Time Series Analysis

시계열 자료분석

류영표
ryp1662@gmail.com

류 영 표

AI & Applied Mathematics Researcher · Lecturer

ryp1662@gmail.com
github/youngpyoryu

학력

- 동국대학교 일반대학원 응용수학 석사 수료 (2016-2019)
- 동국대학교 수학과 졸업 (2010-2016)

심사 및 전문가 자문 활동

- 인공지능 학습용 데이터 구축 사업 품질검증 전문가 자문위원 (2023)
- 2023 신뢰할 수 있는 인공지능 개발 안내서(안)_고도화 전문위원
- 2023 자율주행/의료/공공사회 분야 신뢰할 수 있는 인공지능 개발 안내서(안) 자문위원

연구 및 프로젝트

- 부산대학교 병원: 개인 맞춤형 당뇨병 예방·관리 인공지능 시스템 외주 프로젝트 (2022)
- 연세대학교 병원: 영상의학과 FLASK 고도화 연구 (2023, 2024)
- 산업수학 스터디 그룹 (국가수리과학연구소): 피부암/유방암 분류, 유전자 정보 분석
- 폐플라스틱 이미지 객체 검출 경진대회 3위 (2021)
- 동국대학교 수학과 연구원(2025.10 ~ / 양자 AI, XAI 연구 진행 중)

논문

- Quantum computing for the optimization of CT image reconstruction, IEEE ICTC 2022

강의 및 교육 활동

대학 강의

- 고려대학교, 인천대학교, 동양미래대학교, 목포대학교
- 이젠 아카데미: 파이썬/ADSP 과정

기업 교육

- 현대자동차 연구원
- 한화, 하나금융, 한전 KDN, IBK 기업은행, 아가방앤컴퍼니, IM증권

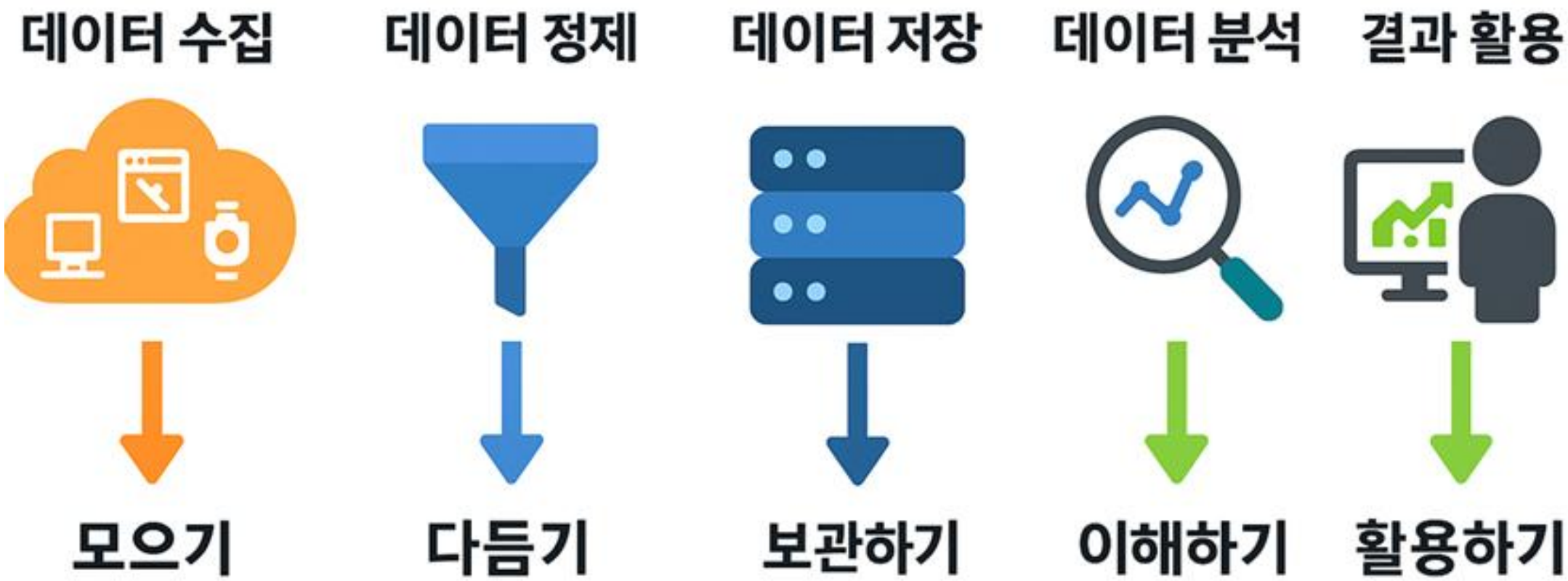
공공/국비 과정

- LG 헬로비전 DX 데이터 스쿨 (4기)
- 보건·바이오 AI 활용 전문가 양성과정 (1-3기)
- 공공데이터 청년 인턴 멘토
- 충청 ICT 취창업 역량강화 프로그램 (SQLD)
- 서울아이티고등학교

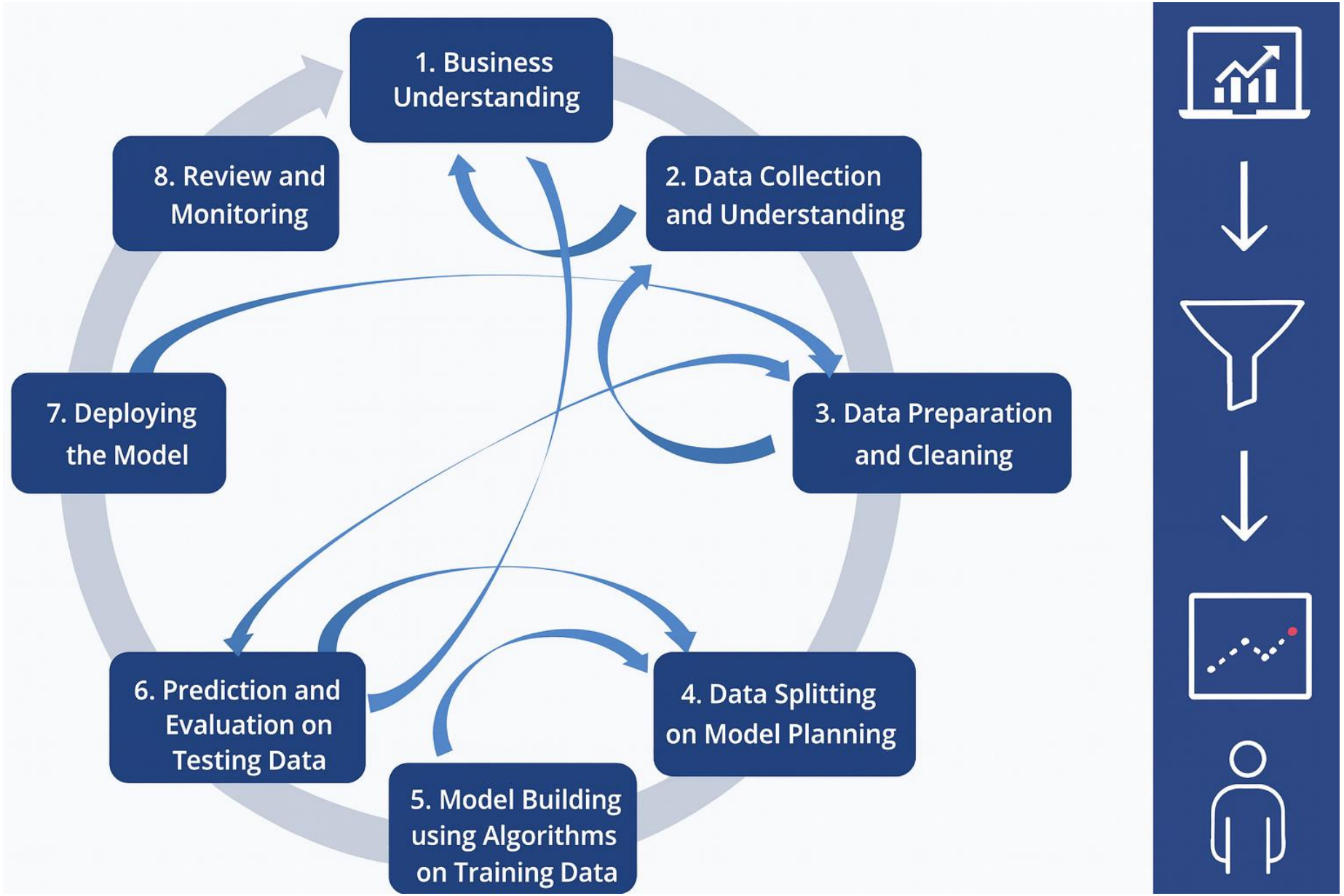
아카데미 & 부트캠프

- 모두의연구소 Aiffel 퍼실리테이터 (1기)
- Upstage AI Boostcamp 멘토 (1-6기)
- 패스트캠퍼스/멀티캠퍼스 조교 및 멘토
- 공개SW개발자대회 멘토

데이터 처리와 분석의 전체 과정



데이터 사이언스 프로젝트의 전체 생명주기



시계열 분석

- 시계열 자료

  : 시간의 흐름에 따라 관찰된 데이터 ex) 일별 기온, 주가, 전력 사용량, 자전거 대여량

  : 시간 순서에 따른 패턴(경향, 계절성, 주기, 불규칙성)를 분석해 미래를 예측하거나 변동 원인을 이해하는 통계 방법

구성요소	의미	예시
경향(Trend)	장기적 증가, 감소	전력 사용량 지속 증가
주기(Cycle)	비정기적 반복	경기순환, 부동산 경기
계절성(Seasonality)	일정 주기 반복	여름 전력 피크, 겨울 매출 감소
불규칙성(Irregular)	예측 불가능 요인	천재지변, 이벤트 이슈

- 시계열 분석

  : 과거 데이터를 이용해 시간의 흐름에 따른 변화 패턴을 모델링

  : 그 모델을 바탕으로 미래 값을 예측

  : 일반 회귀분석은 $X-Y$관계를 본다면, 시계열 분석은 시간(Time) 자체를 독립변수로 고려

데이터 관점에 따른 분류

	횡단면 데이터 (Cross Sectional)	시계열 데이터 (Time Series)	패널 데이터 (Panel)
정의	특정시점 + 다수독립변수 (여러 변수의 관측치)	다수시점+특정독립변수 (여러 시점에 대해 관측한 자료)	다수독립변수+다수시점 (여러 시점에 대해 여러 변수 자료)
예시	2000년의 각 기업의 매출액	기업 A의 연도별 매출액	기업 A,B,C,D의 2000년부터 2004년까지의 관측된 모든 매출액 자료
특징	값 독립적, 모집단 중 특정 시점 표본추출	값 Serial-correlation /Trend/Seasonality 등	시점/변수 일치로 연구자들이 가장 선호

Sales

Annual sales figures for each company in millions of KRW.

