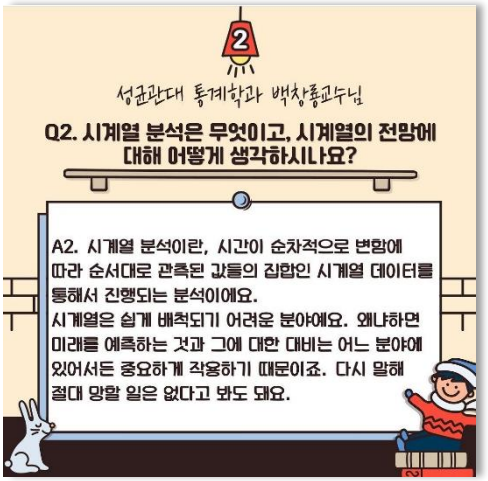
TIME SERIES ANALYSIS

시계열 공부에 필요한 마음가짐><

CONTENTS



출처: 통계학과 학생회 인스타그램

I. 시계열 알아가기

1. 시계열자료란?

시계열자료분석팀에 오신 걸 환영합니다! 시계열자료란 도대체, 무엇일까요? 시간에 따라 관측된 자료를 **시계열(time series)**자료라고 하며 시계열 분석은 시간 순서대로 정렬된 데이터에서 의미 있는 요약과 통계 정보를 추출하는 것을 의미합니다. 일반적으로 시간 t를 이용하여 다음과 같이 표현합니다.

$$\{X_t\ , t = 1,2, \ldots\}$$

여기서 시간 t가 이산형인지 연속형인지에 따라 이산형 시계열자료와 연속형 시계열자료로 구분할 수 있습니다.

2. 시계열자료분석의 목적

시계열자료분석의 목적은 두가지로 나눌 수 있습니다. 또한 시계열 자료는 시간에 따라 관측되는 관계로 대게는 서로 독립이 아니라는 특성을 갖고 있으므로 이에 맞는 분석방법이 필요합니다.

1) 예측을 위한 분석
   → 추세분석, 평활법, 분해법, 자기회귀누적이동평균(ARIMA) 모형 등등…

2) 시스템을 이해하고 제어하기 위한 분석
   → 스펙트럼 분석, 개입분석, 전이함수 모형 등등…

3. 시계열자료의 구성요소

시계열자료분석의 전통적인 방법 중 분해법은 시계열이 여러 성분 또는 요인으로 구성되어 있다고 보아 이를 분해한 후 성분들을 각각 추정하여 원래의 시계열을 해석하려는 방법입니다. 우리가 배우게 될 ARIMA 등이 이 방법을 사용하고 있으며 시계열의 특성을 이해하는데 도움이 될 수 있으니 살펴봅시다!

① **추세변동(Trend)**
   → 시간이 경과함에 따라 관측 값이 증가하거나 감소하는 추세를 갖는 경우의 변동

② **순환변동(Cycle)**
   → 주기적인 변화를 가지나 그 변화가 계절에 의한 것이 아니고, 주기가 긴 경우의 변동
   ex) 태양의 흑점 수의 변화, 남극 빙하의 변화, …

③ **계절변동(Seasonal variation)**
   → 주별·월별·계절별과 같은 주기적인 성분에 의한 변동

④ **우연변동/불규칙 성분(random fluctuation)**
   → 시간에 따른 규칙적인 움직임과 무관하게 랜덤한 원인에 의해 나타나는 변동성분

⇨ ①~③은 시간의 영향을 받는 체계적 성분이고 ④는 시간의 영향을 받지 않는 불규칙성분입니다.

II. 정상성(Stationarity)

1. 정상성은 왜 필요할까?

정상성이 무엇인지에 알아보기 전에, 정상성이라는 가정이 시계열에서 왜 필요한 것인지 먼저 이야기해 보겠습니다! 각 시점 t에서 확률변수$\{X_t\}$의 관측 값인 $\{x_t\}$의 **시계열 모델**(time series model)은 결합분포의 구체화입니다. 즉, 미래의 값을 예측하기 위해서는 무한한 시점들의 결합분포를 고려해야합니다. 그러나 이는 현실적으로 매우 복잡하기에, 몇 가지의 가정을 통하여 이를 대체하는 것이 **정상성**입니다. 대부분의 시계열 이론들이 정상성을 가정하고 전개되고 있습니다.

2. 정상성이란?

정상성이란 시계열의 확률적 성질이 **시간에 흐름에 영향을 받지 않는 것**(time-invariant)을 의미합니다. 즉, 평균, 분산 등에 변화가 없는 것을 의미합니다.

1) 강정상성(strict stationarity)

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

모든 h 와 n>0 에 대하여 시계열 $\{X_t, t \in \mathbb{Z}\}$가 위의 조건을 만족할 때 강정상성을 만족하는 시계열이라고 합니다. 즉, 모든 n 에 대하여 **결합확률밀도가 시간대(t)를 바꾸어도 동일해야 함**을 의미합니다. 그러나 여전히 이를 만족하는 것은 현실적으로 어려우며 매우 복잡합니다.

2) 약정상성(weak stationarity)

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

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

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

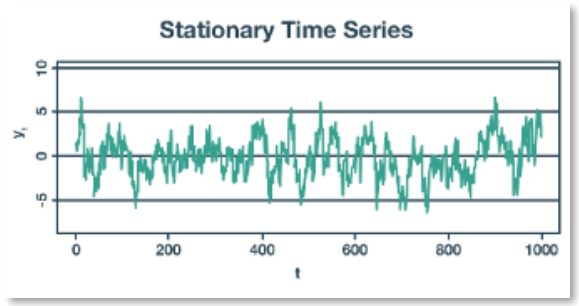
약정상성은 강정상성보다 완화된 조건으로, 시계열 $\{X_t, t \in \mathbb{Z}\}$가 위의 세 조건을 만족할 때 약한 의미의 정상성을 만족한다고 합니다. 위의 세 조건을 조금 더 자세히 살펴보겠습니다.

i. **2 차 적률**(분산과 관련)이 존재하고 시점 t 에 관계없이 일정하며

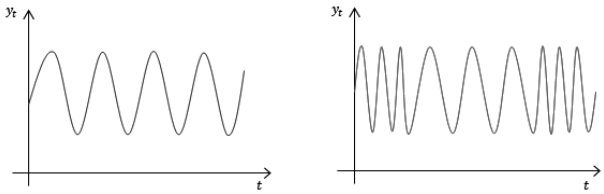
ii. **평균**이 상수로 시점 t 에 관계없이 일정하며

iii. **자기공분산**은 시차 h 에만 의존하고 시점 t 와는 무관해야 합니다.

