시계열자료분석팀

5팀

김민주

박준영

곽동길

강서진

황호성

INDEX

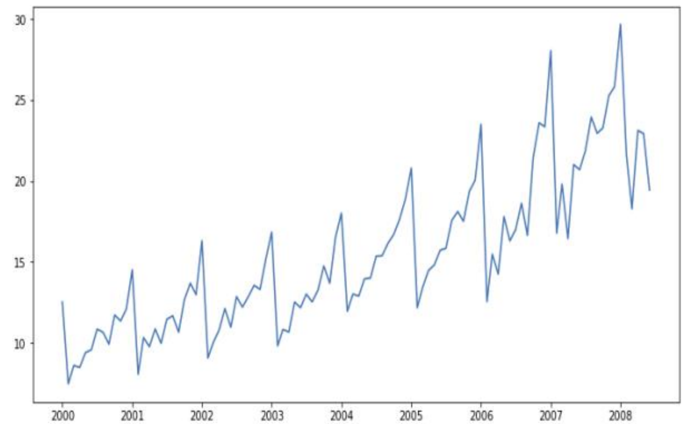
1

시계열 자료 분석

시계열 자료의 정의

시계열 자료 Time series Data

시간 순서에 따라 관측된 자료의 집합



$$\{X_t, t = 1,2,3, \dots\}$$

t : 시점, t의 종류에 따라 연속형, 이산형 자료로 분류됨

시계열 자료의 특징

관측치 간

연관성(dependency)

→

관측치 집합을 고려한

결합분포(joint distribution)

시계열 자료의 특징

관측치 간 연관성(dependency)

관측치 집합을 고려한 결합분포(joint distribution)

지금까지 흔히 다뤄왔던 자료들과 달리

독립성을 만족하지 않기 때문에

선형회귀와 같은 일반적인 방법 사용 X

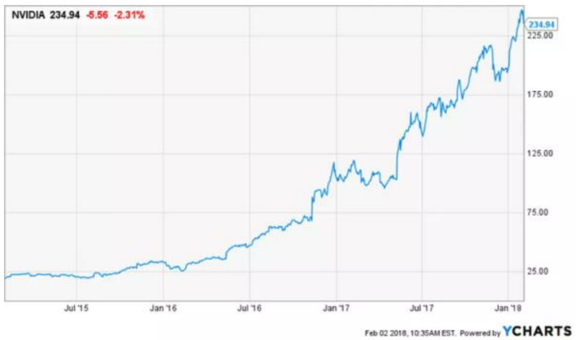
⋮

데이터의 특성을 반영할 수 있는 **시계열 분석**이 필요

시계열 자료의 구성요소 | (1) 규칙요소

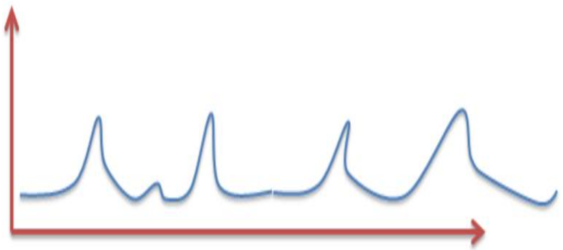
추세 변동 (Trend)

- 시간의 흐름에 따라 증가하거나 감소하는 추세를 갖는 변동
- 특별한 충격이 없는 한 지속



순환 변동 (Cycle)

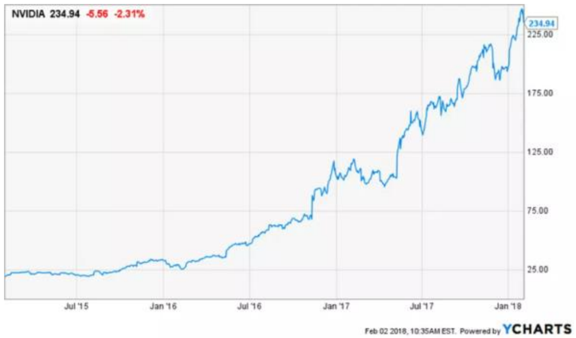
- 일정한 주기를 가지지만 규칙적으로 발생하지 않는 변동
- 경제, 사회적 요인 같은 외부요인으로 발생해 예측이 어려움



시계열 자료의 구성요소 | (1) 규칙요소

추세 변동 (Trend)

- 시간의 흐름에 따라 증가하거나 감소하는 추세를 갖는 변동
- 특별한 충격이 없는 한 지속



순환 변동 (Cycle)

- 일정한 주기를 가지지만 규칙적으로 발생하지 않는 변동
- 경제, 사회적 요인 같은 외부요인으로 발생해 예측이 어려움

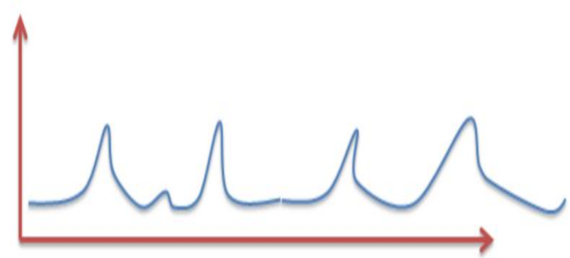
시계열 자료의 구성요소 | (1) 규칙요소

추세 변동 (Trend)

- 시간의 흐름에 따라 증가하거나 감소하는 추세를 갖는 변동
- 특별한 충격이 없는 한 지속

순환 변동 (Cycle)

- 일정한 주기를 가지지만 규칙적으로 발생하지 않는 변동
- 경제, 사회적 요인 같은 외부요인으로 발생해 예측이 어려움



1 시계열 자료 분석

시계열 자료의 구성요소 | (1) 규칙요소

계절 변동 (Seasonal Variaition)

- 규칙적인 주기를 가지고 발생하는 변동
- 주별, 월별, 계절별 등의 시간 간격을 가지고 반복
- 환경적인 요인에 발생하기 때문에 예측 및 처리에 용이함

시계열 자료의 구성요소 | (2) 불규칙요소

우연 변동 (Random Fluctuation)

- 무작위한 원인에 의해 나타나 규칙성을 인지할 수 없는 변동
- 불규칙 성분이라고 불리기도 함

1 시계열 자료 분석

시계열 분해(Times Series Decomposition)

시계열 자료를 비정상 부분(non-stationary part)과
정상 부분(stationary part) 으로 분해하는 과정

추세(m_t)와 계절성 (s_t)은 비정상 부분, 오차(Y_t) 는 정상 부분

덧셈 분해

$$X_t = m_t + s_t + Y_t$$

곱셈 분해

$$X_t = m_t * s_t * Y_t$$

시계열 분해(Times Series Decomposition)

시계열 자료를 비정상 부분(non-stationary part)과
정상 부분(stationary part) 으로 분해하는 과정

추세(m_t)와 계절성 (s_t)은 비정상 부분, 오차(Y_t) 는 정상 부분

데이터에 0이 포함되는지
반드시 확인해야 함
만약 0이 존재한다면 곱셈 분해 사용 X

곱셈 분해

$$X_t = m_t * s_t * Y_t$$

시계열 분해(Times Series Decomposition)

시계열 자료를 비정상 부분(non-stationary part)과
정상 부분(stationary part) 으로 분해하는 과정

추세(m_t)와 계절성 (s_t)은 비정상 부분, 오차(Y_t) 는 정상 부분

덧셈 분해

$$X_t = m_t + s_t + Y_t$$

$$X_t - m_t - s_t = Y_t$$

데이터에 0이 포함되는지 반드시 확인해야 함.

만약 0이 존재한다면 곱셈 분해 사용 X

추세와 계절성을 제거한 후 남은 오차를 이용해서 예측 모델링을 진행!

시계열 분해(Times Series Decomposition)

시계열 자료를 비정상 부분(non-stationary part)과
정상 부분(stationary part) 으로 분해하는 과정

추세(m_t)와 계절성 (s_t)은 비정상 부분, 오차(Y_t) 는 정상 부분

덧셈 분해

$$X_t = m_t + s_t + Y_t$$

곱셈 분해

이번 클린업에서는
덧셈 분해를 주로 다뤄볼 예정!

