Teaching Session | Time Series Forecasting

Tamojit Maiti
Masters in Applied Statistics and Operations Research, ISI Kolkata
Data Scientist at Sixt R&D, previously at Rapido & AB InBev

Agenda

- Time Series
  - What and Why?
  - Difference with Regression
- Terminology
  - Stationarity
  - Time Series Decomposition Components
- Forecasting Techniques
  - Naïve Family
  - Moving Average Family
  - ARIMA family
  - Formulating it as a regression problem
- Evaluation metrics for forecasts
  - MAE/MAPE/SMAPE

Time Series

Time Series

- A sequence of values, indexed by time
- Occur universally everywhere, from business financials to supply and demand of practically everything
- Almost always have a pattern, which if deciphered can help us predict the future
- Prediction of future in some capacity allows businesses to plan accordingly and mitigate risks
- For example, demand prediction of retail goods helps keep the good in stock and avoid them running out

t	y
1	2
2	5
3	10
4	17
5	26
6	?

Time Series | Difference with Regression

- A (univariate) time series can be generally characterized by

$$y_t = \alpha y_{t-1} + \beta y_{t-2} + \gamma y_{t-3} + \cdots + \epsilon_t$$

which looks a lot like a regression formulation

- But in a time-series, all the values of y at time t depend on its previous values
- Whereas in regression, no value of y depends on any other value of y
- In other words, order is important in time series, and this time dependence is referred to as autocorrelation
- Whereas for regression, the values of y are said to be independent and identically distributed

Definitions

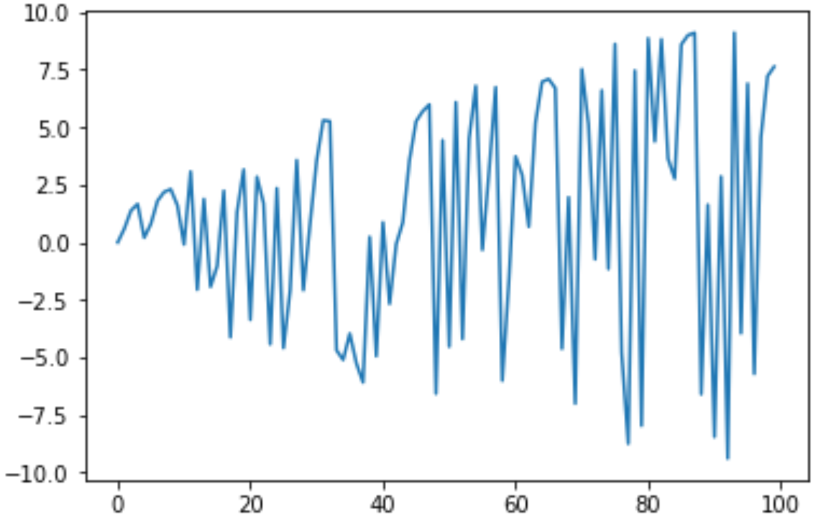
Time Series | Stationarity

- A (univariate) time series $\{y_t\}$ is said to be stationary if it has
  - Finite and independent of time Mean $E(Y) = \mu \neq f(t)$
  - Finite and independent of time Variance $E(Y^2) < \infty \neq f(t)$
  - Absence of seasonality $\gamma(s,t) = \gamma(s-h, t-h) \quad \forall \ h, s, t$

