클린업&리드오프 1주차

시계열 자료 분석

4팀

문근영
임하경 김나희
김다희 이원준
최가연 박혜상

0 시계열 팀의 클린업&리드오프 살펴보기

1주차 Time Series

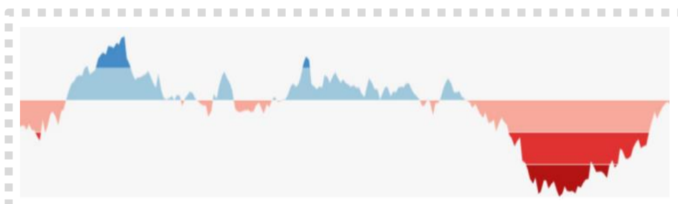
시계열 정의, 정상성,
시계열의 정상화

2주차 Time Series

ACF, AR, MA, ARMA

3주차 Time Series

단위근 검정, ARIMA, 모형 평가



INDEX

1

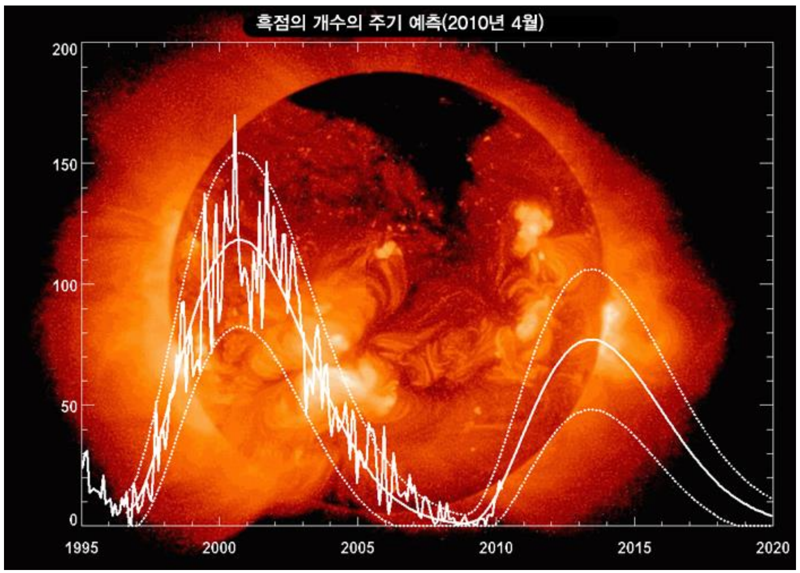
시계열이란??!

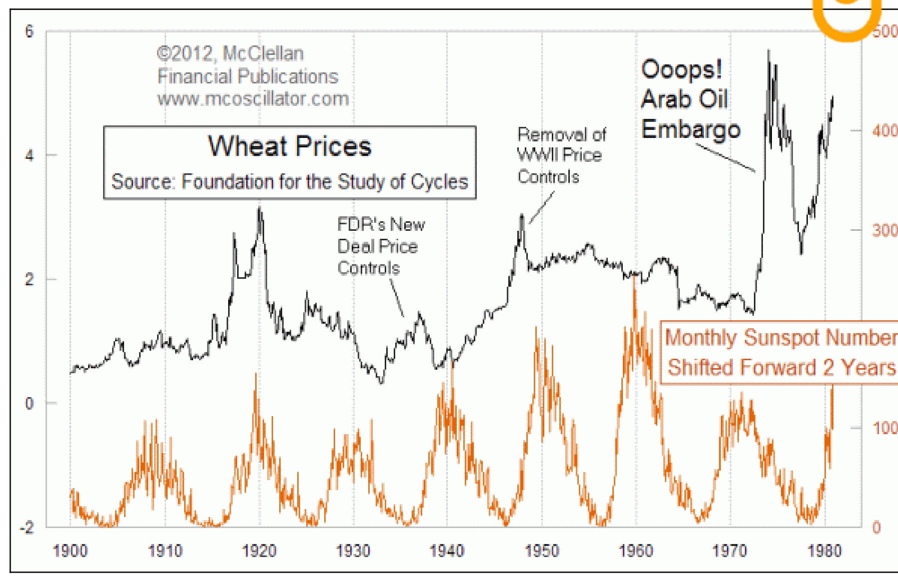
학습목표

시계열 분석의 정의와 목적을 알아보고,
회귀분석과 비교하여 그 특징을 이해한다

1 시계열 분석

시계열 분석의 역사





태양의 흑점 자료

밀 가격지수

1 시계열 자료의 정의

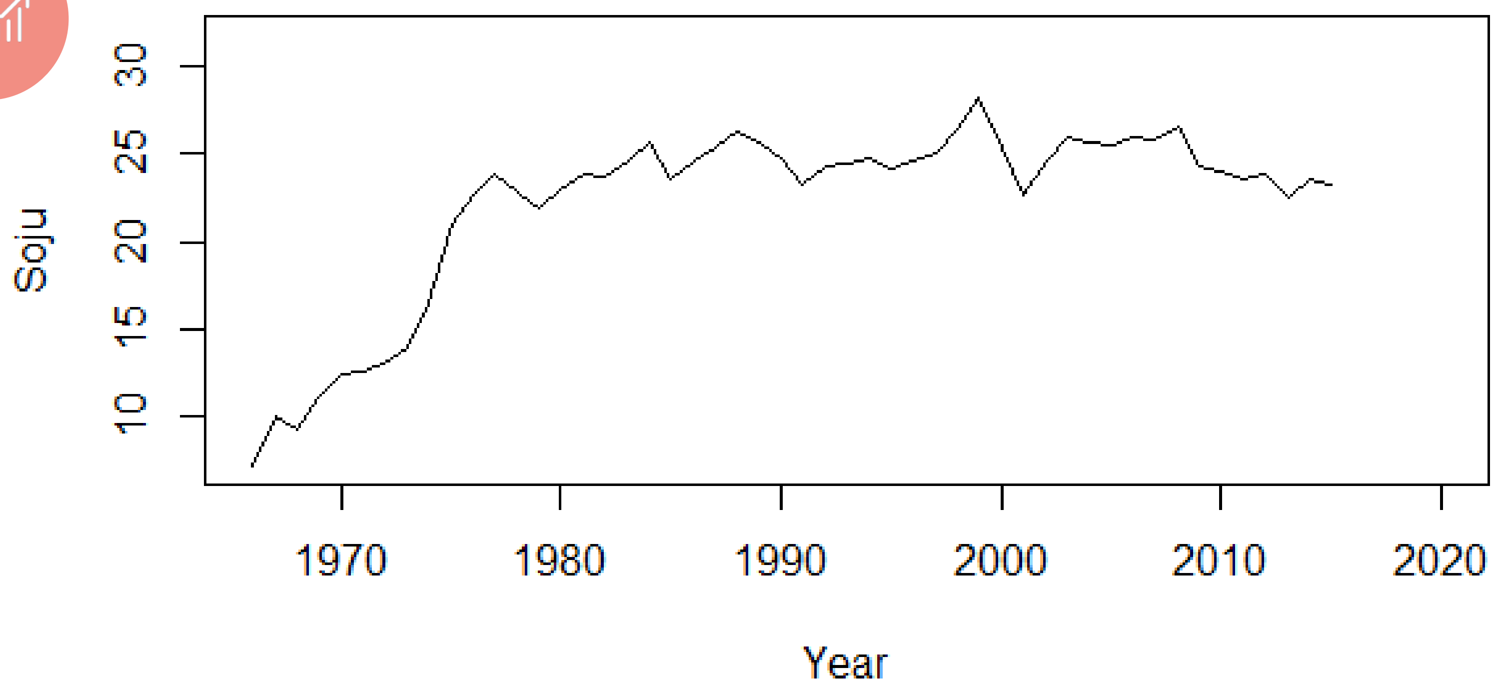
관측치 또는 통계량의 변화를 '시간의 흐름'에 따라서 포착한 자료

예시1) [한국일보] 주가 대세 상승기? "삼성전자 착시" VS "경제 선반영"

