데이터 분석 전처리 (판다스)

시계열 분석

시계열 데이터의 특성

시간이라는 독립변수

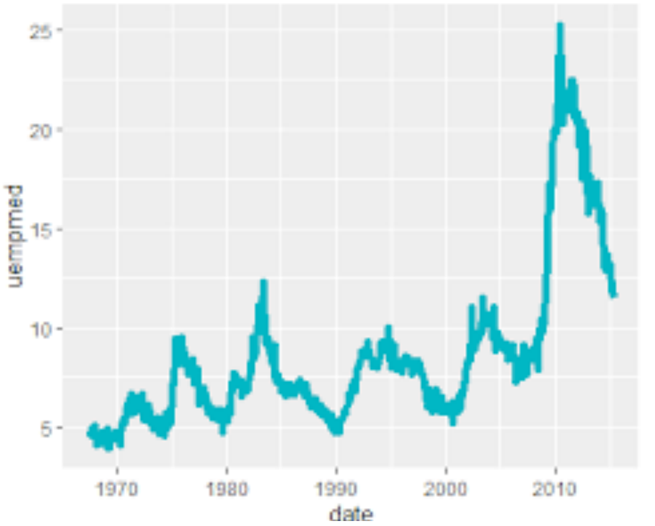
기존 데이터의 특성과는 다른

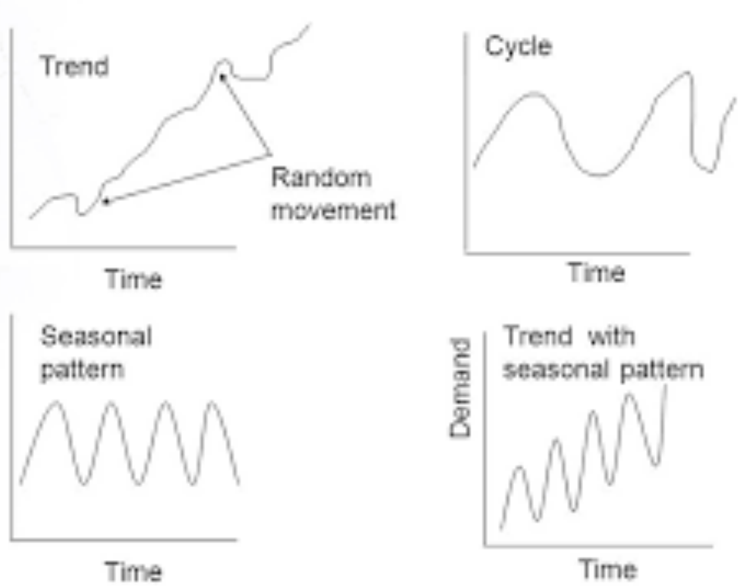
시계열만의 특성

트렌드 성분

주기 변동 성분

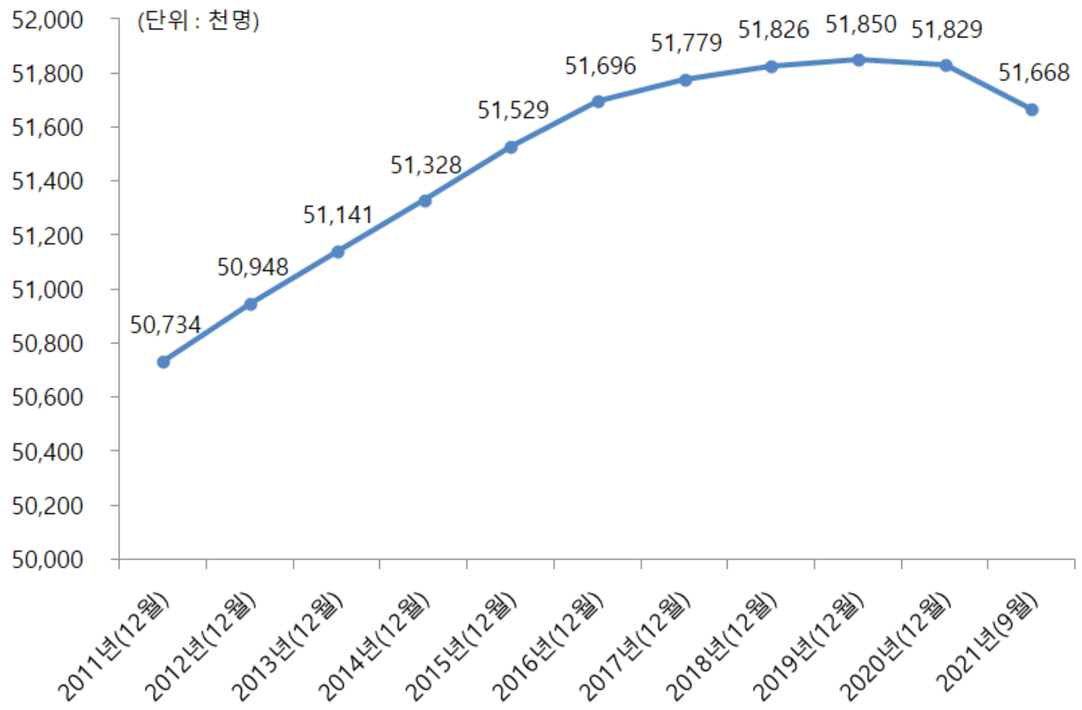
노이즈 성분





자기 상관 관계 (Auto Correlation)