→ 분포 전체가 동일해야 하는 강정상성과는 달리 시계열이 약정상성을 만족하는지 확인하기 위해서 $E[X_t]$(1 차 적률) 및 $\gamma(h)$만 고려하면 된다는 점에서 훨씬 간단하지 않나요? 앞으로 등장하는 "정상성"은 위의 약정상성을 의미합니다.



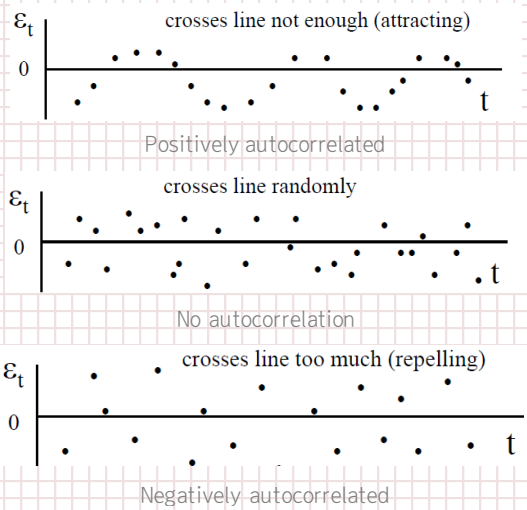
정상성을 만족하는 시계열로, x축은 시간을, y축은 값을 나타내는 시계열 그래프입니다.
평균 및 분산이 일정함을 시각적으로 확인할 수 있습니다.

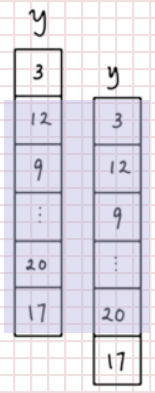


공분산이 시간에 의존하지 않는 경우(좌)와 공분산이 시간에 의존하는 경우

여기서 잠깐! 자기공분산(autocovariance)/자기상관(autocorrelation)이란?

공분산(covariance)와 상관계수(correlation coefficient)는 다들 들어 보셨죠? 그런데, 여기에 'auto'가 왜 붙은 것일까요? 먼저 공분산과 상관계수의 의미부터 떠올려봅시다. 공분산은 두 변수 사이의 관계, 그리고 상관계수는 이를 각 변수의 표준편차로 나누어주어 두 변수의 선형관계를 -1 에서 1 사이의 값으로 표현한 값입니다. 시계열에서는 두 변수가 아닌 자기 자신과 몇 시점 떨어진 자기 자신사이의 공분산 및 상관계수를 구함으로써 해당변수가 시차를 두고 어느정도 연관되어 있는지를 확인합니다. 이제 왜 "auto"혹은 "자기"라는 말이 붙었는지 아시겠죠? 자기공분산/자기상관에 대해 더 자세히 배울 예정이니 지금은 여기까지만 !!

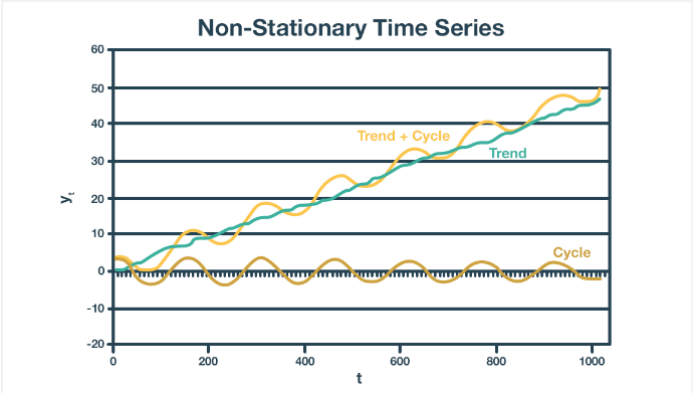




⇒ y와 한 시점 뒤 y와의 상관을 구한다! ^_^

III. 정상화

앞서 정상성 개념을 소개할 때, 여러 시계열 모델들이 정상성을 가정하고 전개된다고 하였습니다. 그러나 현실의 대부분의 시계열은 정상성을 만족하지 않습니다. 그렇다면, 정상시계열이 아닌 시계열의 경우 이에 해당하는 모델에 적합하여 예측하기 위해선 정상시계열로 변환해 줘야하겠지요? (마치 회귀분석에서 회귀의 기본 가정이 맞지 않을 때 몇가지의 변형을 통해 맞춰주는 것과 비슷하다고 생각하세요^_^)



1. 시계열 자료의 탐색

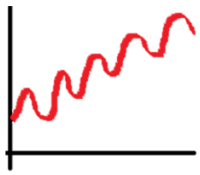
그럼 우리 먼저 주어진 시계열이 정상성을 만족하는지 만족하지 않는지 확인부터 해야 하겠지요? 시계열 플랏을 통해 확인할 수 있습니다. 시계열 플랏은 x축은 time point t, y축은 각각의 시간에 대응하는 관측값으로 구성되어 있습니다. (시계열에 따라 x축의 범위를 잘 조정하세요!) 그 다음 추세가 존재하는지, 돌발적인 변화가 있는지 혹은 이상치들이 존재하는지 등을 파악합니다.

옆의 예시 그림을 살펴봅시다! 예시의 데이터는 월별로 기록된 데이터로, x축은 데이터가 기록된 일시, 즉 해당월이고 y축은 TB rate (미국 재무부 등에서 발행하는 증권의 시중 할인율) 입니다. 위의 그림 같은 경우, 수준(평균)의 변화가 있어보이지요? 분산의 변동은 없어보이네요!!