1 시계열 자료의 정의

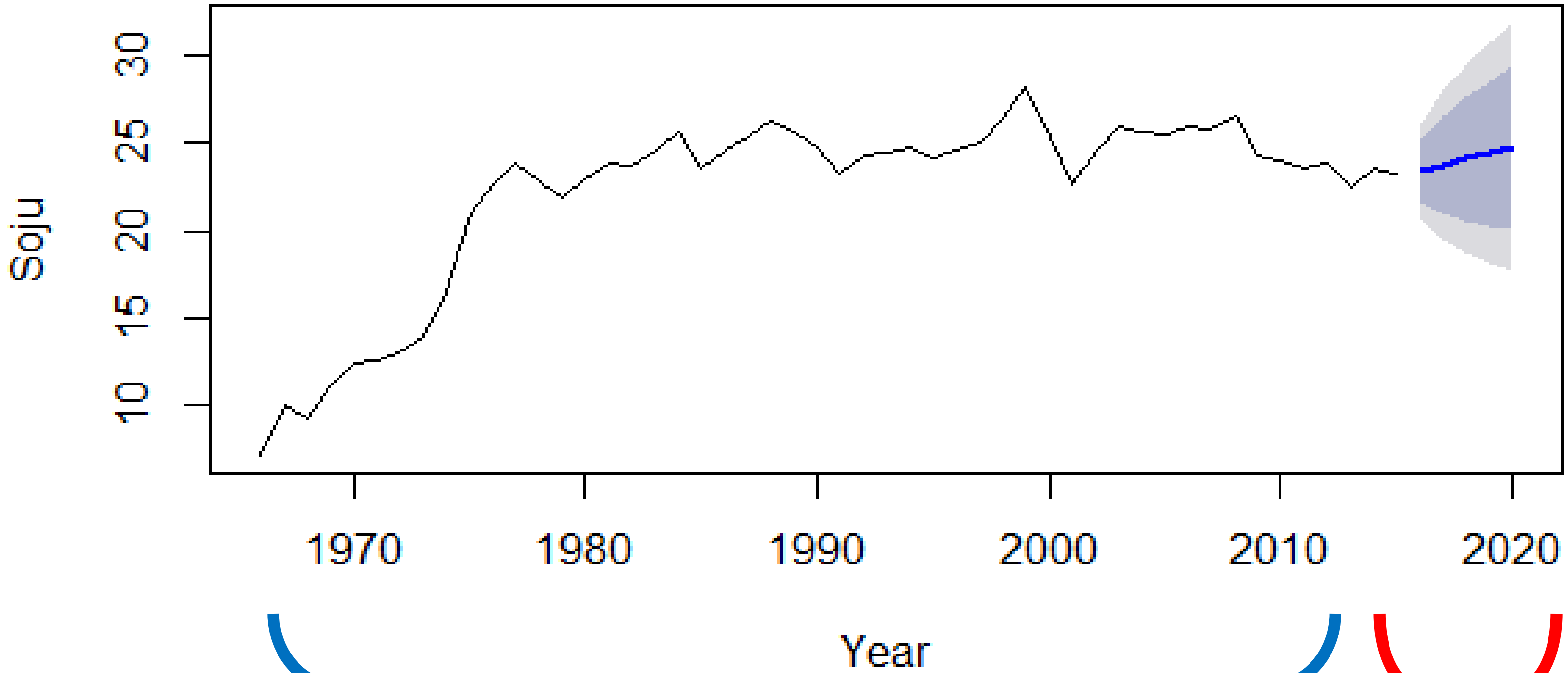
관측치 또는 통계량의 변화를 '시간의 흐름'에 따라서 포착한 자료

예시 2) 소주 판매량



1 시계열 분석의 목적

Forecasting Sales of Soju Using ARIMA



과거의 자료를 통해 → 미래의 자료를 예측

회귀 분석의 세 가지 가정

가정	가정 무너졌을 때
정규성	자료의 개수(n) 늘리기
등분산성	변수 변환
독립성	시계열로 간다

1 시계열과 회귀분석의 비교

회귀분석

$$\varepsilon_i \sim N(0, \sigma^2)$$

- 종속변수와 독립변수의 관계에서 모델 도출
- 에러가 정규분포를 따르고 서로 독립

시계열

$$\varepsilon_i \nsim \mathrm{iid}$$

- 변수 자체의 시간의 흐름에 따른 특성에서 모델 도출
- 에러가 서로 독립이 아니고 상호의존적

2

정상성

학습목표

시계열 분석에서 중요한 가정인 정상성의 정의를 알아보고,
IID와 WN에 대해 배워보자

2 정상성

시계열을 예측하는 원리?!

오늘의 나와 내일의 나는 게으름의 정도가 같다

2 정상성

- 과거에 있었던 패턴이 지금도 비슷할 것!
- 수십~수백 개의 지점에 서의 확률을 전부 계산하는 것은 심각하게 비현실적

정상성
(Stationarity)

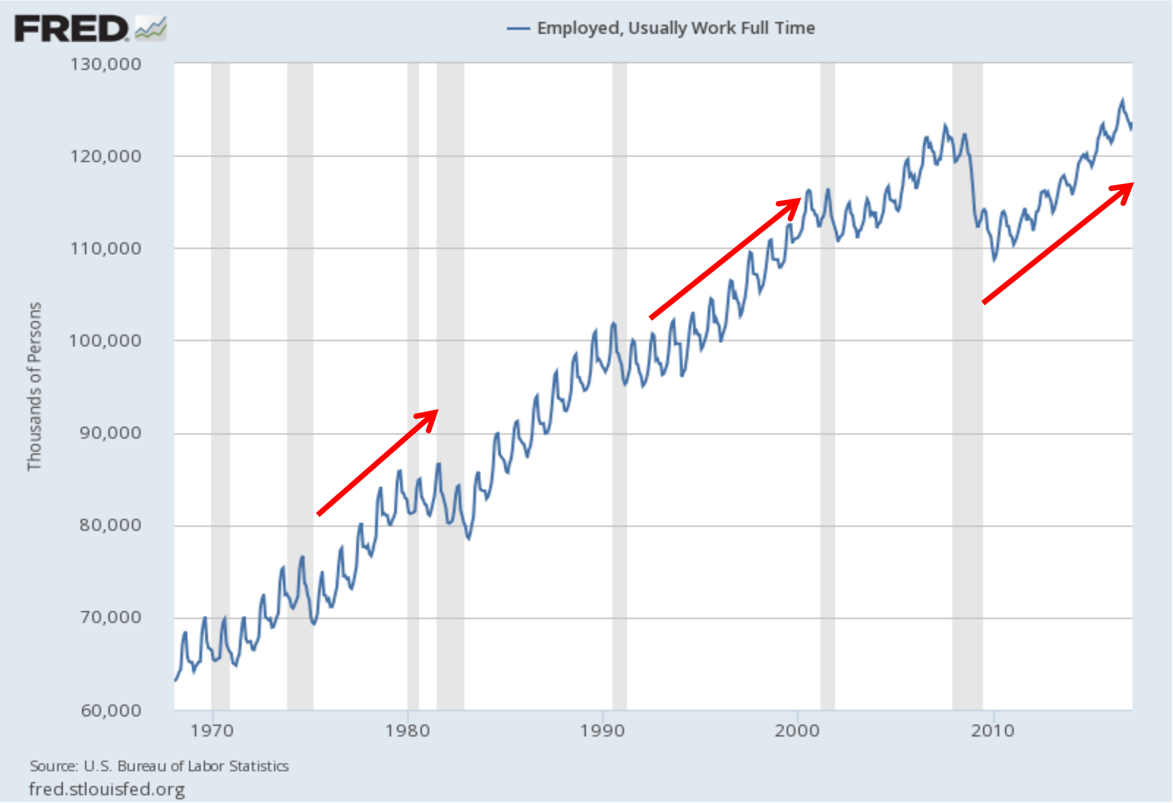
정상성이란?

