시계열 분석

Time series Analysis

시계열 자료 분석 목적

- 주어진 시계열 자료들의 생성구조를 이해하고 기술(description)
- 현재까지 관측된 값으로부터 미래의 값을 예측(forecasting)
- 생성된 시스템을 제어(control)

-> 시계열자료를 적합할 수 있는 이론적인 수학적 모형을 선택

-> 모형의 모수를 추정한 후 모형의 적합성을 검토해 선택된 모형을 시계열의 생성체계를 이해하는데 사용

시계열분석 vs. 회귀 분석

- 회귀분석은 시점을 고려하지 않지만, 시계열 분석은 시간을 고려함.
- 데이터 획득 시간을 알고 있다면 하나의 변수만 있어도 분석 가능.
    - ex) 매일 카카오 주가를 기록한다면 결과적으로 시간/주가의 두개의 데이터를 가지는 것.

1. 평균기법과 평활모형

평균기법

- 단순평균, 단순이동평균, 이중이동평균, 가중이동평균으로 시계열을 평활
    - 단순평균 : 관측값 전체에 동일한 가중치 부여. 추세가 없는 시계열을 묘사
    - 이동평균 : 시간이 경과함에 따라 수준이 변하면 일부 관측값에만 동일한 가중치를 부여

- 많은 시계열 자료들은 계절성분을 포함하거나 불규칙 변동을 포함해 추세를 정확히 파악하기 어려움
    - 이동평균법은 예측의 목적보다는 시계열 자료를 평활하여 계절변동 및 불규칙 변동을 제거해 전반적 추세 파악

단순이동평균(Simple Moving Average, SMA)

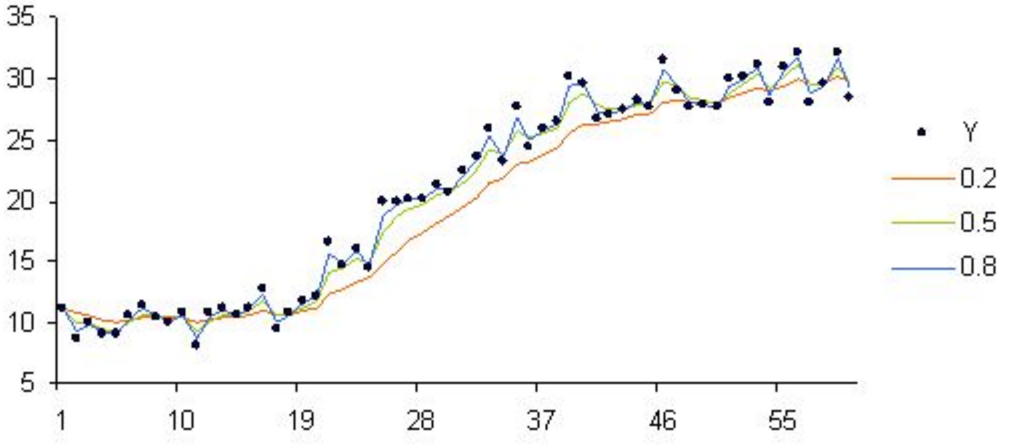
- 단순이동평균의 효과는 시계열의 불규칙 변동을 평활시킴
- 평활의 정도는 m에 의존. 아래 예시에서 m을 50, 200에 따라 평활의 정도가 달라짐
- m이 크면 불규칙 변동이 더 많이 평활되어 예측선은 고르지만, Y가 실제 변화에 더디게 반응
- 예측의 안정성과 변화에 대한 반응도의 상충관계를 고려해 m을 선택

$$\hat{T}_t = \frac{1}{m} \sum_{j=-k}^{k} y_{t+j},$$

http://kko.to/jlWBILpfo

지수평활법(Exponential Smoothing)

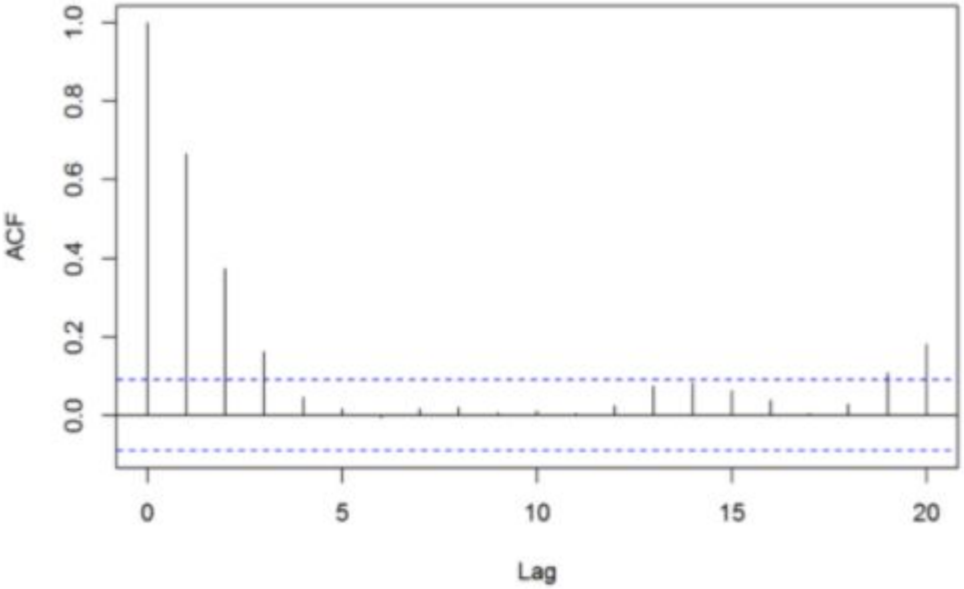
- 시계열의 수준이 t에 의존하여 느리게 변동하는 경우 미래값을 예측하기 위해 가장 최근의 관측값에 가장 큰 가중치를 부여하고 과거로 감에 따라 작은 가중치를 부여하는 평활함수를 정의해 미래 예측값을 최신화
- 단순 지수평활법은 다음 예측치 (St)를 현재 값 (yt-1)과 이전 예측치(St-1)의 합산으로 계산. 알파 (α)는 0보다 크고 1보다 작은 스무딩 매개변수 : St = α yt-1 + (1-α) St-1
- alpha가 크면 시계열 변화에 더 반응
- 작으면 평활의 효과가 커짐