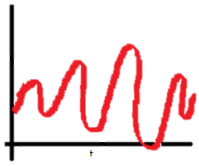
2. 비정상 시계열

본격적으로 정상화에 들어가기 앞서 정상성을 다시 떠올리며 비정상 시계열이 될 수 있는 경우는 어떤 것들이 있는지 짚고 넘어가겠습니다.

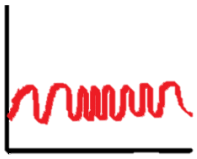
[평균이 일정하지 않은 경우]



[분산이 일정하지 않은 경우]



[공분산이 시점에 의존하는 경우]



3. 정상화 과정

1) 분산이 일정하지 않은 경우

시간의 흐름에 따라 변동폭이 커지는 자료들이 존재합니다. 이러한 경우에 **분산안정화 변환**(variance stabilizing transformation)을 통하여 시간의 흐름에 따라 분산이 변하지 않고 일정하도록 해주어야 합니다. 이러한 현상을 이분산(heteroscedasticity)이라고도 합니다.

- Box-Cox transformation

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

- Log-transformation

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

- Square root transformation

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

2) 평균이 일정하지 않은 경우

$$X_t = m_t + s_t + Y_t$$

m_t : 추세, s_t : 계절성, Y_t : 정상성을 만족하는 오차

평균이 일정하지 않은 경우, 추세가 존재하거나 계절성이 존재하거나 혹은 추세와 계절성 모두 존재하여 평균이 일정하지 않을 수 있습니다. 따라서 각각의 경우에 어떤 방식으로 추세 및 계절성을 제거하여 정상시계열로 변환할 수 있는지 알아보겠습니다.

i. 회귀

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.

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

[4] 추정한 추세를 시계열에서 제거하면 정상시계열이 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, OLS 를 통하여 a_j와 b_j를 추정합니다.

참고

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

[1] $f_1 = [n/d]$(n = data 수, d = 주기) → $f_j = j f_1$ [2] $\lambda_j = f_j(2\pi/n)$

k는 주로 1~4 사이의 값을 사용합니다.

[4] 추정한 계절성을 시계열에서 제거하면 정상시계열이 됩니다.

C – [추세 및 계절성이 동시에 존재하는 경우]

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

[A], [B]를 차례대로 해줍니다.

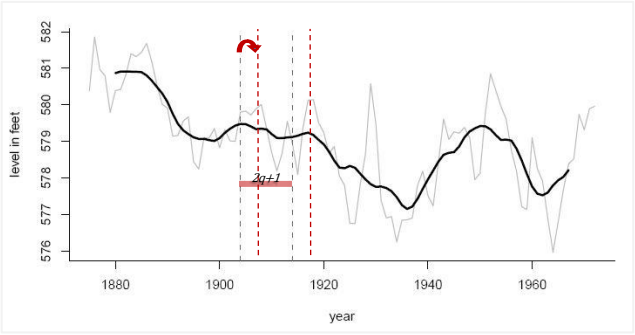
이 후에도 여전히 추세가 남아있다면 다시 추제를 제거해줍니다.

⇨ **회귀방법의 단점**: 기본적으로 최소제곱법은 오차항의 독립성을 가정하고 전개되는데, 시계열의 오차항은 독립성을 가정하지는 않아 추정이 정확하지 않을 수 있다. 특히 분산을 계산할 때 만약 오차항이 연관되어(correlated) 있음에도 불구하고 독립을 가정한 상태로 계산한 분산은 틀릴 수 있습니다. 이에 따라 신뢰 구간의 계산까지 오류가 생길 수 있다는 단점이 존재합니다.

ii. 평활

A1 – [추세만 존재하는 경우] 이동평균 평활법(Moving Average Smoothing)

이동평균 평활법은 일정기간마다 평균을 계산하여 추세를 추정하는 방식입니다.



[1] 길이가 $2q+1$인 구간의 평균을 구합니다.

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

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

[2] 위의 식에 추세 성분 m_t를 대입합니다. (추세는 linear 하다고 가정)

$$m_t = c_0 + c_1 t$$

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

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

[3] 위의 과정을 통하여 추세부분만 남은 W_t를 X_t에서 제거하여 줍니다.

⇨ 위와 같은 이동평균 평활법은 2q+1 의 구간 내에서는 모두 동일하게 $\frac{1}{2q+1}$만큼의 가중치를 가집니다. 또한 현실에서 t 시점을 예측할 때, t 시점 이후의 데이터 또한 활용할 수 있는 경우는 거의 불가능합니다. 따라서 과거의 데이터에만 의존하여 추정하는 지수평활법에 대하여 알아보겠습니다.

A2 - [추세만 존재하는 경우] 지수평활법(Exponential Smoothing)

지수평활법은 추세 $\hat{m}_t$를 시점 t까지의 관찰값을 이용하여 추정하는 방법입니다. 즉, 과거시점의 데이터를 이용하여 추세를 추정한다는 점에서 이동평균평활법과 차이가 있습니다. 과정은 다음과 같습니다.

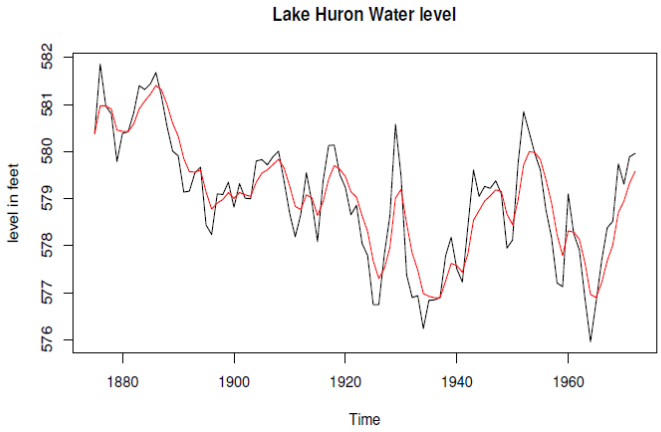
[1] 다음과 같이 추세를 추정합니다.

$$\hat{m}_1 = X_1$$
$$\hat{m}_2 = aX_2 + (1-A)\hat{m}_1 = aX_2 + (1-a)X_1$$
$$\hat{m}_3 = aX_3 + (1-a)\hat{m}_2 = \square$$
$$\vdots$$
$$\hat{m}_t = aX_t + (1-a)\hat{m}_{t-1} = \sum_{j=0}^{t-2} a(1-a)^j X_{t-j} + (1-a)^{t-1}X_1$$

[2] 추정한 추세를 시계열에서 제거합니다.