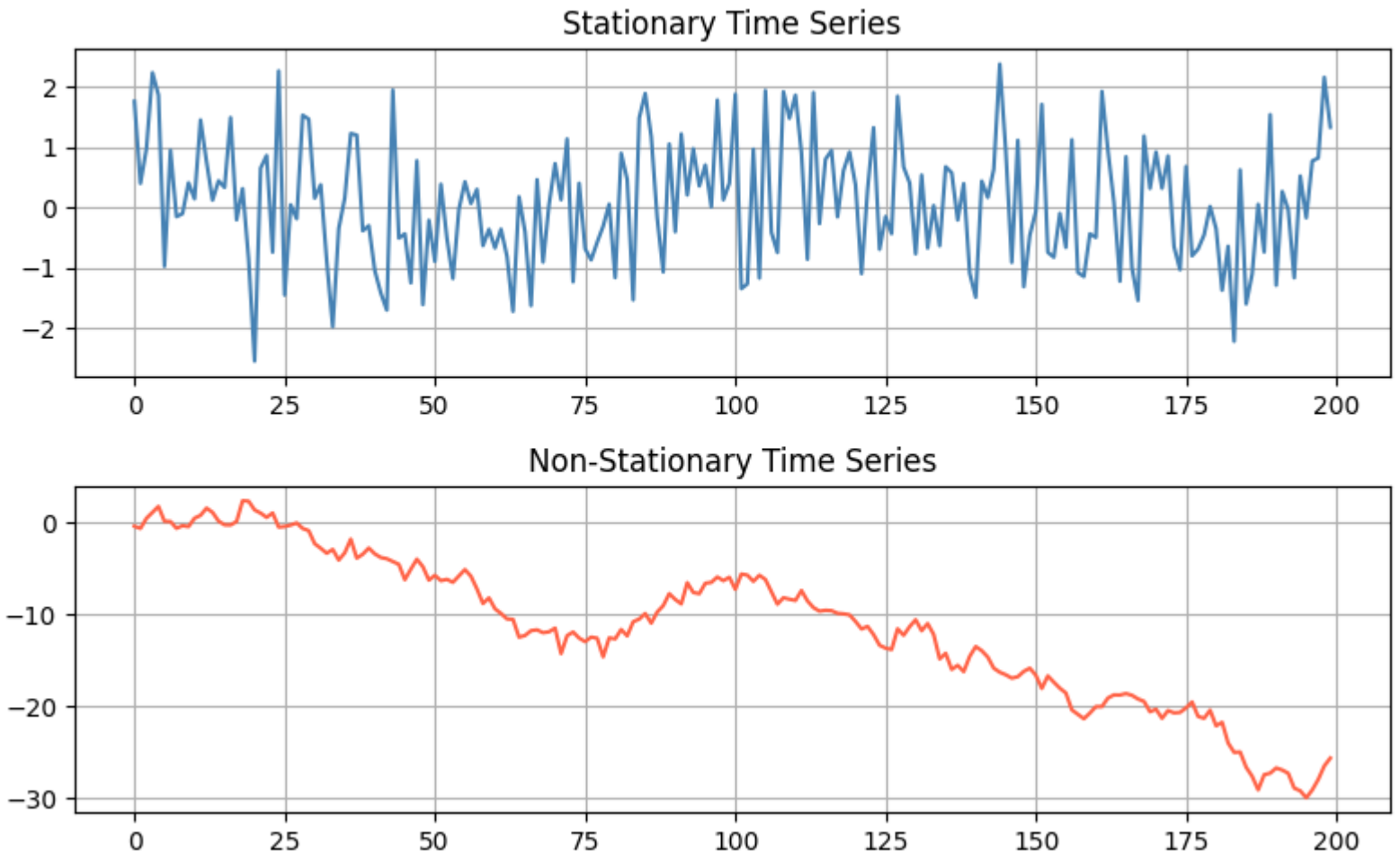
패널

Year	A	시계열 B	C	D 횡단면
2000	1,881	11,296	24,855	6,929
2001	1,900	12,007	23,130	5,693
2002	1,994	12,659	23,519	6,145
2003	15,24	13,091	20,761	6,769
2004	2,107	13,636	22,505	7,902

출처 : http://triki.net/study/3343

정상성과 비정상상의 구분

- 시계열 데이터는 시간에 따라 평균이나 분산 같은 통계적 특성이 변할 수 있다.
- 이러한 변화가 없으면 정상성(Stationary), 변화가 있으면 비정상성(Non-Stationary)이라 한다.



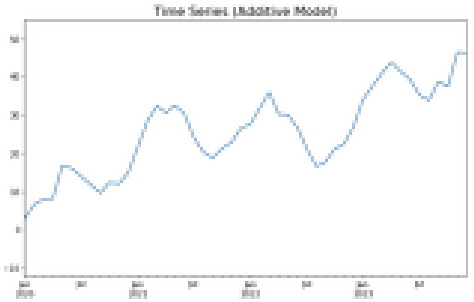
시계열 구성 요인

시계열 구성 요인
(Time Series Component Factors)

시계열 분해
(de-composition)

시계열 구성
(composition)

시계열 가법 모형
(Timeseries additive model)



추세 요인
(trend factor)

+

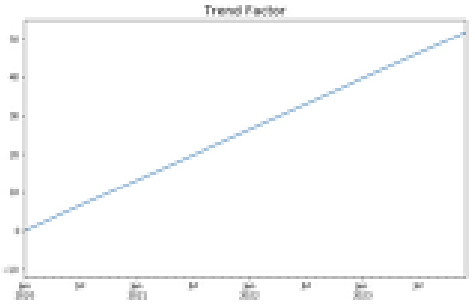
순환 요인
(cycle factor)

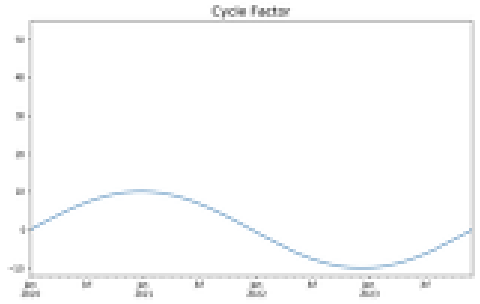
+

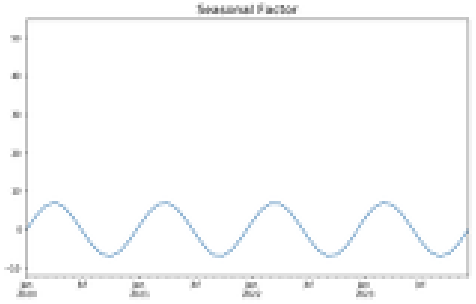
계절 요인
(seasonal factor)

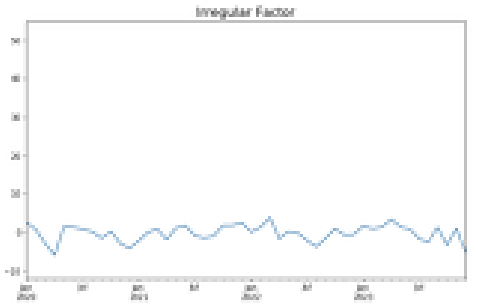
+

불규칙 요인
(irregular factor)









R, Python 데이터 분석과 프로그래밍의 친구 **http://rfriend.tistory.com**

정상성과 비정상성의 비교(Stationary VS Non-Stationary)

- **정상성(Stationary)**

  : 시계열의 평균, 분산이 시간에 따라 변하지 않음

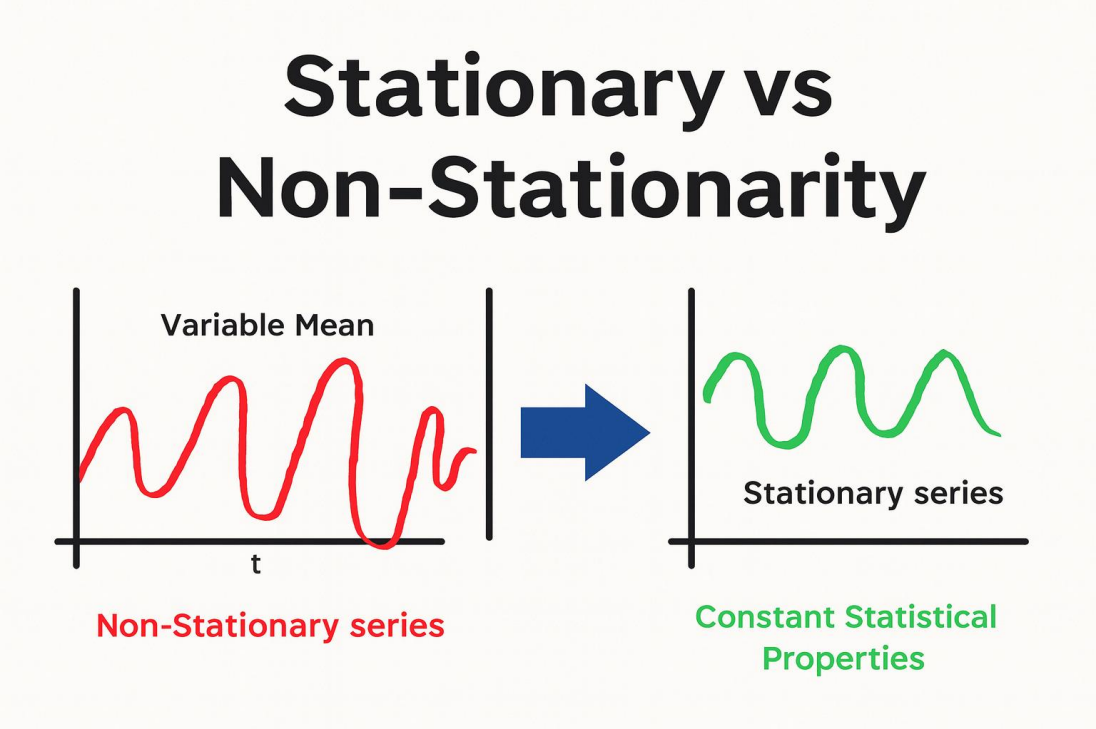
  : 주기적 변동이 없고, 과거와 미래의 통계적 성질이 동일함.

- **정상 시계열의 조건**

  : 평균은 모든 시점(시간 t)에 대해 일정하다.

