DATA SCIENCE INTERVIEW PREPARATION (30 Days of Interview Preparation)

Day13

Q1. What is Autoregression?

Answer:

The autoregressive (AR) model is commonly used to model time-varying processes and solve problems in the fields of natural science, economics and finance, and others. The models have always been discussed in the context of random process and are often perceived as statistical tools for time series data.

A regression model, like linear regression, models an output value which are based on a linear combination of input values.

Example: y^ = b0 + b1*X1

Where **y^** is the prediction, **b0** and **b1** are coefficients found by optimising the model on training data, and X is an input value.

This model technique can be used on the time series where input variables are taken as observations at previous time steps, called lag variables.

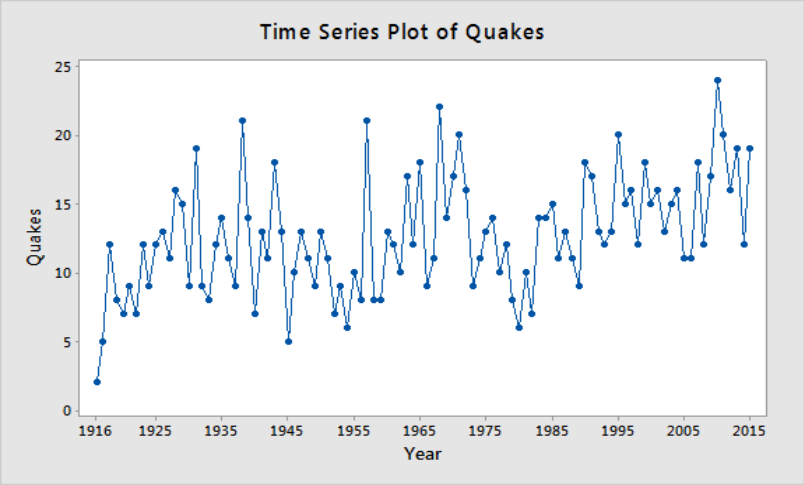
For example, we can predict the value for the next time step (t+1) given the observations at the last two time steps (t-1 and t-2). As a regression model, this would look as follows:

X(t+1) = b0 + b1*X(t-1) + b2*X(t-2)-

Because the regression model uses the data from the same input variable at previous time steps, it is referred to as an autoregression.

The notation AR(p) refers to the autoregressive model of order p. The AR(p) model is written

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t.$$



Q2. What is Moving Average?

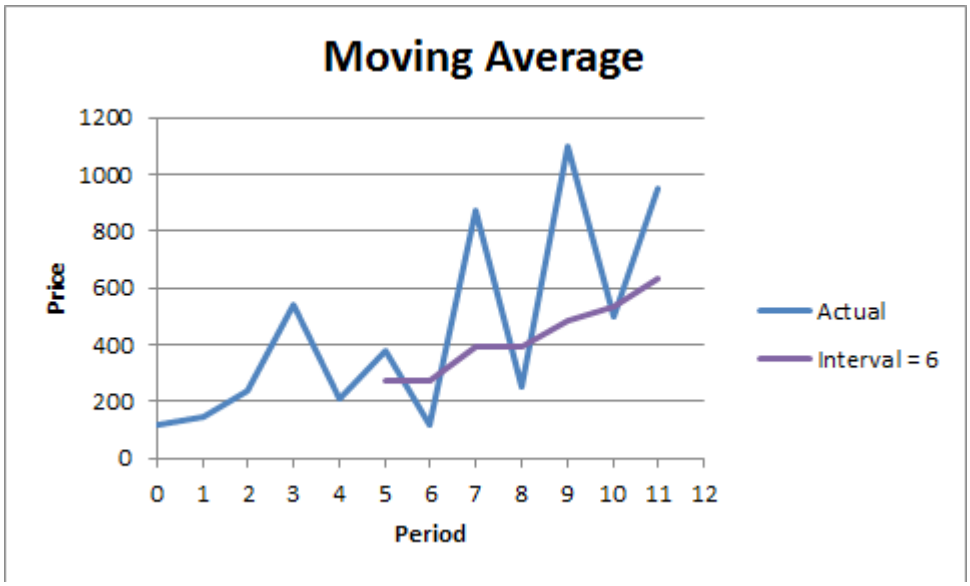
Answer:

Moving average: From a dataset, we will get an overall idea of trends by this technique; it is an average of any subset of numbers. For forecasting long-term trends, the moving average is extremely useful for it. We can calculate it for any period. For example: if we have sales data for twenty years, we can calculate the five-year moving average, a four-year moving average, a three-year moving average and so on. Stock market analysts will often use a 50 or 200-day moving average to help them see trends in the stock market and (hopefully) forecast where the stocks are headed.

