17기 정규세션

ToBig's 16기 강의자

이예림

Time Series Analysis

시계열분석

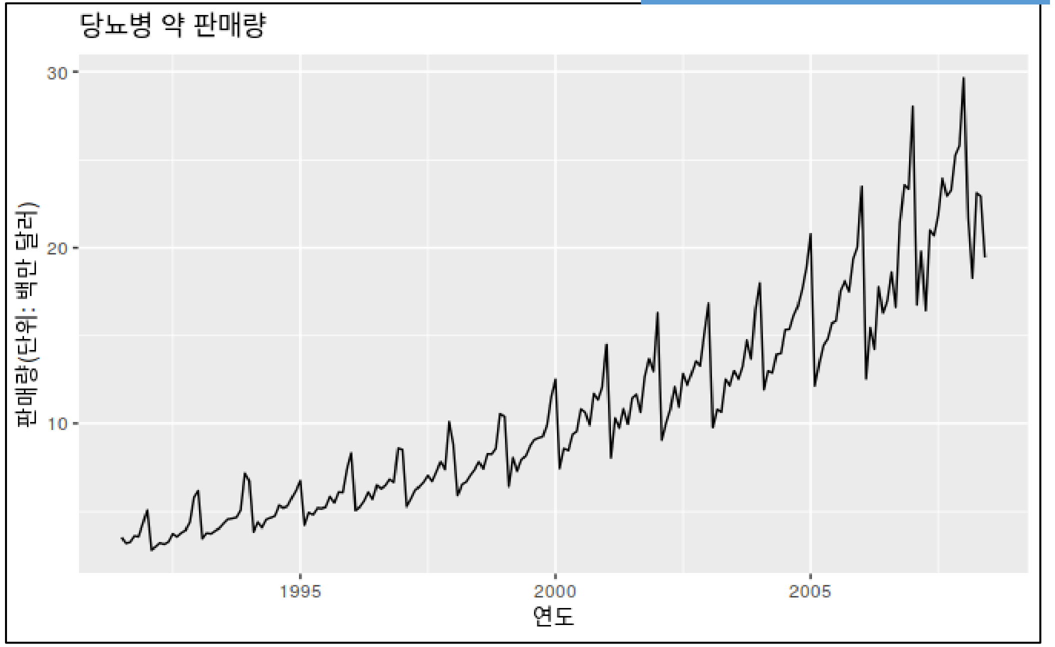
Contents

Contents

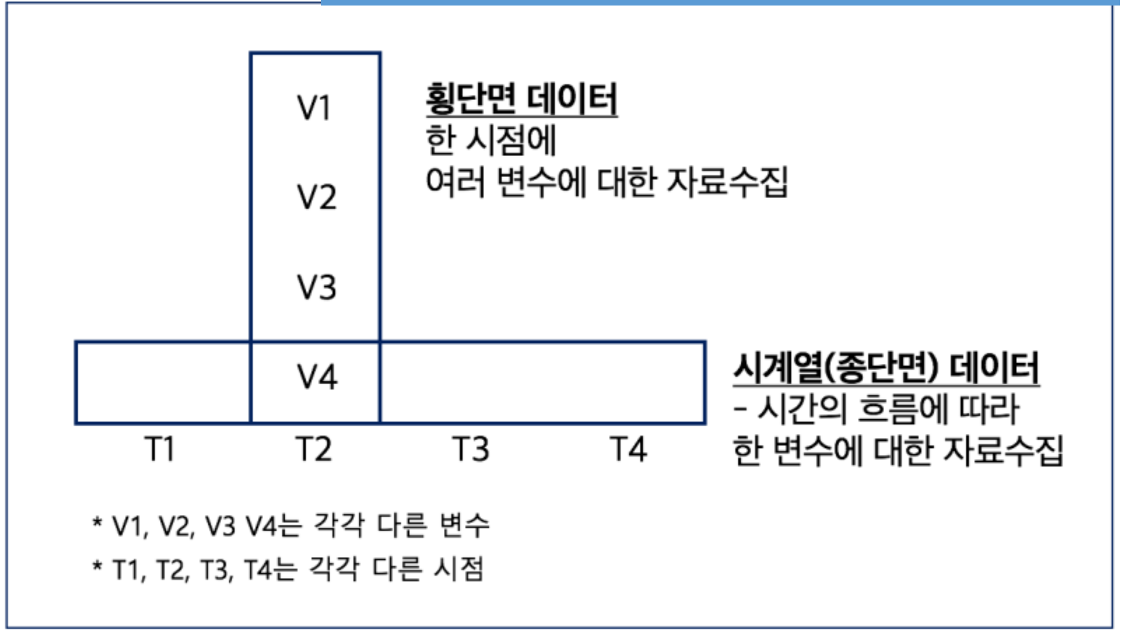
시계열 데이터 (Time Series Data)

- 시간의 흐름에 따라 일정한 간격으로 기록된 종단면 데이터
  - Ex. 주가지수, 일일 기온, 월별 교통사고 건수, 연간 광고 비용 등

시계열 데이터 예시



(참고) 횡단면 데이터 VS 종단면 데이터



시계열의 구성

추세 (Trend)	계절성 (Seasonal Variation)	순환 (Cycle Variation)	불규칙 요인 (Irregular Movements)
• 시계열의 장기적 경향 • 상승 경향 or 감소 경향 or 안정적	• 1년 주기 • 단기적 변동 • Ex. 여름에는 냉방기기 판매량이 높고, 겨울에는 판매량이 낮음	• 1년보다 긴 주기(2년 ~10년)를 갖고 순환 • 중기적 변동 • Business Cycle	• 예측 및 제어가 불가능 • 우연적 • Noise

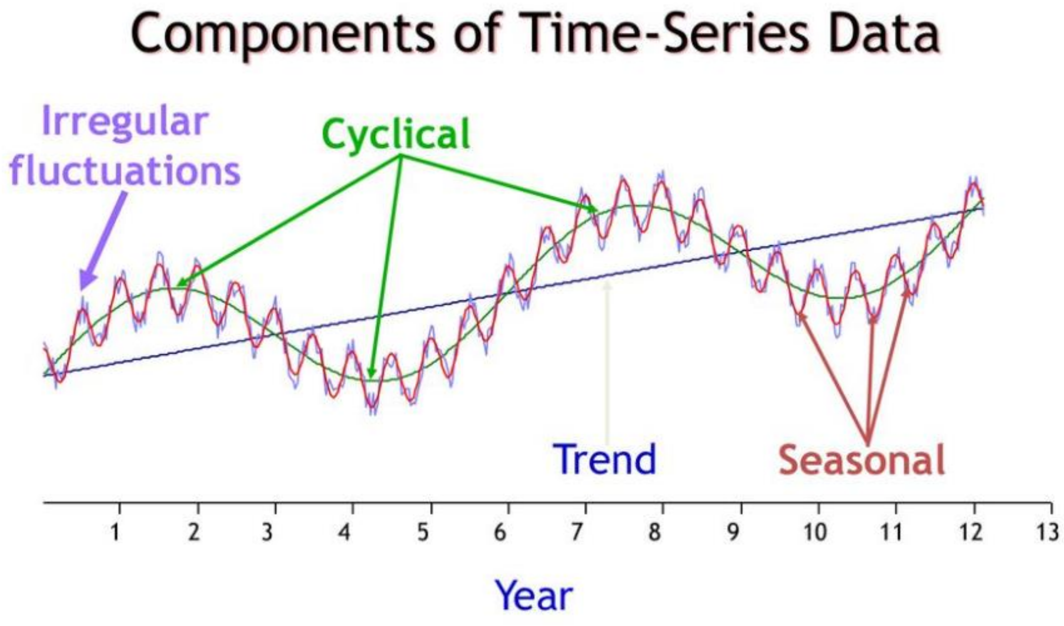
시계열 모형화

- 시계열을 구성하는 4가지 요소는 단일로 시계열을 구성하기도, 가법 or 승법으로 구성하기도 한다.

- 가법 모형 (Additive Model)
  - 구성요소 간 독립을 가정
  - 각 구성요소를 더하는 모형

- 승법 모형 (Multiple Model)
  - 구성요소 간 상호작용을 가정 (독립 X)
  - 각 구성요소를 곱하는 모형

- 시계열 분해 : 하나의 시계열을 구성요소별로 분해



시계열 분석 목적

기술
- 시도표(time plot)나 자기상관함수를 사용해 자료의 특성 파악
- 추세, 계절성, 특이점, 변화점 등을 파악

예측
- 시계열 분석의 가장 중요한 목적
- 특정 데이터가 과거 변화 패턴과 비슷하게 현재 시점 이후에도 변화할 것이라는 전제 하에, 미래의 값을 예측

설명
- 특정 시스템의 시계열적 패턴을 설명하기 위한 과학적 노력
- Ex. 경제 시계열 분석을 통해 경기 주기가 존재함을 입증

제어
- 특정 시스템에 인위적인 조작을 가해 원하는 목표에 부응하도록 유도
- Ex. 극심한 경기 변동이 우려되는 경우, 정부가 중앙은행을 통해 정책적으로 개입

Contents

평활법 (Smoothing Method)

- 평활법 : 과거 및 현재 데이터의 불규칙 변동을 부드럽게 평활(smoothing)하여 미래 값 예측

- 평활법의 종류
  - 목측법(Eye-measurement) : 눈짐작으로 선 긋기. 정확성은 낮으나 신속한 경향 파악 가능.
  - 이동평균법(Moving Average Method) : 최근 특정 기간의 관측치 몇 개를 산술평균한 값을 이용해 미래 시점의 예측치로 사용
  - 지수평활법(Exponential Smoothing) : 모든 데이터를 활용하되, 최근 관측치에 더 큰 가중치를 준 가중평균 값을 미래 시점의 예측치로 사용
  - 최소제곱법 : 편차의 제곱합을 최소로 하는 경향선을 구해 미래 값 예측 (회귀분석)