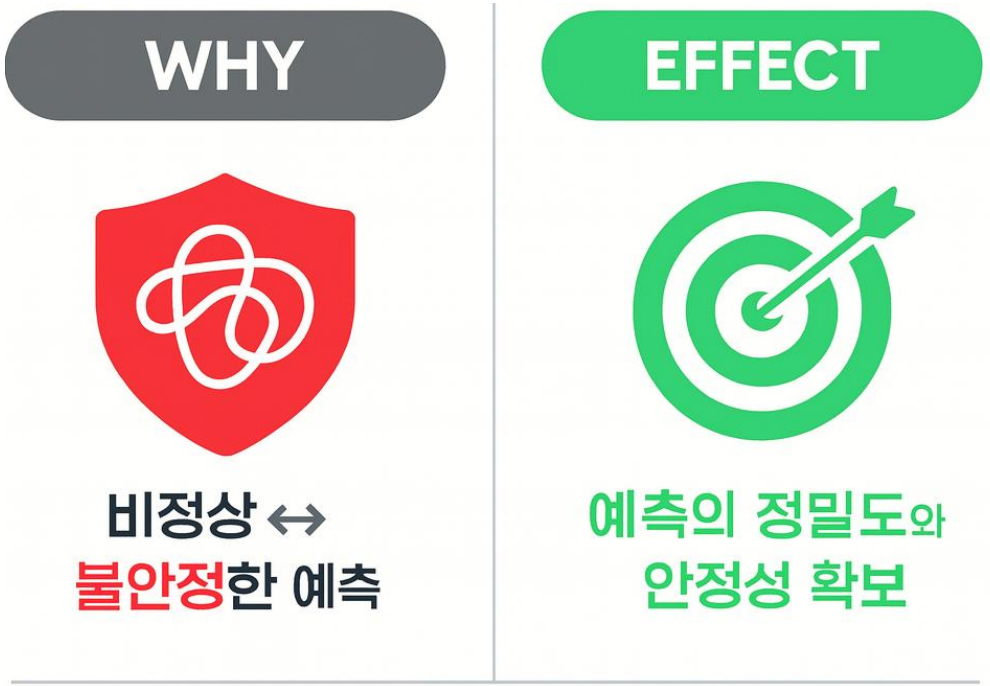
  : 분산은 모든 시점(시간 t)에 대해 일정하다.

  : 공분산은 시점(시간 t)에 의존하지 않고, 단지 **시차(lag)에만 의존**한다.



왜 정상성을 확보할까?

- 비정상 데이터를 정상화해 예측하면 모델이 안정적으로 동작함.
- 정상성 확보의 목적 : 예측값이 특정 범위 내에서 일정하게 유지되도록 함.
- **즉, 매출이 아니라 변동률을 예측하는 과정과 유사**
- 넓은 범위를 좁은 범위로 바꿔 예측의 정밀도를 높이는 것



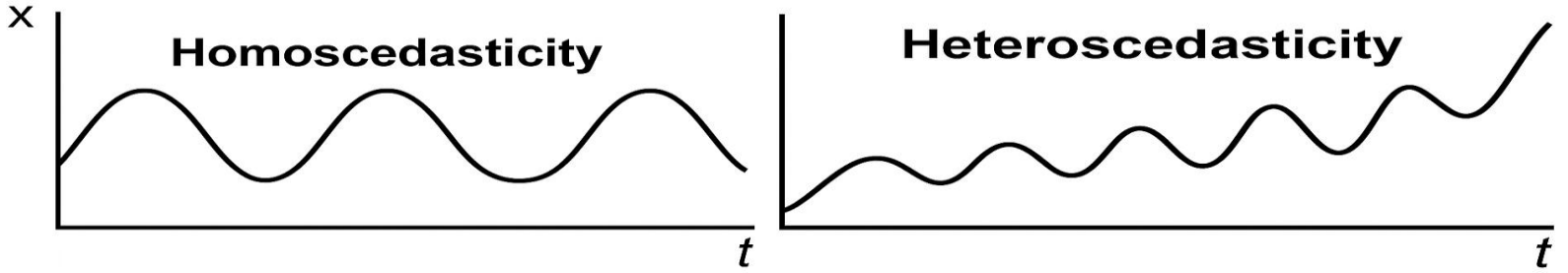
1. Stationary 가정 : 데이터가 정상성일수록 분석효과 ↑, 파라미터 ↓
2. 백색잡음(White noise) : 잔차 검증 역시 정상성을 전제로 함

1. 모델 단순화 → 분석 효율 ↑
2. 과적합 방지 → 일반화 성능 ↑
3. 잔차(백색잡음) 또한 정상성을 만족해야 신뢰 가능

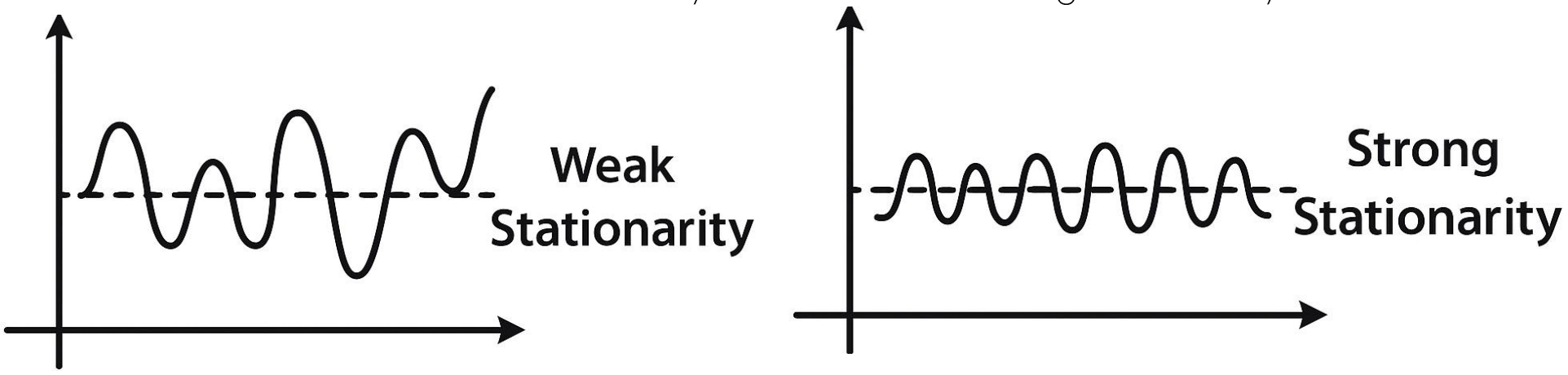
통계적 특성(Statistical Properties)

- 평균, 분산, 공분산을 포함한 모든 분포적 특성(Distributional characteristics)을 총칭

구분	의미
Homoscedasticity(등분산성)	분산이 일정하고 발산하지 않음(유한함)
Heteroscedasticity(이분산성)	분산이 일정하지 않거나 발산함



- 확률과정(Stochastic Process) : 시간의 흐름에 따라 확률적으로 변하는 구조
- 확률과정의 특성에 따라 약정상성(Weak Stationary) 또는 강정상성(Strong Stationary) 상태를 가질 수 있음.



확률과정(Stochastic Process)

- 시간 t에 따라 확률적으로 변하는 변수들의 집합 (예 : 주가, 온도, 수요량)

$$\{X_t : t \in T\}$$

- 각 시점이 값이 **확률변수**이므로, 전체는 **확률 과정**

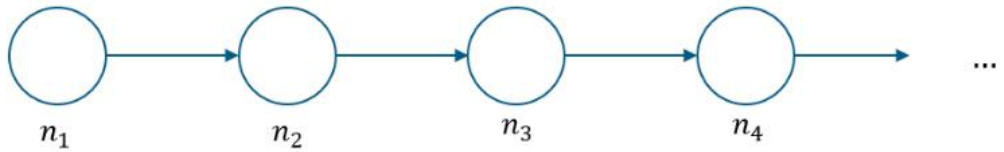
- 독립과정 VS 의존과정
  - 독립과정(Independ Process)

    : X_t가 과거 $X_{t-1}, X_{t-2}, \cdots$와 무관

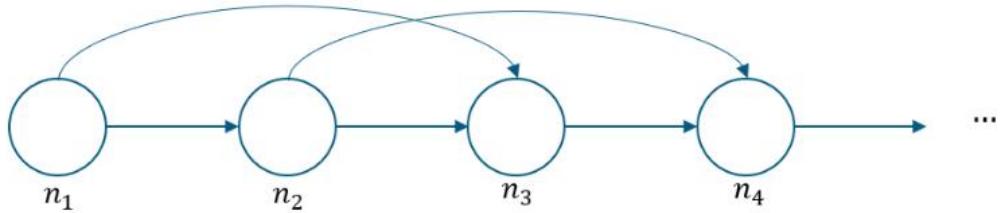
    : 현실 데이터(주가, 날씨)는 거의 해당되지 않음
  - 의존과정(Dependent Process)

    : 과거가 현재에 영향을 미침.

    : 이를 수학적으로 표현하는 대표적 개념이 마르코프 프로세스(Markov Process)



1차 마르코프 모델



2차 마르코프 모델

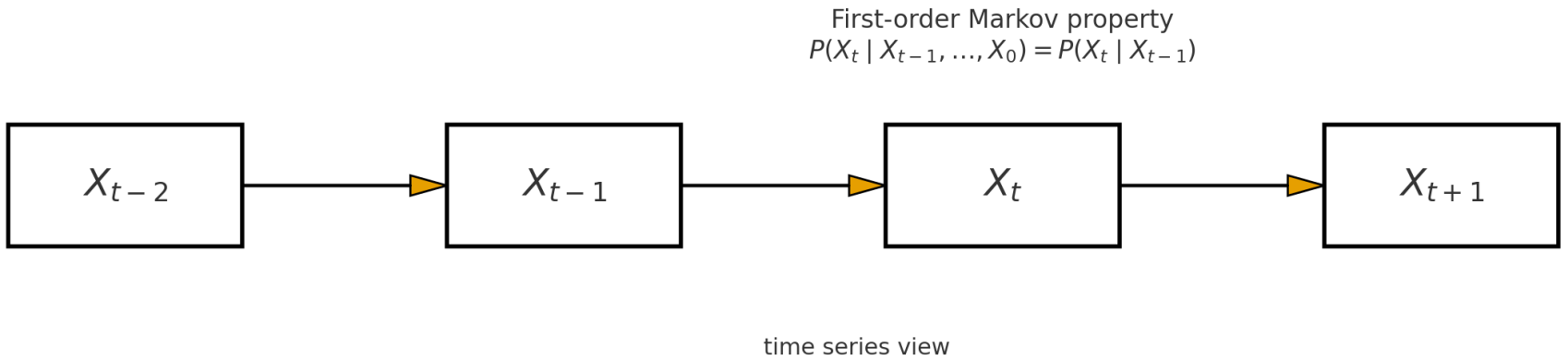
출처 : https://khuda.notion.site/2-Machine-Learning-08-10-cebf1c5ed5ee40b39aee15499729d814

Markov Process & Markov Chain

$$P\left(x^{(t+1)}\middle|x^{(0)},\cdots,x^{(t-1)},x^{(t)}\right)=P\left(x^{(t+1)}\middle|x^{(t)}\right)$$

$$\Rightarrow x^{t+1}\ is\ \ indepent\ of\ \left\{x^{(0)},\cdots,x^{(t-1)}\right\}\ \ given\quad x^{(t)}$$

즉, 현재 상태가 주어지면 미래는 과거와 독립이다 -> 현재 상태만으로 미래를 예측



- 시간과 상태가 이산이면 Markov Chain, 시간이 연속적이거나 상태공간이 연속이면 Markov Process라고 부른다.
- 상태공간(State space) : Markov ProcessL나 Markov Chain에서 시스템이 가질 수 있는 모든 가능한 상태들의 집합.

$$S=\{s_1,s_2,\cdots,s_n\}$$

여기서 S는 상태공간, 각 s_i는 시스템이 취할 수 있는 하나의 상태(state)

-> 시계열에서 "상태공간"은 시점별 관측값 또는 잠재변수(Latent variable)가 가질 수 있는 모든 가능성 범위.