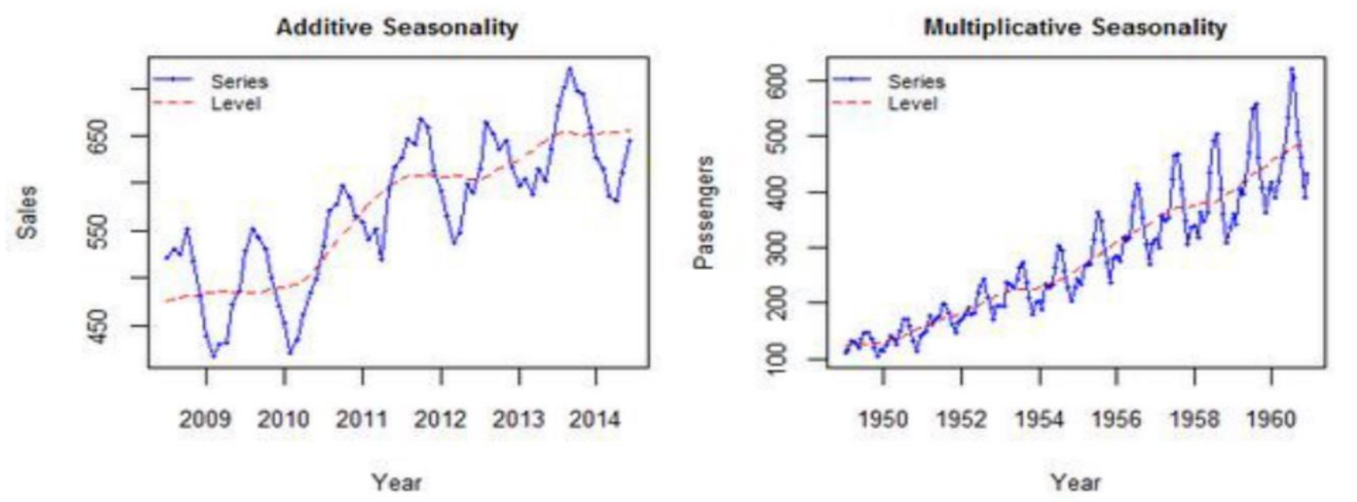
데이터에 0이 포함되는지 반드시 확인해야 함.
만약 0이 존재한다면 곱셈 분해 사용 X

시계열 분해(Times Series Decomposition)

덧셈 분해 VS 곱셈 분해

추세와 계절성이 서로 영향을 미치지 않는다 → 덧셈 분해

추세에 따라 계절성이 변화한다 → 곱셈 분해



시계열 분해(Times Series Decomposition)

시계열 모형은 정상성을 가정하는데,

시간에 따라 변동폭이 일정하지 않으면

정상성을 만족하지 못하게 됨

⋮

이런 경우, **곱셈 분해식에 로그를 취해**

덧셈 분해로 나타낸 후 계산

2

정상성

정상성의 정의

정상성 Stationarity

시계열 자료의 확률적 성질이

시점에 의존하지 않고 시차에만 의존하는 특성

시간의 흐름에 따라 평균, 분산이 변하지 않아 결합분포를 쉽게 구할 수 있음

강정상성

(Strict Stationarity)

약정상성

(Weak Stationarity)

2 정상성

강정상성

강정상성 Strict Stationarity

일정한 시차 간격(h)을 가지는 관측치들이 모두 같은 분포를 따른다는 특성

⋮

$$(X_{t_1}, \ldots, X_{t_n}) \overset{d}{=} (X_{t_1+h}, \ldots, X_{t_n+h}), \ \ \text{h = lag}, \forall n \ \geq 0$$

2 정상성

강정상성

강정상성 Strict Stationarity

일정한 시차 간격(h)을 가지는 관측치들이 모든 같은 분포를 따른다는 특성

But, 이렇게 엄격한 조건을 만족하는 데이터는 많지 않음

→ 조건을 완화하기 위한 **정규성(Gaussianity) 가정**

$$(X_{t_1}, \ldots, X_{t_n}) \overset{d}{=} (X_{t_1+h}, \ldots, X_{t_n+h}),\ \ h = \text{lag},\ \forall n \geq 0$$

정규성을 가정하는 강정상성

$$(X_{t_1}, \ldots, X_{t_n}) \sim MVN(\mu, \Sigma)$$

평균벡터 μ와 공분산 행렬 Σ을 추정해서 전체 분포를 구할 수 있음!

확률변수의 기댓값은
상수(constant)

$$E[X_t] = m, \forall t \in Z$$

확률변수의 공분산은
시차에 의존

$$Cov(X_r, X_{r+h}) = \sigma_h^2, \forall r, h \in Z$$

2 정상성

정규성을 가정하는 강정상성

$$(X_{t_1}, \ldots, X_{t_n}) \sim MVN(\mu, \Sigma)$$

평균벡터 μ와 공분산 행렬 Σ을 추정해서 전체 분포를 구할 수 있음!

확률변수의 기댓값은
상수(constant)

$$E[X_t] = m, \forall t \in Z$$

확률변수의 공분산은
시차에 의존

$$Cov(X_r, X_{r+h}) = \sigma_h^2, \forall r, h \in Z$$

정규성을 가정하는 강정상성

$(X_{t1}, \dots, X_{tn}) \sim MVN(\mu, \Sigma)$

하지만, 이 역시 분포에 대한 조건이 포함되어 있는 **엄격한 가정**

평균벡터 μ와 공분산 행렬 Σ을 추정해서 전체 분포를 구할 수 있음!

현실에서는 **약정상성**의 개념을 사용

확률변수의 기댓값은 상수(constant)

$E[X_t] = m, \forall t \in Z$

확률변수의 공분산은 시차에 의존

$Cov(X_r, X_{r+h}) =$

2 정상성

약정상성

약정상성의 3가지 조건

1. $E[|X_t|]^2 < \infty, \ \forall t \in Z$

: 2차 적률(분산 관련)이 존재하고 시점 t에 관계없이 일정하다.

2. $E[X_t] = m, \ \forall t \in Z$

: 평균이 상수로 시점 t에 관계없이 일정하다.

3. $\gamma_x(r,s) = \gamma_x(r+h, s+h), \ \forall r,s,h \in Z$

$\gamma_x\ (r,s) \coloneqq Cov(X_r, X_s))$

: 공분산은 시차 h에 의존하며 시점 t와 무관하다.

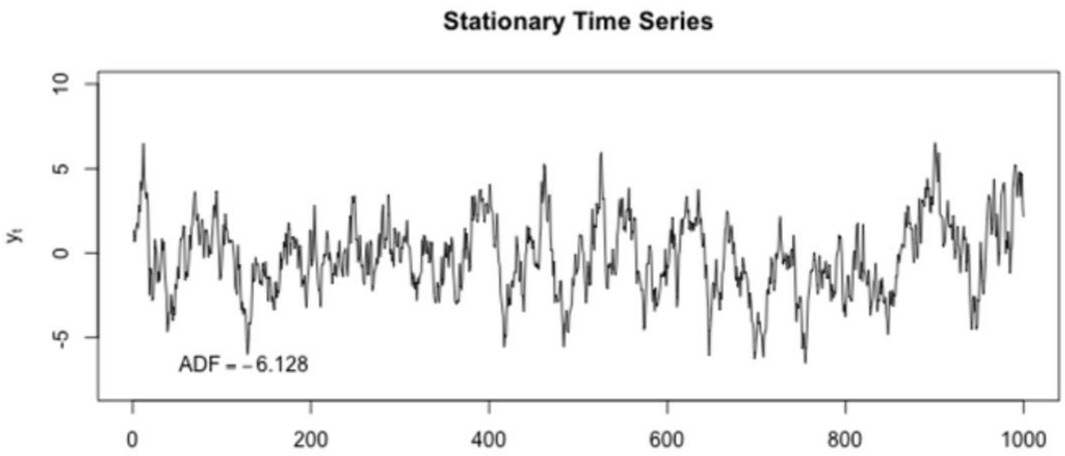
3

정상화

3 정상화

정상 시계열

시계열 Plot을 통해 식별 가능



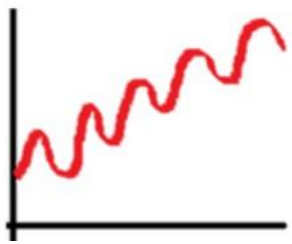
- ✓ 특별한 추세 X
- ✓ 계절성 X
- ✓ 일정한 평균과 분산

→ 정상 시계열

3 정상화

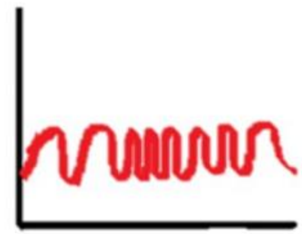
비정상 시계열

평균 일정 X



분산 일정 X

공분산 시점 의존



정상화 과정을 통해 정상 시계열로 변환 필요

3 정상화

정상화가 필요한 이유

선형 과정 시계열 모형으로의 적용

시계열 자료의 오차는 독립성 조건 만족 ✕

→ 정상화를 통해 정상성 조건을 만족해야

데이터를 선형 과정으로 표현할 수 있음!

안정적이고 정확한 예측

모델의 정확도가 시점에 따라 달라지는 것을 방지

데이터를 선형 과정으로 표현하는 이유와 방법은 2주차에서 다룰 예정!

3 정상화

정상화가 필요한 이유

선형 과정 시계열 모형으로의 적용

시계열 자료의 오차는 독립성 조건 만족 ✕

→ 정상화를 통해 정상성 조건을 만족해야

데이터를 선형 과정으로 표현할 수 있음!

안정적이고 정확한 예측

모델의 정확도가 시점에 따라 달라지는 것을 방지

데이터를 선형 과정으로 표현하는 이유와 방법은 2주차에서 다룰 예정!

