시계열분석팀 1주차 교안 ⏳

[목차]

1. 시계열 자료 및 분석

(1) 시계열 자료란?

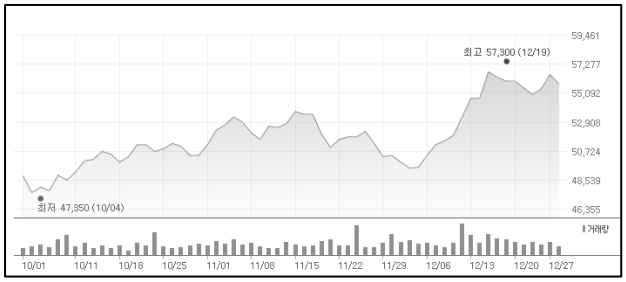
연도별, 계절별, 월별, 일별, 또는 작은 시간대로서 시, 분, 초별로 시간의 흐름에 따라 순석대로 관측되는 자료

시간의 영향을 받음

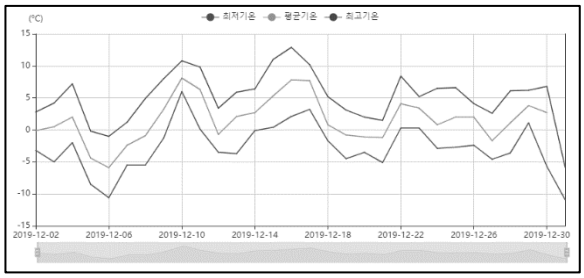
자료가 생성되는 특성에 따라 이산시계열(discrete time series), 연속시계열로(continuous time series) 나뉘지만 실제로 접하는 데이터는 대부분 이산시계열

Ex) 월 상품 매출, 강우량, 경제지표 데이터

[시계열 자료 예시 - 주가지수 데이터]



[시계열 자료 예시 - 날씨 데이터]



(2) 시계열 분석의 목적

1. 시간 변수의 흐름에 따른 종속 변수의 패턴을 예측
2. 시계열 자료가 생성된 시스템을 이해하고 제어(control)

(3) 시계열 분석

말그대로 시계열 자료를 분석하는 것

목적에 따라 분석 방법이 달라짐

1) 예측이 목적 : 추세분석(trend analysis), 평활법(smoothing method), 분해법(decomposition method), 자기회귀누적이동평균(ARIMA)
2) 시스템을 이해, 제어가 목적 : 스펙트럼분석(spectral analysis), 개입분석(intervention analysis), 전이함수모형(transfer function model)

➔ 우리는 예측이 목적인 모델에 초점을 맞추어 공부할 예정

(4) 회귀분석과의 차이

회귀 모델 : $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

시계열 모델 : $y_t = \delta + \phi_1 y_{1-1} + \varepsilon_t$

➔ 회귀분석을 통한 예측은 설명변수를 통해 예측을 하는 반면 시계열은 과거 자료로부터 시계열에 존재하는 패턴을 찾아 예측함

➔ 또한 시계열분석은 자기상관성을 가지는 것을 전제로 하여 과거의 데이터가 미래의 데이터에 영향을 미친다고 보고 분석을 하기 때문에 순서가 중요한 반면, 회귀분석은 독립성을 전제로 하기 때문에 순서를 신경 쓰지 않음

(5) 시계열 자료의 특성

1) 추세변동 (Trend Variation : T) : 상승과 하락의 경향(장기 변동 요인)
2) 순환변동 (Cyclical Variation : C) : 2~10년의 주기에서 동안 상승과 하락을 반복하는 요소(중, 장기적 변동요인)
3) 계절변동 (Seasonal Variation : S) : 일정한 기간(월, 요일, 분기) : 1년단위 반복적 변동 요소 (단기 변동요인)
4) 불규칙적변동 (Irregular Variation : I) : 어떤 규칙없이 예측 불가능한 변동요인 (설명할 수 없음) – White Noise

- 불규칙적 성분(irregular component) 과 체계적 성분(systematic component)으로 나눌 수 있음
- 체계적 성분에 추세 성분, 순환성분, 계절성분이 있음. 불규칙 성분이 백색잡음이며, 앞으로 배

울 정상적 시계열 중 가장 대표적인 예시임.