정상성(Stationary)

- 약정상성(Weak Stationarity)

  : 평균과 분산, 공분산이 일정하다면 약정상성을 만족함

  : 대부분의 시계열 분석에서는 이 조건만 충족해도 충분함

  1. $E[X_t] = \mu \rightarrow$ 평균일정
  2. $Var(X_t) = \sigma^2 < \infty \rightarrow$ 분산 유한, 고정
  3. $Cov(X_{t+h}, X_t) = \gamma(h) \rightarrow$ 공분산 시점이 아닌 시차 h에만 의존

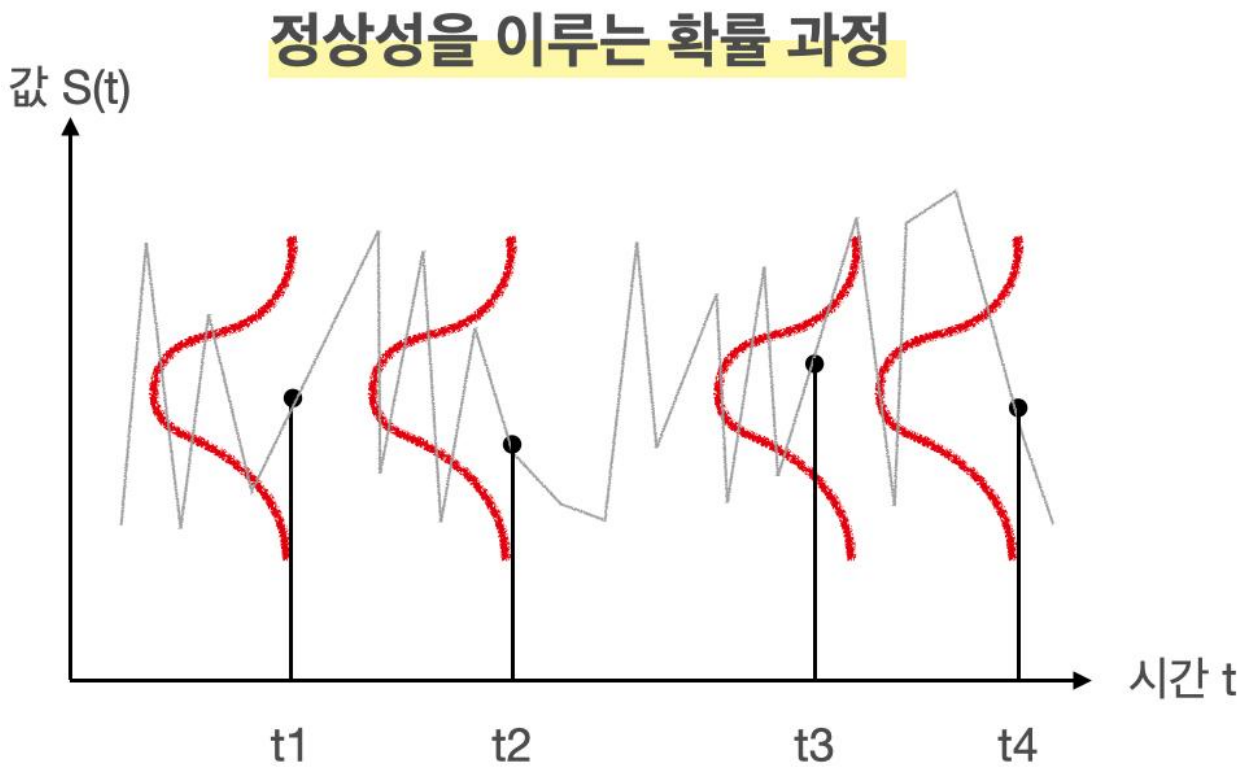
- 강정상성(Strong Stationarity)

  : 모든 시점 t에서 확률분포 전체가 동일한 경우

  : 결합분포 $(X_t, X_{t+1}, \cdots, X_{t+k})$가 시간 이동에 대해 동일

  : 약정상성보다 더 강한 제약

  : 대부분의 실무에서는 약정상성만 가정



출처 : https://velog.velcdn.com/images/knh19950524/post/f4ac60a9-b917-4a40-ac76-a4dfe0a2f765/image.png

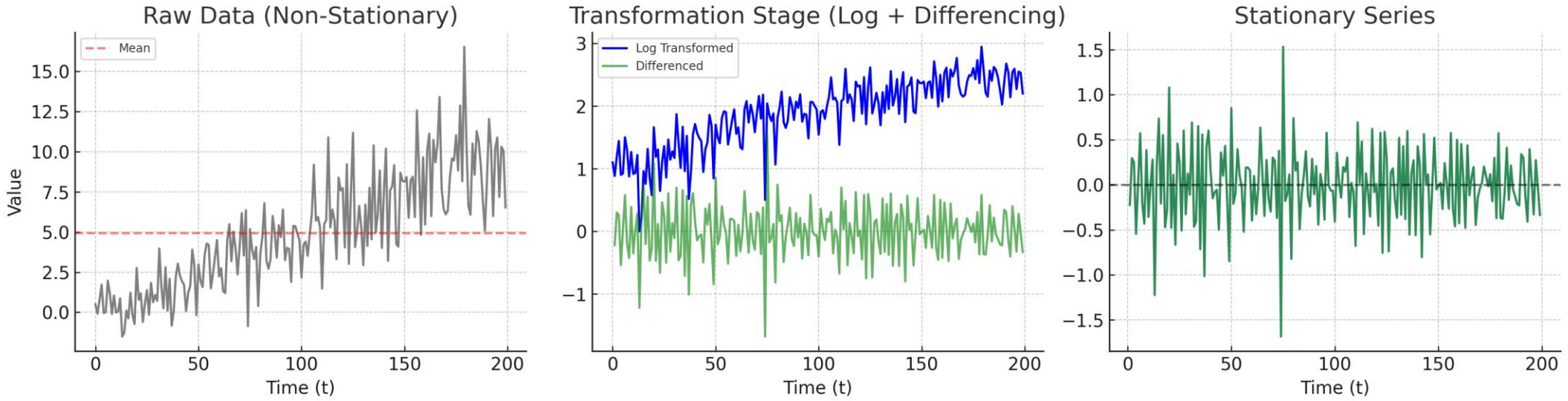
정상 시계열 전환의 목적

- 비정상 데이터를 정상성 만족 형태로 변환하여 분석 수행

비정상 원인	변환 방법	설명
평균이 일정하지 않음	차분(Differencing)	시점 간 차이를 계산하여 추세 제거
계절성이 존재함	계절 차분(Seasonal Differencing)	주기적 패턴 제거
분산이 일정하지 않음	로그 변환(log transform)	데이터의 스케일 안정화

- 차분 ($\nabla X_t = X_t - X_{t-1}$)

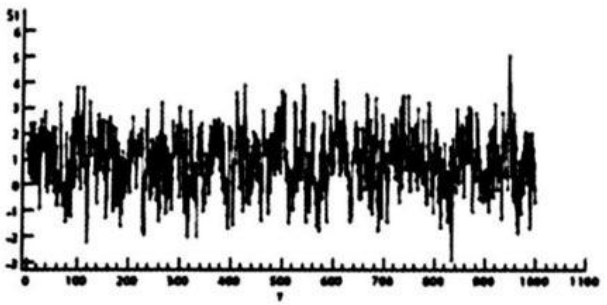
  : 현 시점의 자료 값에서 전 시점의 자료 값을 빼 주는 것 의미함. / 현재시점(t_i)의 자료에서 인접시점(t_{i-1})의 자료를 차감하는 것.



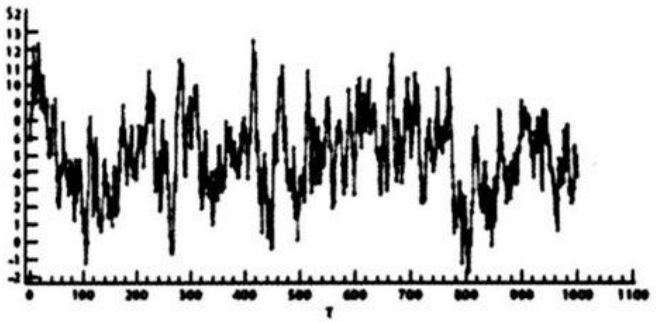
정상 시계열의 특성(Properties of Stationary Series)

정상 시계열의 모습

$y_t = 0.5 + 0.5\, y_{t-1} + e_t,\ e_t \sim iid\ N(0,1)$



$y_t = 0.5 + 0.9\, y_{t-1} + e_t$



$E(y_t) = \mu$ [일정한 평균]

$var(y_t) = \sigma^2$ [일정한 분산]

$cov(y_t, y_{t+s}) = cov(y_t, y_{t-s}) = \gamma_s$ [공분산은 t가 아닌 s에 의존함]

- 정상 시계열은 평균·분산이 일정하고, 자기공분산이 시차에만 의존한다.
- 항상 평균으로 회귀하려는 경향이 있으며, 변동 폭은 일정하다.
- 비정상 시계열은 시기별 통계 특성이 달라 일반화가 어렵다.

- 시차(Lag)
  - 시계열 데이터에서 서로 다른 두 시점의 간격
  - $ex)\ h = 1$ 이면 한시점 뒤, h=2이면 두 시점 뒤 데이터를 의미

자기공분산(Autocovariance, $\gamma(h)$)

- 자기공분산은 시차 h만큼 떨어진 두 시점 y_t와 y_{t+h} 사이의 함께 변하는 정도를 나타냄.
- 즉, "지금 값과 과거 값이 얼마나 비슷한 방향으로 움직이느냐"를 수치를 표현한 것

$$\gamma(h) = E[(y_t - \mu)\,(y_{t+h} - \mu)]$$

시차 h	변환 방법	직관적 의미
$\boldsymbol{h = 0}$	$\gamma(0) = \mathrm{Var}(y_t)$	자기 자신과 공분산 -> 분산
$\boldsymbol{h > 0}$	$\gamma(h) = \mathrm{Cov}(y_t, y_{t+h})$	h 시점 떨어진 두 값의 유사도
$\boldsymbol{h < 0}$	$\gamma(-h) = \gamma(h)$	대칭성 (공분산은 순서 무관)

- $\gamma(h) > 0$: 두 시점이 같은 방향으로 움직임(양의 상관)
- $\gamma(h) < 0$: 반대 방향으로 움직임(음의 상관)
- $|\gamma(h)|$ 가 클수록 강한 관계, 0에 가까울수록 무관계
- 이론적으로 범위의 제한은 없다.

자기상관계수(Autocorrelation Coefficient, $\rho(h)$)

- 자기 공분산을 표준화(normalization)해서 단위에 상관없이 비교할 수 있게 만든 값

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}$$

- $\gamma(0)$은 분산이므로, 항상 양수
- 결과적으로 $\rho(h)$의 값은 $-1 \sim +1$ 사이로 제한됨.