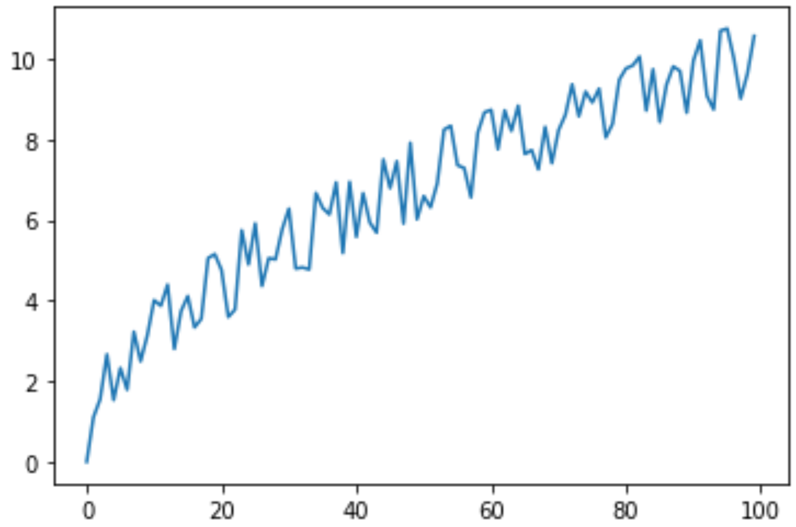
- Why?
  - Certain family of forecasting methods (ARIMA) have stationarity as a requirement before they can be applied

- Checking for Stationarity
  - Visually
  - Global vs Local Tests
  - Augmented Dickey Fuller Test

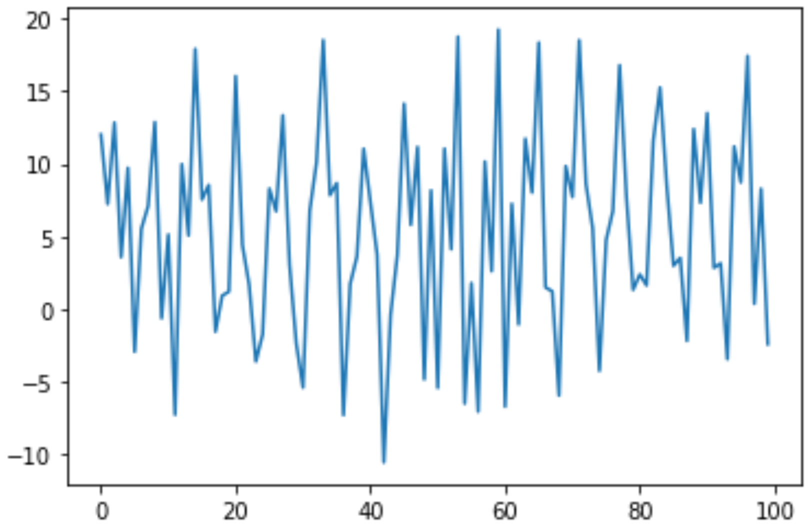
Time Series | Non – Stationary Time Series Examples



Variance is a function of time



Mean is a function of time



Seasonality is present

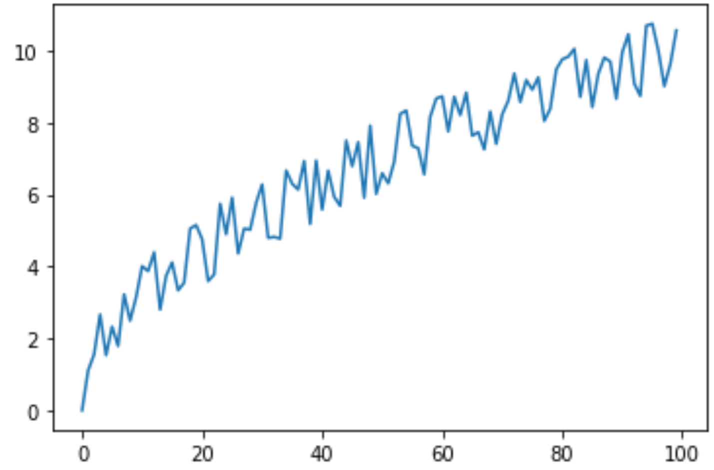
Time Series | Make a time series stationary

- If the mean is changing with time, try differencing it once, or in rare cases, twice