- *Cf)* 계절변동은 보통 주별, 월별, 계절별이지만 순환변동은 이보다 주기가 김. 태양 흑점수의 변화를 순환주기를 가진 시계열자료라 할 수 있음

2. 정상성

(1) 정상성 (Stationarity)

- 시계열의 확률적 성질 (평균, 분산 등) 들이 시간의 흐름에 따라 변하지 않는 성질
- '안정성'이라고도 불림

정상성의 필요성 : 시계열의 각 시점은 확률변수이기 때문에 각 시점별로 확률분포를 가진다. 시계열 분석에서 미래 예측을 위해서는 무한한 시점들의 결합분포를 고려해야 하지만, 다수의 시점들에 대한 결합분포를 파악하는 것은 너무나 복합하기에 이를 간략한 방법으로 대체하기 위해서 '정상성'이라는 가정을 필요로 한다.

(2) 강정상성

$$F(\mathrm{X_t}, \dots, \mathrm{X}_{t+p}) = F(X_{t+k}, \dots, X_{t+k+p})$$

- 동일한 기간의 시계열에 대한 결합확률분포가 모든 시계열 구간에서 동일하게 나타나는 경우
- 지나치게 엄격한 가정으로 이를 만족하는 시계열 자료를 구하는 것은 불가능
- 따라서 약정상성가정이 필요

(3) 약정상성

1) $\mathrm{E}(\mathrm{y_t}) = \mu < \infty$: 평균
2) $\mathrm{Var}(\mathrm{y_t}) = E(y_t - \mu)^2 = \gamma_0 < \infty$: 분산
3) $\mathrm{Cov}(\mathrm{y_t}, \mathrm{y_{t+k}}) = E(y_t - \mu)(y_{t+k} - \mu) = \gamma_k < \infty \quad \forall k$: 자기공분산

- 동일한 기간의 시계열에 대한 1, 2차 Moment가 동일하게 나타나는 경우
- 분포의 동일성이 만족되지 않더라도 특성치의 동일성이 만족하는 경우
- 위의 세가지를 만족하면 정상적 시계열로 간주
- 시간t에 관계없이 평균과 분산이 일정함

- 자기공분산 γ_k가 시차(time lag) k에만 의존하고, 시점 t와는 무관함
- 강정상성을 만족하는 시계열은 거의 없으므로, 약정상성을 만족하면 정상적 시계열로 간주
- 2차정상성(second-order stationarity), 공분산정상성(covariance stationarity)라고도 함

3. 정상화

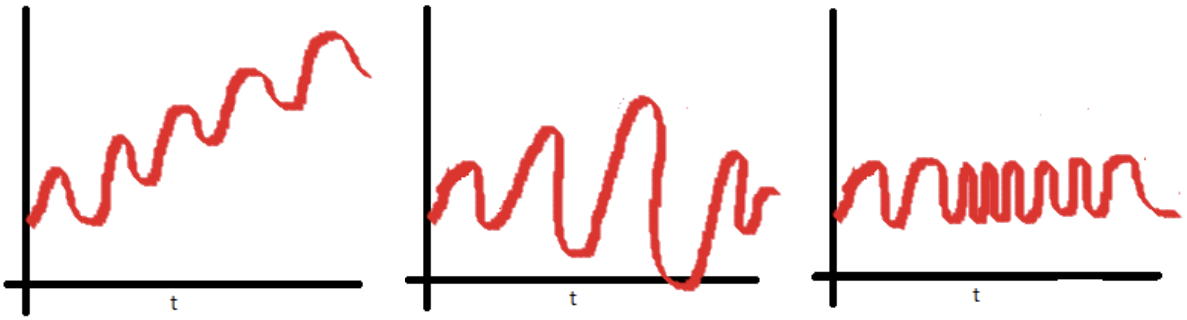
(1) 비정상시계열

대부분의 시계열은 불안정 시계열이다

비정상 시계열의 특징

1) 추세가 있는 경우 (평균이 일정하지 않을 때)
2) 분산이 시간대에 따라 변하는 경우
3) 시점에 공분산이 의존하는 경우
4) 계절변동이 있는 경우

[추세가 존재] [분산이 일정하지 않음] [공분산이 일정하지 않음]



➔ 시계열의 정상성을 확보 후 분석이 필요!!

