시계열자료분석 클린업 1 주차

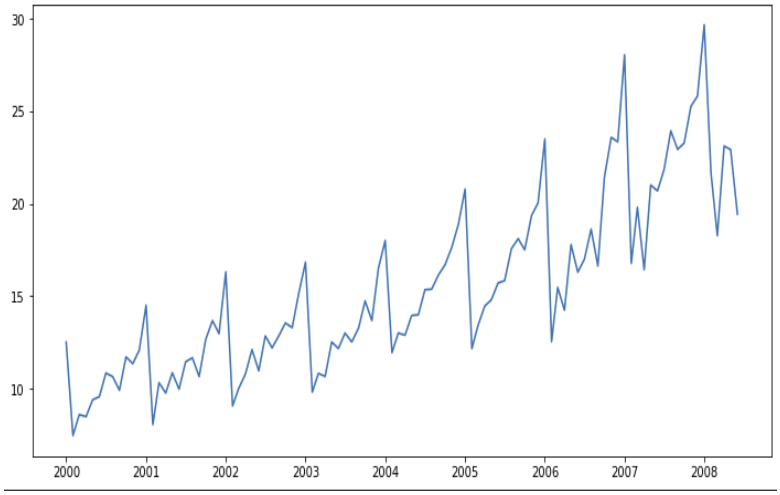
[목차]

시계열자료분석 1주차 클린업에 오신 걸 환영합니다! 3주 동안 시계열 클린업을 진행할 장다연입니다~ 여기 계신 분들 모두가 시계열에 흥미를 가지고 와주셨으니 보답하는 마음으로 최선을 다해, 쉽게 설명해보도록 하겠습니다. 유익한 한 학기, 유익한 클린업이 되기를 바라며 출발해볼까요?!

1 시계열자료분석 알아보기

1.1 시계열 자료란?

시계열 자료(time series)란 시간 순서에 따라 관측된 자료의 집합을 의미합니다. 일반적으로 $\{X_t,\ t = 1,2,3,\dots\}$로 표현하며 시간 t가 이산형인 경우 이산형 시계열 자료, 연속형인 경우 연속형 시계열 자료로 구분합니다. 대표적 시계열 자료인 주식으로 예를 들어보면, 매일매일의 주식 종가는 이산형 시계열 자료, 하루의 주식 가격은 연속형 시계열 자료라고 할 수 있습니다.



1.2 시계열 자료의 특징

시계열 자료는 시간에 따라 관측된 자료이기 때문에 시간의 흐름이 반영되어 관측치(observation)들 사이에 **연관성(dependency)**이 존재합니다. 따라서 특정 시점에 대한 확률변수 X_t의 분포는 하나의 관측치만을 고려하는 것이 아니라, 전체 시점에서의 관측치 집합 $\{x_1, x_2, \dots\}$을 모두 고려한 **결합 분포(joint distribution)**입니다. 지금까지 우리가 다뤄왔던 자료들과 달리 독립성 조건을 만족하지 않는 것입니다. 이러한 이유로 시계열 자료는 일반 선형회귀를 사용할 수 없고, 데이터의 특성을 반영할 수 있는 특별한 분석법이 필요합니다. 그 분석법이 바로 우리가 공부할 시계열 분석입니다.

시계열 분석이란 시계열 자료와 추세 분석을 다루는 통계 기법으로, 시간 순으로 정렬된 데이터에서 관계를 찾아내고 의미 있는 요약과 통계 정보를 추출하는 과정을 의미합니다. 시계

열 분석의 목적은 여러 가지가 있지만, 클린업에서는 미래 예측을 중심으로 다루려고 합니다. 그럼 지금부터 시계열에 대해 더 자세히 알아볼까요?!

1.3 시계열 자료의 구성요소

1) 추세 변동(Trend)

- 시간이 흐름에 따라 관측치가 증가하거나 감소하는 추세를 갖는 변동
- 특별한 충격이 없는 한 지속되는 특성이 있음

2) 순환 변동(Cycle)

- 일정한 주기를 가지고 변화하지만 규칙적으로 발생하지는 않는 변동
- 경기 침체와 회복과 같이 경제적, 사회적 요인에 의해 발생해 예측 어려움

3) 계절 변동(Seasonal variation)

- 규칙적인 주기를 가지고 발생하는 변동
- 주별, 월별, 계절별과 같이 특정 시간 간격을 가지고 반복됨
- 환경적인 요인에 의해 발생해 예측 및 처리에 용이함

4) 우연 변동 / 불규칙 성분 (Random fluctuation)

- 무작위 원인에 의해 나타나 일정한 규칙성을 인지할 수 없는 변동

추세 변동, 순환 변동, 계절 변동을 규칙 요소라고 하며, 우연 변동과 불규칙 성분을 불규칙 요소라고 이야기합니다. 3주에 걸친 클린업을 통해 더 자세히 배우겠지만, 시계열 분석에서는 위 4가지 구성요소를 분해하여 미래를 예측하고자 합니다.

1.4 시계열 분해(Time Series Decomposition)

시계열 분석의 이해를 위해 시계열 분해에 대해 간단히 공부해보겠습니다.

시계열 자료의 분해란 시계열 자료를 비정상 부분(non-stationary part)와 정상 부분(stationary part)으로 분해하는 작업입니다. 추세(m_t)와 계절성(s_t)을 비정상 부분, 오차(Y_t)를 정상 부분이라고 합니다.