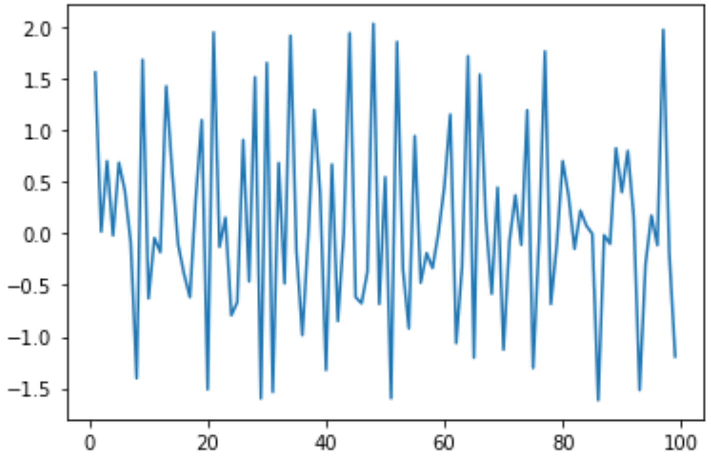
$$\Delta_t^1 = y_t - y_{t-1}$$

Differencing is akin to modelling the rate of change of y instead of modelling y natively

Physical processes rarely follow higher than second order differential equations, which is why differencing it once usually gets rid of the change in mean



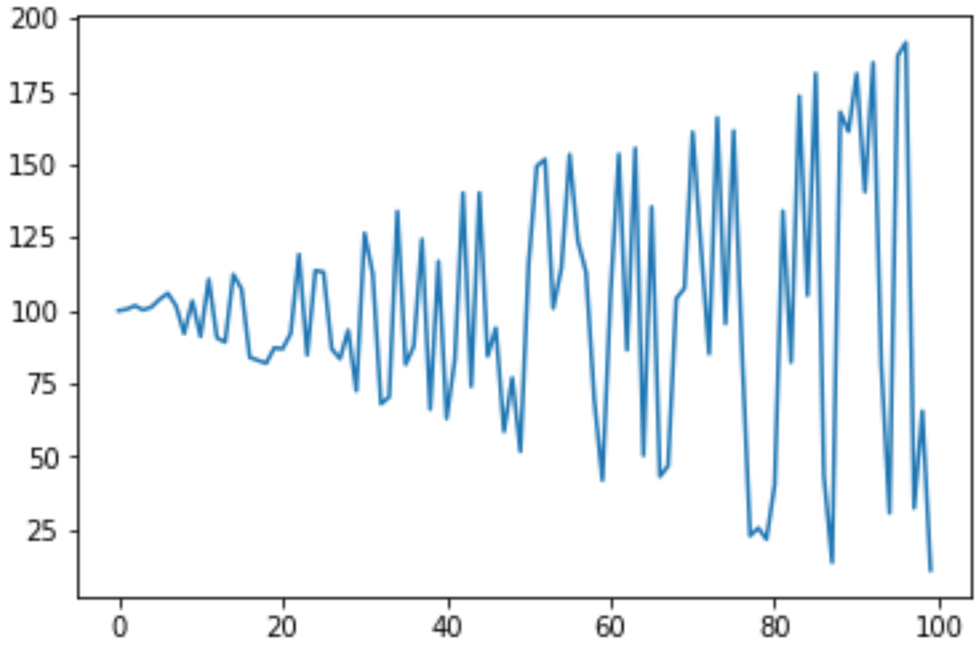
Before differencing – mean changes with time