(2) 정상화 방법

① 분산이 일정하지 않은 경우 분산안정화 (로그변환, 제곱근 변환, Box-cox변환)

② 분산 변환 후 추세/계절성을 갖는다면 회귀/평활/차분을 통해 정상화

③ 비정상을 유발하는 성분인 체계적 성분(추세, 계절성)과 불규칙성분(백색잡음)을 분해

$$Y_t = T_t + S_t + I_t$$

T_t: 추세성분, S_t : 계절 성분, I_t : 불규칙 성분

➔ 비정상을 유발하는 T_t, S_t 제거하고 I_t를 찾아 정상시계열(White Noise)인지 검정

➔ 순환성분은 계절성분보다 더 긴 주기를 갖는 성분임. 따라서 관측자료를 통해 주기를 추정하나 순환성분은 그 주기를 찾아 모형에 반영하기 쉽지 않기에 분해법에서는 순환성분을 일반적으로 고려하지 않음

1) **분산**이 일정하지 않은 경우 – 분산안정화

① 로그변환

$f(x) = \ln(x)$

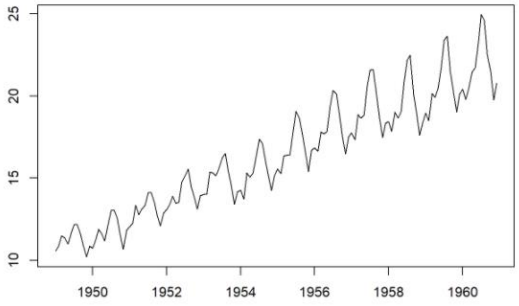
② 제곱근 변환

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

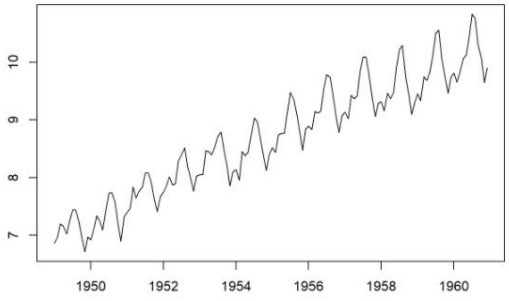
③ Box-cox 변환

$$f(x;\lambda) = \frac{x^{\lambda} - 1}{\lambda}$$

[분산안정화 전]



[분산안정화 후]



2) **추세/계절성**을 가질 경우 – 회귀, 평활, 차분

시계열의 유형과 방법에 따라 정상화 과정이 달라짐

✓ 시계열 유형 : 1) 추세만 있는 시계열, 2) 계절성만 있는 시계열, 3) 추세와 계절성 모두 있는 시계열

✓ 정상화 방법 : 1) 회귀, 2) 평활, 3) 차분

① 회귀

(1) 추세만 있는 시계열

Step1. 추세성분과 불규칙성분만 있는 시계열 모형을 가정한다.

$$Y_t = T_t + I_t, E(I_t) = 0$$

Step2. 추세성분 T_t 를 t에 대한 선형회귀식으로 나타낸다.

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

Step3. 나타낸 선형회귀식을 최소제곱법(OLS)를 통해 각 계수를 추정한다.

$$(\hat{c_0}, \hat{c_1}, \ldots, \hat{c_p}) = argmin \sum (Y_t - T_t)^2$$

Step4. 추정한 추세를 시계열에서 제거한다.

(2) 계절성만 있는 시계열

Step1. 계절성분과 불규칙 성분이 있는 모형을 가정한다.

$$Y_t = S_t + I_t, E(I_t) = 0 \text{ 이때, } S_t = a_0 + \sum_{j=1}^{k} a_j \cos(\lambda_j t) + b_j \sin(\lambda_j t))$$

Step2. λ_j, k 를 선택 후 하나의 회귀식으로 보고 최소제곱법을 적용시켜 각 계수 (a_j, b_j) 를 추정한다

Step3. 추정된 계절성을 시계열에서 제거한다.

(3) 추세, 계절성 모두 있는 시계열

Step1. (1), (2)를 차례대로 해준다.