2. 자기상관함수 / 부분자기상관함수

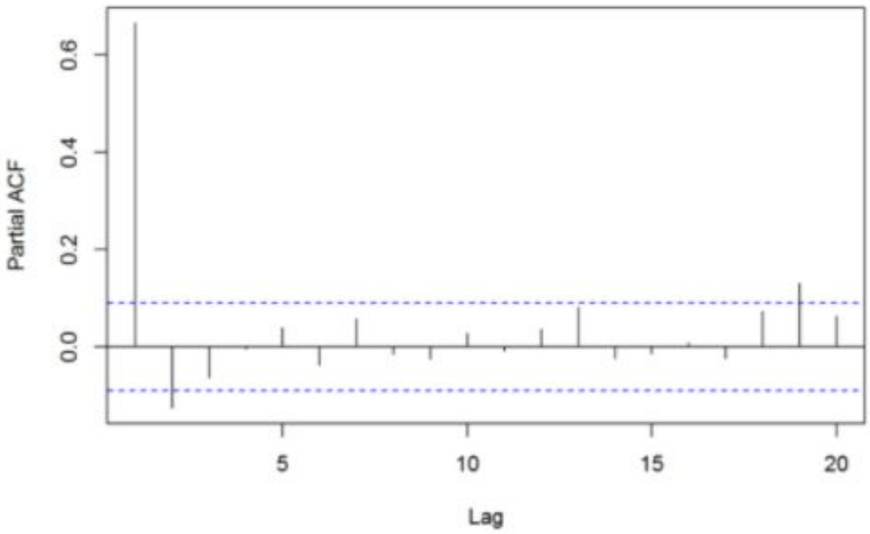
자기상관함수(ACF, Auto-Correlation Function)

- k시간 단위로 구분된 시계열 관측치 간 상관 관계 함수
- k가 1,2,3,... 일 때, k단계 떨어진 데이터 점 쌍 간들간의 상관관계
- 점선으로 유의미한 상관과 유의미하지 않은 상관을 확인 가능. 선 위쪽이 유의미한 값



부분자기상관함수(PACF, Partial Auto-Correlation Function)

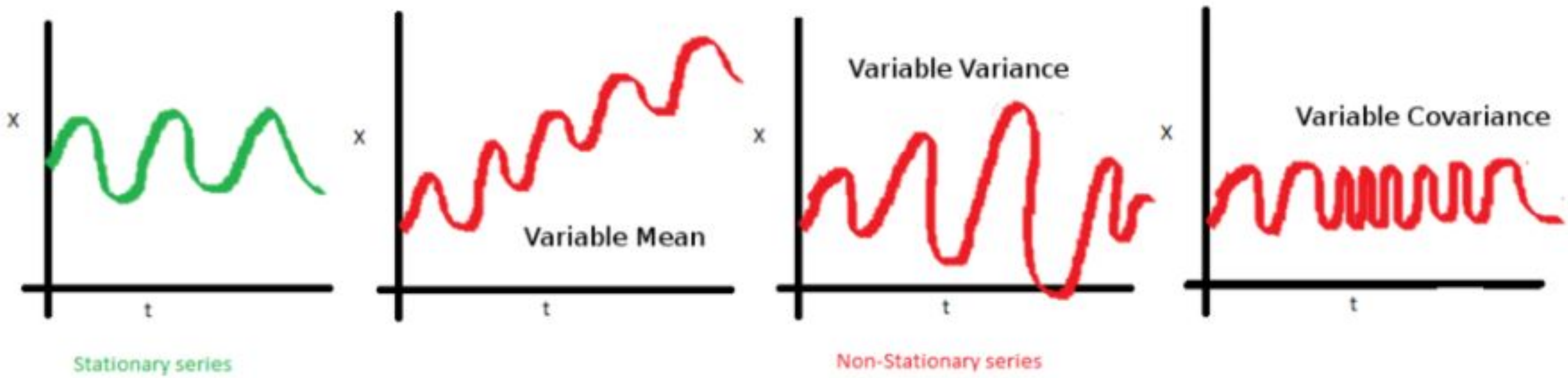
- 부분 상관이란 두 확률변수 X와 Y에 의해 다른 모든 변수들에 나타난 상관 관계를 설명하고 난 이후에도 여전히 남아있는 상관관계
- PACF는 ACF와 마찬가지로 시계열관측치 간 상관관계 함수이고, 시차 k에서의 k단계만큼 떨어져 있는 모든 데이터 점들간의 상관관계
- PACF에서도 선 위쪽이 유의미한 값