분산이 일정하지 않은 경우의 정상화

'분산이 시점에 의존하지 않고 일정해야 한다'는

약정상성 조건에 위배되는 상황

분산 안정화 변환

(Variance Stabilizing Transformation)

로그 변환

제곱근 변환

Box-cox 변환

3 정상화

분산이 일정하지 않은 경우의 정상화

'분산이 시점에 의존하지 않고 일정하다'는

약정상성 조건에 위배되는 상황

분산 안정화 변환

(Variance Stabilizing Transformation)

로그 변환

제곱근 변환

Box-cox 변환

3 정상화

분산이 일정하지 않은 경우의 정상화

$$f(X_t) = \log(X_t)$$

$$f(X_t) = \sqrt{X_t}$$

$$f_\lambda(X_t) = \begin{cases} \dfrac{X_t^\lambda - 1}{\lambda} & if\ \lambda \geq 0 \\ logX_t & if\ \lambda = 0 \end{cases}$$

로그 변환

제곱근 변환

Box-cox 변환

3 정상화

평균이 일정하지 않은 경우의 정상화

평균이 일정하지 않은 총 3가지 경우

- **추세만** 존재하는 경우
- **계절성만** 존재하는 경우
- **추세와 계절성이 모두** 존재하는 경우

평균이 일정하지 않은 경우의 정상화

평균이 일정하지 않은 경우 정상화 방법

회귀 (Regression)

평활 (Smoothing)

차분 (Differencing)

비정상 부분을 추정하고 제거하여 정상화

3 정상화

평균이 일정하지 않은 경우의 정상화 | (1) 회귀 (Regression)

a. 추세만 존재하는 경우 : Polynomial Regression

[1] 추세만 존재하는 시계열 정의

$$X_t = m_t + Y_t, \quad E(Y_t) = 0$$

[2] 시간 t에 대한 추세성분 m_t의 선형회귀식

$$m_t = c_0 + c_1 t + c_2 t^2 + \cdots + c_p t^p$$

평균이 일정하지 않은 경우의 정상화 | (1) 회귀 (Regression)

a. 추세만 존재하는 경우 : Polynomial Regression

[1] 추세만 존재하는 시계열 정의

$$X_t = m_t + Y_t, \quad E(Y_t) = 0$$

[2] 시간 t에 대한 추세성분 m_t의 선형회귀식

$$m_t = c_0 + c_1 t + c_2 t^2 + \cdots + c_p t^p$$

평균이 일정하지 않은 경우의 정상화 | (1) 회귀 (Regression)

a. 추세만 존재하는 경우 : Polynomial Regression

[3] 최소제곱법(OLS)을 통한 선형회귀식 계수 추정

$$\left(\hat{c}_0, \hat{c}_1, \cdots, \hat{c}_p\right) = \arg\min_{c} \sum_{t=1}^{n} (X_t - m_t)^2$$

[4] 시계열에서 추정한 추세를 제거

$$X_t - \widehat{m}_t \approx Y_t$$

평균이 일정하지 않은 경우의 정상화 | (1) 회귀 (Regression)

a. 추세만 존재하는 경우 : Polynomial Regression

[3] 최소제곱법(OLS)을 통한 선형회귀식 계수 추정

$$\left(\hat{c}_0, \hat{c}_1, \cdots, \hat{c}_p\right) = \arg\min_{c} \sum_{t=1}^{n} (X_t - m_t)^2$$

[4] 시계열에서 추정한 추세를 제거

$$X_t - \widehat{m}_t \approx Y_t$$

평균이 일정하지 않은 경우의 정상화 | (1) 회귀 (Regression)

b. 계절성만 존재하는 경우 : Harmonic Regression

[1] 주기 = d인 계절성만을 가지는 시계열 가정

$$X_t = s_t + Y_t, \quad E(Y_t) = 0$$

[2] 시간 t에 대한 계절성분 s_t의 회귀식

$$s_t = a_0 + \sum_{j=1}^{\infty} \left(a_j \cos(\lambda_j t) + b_j \sin(\lambda_j t)\right)$$

평균이 일정하지 않은 경우의 정상화 | (1) 회귀 (Regression)

b. 계절성만 존재하는 경우 : Harmonic Regression

[1] 주기 = d인 계절성만을 가지는 시계열 가정

$$X_t = s_t + Y_t, \quad E(Y_t) = 0$$

[2] 시간 t에 대한 계절성분 s_t의 회귀식

$$s_t = a_0 + \sum_{j=1}^{\infty} (a_j \cos(\lambda_j t) + b_j \sin(\lambda_j t))$$

평균이 일정하지 않은 경우의 정상화 | (1) 회귀 (Regression)

b. 계절성만 존재하는 경우 : Harmonic Regression

[3] 적절한 λ_j와 k 선택 및 OLS를 통한 a_j와 b_j 추정

$$s_t = a_0 + \sum_{j=1}^{k} \left(a_j \cos(\lambda_j t) + b_j \sin(\lambda_j t)\right)$$

푸리에 급수 이용하여 계산

[4] 추정한 계절성 제거

$$X_t - \hat{s}_t \approx Y_t$$

평균이 일정하지 않은 경우의 정상화 | (1) 회귀 (Regression)

b. 계절성만 존재하는 경우 : Harmonic Regression

[3] 적절한 λ_j와 k 선택 및 OLS를 통한 a_j와 b_j 추정

$$s_t = a_0 + \sum_{j=1} \left(a_j \cos(\lambda_j t) + b_j \sin(\lambda_j t)\right)$$

푸리에 급수 이용하여 계산

적절한 λ_j와 k는?

(1) λ_j는 주기가 2π인 함수의 주기와 데이터 주기를 맞춰 주기 위한 값

1. 주기 반복 횟수 $f_1 = \left[\frac{n}{d}\right] \rightarrow f_j = jf_1$

2. $\lambda_j = f_j\,(2\pi/n)$

(2) k는 1~4 사이의 값 사용 (forward, backward, BIC, ⋯ 기준으로 정함)

ex. $n = 72,\ d = 12 \rightarrow f_1 = \left[\frac{72}{12}\right] = 6$

$\lambda_j = j \times 6 \times 2\pi/72$

(n = 데이터 개수, d = 주기)

3 정상화

평균이 일정하지 않은 경우의 정상화 | (1) 회귀 (Regression)

c. 추세와 계절성이 모두 존재하는 경우

Polynomial Regression과 Harmonic Regression 차례대로 진행

↓

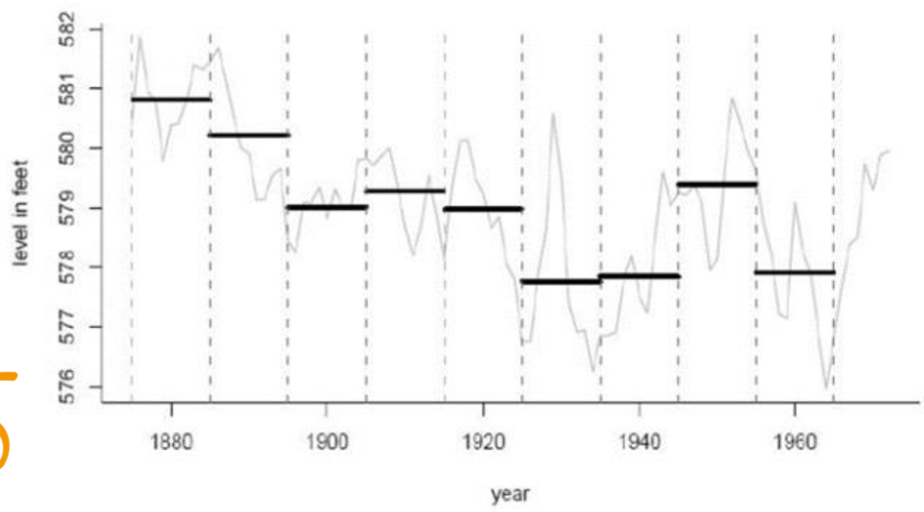
이후에도 남아있는 추세가 보인다면, 같은 과정을 반복해 제거

3 정상화

평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

평활

시계열 자료를 여러 구간으로 나눈 후, 구간의 평균들로 추세를 추정하는 방법



전체 시계열 자료와 구간 평균의 움직임은 비슷한 것이라는 아이디어 이용

평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

회귀와 평활의 비교

회귀

전체 데이터를 한 번에 추정

평활

구간을 나눠서 국소적으로 추정

국소적 변동에 주목해야 하는 경우 평활 사용

평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

a. 추세만 존재하는 경우: Moving Average Smoothing(MA)

[1] 길이가 $2q+1$인 구간의 평균

$$W_t = \frac{1}{2q+1}\sum_{j=-q}^{j=q}(m_{t+j} + Y_{t+j})$$

$$= \frac{1}{2q+1}\sum_{j=-q}^{j=q} m_{t+j} + \frac{1}{2q+1}\sum_{j=-q}^{j=q} Y_{t+j}$$

