

Time Series

ARIMA

Agenda

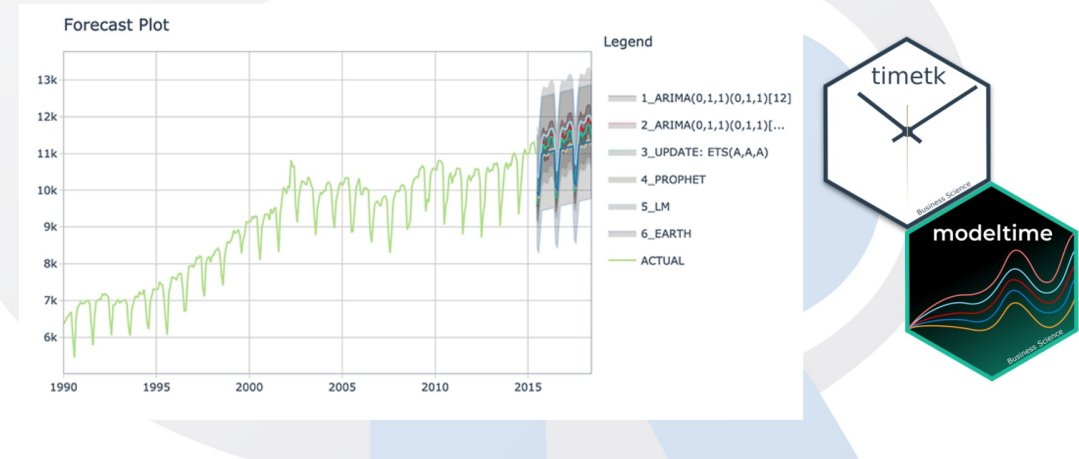
- Introduction
- White noise
- Random Walk
- Stationarity
- Seasonality
- Holt Winter – Exponential Smoothing
- Recap AR and MA
- Disadvantages of AR and MA
- ARIMA
- ARIMAX
- SARIMAX

Time Series Introduction

- We're going to discover how it differs from other types of data you've previously encountered and why
- Time series is a sequence of information which attaches a time period to each value
- The value can be pretty much anything measurable.
- It depends on time in some way, like prices, humidity or number of people.
- As long as the values we record are unambiguous, any medium could be measured with Time series.
- There aren't any limitations regarding the total time span of our Time series.
- It could be a minute, a day, a month or even a century.
- All we need is a starting and an ending point.

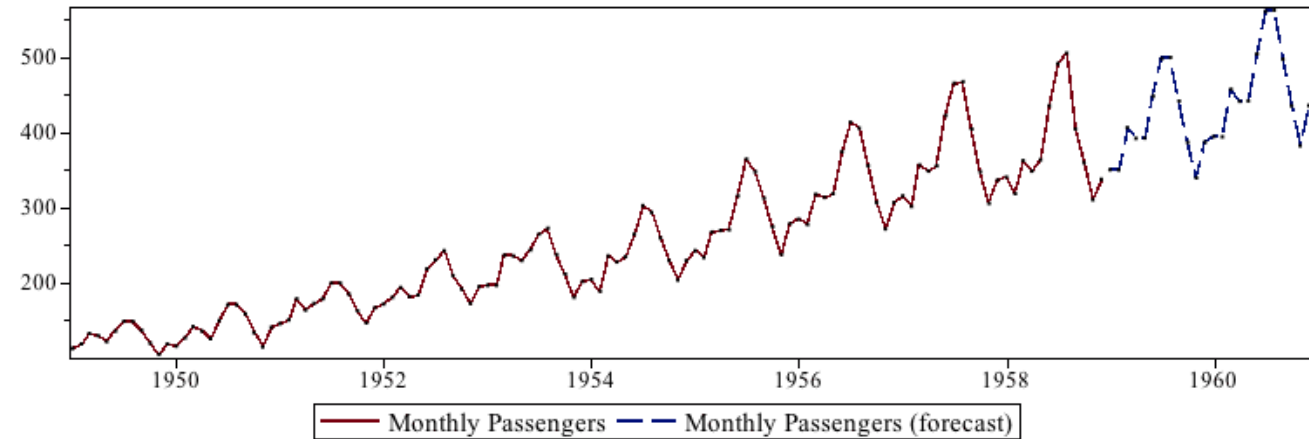
Time Series Modeling

Visualize, wrangle, and preprocess time series data



Time Series Basics

- Chronological Data
- Cannot be shuffled
- Each row indicate specific time record
- Train – Test split happens chronologically
- Data is analyzed univariately (for given use case)
- Nature of the data represents if it can be predicted or not