과거의 데이터가 현재의 데이터에 영향을 주는 경향을 자기 상관

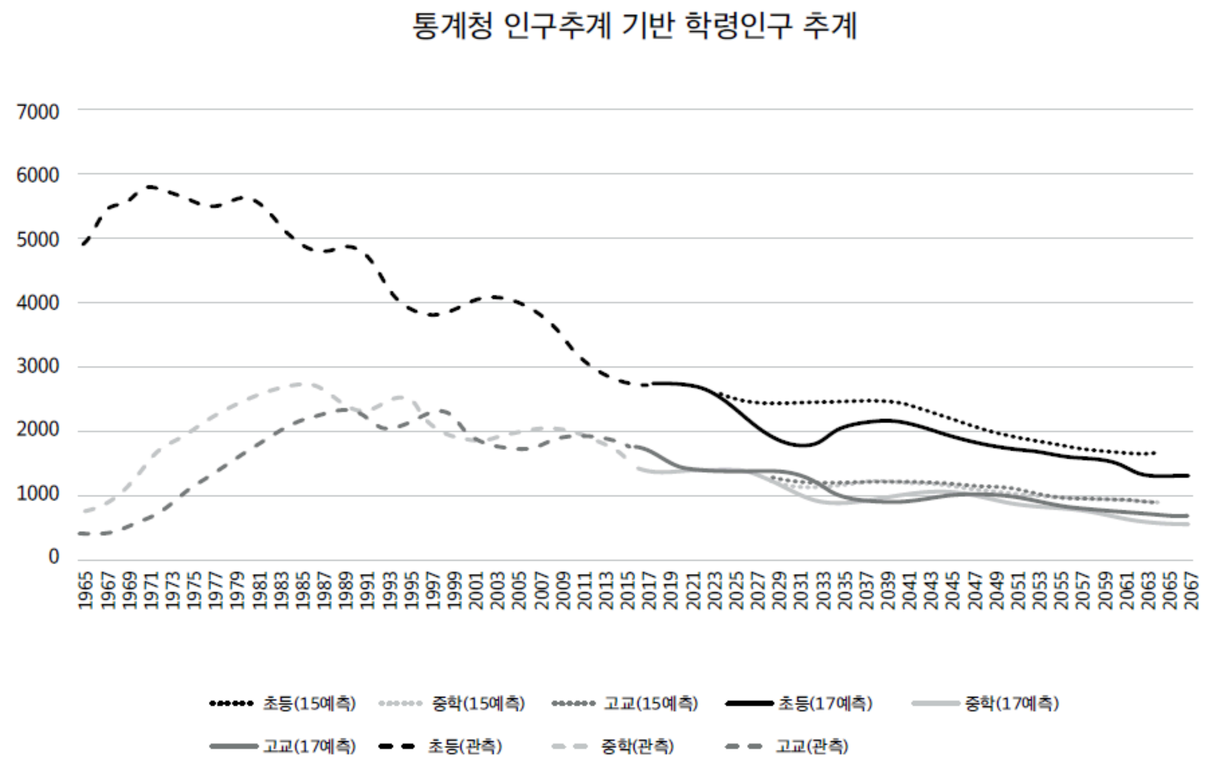


자기 자신의 n번째 과거 데이터와 현재 데이터간의 상관관계

우리나라 인구수 자기상관관계가 매우 높음

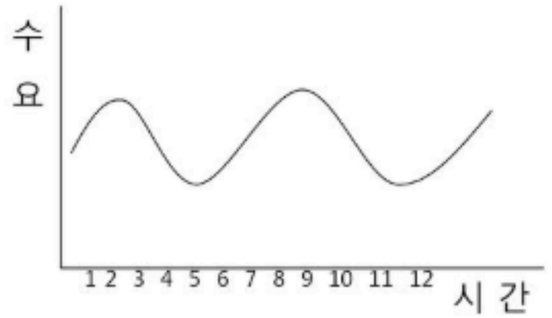
추세 경향성 (Trend)

시계열 데이터의 장기적으로 점차 증가 또는 감소 추세 경향성