길이가 $2q+1 \rightarrow$ 앞뒤로 q 개 의미

평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

a. 추세만 존재하는 경우: Moving Average Smoothing(MA)

[2] 위의 식에 추세 성분 m_t 대입, 추세는 Linear함을 가정

$$m_t = c_0 + c_1 t,\ E(Y_t) = 0$$

$$W_t = \frac{1}{2q+1}\sum_{j=-q}^{j=q} m_{t+j} + \frac{1}{2q+1}\sum_{j=-q}^{j=q} Y_{t+j} = m_t$$

$$\frac{1}{2q+1}\sum_{j=-q}^{j=q} Y_{t+j} \approx E(Y_t) = 0\ (by\ WLLN)$$

WLLN은 약대수의 법칙

[3] 추세 부분만 남은 W_t를 X_t에서 제거

$$X_t - \widehat{m}_t \approx Y_t$$

평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

a. 추세만 존재하는 경우: Moving Average Smoothing(MA)

[2] 위의 식에 추세 성분 m_t 대입, 추세는 Linear함을 가정

$$m_t = c_0 + c_1 t,\ E(Y_t) = 0$$

$$W_t = \frac{1}{2q+1} \sum_{j=-q}^{j=q} m_{t+j} + \frac{1}{2q+1} \sum_{j=-q}^{j=q} Y_{t+j} = m_t$$

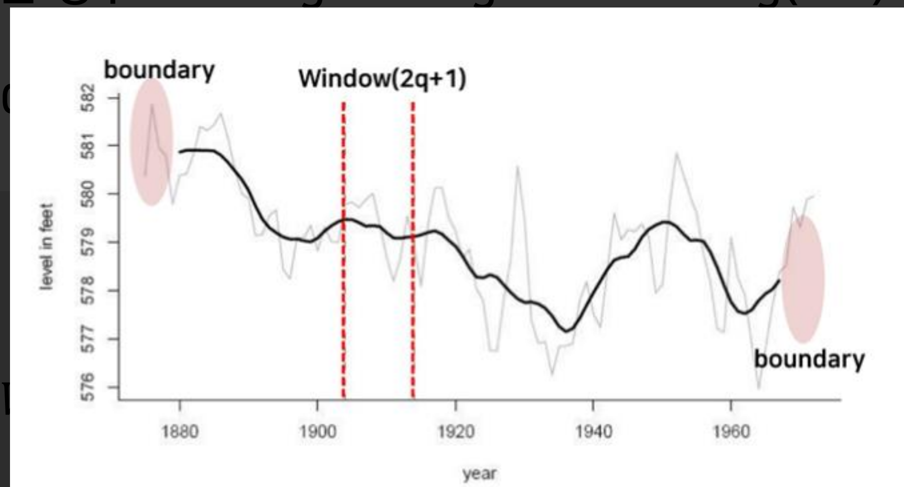
$$\frac{1}{2q+1} \sum_{j=-q}^{j=q} Y_{t+j} \approx E(Y_t) = 0\ (by\ WLLN)$$

구간 평균 W_t는 근사적으로 추세 m_t와 같아짐

[3] 추세 부분만 남은 W_t를 X_t에서 제거

a. 추세만 존재하는 경우: Moving Average Smoothing(MA)

Moving Average Smoothing의 약점



1. 국소적 변동은 잘 설명할 수 있으나,

데이터의 맨 앞 q개와 맨 뒤 q개의 boundary는 추정할 수 없음

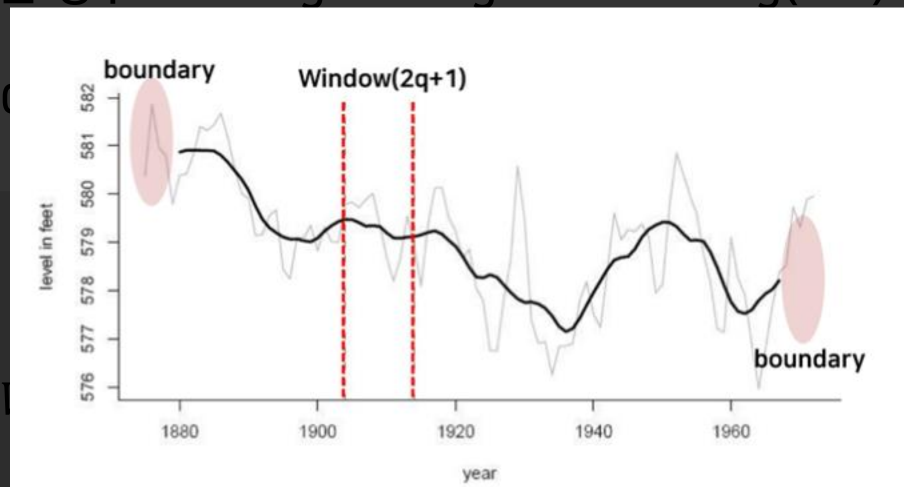
2. 현실에서는 미래의 관측값을 사용할 수 없음

Moving Average Smoothing의 약점

평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

a. 추세만 존재하는 경우: Moving Average Smoothing(MA)

[2] 위의 식



$$\frac{1}{2q+1}\sum_{j=-q}^{j=q} Y_{t+j} \approx E(Y_t) = 0 \ (by\ WLLN)$$

1. 국소적 변동은 잘 설명할 수 있으나,

2. 데이터의 맨 앞 q개와 맨 뒤 q개의 **boundary**는 추정할 수 없음

2. 현실에서는 미래의 관측값을 사용할 수 없음

[3] 추세 부분만 남은 W를 X에서 제거!

$$X_t - \hat{m}_t \approx Y_t$$

→ 과거의 데이터만을 가지고 추세를 제거할 수 있는 **지수 평활법** 사용

평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

a. 추세만 존재하는 경우: Exponential Smoothing

[1] 추세 추정

$$\hat{m}_1 = X_1$$

$$\hat{m}_2 = aX_2 + (1-a)\hat{m}_1$$

$$\vdots$$

$$\hat{m}_t = aX_t + (1-a)\hat{m}_{t-1} = \sum_{j=0}^{t-2} a(1-a)^j X_{t-j} + (1-a)^{t-1} X_1$$

a는 과거 관측치에 대한 가중치, $a \in [0,1]$

[2] 추정한 추세를 시계열에서 제거

$$X_t - \hat{m}_t \approx Y_t$$

평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

a. 추세만 존재하는 경우: Exponential Smoothing

[1] 추세 추정

$$\hat{m}_1 = X_1$$

$$\hat{m}_2 = aX_2 + (1-a)\hat{m}_1$$

$$\vdots$$

$$\hat{m}_t = aX_t + (1-a)\hat{m}_{t-1} = \sum_{j=0}^{t-2} a(1-a)^j X_{t-j} + (1-a)^{t-1} X_1$$

a는 과거 관측치에 대한 가중치, $a \in [0,1]$

[2] 추정한 추세를 시계열에서 제거

과거의 관측치일수록 가중치의 값이 지수적으로 감소하는 방식으로 추세 추정

$$X_t - \hat{m}_t \approx Y_t$$

평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

a. 추세만 존재하는 경우: Exponential Smoothing

✓ 이동평균 평활법과 지수 평활법 모두 추세 외에 파라미터 q와 a 추정 필요

⋮

Why?

q가 작으면 작은 변화들도 잘 잡아내지만, 변동성이 심해짐

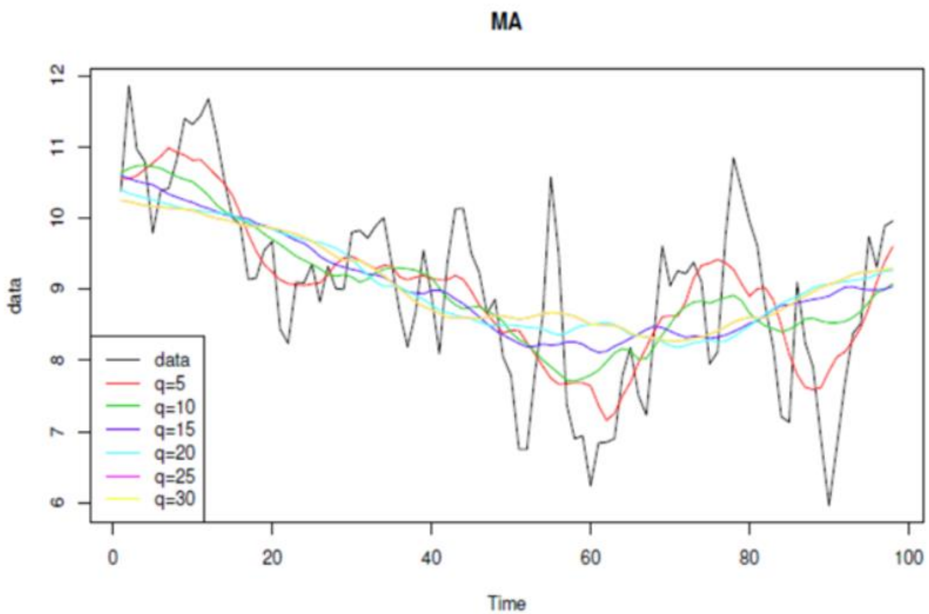
반대로 q가 클 경우 변동성은 줄어들지만, 작은 변화들을 잡아내지 못함

→ bias-variance trade off 발생 !