이동평균법 (1) – 단순 이동평균법

- 이동평균법 : 최근 특정 기간의 관측치 몇 개를 산술평균한 값을 이용해 미래 시점의 예측치로 사용
- 단순 이동평균법 : 가장 최근에 얻어진 관측값 m개의 산술평균값을 이용
  - 시계열 $\{Z_n\}$에서, 시점 t 에서의 관측값을 Z_t 라 할 때, 시점 n 에서의 단순 이동평균

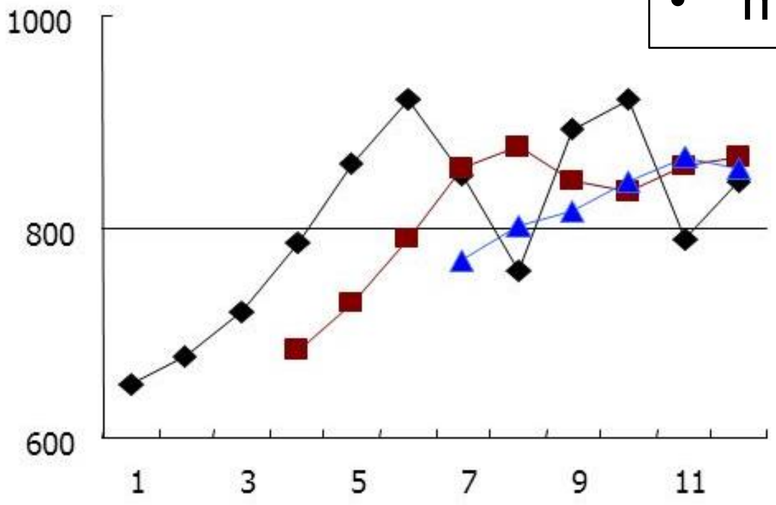
    : $MA_n = \frac{Z_n + Z_{n-1} + \cdots + Z_{n-m+1}}{m} = \frac{1}{m}\sum_{t=n-m+1}^{n} Z_t$
  - 시점 n 에서 추정한 시점 $n+1$ 의 예측값을 F_{n+1} 이라고 하면, $F_{n+1} = MA_n$

이동평균법 (1) – 단순 이동평균법

Week	Demand	3-week MA	6-week MA
1	650		
2	678		
3	720		
4	785	683	
5	859	728	
6	920	788	
7	850	855	769
8	758	876	802
9	892	843	815
10	920	833	844
11	789	857	867
12	844	867	855

= (650+678+720)/3

= (650+678+720+785+859+920)/6



- 데이터 변동이 클수록, 최근 데이터의 경향을 많이 반영해야 함
  → m이 작아야 함
- 데이터의 변동이 완만하다면,
  → m이 커도 됨
- m의 선택 : MSE를 최소로 하는 m

이동평균법 (2) – 이중(선형) 이동평균법

- 선형 추세가 있는 시계열에 단순 이동평균법을 적용하면, 추세를 늦게 따라가는 문제 발생
- 이중(선형) 이동평균법 : 이동평균을 두 번 적용한 방법

- 시계열 $\{Z_n\}$에서, 시점 t 에서의 관측값을 Z_t 라 할 때, 시점 n 에서의 단순 이동평균

$$: MA_n = \frac{Z_n + Z_{n-1} + \cdots + Z_{n-m+1}}{m} = \frac{1}{m}\sum_{t=n-m+1}^{n} Z_t$$

- 시계열 $\{Z_n\}$에서, 시점 t 에서의 관측값을 Z_t 라 할 때, 시점 n 에서의 이중(선형) 이동평균

$$: MA'_n = \frac{1}{m}(MA_n + MA_{n-1} + \cdots + MA_{n-m+1})$$

이동평균법 (2) – 이중(선형) 이동평균법

시계열 $\{Z_t\}$이 다음과 같은 선형추세를 갖는다고 할 때

$$Z_t = a + bt + e_t$$

- 시점 t 에서의 한 단계 이후 $t+1$ 시점의 예측치

$$F_{t+1} = E[Z_{t+1} \backslash Z_t, Z_{t-1}, \cdots, Z_1] = a + b(\mathrm{t}+1)$$

$$\widehat{F_{t+1}} = \hat{a} + \hat{b}(t+1) = 2MA_t\text{-}MA'_t + 1\hat{b}, \qquad \hat{b} = \frac{2}{m-1}(MA_t\text{-}MA'_t)$$

- 시점 t 에서의 l 단계 이후 예측치

$$\widehat{F_{t+l}} = \hat{a} + \hat{b}(t+l) = (2MA_t\text{-}MA'_t) + \hat{b}l, \qquad \hat{b} = \frac{2}{m-1}(MA_t - MA'_t)$$

참고) 이동평균 계산

- 시계열 $\{Z_n\}$에서, 시점 t 에서의 관측값을 Z_t 라 할 때, 시점 n 에서의 단순 이동평균

  $: MA_n = \frac{Z_n + Z_{n-1} + \cdots + Z_{n-m+1}}{m} = \frac{1}{m}\sum_{t=n-m+1}^{n} Z_t$

- 시계열 $\{Z_n\}$에서, 시점 t 에서의 관측값을 Z_t 라 할 때, 시점 n 에서의 이중(선형) 이동평균

  $: MA'_n = \frac{1}{m}(MA_n + MA_{n-1} + \cdots + MA_{n-m+1})$

이동평균법 (2) – 이중(선형) 이동평균법

표 3.4 가상자료에 대한 선형이동평균 예측

t	원계열(Z_t)	이동평균(MA_t)	이중이동평균(MA_t')	예측값(F_t)	예측오차(E_t)
1	10	–	–	–	–
2	18	–	–	–	–
3	20	16	–	–	–
4	28	22	–	–	–
5	30	26	21.33	–	–
6	38	32	26.67	35.34	2.66
7	37	35	31.00	42.66	−5.66

$$F_{5+1} = (2MA_5 - MA_5') + \frac{2}{3-1}(MA_5 - MA_5') \cdot 1 = 2 \times 26 - 21.33 + \frac{2}{3-1}(26 - 21.33) \cdot 1 = 35.34$$

$$F_{6+1} = (2MA_6 - MA_6') + \frac{2}{3-1}(MA_6 - MA_6') \cdot 1 = 2 \times 32 - 26.67 + \frac{2}{3-1}(32 - 26.67) \cdot 1 = 42.66$$

지수평활법 (1) – 단순 지수평활법

- 지수평활법 : 모든 데이터 활용, 최근 관측치에 더 큰 가중치를 준 가중평균 값을 미래 시점의 예측치로 사용
- 시계열 구조 변화에 유연하게 대응 가능, 실제 활용 시 정확도가 높아 단기 예측에 많이 사용

- 단순 지수평활법 : 추세나 계절성이 없는, 수평적 패턴의 데이터에 유용
  - 시계열 $\{Z_n\}$의 시점 n 에서의 지수평활치

    : $S_n = \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + \alpha(1-\alpha)^2 Z_{n-2} + \cdots$
  - 시계열 $\{Z_n\}$의 시점 $n+1$ 에서의 지수평활치 :

    : $S_{n+1} = \alpha Z_{n+1} + (1-\alpha)S_n$

Notation

Z_n : 시점 n 에서의 관측값
α : 지수평활상수 (Smoothing Constant)

지수평활법 (1) – 단순 지수평활법

- 지수평활상수 α 선택
  - MSE를 최소로 하는 α (일반적으로, 0.05 ~ 1)
  - 1에 가까울수록 최근 추세 많이 반영, 변화에 민감하게 반응
  - 0에 가까울수록 과거 데이터에도 영향 많이 받고, 전체 평균으로 예측

- 초기 평활값 S_0 선택
  - Brown(1962), Montgomery & Johnson(1976) : $\bar{Z} = \frac{1}{n}\sum_{t=1}^{n} Z_t \ \ (n = 6 \ \text{or} \ n = \frac{T}{2})$
  - Makridakis & Wheelwright(1978) : Z_1

지수평활법 (2) – 이중 지수평활법

- 선형 추세가 있는 시계열에 단순 지수평활법을 적용하면, 추세를 늦게 따라가는 문제 발생

- 이중 지수평활법 : 지수평활을 두 번 적용, 시계열이 선형 추세가 있는 경우에 적합
  - 지수평활값과 이중지수평활값:

  : $\mathrm{SM}_{\mathrm{n}} = \alpha \mathrm{Z}_{\mathrm{n}} + (1-\alpha)\mathrm{SM}_{\mathrm{n}-1}$ (지수평활값)

  : $\mathrm{SM'}_{\mathrm{n}} = \alpha \mathrm{SM}_{\mathrm{n}} + (1-\alpha)\mathrm{SM'}_{\mathrm{n}-1}$ (이중지수평활값)

지수평활법 (2) – 이중 지수평활법

- 시점 t 에서의 한 단계 이후 $t+1$ 시점의 예측치

$$F_{t+1} = E[Z_{t+1} \backslash Z_t, Z_{t-1}, \cdots, Z_1] = a + b(\mathrm{t}+1)$$

$$\widehat{F_{t+1}} = \hat{a} + \hat{b}(t+1) = (2SM_t\text{-}SM_t')+1\hat{b}, \qquad \hat{b} = \frac{\alpha}{1-\alpha}(SM_t\text{-}SM_t')$$