시계열 자료의 변동이 과거와 미래가 큰 차이가 없다는 가정

시간 경과(시차)에 따라 거의 규칙적으로 변동할 것이라는 가정

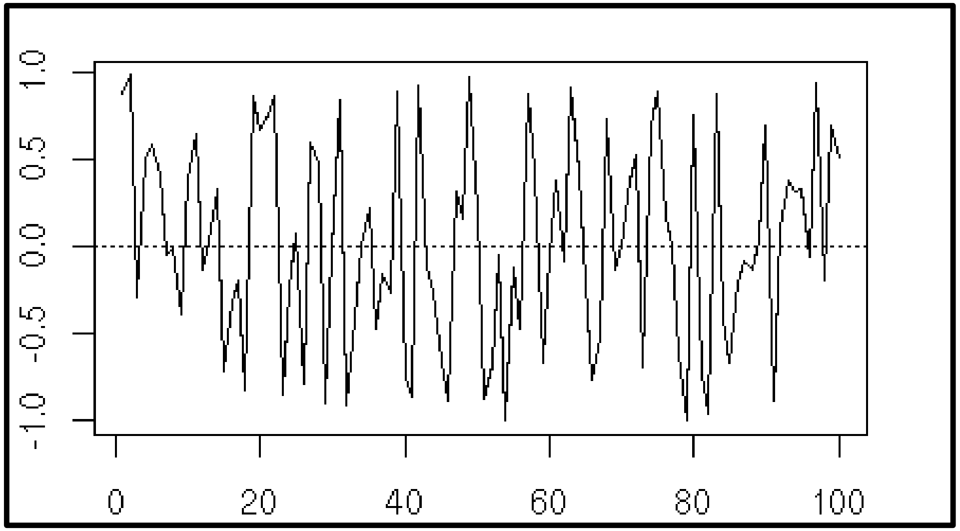
즉,
미래 자료를 예측할 수 있다!

정상성 예시



자료의 변동 특징이 같다!

정상 시계열이란?



✓ 시점이 변하더라도 시차에 따라 확률분포가 일정하다

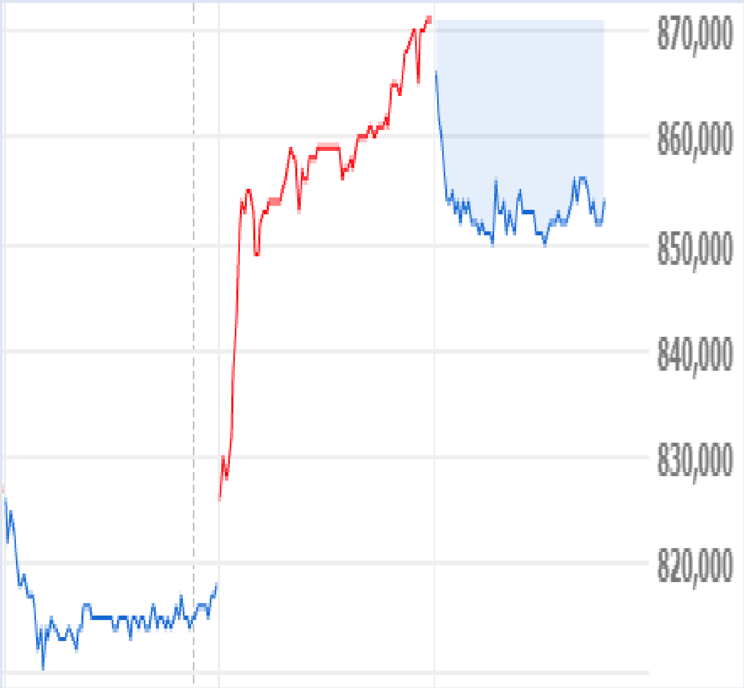
✓ 경제 성장률, White Noise

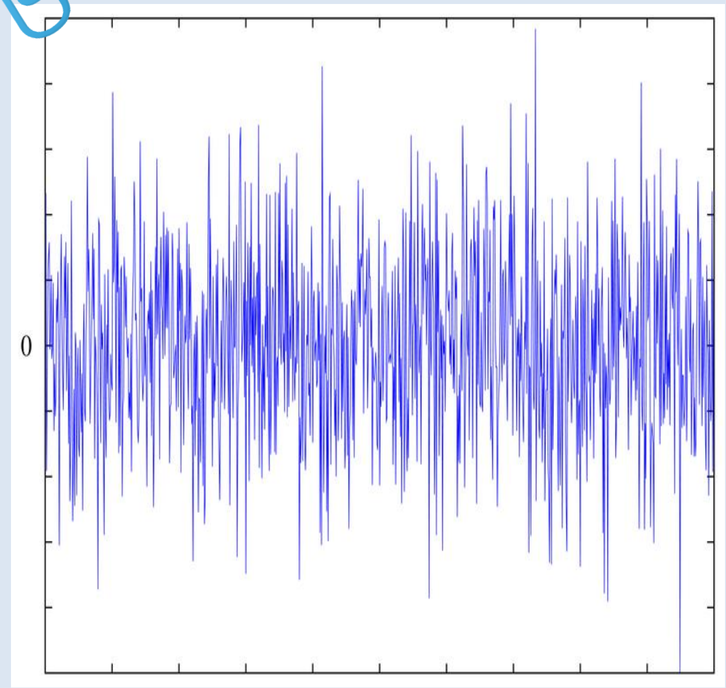
과거의 변동이 미래에도 이어지므로
과거 데이터를 바탕으로 미래의 결과를 예측할 수 있다.

2 정상성

시계열 예시

정상 시계열





정상시계열이 비정상시계열에 비해 모델을 통한
미래 예측이 쉽다!

정상 시계열의 성질

$$E(Y_t) = \mu$$

$$Var(Y_t) = \sigma^2$$

$$Cov(Y_t, Y_{t+h}) = \gamma_h$$

일정한 평균

분산 값 존재

공분산은 시점이 아니라 시차에 따라 정해진다.