평균이 일정하

a. 추세만 존

✓ 이동평균 … 추정 필요



반대로 q가 클 경우 변동성은 줄어들지만, 작은 변화들을 잡아내지 못함

→ bias-variance trade off 발생 !

How?

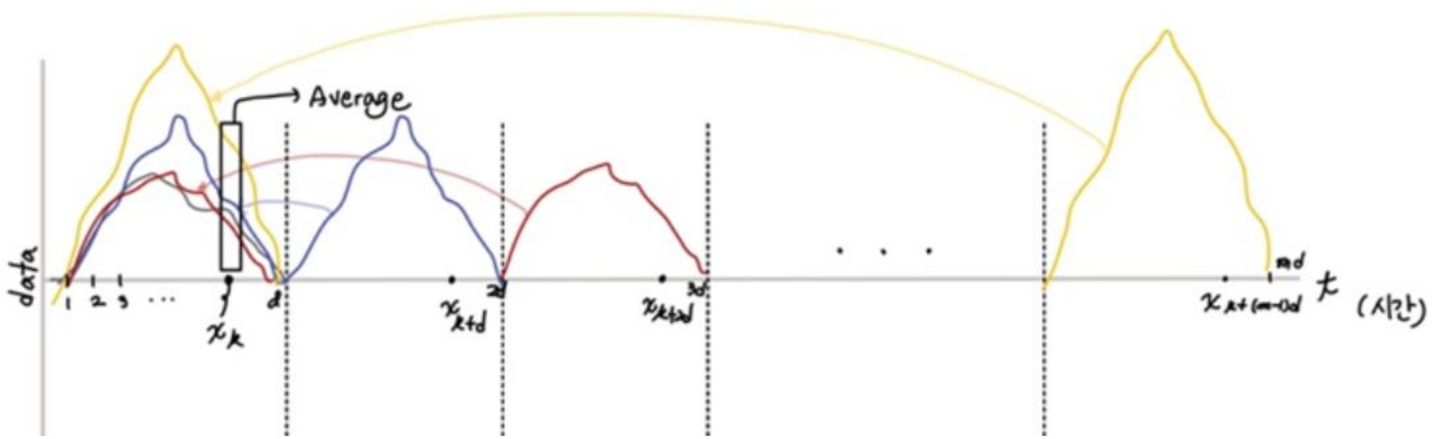
Cross-validation(CV)를 통해 MSE 추정하여 최적의 파라미터 선택

평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

b. 계절성만 존재하는 경우: Seasonal Smoothing

Seasonal Smoothing

주기가 d인 시계열 자료에서 주기만큼의 데이터들을 모두 겹친(overlay) 후,
겹쳐진 값들의 평균으로 계절성 추정



평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

b. 계절성만 존재하는 경우: Seasonal Smoothing

[1] $k = 1, \cdots, d$에 대한 계절성분($\hat{s}_k$) 추정

$$\hat{s}_k = \frac{1}{m}(x_k + x_{k+d} + \cdots + x_{k+(m-1)d}) = \frac{1}{m}\sum_{j=0}^{m-1} x_{k+jd}$$

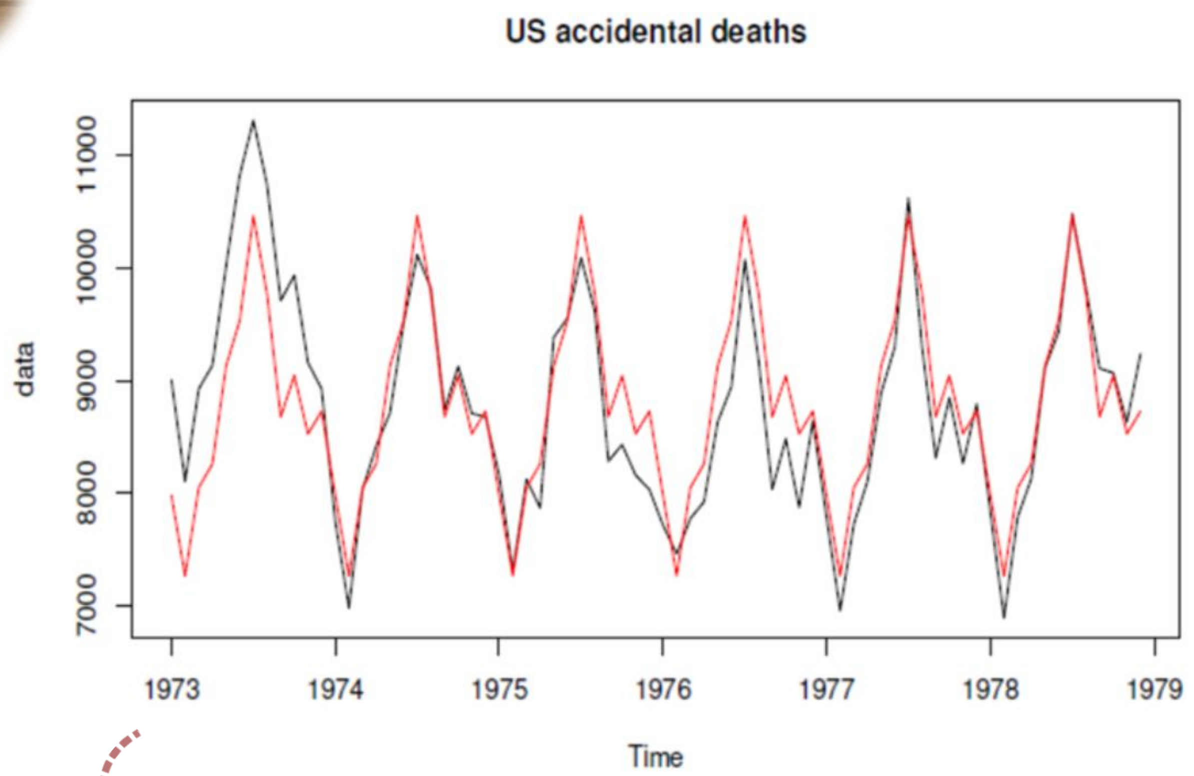
$$\hat{s}_k = \hat{s}_{k-d},\ \ if\ k > d$$

↓

[2] 추정된 계절 성분을 다른 주기에도 적용해 전체 계절성을 추정하고 제거

정상화

b. 계절



[2] 추정된 계절 성분을 다른 주기에도 적용해 전체 계절성을 추정하고 제거

빨간색으로 그려진 계절 성분은 모든 주기에서 동일하게 반복

평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

c. 추세와 계절성이 모두 존재하는 경우: Classical Decomposition Algorithm

[1] MA filter를 이용하여 추세를 추정

$if\ d = 2q$ (짝수),

$$\hat{m}_t = \frac{0.5X_{t-q} + X_{t-q+1} + \cdots + X_{t+q-1} + 0.5X_{t+q}}{2q}$$

$if\ d = 2q+1$ (홀수),

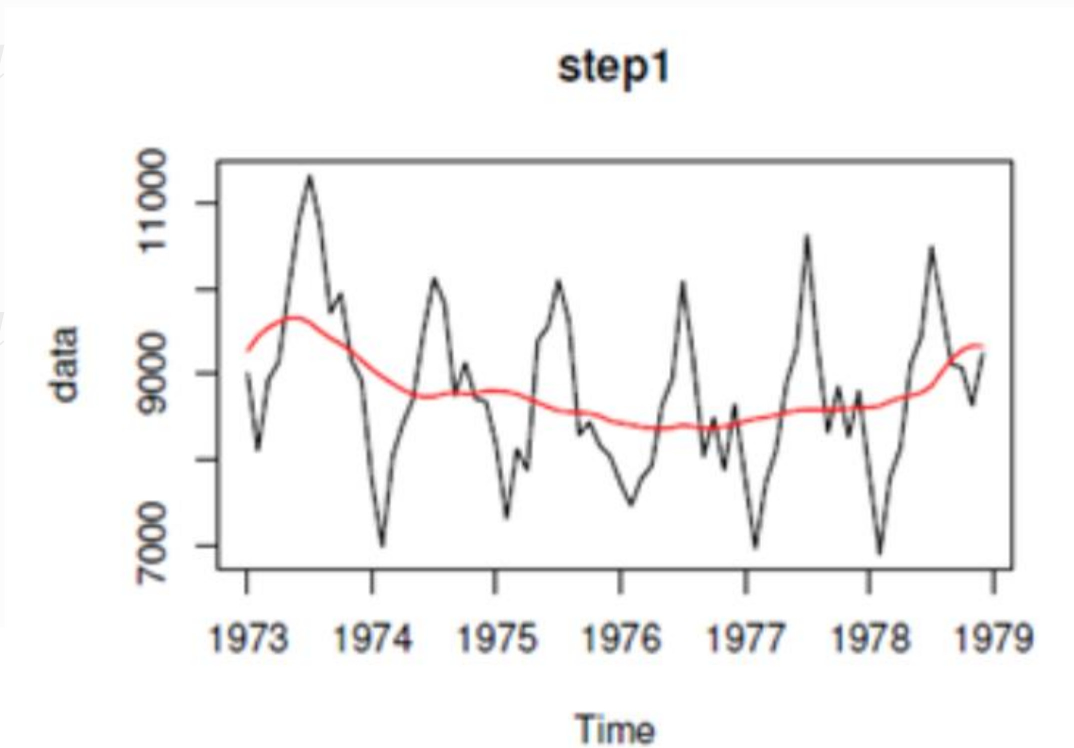
$$\hat{m}_t = \frac{X_{t-q} + X_{t-q+1} + \cdots + X_{t+q-1} + X_{t+q}}{2q}$$

