시계열자료분석팀 1주차

[목차]

1 시계열자료분석 알아보기

1.1 시계열 자료란?

시계열 자료란 시간에 따라 관측된 자료의 집합을 시계열 자료(time series)라고 합니다. 일반적으로 $X_t, t = 1,2, \ldots$ 와 같이 시계열 자료를 표현할 수 있습니다. 이 때 시간 t가 이산형이면 이산형 시계열 자료, 연속형이라면 연속형 시계열 자료로 구분할 수 있습니다.

시계열 자료의 대표적인 특성은 '**dependency**'입니다. 시계열 데이터는 시간에 따른 자료이므로 관측치(observations)들 사이에 연관성이 존재합니다. 즉, 잔차의 독립성 가정을 만족하지 않습니다.

1.2 시계열 자료의 구성 요소

1) 추세 변동 (Trend)

➔ 시간이 경과함에 따라 관측치가 증가하거나 감소하는 추세를 갖는 경우의 변동

2) 순환 변동 (Cycle)

➔ 주기적인 변화가 있지만 계절에 의한 것이 아니며, 주기가 긴 경우의 변동

3) 계절 변동 (Seasonal variation)

➔ 주별·월별·계절별과 같은 주기적인 성분에 의한 변동

4) 우연 변동 / 불규칙 성분 (Random fluctuation)

➔ 시간에 따른 규칙적인 움직임과 무관하게, 무작위 원인에 의해 나타나는 변동

1.3 시계열 모형이란?

시계열 모형(time series model)이란, 시계열 데이터 x_t가 어떤 확률 구조(=확률 분포)에 따라 생성되는지를 묘사하는 확률적 모형입니다. 확률적 모형이란, 미래에 대한 정보가 불확실하여 확률 또는 확률분포로써 표현되는 모형을 말합니다.

이 때, 특정 시점에 대한 확률 변수인 X_t는 하나의 관측치(x_t)만 고려하는 것이 아니라, 전체 시점에서의 관측치 집합 $\{x_1, x_2, \ldots\}$을 모두 고려한 결합 분포입니다.

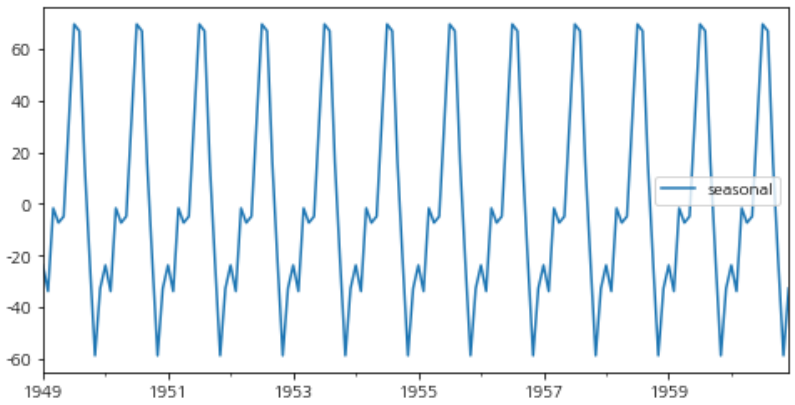
2 정상성 (Stationarity)

2.1 정상성(Stationarity)이란?

시계열자료분석에 있어서 가장 핵심적인 개념이 바로 정상성입니다. **정상성**이란 **시계열 자료의 확률적 성질이 시점(t)에 의존하지 않고**, 대신에 **시차(lag)에만 의존하는 특성**을 의미합니다. 즉, 시간의 흐름에 따라 평균과 분산이 변하지 않는 것을 의미합니다.

1.3에서 공부한 것처럼 시계열 모형은 전체 시점의 관측치에 대한 결합 분포인 확률 변수들로 구성됩니다. 이 때, 미래(X_{t+1})에 대해 예측하고자 한다면, 우리가 관측하지 못한 미래 값(x_{t+1})까지 포함된 결합 분포를 고려해야 합니다. 각 확률 변수들은 각각의 확률 분포를 가지는데, 각 시점마다의 모든 확률 분포를 구하는 것은 사실상 불가능합니다.

바로 여기서 정상성의 필요성이 등장합니다. 정상성을 가정하면 위와 같은 조건을 완화할 수 있습니다. 각 확률적 성질이 시점에 의존하지 않고 lag에만 의존하기 때문에, lag 내에서의 분포를 구하면 해당 분포가 간격마다 반복된다고 볼 수 있기 때문입니다.