값의 범위	해석	의미
$\rho(h) = 1$	완전 양의 자기상관	두 시점이 동일한 움직임
$\rho(h) = 0$	자기상관 없음	독립, 무작위(white noise)
$\rho(h) = -1$	완전 음의 자기상관	정반대 방향으로 움직임

시계열자료 분석방법

1) 분석 접근법

구분	주요 방법	특징
이론 기반	회귀분석(계량경제), Box-Jenkins	통계 모델 중심의 구조적접근
직관적 접근	지수평활법, 시계열 분해법	시간 흐름에 따른 변동 패턴 분석
장기 예측	회귀모형 기반	추세(Trend) 중심의 분석
단기 예측	Box-Jenkins, 지수 평활, 분해법	최근 변동 반영, 단기 변동성 대응

2) 자료 형태에 따른 분석 방법

구분	주요 방법	특징
단일 시계열 분석	Box-Jenkins(ARIMA), 지수평활, 시계열 분해법	한 변수의 시간 변화 패턴 분석
시간(t) 포함 회귀형 분석	회귀 모형, 경기 · 소매지수 등	독립변수 중 하나가 '시간'일 때 사용
다중 시계열 분석	VAR, 상태공간모형, 다변량 ARIMA	여러 변수의 상호작용 분석 (ex. 경기 · 소비 · 금리)

시계열자료 분석방법

3) 이동평균법(Moving Average Method)

: 일정 기간의 평균값을 계산하여 노이즈를 줄이고 추세를 예측하는 기법

: 시계열 자료에서 계절변동과 불규칙변동을 제거하여 추세변동과 순환변동만 가진 시계열로 변환하는 방법으로 사용.

$$F_{n+1} = \frac{1}{n}\left(Z_n + Z_{n-1} + \dots + Z_{n-m-1}\right) = \frac{1}{m}\sum_{t}^{n} Z_t, \quad t = n-m+1$$

- m은 이동평균한 특정기간, Z_n은 가장 최근 시점의 데이터

  -〉 n개의 시계열 데이터를 m기간으로 이동평균하면 n-m+1개의 이동평균 데이터가 생성된다.

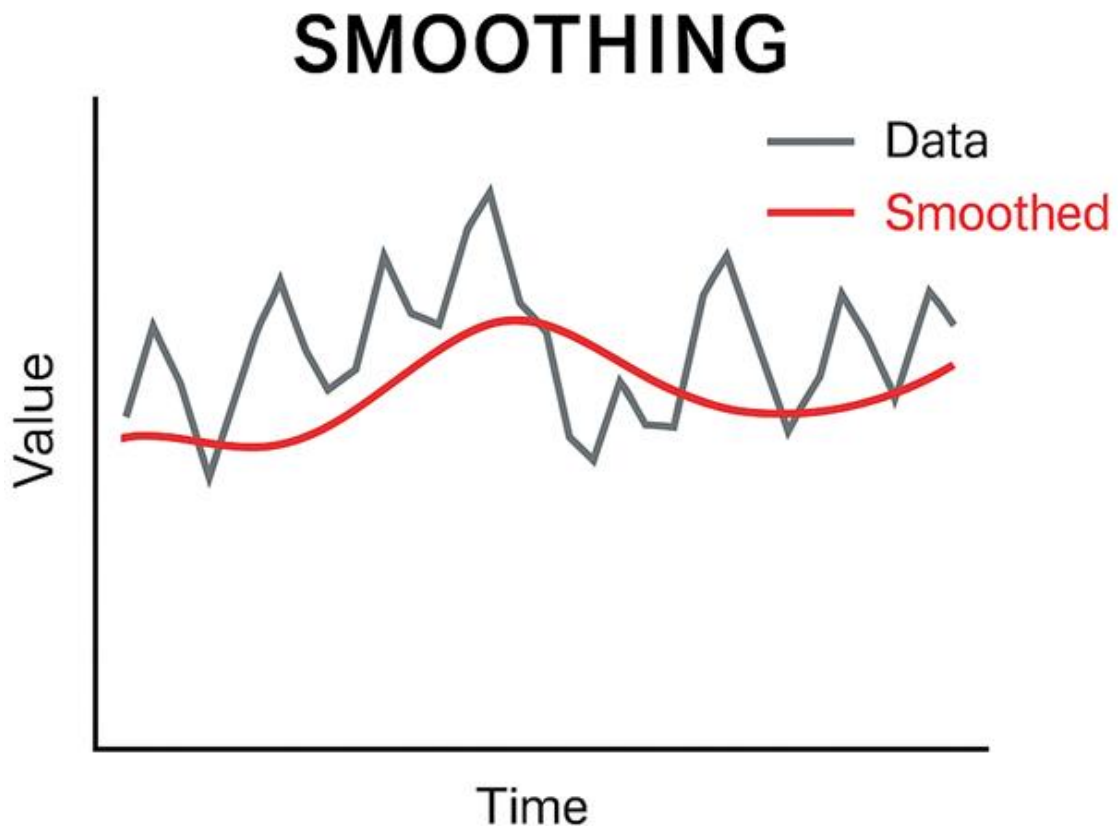
[특징]

- 간단하고 쉽게 미래를 예측할 수 있으며 자료의 수가 많고 안정된 패턴을 보이는 경우 예측의 품질이 높음.
- 특정 기간 안에 속하는 시계열에 대해서는 동일한 가중치를 부여함
- 일반적으로 시계열 자료가 뚜렷한 추세가 있거나 불규칙변동이 심하지 않은 경우에는 짧은 기간(m개의 개수가 적음)의 평균을 사용, 반대로 불규칙변동이 심한 경우 긴 기간(m개의 개수가 많음)의 평균을 사용.
- 가장 중요한 것은 적절한 기간을 사용하는 것, 즉 적절한 n의 개수를 결정하는 것임.

평활화(Smoothing)

- 시계열 데이터에는 불규칙한 단기 변동(Noise)이 존재.
- 평활화(Smoothing)은 이러한 변동을 줄이고 전체적인 추세(Trend)나 패턴을 드러내는 기법
- -〉 데이터의 “흐름만 남기고 잡음은 제거”

  -〉 과거 데이터를 일정 규칙으로 평균화하여 미래의 경향을 파악한다.



시계열자료 분석방법

4) 가중이동평균(Weighted Moving Average))

: 일정 기간의 관측값에 서로 다른 가중치를 부여하고 평균을 계산하는 방법.

: 최근 데이터일수록 더 큰 가중치를 주어 변화에 빠르게 반응하면서도 단기 변동을 평활화함.

: Window 크기 n과 가중치 벡터 $w = (w_0, w_1, \cdots, w_{n-1})$가 필요하며, 일반적으로 $w_0 < w_1 < \cdots < w_{n-1}$

$$\bar{Y} = \frac{\sum_{i=0}^{n-1} w_i Y_{t-i}}{\sum_{i=0}^{n-1} w_i}$$

Y_t : 시점 t의 실제 관측값

w_i : 각 시점에 부여된 가중치(최근 데이터일수록 큼)

n : 이동평균에 포함되는 데이터 개수

[특징]

- 단기 노이즈를 줄이고 데이터의 흐름을 부드럽게 표현
- 단순 이동평균보다 반응 속도가 빠름
- Window이 커질수록 더 부드럽지만, 추세 반응이 느려짐
- 최근 데이터에 더 큰 가중을 주면 급격한 변화를 빠르게 포착

시계열자료 분석방법

5) 지수평활법(Exponential Smoothing)

: 모든 과거 데이터를 사용하되, 오래된 데이터의 영향을 지수적으로 감소시키는 방식.

: 과거 정보의 영향을 점진적으로 줄이면서 단기 예측까지 포함하는 평활법

$$
\begin{aligned}
F_{n+1} &= \alpha Z_n + (1-\alpha)F_n \\
&= \alpha Z_n + (1-\alpha)[\alpha Z_{n-1} + (1-\alpha)F_{n-1}] \\
&= \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + (1-\alpha)^2 F_{n-1} \\
&= \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + (1-\alpha)^2[\alpha Z_{n-2} + (1-\alpha)F_{n-2}] \\
&\vdots \\
&= \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + \alpha(1-\alpha)^2 Z_{n-2} + \alpha(1-\alpha)^3 Z_{n-3} + \cdots
\end{aligned}
$$

F_{n+1}은 n시점 다음의 예측값

α는 지수평활계수

Z_n은 n시점의 관측값

->지수평활계수가 과거로 갈수록 지수형태로 감소하는 형태, 최근에 더 많은 가중치

[특징]

- 단기간에 발생하는 불규칙변동을 평활하는 방법
- 자료의 수가 많고 안정된 패턴을 보이는 경우일수록 예측 품질이 높음
- 지수평활계수는 과거로 갈수록 지속적으로 감소함
- 지수평활법은 불규칙변동의 영향을 제거하는 효과가 있으며, 중기 예측 이상에 주로 사용됨.

(단, 단순지수 평활법의 경우, 장기추세나 계절변동이 포함되는 시계열의 예측에는 적합하지 않음)

시계열모형

1) 자기회귀 모형(AR 모형, autoregressive model)

: 과거와 현재의 자신과의 관계를 정의 한 것.

: 이전 관측값(과거)이 이후 관측값(현재)에 영향을 주는 원리를 사용하기 때문에 Z를 활용함.

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)

-> p에 특정한 숫자 n을 대입하여 n 시점부터 회귀를 하고 싶으면 AR(n)으로 쓴다.

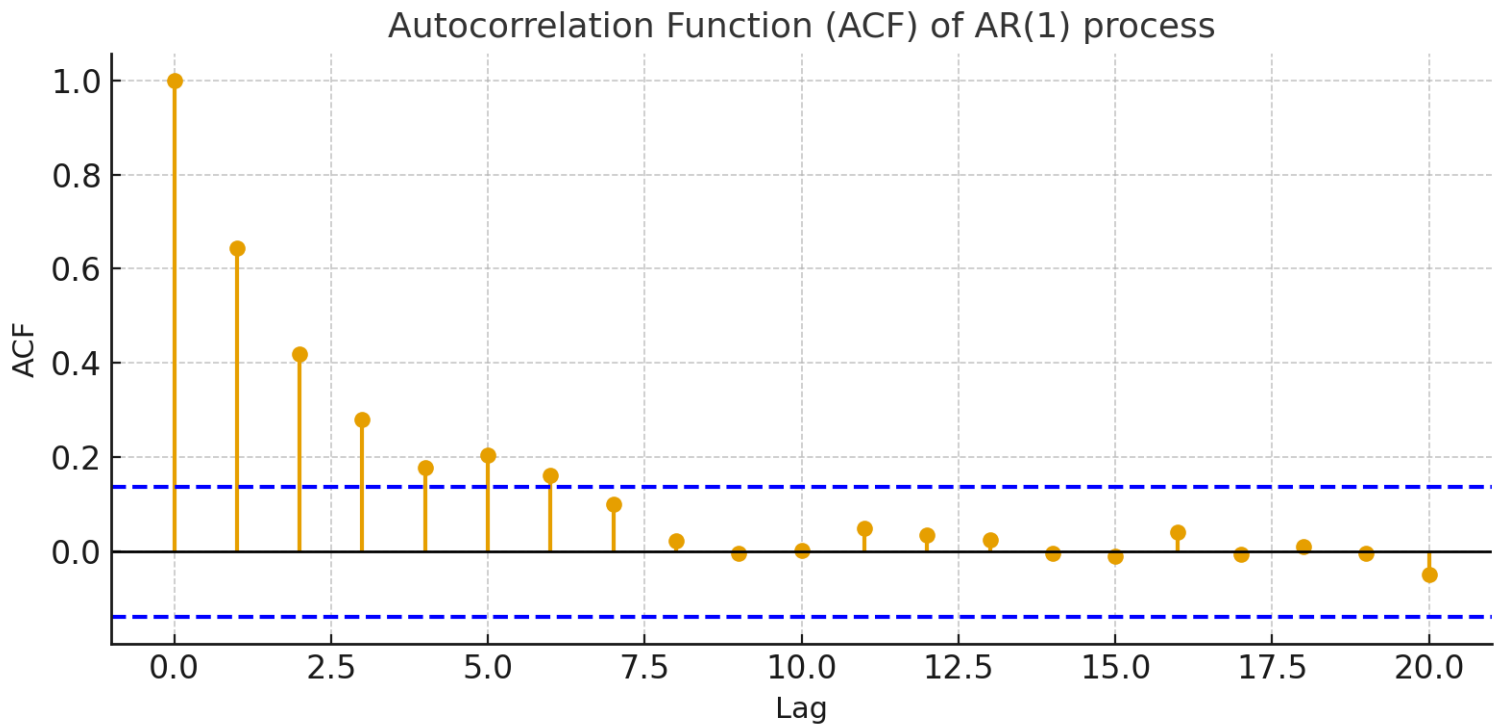
-> 평균이 0, 분산이 σ^2, 자기공분산이 0인 경우, 시계열간 확률적 독립인 경우 강(Strictly)백색잡음과정이라고 한다.