Window를 주기 d와 같게 하는 이유는 계절성의 영향을 피하기 위함

평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

c. 추세와 계절성이 모두 존재하는 경우: Classical Decomposition Algorithm

[1] MA filter를 이용하여 추세를 추정

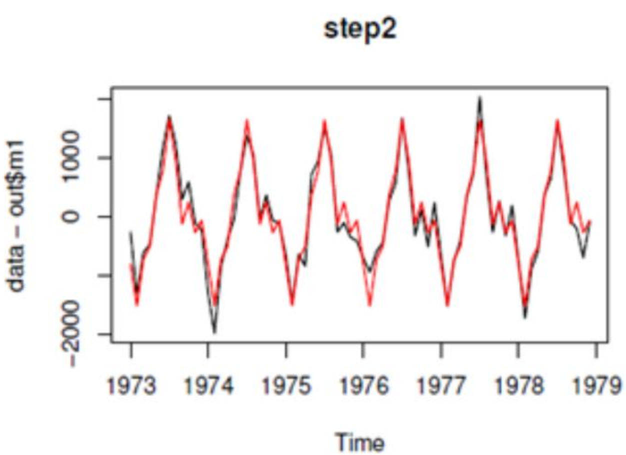


평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

c. 추세와 계절성이 모두 존재하는 경우: Classical Decomposition Algorithm

[2] 추정한 추세를 제거한 후,

Seasonal smoothing으로 계절성 추정



[3] 추정한 계절성을 제거한 후,

OLS를 활용하여 다시 추세 추정

평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

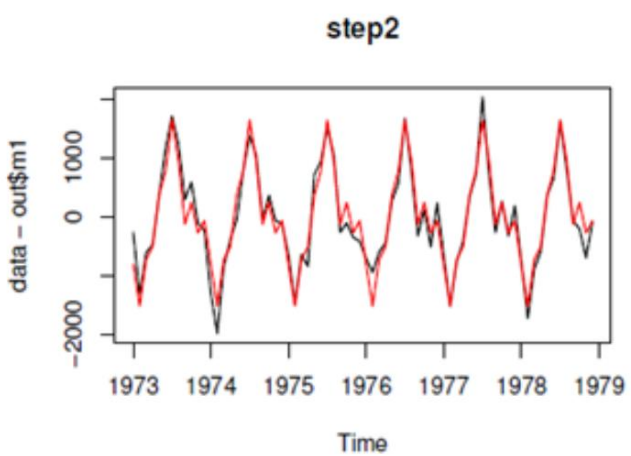
c. 추세와 계절성이 모두 존재하는 경우: Classical Decomposition Algorithm

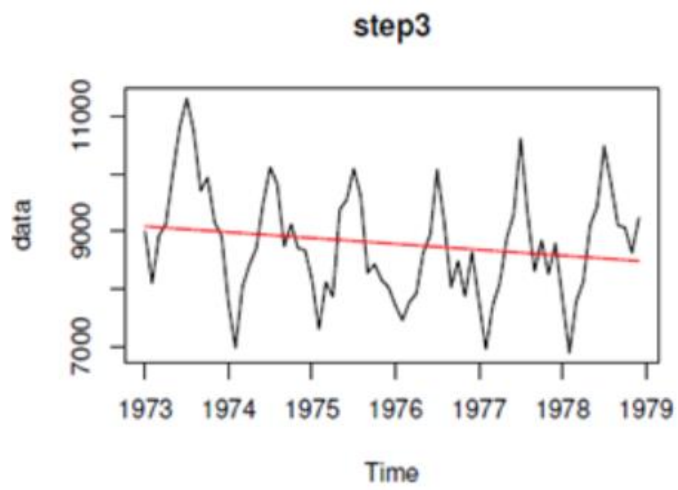
[2] 추정한 추세를 제거한 후,

Seasonal smoothing으로 계절성 추정

[3] 추정한 계절성을 제거한 후,

OLS를 활용하여 다시 추세 추정





평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

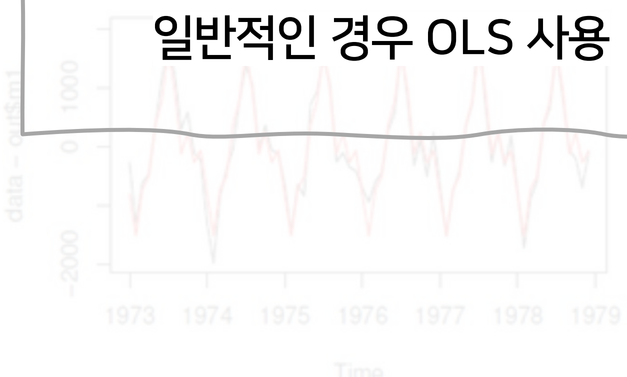
c. 추세와 계절성이 모두 존재하는 경우: Classical Decomposition Algorithm

[2] 추정한 추세를 제거한 후,

Seasonal smoothing으로 계절성 추정

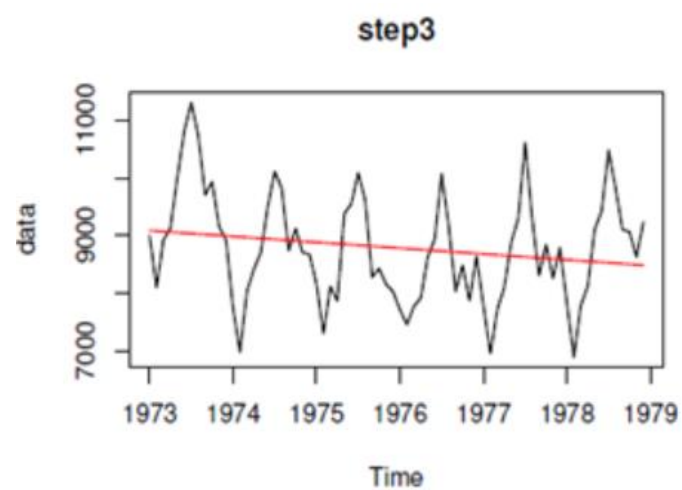
OLS 대신 Smoothing 사용 가능,

일반적인 경우 OLS 사용



[3] 추정한 계절성을 제거한 후,

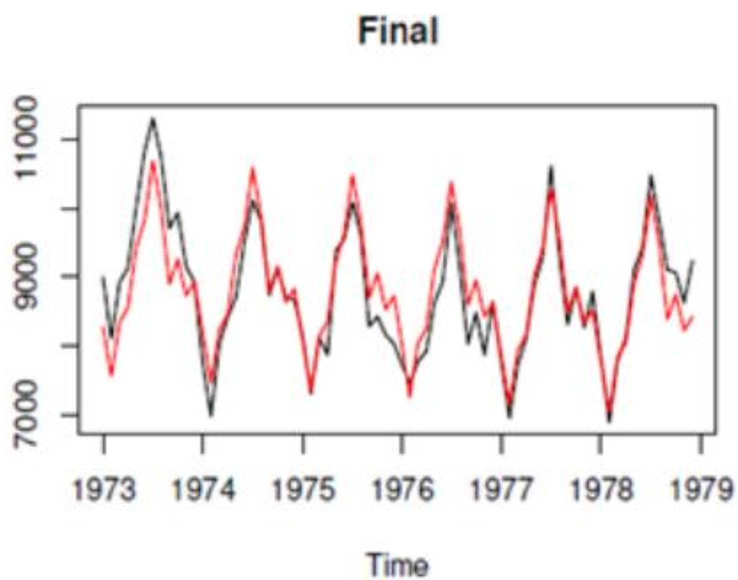
OLS를 활용하여 다시 추세 추정



평균이 일정하지 않은 경우의 정상화 | (2) 평활 (Smoothing)

c. 추세와 계절성이 모두 존재하는 경우: Classical Decomposition Algorithm

[4] 다시 추정한 추세를 제거



⋮

이후에도 추세 혹은 계절성이 존재한다면 [1]~[4] 과정 반복

평균이 일정하지 않은 경우의 정상화 | (3) 차분

차분(Differencing)

이 '차이'를 이용하여 추세와 계절성 제거 !