3. ARIMA 모델

정상 시계열 모형

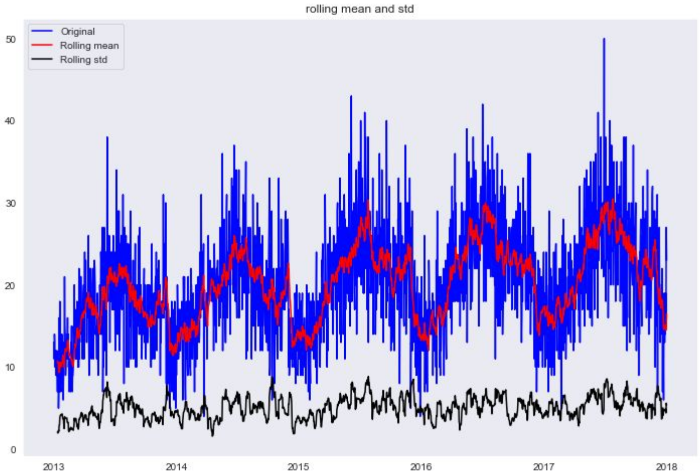
- 정상성(Stationarity)
    - 통계적 특징이 시간에 따라 변하지 않음
    - 주기적인 변동이 없다는 것으로 미래는 확률적으로 과거와 동일
- 시계열분석에서는 주로 통계 검정 및 모델이 정상성을 가정함. 정상성을 만족하지 못하다면 데이터를 정상화 시킨 뒤 분석을 해야함.



정상성을 확보하는 방법

- 1) 시각화 하여 확인
- 2) Dickey-Fuller Test : 정상성을 확인해주는 검정법(단위근 검정)
- 차분(Differencing) 및 변환(Transformation)
  - 시계열 데이터의 평균을 안정화 : 차분
  - 시계열 데이터의 분산 안정화 : 변환

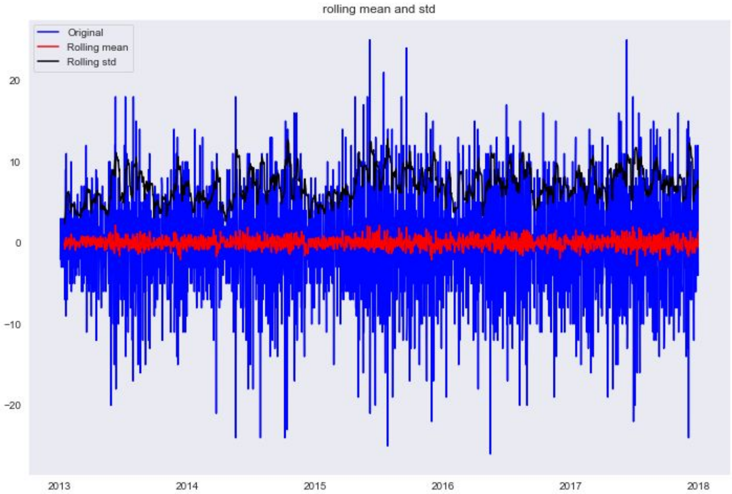
정상성을 확보하는 방법



Results of Dickey-Fuller Test
p-value = 0.0361. The series is likely non-stationary.
Test Statistic -2.987278
p-value 0.036100
#Lags Used 20.000000
Number of Observations Used 1805.000000
Critival Value(1%) -3.433978
Critival Value(5%) -2.863143
Critival Value(10%) -2.567623

->

1차 차분



Results of Dickey-Fuller Test
p-value = 0.0000. The series is likely stationary.
Test Statistic -1.520810e+01
p-value 5.705031e-28
#Lags Used 2.000000e+01
Number of Observations Used 1.804000e+03
Critival Value(1%) -3.433980e+00
Critival Value(5%) -2.863143e+00
Critival Value(10%) -2.567624e+00
dtype: float64

추세, 계절성, 주기성

- 추세(trend)
  - 데이터가 장기적으로 증가하거나 감소.
  - 선형적일 필요는 없음
- 계절성(seasonality)
  - 해마다 특정한 때, 1주일마다 특정 요일에 같은 패턴이 발생하는 경우
  - ex) 여름마다 에어컨 소비 증가에 따른 전력 수요량의 증가
- 주기성(cycle)
  - 고정된 빈도가 아닌 형태로 증가하거나 감소하는 모습

AR(AutoRegression) 모형

- 자기 회귀 모델로 자기 자신의 과거를 사용. 이전의 자신의 관측값이 이후의 자신의 관측값에 영향을 준다는 아이디어의 모형.
- $X(t) = (X_{t-1} * w) + b + (e_t * u)$
- 이전의 자기 상태에 w를 곱하고 b를 더한 것에 (e(t)*u)라는 특수한 값을 더함
- e(t)는 white-noise(백색 잡음)으로 N(0,1)을 따르는 random noise
- AR 모형에서는 불확실성을 포함하기 위해, 불규칙한 데이터를 잡아주기 위해 노이즈를 사용

AR(AutoRegression) 모형

t를 현재 시점, p를 과거 시점이라고 할 때,

Z = 시계열 자료, Φ = 모수, α = 오차항

$$Z_t = \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \cdots + \Phi_p Z_{t-p} + \alpha_t$$

시계열 자료 현재 시점

과거가 현재에 미치는 영향을 나타내는 모수 × 시계열 자료 과거 시점

오차항 (백색 잡음 과정)

- AR(1) : AR모형의 가장 간단한 형태. 바로 직전의 데이터가 다음 데이터에 영향을 준다고 가정.
- AR(p) : p이전의 시점부터 자기회귀
  - AR(1)은 1시점 전에 의해 현재 시점이 영향을 받음.
  - AR(3)는 3시점 전까지에 의해 영향을 주는 모형.