2.2 강정상성 (Strict Stationarity)

$$\left(X_{t_1}, \dots, X_{t_n}\right) \overset{d}{=} \left(X_{t_1+h}, \dots, X_{t_n+h}\right)$$

모든 h와 양수의 n에 대하여 시계열 $\{X_t, t \in \mathbb{Z}\}$가 위의 조건을 만족할 때, 강정상성을 만족하는 시계열이라 합니다. 이때 h는 시차 lag을 의미합니다. t_1부터 t_n까지 n 기간만큼의 시계열에 대한 결합 분포는, 시점을 h만큼 옮겨도 동일한 기간에 대해서는 같은 결합 분포를 가져야 함을 의미합니다. 하지만 강정상성 역시 여전히 지나치게 엄격한 가정이므로 더 완화된 가정이 필요합니다.

$$\left(X_{t_1}, \dots, X_{t_n}\right) \sim MVN(\mu, \Sigma)$$

조건 완화를 위해 강정상성에 정규성(Gaussianity)을 가정한다면, 분포를 추정하기 위해 우리가 해야 할 것은 평균 벡터인 μ와 공분산 행렬인 Σ을 구하는 문제로 가벼워집니다. 정규성 가정으로부터 **1) 각 확률변수의 기댓값은 상수(constant)**며, **2) 각 확률변수의 공분산은 시점이 아닌 시차에만 의존**하게 됩니다. (공분산 행렬은 대칭 행렬이기 때문)

2.3 약정상성 (Weakly Stationarity)

정규성 가정으로부터 나오는 조건들을 더 확장하여 완화한 것이 바로 약정상성입니다.

i) $\boldsymbol{E[|X_t|]^2 < \infty, \quad \forall t \in \mathbb{Z}}$

= 2차 적률(분산)이 존재하고, 시점에 관계없이 일정하다.

ii) $\boldsymbol{E[X_t] = m, \ \forall t \in \mathbb{Z}}$

= 평균은 상수로, 시점에 관계없이 일정하다.

iii) $\boldsymbol{\gamma_X(r,s) = \gamma_X(r+h, s+h), \ \forall r,s,h \in \mathbb{Z}}$

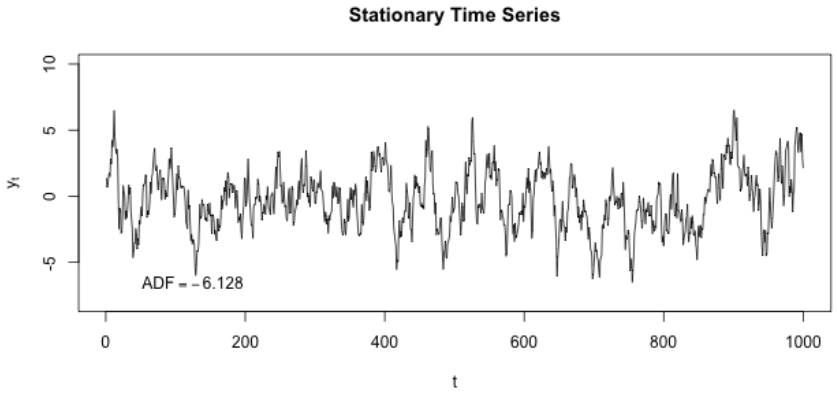
$(\boldsymbol{\gamma_X(r,s) := Cov(X_r, X_s)})$

= 자기공분산은 시차 h에만 의존하고, 시점 t와는 무관하다.

시계열 $\{X_t, t \in \mathbb{Z}\}$ 가 위의 세 가지 조건을 만족할 때, 약정상성을 만족하는 시계열이라 합니다. 강정상성과는 다르게 약정상성은 분포의 동일성을 가정하지는 않습니다. 앞으로 다루게 될 모든 정상 시계열은 이 약정상성을 만족하는 시계열을 의미합니다.

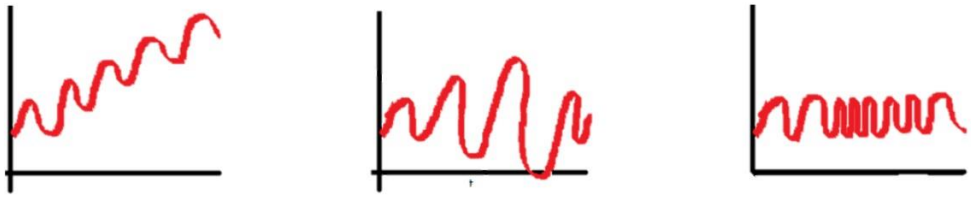
3 정상화

3.1 정상 시계열과 비정상 시계열



위의 시계열은 정상성을 만족하는 시계열로, 특별한 추세나 계절성 없이 평균과 분산이 일정해 보입니다.