⇨ $a \in [0,1]$ 이며, 과거 관측치에 대한 가중치입니다.

⇨ 더 과거의 값일수록 가중치의 값이 지수적으로 줄어드는 것을 알 수 있습니다.

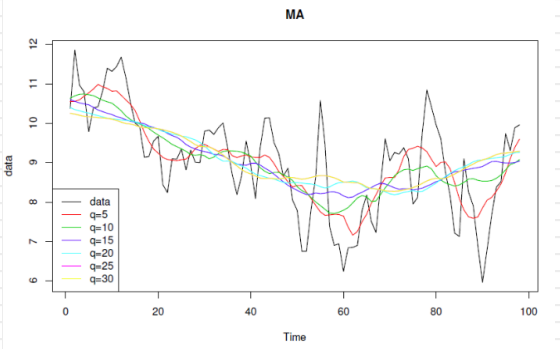


a = 0.4일 때 지수평활법으로 추정된 추세

여기서 잠깐!

평활법에서 q와 a의 선택

평활법은 tuning parameter를 선택해줘야 한다는 단점이 있습니다. 선택해줘야 하는 a와 q는 각각 이동평균평활법과 지수평활법의 결과에 중요한 영향을 미치기에, 이를 신중히 선택해주어야 합니다. 가장 일반적인 방법은 cross-validation (CV)를 통하여 MSE를 추정하여 최적의 파라미터를 선택하는 방법입니다.



MA(moving average)의 경우를 통해 parameter q의 크기가 어떤 영향을 미치는지 구체적으로 알아보겠습니다.

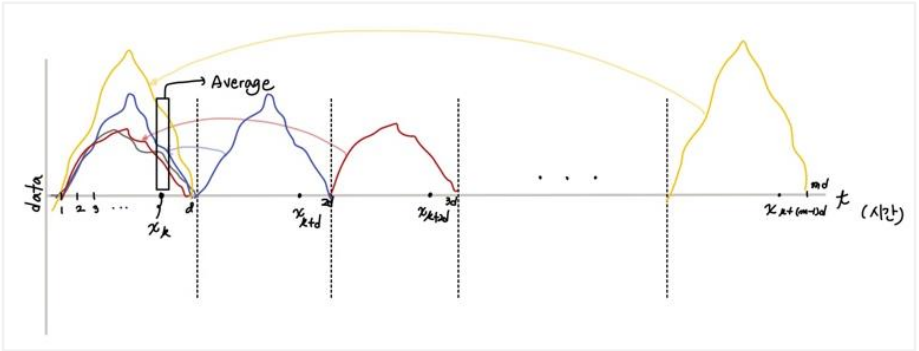
$$W_t = \underbrace{\frac{1}{2q+1}\sum_{j=-q}^{q} m_{t+j}}_{\approx m_t \ if\ q\ is\ small} + \underbrace{\frac{1}{2q+1}\sum_{j=-q}^{q} Y_{t+j}}_{\approx 0\ if\ q\ is\ large}$$

위의 식에서 q가 작은 경우에 bias는 감소하지만 variance는 증가하는 반면, q가 큰 경우 bias는 증가하지만 variance는 감소함을 알 수 있습니다. 위와 같은 bias와 variance의 관계를 *bias-variance trade off* 라고 하며 이러한 편향과 분산의 관계가 존재하기에 적절한 parameter를 찾는 것이 중요합니다.

CV와 bias -variance trade off에 대해 더 자세히 알고싶다면 데마 1주차 참고><

B - [계절성만 존재하는 경우] Seasonal Smoothing

IDEA: 주기가 d인 관측치를 한 주기 안에 모두 겹친 후 평균내자!

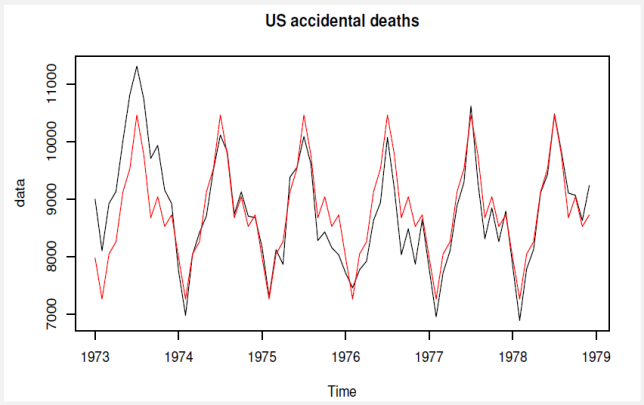


한땀한땀,,,,그렸습니다,,,~~(아무도 그럴 생각없겠지만 무료배포 허용)~~

[1] k = 1, ⋯ , d 에 대하여 계절성분을 다음과 추정합니다.

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

$$m = \#\ of\ obs.\ in\ kth\ seasonal\ component$$



⇨ 첫번째 주기의 패턴이 반복되는 것을 알 수 있습니다.

C – [추세 및 계절성이 동시에 존재하는 경우] classical decomposition algorithm

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

[1] 먼저 MA filter 를 이용하여 추세를 예측합니다. 이때, $\Sigma_{k=1}^{d} s_j = 0$

Why? 위 조건이 있기에 계절성이 존재함에도 추세를 이동평균법으로만 예측해도 충분하기 때문입니다. 예시를 보면 무슨 말인지 알 수 있을꺼에요 !!