정상성

강정상성

약정상성

2 정상성: 강정상성

강정상성 : 일정 시차의 두 시계열의 Joint Distribution이 동일하다는 가정

IID

(Independent, Identically Distributed)

$$E(Y_t) = 0$$

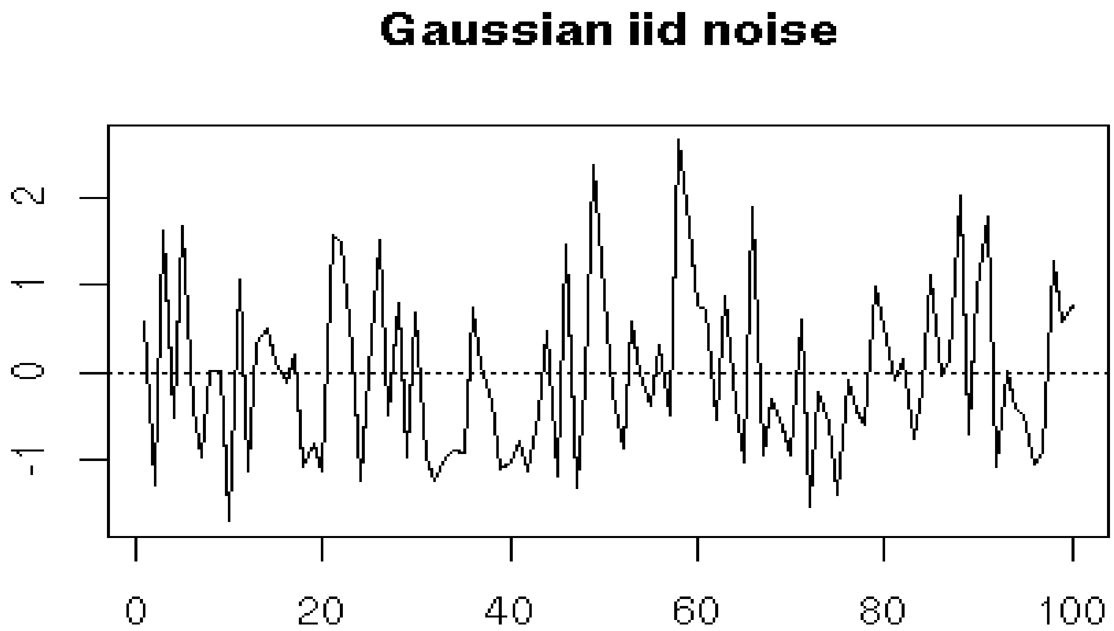
$$E\left(Y_t^2\right) = \sigma^2$$

$$Cov(Y_t, Y_{t+h}) = 0$$ (시점이 서로 다를 때)

$$F(X_t, X_{t+1}, \cdots X_{t+s}) = F(X_{t+h}, X_{t+1+h}, \cdots X_{t+s+h}) \text{ for } \forall\, t, s, h$$

시점에 상관 없이 같은 시차(lag=h)에서 같은 분포가 반복된다

강정상성 (Strict Stationarity)



과거 추세 및 변동을 바탕으로 미래를 예측할 수 있지만
시차마다 모든 분포가 같기는 매우 힘들다
→ 비현실적

약정상성 (Weakly Stationarity)

: 시점에 상관없이 평균과 분산이 일정하다
(시계열의 기본 가정!)

조건

$E(X_t) = \mu, \ \forall t \in Z$: 평균이 모든 점에서 같다

$E(X_t^2) < \infty, \forall t \in Z$: 분산이 같은 값으로 존재

$Cov(X_t, X_{t+h}) = Cov(X_0, X_h) \text{ for } \forall t, h$

강정상성 (Strict Stationarity) $\rightarrow$ 약정상성 (Weak Stationarity)

약정상성 (Weak Stationarity) $\not\rightarrow$ 강정상성 (Strict Stationarity)

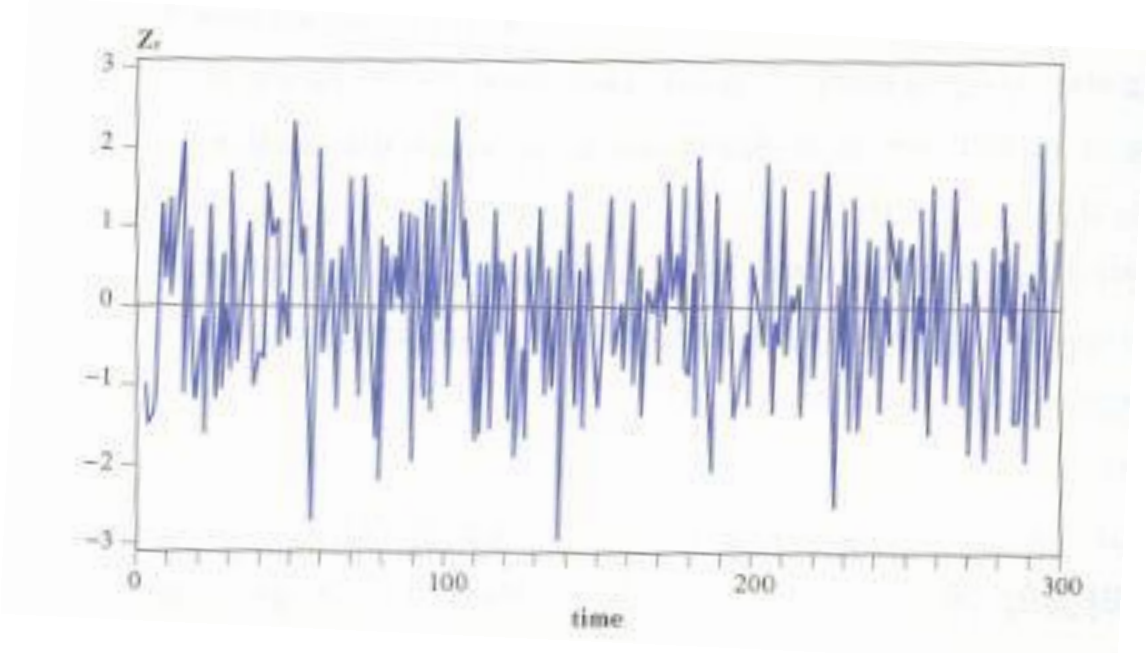
2 정상성: 약정상성

약정상성을 띄는 시계열 – 백색소음 (White Noise)

- 평균이 0
- 분산이 일정
- 서로 다른 지점에서의 공분산이 0

$$X_t \sim WN(0, \sigma^2)$$

$$X_t = \varepsilon_t,\ t = 1, 2, \ldots$$



IID Process & White Noise

시계열	IID Process	White Noise
특징	$\boldsymbol{X_1, X_2}, \cdots, \boldsymbol{X_n} \sim iid,$ $E(X_t) = 0, \quad E\left(X_t{}^2\right) = \sigma^2$ $Cov\left(X_{t,} X_s\right) = 0 \ (t \neq s)$	$E(X_t) = 0, \quad E\left(X_t{}^2\right) = \sigma^2$ $Cov\left(X_{t,} X_s\right) = 0 \ (t \neq s)$
정상성	강정상성/ 약정상성 만족	약정상성 만족, 강정상성 불만족

강정상성

(Strict stationary time series)

$[f(X_t, X_{t+1}, \cdots, X_{t+s}) =$

$f(X_{t+h}, X_{t+1+h}, \cdots, X_{t+s+h})$

for ∀t, s, h

어느 시점을 잡아도 시점 간
Joint Distribution이 항상 같다

약정상성

(Weak stationary time series)