[평균이 일정하지 않은 경우] [분산이 일정하지 않은 경우] [공분산이 시점에 의존하는 경우]



[출처] 2022-1 시계열자료분석팀 1주차 교안 by 주혜인님

비정상 시계열은 약정상성 조건을 만족하지 못하는 시계열로, 평균 또는 분산이 일정하지 않은 경우, 공분산이 시점에 의존하는 경우가 있습니다. 이러한 비정상 시계열은 정상화를 통해 정상 시계열로의 변환이 필요합니다.

이제부터 비정상 시계열을 정상화하는 방법에 대해 알아보겠습니다. 구체적으로 알아보기 전에 정상화 과정을 간략하게 정리해보자면 다음과 같습니다.

① 분산이 일정하지 않은 경우 변환을 통해 분산 안정화

② 비정상 시계열 X_t를 추세 m_t, 계절성 s_t, stationary error Y_t로 분해

③ 추세 또는 계절성을 추정하여 제거 (결과적으로, stationary error만 남음)

3.2 분산이 일정하지 않은 경우의 정상화

시간의 흐름에 따라 분산이 달라지는 이분산성을 띄는 경우, 분산 안정화 변환 (Variance Stabilizing Transformation, VST)을 통해 분산이 시간의 흐름에 의존하지 않고 일정하도록 변환할 수 있습니다.

- 로그 변환

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

- 제곱근 변환

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

- Box-Cox 변환

$$f_\lambda(X_t) = \begin{cases} \frac{X_t^\lambda - 1}{\lambda}, X_t \geq 0, \lambda > 0 \\ logX_t, \lambda = 0 \end{cases}$$

3.3 평균이 일정하지 않은 경우의 정상화

$$X_t = m_t + s_t + Y_t$$

$\mathrm{m_t}$: 추세(trend)

$\mathrm{s_t}$: 계절성(seasonality)

$\mathrm{Y_t}$: 정상성을 만족하는 오차($stationary\ error$)

추세와 계절성을 비정상 부분(non-stationary part)이라고 하며, 정상성을 만족하는 오차를 정상 부분(stationary part)이라고 합니다.

평균이 일정하지 않게 되는 원인으로는 추세만 존재하는 경우, 계절성만 존재하는 경우, 추세와 계절성이 모두 존재하는 경우가 있습니다. 지금부터는 비정상 부분을 추정하여 제거하는 방법에 대해 알아보겠습니다.

3.3.1 회귀 (Regression)

1) 추세만 존재하는 경우 ; Polynomial Regression

[1] 시계열을 다음과 같이 가정합니다.

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

[2] 추세 성분 m_t를 다음과 같이 시간 t에 대한 선형회귀식으로 나타냅니다.

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

[3] 위 선형회귀식의 계수를 최소제곱법(OLS)를 통하여 추정합니다.

$$(\widehat{c_o}, \ldots, \widehat{c_p}) = \underset{c}{\mathrm{argmin}} \sum_{t=1}^{n} (X_t - m_t)^2$$

[4] 추정한 추세를 시계열에서 제거하면 정상시계열이 됩니다.

2) 계절성만 존재하는 경우 ; Harmonic Regression

[1] 시계열을 다음과 같이 주기가 d 인 계절성만을 가진다고 가정합니다.

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

$$where\ s_{t+d} = s_t = s_{t-d}$$

[2] 계절 성분 s_t를 다음과 같이 시간 t에 대한 회귀식으로 나타냅니다.

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

[3] 적절한 λ_j와 k를 선택한 후, OLS를 통하여 a_j와 b_j를 추정합니다.

(참고)

λ_j는 주기가 2π인 함수의 주기와 데이터의 주기를 맞춰 주기 위한 값으로,

1) 주기 반복 횟수 $f_1 = [n/d]$ (n = 데이터 개수, d = 주기) → $f_j = jf_1$

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

k는 주로 1~4 사이의 값을 사용합니다.

[4] 추정한 계절성을 시계열에서 제거하면 정상시계열이 됩니다.