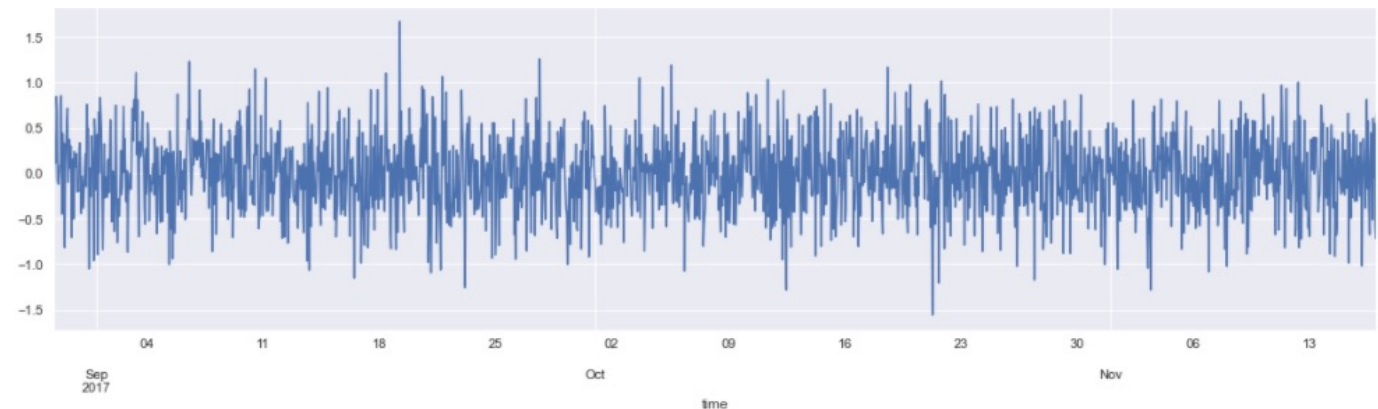


	IOT_Sensor_Reading	Error_Present	Sensor_2	Sensor_Value
time				
2017-08-29 11:00:00	-0.015871	0.353986	-0.787655	0.008144
2017-08-29 12:00:00	-0.101576	0.353986	-0.787655	-0.029860
2017-08-29 13:00:00	-0.118241	0.353986	-0.787655	-0.021717
2017-08-29 14:00:00	-0.214262	0.353986	-0.787655	0.008144
2017-08-29 15:00:00	-0.249972	0.353986	-0.787655	-0.108583

Data Examining & Preprocessing

- Reading Data using Pandas – describe, head
- Check for null values
- Line plot for each feature
- Convert Date from String to Date time feature
- Setting the desired frequency
- Handling missing Values
- QQ plots

	IOT_Sensor_Reading	Error_Present	Sensor_2	Sensor_Value
time				
2017-08-29 11:00:00	-0.015871	0.353986	-0.787655	0.008144
2017-08-29 12:00:00	-0.101576	0.353986	-0.787655	-0.029860
2017-08-29 13:00:00	-0.118241	0.353986	-0.787655	-0.021717
2017-08-29 14:00:00	-0.214262	0.353986	-0.787655	0.008144
2017-08-29 15:00:00	-0.249972	0.353986	-0.787655	-0.108583



White Noise

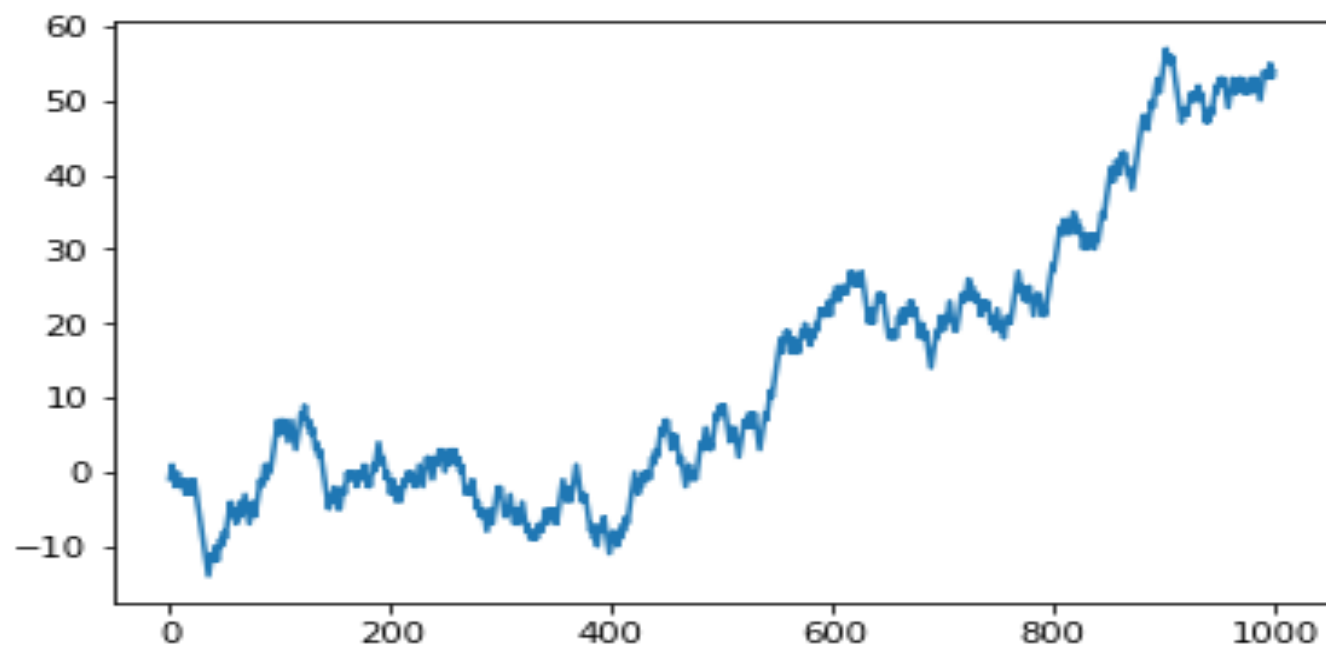
A time series is not white noise if one or more of the following conditions are true:

- Is the mean/level non-zero?
- Does the mean/level change over time?
- Does the variance change over time?
- Do values correlate with lag values?

White noise time series is defined by a **zero mean, constant variance, and zero correlation**. If our time series is white noise, it cannot be predicted/modeled.

Random Walk

A *random walk* is a time series model x_t such that $x_t = x_{t-1} + w_t$, where w_t is a discrete white noise series.



- A random walk is another time series model where the current observation is equal to the previous observation with a random step up or down.
- Random walk is not Stationary
- If we plot the first-order difference of a time series and the result is white noise, then it is a random walk.

Stationarity in Time Series

The stationary time-series is the one with :

- no trend
- Fluctuates around the constant mean and has constant variance.
- Covariance between different lags is constant, it doesn't depend on absolute location in time-series.
- Inferences become easy
- Is **strong stationary** if the distribution of a time-series is exactly the same trough time

How to check for stationarity

- Look at Plots: We can review a time series plot of the data and visually check if there are any obvious trends or seasonality.
- Summary Statistics: We can review the summary statistics for data for seasons or random partitions and check for obvious or significant differences.
- Statistical Tests: We can use statistical tests to check if the expectations of stationarity are met or have been violated. Two mainly used statistical tests are:
 - Augmented Dickey-Fuller test
 - The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test

Augmented Dickey-Fuller test(ADFFuller)

- The Augmented Dickey-Fuller test is a type of statistical test called a unit root test.
- The intuition behind a unit root test is that it determines how strongly a time series is defined by a trend.
- In python, the statsmodel package is used for the implementation of AD Fuller and KPSS test.
- The null hypothesis of the test is that the time series can be represented by a unit root, that it is not stationary.
- The alternate hypothesis (rejecting the null hypothesis) is that the time series is stationary.
- Result is interpreted using the p-value from the test.
 - **p-value > 0.05: Fail to reject the null hypothesis (H_0), the data has a unit root and is non-stationary.**
 - **p-value \leq 0.05: Reject the null hypothesis (H_0), the data does not have a unit root and is stationary.**

Patterns in Time- Series

- Any time series may be split into the following components:

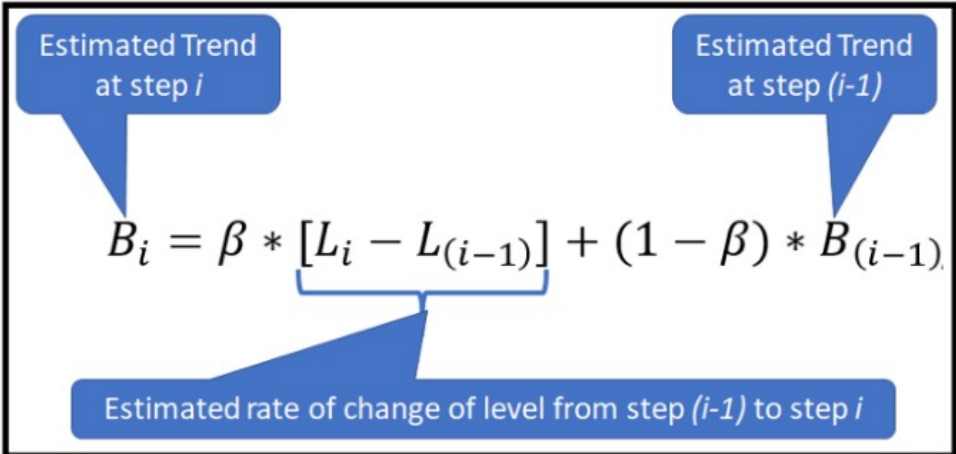
$$\text{Base Level} + \text{Trend} + \text{Seasonality} + \text{Error}$$