⇨ ex) d = 3

$$\frac{X_{t-1}+X_t+X_{t+1}}{3} = \frac{m_{t-1}+m_t+m_{t+1}}{3} + \frac{s_{t-1}+s_t+s_{t+1}}{3} + err = \frac{m_{t-1}+m_t+m_{t+1}}{3} + err$$

[2] 위에서 추정한 추세를 제거해준 후 계절성분을 seasonal smoothing으로 추정합니다.

$$z_t = X_t - \widehat{m_t} \approx s_t + Y_t$$

$$\hat{s}_t^* = \frac{1}{m}\Sigma_{k=0}^{m-1} z_{t+kd} \ , k = 1, \dots, d$$

$$\hat{s}_t = \hat{s}_t^* - \widehat{\bar{s}^*}$$

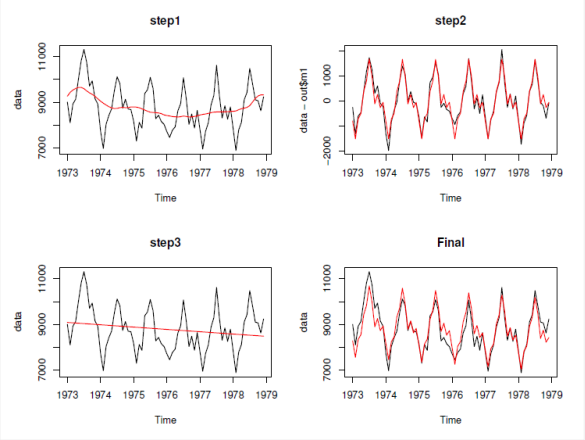
[3] 계절성이 제거된 시계열에서 다시 추세를 OLS를 통해 추정합니다.

$$\widehat{m}_t^{new} = \underset{c}{\operatorname{argmin}} \Sigma_{t=2}^{n} \left(X_t - \hat{s}_t - c_0 - c_1 t - c_2 t^2 - \cdots - c_p t^p\right)^2$$

[4] 새롭게 추정한 추세와 계절성을 다음과 같이 제거해주어 오차를 추정합니다.

$$\hat{e}_t = X_t - \widehat{m}_t - \hat{s}_t$$

⇨ 위의 과정을 거치고 난 후에 또 다시 추세가 있다면 [1]~[4] 반복!



iii. 차분(Differencing)

차분에서는 새로운 연산자가 등장합니다! 바로 한 시점 전으로 돌려주는 작용을 하는 **후향연산자(Backshift Operator)**입니다.

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

차분은 뭘까요? 차분은 관측값들의 차이를 구하는 것입니다! 후향연산자를 사용하여 다음과 같이 표현할 수 있습니다.

[1 차 차분]

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

[2 차 차분]

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

A – [추세만 존재하는 경우] Differencing

그렇다면 차분이 추세를 어떻게 제거할 수 있다는 걸까요?

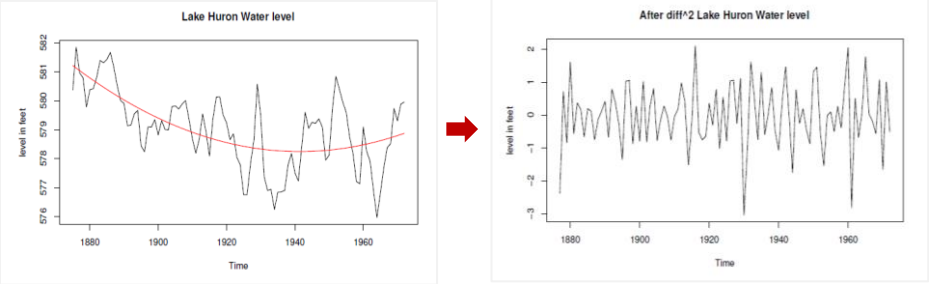
추세를 $m_t = (c_0 + c_1 t)$ 라고 가정해볼께요 !!

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

짜잔~~~~~~ 시간에 영향을 받지 않는 상수만 남았습니다!

일반적으로 k 차 차분을 하면, k 차 추세(k – th order polynomial trend)를 제거할 수 있습니다.

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



B - [계절성만 존재하는 경우] 계절차분(Seasonal differencing)

계절성이 존재하는 경우에는 lag-d differencing 을 통해 계절성을 제거합니다.

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

주의

$$\nabla^d = (1 - B)^d \ (d - th\ order\ differencing)$$

만약 $s_t = s_{t+d}$ 라고 가정한다면,

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

⇨ 오차항만 남아 계절성을 제거할 수 있습니다.

C - [추세 및 계절성이 동시에 존재하는 경우]

추세 및 계절성이 동시에 존재할 때, **계절차분 + p 차 차분**을(이때, p 는 추세의 차수) 하면됩니다.

[1] 그럼! 먼저 계절 차분을 해보자구요! 그럼 계절성이 사라지고 다음과 같이 추세만 남겠지요?

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

[2] 이제 남아있던 추세($m_t - m_{t-d}$)를 제거해주기 위해 차분을 합니다.

여기서 잠깐~~ 아래를 잠깐 볼까요?

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

다음과 같이 표현이 가능하죠! 계절차분이 1 차 차분을 포함하고 있음을 의미합니다.

그래서 추세가 $m_t = c_0 + c_1 t + \cdots + c_p t^p$ 인 경우엔 $\nabla^{p-1}\nabla_d X_t$로 차분을 해준다는 점!

과대차분(overdifferencing)을 조심하세요

과대차분이란 말 그대로 차분을 과하게 하는 것입니다. 이미 정상화가 되어있음에도 불구하고 차분을 또 시도하는 경우, 정상시계열의 선형 결합은 다시 정상 시계열이 되어 "정상성"자체에는 문제가 없습니다. 그러나, 지나친 차분은 ACF를 복잡하게 만들거나 분산이 커지는 문제가 생깁니다.

⇨ 이 성질은 모형 선택시에도 유용하게 사용됩니다. (분산의 크기 비교를 통해 적절한 모형 선택)

여기까지 오시느라 다들 수고가 많으셨지만 조금만 더 힘을 내보자구요!!☞☞

IV. 정상성 검정

만약 성공적으로 추세와 계절성을 제거했다면, 정상성을 만족하는 오차만 남아야 합니다. 과연 지금 남은 오차가 정상성을 만족하는지를 확인해야겠지요??