- 시점 t 에서의 l 단계 이후 예측치

$$\widehat{F_{t+l}} = \hat{a} + \hat{b}(t+l) = (2SM_t\text{-}SM_t')+\hat{b}l$$

지수평활법 (2) – 이중 지수평활법

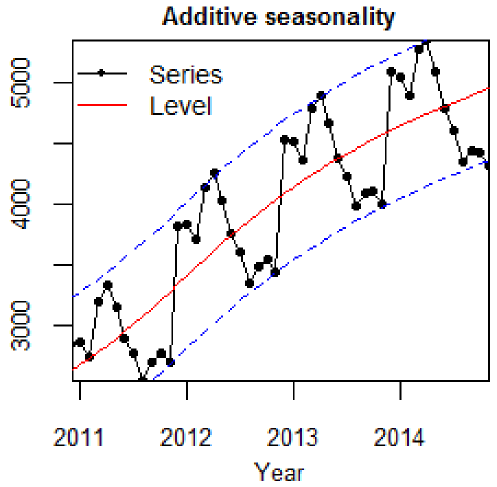
- 지수평활상수 $\alpha = 0.2$, 초기값 $\mathrm{SM}_0 = \mathrm{Z}_1 = 10$ 일 때, 이중 지수평활법 기반 예측값

표 3.8 가상자료에 대한 브라운의 선형지수평활예측값

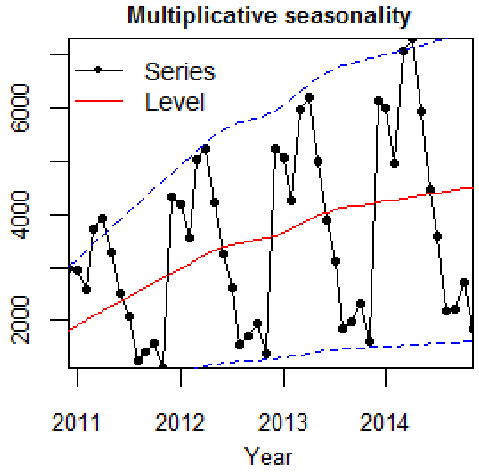
t	원계열(Z_t)	지수평활값 (SM_t)	이중지수평활값 (SM_t)	예측값(F_t)
1	10	10.000	10.000	$F_{1+1}=2\times10-10+\frac{0.2}{0.8}\times(10-10)=10$
2	18	11.600	10.320	$F_{2+1}=2\times11.60-10.32+\frac{0.2}{0.8}\times(11.60-10.32)=13.2$
3	20	13.280	10.912	$F_{3+1}=2\times13.28-10.912+\frac{0.2}{0.8}\times(13.28-10.912)=16.24$
4	28	16.224	11.974	$F_{4+1}=2\times16.224-11.794+\frac{0.2}{0.8}\times(16.224-11.794)=21.53$
5	30	18.379	13.255	$F_{5+1}=2\times18.379-13.255+\frac{0.2}{0.8}\times(18.379-13.255)=24.78$
6	38	22.303	15.065	$F_{6+1}=2\times22.303-15.065+\frac{0.2}{0.8}\times(22.303-15.065)=31.35$
7	37	25.242	17.101	

지수평활법 (3) – Holt-Winters 계절 지수평활법

- 월별, 분기별 자료 : 계절변동(추세+계절성+수준(자료변동)) 존재
- 계절변동의 두 가지 형태
  - 가법적 계절변동 : 시간에 따른 계절적 진폭의 크기 일정
  - 승법적 계절변동 : 계절적 진폭의 크기에 평균 수준에 비례하여 점차적으로 증가 or 감소



가법적 계절변동



승법적 계절변동

지수평활법 (3) – Holt-Winters 계절 지수평활법

- 가법적 계절 지수평활법 : 가법적 계절변동이 있을 때, 가법적으로 각 성분(변동) 갱신
- 주기가 p 인 시계열 $\{y_t\}$ 가 있다고 할 때, 갱신 알고리즘은 아래와 같다.

$$\hat{y}_{t+h|t} = \hat{\mu}_t + h \cdot b_t + s_{t+h-p},\ h \le p$$

(p : 계절성의 주기)

여기서 μ_t, b_t, s_t는 시점 t에서 추정된 수준, 기울기, 계절효과를 의미한다.

$\cdot\ \hat{\mu}_t = \alpha(y_t - s_{t-p}) + (1-\alpha)(\hat{\mu}_{t-1} + b_{t-1})$

$\cdot\ b_t = \beta(\hat{\mu}_t - \hat{\mu}_{t-1}) + (1-\beta)b_{t-1}$

$\cdot\ s_t = \gamma(y_t - \hat{\mu}_t) + (1-\gamma)s_{t-p}$

$\cdot\ \alpha, \beta, \gamma$: 전체, 추세, 계절조정을 위한 평활모수

출처 : https://be-favorite.tistory.com/62

지수평활법 (3) – Holt-Winters 계절 지수평활법

- 승법적 계절 지수평활법 : 승법적 계절변동이 있을 때, 승법적으로 각 성분(변동) 갱신
- 주기가 p 인 시계열 $\{y_t\}$ 가 있다고 할 때, 갱신 알고리즘은 아래와 같다.

$$\hat{y}_{t+h|t} = (\hat{\mu}_t + h \cdot b_t) \times s_{t+h-p},\ h \le p$$

(p : 계절성의 주기)

여기서 μ_t, b_t, s_t의 의미는 위와 같으나, 수준과 계절효과의 세부적인 식이 다르다.

$$\cdot\ \hat{\mu}_t = \alpha\left(\frac{y_t}{s_{t-p}}\right) + (1-\alpha)(\hat{\mu}_{t-1} + b_{t-1})$$

$$\cdot\ b_t = \beta(\hat{\mu}_t - \hat{\mu}_{t-1}) + (1-\beta)b_{t-1}$$

$$\cdot\ s_t = \gamma\left(\frac{y_t}{\hat{\mu}_t}\right) + (1-\gamma)s_{t-p}$$

$\cdot\ \alpha, \beta, \gamma$: 전체, 추세, 계절조정을 위한 평활모수

출처 : https://be-favorite.tistory.com/62

지수평활법 (3) – Holt-Winters 계절 지수평활법

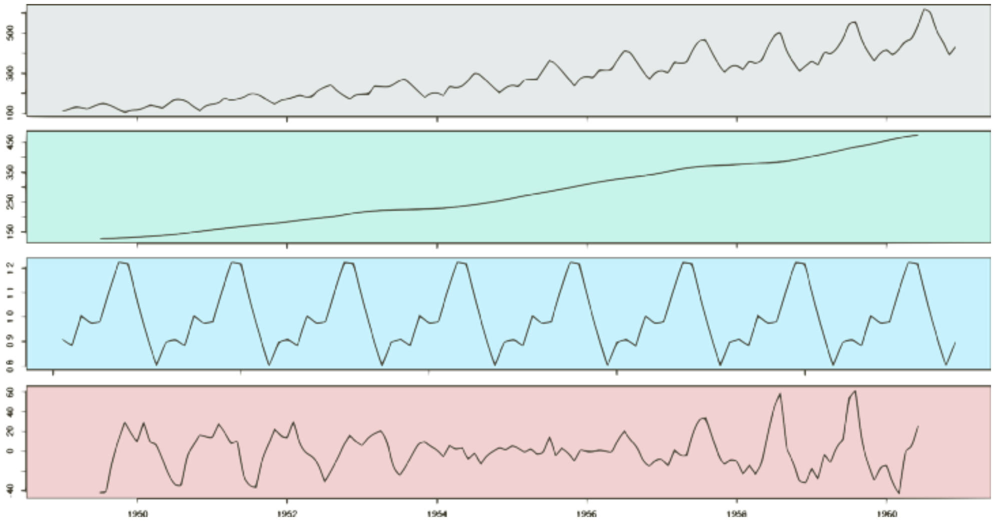
- 시계열 $\{Z_n\}$ 에서의 시점이 n 인 경우, l 시점 후의 예측값 F_{n+l}

  : $F_{n+l} = (\mu_n + b_n \cdot l) \cdot s_{n+l-p}$

- 마찬가지로, 평활상수는 MSE를 최소화하는 값으로 선택한다.

분해법 (Decomposition Method)

- 분해법 : 시계열을 개별 변동성분(추세, 계절성, 순환성)으로 각각 분해한 후, 분해된 각 성분을 개별적으로 예측 → 이후 이를 다시 결합하여 예측
  - 오래된 예측법 중 하나, 최근에는 경제·경영 분야에서 다양하게 응용
  - 장점 : 직관적으로 이해하고 구현하기에 쉬움, 분해를 통해 변동의 원인 파악 가능
  - 단점 : 시계열에 이상점이 있는 경우 부적절



대칭 이동평균 & 중심화 이동평균

- 시점 t 에서의 관측값을 Z_t 라 하고, 이동평균을 MA_t 라 하면,
  - 대칭 이동평균 (m이 홀수) : 해당 시점과 인접한 전후 시점들의 평균

    : $MA_t = \frac{1}{m}\sum_{j=-(m-1)/2}^{(m-1)/2} Z_{t+j}$

  - 중심화 이동평균 (m이 짝수) : 짝수 기간의 이동평균은 두 시점 사이의 값에 대응 → 재차 이동평균

    : $MA_t = \frac{1}{2}\left(MA_{t-\frac{1}{2}} + MA_{t+\frac{1}{2}}\right)$

    $MA_{t-1/2} = \frac{1}{m}\sum_{j=-m/2}^{m/2-1} Z_{t+j}$ $MA_{t+1/2} = \frac{1}{m}\sum_{j=-m/2+1}^{m/2} Z_{t+j}$

대칭 이동평균 & 중심화 이동평균

가상적 자료에 대한 이동평균

출처: 시계열분석과 예측, 이누리 저, 제 2판 p77

기간	관측값$\{Z_t\}$	$3MA$	관측값$\{Z_t\}$	$4MA$	$2\times4MA$
1	20	—	20		–
2	10	20 MA_2	10	–	–
3	30	25 MA_3	30	23.75 $MA_{2.5}$	22.500 MA_3
4	35	25	35	21.25 $MA_{3.5}$	27.500
5	10	35	10	33.75	36.250
6	60	40	60	38.75	39.375
7	50	50	50	40.00	50.000
8	40	60	40	60.00	60.625
9	90	65	90	61.25	61.875
10	65	70	65	62.50	72.500
11	55	80	55	82.50	–
12	120	—	120	–	