The notation MA(q) refers to the moving average model of order q:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

where the $\theta_1, \ldots, \theta_q$ are the parameters of the model, μ is the expectation of X_t (often assumed to equal 0), and the $\varepsilon_t, \varepsilon_{t-1}, \ldots$ are again, white noise error terms.



The notation MA(q) refers to the moving average model of order q:

Q3. What is Autoregressive Moving Average (ARMA)?

Answer:

ARMA: It is a model of forecasting in which the methods of autoregression (AR) analysis and moving average (MA) are both applied to time-series data that is well behaved. In ARMA it is assumed that the time series is stationary and when it fluctuates, it does so uniformly around a particular time.

AR (Autoregression model)-

Autoregression (AR) model is commonly used in current spectrum estimation.

The following is the procedure for using ARMA.

- Selecting the AR model and then equalizing the output to equal the signal being studied if the input is an impulse function or the white noise. It should at least be good approximation of signal.
- Finding a model's parameters number using the known autocorrelation function or the data .
- Using the derived model parameters to estimate the power spectrum of the signal.

Moving Average (MA) model-

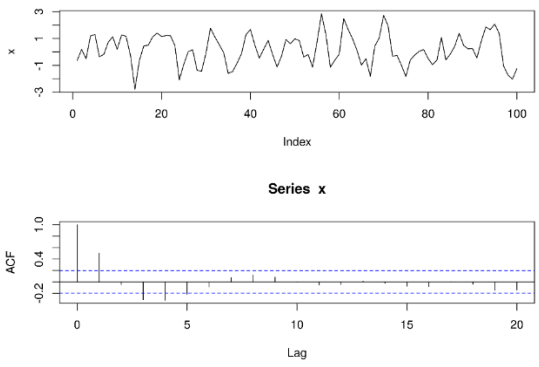
It is a commonly used model in the modern spectrum estimation and is also one of the methods of the model parametric spectrum analysis. The procedure for estimating MA model's signal spectrum is as follows.

- Selecting the MA model and then equalising the output to equal the signal understudy in the case where the input is an impulse function or white noise. It should be at least a good approximation of the signal.
- Finding the model's parameters using the known autocorrelation function.
- Estimating the signal's power spectrum using the derived model parameters.

In the estimation of the ARMA parameter spectrum, the AR parameters are first estimated, and then the MA parameters are estimated based on these AR parameters. The spectral estimates of the ARMA model are then obtained. The parameter estimation of the MA model is, therefore often calculated as a process of ARMA parameter spectrum association.

The notation ARMA(p, q) refers to the model with p autoregressive terms and q moving-average terms. This model contains the AR(p) and MA(q) models,

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}.$$



Q4. What is Autoregressive Integrated Moving Average (ARIMA)?

Answer:

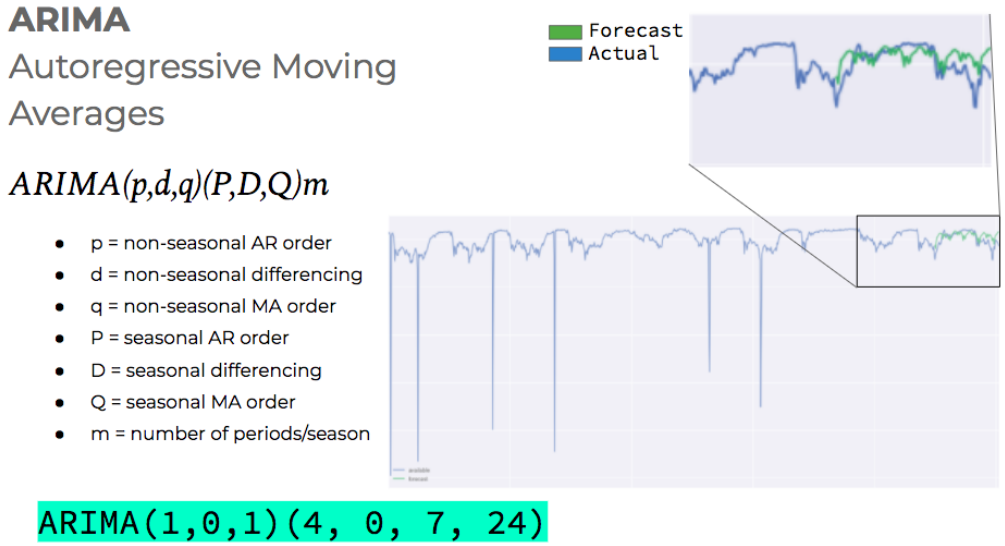
ARIMA: It is a statistical analysis model that uses time-series data to either better understand the data set or to predict future trends.