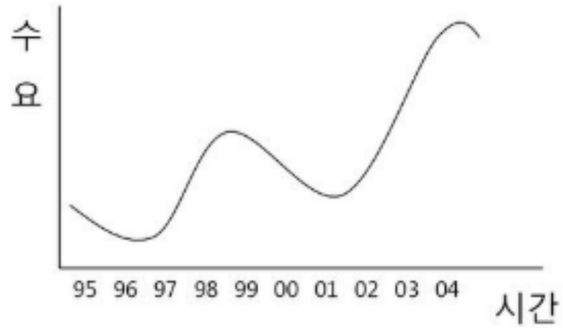


계절성(Seasonality), 순환성(Cycle)

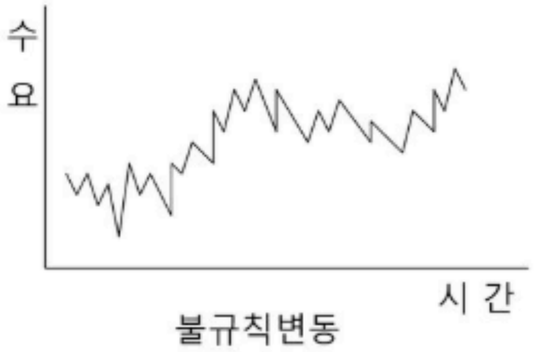
중장기적인 추세 경향성 외에 날짜나 기간에 따른 주기적으로 변화하는 것 추세가 반복되는 변동성들



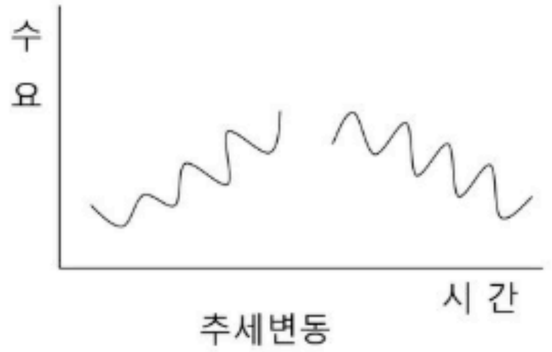
계절변동



순환변동



불규칙변동



추세변동

불확실성

날씨 예측을 생각해보면?