- A **Trend** is observed when there is an increasing or decreasing slope observed in the time series.
- **Seasonality** is observed when there is a distinct repeated pattern observed between regular intervals due to seasonal factors. It could be because of the month of the year, the day of the month, weekdays or even time of the day.
- A time series can be modeled as an additive or multiplicative, wherein, each observation in the series can be expressed as either a sum or a product of the components:
 - Additive time series: $\text{Value} = \text{Base Level} + \text{Trend} + \text{Seasonality} + \text{Error}$
 - Multiplicative Time Series: $\text{Value} = \text{Base Level} \times \text{Trend} \times \text{Seasonality} \times \text{Error}$

- We can do decomposition of a time series by considering the series as an additive or multiplicative combination of the base level, trend, seasonal index and the residual.
 - **The seasonal_decompose in statsmodels** is used to implement this and we can extract the components.
- * We are using Additive Decomposition in our project**

Holt Winter Exponential Smoothing

- Holt-Winters Exponential Smoothing is used for forecasting time series data that exhibits both a trend and a seasonal variation
- The Exponential Smoothing (ES) technique forecasts the next value using a weighted average of all previous values where the weights decay exponentially from the most recent to the oldest historical value.
- When you use ES, you are making the crucial assumption that recent values of the time series are much more important to you than older values.
- The ES technique has two big shortcomings: It cannot be used when your data exhibits a trend and/or seasonal variations.
- The Holt-Winters ES modifies the Holt ES technique so that it can be used in the presence of both trend and seasonality.

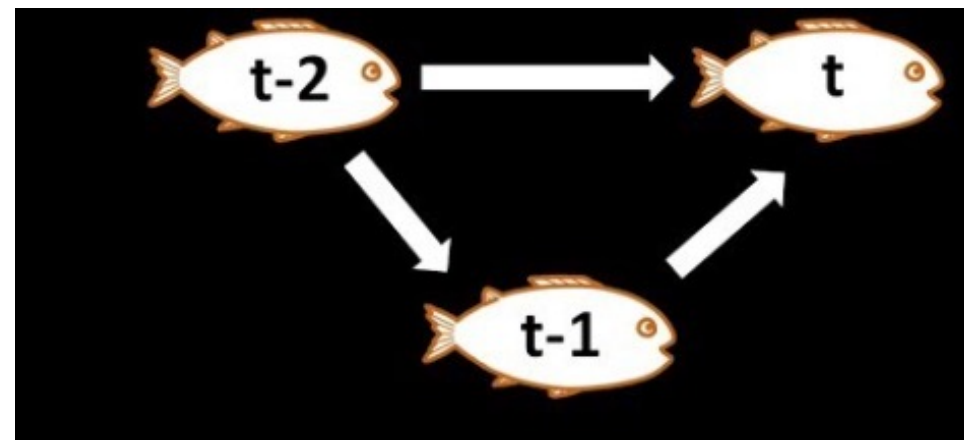


The diagram illustrates the formula for the estimated trend at step i . It features a central equation with callouts identifying its components. A blue callout at the top left points to B_i and is labeled "Estimated Trend at step i ". A blue callout at the top right points to $B_{(i-1)}$ and is labeled "Estimated Trend at step $(i-1)$ ". A blue callout at the bottom points to the term $[L_i - L_{(i-1)}]$ and is labeled "Estimated rate of change of level from step $(i-1)$ to step i ".