AR(AutoRegression) 모형

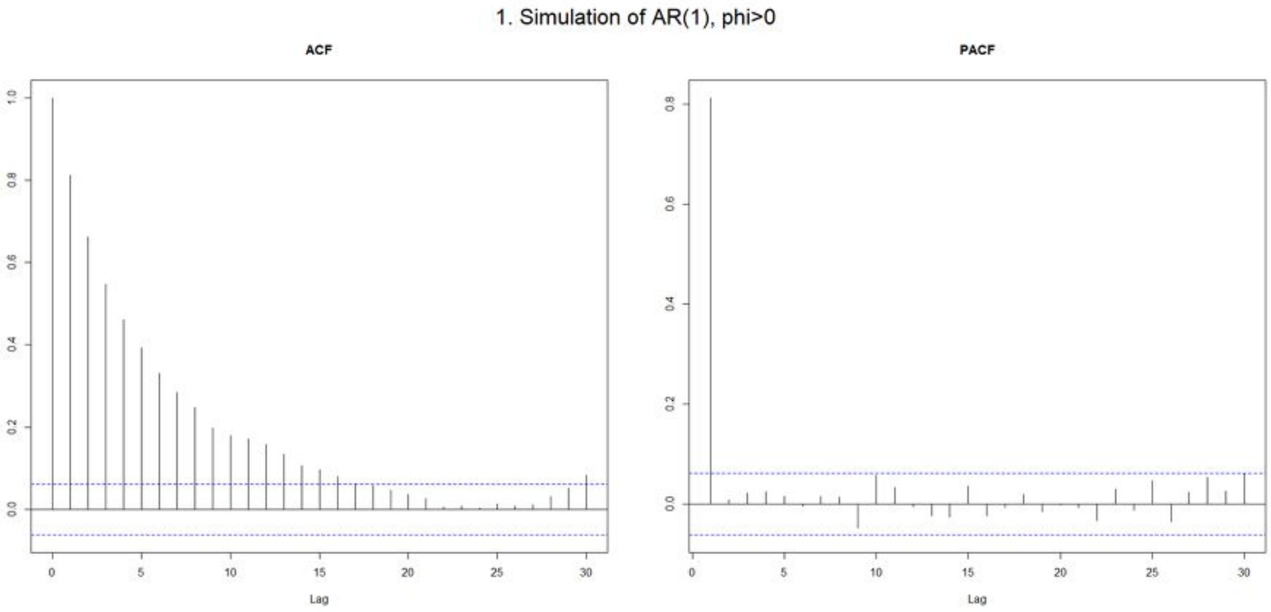
- 자기회귀모형인지 판단하기 위해서는 자기상관함수(ACF, Auto-Correlation Function)와 부분자기상관함수(PACF, Partial Auto-Correlation Function)을 이용해 식별
- 일반적으로 ACF는 시차가 증가함에 따라 점차적으로 감소하고, 부분자기상관함수는 p+1 시차 이후 급격히 감소하여 절단된 형태이며 이를 AR(p) 모형이라고 판별

AR(AutoRegression) 모형

1. AR(1)

1) 자기회귀계수가 양수(0.8)인 경우

- ACF : 지수함수를 그리며, 서서히 '0'으로 감소하는 형태
- PACF : 1차에 두드러지는 스파이크가 나타나고, 이후 모두 '0'으로 절단