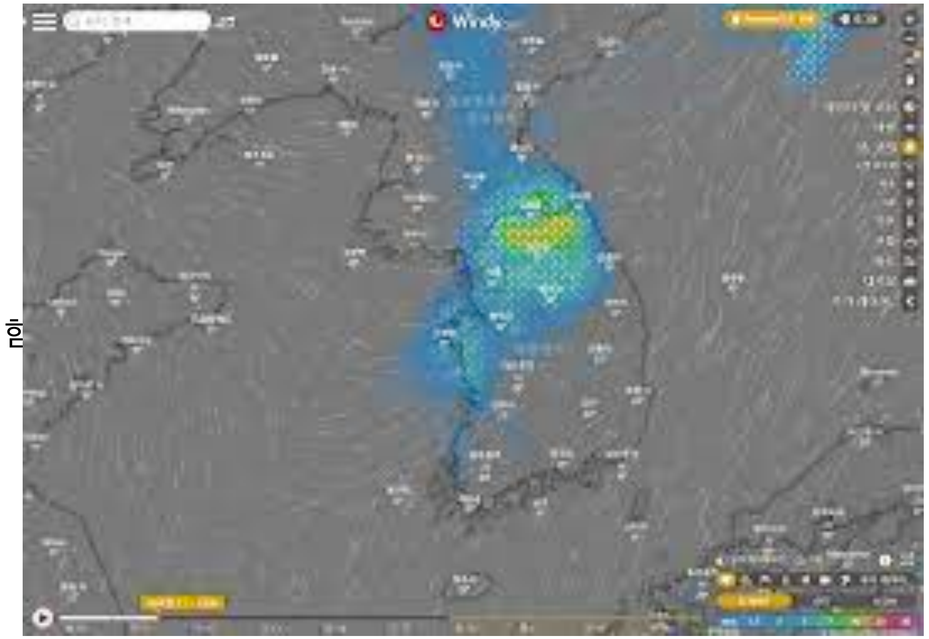
‘미래 예측값은 현재로써 정확히 모른다.

따라서 시계열 모델의 해석에는 반드시 불확실성에 대한 고려가 수반

시계열 분석을 통해 예측된 미래 데이터는 사용된 모델에 적합한 확률 분포에 따른 신뢰 구간이 같이 제공되어 평가 되어야 함

일반적으로 미래 데이터를 말할 때는 앞서 말한 확률분포의 평균을 말하는 경우가 많음

대부분의 시계열 모델은 불확실성을 표현하기 위해 90% , 95% 와 같이 확률 분포에 따른 신뢰 구간을 제공



시계열 예측의 신뢰 구간에서 가장 먼저 이해할 것은 시계열 예측의 예측 기간이 길어질수록 예측 신뢰구간 데이터 분포가 점점 넓어진다는 것

결국 먼 미래 예측은 더 힘들고 어렵다

시계열 분석

탐색 목적 : 외부 인자와 관련된 계절적 패턴, 추세 설명 인과관계 규명

예측 목적 : 과거 데이터 패턴을 통해 미래 값 예측

시계열 데이터는

Xt = St(신호) + At(잡음)

잡음을 전처리 하여 신호만 잘 학습시킬 수 있도록

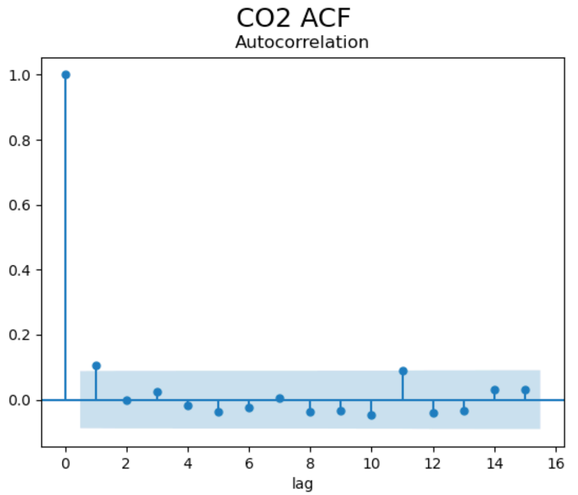
선형 회귀 기반 모델

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t.$$

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t,$$

ACF, PACF

자기상관함수 (ACF) 시계열 데이터 주기성 수치적 확인

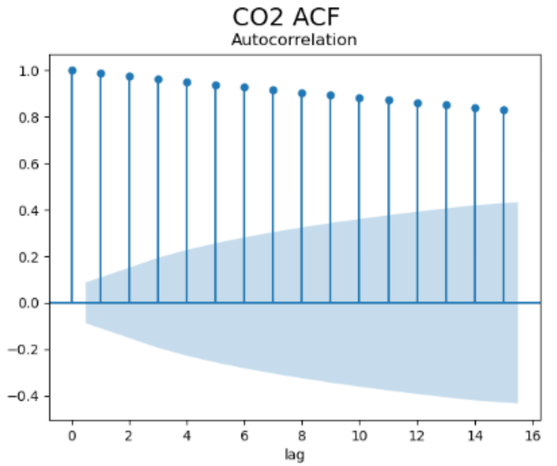


ACF(Autocorrelation function)