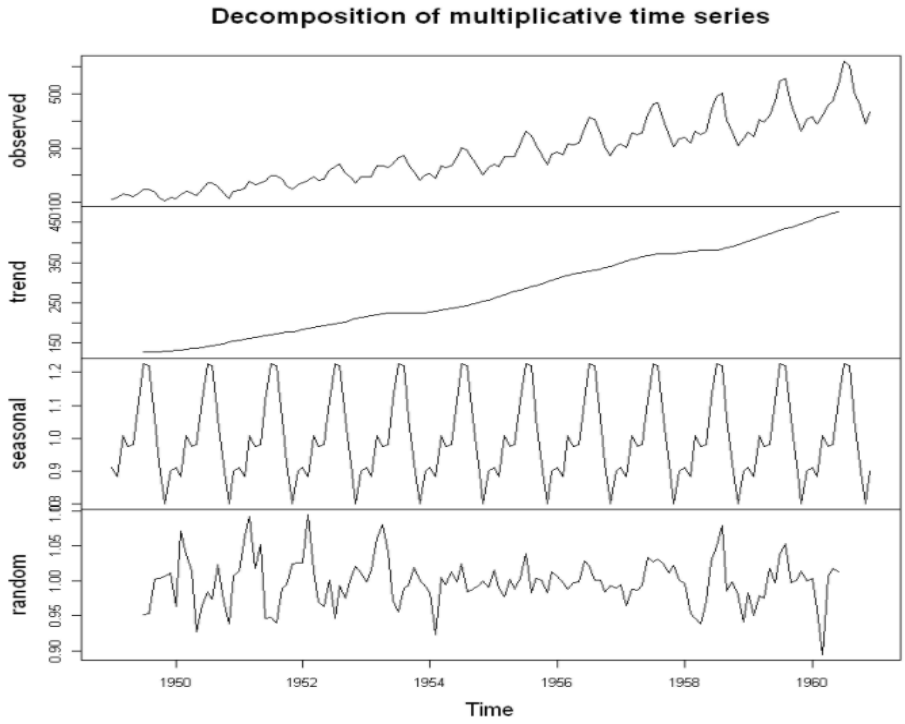
시계열 분해는 크게 덧셈 분해와 곱셈 분해로 나눌 수 있습니다.

1) 덧셈 분해(additive decomposition)

$$X_t = m_t + s_t + Y_t$$

m_t : 추세(trend), s_t : 계절성(seasonality), Y_t : 오차(stationary error)

덧셈 분해는 위 식과 같이 시계열 자료를 구성 요소로 분해하는 것을 의미합니다. (아직 정상성을 배우지 않아 Y_t를 오차라고만 표기하였지만, 원칙상으로 Y_t는 정상성을 만족해야 합니다.) 정상성을 배운 후 더 자세히 배우겠지만, 시계열 분석에서는 m_t (추세)와 s_t(계절성)을 제거한 후 정상성을 만족하는 오차만을 이용해 예측 모델링을 진행합니다. 덧셈 분해의 과정은 아래 그림을 통해 시각적으로 확인할 수 있습니다.



2) 곱셈 분해(multiplicative decomposition)

$$X_t = m_t * s_t * Y_t$$

m_t : 추세(trend), s_t : 계절성(seasonality), Y_t : 정상성을 오차(stationary error)

위와 같이 구성 요소의 곱으로 시계열 자료를 분해하는 방법이 곱셈 분해입니다. 곱셈 분해를 사용하기 위해서는 데이터에 0이 포함되는지를 확인해야 합니다. 만약 존재한다면 곱셈 분해를 사용할 수 없습니다.

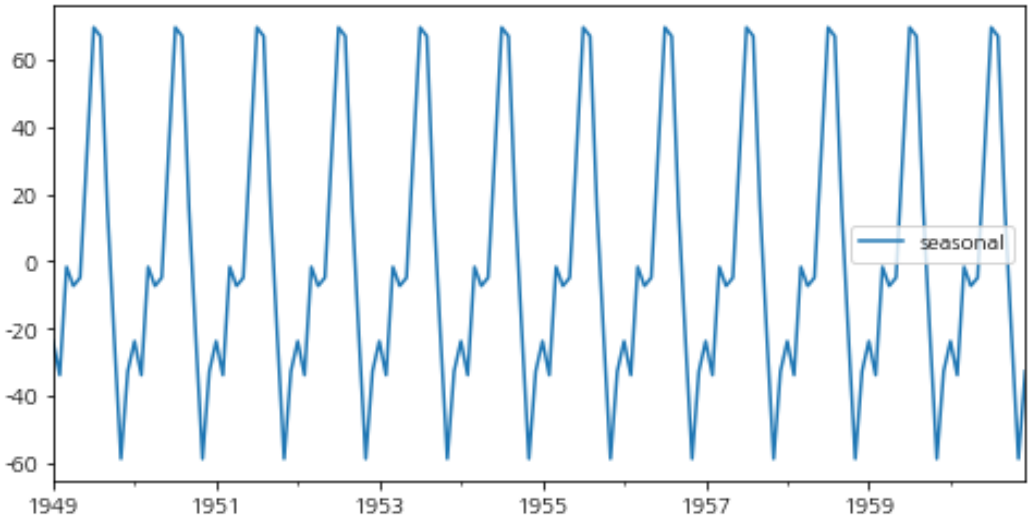
덧셈 분해와 곱셈 분해의 차이점은 추세와 계절성의 관계를 통해 정의할 수 있습니다. 덧셈 분해는 추세와 계절성을 별개의 구성 요소로 보지만, 곱셈 분해는 추세에 따라 계절성이 변화함을 가정합니다. 클린업에서는 덧셈 분해를 이용한 시계열 분석법에 대해 공부하도록 하겠습니다.