3) 추세와 계절성이 모두 존재하는 경우

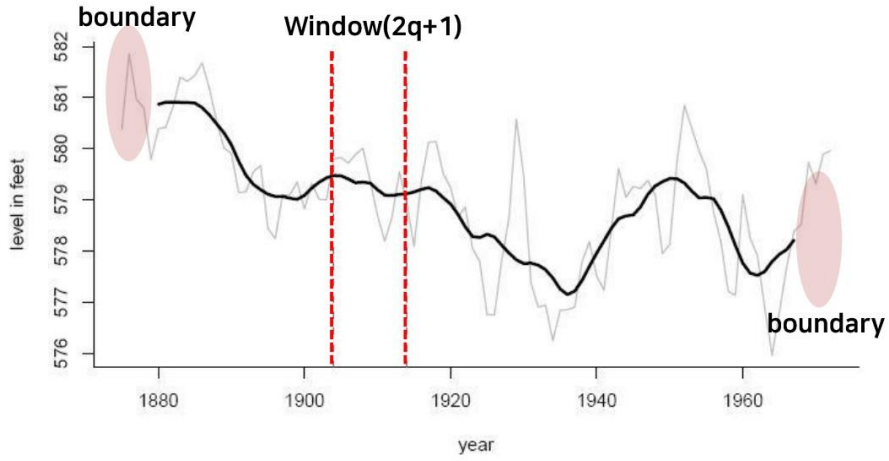
1)과 2)의 과정을 차례대로 진행합니다. 이후에도 남아있는 추세가 보인다면, 다시 추세를 추정하여 제거합니다.

지금까지 회귀를 활용하여 비정상 부분(non-stationary part)을 제거하는 방법에 대해 알아봤습니다. 회귀를 통해 추정한 LSE 자체는 불편추정량이므로 문제가 없지만, OLS와는 다르게 **시계열 자료는 오차항의 독립성을 가정하지 않기 때문에, 추정이 부정확할 수 있다는 문제점**이 존재하다는 것을 기억해야 합니다.

3.3.2 평활 (Smoothing)

회귀는 전체 데이터를 한 번에 추정하기 때문에, **국소적 변동**(local fluctuation)을 표현하기엔 부적절합니다. 이러한 경우 평활 방법을 사용할 수 있습니다.

1) 추세만 존재하는 경우 ; Moving Average Smoothing



[출처] 2022-2 시계열자료분석팀 1주차 교안 by 조웅빈님

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

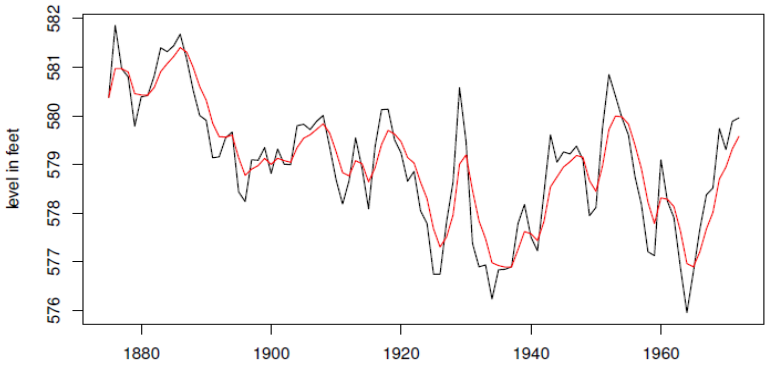
이동평균 평활법은 길이가 2q+1인 구간의 평균을 t 시점에 대한 추정량인 W_t로 사용합니다. 이 때, 파라미터 q로부터 bias-variance trade-off가 발생합니다. 파라미터 q는 cross validation을 통해 찾는 것이 가장 보편적입니다.

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

위 식을 다음과 같이 분해할 수 있습니다. 이 때, 추세는 선형($m_t = c_0 + c_1 t$)임을 가정합니다. 약대수의 법칙에 의해 비정상 부분이 $E(Y_t) = 0$으로 가기 때문에, 해당 부분은 사라지게 됩니다. 결과적으로 W_t는 stationary error를 제거한 순수한 선형의 추세만을 보존한 추정량이 됩니다. 추정한 추세를 제거하면 정상 시계열이 됩니다.

하지만 한계도 존재합니다. 처음 q개와 마지막 q개 시점에서는 추정량을 구할 수 없는 boundary problem이 있습니다. 또한, 현실에서 t 시점에 대한 예측을 위해 t 시점 이후의 데이터를 활용하는 것은 사실상 불가능하다는 문제가 있습니다.