대칭 이동평균

중심화 이동평균

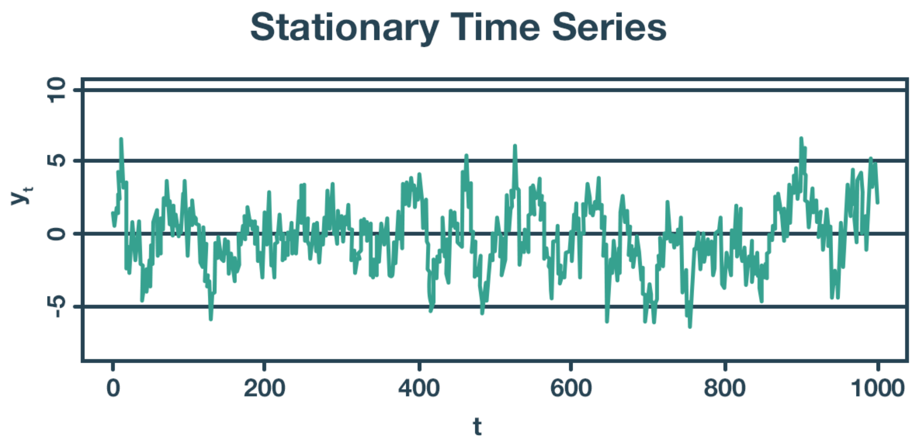
원 계열 – 추세 = 계절성 + 불규칙 성분

분해법에 의한 예측 절차 (가법 모형 전제)

1. 추세 및 추세 조정된 계열의 추정
   - ✓ 원 계열에 추세선을 적합시켜 추세를 추정 → 이 추세를 원 시계열에서 빼서 추세 조정된 계열 얻기
2. 계절성 및 계절 조정된 계열의 추정
   - ✓ 추세 조정된 계열에 대하여 계절성의 길이만큼 이동평균 → 계절성 제거 (불규칙 변동만이 남음)
   - ✓ 다시 추세 조정된 계열에서 추정된 불규칙 변동 제거 → 시점별 계절성 추정
   - ✓ 이렇게 추정된 계절성별로 계절별 평균을 구하고, 계절평균의 합이 0이 되도록 조정 → 계절지수
3. 예측
   - ✓ 분해된 성분들을 개별적으로 예측하여 원 계열의 미래값 예측

Contents

정상 시계열 (Stationary Time Series)

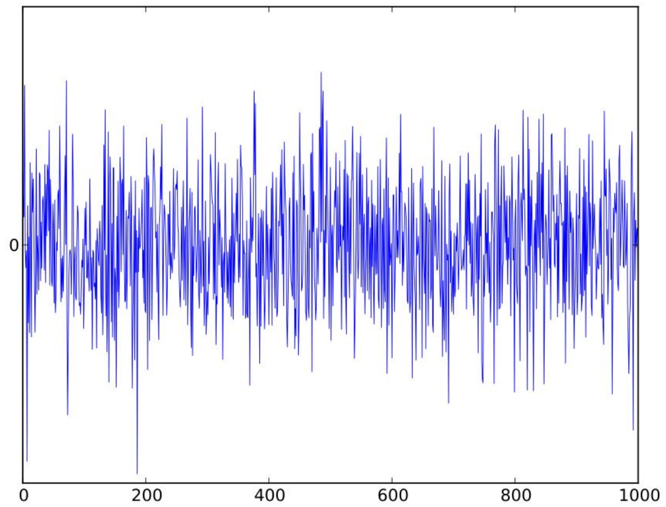


- 정상 시계열 : 시간에 관계없이 평균과 분산이 일정한 시계열
  - 모든 t 에 대해서 $E(Z_t) = \mu$, $Var(Z_t) = \sigma^2$
  - 두 시점 사이의 자기공분산은 시점 t 와 무관하며, 시차(time lag) k 에만 의존 → 시간 불변성

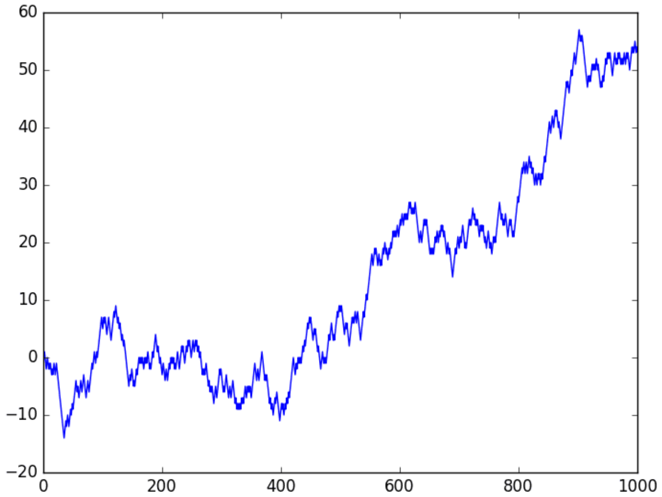
- 정상 시계열의 형태
  - 추세, 계절성, 순환성 등의 패턴이 보이지 않음
  - 자료 변화의 폭이 일정함
  - 시간에 따라 달라지는 자기상관적 패턴을 나타내는 구간이 없음

정상 시계열, 비정상 시계열 예시

- 정상시계열 : 백색잡음과정 (White Noise)
  - ✓ 서로 독립이면서 동일한 분포를 따르는(i.i.d.) 확률변수들의 계열로 구성
  - ✓ $\{\varepsilon_t\}$: $\mathrm{NW}(0,\ \sigma_\varepsilon^2)$, $Z_t = \varepsilon_t$



- 비정상시계열 : 확률보행과정 (Random Walk)
  - ✓ $\{\varepsilon_t\}$: $\mathrm{NW}(0,\ \sigma_\varepsilon^2)$, $Z_t = Z_{t-1} + \varepsilon_t$
  - ✓ 평균은 0으로 일정하나, 분산이 시간에 종속



정상 시계열 여부 판단

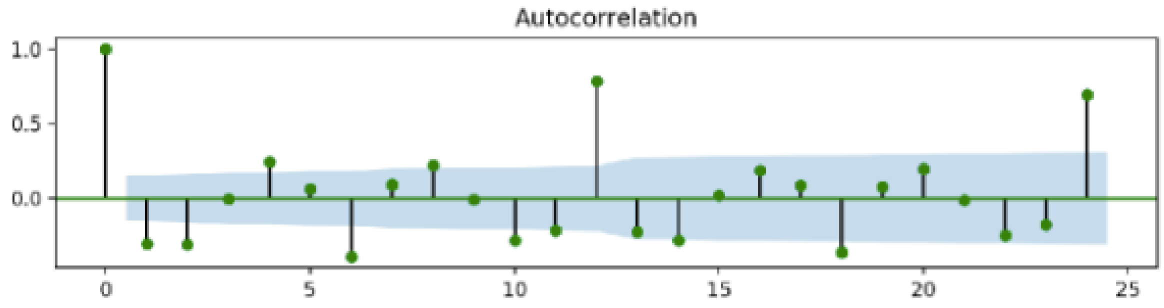
- 판단 기준
  - ✓ X축을 lag(현재 데이터와의 시점 차)로, Y축을 ACF(Autocorrelation Function)로 시각화
  - ✓ 특정 패턴이 없다면 정상 시계열, 있다면 비정상 시계열

- ✓ [참고] Autocorrelation Function (ACF, 자기상관함수)
  - ✓ Correlation(두 변수 간 관계를 -1~1 값으로 표현하는 척도)에 Auto 개념이 추가된 것
  - ✓ 시계열적 관점에서, time shifted된 자기 자신과의 correlation

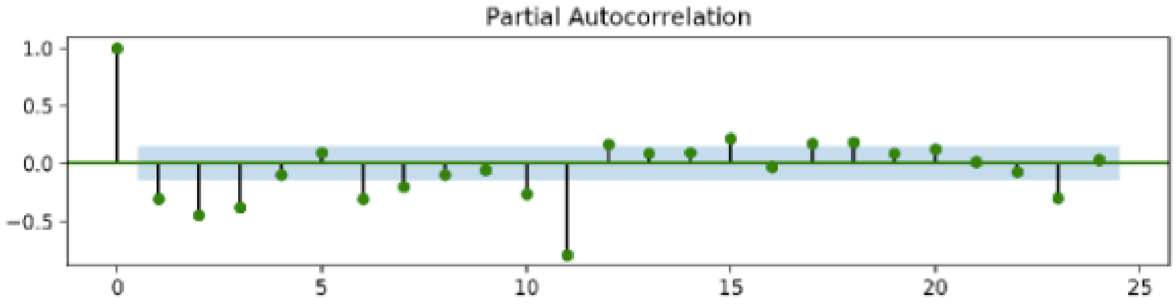
ACF & PACF

- ACF (자기상관함수)
  - ✓ 시차 k 에 대해서, 관측치 간 상관계수 함수

$$ACF(k) = \frac{\sum_{t=1}^{N-k} (y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{N} (y_t - \bar{y})^2}$$



- PACF (Partial ACF, 편자기상관함수)
  - ✓ 시차 k 에 대해서, $\mathrm{Z_t}$와 $\mathrm{Z_{t+k}}$ 사이에 있는 관측치와의 상호 의존성을 제거한 함수
  - ✓ 순수한 상관관계 파악 가능



Box-Jenkins 세 가지 모형

- 자기회귀모형 (Auto Regressive) : AR(p) 모형

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + \cdots + \phi_p Z_{t-p} + a_t, \qquad a_t\text{는 } iid\text{인 } white\ noise$$