-> 백색잡음 과정이 정규분포를 따를 경우 이를 가우시안(Gaussian) 백색잡음과정이라고 한다.

출처 : https://blog.naver.com/tjgml1343/222077619748

자기 상관 함수(ACF, AutoCorrelation Function)

- 과거 시점의 데이터가 현재 데이터에 얼마나 영향을 미치는지 측정하는 함수



ACF(lag k)

$$\rho_k = \frac{Cov(X_t, X_{t-k})}{Var(X_t)}$$

시점 간의 상관관계 0~1 사이의 값

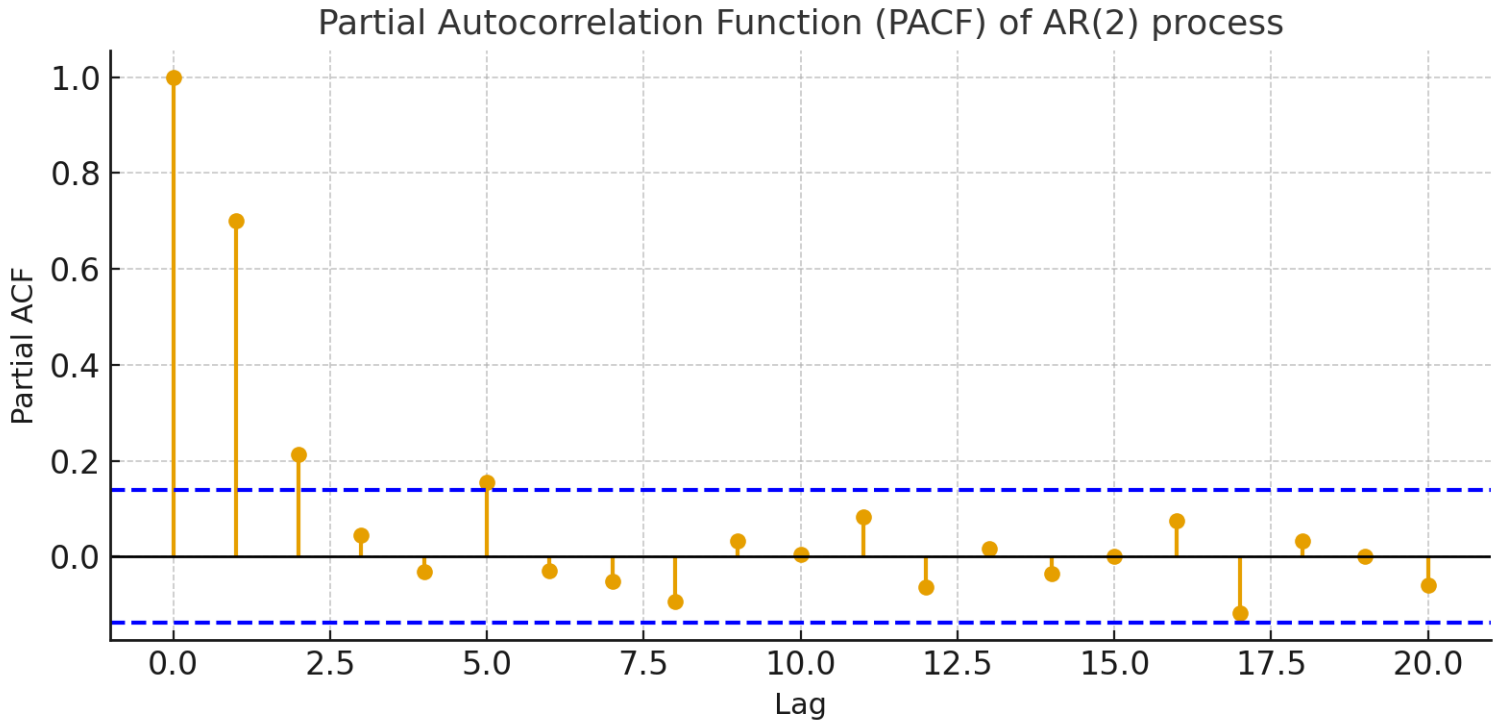
점선(신뢰구간)안 -〉 유의하지 않음.

- ACF가 점차 감소하는 형태 -〉 자기회귀(AR) 모형 특성, 특정 시점 이후 0에 가까워지면 시계열의 정상성(Stationarity)을 시사

$X_t = \emptyset X_{t-1} + \epsilon_t, \rho = \emptyset^k$, lag가 커질수록 상관이 지수적으로 줄어든다.

부분(편) 상관 함수(PACF, Partial AutoCorrelation Function)

- $\emptyset^k$ = 상위 시차의 영향을 제거한 후 남은 X_t, X_{t-k}간 상관
  - ACF는 전체 누적된 자기상관을 보여줌.
  - PACF는 직접적인 상관만 남기므로, AR(p)에서는 p시점 이후 급격히 0으로 절단(cut-off) 된다.



- lag=1, lag=2는 점선(신뢰구간) 밖에 있음 -〉 유의미한 상관, lag≥ 3에서는 대부분 점선 안 -〉 유의하지 않음.

$$X_t = \emptyset_1 X_{t-1} + \emptyset_2 X_{t-t} + \epsilon_t \quad , \epsilon_t \sim N(0, \sigma^2)$$

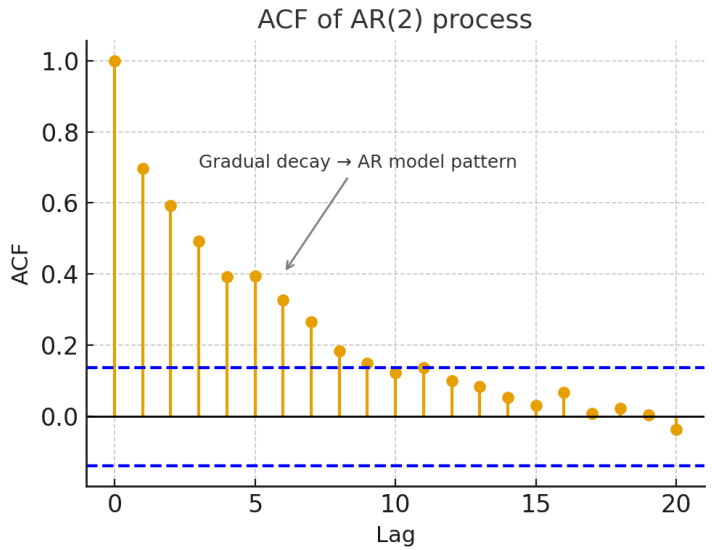
PACF는 다른 시점(lag)의 영향을 제거하고, “직접적으로“ 시점 t와 $t-k$사이의 순수 상관만을 계산함.

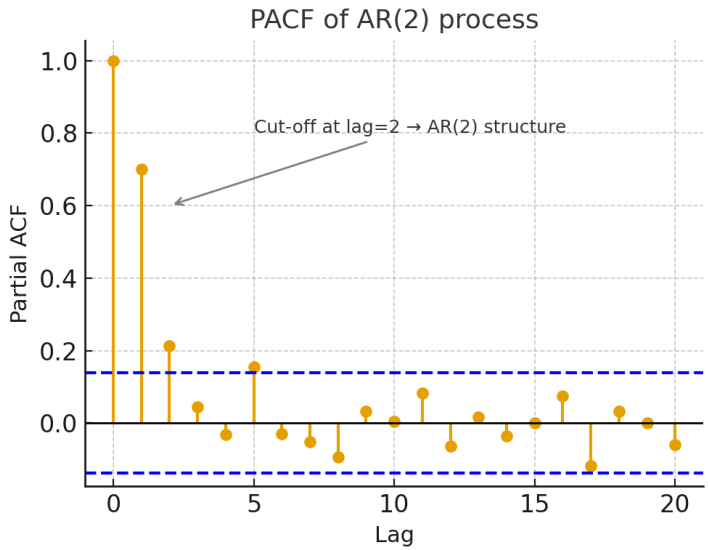
시계열모형

1) 자기회귀 모형(AR 모형, autoregressive model)

: 현재 값 $Y_t = \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \cdots + \emptyset_p Y_{t-p} + \epsilon_t$

함수	의미	AR(p)에서의 패턴
ACF	전체 시점 간의 누적 상관	점차 감소(exponential decay)
PACF	다른 시점 효과 제거 후 순수 상관	P시점 이후 급격히 절단(cut-off)





- ACF는 점차 감소, PACF lag=2이후 절단-〉AR(2) 모형임을 시사

시계열모형

2) 이동평균 모형(MA 모형, Moving Average model)