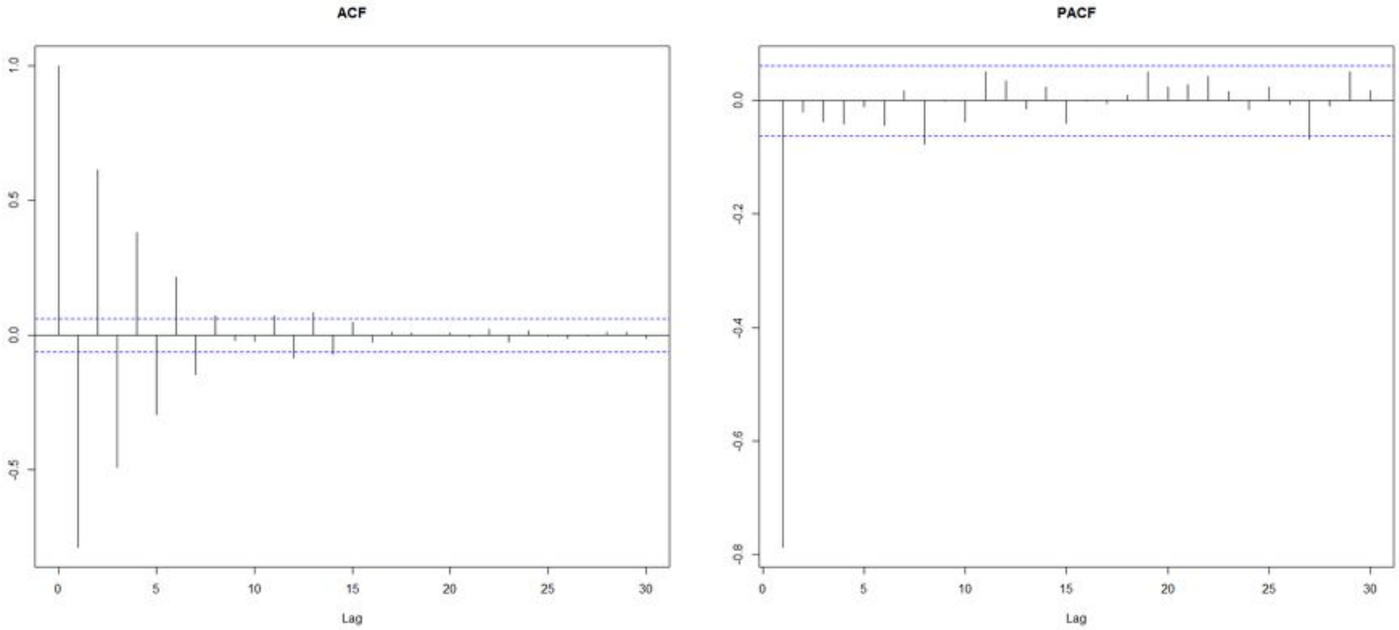


AR(AutoRegression) 모형

2) 자기회귀계수가 음수(-0.8)인 경우

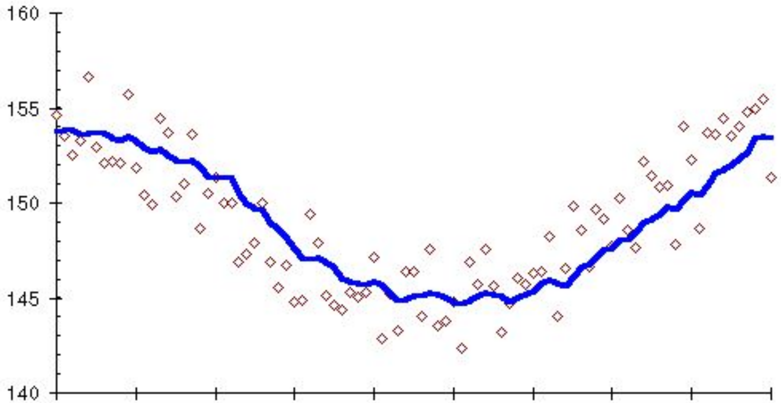
- ACF : 사인함수를 그리며, 서서히 '0'으로 감소하는 형태
- PACF : 1차에 두드러지는 스파이크가 나타나고, 이후 모두 '0'으로 절단

2. Simulation of AR(1), phi<0



MA(Moving Average) 모형

- 시간이 지날수록 어떠한 Random Variable의 평균값이 지속적으로 증가하거나 감소하는 경향.
- 예를들어 봄에서 여름이 되면 전기 수요량이 증가하고, 여름에서 겨울로 가면 감소하는 경향이 있음.
    - 이 경우 전월의 전기 사용량이 다음월 전기 사용량에 상관을 주지 않는다고 가정할 수 있음
- 관측값이 이전의 연속적인 오차항의 영향을 받는다는 모형으로 데이터의 평균값 자체가 시간에 따라 변화하는 경향을 시계열 모형으로 구성 $X(t) = (e_{t-1} * w) + b + (e_t * u)$



데이터의 평균값 자체가 시간에 따라 변화하는 경향이 Moving Average이다.