ACF는 k시간 단위로 구분된 시계열 관측치 간의 y_t와 y_{t+k}간 상관 관계를 측정하는 것이다. ACF의 반환값의 절대값이 커질수록 시차 시계열 데이터의 상관성이 크다고 할 수 있다. 이 크다는 것의 기준은 p-value와 같이 95% 근사 신뢰구간으로 정할 수 있다. 또는 $\pm 1.96 * \frac{SE_k}{\sqrt{n}}$를 신뢰구간으로 적용할 수 있는데 이때,

$SE_k = \sqrt{1 + 2\Sigma_{j=1}^{k} |\hat{\gamma_j^2}|}$이다.

ACF는 정상성 데이터에 대해서는 0으로 빠르게 떨어지고 비정상성 데이터에는 천천히 수치가 떨어진다.



시계열 데이터의 주기성을 수치적으로 확인하고 어떠한 특정 시차가 어떠한 영향을 주는지 알 수 있음

상관관계를 0.05 유의수준 안에서 나타내는 것

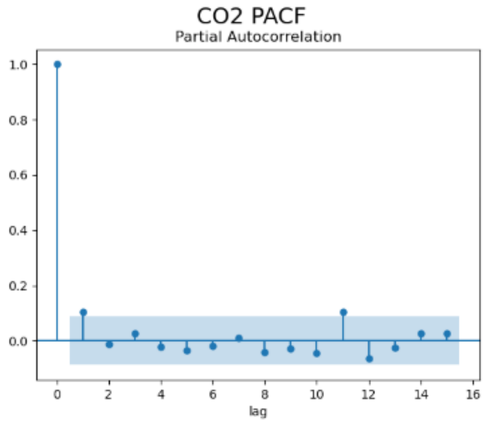
PACF

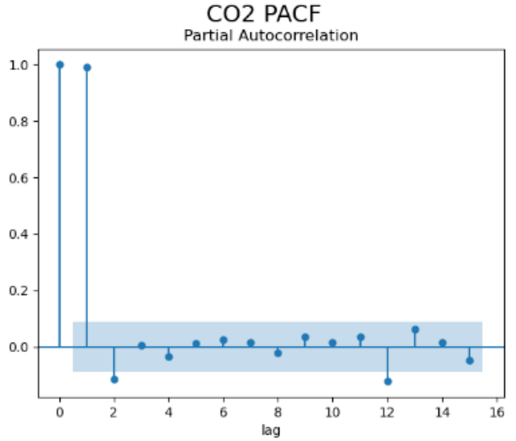
편자기상관함수(Partial autocorrelation function)

y의 시점과 특정 t-k의 시점 이외의 모든 시점과의 영향력을 배제한 순수한 영향력 나타내는 척도

PACF(Partial Autocorrelation function)

PACF는 ACF가 모든 시계열 데이터의 특성을 분석하는 것에 한계가 있기에 추가적인 분석의 필요에 따라 사용된다. 예를 들어 Y_t와 Y_{t-1}이 관계가 있다면 Y_{t-1}와 Y_{t-2}도 상관이 있고 Y_t와 Y_{t-2}는 Y_t와 Y_{t-1}간의 유의미한 관계가 있음으로 인해 상관있다는 결과가 도출될 수 있다는 점이다. 이 추가적인 분석에선 부분 상관이란 확률 변수인 X와 Y에 의해 모든 변수들에 대한 상관 관계를 분석한 후에도 아직 남아있는 상관관계를 해석한다. PACF는 y_t와 y_{t+k}간 상관 관계를 도출하는 것은 같지만 t와 $t+k$ 간 다른 y값의 영향력은 배제하고 측정하는 방식이다. 즉 시차 k에서의 k단계만큼 떨어져 있는 모든 데이터들 간의 상관 관계를 말한다.
PACF가 시차 n에서 떨어지는 경우 AR(n)을 사용하고 하락이 점진적이라면 MA를 사용한다.