관측값들의 **차이**를 구하는 것

후향연산자를 사용하여 차분을 표현

후향연산자

$$BX_t = X_{t-1}$$

관측값을 한 시점 전으로 돌려주는 역할을 함

평균이 일정하지 않은 경우의 정상화 | (3) 차분

차분(Differencing)

이 '차이'를 이용하여 추세와 계절성 제거 !

관측값들의 **차이**를 구하는 것

후향연산자를 사용하여 차분을 표현

1차 차분

$$\nabla X_t = X_t - X_{t-1}$$
$$= (1 - B)X_t$$

후향연산자

$$BX_t = X_{t-1}$$

관측값을 한 시점 전으로 돌려주는 역할을 함

2차 차분

$$\nabla^2 X_t$$
$$= \nabla(\nabla X_t) \ = \nabla(X_t - X_{t-1})$$
$$= X_t - 2X_{t-1} + X_{t-2}$$
$$= (1 - B)^2 X_t$$

평균이 일정하지 않은 경우의 정상화 | (3) 차분

a. 추세만 존재하는 경우: Differencing (Regular Differencing)

추세를 선형이라고 가정

$$m_t = c_0 + c_1 t$$

추세의 1차 차분

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

시간 t 에 영향을 받지 않는 상수만 남음

→ 추세가 제거된 것 !

평균이 일정하지 않은 경우의 정상화 | (3) 차분

a. 추세만 존재하는 경우: Differencing (Regular Differencing)

$$\nabla^2 m_t = \nabla\{(c_0 + c_1 t + c_2 t^2) - (c_0 + c_1(t-1) + c_2(t-1)^2)\}$$
$$= \nabla(c_1 - c_2 + 2c_2 t) = 2c_2$$
$$\vdots$$
$$\nabla^k m_t = \cdots = k!\, c_k$$

추세가 k차라면,

k차 차분을 진행해서 추세를 제거할 수 있음

평균이 일정하지 않은 경우의 정상화 | 차분

a. 추세만 존재하는 경우: Differencing (Regular Differencing)

차분으로 추세를 제거하는 방법은 직관적이지만,

아래 식처럼 오차까지 차분되어 식이 복잡해진다는 단점이 있음.

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

추세가 k차라면,

k차 차분을 진행해서 추세를 제거할 수 있음

평균이 일정하지 않은 경우의 정상화 | (3) 차분

b. 계절성만 존재하는 경우: Seasonal Differencing

"lag-d differencing"을 통해 계절성을 제거

lag-d 차분

d시점 전과의 관측값 차이

$$\nabla_d X_t = (1 - B^d) X_t$$

일반적인 차분: 직전 시점과의 관측값 차이

평균이 일정하지 않은 경우의 정상화 – 차분

b. 계절성만 존재하는 경우: Seasonal Differencing

차분의 표현법 차이

"lag-d differencing"을 통해 계절성을 제거

d차 차분

$\nabla^d = (1 - B)^d$

lag-d 차분

$\nabla_d = (1 - B^d)$

lag-d 차분

d시점 전과의 관측값 차이

$$\nabla_d X_t = (1 - B^d) X_t$$

일반적인 차분: 직전 시점과의 관측값 차이

평균이 일정하지 않은 경우의 정상화 | (3) 차분

b. 계절성만 존재하는 경우: Seasonal Differencing

$s_t = s_{t-d}$로 계절성 가정 후, lag-d 차분 적용

$$\nabla_d X_t = s_t - s_{t-d} + Y_t - Y_{t-d} = 0 + error$$

차분 결과 오차항만 남음

⇢ 계절성이 제거된 것 !

평균이 일정하지 않은 경우의 정상화 | (3) 차분

c. 추세와 계절성이 모두 존재하는 경우 – lag-d 차분 + p차 차분

1단계: 계절차분

$s_t = s_{t-d}$ 이므로 0이 됨

$$\nabla_d X_t = m_t - m_{t-d} + s_t - s_{t-d} + Y_t - Y_{t-d}$$

2단계: 차분 (남아있는 추세 제거)

추세가 p차라면, p-1차 차분을 진행 (∇^{p-1})

평균이 일정하지 않은 경우의 정상화 | (…) 차분

c. 추세와 계절성이 모두 존재하는 경우 – lag-d 차분 + p차 차분

2단계에서 p-1차 차분을 하는 이유

1단계: 계절차분

$s_t = s_{t-d}$ 이므로 0이 됨

$$\nabla_d X_t = m_t - m_{t-d} + s_t - s_{t-d} + Y_t - Y_{t-d}$$

$$\nabla_d = (1 - B^d)(1 - B)(1 + B + \cdots + B^{d-1})$$

1단계 계절차분에 $(1 - B)$ 가 이미 포함되어 있기 때문

2단계: 차분 (남아있는 추세 제거)

추세가 p차라면, p-1차 차분을 진행

$$\nabla^{p-1}(\nabla_d X_t) = (1 - B)^{p-1}(1 - B)\left(1 + B + \cdots + B^{d-1}\right)X_t$$

⇢ P차 추세 제거

4

정상성 검정

정상성 검정

시계열 자료의 비정상 부분을 제거하였다면,

정상성을 만족하는 오차(Y_t)만 남아있어야 함

1. 자기공분산함수 (ACVF)
2. 자기상관함수 (ACF)

위 두 함수로 오차의 정상성 만족 여부 검정 !

정상성 검정

시계열 자료의 비정상 부분을 지금까지의 과정을 통해 제거하였다면

이제 정상성을 만족하는 오차(Y_t)만 남아 있음

정상성 검정을 통해

시계열의 확률적 성질이 시간 t에 의존하는 정도

즉, 시간에 따른 상관 정도를 파악하고자 함이 최종 목적 !

1. 자기공분산함수 (ACVF)
2. 자기상관함수 (ACF)

위 두 함수로 오차의 정상성 만족 여부 검정 !

정상성 검정

[1] 자기공분산함수 (ACVF)

$$\gamma_X(h) = Cov(X_t, X_{t+h}) = \mathbb{E}[(X_t - \mu)(X_{t+h} - \mu)]$$

[1-1] 표본자기공분산함수 (SACVF)

$$\widehat{\gamma_X}(h) = \frac{1}{n}\sum_{j=1}^{n-h}(X_j - \bar{X})(X_{j+h} - \bar{X})$$

원칙상 $n-h$이어야 하지만, Non-negative Definite 성질을 만족시키기 위한 것임

정상성 검정

[1] 자기공분산함수 (ACVF)

$$\gamma_X(h) = Cov(X_t, X_{t+h}) = \mathbb{E}[(X_t - \mu)(X_{t+h} - \mu)]$$

[1-1] 표본자기공분산함수 (SACVF)

약정상성을 만족한다는 조건 하에 공분산은 시차에만 의존하므로,

$$\widehat{\gamma_X}(h) = \frac{1}{n}\sum^{n-h}(X_t - \bar{X})(X_{t+h} - \bar{X})$$

$$\gamma_X(0) = Var(X_t)$$

원칙상 $n - h$이어야 하지만,

Non-negative Definite 성질을 만족시키기 위한 것임

정상성 검정

[2] 자기상관함수 (ACF)

$$\rho_X(h) = Corr(X_t, X_{t+h}) = \frac{Cov(X_t, X_{t+h})}{\sqrt{Var(X_t)}\sqrt{Var(X_{t+h})}} = \frac{\gamma_X(h)}{\gamma_X(0)}$$

[2-1] 표본자기상관함수 (SACF)

$$\widehat{\rho_X}(h) = \frac{\widehat{\gamma_X}(h)}{\widehat{\gamma_X}(0)}$$

※ 표본자기공분산함수와 표본자기상관함수는

시계열 데이터에서 샘플링 된 데이터의 정상성 검정을 위해 사용하는 함수

정상성 검정

[2] 자기상관함수 (ACF)

$$\rho_X(h) = Corr(X_t, X_{t+h}) = \frac{Cov(X_t, X_{t+h})}{\sqrt{Var(X_t)}\sqrt{Var(X_{t+h})}} = \frac{\gamma_X(h)}{\gamma_X(0)}$$