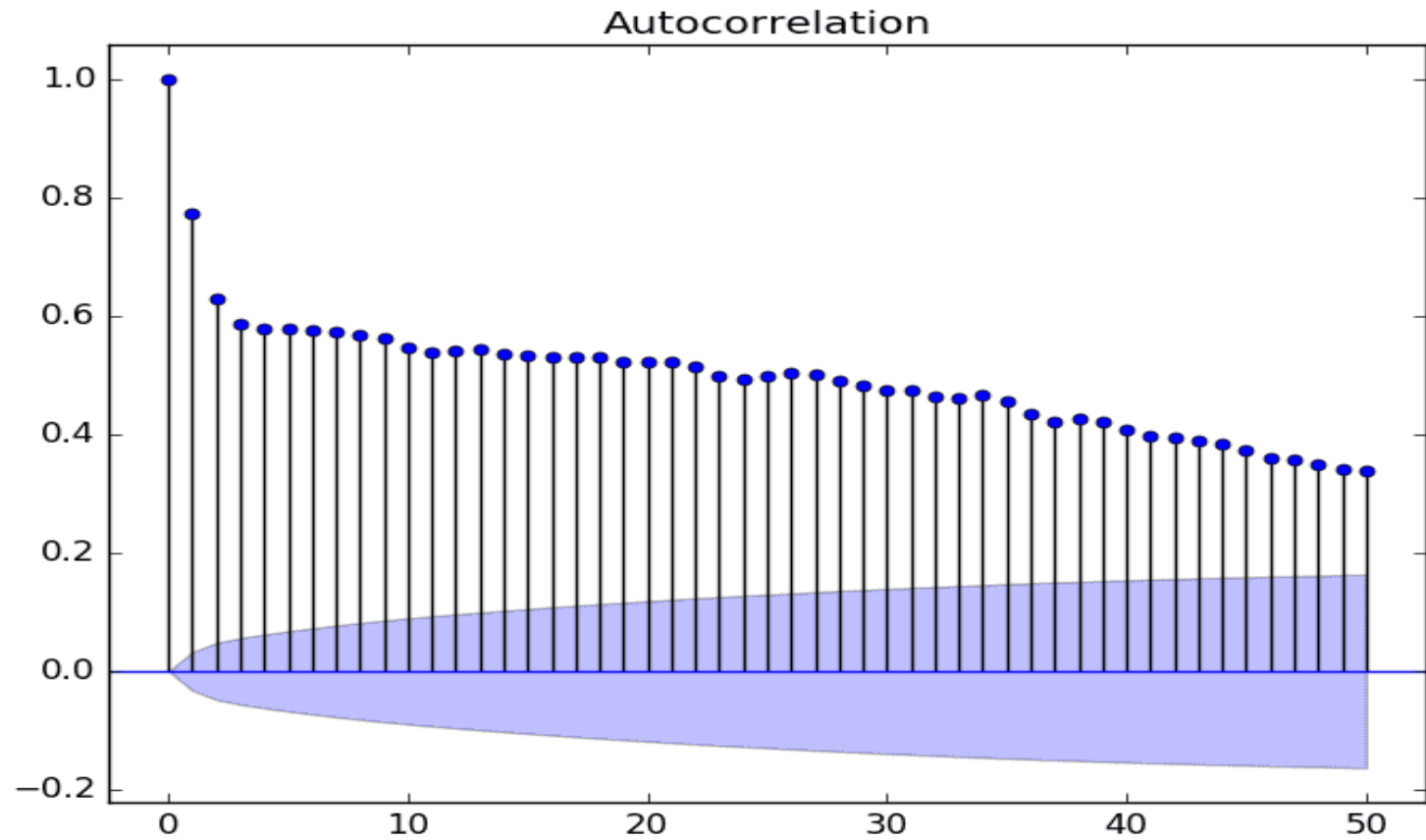
$$B_i = \beta * [L_i - L_{(i-1)}] + (1 - \beta) * B_{(i-1)}$$

ACF

- ACF is an (complete) auto-correlation function which gives us values of auto-correlation of any series with its lagged values
- It describes how well the present value of the series is related with its past values.
- A time series can have components like trend, seasonality, cyclic and residual. ACF considers all these components while finding correlations hence it's a 'complete auto-correlation plot'.

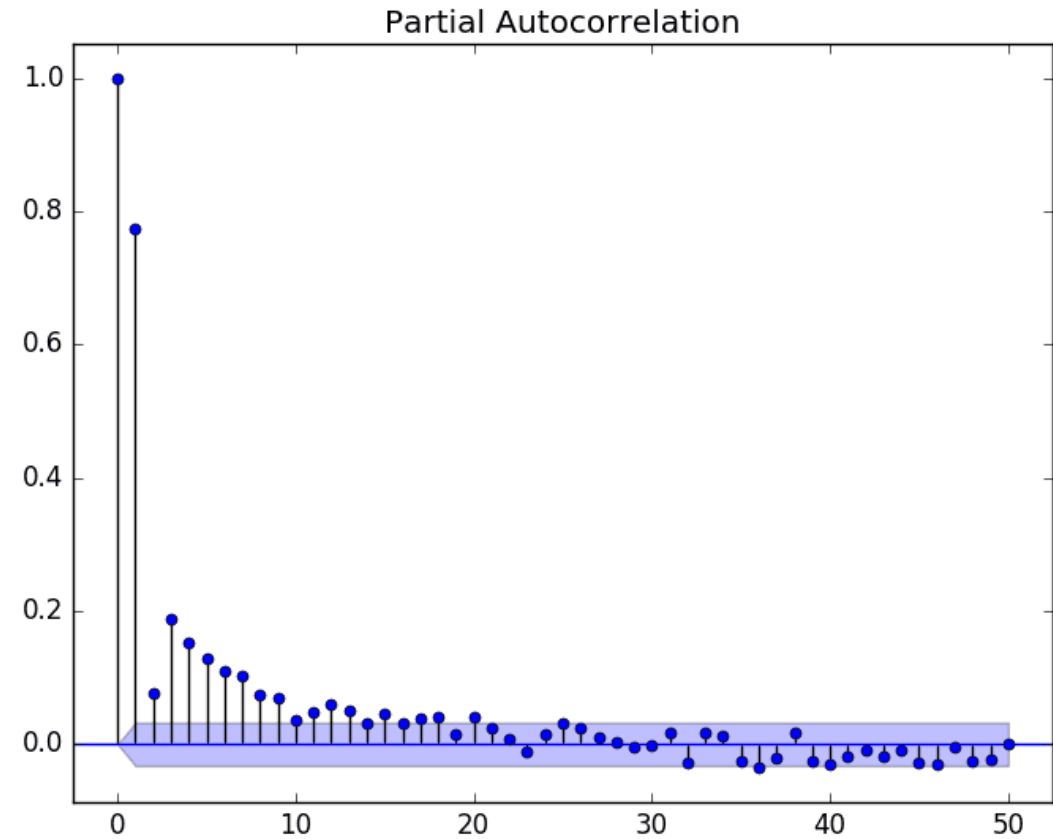


ACF Plot Example:



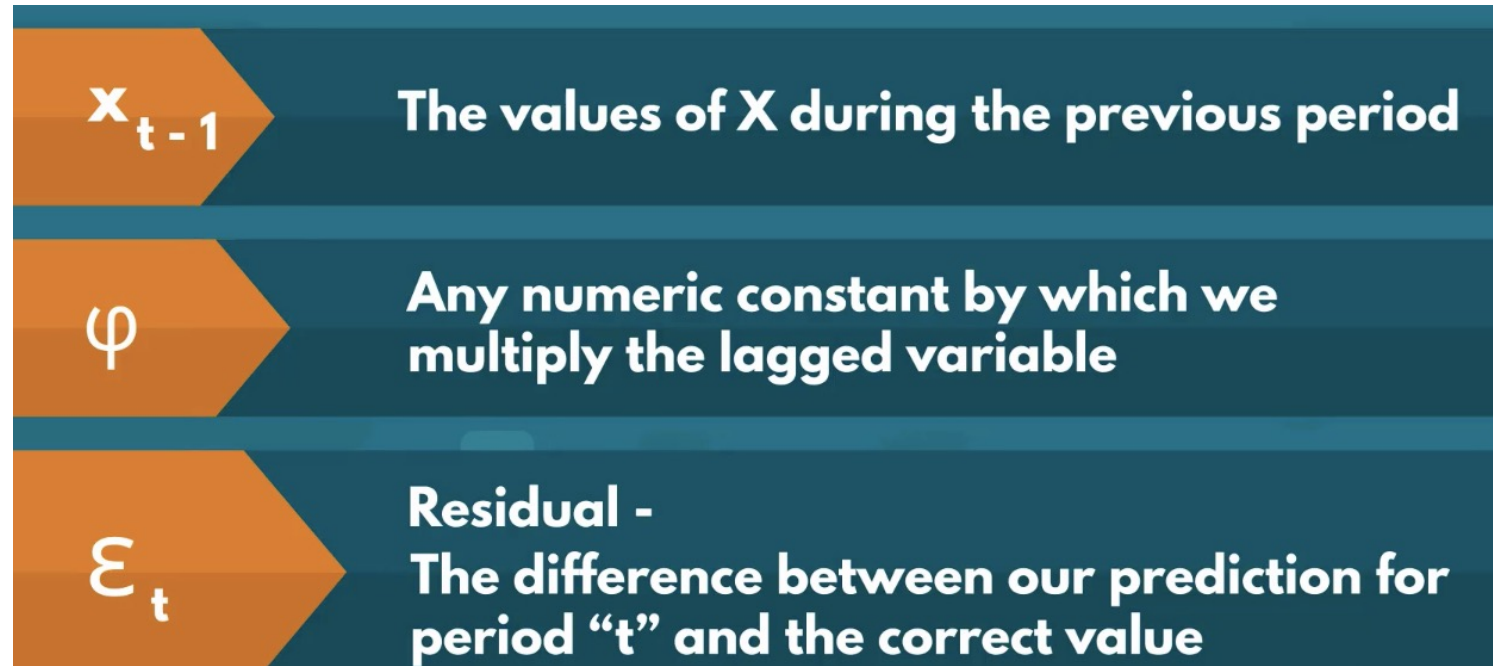
PACF

- It finds correlation of the residuals ,which remain after removing the effects which are already explained by the earlier lags.



Autoregression AR Time Series

- By knowing the prices today, we can often make a rough prediction about prices tomorrow.
- The autoregressive model or a model for short, relies on past period values and past periods only to predict current period values.
- It's a linear model where current period values are a sum of past outcomes multiplied by a numeric factor.
- Value of phi lies between -1 and 1