: 유한한 개수의 백색잡음의 결합이므로 언제나 정상성을 만족

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

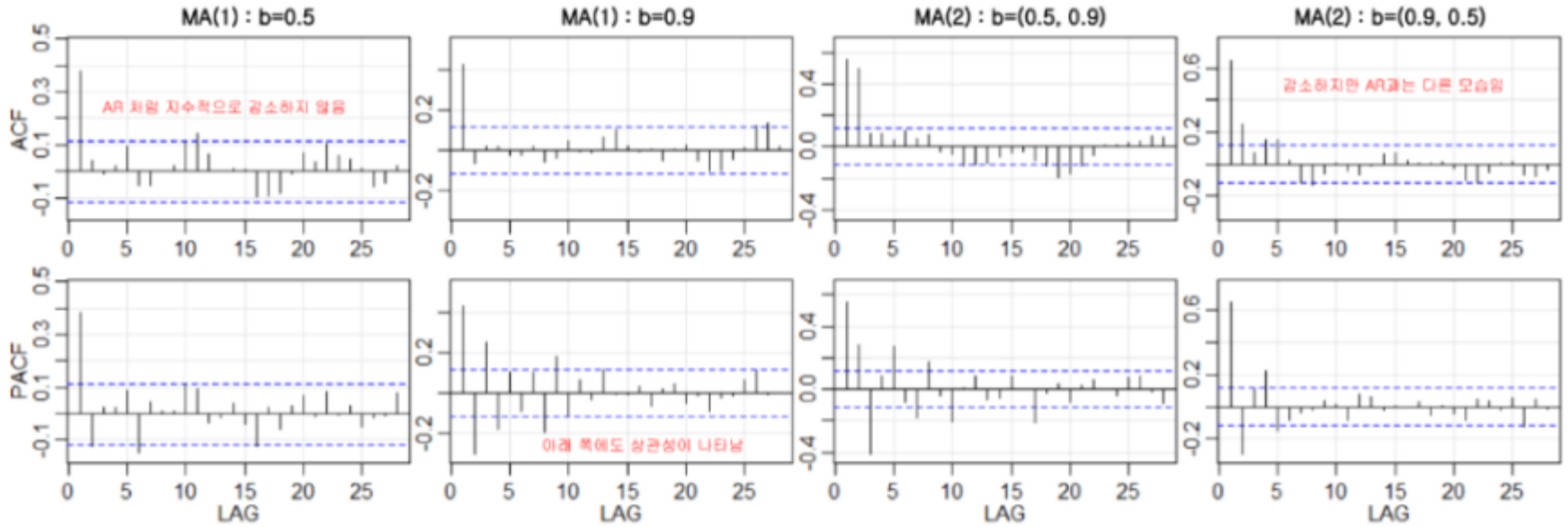
[MA1 모형] 1차 이동평균모형 : 이동평균모형 중에서 가장 간단한 모형으로 시계열이 같은 시점의 백색잡음과 바로 전 시점의 백색잡음의 결합으로 이뤄진 모형

$$Y_t = \alpha_t - \theta_1 \alpha_{t-1}$$

[MA2 모형] 2차 이동평균모형 : 바로 전 시점의 백색잡음과 시차가 2인 백색잡음의 결합으로 이뤄진 모형

$$Y_t = \alpha_t - \theta_1 \alpha_{t-1} - \theta_2\, \alpha_{t-2}$$

-〉 AR모형과 반대로 ACF에서 절단점을 갖고, PACF가 빠르게 감소



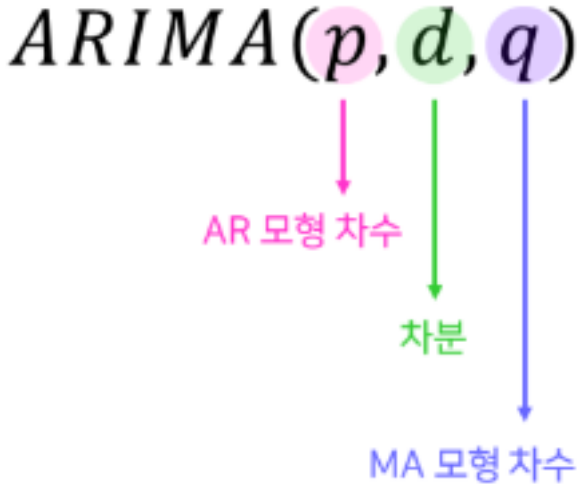
시계열모형

3) **자기회귀누적이동평균 모형(ARIMA(p, d, q) 모형**, autoregressive integrated moving average model)

: ARIMA 모형은 비정상시계열 모형이다.

: ARIMA 모형은 차분이나 변환을 통해 AR모형이나 MA모형, 이 둘을 합친 ARMA 모형으로 정상화 할 수 있다.

- P는 AR모형, q는 MA모형과 관련이 있는 차수이다.
- 시계열 $\{Z_t\}$의 d번 차분한 시계열이 ARMA 모형이면, 시계열 $\{Z_t\}$는 차수가 p, d, q인 ARIMA 모형, 즉 ARIMA(p,d,q)모형을 갖는다고 한다.



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(1,1,2)의 경우에는 1차분 후 AR(1) MA(2) ARMA(1,2) 선택 활용

  -〉 이런 경우 가장 간단한 모형을 선택하거나 AIC를 적용하여 점수가 가장 낮은 모형을 선택한다.

출처 : https://blog.naver.com/tjgml1343/222077619748

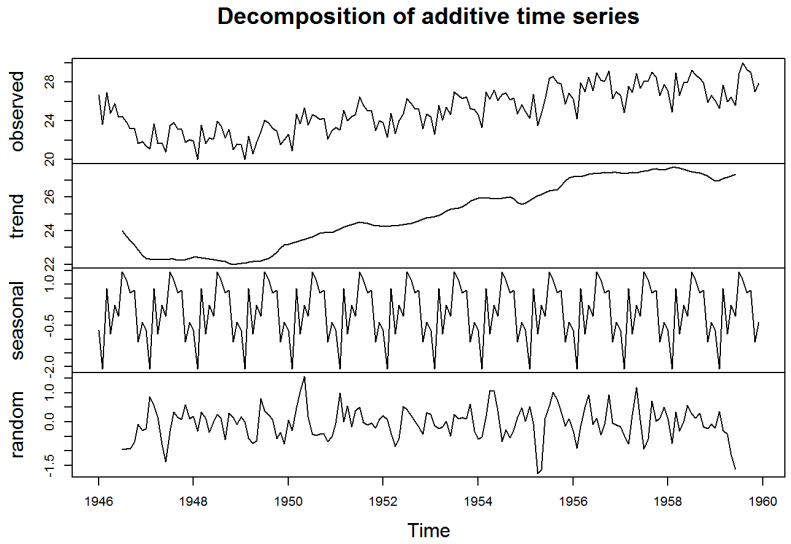
시계열모형

4) 분해시계열

: 시계열에 영향을 주로 일반적인 요인을 시계열에서 분리해 분석하는 방법. 회귀분석적인 방법을 주로 사용

$$Z_t = f(T_t, S_t, C_t, I_t)$$

- T_t : 경향(추세)요인 : 자료가 오르거나 내리는 추세, 선형, 이차식 형태, 지수적 형태 등
- S_t : 계절요인 : 요일, 월, 사계절 각 분기에 의한 변화 등 고정된 주기에 따라 자료가 변하는 경우
- C_t : 순환요인 : 경제적이나 자연적인 이유 없이 알려지지 않은 주기를 가지고 변화하는 자료
- I_t : 불규칙 요인 : 위의 세가지 요인으로 설명할 수 없는 오차에 해당하는 요인



시계열 분석 참고 : R 시계열 분석 Time Series ARIMA(woosa7.github.io/R-시계열분석-Time Series-ARMIA/)

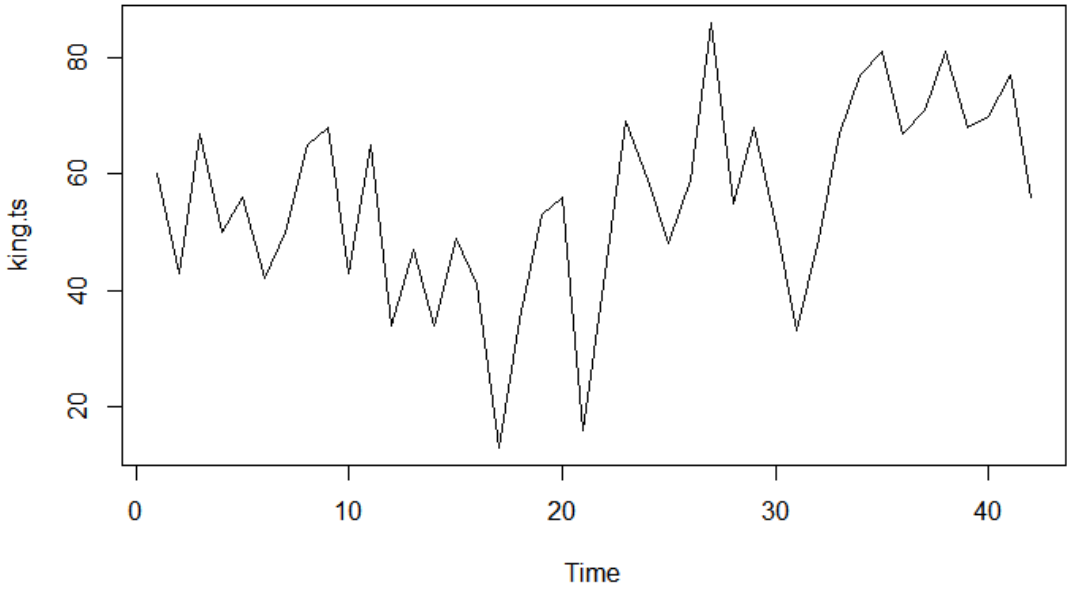
출처 : https://woosa7.github.io/R-%EC%8B%9C%EA%B3%84%EC%97%B4%EB%B6%84%EC%84%9D-Time-Series-ARIMA/

R을 이용한 시계열분석

- 영국 왕들의 사망 시 나이 데이터를 이용한 시계열 분석

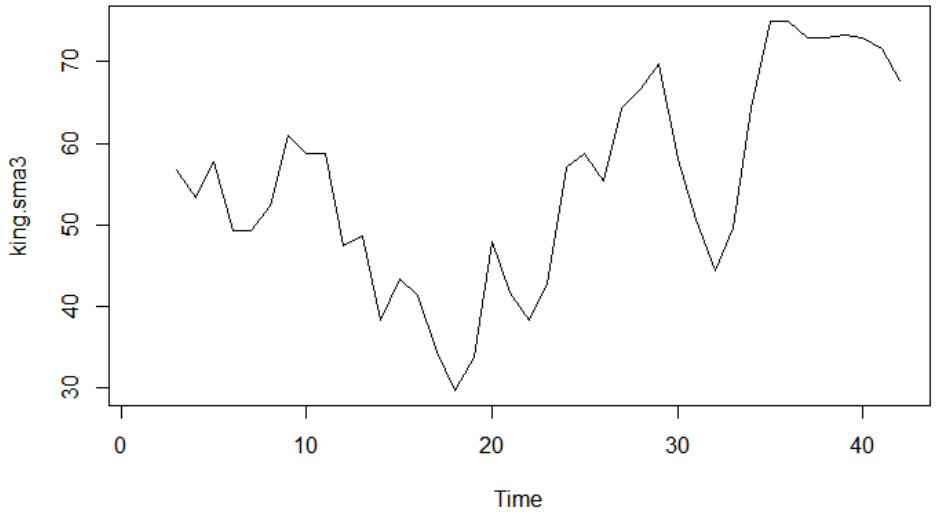
  - 영국 왕 42명의 사망 시 나이 예제는 비 계절성을 띄는 시계열 자료
  - 비계절성을 띄는 시계열 자료는 트렌드 요소, 불규칙 요소로 구성
  - 20번째 왕까지는 38세에서 55까지 수명을 유지하고, 그 이후부터는 수명이 늘어서 40번째 왕은 73세까지 생존

```
library(tseries)
library(forecast)
library(TTR)
king <- scan("http://robjhyndman.com/tsdldata/misc/kings.dat", skip = 3)
king.ts <- ts(king)
plot.ts(king.ts)
```