MA(Moving Average) 모형

t를 현재 시점, p를 과거 시점이라고 할 때,

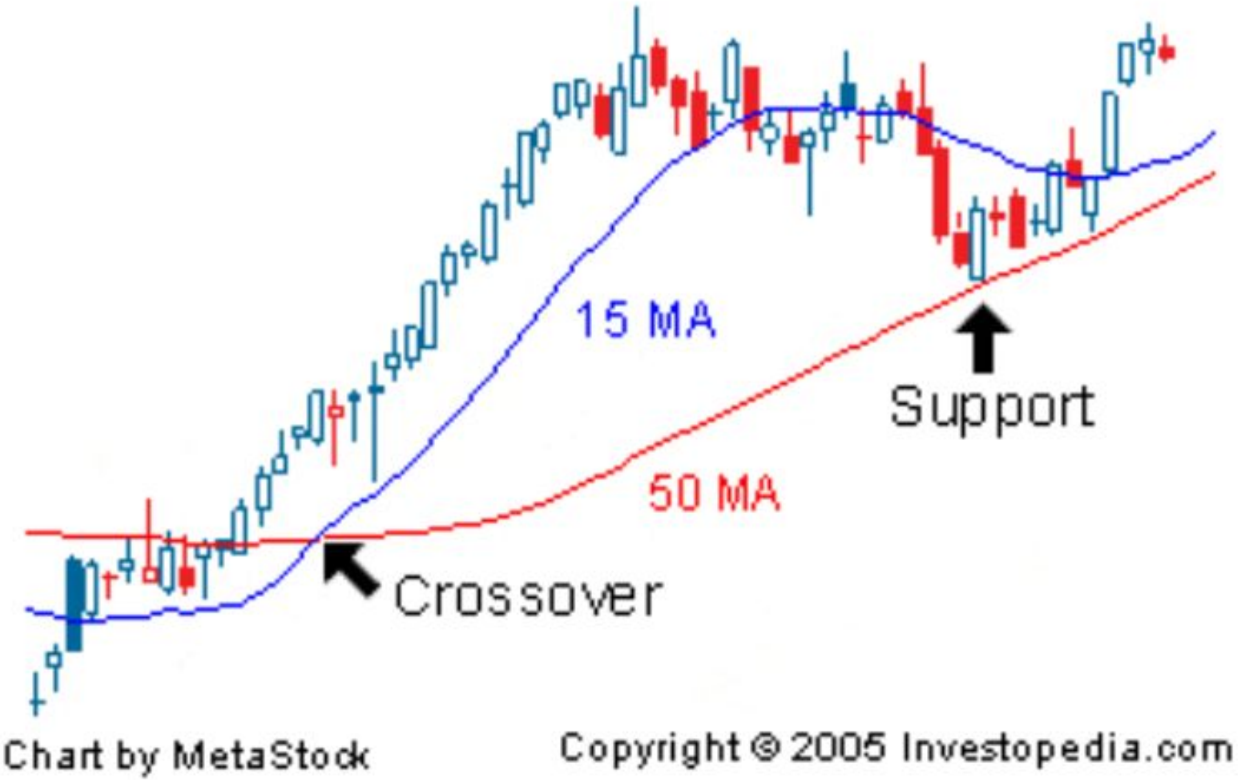
Z = 시계열 자료, θ = 매개변수, α = 오차항

$$Z_t = \theta_1\alpha_{t-1} + \theta_2\alpha_{t-2} + \cdots + \theta_p\alpha_{t-p} + \alpha_t$$

- Z_t: 시계열 자료 현재 시점
- $\theta_1\alpha_{t-1} + \theta_2\alpha_{t-2} + \cdots + \theta_p\alpha_{t-p}$: 매개변수 × 과거 시점의 오차 (백색 잡음)
- α_t: 오차항 (백색 잡음 과정)

- MA(1)
  - MA모형의 가장 간단한 형태.
  - AR(1)은 이전 시점의 관측값이 영향을 미친다고 가정한다면, MA(1)은 이전 시점의 오차(e(t-1))를 이용해 현재를 추론. -> 이전에 발생한 error가 중요하지 이전 관측값은 중요하지 않음.
  - 즉, 변화하는 트렌드를 고려하는 모형.
- MA(q)
  - 더 이전 시점을 모델에 넣고자 하는 경우

MA(Moving Average) 모형



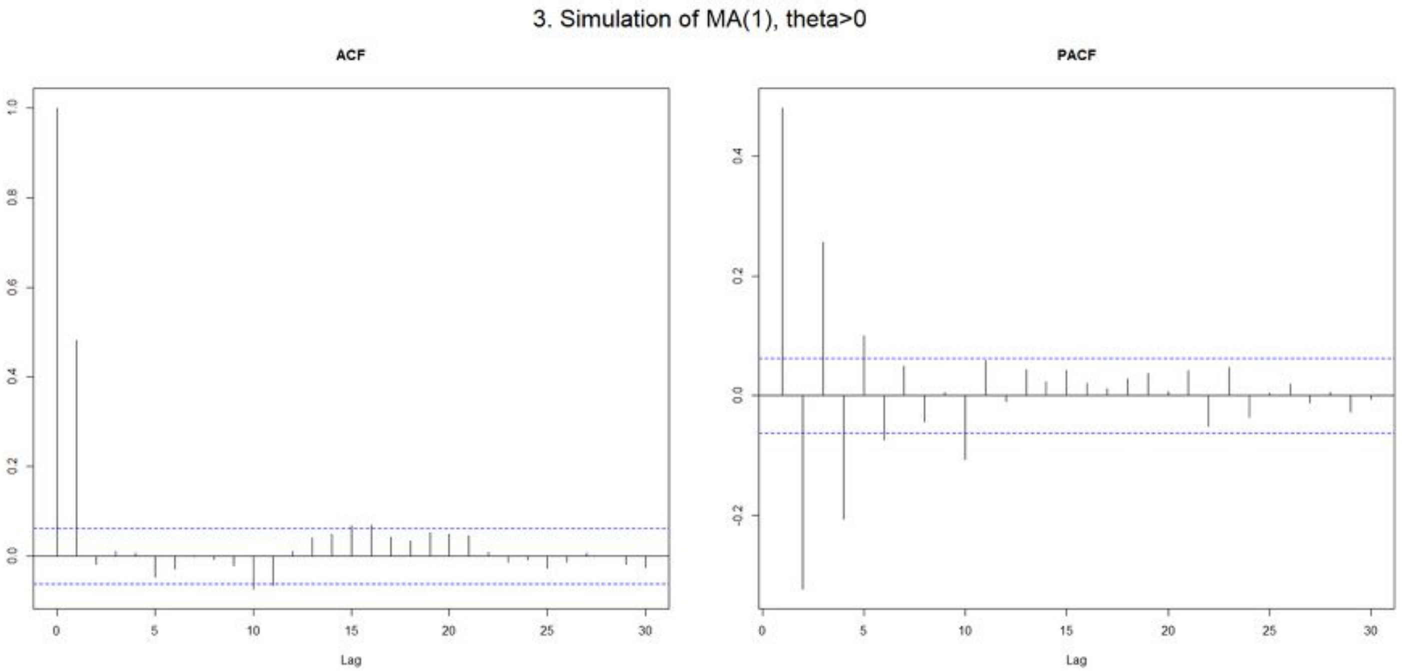
증권가에서 기술적 분석을 할 때 가장 많이 쓰는 것이 MA모형이기도 하다. 예를 들어 최근 50일 평균값보다 최근15일 이동평균값이 커지면 주가가 치솟는다, 즉 골든크로스가 발생한다 같은 접근 말이다.