- 이동평균모형 (Moving Average) : MA(q) 모형

$$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \theta_3 a_{t-3} - \cdots - \theta_q a_{t-q}, \qquad a_t\text{는 } iid\text{인 } white\ noise$$

- 자기회귀이동평균 모형 (Auto Regressive Moving Average) : ARMA(p, q) 모형

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + \cdots + \phi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \theta_3 a_{t-3} - \cdots - \theta_q a_{t-q}$$

Box-Jenkins 모형 (1) – AR(p) 모형

자기회귀모형 (Auto Regressive) : AR(p) 모형

$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + \cdots + \phi_p Z_{t-p} + a_t,\quad a_t$는 iid인 $white\ noise$

- 자기 자신을 종속변수로 하고, 그 이전 시점의 시계열을 독립변수로 하는 모형
  - 즉, 시계열 Z_t 를 그 이전 시점의 계열로 회귀
  - p차 AR 모형은 Z_t 가 $Z_{t-1}, Z_{t-2}, \cdots, Z_{t-p}$ 의 선형결합임을 가정, 항상 가역성 만족
- p는 하이퍼파라미터 (ACF, PACF 확인 후 결정)

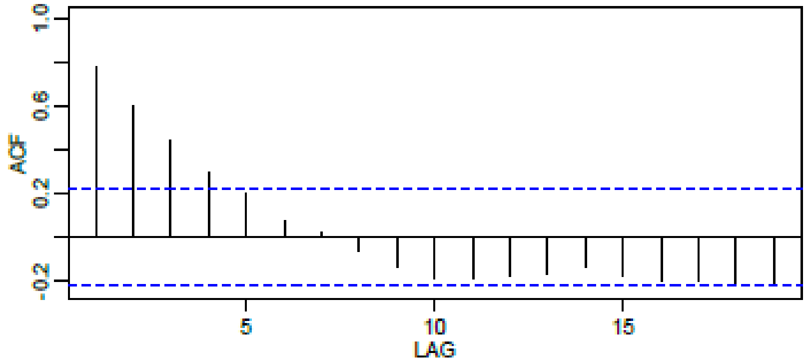
- 회귀모형과 달리,
  - ✓ 독립변수 간 독립성 가정 없음
  - ✓ 독립변수 개수가 모형의 추정 과정에서 결정

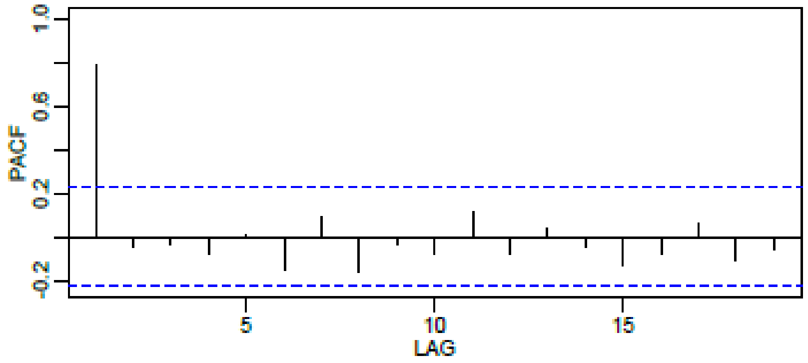
AR(p) 예시 : AR(1), AR(2)

자기회귀모형 (Auto Regressive) : AR(p) 모형

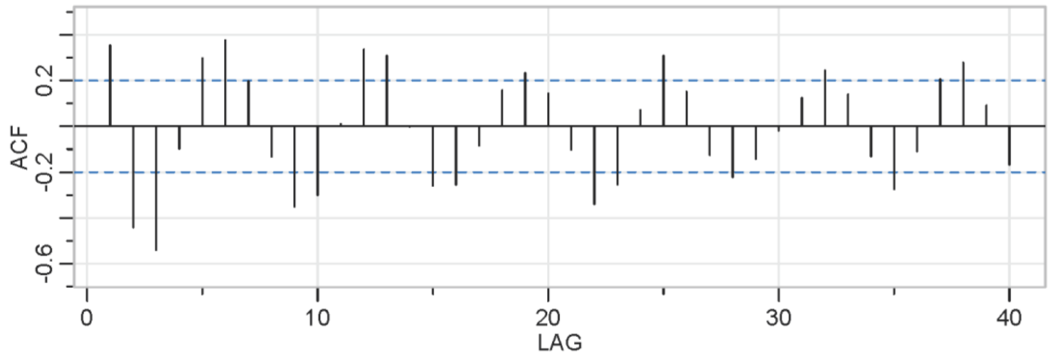
$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + \cdots + \phi_p Z_{t-p} + a_t,\quad a_t$는 iid인 $white\ noise$

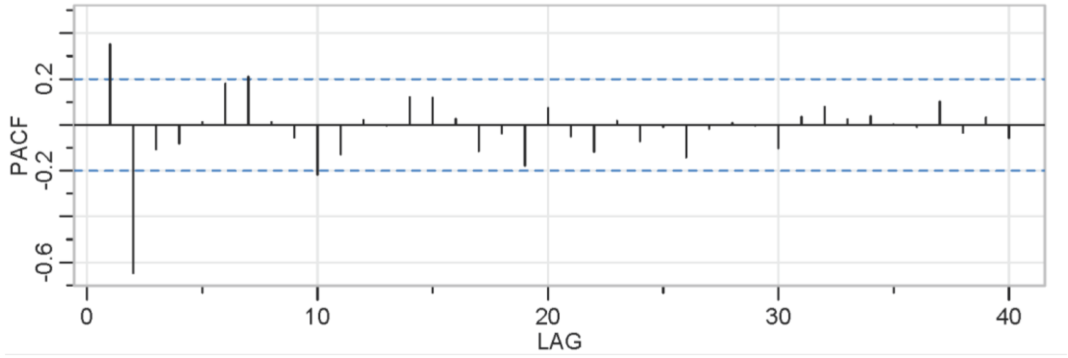
AR(1) : $Z_t = \emptyset_1 Z_{t-1} + a_t$





AR(2) : $Z_t = \emptyset_1 Z_{t-1} + \emptyset_2 Z_{t-2} + a_t$





Box-Jenkins 모형 (2) – MA(q) 모형

이동평균모형 (Moving Average) : MA(q) 모형

$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \theta_3 a_{t-3} - \cdots - \theta_q a_{t-q},$ a_t는 iid인 $white\ noise$

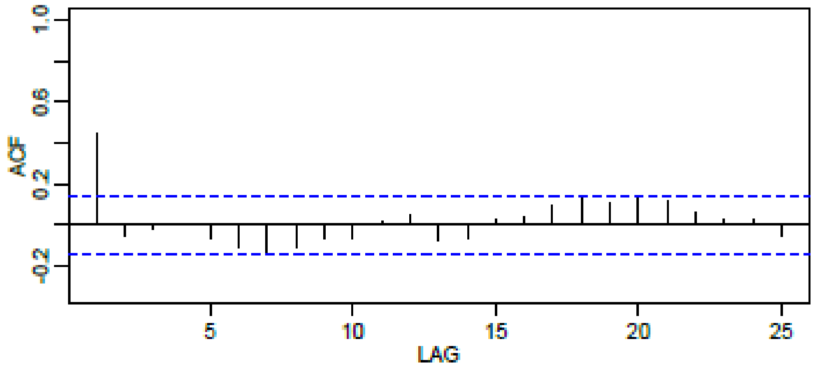
- 자기 자신을 종속변수로 하고, 그 이전 시점의 white noise process를 독립변수로 하는 모형
  - 즉, 시계열 Z_t 를 그 이전 시점의 백색잡음 계열$\{a_t, a_{t-1}, a_{t-2}, \cdots\}$로 회귀
  - q차 MA 모형은 항상 정상성 만족
- q는 하이퍼파라미터 (ACF, PACF 확인 후 결정)

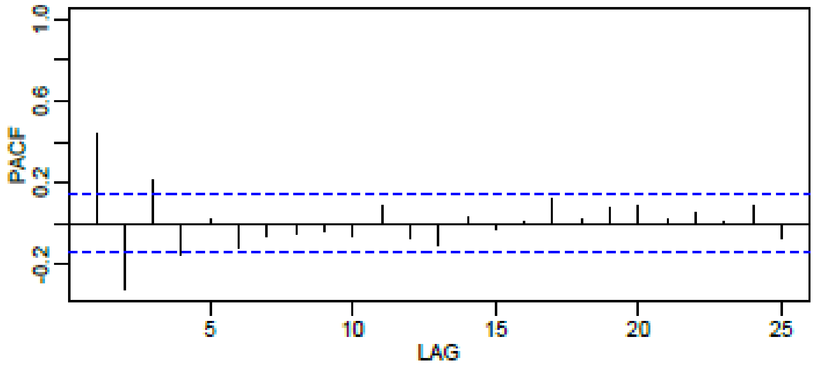
MA(q) 예시 : MA(1), MA(2)