2 정상성(Stationarity)

2.1 정상성이란?

시계열 분석을 진행할 때 가장 먼저 고려해야 하는 것이 바로 **정상성(stationarity)**입니다. 정상성이란 시계열 자료의 확률적 성질이 시점 t에 의존하지 않고 시차 lag에만 의존하는 특성을 의미합니다. 쉽게 이야기하자면 시간의 흐름에 따라 평균과 분산이 변하지 않는 것입니다.

그렇다면 시계열자료분석에서 정상성이 왜 중요한 것일까요? 1.2에서 배운 것처럼 시계열 자료는 하나의 시점만을 고려하는 것이 아니라 모든 관측치를 고려하는 **결합 분포**(joint distribution)입니다. 따라서 우리가 미래 X_{t+1}을 예측하고자 한다면 우리가 관측하지 못한 x_{t+1}까지를 포함한 $\{x_1, x_2, x_3, \dots, x_t, x_{t+1}\}$의 결합 분포를 알아야 합니다. 하지만 현실적으로 모든 시점에 대해 각각의 결합 분포를 구하는 것은 쉽지 않겠죠? 이때 정상성을 가정한다면 더욱 쉽게 구할 수 있습니다. 아래 그림을 통해 자세히 알아보도록 하겠습니다.



위 그림은 일반적은 시계열 그래프입니다. 이렇게 일정한 간격을 두고 반복되는 데이터라면, 즉 확률적 성질이 시차에만 의존한다면 시차 내에서의 분포를 구해 편리하게 전체 분포를 예측할 수 있습니다.

2.2 강정상성(Strict stationarity)

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

위 식처럼 일정한 시차 간격을 가지는 관측치 집합들이 모두 같은 분포를 따른다는 조건을 만족한다면 해당 시계열자료는 **강정상성**을 만족합니다. 다시 말해 강정상성이란, 위에서 설명했던 확률적 성질이 시차에만 의존한다는 정상성의 정의를 만족하는 정상성입니다. 하지만 현실에서 이러한 지나치게 엄격한 조건을 만족하는 데이터는 많지 않습니다.

그렇다면 강정상성에 **정규성**(Gaussianity)을 가정하면 어떨까요?

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

시계열 자료가 정규분포를 따른다고 가정하면, 우리는 평균 벡터인 μ와 공분산 행렬 Σ만 추정해서 전체 분포를 구할 수 있습니다. 이 가정을 통해 2가지를 알아낼 수 있습니다.

1) 확률변수의 기댓값은 상수(constant)이다.

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

2) 확률변수의 공분산은 시차에 의존한다.

➔ $Cov(X_r, X_s) = Cov(X_{r+h}, X_{s+h}), \ \forall r, s, h \in \mathbb{Z}$

위 과정을 통해 강정상성을 조건을 완화했지만, 이 역시 여전히 엄격한 가정이기 때문에 현실에서는 약정상성 개념을 사용합니다.

2.3 약정상성(Weak stationarity)

약정상성은 강정상성과 정규성 가정에서 나온 조건들을 더 확장하고 완화한 정상성입니다. 아래의 3가지 조건을 모두 만족한다면 약정상성을 만족하는 시계열입니다.

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

➔ 2차 적률(분산 관련)이 존재하고 시점 t에 관계없이 일정하다

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

➔ 평균이 상수로 시점 t에 관계없이 일정하다

3) $\gamma_X(r, s) = \gamma_X(r + h, s + h), \ \forall r, s, h \in \mathbb{Z}, \quad (\gamma_X(r, s) := Cov(X_r, X_s))$

➔ 공분산은 시차 h에 의존하며 시점 t와 무관하다

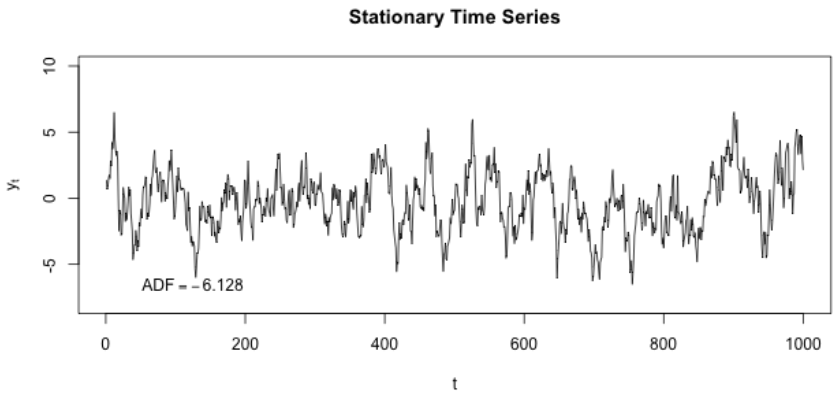
앞으로 우리가 클린업에서 다룰 시계열은 모두 약정상성을 만족하는 시계열입니다.