1. 자기공분산함수, 자기상관함수

정상성을 만족하지 않는 다는 것은 곧 확률적 성질이 시간에 의존하게 되는 것입니다. 정상성을 확인하기 위하여 평균, 분산 뿐만 아니라 시간에 따른 상관 정도를 나타내기 위한 자기상관함수(autocovariance function) 또는 자기상관계수(autocorrelation function)를 확인해야 합니다.

- 자기공분산함수 ACVF (autocovariance function)

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

- 표본자기공분산함수 SACVF (sample autocovariance function)

$$\hat{\gamma}_k = \frac{1}{T}\Sigma_{j=1}^{T-k}(X_j - \bar{X})(X_{j+k} - \bar{X})$$

- 자기상관함수 ACF (autocorrelation function)

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

- 표본자기상관함수 SACF (sample autocorrelation function)

$$\widehat{\rho_k} = \frac{\widehat{\gamma_k}}{\hat{\gamma}_0}, \hat{\rho}(0) = 1$$

2. 백색잡음(White Noise)

자가상관이 없는 시계열을 백색잡음이라고 부릅니다. 백색잡음은 대표적인 정상 확률과정으로 $\{X_t\}$는 상관관계가 존재하지 않고(uncorrelated) 평균이 0, 분산이 $\sigma^2 < \infty$이면 백색잡음이라고 부릅니다. 다음과 같이 표기할 수 있습니다.

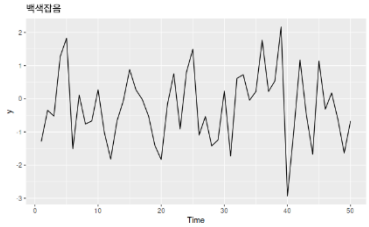
$$\{X_t\} \sim WN(0, \sigma^2)$$

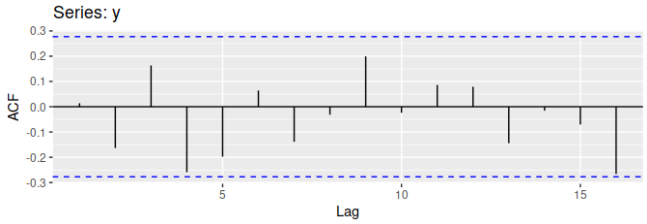
$IID(0, \sigma^2)$는 백색잡음이라고 할 수 있지만 그 역은 true 가 아니라는 사실 주의하세요!! (+IID 와 WN 의 acf 는 같습니다)

Cf) IID Noise

$$\gamma_X(t+h, t) = \begin{cases} \sigma^2, if\ h = 0 \\ 0\ .if\ h \neq 0 \end{cases}$$

$$\{X_t\} \sim IID(0, \sigma^2) \Rightarrow X_t\ are\ independent\ and\ identically\ distributed$$





3. 백색잡음 검정

만약 추세/계정성을 시계열에서 제거한 뒤 남은 오차항이 WN 이거나 IID 라면, σ^2 은 $\gamma(0)$만 추정해주면 됩니다. 그러나 WN/IID 가 아니라면 여전히 존재하는 의존성을 설명해주기 위하여 더 복잡한 시계열 모델을 필요로 합니다.

1) $\hat{\rho}(h)$ correlated vs uncorrelated

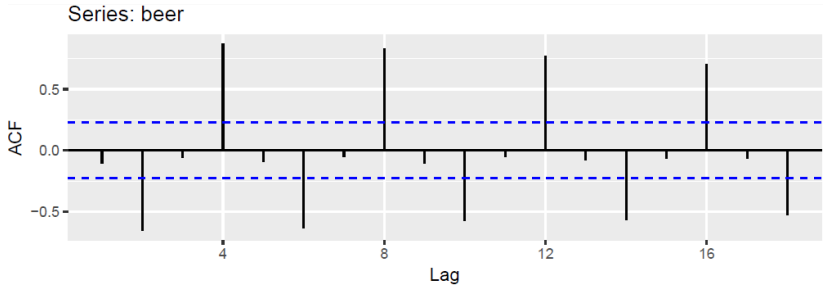
만약, 오차 Y_t 가 백색잡음이라면, $\hat{\rho} \approx \mathcal{N}\left(0, \frac{1}{n}\right)$ (n↑ ⇨ 평균이 0 이고 분산이 1/n 인 정규분포를 따릅니다.)

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 내 있다면, 오차들이 uncorrelated 되어 있다고 할 수 있습니다.

이를 ACF 그래프를 통해 시각적으로도 확인해 볼 수 있습니다.



위 그래프는 ACF 그래프로, x 축의 lag 는 시차, y 축은 acf 를 나타냅니다.
각각의 값은 시차(lag)가 h 일 때의 ACF 값이며, 파란 점선은 신뢰구간을 의미하며
이 두 선 안에 ACF 값이 존재하면 autocorrelation 이 존재하지 않는다고 판단합니다.

- 이 외에도 portmanteau test, Ljung-Box test, McLeod-Li test 등을 이용하여 확인할 수 있습니다.
- R 에서 itsmr 패키지의 test()함수를 사용하여 확인 가능합니다.

```
## Null hypothesis: Residuals are iid noise.
## Test                      Distribution Statistic   p-value
## Ljung-Box Q              Q ~ chisq(20)    107.83         0 *
## McLeod-Li Q              Q ~ chisq(20)     68.71         0 *
## Turning points T     (T-64)/4.1 ~ N(0,1)        40         0 *
## Diff signs S       (S-48.5)/2.9 ~ N(0,1)        50    0.6015
## Rank P         (P-2376.5)/162.9 ~ N(0,1)      2344    0.8419
```

2) 정규성 확인

- H0: 정규성이 존재한다
- QQplot : 시각적으로 정규성을 만족하는지 확인하는 방법
- Kolmogorov-Sminorv test
- Jarque-Bera test

  왜도(Skewness)와 첨도(Kurtosis)가 정규분포(normal distribution)로 보기에 적합한지에 대한 적합도(Goodness-of-fit) 검정하는 방법

3) 정상성 검정

- kpss test

  H0: 정상성을 가진다
- adf test, pp test

  H0: 비정상성을 가진다

V. 모형의 필요성

다음주부터는 본격적으로 시계열의 예측모델에 대해 배우는데요, 배우기 전에 이번주에 그러한 모형들이 왜 필요한지 알아보고 가겠습니다.