Step2. 다시 추세제거를 해준다. (다시 추세가 생길 경우)

회귀방법의 단점 : 회귀분석에서는 잔차에 대한 독립성을 가정하는데, 시계열에서는 잔차의 독립성을 가정하지 않기에 추정이 명확하지 않을 수 있다. 또한 추세와 계절성이 독립이 아니므로 위와 같은 방법으로 분해를 하는 것은 옳지 않을 수도 있다. 더불어, 가장 중요한 가정은 모수들이 시간에 따라 변하지 않는다는 것이다. 따라서 모수가 시간에 따라 변한다면 평활법을 이용하는 것이 더 좋을 수도 있다.

② 평활

(1) 추세만 있는 시계열

✓ 이동평균 평활법(Moving Average Smoothing) : 일정기간마다 평균을 계산

Step1. 시점 t에 대해 t-q시점의 관측치부터 t+q의 시점의 관측치까지의 평균을 구한다.

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

Step2. Y_{t+j}에 추세만 있는 시계열식 대입(추세는 linear하다고 가정)

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

$$= \frac{1}{2q+1}\sum_{j=-q}^{q} T_{t+j} + \frac{1}{2q+1}\sum_{j=-q}^{q} I_{t+j}$$

$$\frac{1}{2q+1}\sum_{j=-q}^{q} T_{t+j} = c_0 + c_1 t = T_t \,, t \in [q+1, n-q]$$

$$\frac{1}{2q+1}\sum_{j=-q}^{q} I_{t+j} \approx E(I_t) = 0$$

Step3. $Y_t - W_t$ 를 해주면 추세를 제거할 수 있다.

이동평균법의 단점 : 최근 자료와 과거자료에 대한 가중치가 동일하게 적용됨

✓ 지수 평활법(Exponential Smoothing) : 과거자료의 가중평균으로 추세를 추정

Step1. 추세를 다음과 같이 구한다.

$$\hat{T}_t = aY_t + (1-a)\widehat{T_{t-1}}$$

시점1 추세 : $\hat{T}_1 = Y_1$
시점2 추세 : $\widehat{T_2} = aY_2 + (1-a)\hat{T}_1 = aY_2 + (1-a)Y_1$
시점3 추세 : $\widehat{T_3} = aY_3 + (1-a)\widehat{T_2} = aY_2 + a(1-a)Y_2 + (1-a)Y_1$
⋮
시점t 추세 : $\hat{T}_t = aY_t + (1-a)\widehat{T_{t-1}} = \sum_{j=0}^{t-2}(a(1-a)^j Y_{t-j} + (1-a)^{t-1}Y_1$

Step2. 추세를 기존관측치에서 제거한다.

- a 는 평활계수로 가까운 과거에 대한 가중치를 나타냄
- 과거시점일수록 더 작은 가중치가 부여된다.($0 \le a \le 1$)
- 평활계수a가 클수록 시계열 변화에 따른 예측값의 변화가 크게 나타나며, 작을수록 시계열 변화에 따른 예측값의 변화가 작게 나타남

➔ 평활법에서 q와 a의 선택 : q와 a는 분석자가 선택하는 값이기 때문에 신중히 선택해야한다. 교차검증을 토해 MSE를 최소화하는 q와 a찾아 보통 선택한다.

(2) 계절성만 있는 시계열

Step1. $\widehat{S_t}$를 같은 주기를 갖는 모든 값들의 평균값으로 대체한다.

계절성 주기가 d일때,

$$\widehat{S_1} = \frac{1}{\mathrm{m}}\left(Y_1 + Y_{1+\mathrm{d}} + Y_{1+2\mathrm{d}} + Y_{1+3\mathrm{d}} + \cdots + Y_{1+(\mathrm{m}-1)\mathrm{d}}\right)$$

$$\widehat{S_2} = \frac{1}{m}(Y_2 + Y_{2+\mathrm{d}} + Y_{2+2\mathrm{d}} + Y_{2+3\mathrm{d}} + \cdots + Y_{2+(\mathrm{m}-1)\mathrm{d}}$$

$$\vdots$$

$$\widehat{\mathrm{S_d}} = \frac{1}{\mathrm{m}}(Y_{\mathrm{d}} + Y_{\mathrm{d}+\mathrm{d}} + Y_{\mathrm{d}+2\mathrm{d}} + Y_{\mathrm{d}+3\mathrm{d}} + \cdots + Y_{\mathrm{d}+(\mathrm{m}-1)\mathrm{d}}$$