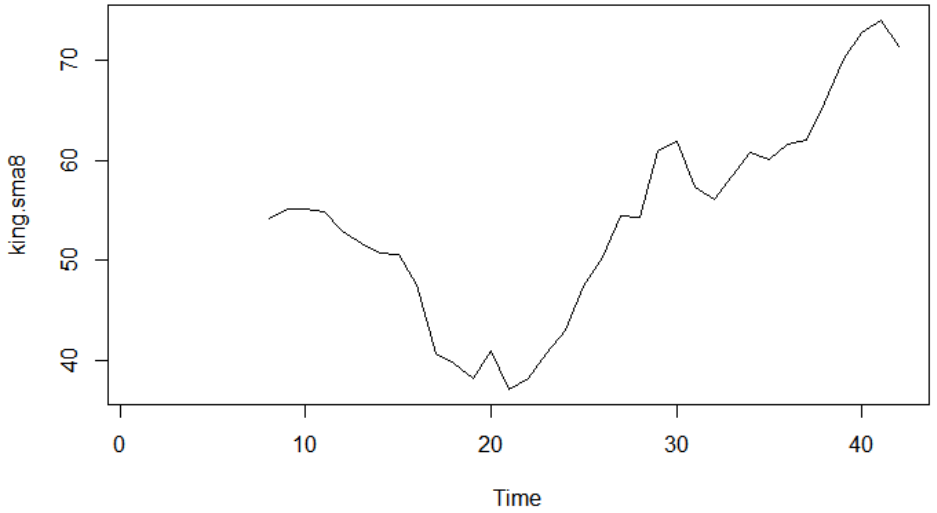


R을 이용한 시계열분석

```
#3년마다 평균을 내서 그래프를 부드럽게 표현
king.sma3 <- SMA(king.ts, n=3)
plot.ts(king.sma3)
```



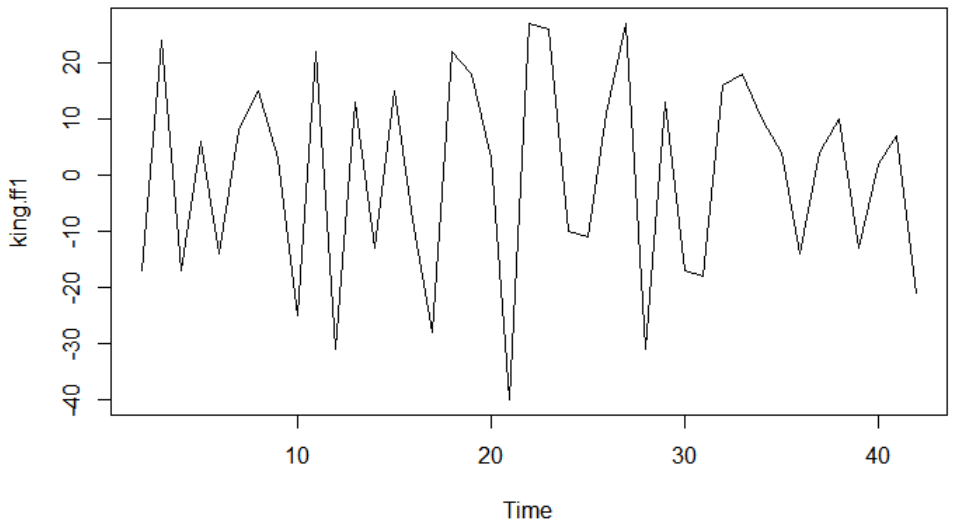

```
#8년마다 평균을 내서 그래프를 부드럽게 표현
king.sma8 <- SMA(king.ts, n=8)
plot.ts(king.sma8)
```



ARIMA 모델

- ARIMA모델은 정상성 시계열에 한해 사용한다
- 비정상 시계열 자료는 차분해 정상성으로 만족하는 조건의 시계열로 바꿔준다
- 분해시계열 그래프에서 평균이 시간에 따라 일정치 않은 모습을 보이므로 비정상 시계열이다
- 1차 차분 결과에서 평균과 분산이 시간에 따라 의존하지 않음을 확인한다
- ARIMA(p,1,q)모델이며 차분을 1번 해야 정상성을 만족한다

```
#ARIMA적용->1차 차분
king.ff1 <- diff(king.ts, difference=1)
plot.ts(king.ff1)
```

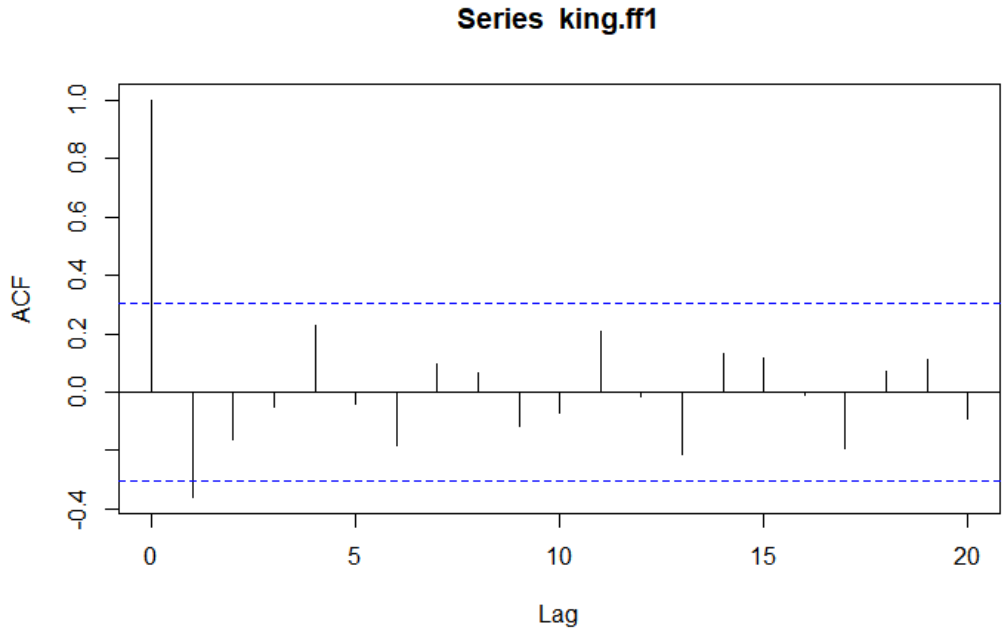


ACF와 PACF를 통한 적합한 ARIMA 모델 결정

① ACF

- lag는 0부터 값을 갖는데, 너무 많은 구간을 설정하면 그래프를 보고 판단하기 어렵다.
- ACF값이 lag 1인 지점 빼고는 모두 점선 구간 안에 있고, 나머지는 구간 안에 있다.

```
#acf를 통해 ARIMA모델 결정
acf(king.ff1, lag.max=20)
acf(king.ff1, lag.max=20, plot=false)
```

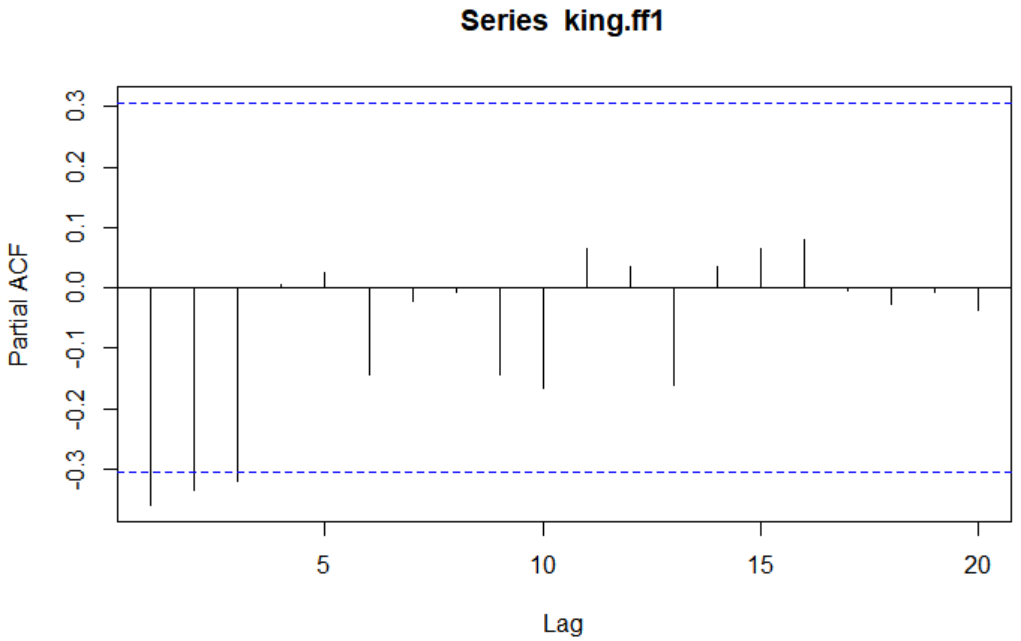


ACF와 PACF를 통한 적합한 ARIMA 모델 결정

② PACF

- PACF 값이 lag 1, 2, 3에서 점선 구간을 초과하고 음의 값을 가지며 절단점이 lag 4이다.

```
#PACF를 통해 lag 절단점 결정
pacf(king.ff1, lag.max=20)
pacf(king.ff1, lag.max=20, plot=FALSE)
```



종합

- 아래와 같이 ARMA 후보들이 생성

  - ARMA(3,0) 모델 : PACF값이 lag4에서 절단점을 가짐, AR(3)모형
  - ARMA(0,1) 모델 : ACF값이 lag2에서 절단점을 가짐, MA(1)모형
  - ARMA(p,q) 모델 : 그래서 AR모형과 MA을 혼합

```
> auto.arima(king)
Series: king
ARIMA(0,1,1)

Coefficients:
          ma1
      -0.7218
s.e.   0.1208

sigma^2 = 236.2:  log likelihood = -170.06
AIC=344.13   AICc=344.44   BIC=347.56
```

```
> #ARIMA 적용
> king.arima <- arima(king, order=c(0,1,1))
> king.forecasts <- forecast(king.arima)
> king.forecasts
   Point Forecast    Lo 80    Hi 80    Lo 95     Hi 95
43       67.75063 48.29647 87.20479 37.99806  97.50319
44       67.75063 47.55748 87.94377 36.86788  98.63338
45       67.75063 46.84460 88.65665 35.77762  99.72363
46       67.75063 46.15524 89.34601 34.72333 100.77792
47       67.75063 45.48722 90.01404 33.70168 101.79958
48       67.75063 44.83866 90.66260 32.70979 102.79146
49       67.75063 44.20796 91.29330 31.74523 103.75603
50       67.75063 43.59372 91.90753 30.80583 104.69543
51       67.75063 42.99472 92.50653 29.88974 105.61152
52       67.75063 42.40988 93.09138 28.99529 106.50596
```

Thank you.

시계열 자료 분석
ryp1662@gmail.com