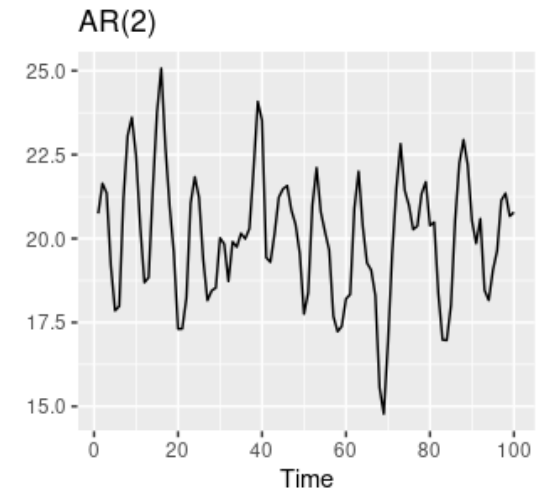
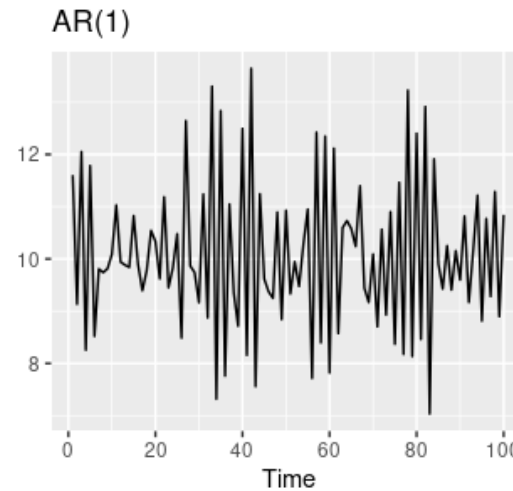
$$x_t = C + \varphi x_{t-1} + \varepsilon_t$$

Moving Average

- Moving averages are a simple and common type of smoothing used in time series analysis and time series forecasting.
- Calculating a moving average involves creating a new series where the values are comprised of the average of raw observations in the original time series.
- A moving average requires that you specify a window size called the window width. This defines the number of raw observations used to calculate the moving average value.

Limitations of AR and MA

- It is assumed that both trend and seasonal components have been removed from your time series.
- This means that your time series is stationary, or does not show obvious trends (long-term increasing or decreasing movement) or seasonality (consistent periodic structure).
- Hence in order to model ARMA models we need to ensure stationarity for better predictions



ARIMA

- Auto Regressive Integrated Moving Average (ARIMA) model is among one of the more popular and widely used statistical methods for time-series forecasting.
- It is a class of statistical algorithms that captures the standard temporal dependencies that is unique to a time series data.
- ARIMA is an acronym for “autoregressive integrated moving average.” It’s a model used in statistics and econometrics to measure events that happen over a period of time.
- The model is used to understand past data or predict future data in a series.
- It’s used when a metric is recorded in regular intervals, from fractions of a second to daily, weekly or monthly periods.



Breaking AR,I,MA

- **Auto Regressive (AR)** regression model is built on top of the autocorrelation concept, where the dependent variable depends on the past values of itself.
- **Integrated(I)**: The integrated part of ARIMA attempts to convert the non-stationarity nature of the time-series data to something a little bit more stationary. By performing prediction on the **difference** between any two pair of observation rather than directly on the data itself.
- **Moving Average(MA)** attempts to reduce the noise in our time series data by performing some sort of aggregation operation to your past observations in terms of residual error.

A single factor of integration is enough to reach stationarity



Integration



Advantages of ARIMA

- The ARIMA is just an integrated version of the ARMA model. What that means is, we simply integrate the data (however many times is needed) to get a stationary set.
- An autoregressive model is a regression, just one in which some of the independent variables are past values of the dependent variable.
- ARIMA combines autoregressive with moving average terms and can adjust to different levels of integration. These are all adjustments you could make within a regression model.
- So there's no real theoretical difference between AR-MA and ARIMA.
- The practical difference is ARIMA packages are built to assume time series data, whereas most regression packages make no special allowances for time-dependent data.

How ARIMA Works

ARIMA (1, 1, 1) $\Delta P_t = c + \varphi_1 \Delta P_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$