이동평균모형 (Moving Average) : MA(q) 모형

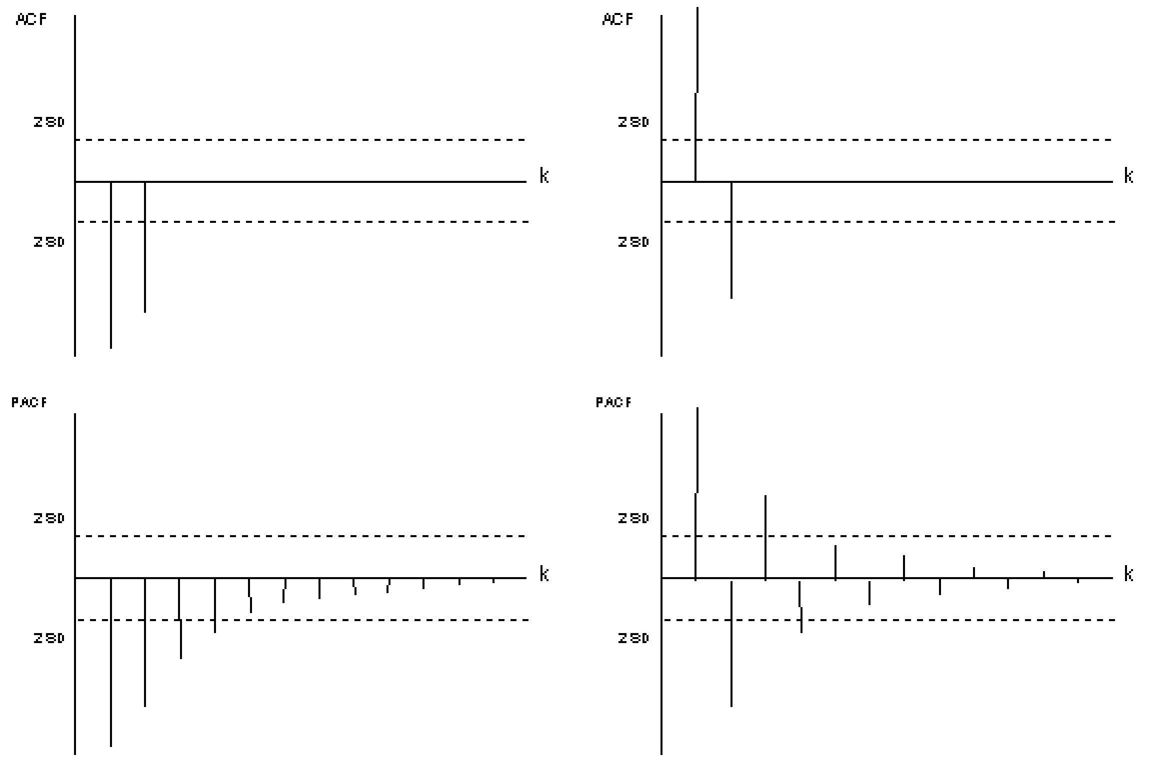
$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \theta_3 a_{t-3} - \cdots - \theta_q a_{t-q},\quad a_t$는 iid인 $white\ noise$

MA(1) : $Z_t = a_t - \theta_1 a_{t-1}$





MA(2) : $Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$



AR(p) 모형, MA(q) 모형 비교

- AR(p), MA(q) 모형의 Theoretical ACF, PACF 특징

	AR(p)	MA(q)
ACF	Die out (지수적 감소, 소멸하는 sine 함수 형태)	Cut off after lag q (q+1 시점부터 0으로 절단)
PACF	Cut off after lag p (p+1 시점부터 0으로 절단)	Die out (지수적 감소, 소멸하는 sine 함수 형태)

Box-Jenkins 모형 (3) – ARMA(p, q) 모형

자기회귀이동평균 모형 (Auto Regressive Moving Average) : ARMA(p, q) 모형

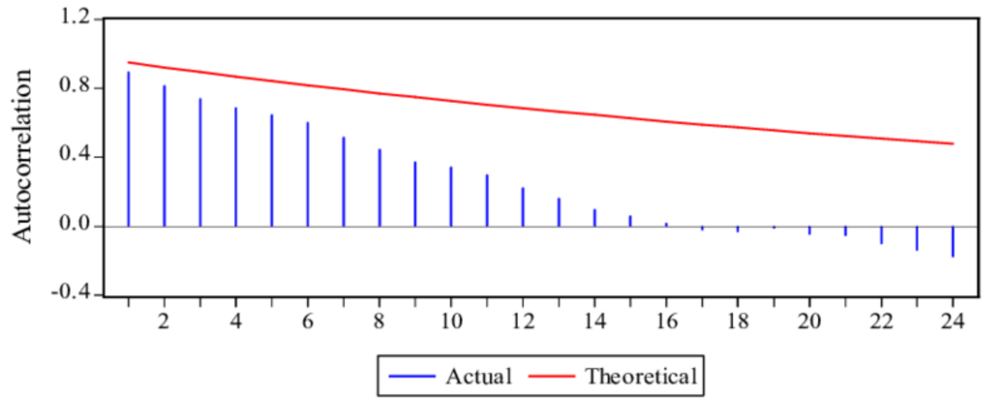
$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + \cdots + \phi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \theta_3 a_{t-3} - \cdots - \theta_q a_{t-q}$$

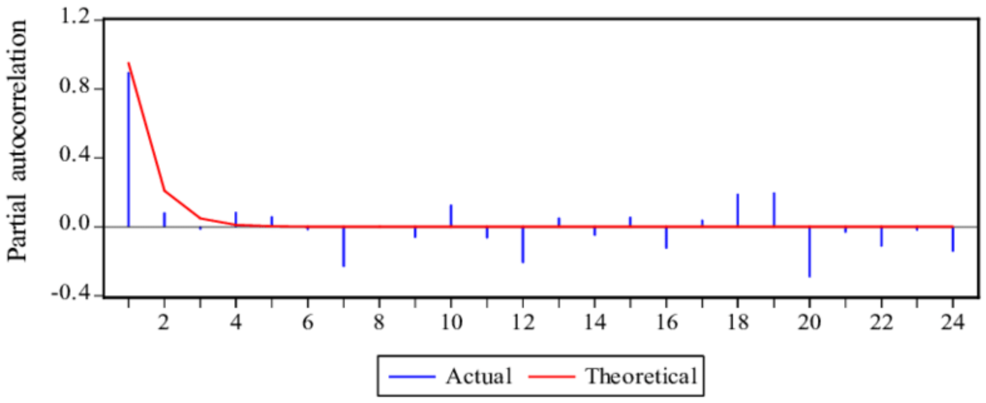
- ARMA(p, q) 모형은 AR(p) 모형과 MA(q) 모형을 결합시킨 모형
- 정상성과 가역성을 만족하기 위해서는,
  - ✓ AR 부분 특성 방정식의 모든 근의 절댓값이 1보다 커야 함 → 정상성 만족
  - ✓ MA 부분 특성 방정식의 모든 근의 절댓값이 1보다 커야 함 → 가역성 만족
- p, q는 하이퍼파라미터 (ACF, PACF 확인 후 결정)

ARMA(p, q) 모형 예시 : ARMA(1,1)

자기회귀이동평균 모형 (Auto Regressive Moving Average) : ARMA(p, q) 모형

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + \cdots + \phi_p Z_{t-p}$$
$$+ a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \theta_3 a_{t-3} - \cdots - \theta_q a_{t-q}$$





- ARMA(1,1) 모형의 ACF와 PACF
  - ACF : AR 모형의 ACF처럼 Die out(지수적 감소)
  - PACF : MA 모형의 PACF처럼 Die out(지수적 감소)

AR(p) 모형, MA(q), ARMA(p, q) 모형 비교

- AR(p), MA(q), ARMA(p, q) 모형의 Theoretical ACF, PACF 특징

	AR(p)	MA(q)	ARMA(p, q)
ACF	Die out (지수적 감소, 소멸하는 sine 함수 형태)	Cut off after lag q (q+1 시점부터 0으로 절단)	Die out (q-p+1 시점부터 지수적 감소, 소멸하는 sine 함수 형태)
PACF	Cut off after lag p (p+1 시점부터 0으로 절단)	Die out (지수적 감소, 소멸하는 sine 함수 형태)	Die out (p-q+1 시점부터 지수적 감소, 소멸하는 sine 함수 형태)

비정상 시계열 – ARIMA(p, d, q) 모형

- 추세나 계절성이 포함된 비정상 시계열의 경우, 해당 시계열을 차분을 통해 정상화할 수 있음
  - 즉, 차분을 통해 추세나 계절성을 제거(감소)할 수 있음

- 참고) 비정상성 판단
  - ✓ Augmented Dickey-Fuller Test
  - ✓ 검정통계량 값(ADF)이 임계값보다 작으면(p-value가 신뢰수준보다 작으면) stationary
  - ✓ [Python] from statsmodels.tsa.stattools import adfuller

비정상 시계열 – ARIMA(p, d, q) 모형

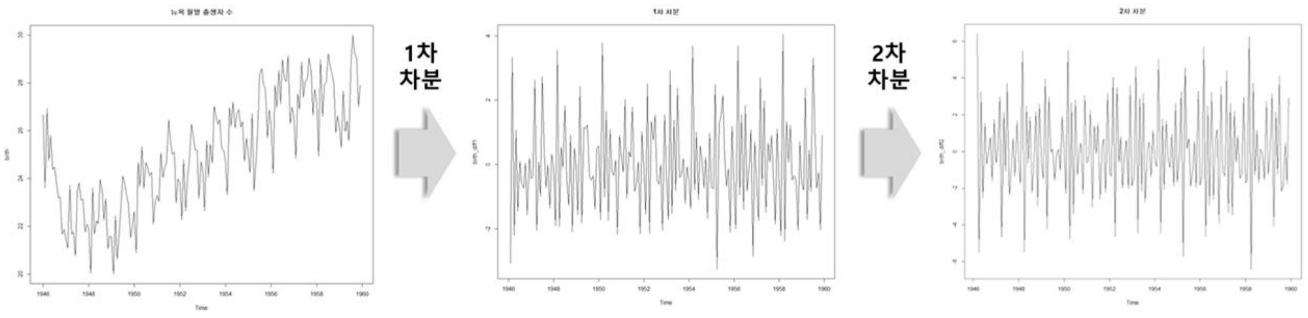
- 차분(Differencing) : 현 시점 데이터에서, d 시점 이전 데이터를 뺀 것
- 차분을 통해 시계열을 정상화할 수 있는데, 과대차분은 ACF를 복잡하게 만들거나 분산을 크게 함