오차항 Y_t의 공분산 행렬은 다음과 같습니다.

$$\Gamma = \begin{pmatrix} \mathrm{Cov}(Y_1, Y_1) & \mathrm{Cov}(Y_1, Y_2) & \dots & \mathrm{Cov}(Y_1, Y_n) \\ \mathrm{Cov}(Y_2, Y_1) & \mathrm{Cov}(Y_2, Y_2) & \dots & \mathrm{Cov}(Y_2, Y_n) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}(Y_n, Y_1) & \mathrm{Cov}(Y_n, Y_2) & \dots & \mathrm{Cov}(Y_n, Y_n) \end{pmatrix}$$

$$= \begin{pmatrix} \gamma(0) & \gamma(1) & \dots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \dots & \gamma(n-2) \\ \vdots & \vdots & \vdots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \dots & \gamma(0) \end{pmatrix}$$

만약, 백색잡음 과정이라면 대각선의 $\gamma(0)$(분산)을 제외한 나머지가 모두 0 이 되어 분산만 추정해주면 되지만, 백색잡음 과정이 아니라면 그 이외의 요소를 모두 추정해주어야 합니다. 이때, 특정 모형을 통해 추정이 가능합니다.

[부록] 시계열자료에 일반 선형회귀를 사용하지 않는 이유

부록으로 써 뒀지만!! 꼭 읽어보세요!! 중요하지 않아서 부록으로 놔둔 것이 아니라는 점!!

- 선형회귀: IID데이터가 있다는 것을 가정하지만 시계열에는 해당되지 않습니다.

  → 시계열 데이터는 시간에 가까운 데이터일수록 서로 강한 관계를 맺는 경향이 존재

  → 다음과 같은 조건이 충족될 때 일반적인 **최소제곱선형회귀(least square linear regression)** 모델을 시계열 데이터에 적용 가능

  1) **시계열의 행동에 대한 가정**

     - 시계열은 예측 변수에 대한 선형적 반응을 보입니다.

     - 입력 변수는 시간에 따라 일정하지 않거나 다른 입력 변수와 완벽한 상관관계를 갖지 않습니다.

     (데이터의 시간 차원을 설명하기 위해 전통적인 선형회귀의 독립변수에 대한 요구 사항을 확장)

  2) **오차에 대한 가정**

     - 각 시점의 데이터에 대해 모든 시기의 모든 설명변수에 대한 오차 값은 0

     - 특정 시기의 오차는 과거나 미래의 모든 시기에 대한 입력과 관련이 없습니다. 따라서 오차에 대한 자기상관 함수 그래프는 어떠한 패턴도 띄지 않습니다.

     - 오차의 분산은 시간으로부터 독립적입니다.

     - 이러한 가정이 성립된다면 OLS는 주어진 입력에 대한 계수의 unbiased estimator가 되며 이는 시계열에서도 마찬가지입니다.

  *** 필수가정을 충족하지 못한 데이터에 선형 회귀를 적용하였을 때**

  1) 계수가 모델의 오차를 최소화하지 않습니다

  2) p-value가 부정확합니다. (계수가 0인지 아닌지에 대한)

R 실습하기!

오늘 배운 내용들을 직접 해보지 않으면 무슨 재미가 있겠어요~~~

어렵지 않게 배운 내용을 구현해보는 것에 의의를 두었으니 기술적인 측면보다 시계열 자료에 익숙해지는 시간이 되길 바라며,,,🌰🌾🍁🍂 가을은 결실의 계절이죠... 배움의 결실을 만들어보자는 의미에서...

0. 준비하기

   다음의 코드를 실행하여 필요한 패키지와 데이터를 불러옵니다.

```
library(itsmr)
library(MASS)
library(tseries)
library(nortest)
data = AirPassengers
#R에 내장되어 있는 시계열데이터로 1949년부터 1960년까지의 월별 국제선 승객의 총합입니다.
```

1. 데이터 살펴보기

   시계열 플랏을 이용해 데이터를 살펴보겠습니다.

   `ts.plot()`함수를 이용해서 플랏을 그려보고, 데이터가 어떤 특징이 있는지 생각해봅시다!

   1) 분산이 일정한가요? 2) 평균이 일정한가요? 3) 계절성이 존재하나요?

2. 분산안정화

   분산을 안정화 시키기 위해 로그 변환, 제곱근 변환, Box-Cox 변환을 각각 시도합니다.

   Hint : `log(data)`, `data^(1/2)`,

   `lambda<-BoxCox.lambda(data)`

   `BoxCox(data, lambda)`

   변환한 결과를 `ts.plot`함수를 이용해 확인해봅니다.

   위의 세 변환 중 로그 변환을 한 데이터를 log_data로 저장합니다.

3. 추세 제거

   분산을 안정화 한 데이터를 이용하여 추세를 추정합니다.

   1) 회귀

      회귀를 이용해 보겠습니다! 시간이 x변수인 회귀식을 설정하면 되겠지요?

      ex) `n=length(log_data)`

      `x=seq(1,n,1)` #1부터 n까지 1씩 증가하는 수열

      `fit = lm(log_data ~ 1+ x)` # $c_0 + c_1 t$인 회귀식 모델링

```
xa = as.vector(time(log_data))
lines(xa, out.lm$fitted.values,col='red')
```

#시각화를 통해 확인

`dtrend <- log_data- fit$fitted.values` #추세를 제거한 식을 만들었습니다.

2) 평활

`smooth.ma()`함수를 이용합니다. q를 조절해서 다양한 시도를 해보세요!

Q) q가 증가함에 따라 어떤 변화가 있나요?