다른 시점들과의 다중공선성을 제거한 단 두 시점과의 관계 수치화

시계열 관측치 간 상관 관계 함수이며, 시차 K에서 K만큼 떨어져 있는 모든 데이터 점들간의 상관 관계

ARIMA

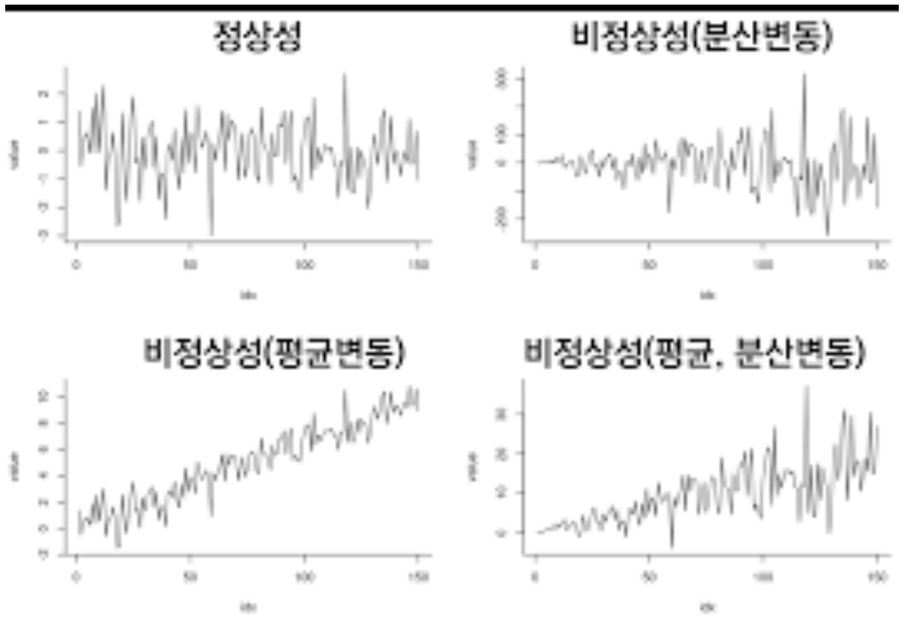
Auto regressive integrated moving average

정상성(Stationarity)?

AR, MA 모형 분석 시작하기 전 정상성 만듦

정상성이란 모든 시점에 대해서 일정한 평균을 갖도록 하는 것, 즉 추세나 계절성이 없는 시계열 데이터로 만들어 주는 것

즉 정상성을 나타내는 시계열은 평균과 분산이 안정되어 있는 상태, **평균이 일정하지 않으면 차분, 분산이 일정하지 않으면 변환**



지연과 차분(Difference)

시계열 데이터에 추세가 있어서 평균이 일정치 않으면 차분

현재 상태의 값에서 바로 전 상태 값을 빼주는 것 -> 모든 기간의 평균을 일정하게 만든다.

추세만 차분하는 것은 1차 차분, 시계열에 계절성도 존재하는 경우에는 계절성의 시차인 n 시점

전 값을 빼주는 2차 차분, 계절 차분 가능

지연은 정해진 시간만큼 앞으로 당기거나 혹은 뒤로 밀어낸 데이터를 말함

변환(Transformation)

시계열 데이터의 분산이 일정치 않을 때는 변환을 해줘야 한다.

