant, here we assume that the <u>average value</u> of the series can be <u>described</u> by a <u>trend line</u>.

- it is also useful for forecasting models

trend

linear rise

Time series with weak stationarity property

- Weak stationarity

Trend Stationarity:

Instead of assuming that the mean is constant, here we assume that the average value of the series can be described by a trend line.

- it is also useful for forecasting models

Trend

linear rise

Converting a non-stationary series to stationary series

- Differencing

Time series value y_i is replaced by the difference between it and the previous value.

10

9

The new value y'_i is computed as $y'_i = \underbrace{y_i - y_{i-1}}$

If the series is stationary after differencing, then an appropriate model for the data is $\boxed{y_{i+1} = y_i + e_{i+1}}$ for the differenced series.

New series

where e_{i+1} corresponds to white noise with zero mean

Converting a non-stationary series to stationary series

- A differenced time series would have $t - 1$ values for a series of length t
- Higher order differencing can be used to achieve stationarity in second order changes.

$$y_i'' = y_i' - y_{i-1}'$$
$$= y_i - 2\,y_{i-1} + y_{i-2}$$

$(y_i - y_{i-1}) - (y_{i-1} - y_{i-2})$

$= y_i - 2y_{i-1} + y_{i-2}$

- The corresponding model of the series is as follows:

$$\boxed{y_{i+1} = y_i + c + e_{i+1}}$$

for the differenced series

The model allows the series to drift over time.

Non-zero constant c accounts for the drift

Converting a non-stationary series to stationary series

sales in Feb 2023
− sales in Feb 2022

- **Seasonal differencing** may be useful for certain series

$$y_i' = y_i - y_{i-m} \qquad m > 1$$

- For **geometrically increasing series**, for example, price of some commodity increases at an approximately constant inflation factor,

  - we can apply the **logarithm function** to the series values, before the differencing operation

Augmented Dickey Fuller Test

$$\text{Non-stationary} \xrightarrow{\text{Diff, Log}} \text{Stationary}$$

Moving Average of Error

- One can also create a regression equation involving forecast errors of past data and use it as a predictor.

$$y_t = L + \underbrace{e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q}}_{\text{previous forecast errors.}}$$

Here e_i is the forecast error of data point i

(AR) Auto regression : regress over the past observations / differences

(MA) of error : regression over the forecast errors.

Autoregressive Integrated Moving Average (ARIMA)

ARIMA

- The ARIMA model is a combination of the differenced autoregressive model with the moving average model.

forecasted error.

α, θ

$$y'_t = L + \underbrace{\alpha_1 y'_{t-1} + \alpha_2 y'_{t-2} + \ldots + \alpha_p y'_{p-1}}_{AR(p)} + \underbrace{e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q}}_{MA}$$

where $y'_t = y_t - y_{t-1}$ or it can be an order d differencing.

- The AR part of ARIMA shows that the time series is regressed on its own past data.
- The MA part indicates that the forecast error is a linear combination of past respective errors.

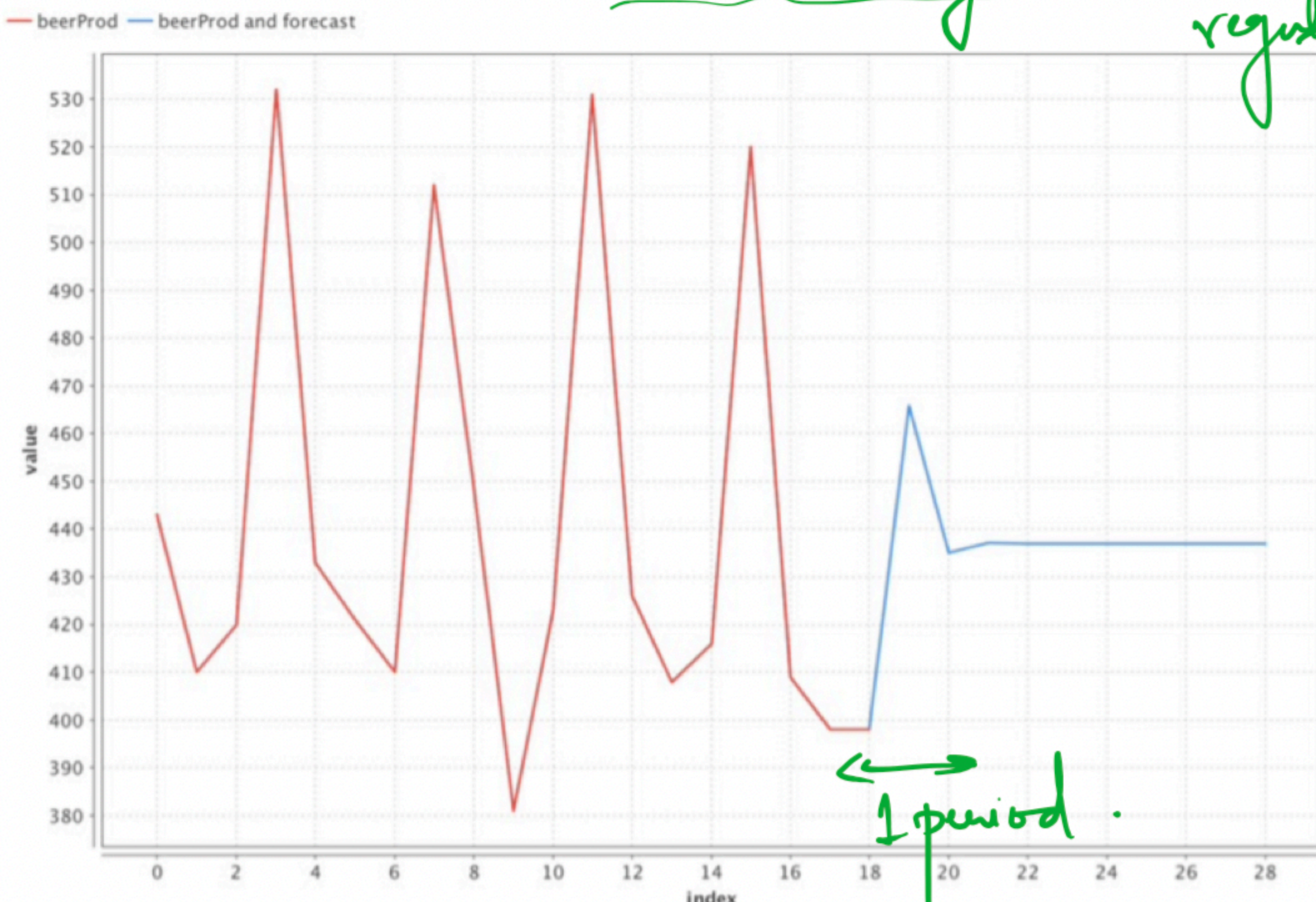
Forecast using ARIMA

with seasonality

seasonal

regular

$\oplus =$



1 period.