2) 추세만 존재하는 경우 ; Exponential Smoothing



지수 평활법은 추세 $\hat{m}_t$를 t 시점까지의 관찰값만을 이용하여 추정하는 방법입니다. 즉, 과거의 데이터만을 이용하여 추세를 추정한다는 점에서 이동평균 평활법보다 현실적인 접근법이라고 할 수 있습니다.

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

$$\hat{m}_2 = aX_2 + (1-a)\hat{m}_1 = aX_2 + (1-a)X_1$$

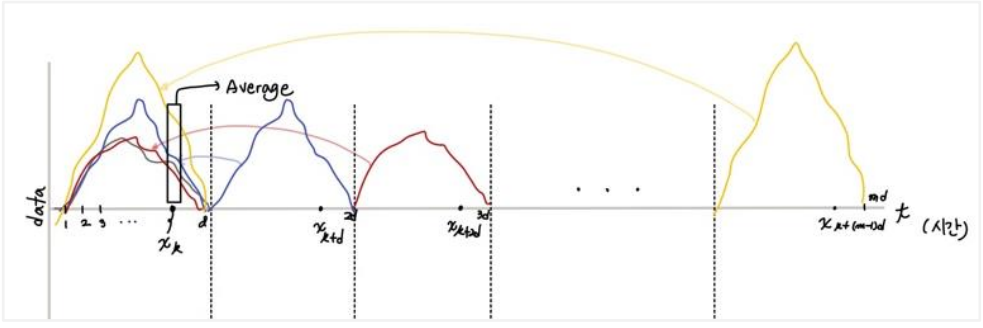
$$\hat{m}_3 = aX_3 + (1-a)\hat{m}_2 = \ aX_3 + a(1-a)X_2 + \ (1-a)^2X_1$$

$$\vdots$$

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

추세에 대한 추정량은 확률변수(X_t)와 이전 시점의 추세 추정량($\hat{m}_{t-1}$)의 가중 평균입니다. 즉, 과거 값에 대한 가중치가 지수적으로 감소합니다. 파라미터 a 역시 cross validation을 통해 가장 적절한 값을 찾아주어야 합니다. 추정된 추세를 제거하면 정상 시계열이 됩니다.

3) 계절성만 존재하는 경우 ; Seasonal Smoothing



[출처] 2022-1 시계열자료분석팀 1주차 교안 by 주혜인님

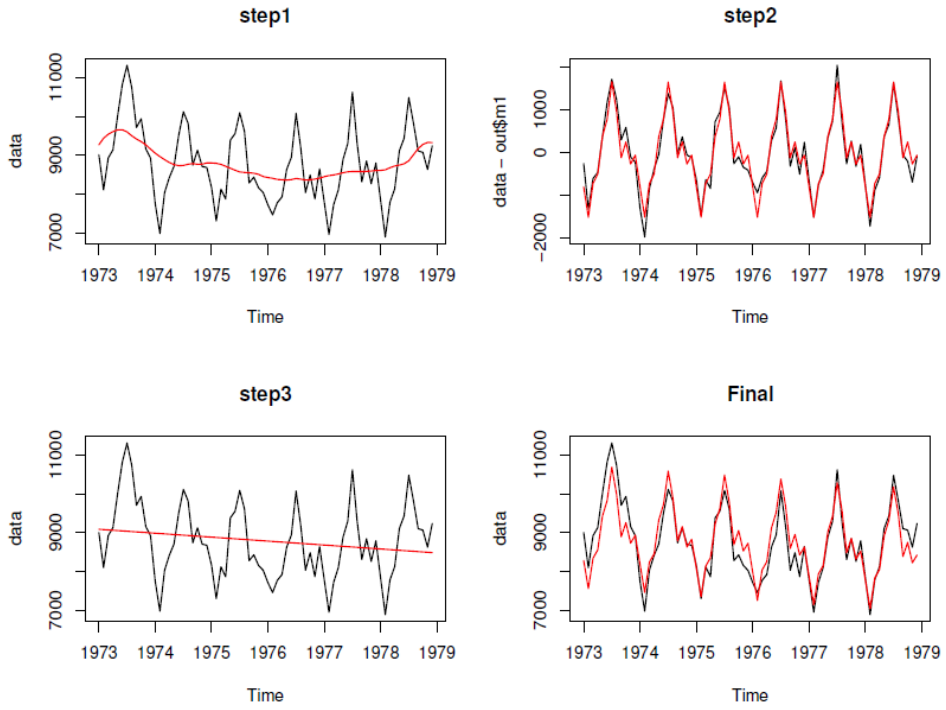
Seasonal 평활법은 동일한 기간 d에 대한 주기의 관측치들을 모두 겹친 다음, 겹쳐진 값들을 각 시점마다 평균 내어 계절성을 추정하는 방식입니다.