An ARIMA model can be understood by the outlining each of its components as follows-

- **Autoregression (AR)***:* It refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.
- Integrated (I): It represents the differencing of raw observations to allow for the time series to become stationary, i.e., data values are replaced by the difference between the data values and the previous values.
- Moving average (MA): It incorporates the dependency between an observation and the residual error from the moving average model applied to the lagged observations.

Each component functions as the parameter with a standard notation. For ARIMA models, the standard notation would be the ARIMA with p, d, and q, where integer values substitute for the parameters to indicate the type of the ARIMA model used. The parameters can be defined as-

- *p*: It the number of lag observations in the model; also known as the lag order.
- *d*: It the number of times that the raw observations are differenced; also known as the degree of differencing.
- q: It the size of the moving average window; also known as the order of the moving average.



Q5.What is SARIMA (Seasonal Autoregressive Integrated Moving-Average)?

Answer:

Seasonal ARIMA: It is an extension of ARIMA that explicitly supports the univariate time series data with the seasonal component.
It adds three new hyper-parameters to specify the autoregression (AR), differencing (I) and the moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

Configuring the SARIMA requires selecting hyperparameters for both the trend and seasonal elements of the series.

Trend Elements
Three trend elements requires the configuration.
They are same as the ARIMA model, specifically-

p: It is Trend autoregression order.
d: It is Trend difference order.
q: It is Trend moving average order.

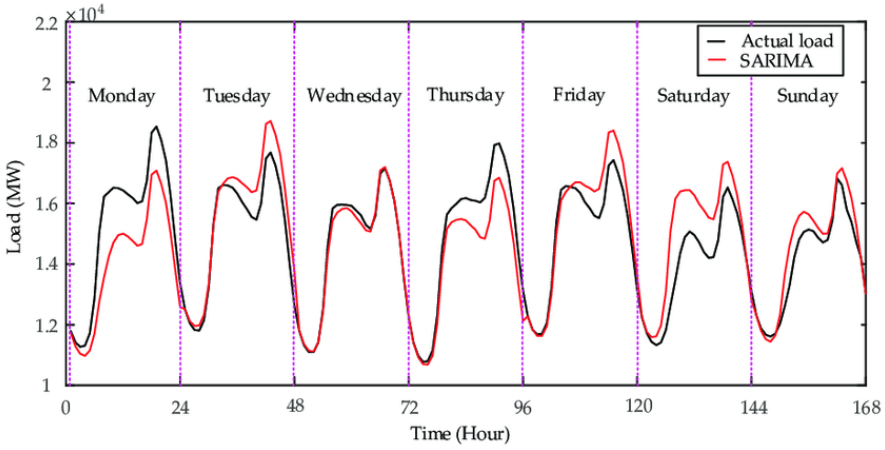
Seasonal Elements-

Four seasonal elements are not the part of the ARIMA that must be configured, they are-
P: It is Seasonal autoregressive order.
D: It is Seasonal difference order.
Q: It is Seasonal moving average order.
m: It is the number of time steps for the single seasonal period.

Together, the notation for the SARIMA model is specified as-

SARIMA(p,d,q)(P,D,Q)m-

The elements can be chosen through careful analysis of the ACF and PACF plots looking at the correlations of recent time steps.



Q6. What is Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors (SARIMAX) ?