Step2. 이 과정을 통해 계절성을 추정함

(3) 추세, 계절성 모두 있는 시계열 (회귀+평활) : Classical decomposition algorithm

Step1. 이동평균평활법을 이용하여 추세를 추정한다 (이때 $\sum_{j=1}^{d} S_j = 0$)

Step2. 관측값에서 추정한 추세를 빼 계절성과 불규칙성분만 남긴다.

$$\mathrm{Y_t} - \widehat{T_t} \approx S_t + I_t$$

Step3. Seaonsal Smoothing을 통하여 계절성을 추정한다

Step4. 관측값에서 추성한 계절성을 빼 추세와 불규칙성분만 남긴다

$$\mathrm{Y_t} - \widehat{S_t} \approx T_t + I_t$$

Step5. (Step4)식의 추세성분$\mathrm{T_t}$를 회귀를 통해 추정한다.

Step6. (Step3)에서 추정했던 계절성$\mathrm{S_t}$ 와 (Step5)에서 새롭게 추정한 추세성분$\mathrm{T_t}$를 관측치에서 제거한다.

* 후항연산자(backshift operator, B) : $\mathrm{BX_t} = X_{t-1}$

③ 차분

- 현재시계열에서 과거 시계열을 빼는 것

✓ 1차차분 : $\nabla \mathrm{X_t} = X_t - X_{t-1}$

✓ 2차차분 : $\nabla^2 \mathrm{X_t} = \nabla(\nabla X_t) = X_t - X_{t-1} - (X_{t-1} - X_{t-2}) = X_t - 2X_{t-1} + X_{t-2}$

✓ 3차차분 : $\nabla^3 X_t = \nabla^2 X_t - \nabla^2 X_{t-1}$

✓ 후항연산자로 표현 : $(1 - \mathrm{B})\mathrm{X_t} = X_t - X_{t-1}$

(1) 추세만 있는 시계열

- 추세가 1차식일 때 : $Y_\mathrm{t} = T_t + I_t = (c_0 + c_1 t) + I_t$

$$\nabla Y_t = (1 - B)Y_t = \mathrm{Y_t} - Y_{t-1}$$
$$= (c_0 + c_1 t + I_t) - (c_0 + c_1(t-1) + I_{t-1})$$
$$= c_1 + I_t - I_{t-1}$$

➔ 상수와 오차항만 남아 추세 제거 가능하다

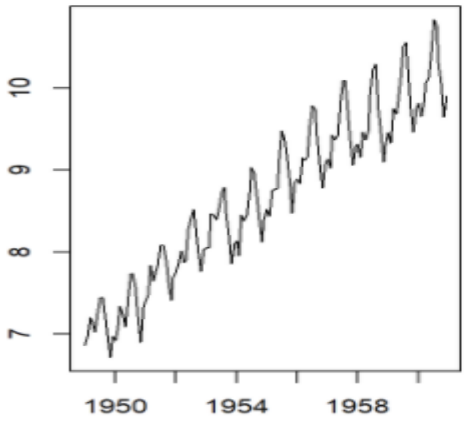
- 추세가 2차식일 때 : $Y_\mathrm{t} = T_t + I_t = (c_0 + c_1 t + c_2 t^2) + I_t$

$$\nabla^2 \mathrm{Y_t} = Y_t - 2Y_{t-1} + Y_{t-2}$$
$$= 2c_2 + (I_t - 2I_{t-1} + I_{t-2})$$

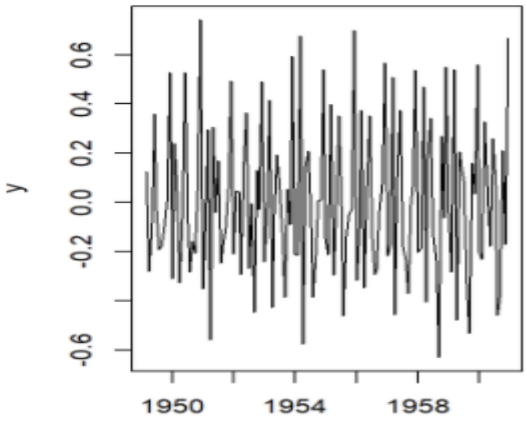
➔ 상수와 오차항만 남아 추세 제거 가능하다

- 일반적으로 추세가 k차 다항식일 경우 k번차분을 하면 추세 제거 가능

[차분 전]