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

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

k 시점의 계절성의 추정량($\hat{s}_k$)은 d 간격만큼 떨어진 데이터들의 평균입니다. 추정된 계절 성분($\hat{s}_k$)은 다른 주기에도 동일하게 적용이 되며, 이를 시계열에서 제거하면 정상 시계열이 됩니다.

4) 추세와 계절성이 모두 존재하는 경우 ; Classical Decomposition Algorithm



[1] 먼저 MA filter를 이용하여 추세를 추정합니다.

$$if\ d = 2q\ (\text{짝수}),\ \hat{m}_t = \frac{0.5X_{t-q} + X_{t-q+1} + \cdots + X_{t+q-1} + 0.5X_{t+q}}{2q}$$

$$if\ d = 2q+1\ (\text{홀수}),\ \hat{m}_t = \frac{X_{t-q} + X_{t-q+1} + \cdots + X_{t+q-1} + X_{t+q}}{2q+1}$$

[2] 위에서 추정한 추세를 제거한 후, seasonal smoothing으로 계절성을 추정합니다.

[3] 추정한 계절성을 제거한 후, OLS를 활용하여 다시 추세를 추정합니다. (추세를 추

정하기 위해 OLS 이외의 방법을 사용해도 됩니다.)

[4] 다시 추정한 추세를 제거합니다. 이러한 과정을 더 반복할 수 있습니다.

지금까지 평활법을 활용하여 비정상 부분(non-stationary part)을 제거하는 방법에 대해 알아보았습니다. 정리하자면, 평활법은 회귀 방법보다 국소적 변동을 잘 표현한다는 장점이 있지만, q와 a 등 파라미터를 결정하는 점이 까다롭다는 단점이 있다는 것을 기억해두면 좋을 것 같습니다.

3.3.3 차분 (Differencing)

차분에서는 '후향 연산자 (B)'라는 새로운 연산자가 등장합니다. 후향 연산자는 식이 복잡해지는 것을 방지하기 위해 과거 시점을 간단하게 표현해주는 역할을 합니다.

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

그렇다면 차분이란 무엇일까요? 차분은 관측값들의 차이를 구하는 것입니다.

[1차 차분] $\boldsymbol{\nabla X_t} = X_t - X_{t-1} = \boldsymbol{(1-B)X_t}$

[2차 차분] $\boldsymbol{\nabla^2 X_t} = \nabla(\nabla X_t) = \nabla(X_t - X_{t-1}) = X_t - 2X_{t-1} + X_{t-2} = \boldsymbol{(1-B)^2 X_t}$

1) 추세만 존재하는 경우 ; Differencing

추세를 $m_t = (c_0 + c_1 t)$과 같이 선형으로 가정하겠습니다.

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

추세에 1차 차분을 적용한 결과, 시간 t에 영향을 받지 않는 상수만 남은 정상 시계열이 됩니다. 일반적으로 k차 차분을 적용하면 k차의 추세를 제거할 수 있습니다.

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

2) 계절성만 존재하는 경우 ; lag-d Differencing

lag-d differencing에 사용되는 연산자는 다음과 같이 정의됩니다.

$$\nabla_d X_t = (1 - B^d) X_t = X_t - B^d X_t, \qquad t = 1, \dots, n$$

d차 차분은 $\nabla^d = (1 - B)^d$이고, lag-d 차분은 $\nabla_d = (1 - B^d)$로 표현하는데, 헷갈릴 수 있으니 주의해야 합니다!

$s_t = s_{t+d}$를 가정하고, 계절성만 있는 비정상 시계열 $X_t = s_t + Y_t$ 에 lag-d 차분을 적용해보겠습니다.

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

위와 같이 계절성이 제거되고 오차항만 남는 정상 시계열이 됩니다.

3) 추세와 계절성이 모두 존재하는 경우 ; lag-d 차분 + p차 차분

추세 및 계절성이 동시에 존재하는 비정상 시계열의 경우, 계절 차분 (lag-d 차분) + p차 차분 (p=추세의 차수)을 적용할 수 있습니다.