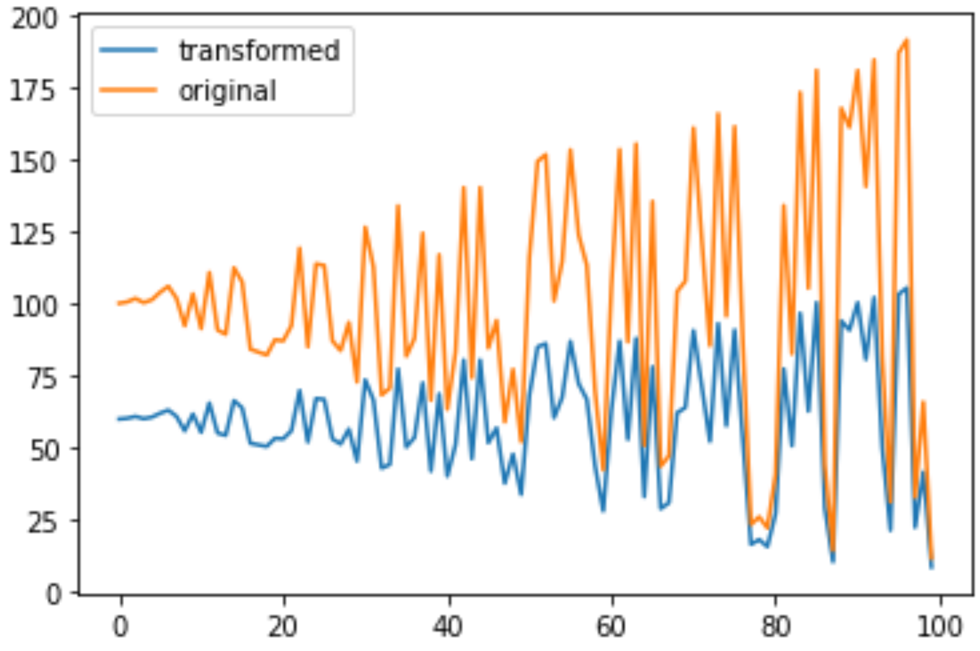
After differencing – constant mean

Time Series | Make a time series stationary

- If the variance is changing with time, try using the Box-Cox transform or the Yeo-Johnson transform. It has the effect of making the variance constant for the time series