P_t, P_{t-1}

Values in the current period and 1 period ago respectively

$\varepsilon_t, \varepsilon_{t-1}$

Error terms for the same two periods

c

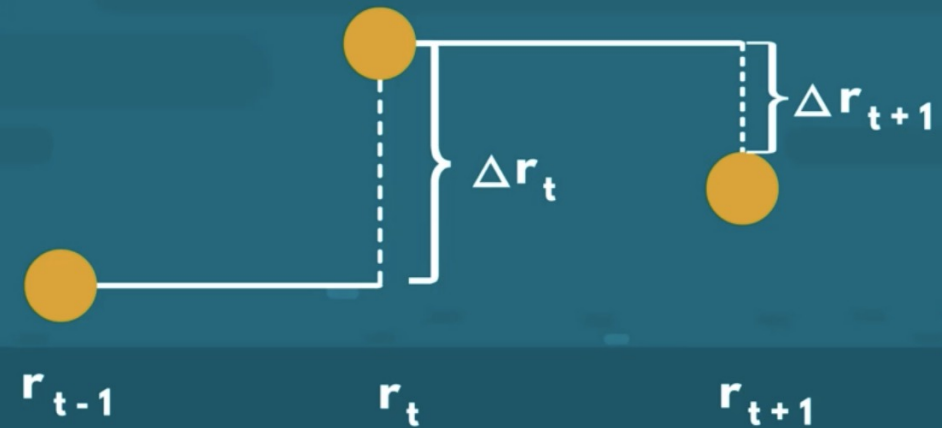
Baseline constant factor

φ_1

What part of the value last period is relevant in explaining the current one

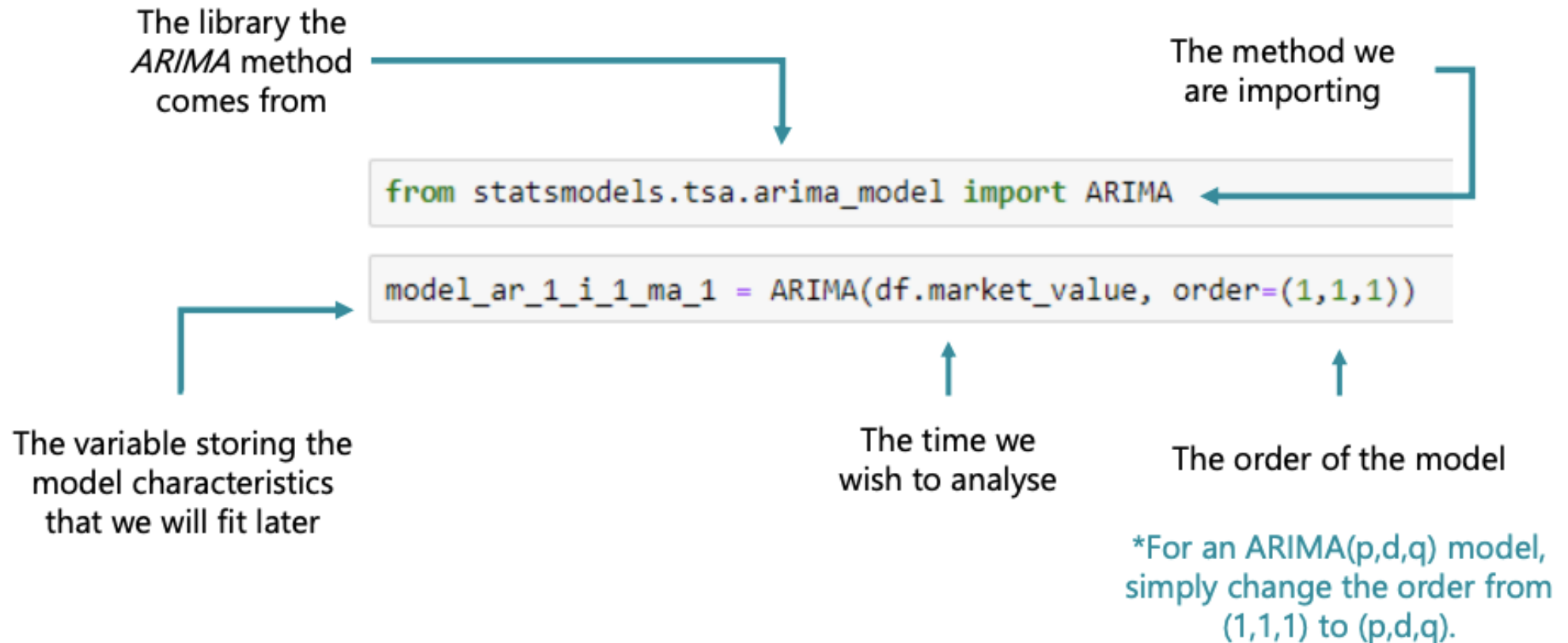
θ_1

What part of the error last period is relevant in explaining the current value



How ARIMA Works

Implementation of the Simple Model in Python:



How ARIMA Works

- We will start with the single level of integration, and fit the basic ARIMA (1, 1, 1).
- Residual of the ARIMA fitting model explains till what number of lags our model is able to predict the series. Hence, residual for near about lags should be insignificant.
- This means that the residual should follow a White noise and in order to analyze that we will plot ACF graph to see the significance of residual till each lag.
- If the near about lag is significant we will increase the number the p (AR) and q (MA) component and check the significance in the same way
- We will also see the significance of nested models just to get the best fit on data

ARIMAX

- The names ARMAX and ARIMAX come as extensions of the ARMA and ARIMA respectively.
- The X added to the end stands for “exogenous”. In other words, it suggests adding a separate different outside variable to help measure our endogenous variable.
- It can be a time-varying measurement like the inflation rate or the price of a different index. Or a categorical variable separating the different days of the week. It can also be a Boolean accounting for the special festive periods. Finally, it can stand for a combination of several different external factors.

SARIMAX

- SARIMAX(Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors) is an updated version of the ARIMA model
- ARIMA includes an autoregressive integrated moving average, while SARIMAX includes seasonal effects and eXogenous factors with the autoregressive and moving average component in the model.
- Another seasonal equivalent model holds the seasonal pattern; it can also deal with external effects. This feature of the model differs from other models.
- For example, in a time series, the temperature has seasonal effects like it is low in winter, high in summers. Still, with the effect of external factors like humidity, the temperature in winter is increased and also due to rain, there is a chance of lower temperature. We can't predict the exact value for these factors if they do not appear in a cyclic or any seasonal behaviour.