3) 차분

분산안정화한 데이터를 이용하여 차분을 시도해 보세요!

Hint: 1차 차분: `diff()` , 2차 차분: `diff(diff())` , …

4. 계절성 제거

1) 평활

추세를 제거한 데이터를 이용하여 계절성분을 제거합니다.

Hint: `season(data, d= )`

2) 차분

Hint: `diff(diff(log_data, lag = 12))`

5. 계절차분 & 1차 차분을 동시에 시도해 보세요!
6. 잔차검정

`acf()`

7. 자기상관검정(독립성검정)

→ `Box.test( , type = “Ljung-Box”)`

8. 정규성검정

→ `qqnorm() ; jarque.bera.test()`

9. 정상성검정

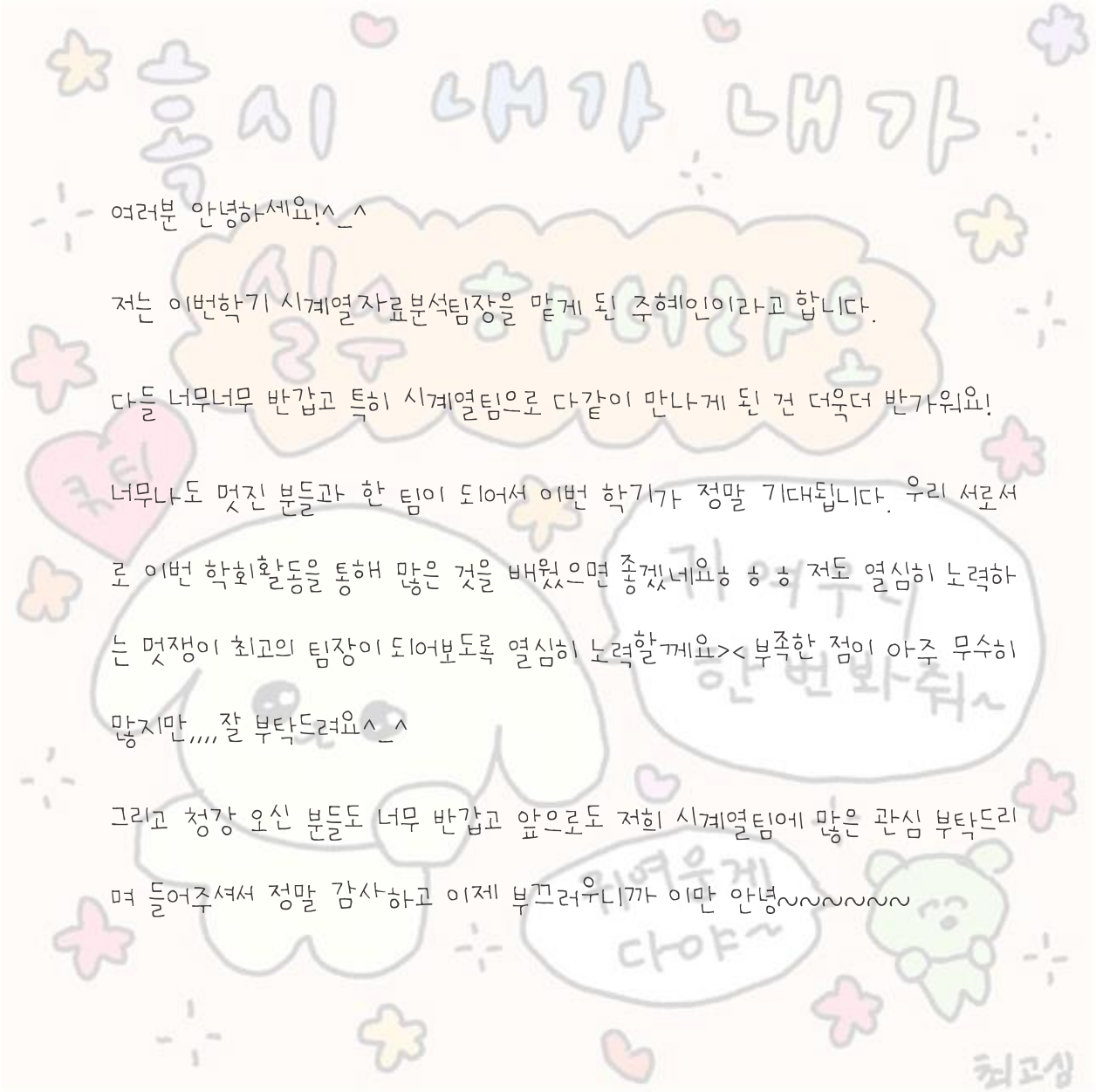
→ `adf.test(), kpss.test()`

10. Classical decomposition

분해법에 대해 배운거 기억나시나요? `decomposition()`이라는 함수를 사용하면 시계열을 여러 성분으로 분해한 플랏을 보여줍니다. 다음과 같이 해보고 눈으로 직접 확인해 보세요!

```
decom = decompose(data)
plot(decom)
```

패키지 과제도 해야 하고 여러모로 바쁘실 테니 따로 제출하거나 하진 않을게요!! 모르는 부분 생기면 바로 물어 봐주세요! 정답은 수요일에 공개합니다><



여러분 안녕하세요!^_^

저는 이번학기 시계열자료분석팀장을 맡게 된 주혜인이라고 합니다.

다들 너무너무 반갑고 특히 시계열팀으로 다같이 만나게 된 건 더욱더 반가워요!

너무나도 멋진 분들과 한 팀이 되어서 이번 학기가 정말 기대됩니다. 우리 서로서로 이번 학회활동을 통해 많은 것을 배웠으면 좋겠네요ㅎㅎㅎ 저도 열심히 노력하는 멋쟁이 최고의 팀장이 되어보도록 열심히 노력할께요><부족한 점이 아주 무수히 많지만,,,, 잘 부탁드려요^_^

그리고 청강 오신 분들도 너무 반갑고 앞으로도 저희 시계열팀에 많은 관심 부탁드리며 들어주셔서 정말 감사하고 이제 뷰끄러우니까 이만 안녕~~~~~~