3 정상화

지금까지 시계열 자료의 기본적인 개념들을 모두 배웠습니다. 이제 본격적으로 시계열 자료를 다루는 방법에 대해 배워보도록 하겠습니다. 분석 또는 예측에 사용할 시계열 자료가 있다면 가장 먼저 정상성을 만족하는 자료인지 파악해야 합니다. 만약 정상성을 만족한다면 바로 분석을 진행할 수 있지만, 만족하지 않을 경우에는 정상화를 진행해야 합니다. 그럼 지금부터 정상성을 만족하는지 판단하는 방법과 정상화 과정에 대해 배워볼까요~?

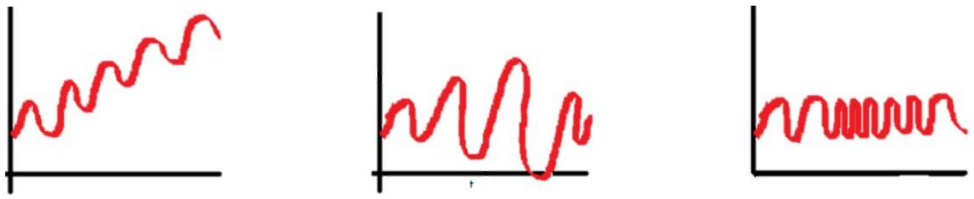
3.1 정상 시계열과 비정상 시계열

시계열 자료가 정상 시계열인지 여부는 시계열 플랏(TS plot)을 통해 시각적으로 확인할 수 있습니다.



위 시계열은 정상성을 만족하는 정상 시계열로, 특별한 추세나 계절성이 보이지 않으며 평균과 분산 역시 일정한 것으로 판단할 수 있습니다.

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



비정상 시계열은 정상성 조건을 만족하지 못하는 시계열입니다. 위 예시와 같이 평균 또는 분산이 일정하지 않거나, 공분산이 시점에 의존하는 시계열이 비정상 시계열에 해당합니다. 비정상 시계열의 경우 바로 분석에 사용할 수 없어 정상 시계열로의 변환인 정상화 과정이 필요합니다.

3.2 정상화가 필요한 이유

'시계열 데이터는 시간의 흐름을 내포하고 있으니까 이를 반영해서 바로 모델링을 진행하면 안 될까?'라는 의문이 생기실 수 있습니다. 정상화 과정을 배우기에 앞서, 정상화가 필요한 이유를 통해 이러한 의문을 해결해보도록 하겠습니다!

정상화가 필요한 이유는 크게 2가지로 볼 수 있습니다.

1) 독립성 조건의 붕괴

1.2에서 시계열 자료는 오차의 독립성 조건을 만족하지 않는다고 배웠습니다. 이럴 경우 우리가 흔하게 알고 있는 큰 수의 법칙, CLT, 선형회귀 등의 방법들을 사용할 수 없습니다. 따라서 시계열의 분해와 정상화를 통해 상관관계가 존재하지 않는, 정상성을 만족하는 오차만 남긴 후 모델링을 진행하는 것입니다.

2) 안정적이고 정확한 예측

정상성의 정의를 다시 떠올려보자면, 자료의 확률적 성질이 시점 t가 아닌 시차 lag에만 의존하는 것이었습니다. 정상성을 만족하지 않는 데이터를 사용한다면 데이터를 설명하는 모델의 정확도가 시점에 따라 달라질 수 있습니다.

간단히 정리하자면, 더욱 안정적이고 정확한 예측을 하기 위해 정상화를 진행한다고 이해하시면 되겠습니다.

3.3 분산이 일정하지 않은 경우

약정상성의 조건 중 분산이 시점 t에 의존하지 않고 일정하다는 조건이 있었습니다. 이 조건을 만족하지 못하는 경우에는 **분산 안정화 변환** (Variance Stabilizing Transformation, VST)를 통해 분산을 안정화해야 합니다.

1) Log-transformation

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

2) Square root transformation

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

3) Box-Cox transformation

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

3.4 평균이 일정하지 않은 경우

평균이 일정하지 않게 되는 원인은 총 3가지가 있는데, 추세만 존재하는 경우, 계절성만 존재하는 경우, 추세와 계절성이 모두 존재하는 경우입니다. 일반적으로 회귀, 평활, 차분의 3가지 방법을 통해 비정상 부분을 추정하고 제거하여 정상화를 진행합니다. 그럼 지금부터 이 3가지 방법을 각각 알아볼까요?

1) 회귀(Regression)

a. 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.

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

[4] 추정한 추세를 제거하면 정상 시계열이 됩니다.

b. 계절성만 존재하는 경우 : Harmonic Regression

[1] 시계열을 다음과 같이 주기가 d인 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.

(예시)

n=72, d=12 ---> f_1 = [72/12] = 6, λ_j = j x 6 x $2\pi/72$

[4] 추정한 계절성을 제거하면 정상 시계열이 됩니다.

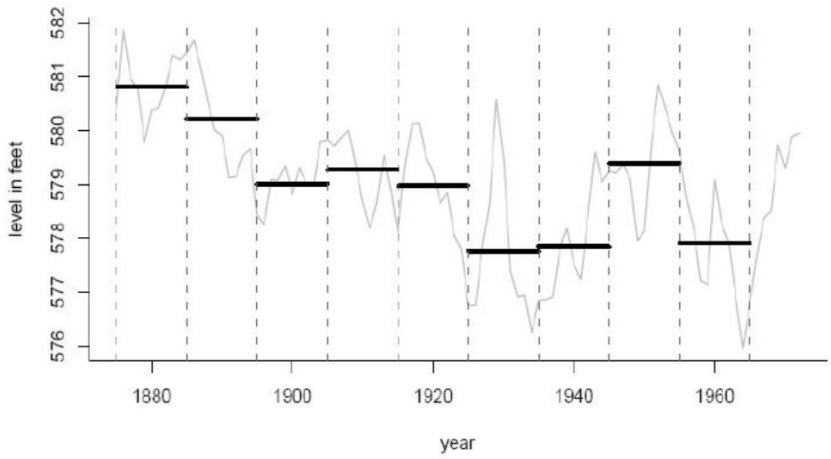
c. 추세와 계절성이 모두 존재하는 경우

a와 b의 과정을 차례대로 진행합니다. 이후에도 남아있는 추세가 보인다면, 같은 과정을 반복해 제거합니다.