$E(X_t) = \mu$

$Var(X_t) < \infty$

$cov(X_t, X_{t+h}) = cov(X_0, X_h)$

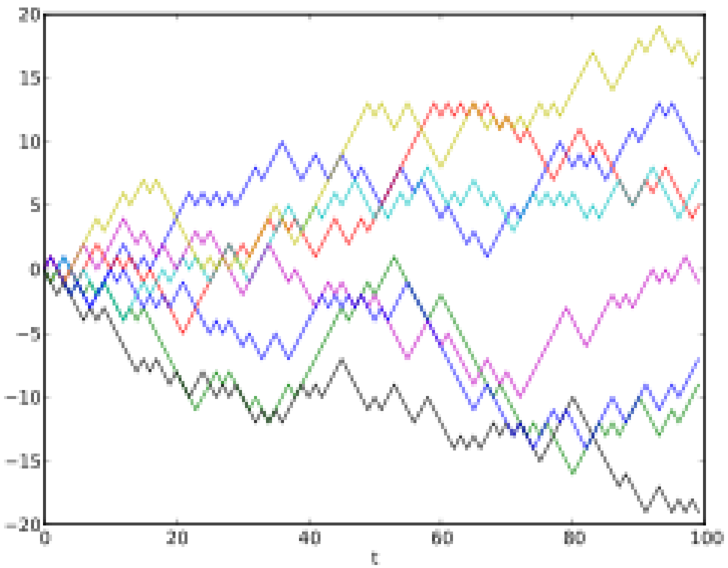
for ∀ t, h

어떤 시점에서도 평균 같다
분산 일정
같은 시차의 공분산이 같다

2 정상성: Random Walk

정상성을 만족하지 않는 시계열 – 확률보행 (Random Walk)

예시) 주가, 액체 · 기체 내 분자의 움직임



$$X_t = X_{t-1} + \varepsilon_t$$

$$\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

$$X_1 = \varepsilon_1$$
$$X_2 = \varepsilon_1 + \varepsilon_2$$
$$X_3 = \varepsilon_1 + \varepsilon_2 + \varepsilon_3$$
$$\cdots$$
$$X_t = \sum_{i=1}^{t} \varepsilon_i \quad [\varepsilon_i \sim iid\ N(0, \sigma^2),\ i=1, 2, \cdots]$$

$$E(X_t) = 0,\ V(X_t) = t\sigma^2$$

분산이 t에 따라 달라지기 때문에 정상성 만족X

위의 그래프처럼 다음 시점에 어떻게 될지 예측 불가능

시계열 $F_x(X_1\ , X_2\ ,\cdots X_n\ )$ 정상화의 필요성

$\sigma_1^2, \sigma_2^2,\cdots \sigma_n^2$ $\mu_1, \mu_2\ ,\cdots \mu_n$

- 평균, 분산 및 자기상관계수들과 시계열 모형의 계수까지 포함한다면 수없이 많은 모수를 추정해야 한다
- n개 시점으로 , 모수를 추정할 경우 자료에 비해 추정치가 많아 추정한다고 해도 신뢰성을 기대할 수 없다.
- 추정 해야할 모수의 개수를 대폭 줄일 수 있다.

3

시계열의 정상화

-직접 해보기-

Classical decomposition과 차분을 중심으로

학습목표

시계열 자료에서 추세와 계절성을 제거하여
정상화된 오차를 구해보자

시계열을 정상화 하는 보편적인 2가지 방법

CLASSICAL DECOMPOSE

자료가 trend, 계절성을 가지는 경우

$$\underbrace{X_t}_{} = \underbrace{M_t}_{\text{trend}} + \underbrace{S_t}_{\text{계절성}} + \underbrace{Y_t}_{\text{오차}}$$

차분

자료가 trend를 가지는 경우

$$X_t = \underbrace{M_t}_{\text{trend}} + \underbrace{Y_t}_{\text{오차}}$$

Classical Decomposition :

Trend와 계절성(seasonality)를 동시에 제거하고 정상성을 가진 오차만을 추출하는 방법

$$X_t = \underbrace{M_t}_{\text{trend}} + \underbrace{S_t}_{\text{계절성}} + \underbrace{Y_t}_{\text{오차}}$$

1. MA filter로 trend 추정
2. 추정된 trend 제거 후 seasonal averaging
3. 단위근 검정을 통해 정상성을 판단한다.

3 시계열의 정상화

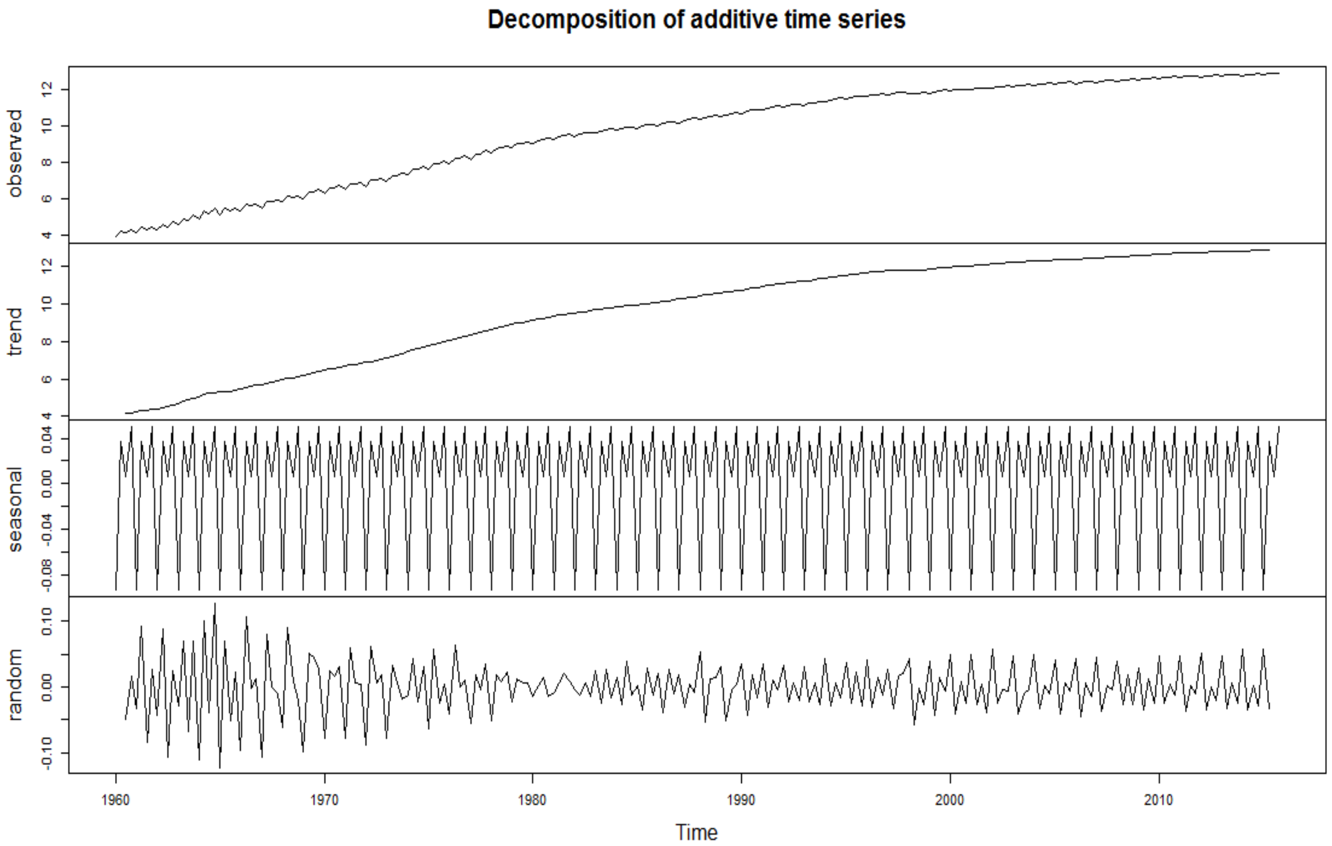
R에서 decompose을 사용한 CLASSICAL DECOMPOSITION

데이터

trend

계절성

오차



3 시계열의 정상화; Classical decomposition

Classical Decomposition :

1. MA Filter

- 일정 간격으로 지점을 잡은 후,
  지점 주변의 평균으로 빼 준다.