[2차차분 후]



(2) 계절성만 있는 시계열

✓ 계절차분(Seasonal differencing)을 이용한다

✓ 주기가 d인 계절차분 : $\nabla_{\mathrm{d}} X_t = \left(1 - B^d\right) X_t = X_t - X_{t-d}$

✓ Cf) d차차분 : $\nabla^{\mathrm{d}} X_t = (1 - B)^d X_t$

- 주기가 4인 시계열

$$\nabla_4 Y_t = (1 - B^4) Y_t$$
$$= (1 - B^4)(S_t + I_t)$$
$$= (S_t + I_t) - (S_{t-1} + I_{t-1})$$
$$= I_t - I_{t-1}$$

➔ 오차항만 남아 계절성제거 가능

(3) 추세와 계절성 모두 있는 시계열

Step1. 추세와 계절성이 있는 모형을 가정

$$\mathrm{Y_t} = T_t + S_t + I_t = (c_0 + c_1 t) + S_t + I_t$$

Step2. 1차차분과 주기가 d인 차분 동시에 적용

$$\nabla\nabla_{\mathrm{d}} Y_t = (1 - B)(Y_t - Y_{t-d})$$
$$= (1 - B)(c_0 + c_1 t + S_t + I_t - c_0 - c_1(t - d) - S_{t-d} - I_{t-d})$$
$$= (1 - B)(c_1 d + S_t - S_{t-d} + I_t - I_{t-d})$$
$$= I_t - I_{t-1} - I_{t-d} - I_{t-d-1}$$

➔ 오차항만 남아 추세와 계절성제거 가능

➔ 계절차분과 1차차분을 바꾸어도 결과는 같음

4. 정상성 검정

추세와 계절성을 제거한 뒤 남은 오차항이 정상성을 따르는지 검정하는 과정

(1) 자기공분산함수, 자기상관함수

일반적으로 시계열자료은 현재의 상태가 과거 및 미래의 상태와 밀접한 관계를 갖고 있다. 따라서 시간의 흐름에따라 독립적이지 않다. 이러한 경우 시계열은 자기상관관계를 갖는다고 하며, 시간에 따른 상관정도를 나타내기 위하여 다음과 같은 것을 사용한다.

➔ 정상성을 따르는지 알아보기 위해 오차항의 자기공분산을 확인

① 자기공분산함수(Auto-Covariance Function)

$$\gamma(\mathrm{k}) = \mathrm{Cov}(\mathrm{X_t}, \mathrm{X_{t+k}}) = E[(X_t - \mu)(X_{t+k} - \mu)]$$

② 표본자기공분산함수:

$$\hat{\gamma}_k = \frac{1}{T}\sum_{j=1}^{T-k}(X_j - \bar{X})(X_{j+k} - \bar{X}), k = 1,2,3, \ldots$$

③ 자기상관계수(Auto-Correlation Function) :

$$\rho(\mathrm{k}) = Corr(X_t, X_{t+k}) = \frac{Cov(X_t, X_{t+k})}{\sqrt{Var(X_t)}\sqrt{Var(X_{t+k})}} = \frac{\gamma(k)}{\gamma(0)}, \text{이때} \gamma_0 = Var(X_t)$$
$$= E[(X_t - \mu)^2]$$

④ 표본자기상관계수(Sample Auto Correlation Function) :

$$\widehat{\rho_k} = \frac{\widehat{\gamma_k}}{\widehat{\gamma_0}},$$

(2) 백색잡음 (White Noise)

- 서로 독립이고 동일한 분포를 따르는(iid) 확률변수들의 계열로 구성된 확률과정
- $\mathbf{Y_t} \sim \boldsymbol{WN}(\mathbf{0}, \boldsymbol{\sigma_Y^2})$

① $\mathrm{E}(\mathrm{Y_t}) = 0$

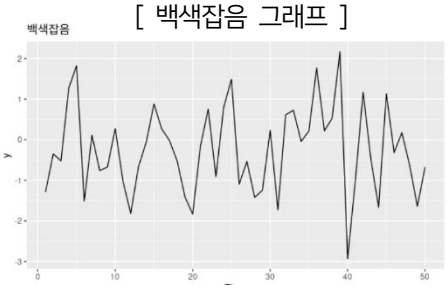
② $\mathrm{Var}(\mathrm{Y_t}) = \sigma^2$