지금까지 회귀를 이용하여 비정상 부분을 제거하는 방법에 대해 알아봤습니다. 하지만 계속해서 언급한 것처럼 시계열 자료는 독립성을 가정하지 않기 때문에 회귀분석의 기본 가정을 만족하지 않아 추정이 부정확할 수 있다는 것을 기억해야 합니다.

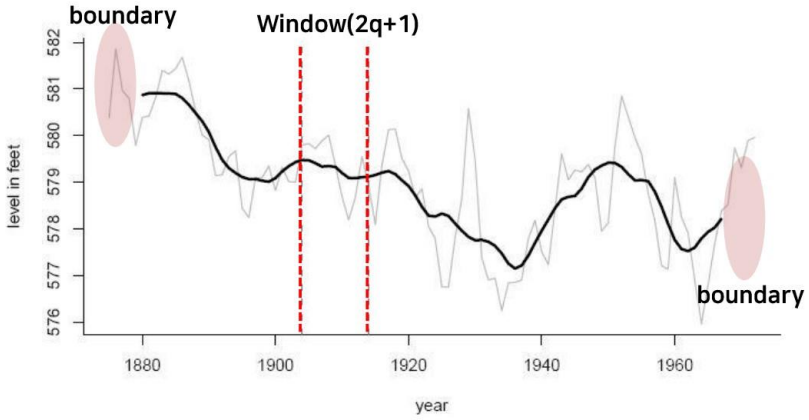
2) 평활 (Smoothing)

회귀는 전체 데이터를 한 번에 추정하는 방법이기 때문에, 국소적 변동(local fluctuation)이 존재하는 경우에 사용하기에는 부적절합니다. 이처럼 국소적 변동에 주목해야 하는 경우에는 평활 방법을 사용할 수 있습니다.



평활법은 시계열 자료를 여러 구간으로 나눈 후, **구간의 평균들로 추세를 추정**하는 방법입니다. 위 그림을 통해 알 수 있는 것처럼, 전체 시계열 자료와 구간 평균의 움직임은 비슷할 것이라는 아이디어를 이용한 방법입니다.

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



[1] 길이가 2q+1인 구간의 평균을 구합니다.

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

[2] 위 식에 추세 성분 m_t를 대입합니다. 이때 추세는 linear함을 가정하겠습니다.

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

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

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

위 식에 의해 구간의 평균 W_t는 근사적으로 추세 m_t와 같아집니다.

[3] 위 과정을 통해 추세부분만 남은 W_t를 X_t에서 제거합니다.

이동평균 평활법은 국소적 변동을 설명할 수는 있지만, 데이터의 맨 앞 q개와 맨 뒤 q개의 **boundary**는 값을 추정할 수 없다는 큰 약점을 가진 방법입니다. . 또한 현실에서는 미래의 관측값을 사용할 수 없기 때문에, 과거의 데이터만을 가지고 추세를 제거할 수 있는 방법인 지수 평활법에 대해 알아보도록 하겠습니다.

b. 추세만 존재하는 경우 : Exponential Smoothing

지수 평활법은 미래의 데이터를 활용하지 않고, t 시점까지의 관측값만을 이용하여 추세를 추정하고 제거하는 보다 더 현실적인 방법입니다.

[1] 다음 과정을 통해 추세를 추정합니다.

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

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

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

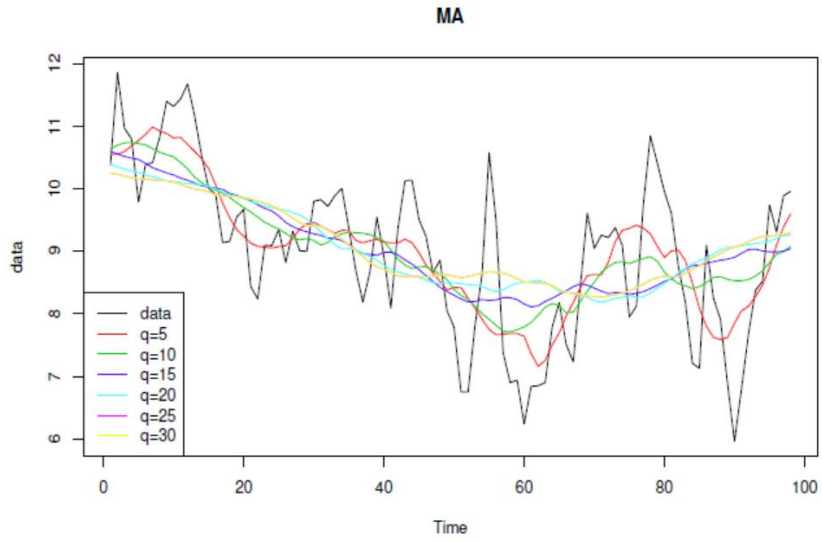
$$\vdots$$

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

[2] 추정한 추세를 시계열에서 제거합니다.