Variance is a function of time

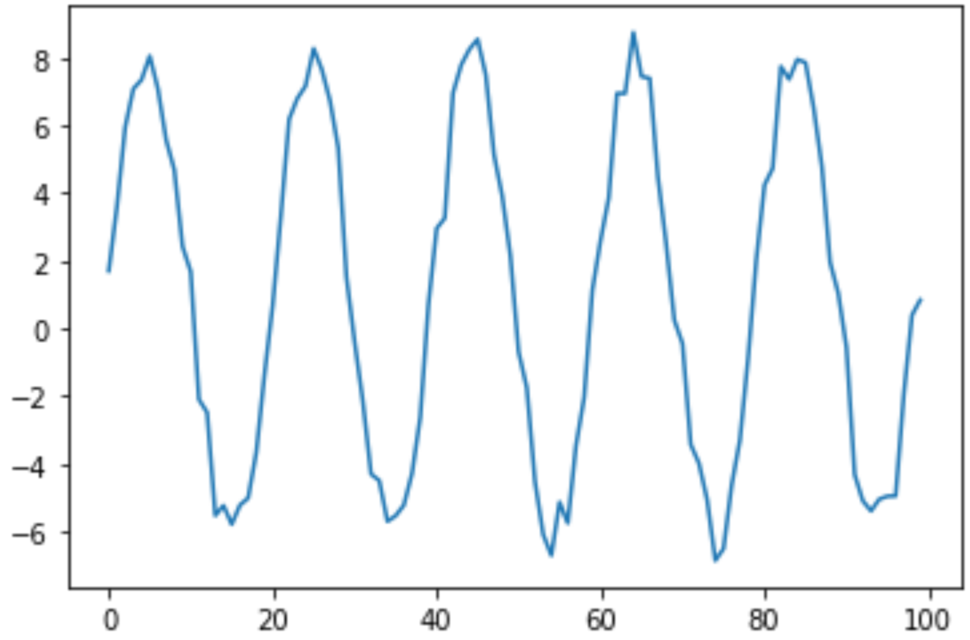


Magnitude of variance reduced, although not eliminated

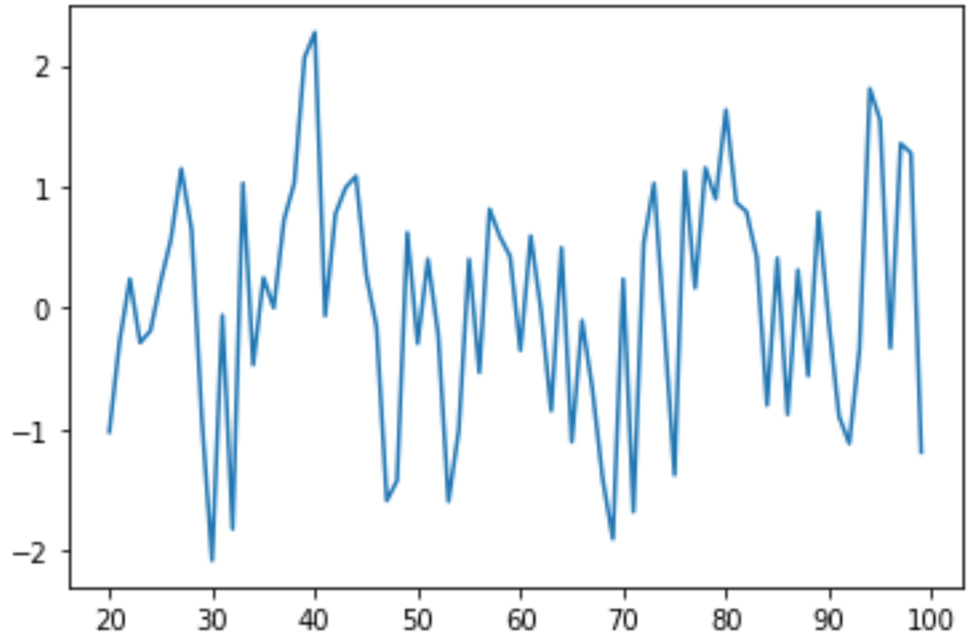
Time Series | Make a time series stationary

- If the time series has seasonality, it can be depersonalized by seasonal differencing

$$s_t = y_t - y_{t-h}$$



Before seasonal differencing



After seasonal differencing

Time Series | Components

- A time series is usually comprised of
  - Level Base value of the time series
  - Trend Long-term rate of change of the time series
  - Seasonality How frequently and strongly do the values repeat
  - Exogenous Influences Are there other time-series that affect my time series?
  - Residuals Whatever could not be modelled

$$y_t = \mu + T(t) + S(t) + WX_t + \epsilon_t$$

Time Series | Estimating Trend

- Trend is estimated by
  - Simple Moving Average
  - Exponentially Weighted Moving Average
  - Locally Weighted Scatterplot Smoothing (LOESS)

Time Series | Estimating Seasonality

- Seasonality is estimated by
  - Fourier Decomposition
  - Autocorrelation Function

Time Series | Residual Analysis

- We want to make sure that no information is left in the time series after we have extracted the various components of the time series
- We want our residuals to be as close to noise as possible
- We can test whether the residuals contain any useful information or not via a statistical test, called the Augmented Dickey Fuller Test
- Ideally, you should keep on stacking time-series models until the residuals resemble white noise

Forecasting Models

Time Series | Forecasting Models | Naive

- Naïve Forecast $y_t = y_{t-1}$

What happens tomorrow is the same as what happens today

Used as a first benchmark forecast, every model should beat the naïve forecast model

- Seasonal Naïve Forecast $y_t = y_{t-s}$

What happens tomorrow is the same as what happened last season (week, month, year)

Used as a second benchmark forecast, every model should beat the seasonal naïve forecast model

- Simple Moving Average $$y_t = \frac{1}{k}\sum_{i=1}^{k} y_{t-i}$$

What happens tomorrow is an average of what happened in the last k days

This is usually a good model if your time series contains only level and trend

- Exponentially Weighted Moving Average $$y_t = \frac{1}{k}\sum_{i=1}^{k} \alpha^i y_{t-i}$$

What happens tomorrow is a weighted average of what happened in the last k days, the weights form a geometric series

Recent points are given more weightage in computing final forecast

- AR(p) Process

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \epsilon_t$$

An AR(1) process is simply

$$y_t = \alpha_1 y_{t-1} + \epsilon_t$$

with

$$|\alpha_1| < 1 \qquad \text{(Why?)}$$

$$E(\epsilon_t) = 0 \qquad \text{(Why?)}$$

$$E(\epsilon_t \epsilon_{t-i}) = 0 \; \forall \, i \in \mathrm{N} \qquad \text{(Why?)}$$