분산이 커지는 경우는 각 시점의 값에 로그나 루트 씌워 분산 크기 완화 (감소)

AR MA

AR 자기회귀모델

회귀 기반의 시계열 분석 시차변수만 사용 개념

이전 관측 값이 이후 관측 값에 영향을 준다는 아이디어

$$AR(1) \ : \ X_t = a X_{t-1} + e_t \qquad (-1 < a < 1)$$

- X_t : 현재 시점의 정상 시계열
- a : 상수
- X_{t-1} : 이전 시점의 정상 시계열
- e_t : 현재 시점의 잔차

현재 시점의 정상 시계열
= a * 이전 시점의 정상 시계열 + 현재의 잔차

t를 현재 시점, p를 과거 시점이라고 할 때,

Z = 시계열 자료, Φ = 모수, α = 오차항

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

- Z_t : 시계열 자료 현재 시점
- $\Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \cdots + \Phi_p Z_{t-p}$: 과거가 현재에 미치는 영향을 나타내는 모수 × 시계열 자료 과거 시점
- α_t : 오차항 (백색 잡음 과정)

AR MA

MA모형은? 관측값 이전 시점의 연속적인 예측 오차의 영향을 이용하는 방법

$$y_t = \theta_0 + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

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)

트렌드가(평균 혹은 시계열 그래프에서 y값) 변화하는 상황에 적합한 회귀 모델.

시계열을 따라 윈도우 크기 만큼 슬라이딩하여 이동 평균 모델이라 함

MA모형은 관측값의 이전 시점의 연속적인 예측 오차의 영향을 이용하는 방법이다.

y값은 해당 시점 오차항에 n 시점 이전의 오차항에 이동 평균 계수를 곱한 값들을 더해준 것

이전 시점의 상태를 이용하여 현재를 예측하는 방식이 아니다. **이전 시점의 변동값과 오차항을 이용하여 현재 상태를 추론**

ARMA

- **AR 모형**
  - AR 표현방식이며 유한 시차로 구성
  - AR(1): 시차 1 변수 포함

$$Z_t = \phi_1 Z_{t-1} + a_t$$

- **MA 모형**
  - MA표현방식이며 유한 시차로 구성
  - MA(1): 시차 1 백색잡음 포함

$$Z_t = a_t - \theta a_{t-1}$$

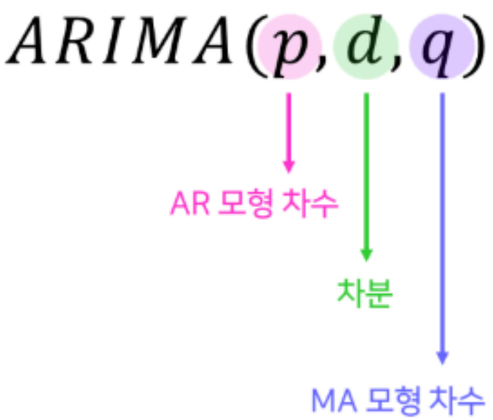
- **ARMA 모형**
  - AR방식과 MA방식이 결합된 형태
  - ARMA(1,1): 시차1의 변수와 시차1의 백색잡음 포함

$$Z_t = \phi_1 Z_{t-1} + a_t - \theta a_{t-1}$$

자동 회귀 이동 평균

AR과 MA를 섞은 모델로 두 가지 관점에서 과거 데이터 사용하는 것

ARIMA



ARIMA는 차분, 변환을 통해
AR, MA, ARMA로 정상화

- p=0이면 IMA(d,q) -> d번 차분하면 MA(q)
- d=0이면 ARMA(p,q) -> 정상성 만족
- q=0이면 ARI(p,d) -> d번 차분하면 AR(p)

ARIMA 모델은 **ARIMA(p, d, q)**으로 표현하며, 순서대로 **AR 모형 차수**, **차분**, **MA 모형 차수**를 의미한다.
이 차수의 최적 차수는 **자기 상관 함수 (ACF)**, **편 자기 상관 함수 (PACF)**를 사용하여 찾아야 한다.

하지만 보통 시계열 데이터는 추세를 가지고 있으며, 일정한 패턴이 있음 (대부분)

따라서 대부분 불안정 (Non - stationary) 한 패턴

그래서 모델 자체의 불안정성을 제거하는 기법인 ARIMA

ARIMA는 모델의 과거 데이터가 가지고 있던 추세까지 반영

시계열의 비정상성을 설명하기 위해 시점간의 차분을 사용하는 것

AR 모델의 자기 회귀 부분의 차수, MA 모델의 이동 평균 부분의 차수

그리고 1차 차분이 포함된 정도를 포함하여 ARIMA(p,d,q) 표현

요약하면?