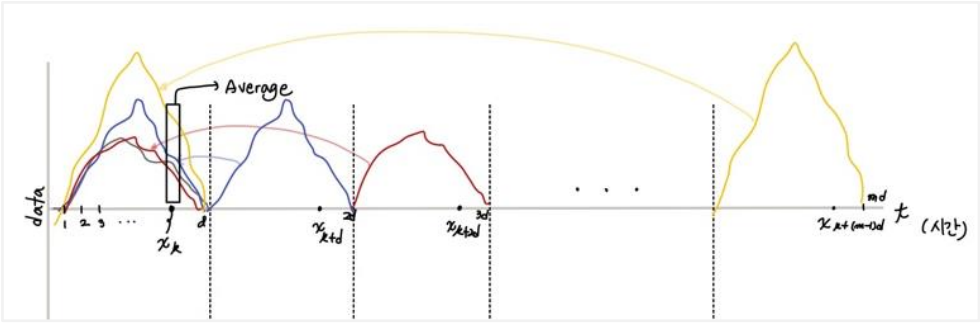
위 과정에서 $a \in [0,1]$ 이며, 과거 관측치에 대한 가중치입니다. 식을 통해 알 수 있는 것처럼 과거의 관측치일수록 가중치의 값이 지수적으로 감소하는 방식으로 추세를 추정하는 방법입니다.

이동평균 평활법과 지수 평활법 모두 추세 외에도 파라미터 q와 a를 추정해야 합니다. 일반적으로 q가 작으면 작은 변화들도 잘 잡아낼 수 있지만, 그만큼 변동성이 심해집니다. 반대로 q가 클 경우 변동성은 줄어들지만, 작은 변화들을 잡아내지 못합니다. 즉 **bias-variance trade off**가 발생하는 것입니다. 따라서 cross-validation(CV)를 통해 MSE를 추정하여 최적의 파라미터를 선택해야 합니다.



c. 계절성만 존재하는 경우 : Seasonal Smoothing

Seasonal Smoothing은 주기가 d인 시계열 자료가 있을 때 주기만큼의 데이터들을 모두 겹친(overlay) 후, 겹쳐진 값들의 평균으로 계절성을 추정하는 방법입니다.

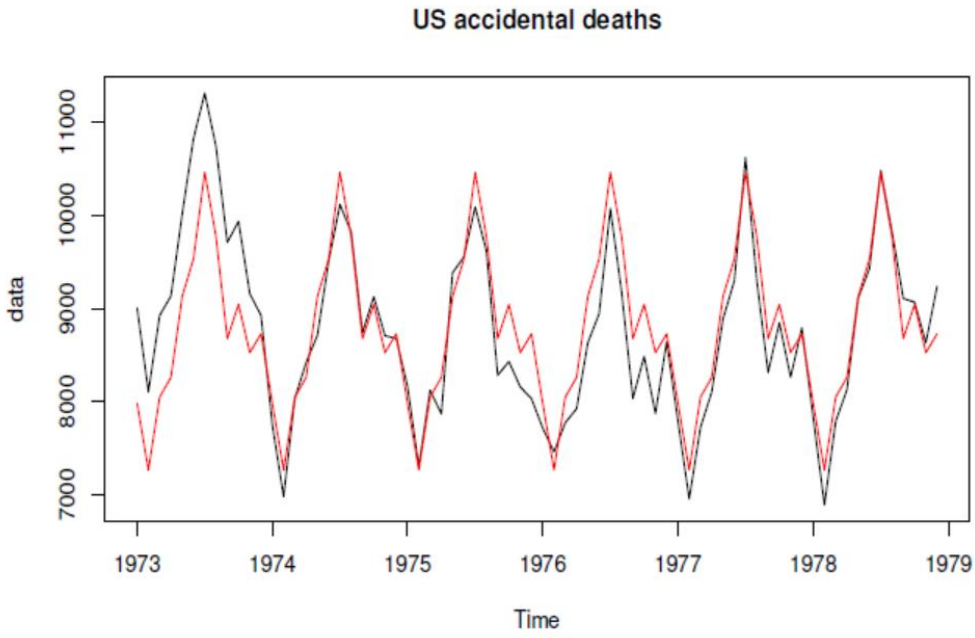


[1] 아래 식을 통해 $k = 1, \cdots, d$에 대한 계절성분$(\hat{s}_k)$을 추정합니다.

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

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

[2] 추정된 계절 성분을 다른 주기에도 적용해 전체 계절성을 추정하고, 이를 시계열 자료에서 제거합니다.



위 그림처럼 빨간색으로 그려진 계절 성분은 모든 주기에서 동일하게 반복되고 있습니다.

d. 추세와 계절성이 모두 존재하는 경우 : Classical Decomposition Algorithm

[1] 먼저 MA filter를 이용하여 추세를 추정합니다.

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

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

(Window를 주기 d와 같게 하는 이유는 계절성의 영향을 피하기 위함!)