MA(Moving Average) 모형

2. MA(1)

1) 이동평균계수가 양수(0.8)인 경우

- ACF : 1차에 두드러지는 스파이크가 나타나고, 이후 모두 '0'으로 절단
- PACF : 사인함수를 그리며, 서서히 '0'으로 감소하는 형태

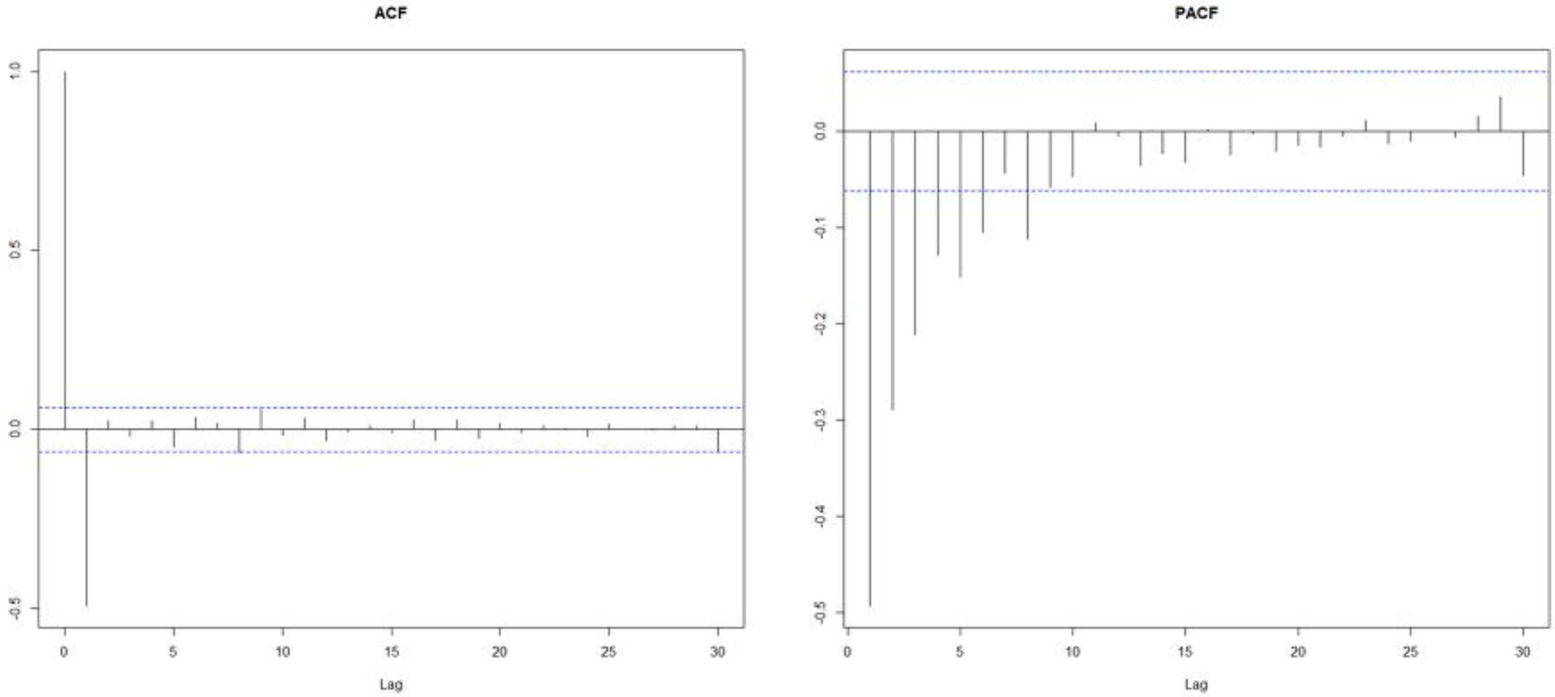


MA(Moving Average) 모형

2) 이동평균계수가 음수(-0.8)인 경우

- ACF : 1차에 두드러지는 스파이크가 나타나고, 이후 모두 '0'으로 절단
- PACF : 지수함수를 그리며, 서서히 '0'으로 감소하는 형태