X		
2	X	Y
7	2	5
10	7	3
5	10	-5
8	5	3
	8	-

1차 차분: $Y_t = X_t - X_{t-1} = \nabla X_t$

X		
2		
7	X	Y
10	2	8
5	7	-2
8	10	-2
	5	-
	8	-

2차 차분: $Y_t^{(2)} = X_t - X_{t-2} = \nabla^{(2)} X_t$



비정상 시계열 – ARIMA(p, d, q) 모형

- ARIMA(p, d, q) : 누적(Integrataed) 자기회귀이동평균 모형
  - ✓ d차 차분으로 변환된 시계열이 ARMA(p, q)를 따르면, 원 시계열은 ARIMA(p, d, q)를 따른다.
- p, d, q는 하이퍼파라미터

- 참고) 계절성 시계열 데이터
  - ✓ 계절성이 있는 시계열 데이터는 비정상
  - ✓ 단순 차분을 반복하여 적용함으로써 정상 시계열로 전환되기도 하나, 과대차분의 위험
  - ✓ 일반적으로 계절 차분이 효과적 → $Z_t \sim \mathrm{SARIMA}(p, d, q) \times (P, D, Q)_S$

튜토리얼 : https://hyperconnect.github.io/2020/03/09/prophet-package.html

참고) 시계열 예측 패키지 Prophet

- Prophet : 페이스북에서 공개한 시계열 예측 라이브러리
  - ✓ 정확도가 높고 빠름
  - ✓ 직관적인 파라미터 → 모델 수정에 용이

- 모델의 주요 구성 요소 : Trend, Seasonality, Holiday
  - $y(t) = g(t) + s(t) + h(t) + \epsilon_i$
    - $g(t)$: Trend. 부분적으로 선형 or 로지스틱 곡선으로 구성.
    - $s(t)$: Seasonality. Trend와 달리 주기적으로 나타나는 패턴 포함.
    - $h(t)$: Holiday. 휴일과 같이 불규칙한 이벤트 나타냄.

Contents

Box-Jenkins 방법론

모형 식별

- 시계열 자료를 보고 잠정적인 ARIMA 모형 찾기
- 모수 절약의 원칙 : 가급적 모수는 적게! (AIC, BIC)

모수 추정

- 최소제곱법 (AR 모형), 비선형 최소제곱법
- 최대우도추정법

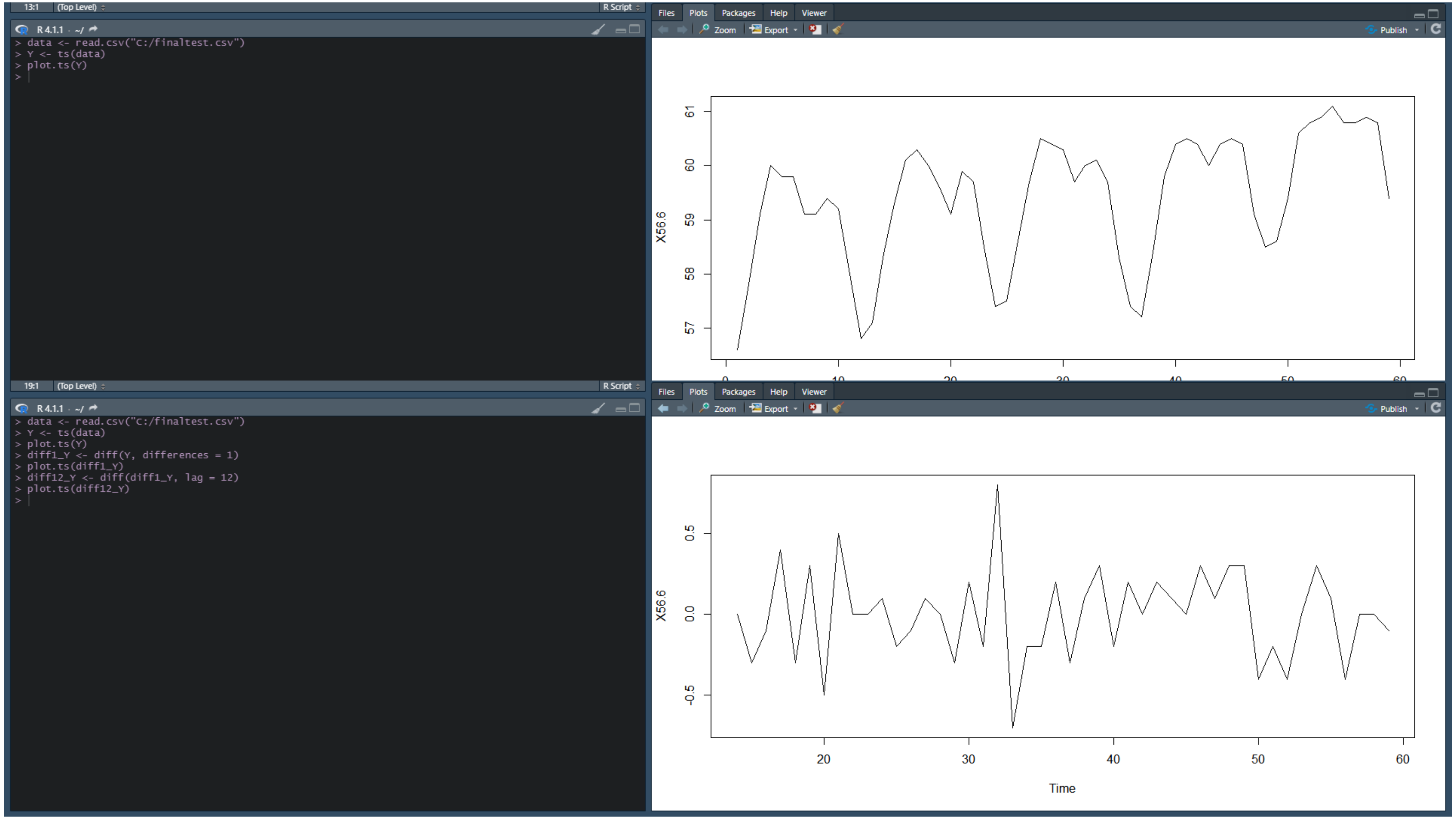
모형 적합성 진단

- 모형의 오차항에 대한 가정 검토
- 잔차의 정규성, 등분산성 및 패턴 유무, 랜덤성 (포트만토 검정)

모형 확정 및 예측

- 예측값 구하기

Box-Jenkins 방법론 실습 - 1. 모형 식별

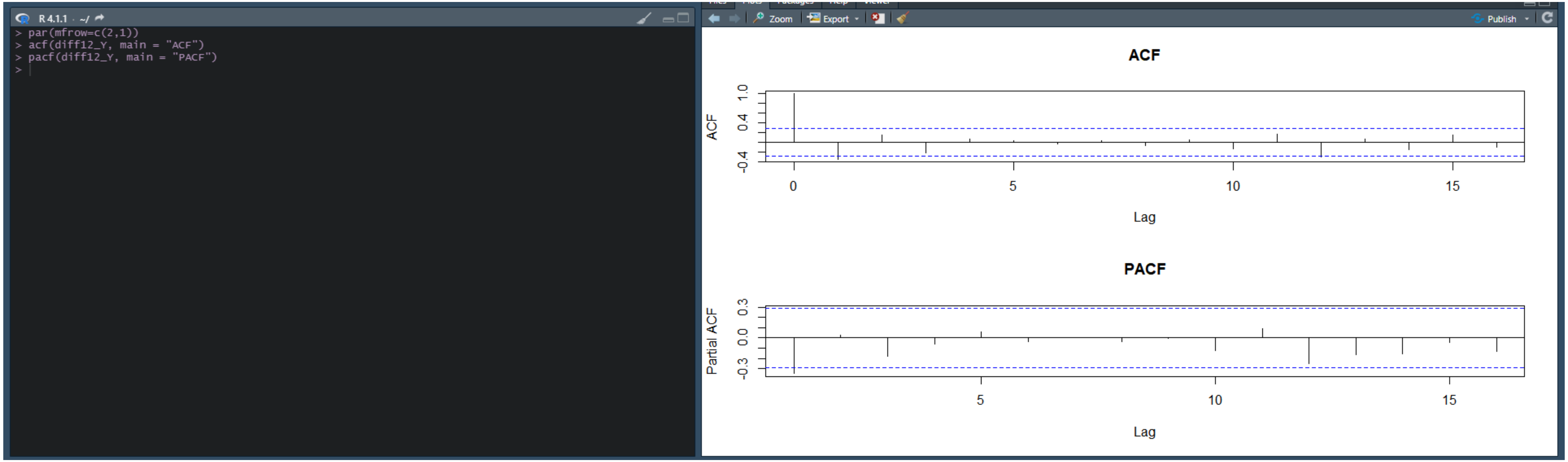


- csv 파일을 불러와 시도표 확인(plot.ts)

- 분산은 일정한 것으로 보임
- 추세 및 계절성이 있다고 판단 : 차분을 수행하되, 1차 차분(differences=1) 및 계절 차분(lag=12)을 순서대로 진행

- 만들어진 정상 시계열이 'diff12_Y'

Box-Jenkins 방법론 실습 - 1. 모형 식별



- ‘diff12_Y’의 ACF와 PACF를 확인한 후, 3가지 후보 모형 선정

1) SARIMA(1, 1, 1) X (1, 1, 0) : ‘fit1’
2) SARIMA(0, 1, 1) X (1, 1, 0) : ‘fit2’
3) SARIMA(2, 1, 1) X (1, 1, 0) (auto.arima 결과 활용) : ‘fit3’