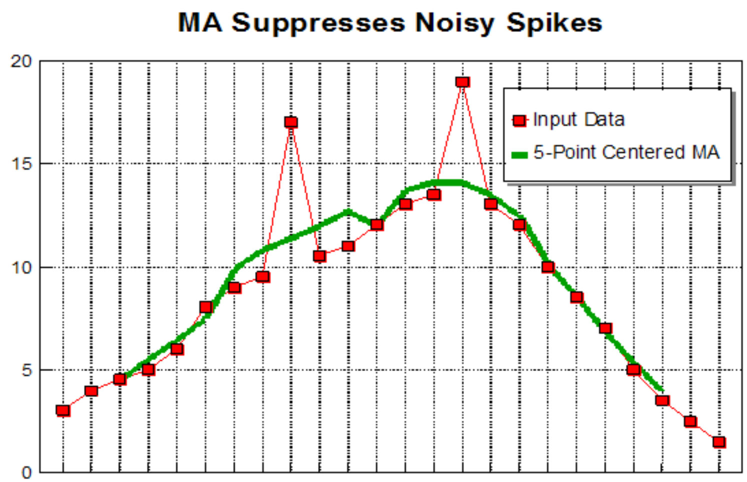


Figure 1

5개 지점을 예시로 잡으면,

X_1부터 X_5까지 지점의 값은 $\frac{X_1+X_2+X_3+X_4+X_5}{5}$로

X_6부터 X_{10}까지 지점의 값은 $\frac{X_6+X_7+X_8+X_9+X_{10}}{5}$로

각각 빼주는 것을 반복한다.

3 시계열의 정상화; Classical decomposition

- 필터의 길이는 몇 개의 지점이 시계열의 한 주기인지에 따라 다르다.

1주일이 1주기라면 (7지점 1주기), 7개 지점의 평균을 정한다.

$$\frac{X_1+X_2+X_3+X_4+X_5+X_6+X_7}{7}, \frac{X_8+X_9+X_{10}+X_{11}+X_{12}+X_{13}+X_{14}}{7}$$ 등으로 빼준다.

분기별 데이터라 4지점 1주기 등이라면, 홀수로 맞추기 위해
양끝에 0.5를 곱해서 필터를 결정한다.

$$\frac{0.5X_1+X_2+X_3+X_4+0.5X_5}{4}, \frac{0.5X_5+X_6+X_7+X_8+0.5X_9}{4}$$ 등으로 빼준다.

2. Seasonal Average Estimation

- MA Filter를 통해 추세를 제거한 이후에는, **계절성**을 추정한다.

→ 어떤 분기별 시계열 자료의 추세 제거 이후의 값이
다음과 같다고 가정한다.

t	1	2	3	4	5	6	7	8	9	10	11	12
Y_t'	-12	4	14	-6	-19	-4	21	2	-17	0	21	-4

2. Seasonal Average Estimation

- 각 분기의 평균값(S_t)을 추세 제거된 시계열(Y_t')의 값에서 빼 준다.

1분기의 경우, 1분기의 값의 평균인 (−12−19−17)/3=−16을,
2분기의 경우, 2분기의 값의 평균인 (4−4−0)=0 등을 빼준다.

t	1	2	3	4	5	6	7	8	9	10	11	12
Y_t'	−12	4	15	−7	−19	−4	21	2	−17	0	21	−4
S_t	−16	0	19	−3	−16	0	19	−3	−16	0	19	−3
ε_t	4	4	4	−4	−3	−4	2	5	−1	0	1	−1

3. 단위근 검정 (Unit Root Test)

- 빼 주고 남은 오차 ε_t에 대해 단위근 검정을 해서 이 오차가 정상성을 만족하는지 확인한다.

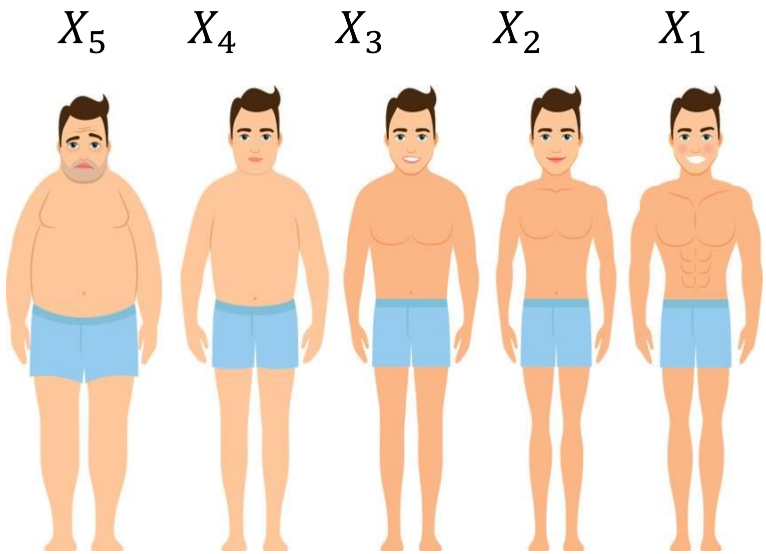
단위근 검정은 Augmented Dickey-Fuller Test를 이용하며, 정상성을 만족하지 못한다는 귀무가설로 가설검정을 한다.

자세한 것은 3주차 참조!

t	1	2	3	4	5	6	7	8	9	10	11	12
ε_t	4	4	4	-4	-3	-4	2	5	-1	0	1	-1

차분:

$$\nabla X = X_t - X_{t-1}$$



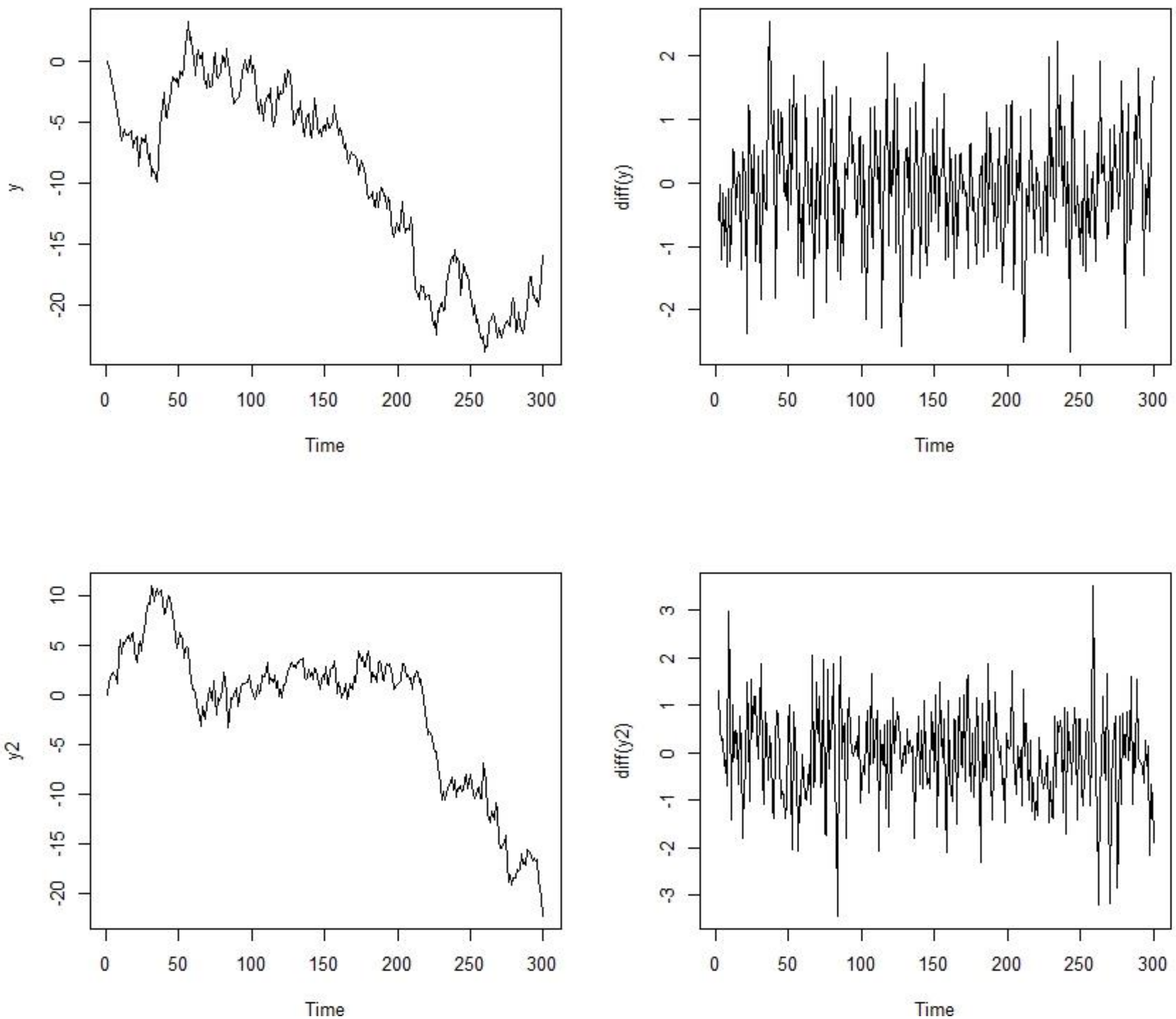
$$\nabla m_t = (c_0 + c_1 t) - \left(c_0 + c_1(t-1)\right) = c_1$$

$$\nabla^k X_t = k!\, c_k + \nabla^k Y_t$$

∇X가 일정한 값으로 얻어진다면
$\nabla X = M_t$
선형의 TREND를 얻는다고 할 수 있다.

이 선형들을 차분으로 제거하고 오차를 구할 수 있다.

R에서 명령어 Diff 을 사용한 차분 정상화



3 시계열의 정상화; 차분

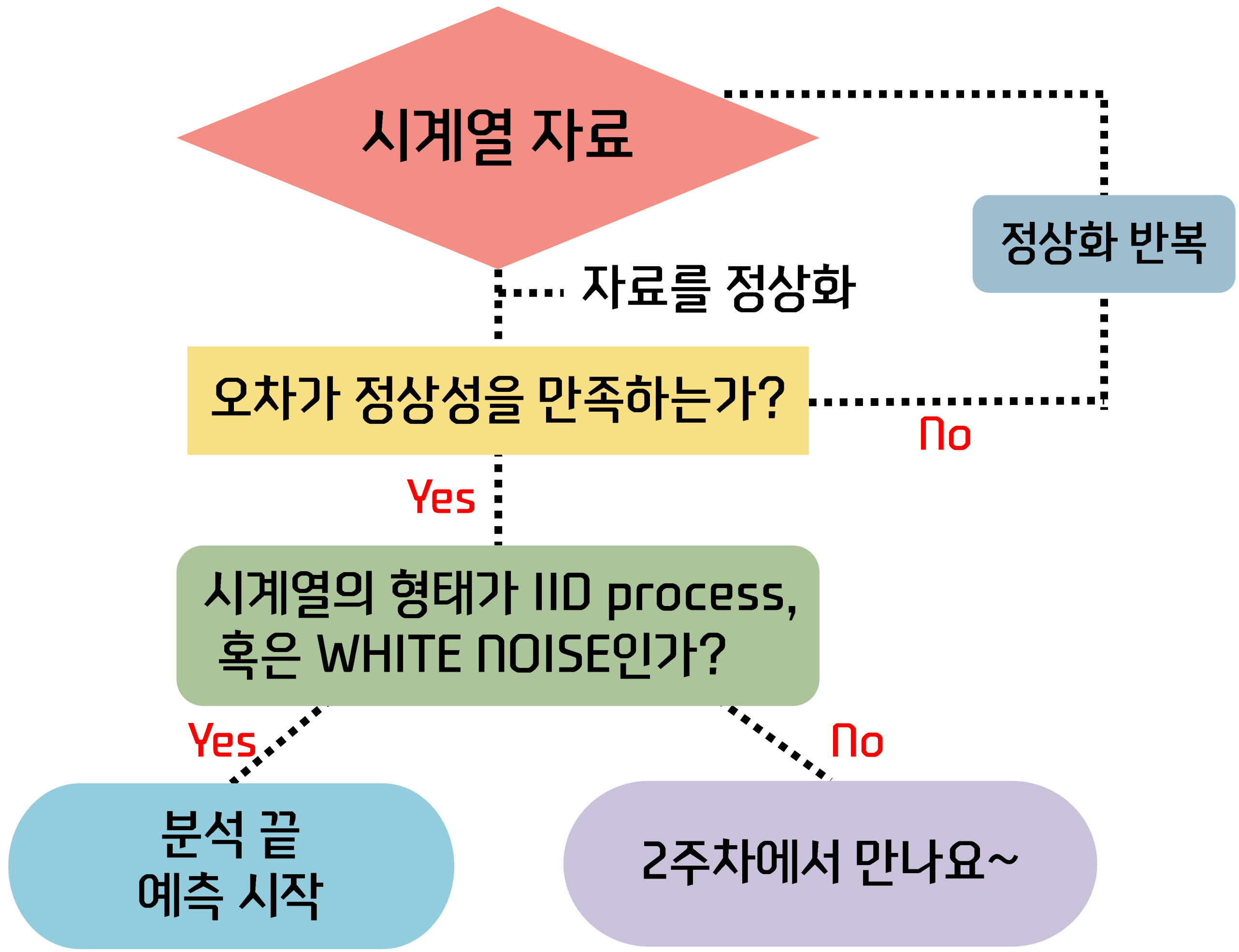
차분의 응용

일반적으로 $\nabla X = X_t - X_{t-1} = \varepsilon_t$ 의 형태로도 충분하나

$\nabla^2 X_t = \nabla X_t - \nabla X_{t-1}$형식으로
$\nabla^k X_t = \nabla^{k-1} X_t - \nabla^{k-1} X_{t-1}$차분을 k 번하는 경우도 있다.

단! 지나친 차분은 정확한 예측을 할 수 없게 만드니 주의!

3 시계열의 정상화



to be continued...