4. Simulation of MA(1), theta<0



AR/MA모형과 ACF/PACF 관계

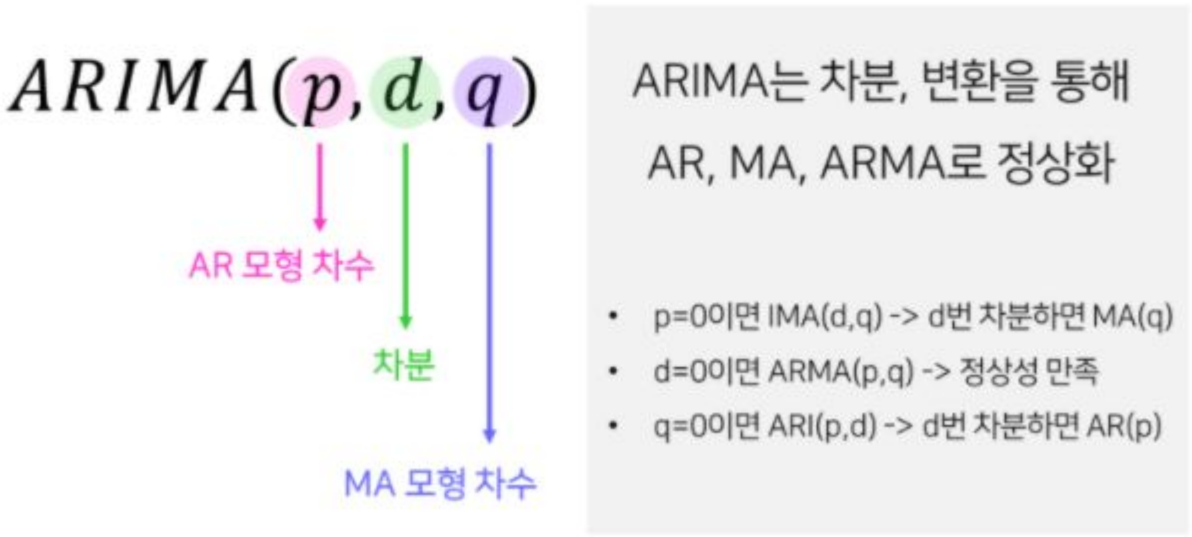
구분	AR(p)	MA(q)
ACF	점차적으로 감소	시차 q 이후에 0
PACF	시차 p 이후에 0	점차적으로 감소

ARMA(AutoRegressive Moving Average) 모형

- AR형태와 MA형태를 동시에 가지고 있는 경우
- 과거의 상태와 오차값을 사용해 현재의 상태를 예측하는 모델
- ARMA(1,1) : $X(t) = (X_{t-1} * w_{11}) + (e_{t-1} * w_{21}) + b + (e_t * u)$
- ARMA(2,2) : $X(t) = (X_{t-1} * w_{11}) + (X_{t-2} * w_{12}) + (e_{t-1} * w_{21}) + (e_{t-2} * w_{22}) + b + (e_t * u)$

ARIMA(AutoRegressive Integrated Moving Average) 모형

- ARMA에서 Integrated라는 개념을 추가한 모델
    - ARMA에서는 불규칙적 시계열 데이터를 제대로 예측하지 못한다는 한계 존재
- ARIMA모델은 관측치 사이의 차분(difference)을 사용해 불규칙적 시계열 데이터를 규칙적으로 활용
- 차분은 거친 결과 변수들이 Whitening되는 효과를 가짐



AR, MA, ARMA, ARIMA

- AR : 현재와 과거의 자신과의 관계 정의
- MA : 현재와 과거 자신의 오차와의 관계 정의
- ARMA : 현재와 과거의 자신 그리고 자신과의 오차를 동시에 고려해 정의
- ARIMA : 현재와 추세(트렌드 변화)간의 관계를 정의

ARIMA 수식

- 소문자 x는 변환된 새로운 현재 상태, 대문자 X는 변환 전 원래의 상태
- X값을 활용해 예측모델을 ARIMA(p,d,q)로 구성 가능.
- ARIMA(1,2,1)이라면 AR과 MA를 1개 만큼 과거를 window로 활용, 차분은 2만큼 활용

$$(d=0): x_t = X_t$$

$$(d=1): x_t = X_t - X_{t-1}$$

$$(d=2): x_t = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2})$$

ARIMA(p,d,q) 차수 결정

$$\hat{y} = \mu + (w_{11} * y_{t-1} + \cdots + w_{1p} * y_{t-p}) - (w_{21} * e_{t-1} + \cdots + w_{2p} * e_{t-p})$$

- 모형식별이란 ARIMA(p,d,q) 모형을 따르는 시계열 Y에 대한 차수 p,d,q를 결정하는 것
- 데이터가 비정상성을 보이면 단위근 검정 후 차분을, 분산이 일정치 않으면 분산 안정화를 위한 변수변환
- ARMA 모형 차수를 결정하는데 AIC, SBC 등이 사용됨(잔차에 근거한 모형 선택 기준통계량)

계절형 ARIMA

- 일정한 시간 간격을 두고 매년 동일한 현상이 반복되는 시계열 데이터
  - 월별 시계열의 경우 계절주기는 12, 분기별 시계열의 경우의 계절주기는 4
  - 계절시계열자료는 분산이 일정하고 추세가 없다 하더라도 정상시계열로 간주하기 어려움
- 추세는 차분하면 제거될 수 있으나, 계절성은 제거되지 않을 수 있음. 이에 계절 시계열 자료에서는 계절차분을 통해 데이터를 정상화함