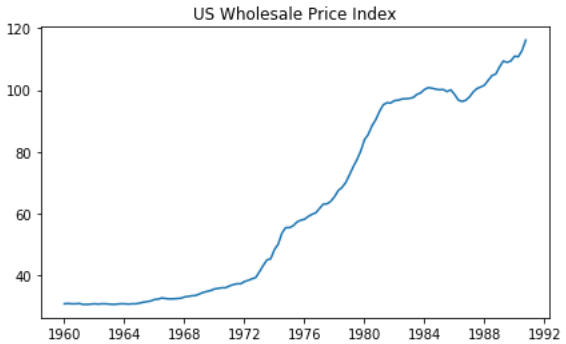
Answer:

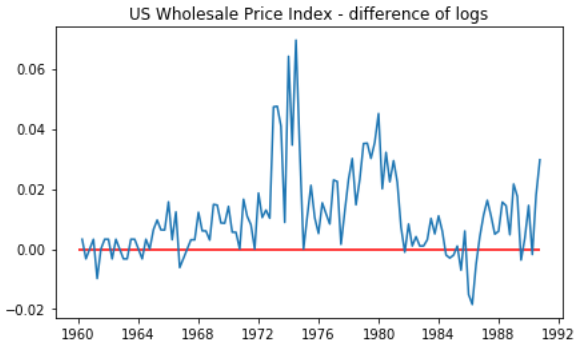
SARIMAX: It is an extension of the SARIMA model that also includes the modelling of the exogenous variables.

Exogenous variables are also called the covariates and can be thought of as parallel input sequences that have observations at the same time steps as the original series. The primary series may be referred as endogenous data to contrast it from exogenous sequence(s). The observations for exogenous variables are included in the model directly at each time step and are not modeled in the same way as the primary endogenous sequence (e.g. as an AR, MA, etc. process).

The SARIMAX method can also be used to model the subsumed models with exogenous variables, such as ARX, MAX, ARMAX, and ARIMAX.

The method is suitable for univariate time series with trend and/or seasonal components and exogenous variables.





Q7. What is Vector autoregression (VAR)?

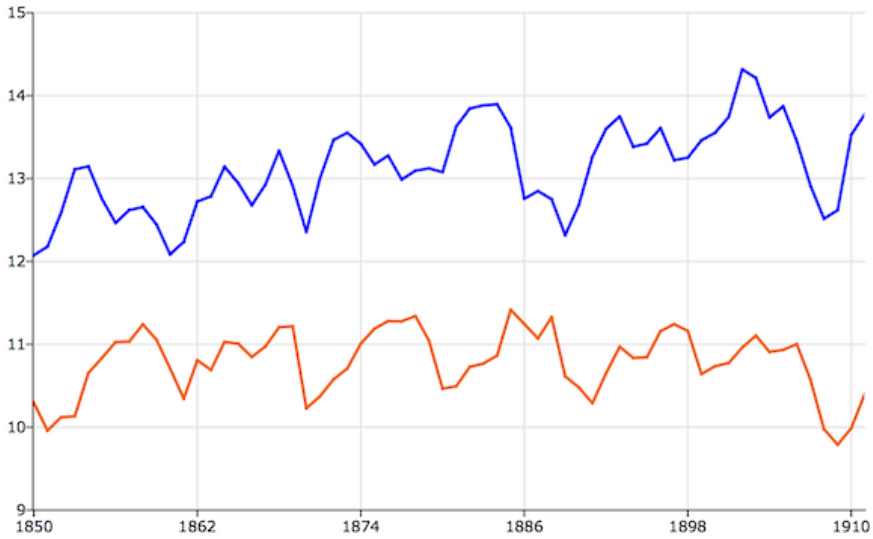
Answer:

VAR: It is a stochastic process model used to capture the linear interdependencies among multiple time series. VAR models generalise the univariate autoregressive model (AR model) by allowing for more than one evolving variable. All variables in the VAR enter the model in the same way: each variable has an equation explaining its evolution based on its own lagged values, the lagged values of the other model variables, and an error term. VAR modelling does not requires as much knowledge about the forces influencing the variable as do structural models with simultaneous equations: The only prior knowledge required is a list of variables which can be hypothesised to affect each other intertemporally.

A VAR model describes the evolution of the set of k variables over the same sample period ($t = 1, ..., T$) as the linear function of only their past values. The variables are collected in the k-vector (($k \times 1$)-matrix) y_t, , which has as the (i^{th})element, $y_{i,t}$, the observation at time t of the (i^{th})variable. Example: if the (i^{th})variable is the GDP, then $y_{i,t}$ is the value of GDP at time *"t"*.