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

하지만 주의해야 할 점이 있습니다. 계절 차분(lag-d 차분)을 인수분해하면 위와 같이 표현할 수 있습니다. 이 때, 계절 차분 자체에 1차 차분이 포함되어 있기 때문에, p차의 추세를 제거하고자 할 때 p-1차의 차분을 적용해야 합니다.

4 정상성 검정

우리는 지금까지 비정상 시계열을 정상화하기 위해, 시계열을 비정상 부분(추세, 계절성)과 정상 부분(stationary error Y_t)으로 분해한 다음, 각 비정상 부분을 제거하는 세 가지 방법인 회귀(regression), 평활(smoothing), 차분(differencing)에 대해 공부했습니다.

비정상 부분을 제대로 제거했다면, 시계열에는 정상성을 만족하는 오차인 stationary error Y_t만 남아있어야 합니다. 다시 한 번 짚고 넘어가자면, 오차가 정상성을 만족하려면 평균과 분산이 일정하고, 시계열의 확률적 성질(공분산 함수)이 시간 t에 의존하지 않고 시차 h에만 의존해야 합니다.

지금부터는 정상화 과정을 거친 다음 오차가 정상성을 만족하는지 검정하는 방법에 대해 알아보겠습니다.

4.1 자기공분산함수(ACVF), 자기상관함수(ACF)

시계열 자료의 공분산 함수가 시차 h에만 의존함을 확인하기 위해서, 시간에 따른 상관 정도를 나타내는 ACVF와 ACF를 확인해야 합니다.

[자기공분산함수 ACVF (auto-covariance function)]

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

ACVF는 symmetric matrix이며, n.n.d를 만족한다는 특성을 기억해두면 좋습니다.

[표본자기공분산함수 SACVF (sample auto-covariance function)]

$$\boldsymbol{\hat{\gamma}_x(h) = \frac{1}{n}\Sigma_{j=1}^{n-h}\left(X_j - \bar{X}\right)\left(X_{j+h} - \bar{X}\right)}$$

[자기상관함수 ACF (auto-correlation function)]

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

[표본자기상관함수 SACF (sample auto-correlation function)]

$$\boldsymbol{\widehat{\rho_x}(h) = \frac{\hat{\gamma}_x(h)}{\hat{\gamma}_x(0)},\ \ \hat{\rho}(0) = 1}$$

4.2 White Noise sequence (백색잡음)

WN sequence에 대해 알아보기 전에 IID process에 대해 먼저 알아보겠습니다. IID는 Independent Identically Distribution으로, 각 확률변수가 독립성을 만족하면서도 동일한 분포를 따른다는 가정입니다. IID process는 IID 가정을 따르면서 평균이 μ이고 분산이 σ^2인 확률변수들의 집합을 의미합니다.

그렇다면 WN sequence란 무엇일까요? WN sequence는 평균이 0이고 분산이 σ^2인 확률변수의 집합이며, IID와 비교하여 독립성 조건을 완화(확률 구조를 알 필요가 없음)하였지만 **확률변수 간에 상관관계가 없어야 합니다**.

$$Cov(X_t, X_s) = 0, if\ t \neq s$$

IID sequence와 WN sequence 모두 대표적인 정상 시계열입니다. IID면 WN이지만, 그 역은 성립하지 않습니다.

4.3 백색잡음 검정

비정상 시계열에 대해 정상화 과정을 거친 다음, 남은 오차항이 WN sequence의 조건을 만족한다면 해당 시계열은 정상 시계열입니다. 이 경우 정상성을 만족하기 때문에 추가적인 모델링 과정을 더 진행할 필요 없이, 분산 $\sigma^2 = r(0)$ 만 추정하면 됩니다. 클린업 2주차에서 더 자세하게 다루겠지만, 만약 오차항이 정상성을 만족하지 않는다면 non-stationary error에 대한 추가적인 모델링이 필요합니다.

따라서 오차항이 WN sequence의 조건을 만족하는지 검정하는 방법을 알아보겠습니다. 백색잡음 검정은 자기상관, 정규성, 정상성에 대한 검정을 거칩니다.