③ $\mathrm{Cov}(\mathrm{Y_t}, \mathrm{Y_{t+k}}) = 0$ (약정상성조건에 '공분산이 0'이라는 조건이 추가됨)

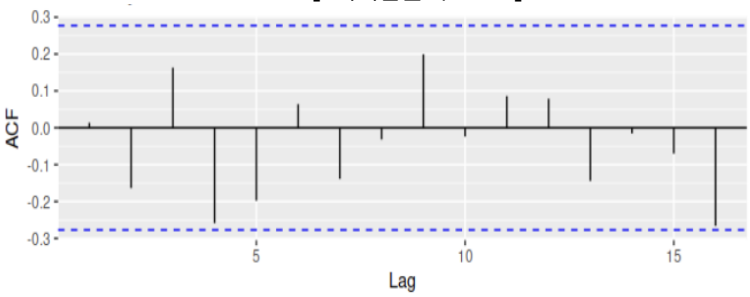
- 대표적인 정상시계열임

- 백색잡음이 정규분포를 따르는 경우 이를 '가우시안 백색잡음'이라도고 함
- 오차항들이 서로 독립이라는 강한 가정대신에 서로 상관관계가 없다(uncorrelated)는 약화된 가정을 만족하는 경우 백색잡음이라고 정의하기도 함
- 추세와 계절성을 제거하고 남은 오차항(잔차)가 백색잡음이라면 그 오차항은 정상성을 따르며 공분상 행렬을 추정할 필요가 없음 -> 추가적인 모델링 불필요

[백색잡음 그래프]



[백색잡음의 ACF]



(3) 백색잡음 검정

- $X_t \sim WN(0,1)$ 이라면 n이 충분히 클 경우, $\hat{\rho}(h) \approx N\left(0, \frac{1}{n}\right)$ 를 따름
- 이 특징을 활용하여 통계적으로 유의한 상관관계가 존재하는지 확인

1) 자기상관관계가 없는지 검정

귀무가설: $H_0: \rho_1 = \rho_2 = \cdots = \rho_k = 0$

① ACF그래프를 통한 검정 : ACF, PACF그래프를 그렸을 때, 각 시점에서 그래프가 신뢰구간안에 존재하고 있을 경우 잔차간의 상관관계가 없다고 판단

② 포트맨토 검정(Portmanteau Test, Ljung-Box 검정) : $\epsilon_t \sim N(0, \sigma^2)$ 가정하에서 검정통계량 $Q' = T(T+2)\sum_{k=1}^{h} \frac{r_k^2}{T-k} \sim \chi(h-K)$ 를 사용하여 검정

: 이때, h는 최대 고려해야할 시차, T는 관측값 개수, K는 모델의 매개변수(원본데이터에서 계산한다면 K=0)

2) 정규성 검정 :

① QQplot확인 : QQplot을 그려서 정규성 가정이 만족되는지 시각적으로 확인

② Kolmogorov-Smirnov test : 자료의 평균/표준편차와 히스토그램을 표준정규분포와 비교하여 적합도를 검정함. 귀무가설은 '정규성을 따른다'이다.

③ Jarque-Bera test : 왜도와 첨도를 정규분포와 비교를 통해서 정규성을 검정함. 귀무가설은 '데이터가 정규분포를 따른다'이다.

3) 정상성 검정 :

① kpss test : 단위근(Unit-root) 검정 방법 중 하나. 귀무가설은 '시계열이 정상(stationary)시계열이다.'이다.

② ADF(Augmented Dickey-Fuller)Test : 정상성을 알아보기위한 단위근 검정방법 중 하나이며, DF검정을 일반화 한것이다. 귀무가설은 '자료에 단위근이 존재한다.'이며, 대립가설은 '자료가 정상성을 만족한다.'이다.

③ PP(Phillips-Perron)Test : 이분산이 있을 경우에도 사용가능한 검정방법. 귀무가설은 '데이터가 비정상이다.'이며, 대립가설이 '자료가 정상성을 만족한다.'이다.

5. R로 실습해보기