- MA(q) Process

$$y_t = \sum_{j=1}^{q} \beta_j \epsilon_{t-j} + \epsilon_t$$

An MA(1) process is simply

$$y_t = \beta_1 \epsilon_{t-1} + \epsilon_t$$

with

$$|\beta_1| < 1 \qquad \text{(Why?)}$$

$$E(\epsilon_t) = 0 \qquad \text{(Why?)}$$

$$E(\epsilon_t \epsilon_{t-i}) = 0 \; \forall \, i \in \mathrm{N} \qquad \text{(Why?)}$$

- ARMA(p,q) Process

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{j=1}^{q} \beta_j \epsilon_{t-j} + \epsilon_t$$

An ARMA(1,1) process is simply

$$y_t = \alpha_1 y_{t-1} + \beta_1 \epsilon_{t-1} + \epsilon_t$$

with

$$|\alpha_1| , |\beta_1| < 1$$

$$E(\epsilon_t) = 0$$

$$E(\epsilon_t \epsilon_{t-i}) = 0 \; \forall \; i \in \mathrm{N}$$

- ARIMA(p,d,q) Process

$$\Delta_t^D = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{j=1}^{q} \beta_j \epsilon_{t-j} + \epsilon_t$$

An ARMA(1,1,1) process is simply

$$\Delta_t^1 = y_t - y_{t-1} = \alpha_1 y_{t-1} + \beta_1 \epsilon_{t-1} + \epsilon_t$$

with

$$|\alpha_1| , |\beta_1| < 1$$

$$E(\epsilon_t) = 0$$

$$E(\epsilon_t \epsilon_{t-i}) = 0 \; \forall \; i \in \mathrm{N}$$

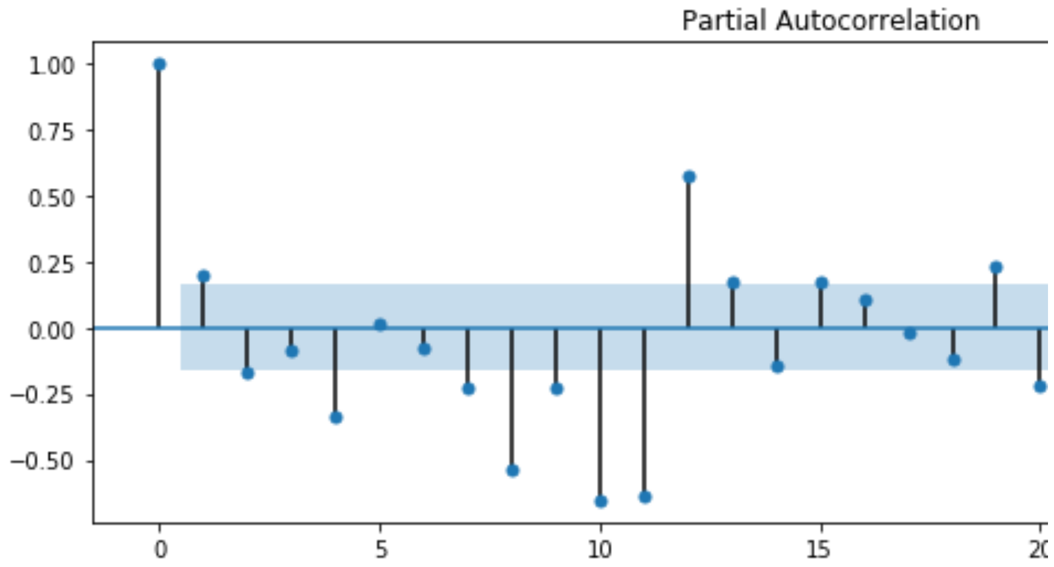
Time Series | Forecasting Models | ARIMA | Finding p

- The value of p represents the number of lagged instances of the time series that affect the present value

- A value of p = 2 means that the value of time series today is influenced not only by the value of the time series yesterday, but also by the value two days back.