Appendix

Methods of Decomposition

R에서 Decompose를 할 수 있는 명령어는 더 존재한다.

decompose()

MA filter를 사용한
Classical Decomposition

모수가 간단한 편이고
추가적인 처리가 필요없다.

stl()

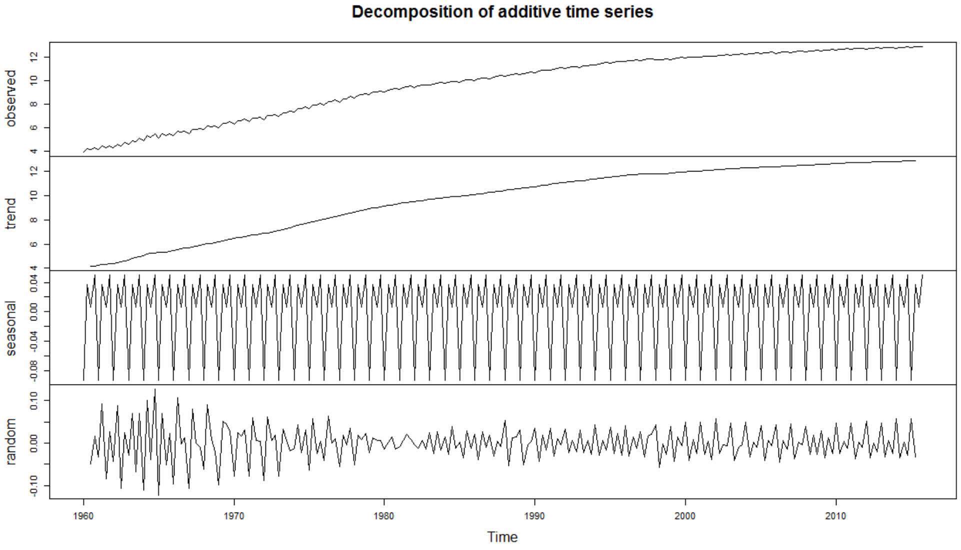
Loess Regression을 사용한
새로운 Decomposition

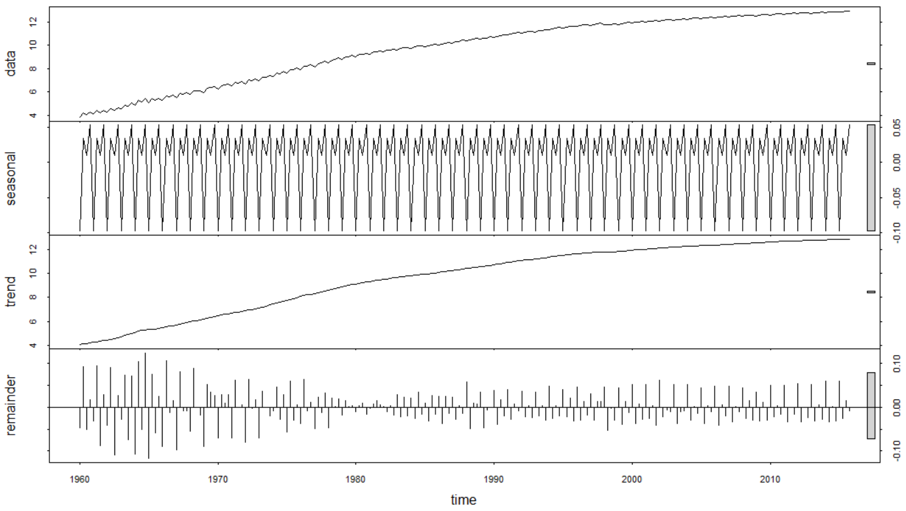
Robust한 추정법이라
이상치에 강한 편이지만
결과가 행렬로 나오므로
추가적인 처리가 필요하다.

Methods of Decomposition

decompose()

stl()





Methods of Decomposition

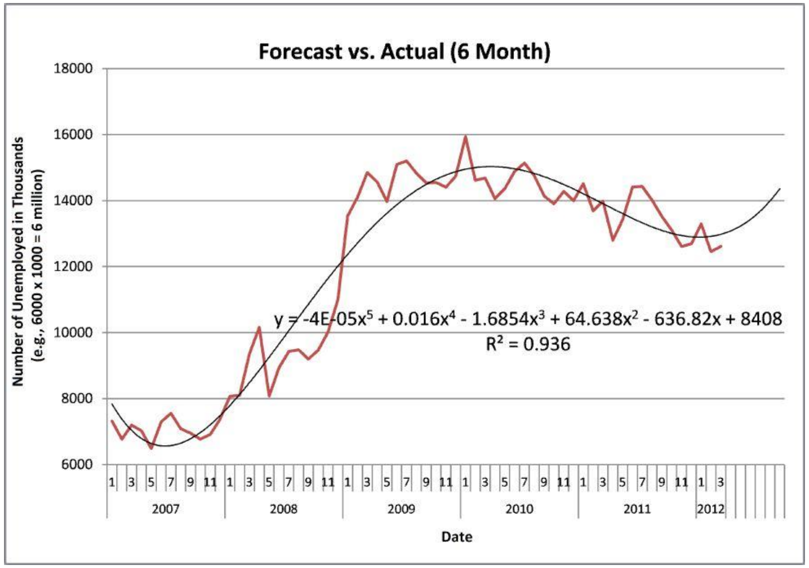
계절성이 없는 자료에서 추세를 직접 알고 싶다면?

Polynomial Regression을 사용하면 된다.

$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_p X^p$의 관계식을 OLS로 추정해서 계수를 통해 추세를 추정한다.

차수가 낮으면
추세를 제대로 알 수 없으며,

차수가 높으면
과적합의 문제가 발생할 수 있다.



R Code

R code (1st week) - 계절성이 있는 자료

```
library(tseries)
GDP=read.csv("GDP.csv")
GDPts=ts(data=GDP$GDP,start=1960,deltat=1/4)
#분기별 자료이니 deltat=1/4. 월별일 경우 1/12

logGDPts=log(GDPts)
GDPcomp=decompose(logGDPts) #시계열 분해
plot(GDPcomp) #분해 결과 확인 가능
plot(GDPcomp$random) #분해 결과 중 오차 표시
random=na.omit(GDPcomp$random)
adf.test(random) #단위근 검정 결과 귀무가설 기각 / 정상화 완료
```

R Code

R code (1st week) - 계절성이 없는 자료

```
library(tseries)
GDP2=read.csv("year.csv")
GDPts2=ts(data=GDP$GDP,start=1953)
#분기별 자료이니 deltat는 사용 불가
logGDPts2=log(GDPts)
#decompose()는 계절성이 없으므로 사용 불가
adf.test(logGDPts2) #단위근 검정 - 정상성을 만족하지 않는다.
dlogGDPts2=diff(logGDPts2) #1회 차분을 해본다.
adf.test(dlogGDPts2) #차분한 시계열의 단위근 검정
#귀무가설을 기각하므로, 시계열이 정상화되었다.
```

☆ R Code

R code (1st week) - 계절성이 없는 자료

```
library(forecast)
ndiffs(logGDPts2) #정상화에 필요한 차분의 횟수를 알 수 있다.
#이 자료는 차분이 2회 필요하다고 나오지만, 판단은 주관적이다.

number=seq(1,63)
GDP2=cbind(number,GDP2)
polylm=lm(log(GDP)~poly(number,3,raw=T),GDP2)
polylm
#Polynomial Regression을 통해 추세를 예측할 수 있다.
#poly() 안의 든 숫자로 차수를 조정할 수 있다
```

☆ Recommended Web Site

시계열 분석을 더 깊이 공부하기 위한 Web Site

https://www.otexts.org/fpp
[Forecasting: principles and practice] – E-book
R을 이용한 회귀분석, 시계열분석 등 수록

http://ecos.bok.or.kr
한국은행 경제통계시스템
우리 나라 경제 관련된 시계열 자료

http://kosis.kr
통계청
연도별, 분기별 자료 등 다양한 시계열 자료

THANK YOU