[2-1] 표본자기상관함수 (SACF)

$$\widehat{\rho_X}(h) = \frac{\widehat{\gamma_X}(h)}{\widehat{\gamma_X}(0)}$$

앞서 본 $\gamma_X(0) = Var(X_t)$ 성질로 인해,

$\rho_X(0) = 1$ 임을 알 수 있음

※ 표본자기공분산함수와 표본자기상관함수는
시계열 데이터에서 샘플링 된 데이터의 정상성 검정을 위해 사용하는 함수

백색잡음

백색잡음 White Noise

자기상관이 존재하지 않는 시계열

시계열 $\{X_t\}$ 가

평균이 0 / 유한한 분산 / 자료 내 상관관계 X

이 세가지 조건을 만족

⋮

$\{X_t\} \sim WN(0, \sigma^2)$로 표현되는 백색잡음

백색잡음

백색잡음 White Noise

자기상관이 존재하지 않는 시계열

시계열 $\{X_t\}$ 가

이 세가지 조건을 만족

$\{X_t\} \sim WN(0, \sigma^2)$로 표현되는 백색잡음

우리가 일반적으로 사용하는 $iid(0, \sigma^2)$는 **백색잡음이지만,**

백색잡음이라고 항상 $iid\ (0, \sigma^2)$는 아님을 주의

iid**에서 독립성 조건이 완화된 것이 백색잡음**

백색잡음 검정

비정상 시계열의 추세(m_t)와 계절성(s_t)을 성공적으로 제거했다면,

남아있는 오차항은 **WN 조건** 혹은 **IID 조건**을 만족함

→ $\{X_t\} \sim WN(0, \sigma^2)$ 를 의미

$\{X_t\}$의 분산인 $Var(X_t) = \sigma^2$을 추정하여

오차항이 WN 조건을 만족하는지 검정

백색잡음 검정

백색잡음 검정

i)
자기상관 검정

ii)
정규성 검정

iii)
정상성 검정

백색잡음 검정 | (1) 자기상관 검정

오차가 백색잡음 $WN(0,1)$을 따른다고 가정하면

표본자기상관함수 $\hat{\rho}(h)$는 평균이 0이고 분산이 1/n인 정규분포로 근사

$$\hat{\rho}(h) \approx \mathrm{N}(0, \frac{1}{n})$$

가설 검정

$$H_0:\ \rho(h) = 0$$

$$vs$$

$$H_1: \rho(h) \neq 0$$

백색잡음 검정 | (1) 자기상관 검정

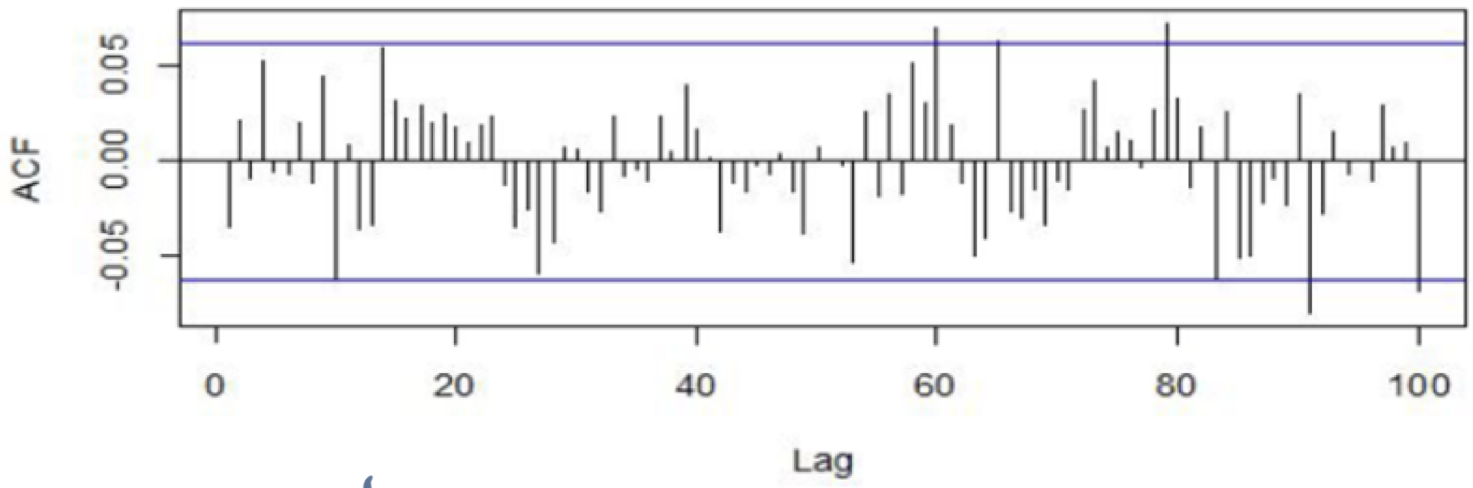
$|\hat{\rho}(h)| < \frac{1.96}{\sqrt{n}}$ 이면

귀무가설 H_0 기각할 수 없음

오차항에 **자기상관이 없다**는 결론

백색잡음 검정 | (1) 자기상관 검정

ACF plot



x축은 시차, y축은 ACF

파란선 안쪽이 자기상관 검정의 신뢰구간

백색잡음 검정 | (1) 자기상관 검정

ACF plot

파란색 신뢰구간 밖으로 벗어난 그래프는

해당 X축 값의 시차만큼 자기상관성이 존재함을 의미

x축은 시차, y축은 ACF

파란선을 통해 신뢰구간 확인

백색잡음 검정 | (2) 정규성 검정

가설 검정

H_0: 정규성이 존재한다

vs

H_1: 정규성이 존재하지 않는다

QQ plot	시각적으로 확인할 수 있는 방법
KS plot	표본과 모집단의 누적확률분포가 얼마나 유사한지 비교하는 방법
Jarque-Bera test	왜도와 첨도를 통해 정규성을 검정하는 방법

백색잡음 검정 | (3) 정상성 검정

<table>
<tr><th></th><th>특징</th><th>H_0</th></tr>
<tr><td>Kpss test</td><td rowspan="2">단위근 검정방법 중 하나</td><td rowspan="3">정상 시계열이다</td></tr>
<tr><td>ADF test</td></tr>
<tr><td>PP test</td><td>이분산이 있는 경우에도
사용가능한 검정방법</td></tr>
</table>

5

1주차 정리

5 1주차 정리

정리

시계열 자료

관측치들 간 dependency 有

규칙요소

추세 / 순환 / 계절성

불규칙요소

우연 변동

덧셈 분해

$$X_t = m_t + s_t + Y_t$$

정리

시계열 자료

관측치들 간 dependency 有

규칙요소

추세 / 순환 / 계절성

불규칙요소

우연 변동

분석을 위해서는,

시계열 자료가 '정상성'을 만족해야 함!

덧셈 분해

$$X_t = m_t + s_t + Y_t$$

5 1주차 정리

정리 | 정상성

정상성

시계열 자료의 확률적 성질이 '시차'에만 의존

현실에서는 주로 약정상성을 이용 !

<약정상성의 조건>

$E[\lvert X_t \rvert]^2 < \infty, \forall t \in Z$	2차적률 有, 시점 t에 관계없이 일정
$E[X_t] = m, \forall t \in Z$	평균 = 상수, 시점 t에 관계없이 일정
$\gamma_x(r,s) = \gamma_x(r+h, s+h),$ $\forall r,s,h \in Z$	공분산은 시차 t에 의존, 시점 t와 무관

정리 | 정상성

정상성

시계열 자료의 확률적 성질이 '시차'에만 의존

현실에서는 주로 약정상성을 이용 !

<약정상성의 조건>

$E[\|X_t\|]^2 < \infty, \forall t \in Z$	2차적률 有, 시점 t에 관계없이 일정
$E[X_t] = m, \forall t \in Z$	평균 = 상수, 시점 t에 관계없이 일정
$\gamma_x(r,s) = \gamma_x(r+h, s+h),$ $\forall r, s, h \in Z$	공분산은 시차 t에 의존, 시점 t와 무관

만족하지 않는다면,

정상화 진행 !

정리 | 정상화

정상화

비정상 시계열을 정상 시계열로 변환

분산 일정 X	평균 일정 X
로그(Log) 변환	회귀 (Regression)
제곱근 (Square Root) 변환	평활 (Smoothing)
Box-Cox 변환	차분 (Differencing)

5 1주차 정리

정리 | 정상화

정상화

비정상 시계열을 정상 시계열로 변환

분산 일정 X

로그(Log) 변환

루트(Square Root) 변환

Box-Cox 변환

평균 일정 X

회귀 (Regression)

평활 (Smoothing)

차분 (Differencing)

남아있는 오차가

정상성을 만족하는지 확인 !

정리 | 정상성 검정

자기공분산함수 (ACVF)

$$\gamma_X(h) = Cov(X_t, X_{t+h}) = E[(X_t - \mu)(X_{t+h} - \mu)]$$

자기상관함수 (ACF)

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = Corr(X_t, X_{t+h}) = \frac{Cov(X_t, X_{t+h})}{\sqrt{Var(X_t)}\sqrt{Var(X_{t+h})}}$$

5 1주차 정리

정리 | 정상성 검정

평균 = 0, 분산 = σ^2

백색잡음 $WN(0, \sigma^2)$

자기상관이 존재하지 않는 시계열

⋮

자기상관 검정	ACF plot
정규성 검정	QQ plot / KS Test / Jarque-Bera Test
정상성 검정	KPSS Test / ADF Test / PP Test

다음 주 예고

1. 모형 식별
2. 선형 과정
3. AR 모형
4. MA 모형
5. ARMA 모형
6. 적합 절차