- We can estimate the value of p from plotting the Partial Autocorrelation Function

- The value of p corresponds to that lag in the PACF plot when it goes to zero



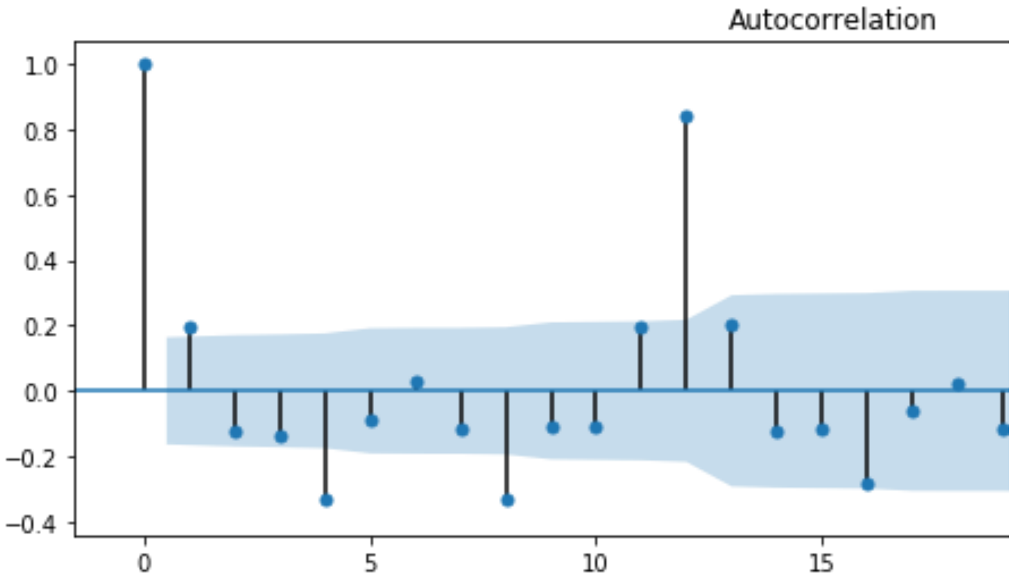
Time Series | Forecasting Models | ARIMA | Finding q

- The value of q represents the number of lagged instances of errors that affect the present value

- A value of q = 2 means that the value of time series today is influenced not only by the error value yesterday, but also by the error value two days back.

- We can estimate the value of q from plotting the Autocorrelation Function

- The value of q corresponds to that lag in the ACF plot when it goes to zero



- The value of d represents the number of differences you need to de-trend the time-series

- A value of d = 1 means that you differenced the time series once, and are now modelling the rate of change of the target variable instead of the target variable itself

- Very rarely does the value of d exceed 2, it is usually either 1 or 0

- Physical processes are typically governed by differential equations no higher than second order, hence a process with d>2 is very rare

- We can use regression models in forecasting, if we preserve time-dependence structure of the variables

- An AR(1) process

$$y_t = \alpha_1 y_{t-1} + \epsilon_t$$

can be thought of as a regression problem as follows

  - The parameter to be estimated is α_1
  - The feature matrix X is the series of lagged values
  - The number of lagged values to take depends on the autoregressive order of the process, determined from the PACF plot

t	y
1	7
2	13
3	14
4	17
5	23
6	34
7	37
8	42

Time Series Problem

t	y	X
1	7	NaN
2	13	7
3	14	13
4	17	14
5	23	17
6	34	23
7	37	34
8	42	37

Regression Formulation

Suppose the time series is such that only the pervious term affects the next term