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + e_t,$$

where the observation y_{t-i} is called the (i-th) ***lag*** of y, c is the k-vector of constants (intercepts), A_i is a time-invariant ($k \times k$)-matrix, and e_t is a k-vector of error terms satisfying.



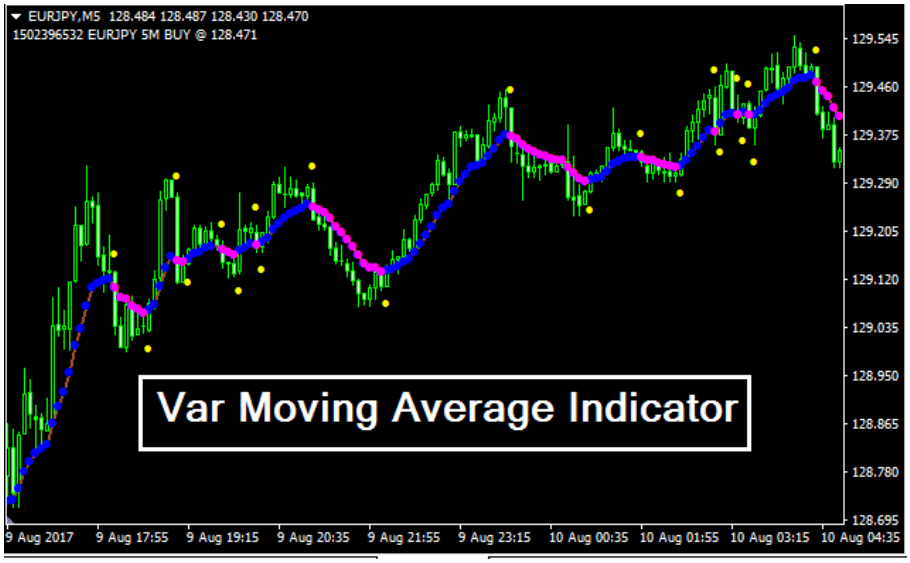
Q8. What is Vector Autoregression Moving-Average (VARMA)?

Answer:

VARMA: It is method models the next step in each time series using an ARMA model. It is the generalisation of ARMA to multiple parallel time series, Example- multivariate time series.

The notation for a model involves specifying the order for the AR(p) and the MA(q) models as parameters to the VARMA function, e.g. VARMA (p, q). The VARMA model can also be used to develop VAR or VMA models.

This method is suitable for multivariate time series without trend and seasonal components.



Q9. What is Vector Autoregression Moving-Average with Exogenous Regressors (VARMAX)?

Answer:

VARMAX: It is an extension of the VARMA model that also includes the modelling of the exogenous variables. It is the multivariate version of the ARMAX method.

Exogenous variables are also called the covariates and can be thought of as parallel input sequences that have observations at the same time steps as the original series. The primary series(es) are referred as the endogenous data to contrast it from the exogenous sequence(s). The observations for the exogenous variables are included in the model directly at each time step and are not modeled in the same way as the primary endogenous sequence (Example- as an AR, MA, etc.).

This method can also be used to model subsumed models with exogenous variables, such as VARX and the VMAX.

This method is suitable for multivariate time series without trend and seasonal components and exogenous variables.

Q10. What is Simple Exponential Smoothing (SES)?

Answer:

SES: It method models the next time step as an exponentially weighted linear function of observations at prior time steps.

This method is suitable for univariate time series without trend and seasonal components.

Exponential smoothing is the rule of thumb technique for smoothing time series data using the exponential window function. Whereas in the simple moving average, the past observations are weighted equally, exponential functions are used to assign exponentially decreasing weights over time. It is easily learned and easily applied procedure for making some determination based on prior assumptions by the user, such as seasonality. Exponential smoothing is often used for analysis of time-series data.

Exponential smoothing is one of many window functions commonly applied to smooth data in signal processing, acting as low-pass filters to remove high-frequency noise.

The raw data sequence is often represented by {xt} beginning at time t = 0, and the output of the exponential smoothing algorithm is commonly written as {st} which may be regarded as a best estimate of what the next value of x will be. When the sequence of observations begins at time t= 0, the simplest form of exponential smoothing is given by the formulas:

$$s_0 = x_0$$
$$s_t = \alpha x_t + (1-\alpha)s_{t-1},\ t > 0$$

where α is the *smoothing factor*, and $0 < \alpha < 1$.