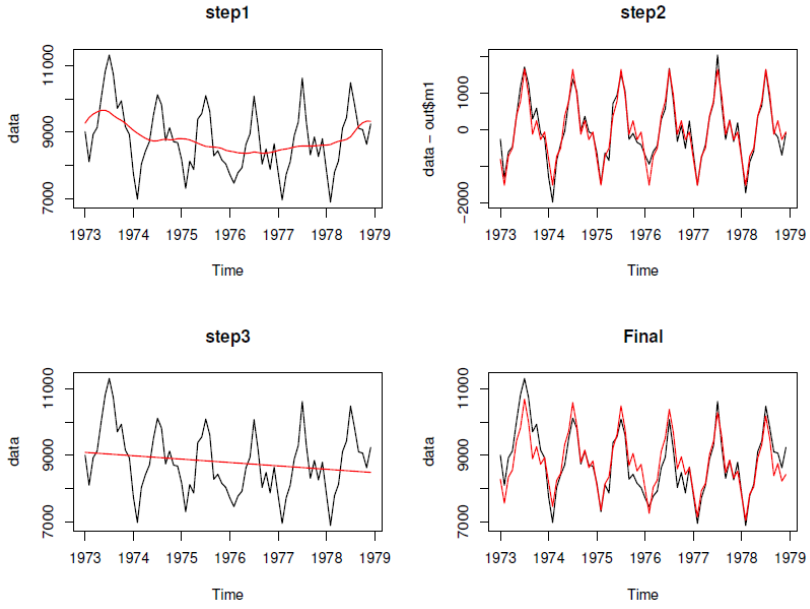
[2] 추정한 추세를 제거한 후, seasonal smoothing으로 계절성을 추정합니다.

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

(OLS 대신 Smoothing을 사용해도 되지만, 일반적으로 OLS를 사용합니다.)

[4] 다시 추정한 추세를 제거합니다.

이후에도 추세 혹은 계절성이 존재한다면 [1]~[4]의 과정을 반복합니다.



3) 차분(Differencing)

차분을 이해하기 위해서 새로운 연산자를 먼저 소개하도록 하겠습니다. 차분에서 사용되는 **후향연산자**(Backshift Operator) $\boldsymbol{B}$는 관측값을 한 시점 전으로 돌려주는 역할을 하는 연산자입니다.

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

후향연산자를 이해했다면, 이제 차분이 무엇인지 알아보겠습니다. 차분이란 관측값들의 차이를 구하는 것입니다. 즉, 차이를 통해 추세, 계절성을 제거할 수 있는 방법입니다. 몇 개의 시점을 이용하는지에 따라 아래와 같이 계산할 수 있습니다.

[1차 차분]

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

[2차 차분]

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

a. 추세만 존재하는 경우 : Differencing

차분을 통해 추세를 제거해보겠습니다. 이때 추세는 선형이라고 가정하겠습니다.

$$m_t = c_0 + c_1 t$$

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

추세를 제거한 결과 시간 t에 영향을 받지 않는 상수만 남아 추세가 제거되었음을 확인할 수 있습니다. 일반적으로 k차 차분을 진행하면, k차 추세 (k-th order polynomial trend)를 제거할 수 있습니다.

차분을 통해 추세를 제거하는 방법은 다른 방법들에 비해 직관적이지만, 아래 식과 같이 오차까지 차분되어 식이 복잡해질 수 있음을 기억해야 합니다.

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

b. 계절성만 존재하는 경우 : Seasonal Differencing

계절성만 존재하는 경우, lag-d differencing을 통해 계절성을 제거합니다. 식으로는 아래와 같이 표현할 수 있습니다.

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

이때 d차 차분와 lad-d 차분의 표현법을 헷갈리지 않도록 주의해야 합니다!

d차 차분은 $\nabla^d = (1-B)^d$로, lad-d 차분은 $\nabla_d = (1-B^d)$로 표현합니다.

$s_t = s_{t+d}$를 가정하고 계절성이 존재하는 시계열에 lad-d 차분을 적용해보겠습니다.

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

$s_t = s_{t+d}$이기 때문에 lad-d 차분 결과 오차항만 남아 계절성이 제거되었습니다.

c. 추세와 계절성이 모두 존재하는 경우 : lag-d 차분 + p차 차분(p=추세의 차수)

[1] 계절차분을 우선적으로 진행합니다. 그 결과 아래의 과정을 거쳐 계절성은 사라지고 추세만 남게 됩니다.

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

[2] 남아있는 추세를 제거하기 위해 차분을 진행합니다.

차분을 통해 계절성과 추세를 모두 제거할 때 주의해야 할 점이 있습니다.

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

위 식과 같이 계절차분에 $(1 - B)$이 포함되어 있기 때문에, **p차 추세를 제거하고자 한다면 (p-1)차 차분을 진행**해주어야 합니다!

4 정상성 검정

시계열 자료에서 비정상 부분을 제거하였다면 **정상성을 만족하는 오차**만 남아야 합니다. 그럼 지금부터 남아있는 오차들이 정상성을 만족하는지 확인하는 과정을 배워봅시다!