So, we shift the y-variable by 1 row, and call it X

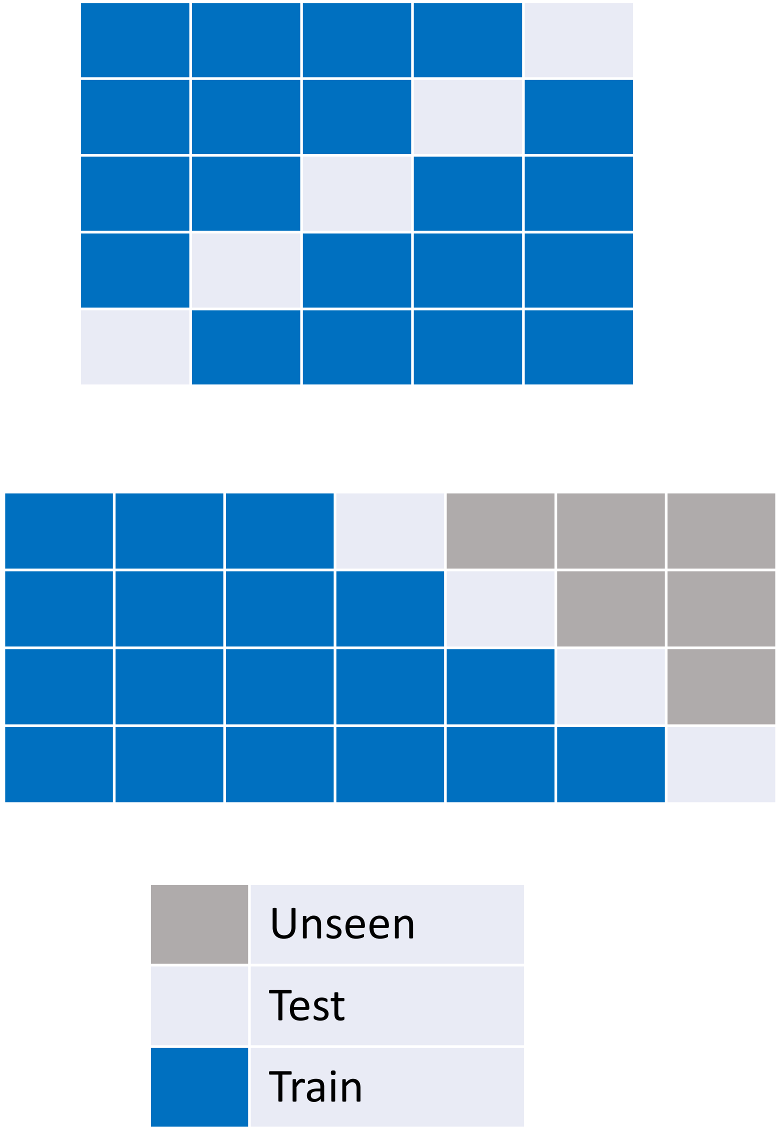
$y = mX$ + c

The regression formulation boils down to finding optimal values of m and c that minimize the MSE

Forecast Evaluation

Time Series | Forecast Evaluation | Cross Validation

- KFold cross-validation used in regression does not work for time series (Why?)
- We need time-aware cross validation strategies
- One such strategy is the Walk-Forward Cross Validation
- It preserves time awareness and eliminates the phenomenon of target leakage



Predicted	12	24	31	19
Actuals	10	25	26	18

- Mean Absolute Error

$$MAE = \frac{1}{4}\{|12-10| + |24-25| + |31-26| + |19-18|\}$$

- Mean Absolute Percentage Error

$$MAPE = \frac{1}{4}\left\{\frac{|12-10|}{10} + \frac{|24-25|}{25} + \frac{|31-26|}{26} + \frac{|19-18|}{18}\right\}\times 100$$

- Symmetric Mean Absolute Percentage Error

$$MAPE = 2\times\frac{1}{4}\left\{\frac{|12-10|}{(10+12)} + \frac{|24-25|}{(25+24)} + \frac{|31-26|}{(26+31)} + \frac{|19-18|}{(18+19)}\right\}\times 100$$

- Mean Absolute Error
  - Simple to calculate and easy to explain
  - Depends on scale of the time series

- Mean Absolute Percentage Error
  - Simple to calculate and easy to explain
  - Asymmetric metric, penalizes errors more when actual values are less for the same absolute deviation
  - Unbounded above, can $\rightarrow \infty$ as denominator $\rightarrow 0$
  - Scale invariant, does not depend on the scale of the time series

- Symmetric Mean Absolute Percentage Error
  - Relatively simple to calculate, tricky to explain
  - Symmetric metric, penalizes errors fairly in both directions for same absolute deviation
  - Bounded at both ends, can vary from $[0\,,200]$
  - Scale invariant since it's a percentage scale

Q & A