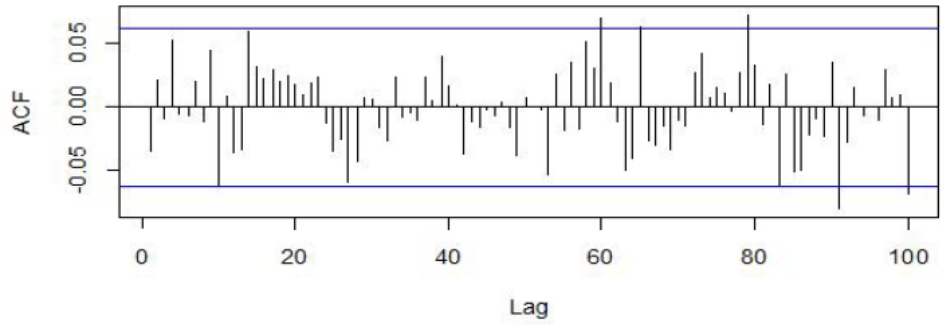
1) 자기상관의 유무 확인

$$\hat{\rho} \approx \mathcal{N}\left(0, \frac{1}{n}\right)$$

오차 Y_t가 백색잡음 WN(0, 1)을 따른다면, 표본자기상관함수 (sample auto-correlation function, SACF) $\hat{\rho}(h)$는 평균이 0이고 분산이 1/n인 정규분포로 근사합니다. ($\hat{\rho}(h)$는 시차 h만큼 떨어진 순서쌍들이 얼마나 선형 상관관계를 보이는지를 설명합니다!) 이러한 사실을 바탕으로 다음의 가설을 테스트합니다.

$$H_0 : \rho(h) = 0 \ vs \ H_1 : \rho(h) \neq 0$$

만약 $|\hat{\rho}(h)|$가 $1.96/\sqrt{n}$ 의 범위 내에 있다면, 귀무가설을 기각하지 못하므로 **오차항에 자기상관이 없다고 할 수 있습니다**. ACF 그래프 (=correlogram)를 이용해 시각적으로 확인 가능합니다.



이 때, h는 최소 n/4보다 크게 설정하는 것이 권장됩니다. h는 시점 사이의 간격을 의미하는데, 간격이 너무 넓으면 SACF에 대한 신뢰성이 떨어지기 때문입니다. 이외에도 Portmanteau 검정, Ljung-Box 검정, McLeod-Li 검정 등을 활용할 수 있습니다.

2) 정규성을 만족하는지 확인

$$H_0 : 정규성이\ 존재한다 \quad vs \quad H_1 : 정규성이\ 존재하지\ 않는다$$

QQ plot : 표본의 quantile과 정규분포의 quantile을 시각적으로 비교

KS test : 표본의 누적확률분포와 모집단의 누적확률분포가 얼마나 유사한지 비교

Jarque-Bera test : 왜도와 첨도의 정규분포로서의 적합도를 검정

위의 검정 결과 오차항이 정규성을 만족하지 않는다면, Box-Cox 변환과 같은 변환 기법을 활용하여 정규분포에 가깝게 만들어줄 수 있습니다. (또는 t분포와 같이 다른 분포를 가정하여 MLE를 구하는 방법도 가능합니다.)

3) 정상성을 만족하는지 확인

Kpss test : 귀무가설로 정상 시계열임을 가정

ADF test : 대립가설로 정상 시계열임을 가정

PP test : 이분산이 있을 때도 사용 가능한 검정법, 대립가설로 정상 시계열임을 가정

고생하셨습니다! (따봉) 어질어질한 시계열자료분석팀 1주차 클린업이 끝났습니다. 오늘 공부한 내용을 간략하게 정리하고 마무리하겠습니다.

시계열 자료는 보통 **dependency**를 가집니다. 시계열 자료의 확률 분포를 보다 쉽게 추정하기 위해, 확률적 성질(공분산)이 시간의 흐름에 의존하지 않고 시차(lag)에만 의존한다는 **정상성 가정**을 합니다. 대부분의 정상성은 **약정상성**을 의미하는데, **평균과 분산이 일정하고, 자기공분산(ACVF)이 시차에만 의존**해야 합니다. 대부분의 시계열은 의존성이 있는 비정상 시계열인데, 이를 nonstationary part (추세, 계절성)과 stationary part (정상성을 만족하는 오차)로 나누어 추세와 계절성을 추정하여 제거해주는 **정상화**를 거쳐야 합니다. 정상화 방법으로는 **회귀, 평활, 차분**이 있습니다. 비정상 부분을 제거한 다음에는 남은 오차항이 **WN sequence**인지 확인하는 **정상성 검정**을 해야 합니다. 특히 자기상관을 확인하기 위해 correlogram (=ACF 그래프)를 확인합니다. (1주차 완!!!)