4.1 자기공분산함수(ACVF)와 자기상관함수(ACF)

오차들이 정상성을 만족하는지는 자기공분산함수와 자기상관함수로 확인할 수 있습니다. 이 두 함수는 시계열의 확률적 성질이 시간 t에 의존하는 정도, 즉 시간에 따른 상관 정도를 확인할 수 있는 함수입니다.

a. 자기공분산함수 ACVF(autocovariance function)

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

b. 표본자기공분산함수 SACVF(sample autocovariance function)

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

(원칙상으로는 n이 아닌 $n-h$로 나눠야 하지만, Non-negative Definite 성질을 만족시키기 위해 n으로 나누어 줍니다!)

c. 자기상관함수 ACF(autocorrelation function)

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

d. 표본자기상관함수 SACF(sample autocorrelation function)

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

* ACVF, ACF의 성질

1) $\gamma_x(0) = Var(X_t) \geq 0 \quad \rightarrow \quad \rho_X(0) = 1$

2) $|\gamma_x(h)| \leq \gamma_x(0)\ for$ all h $\in$ Z $\quad \rightarrow \quad |\rho_X(h)| \leq 1$

3) $\gamma_x(h) = \gamma_x(-h)$ (even function)

4) For any integer n ≥ 1 and vector a $= (a_1, \dots, a_n)' \in R^n$, $\sum_{i,j=1}^{n} a_i \gamma_x(i-j) a_j \geq 0$

(non-negative definiteness)

4.2 백색잡음 (White Noise)

자기상관이 존재하지 않는 시계열을 백색잡음이라고 합니다. 시계열 $\{X_T\}$의 평균이 0, 분산이 $\sigma^2 < \infty$이며 상관관계가 존재하지 않는다면 $\{X_t\}$는 백색잡음이라고 부르며, 아래와 같이 나타냅니다.

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

(우리가 일반적으로 사용하는 $IID(0, \sigma^2)$는 백색잡음이지만, 백색잡음은 IID가 아니라는 사실을 기억해야 합니다!)

4.3 백색잡음 검정

비정상 시계열로부터 정상적으로 추세와 계절성을 제거했다면 남아있는 오차항은 IID 혹은 WN 조건을 만족합니다. 이 경우 우리는 σ^2을 구하기 위해 $\gamma_X(0)$만 추정해주면 됩니다. 아래의 과정을 통해 백색잡음 검정 방법을 알아보겠습니다. 백색잡음 검정은 자기상관, 정규성, 정상성에 대한 검정을 진행합니다.

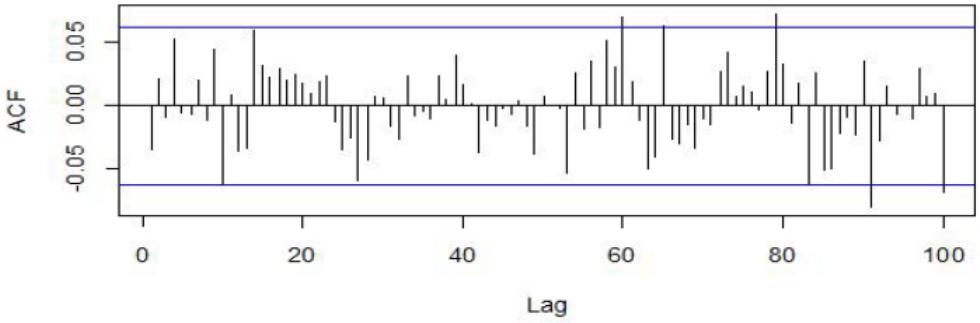
1) 자기상관 검정

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

오차가 백색잡음 WN(0,1)을 따른다면, 표본자기상관함수 $\hat{\rho}(h)$는 평균이 0이고 분산이 1/n인 정규분포로 근사합니다. 이러한 사실을 바탕으로 다음의 가설 검정을 진행합니다.

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

만약 $|\hat{\rho}(h)|$가 $1.96/\sqrt{n}$ 안에 있다면 귀무가설을 기각할 수 없습니다. 즉, 오차항에 자기상관이 존재하지 않는다고 판단합니다. 검정 결과는 ACF 그래프(=correlogram)로도 시각적으로 확인할 수 있습니다.



그래프의 x축은 시차, y축은 acf를 의미하여 파란선을 통해 신뢰구간을 확인할 수 있습니다. 위 그림에서 대부분의 경우 신뢰구간을 벗어나지 않기 때문에 오차항에 자기상관이 존재하지 않는다고 판단합니다.

이 방법 외에도 portmanteau test, Ljung-Box test, McLeod-Li test 등을 통해 자기상관 검정을 진행할 수 있습니다.

2) 정규성 검정

H_0 : 정규성이 존재한다 vs H_1 : 정규성이 존재하지 않는다.

- QQ plot : 시각적으로 확인할 수 있는 방법
- Kolmogorov-Sminorv test(KS test) : 표본의 누적확률분포가 모집단의 누적확률분포와 얼마나 유사한지 비교하는 방법
- Jarque-Bera test : 왜도와 첨도를 통해 정규성을 검정하는 방법

3) 정상성 검정

- Kpss test : 단위근 검정방법 중 하나

  ⇨ 귀무가설 : 정상 시계열이다.

- ADF test : 단위근 검정방법 중 하나

  ⇨ 귀무가설 : 정상 시계열이 아니다.

- PP test : 이분산이 있는 경우에도 사용 가능한 검정 방법

  ⇨ 귀무가설 : 정상 시계열이 아니다.

이것으로 시계열 클린업 1주차를 마치도록 하겠습니다! 익숙하지 않은 개념들과 수식이 많아 힘드셨을 텐데 잘 따라와주셔서 감사합니다^_^

1주차 내용에서는 시계열 자료의 개념과 시계열 분석의 필요성, 정상성의 개념과 정상화의 필요성 위주로 흐름을 복습해주시면 더 많은 도움이 될 것이라 생각합니다. 오늘 배운 이론들을 바탕으로 실제 분석 과정을 R 실습을 통해 확인해보도록 하겠습니다. 수고하셨습니다!