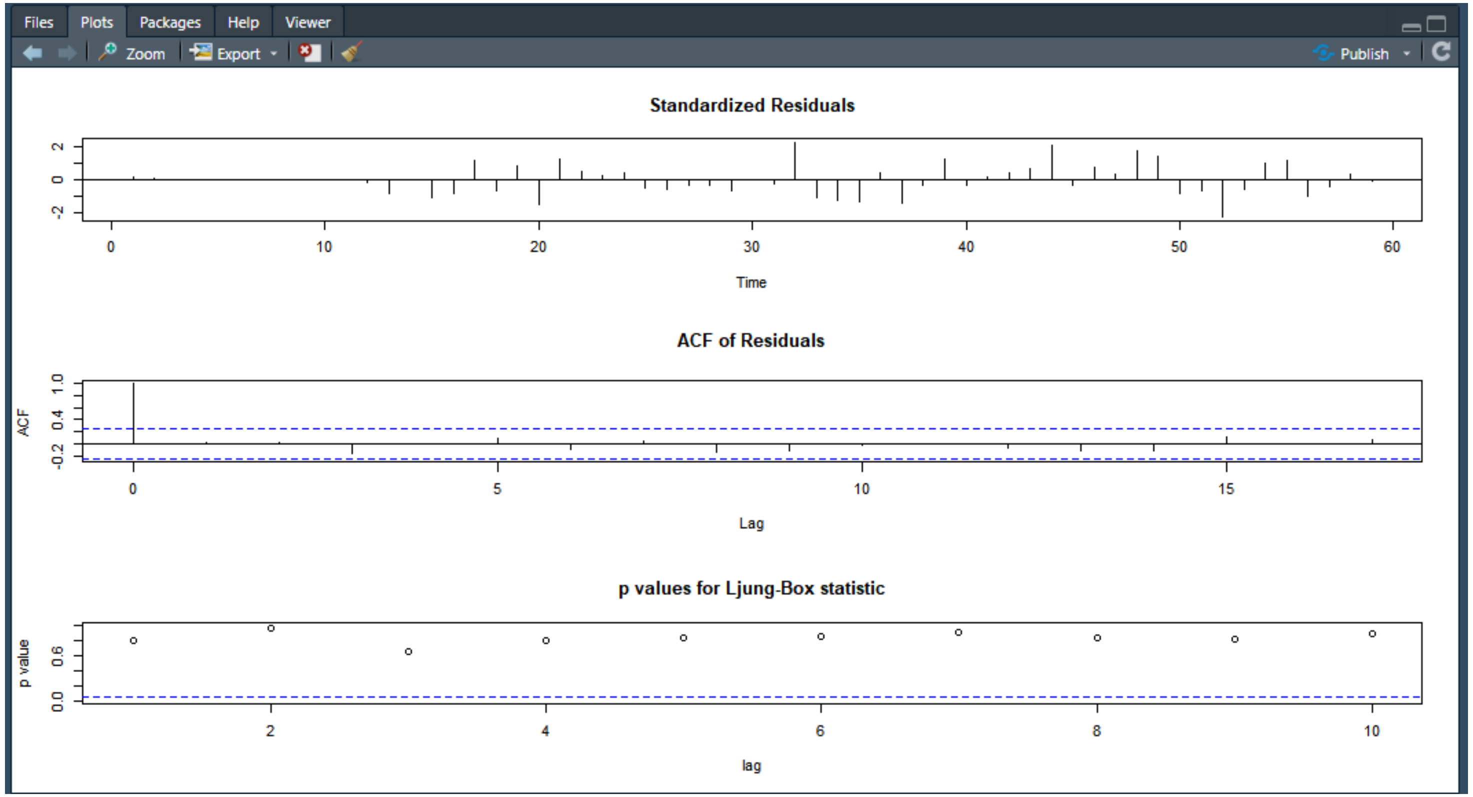
Box-Jenkins 방법론 실습 – 2. 모수 추정

```
> fit1

Call:
arima(x = Y, order = c(1, 1, 1), seasonal = list(order = c(1, 1, 0), period = 12))

Coefficients:
         ar1      ma1      sar1
      0.6091  -0.9513   -0.4285
s.e.  0.3824   0.3755    0.1818

sigma^2 estimated as 0.05561:  log likelihood = -0.78,  aic = 9.56
> fit2

Call:
arima(x = Y, order = c(0, 1, 1), seasonal = list(order = c(1, 1, 0), period = 12))

Coefficients:
          ma1     sar1
      -0.4187  -0.4305
s.e.   0.1580   0.1507

sigma^2 estimated as 0.05759:  log likelihood = -0.94,  aic = 7.89
> fit3

Call:
arima(x = Y, order = c(2, 1, 1), seasonal = list(order = c(1, 1, 0), period = 12))

Coefficients:
         ar1     ar2      ma1     sar1
      0.5279  0.1951  -1.0000  -0.3968
s.e.  0.1459  0.1472   0.0969   0.1520

sigma^2 estimated as 0.05278:  log likelihood = 0.09,  aic = 9.83
>
```

1) SARIMA(1, 1, 1) X (1, 1, 0) : 'fit1'
2) SARIMA(0, 1, 1) X (1, 1, 0) : 'fit2'
3) SARIMA(2, 1, 1) X (1, 1, 0)
(auto.arima 결과 활용) : 'fit3'

3가지 후보 모형을 비교해본 결과,
'fit2'에 대한 AIC 값이 가장 작게 나왔음.
따라서 최종 모형은
SARIMA(0, 1, 1) X (1, 1, 0)으로 결정!

Box-Jenkins 방법론 실습 – 3. 모형 적합성 진단

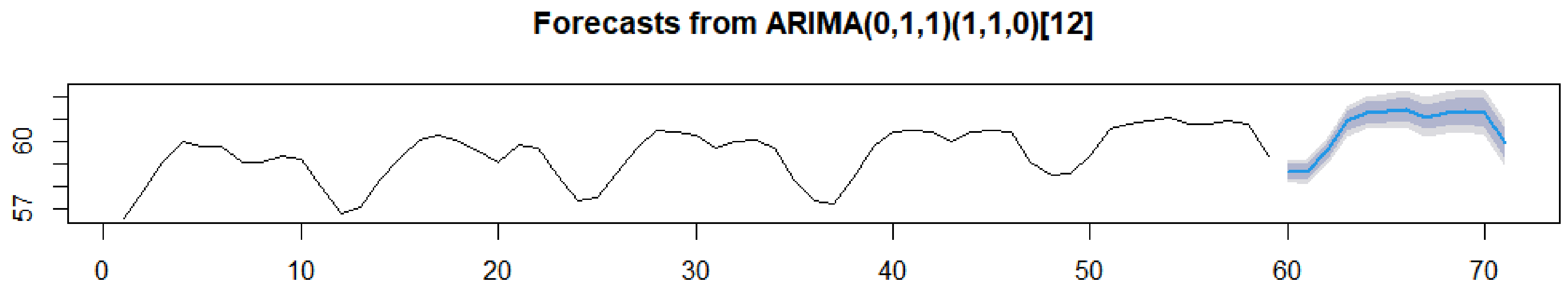


- tsdiag(fit2)의 실행 결과
- 모형이 잘 나온 것으로 보이는데,
  - ✓ 잔차의 ACF에서 자기상관성이 눈에 띄지 않고,
  - ✓ p value가 기준선 위에 위치하기 (높게 나왔기) 때문

Box-Jenkins 방법론 실습 – 4. 모형 확정 및 예측

```
> diff12_Y.forecasts <- forecast(fit2, h = 12)
> diff12_Y.forecasts
   Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
60       58.67934 58.37181 58.98687 58.20901 59.14967
61       58.65019 58.29447 59.00591 58.10617 59.19421
62       59.62239 59.22428 60.02050 59.01353 60.23125
63       60.90849 60.47208 61.34490 60.24106 61.57592
64       61.28069 60.80909 61.75229 60.55944 62.00195
65       61.38069 60.87634 61.88504 60.60936 62.15203
66       61.45154 60.91645 61.98664 60.63318 62.26990
67       61.10849 60.54432 61.67266 60.24567 61.97131
68       61.28069 60.68888 61.87250 60.37559 62.18579
69       61.38069 60.76247 61.99891 60.43520 62.32618
70       61.28069 60.63714 61.92424 60.29646 62.26492
71       59.92374 59.25582 60.59166 58.90225 60.94524
```

- fit2에 기반한 예측값 12개
- forecast 함수를 활용하여 신뢰구간 80%, 95%의 상한과 하한도 함께 출력



과 제

[과제] 시계열 분석 실습

- 데이터 : ‘kingage.csv’
  - ✓ 출처 : http://robjhyndman.com/tsdldata/misc/kings.dat
- Box-Jenkins 방법론(모형 식별, 모수 추정, 모형 적합성 진단, 모형 확정 및 예측)에 따라 시계열 분석을 진행해주세요!
  - ✓ 각 단계별 설명을 주석으로 달아주시면 감사하겠습니다 :)
  - ✓ 최종적으로, 예측값 17개를 출력해주세요.
- Python, R 중 어느 것을 사용하셔도 무관합니다.

Reference

[강의안]
- 투빅스 14기 이원도님 강의안
- 연세대학교 응용통계학과 문지은 교수님 <시계열분석> 강의안

[교재]
- Robert H. Shumway, David S. Stoffer, <Time Series Analysis and Its Applications>

[참고 자료]
- [시계열 분해] Chapter 6 시계열 분해 | Forecasting: Principles and Practice (otexts.com)
- [AR, MA, ARMA] [머신러닝][시계열] AR, MA, ARMA, ARIMA의 모든 것 - 개념편 (velog.io)
- [ARIMA] ARIMA, Python으로 하는 시계열분석 (feat. 비트코인 가격예측) (byeongkijeong.github.io)
- [Prophet] https://hyperconnect.github.io/2020/03/09/prophet-package.html

Q & A

들어주셔서 감사합니다.