7/14 시계열 1주차

1. **시계열 분석이란?**:

   추세나 계절성 등을 요약하고 시간에 따른 패턴을 분석을 분석하는 방법론. 전통적 회귀 모형과는 달리, 예측 변수를 도입하지 않고 그 자체의 과거 패턴이 미래에도 계속된다는 가정 하에 미래값을 예측

- 이동평균: 매 시점에서 직전의 N개의 데이터만의 평균을 산출하여 평활치로 사용
  - 단순이동평균: 시계열 데이터가 수평적 패턴인 경우(상수함수 형태)
  - 이중이동평균: 시계열 데이터가 추세 패턴을 따를 경우(절편과 기울기가 있는 일차함수)
- 단순이동평균법: 시계열 데이터가 수평적 패턴을 띠는 경우 사용.

시점 t에서의 단순 이동평균: $M_t = \frac{1}{N}(X_{t-N+1} + \cdots + X_t)$

- 시점 t+1에서의 이동평균: $M_{t+1} = \frac{1}{N}(X_{t-N+2} + \cdots + X_{t+1})$

$= M_t + \frac{X_{t+1} - X_{t-N+1}}{N}$

-가장 최신 시점의 데이터가 추가되면, 가장 오래된 데이터를 제거하고 평균을 구하는 것

-N이 작을수록 평활효과가 작음(최근의 추세를 많이 반영함
-그러나, 관측된 시점에 관계없이 동일한 가중치 1/N을 부여함

- 이중이동평균법: 시계열이 $X_t = c + bt + a_t$와 같은 선형추세를 가질 때 사용하는 방법
  단순이동평균은 추세를 늦게 따라가므로, 이를 보정하고자 이중이동평균을 활용

  (증명)

$$
\begin{aligned}
M_t &= \frac{1}{N}\cdot\sum_{i=1}^{N} X_{t-N+i} \\
&= \frac{1}{N}\left(X_{t-N+1} + X_{t-N+2} + \cdots + X_t\right) \\
&= \frac{1}{N}\left((+b(t-N+1) + c + b(t-N+2) + \cdots + b(t-N+N)\right) \\
&= \frac{1}{N}\left(N(+Nbt - Nb\cdot N + \frac{N(N+1)}{2}\cdot b\right) \\
&= c + bt - Nb + \frac{b}{2}(N+1) \\
&= (+bt - \frac{N-1}{2}b
\end{aligned}
$$

$$
\begin{aligned}
M_t^{(2)} &= \frac{1}{N}\left(M_{t-N+1} + \cdots + M_t\right) \\
&= \frac{1}{N}\left((+b(t-N+1) - \frac{N-1}{2}b + \cdots + c + b(t-N+N) - \frac{N-1}{2}b\right) \\
&= \frac{1}{N}\left(N(+Nbt - N^2b + \frac{N(N+1)}{2}b - \frac{N(N-1)}{2}b\right) \\
&= \cdots = (+bt + (N-1)b
\end{aligned}
$$

$$
\therefore E(M_t) - E(M_t^{(2)}) = \frac{N-1}{2}b
$$

$$
\Rightarrow b = \frac{2}{N-1}E(M_t) - E(M_t^{(2)})
$$

이를 통해, b의 추정값과 c의 추정값을 구할 수 있음!

- 예측

예측

- 시점 T에서 다음 시점의 예측치 (한단계 이후 예측)

T+1 시점

$$f_{T,1} = E[X_{T+1}|X_T, X_{T-1}, \ldots] = c + b(T+1)$$

$$\hat{f}_{T,1} = \hat{c} + \hat{b}(T+1) = 2M_T - M_T^{(2)} + \hat{b}$$

2. 단순이동 − 이중 이동 + b

$$\hat{b} = \frac{2}{N-1}(M_T - M_T^{(2)})$$

- 성능 지표: e(t,1) = X(t+1) - f(t,1) → t+1 시점에서의 실제값 - t 시점에 예측한 t+1 시점 값

2. **지수평활법**

- 평균이동법의 한계: 전체 데이터에 동일한 가중치를 부여함. 따라서 최근의 데이터에 더 큰 가중치를 부여하는 방법이 필요함
- 최근 자료에 더 큰 가중치를 주고, 과거로 갈수록 지수적으로 감소하는 가중치 사용
  → 단순지수평활, 이중지수평활, 홀트 모형

- 단순지수평활법: 시계열 데이터가 수평적 패턴인 경우 사용

- 시점 t에서의 지수평활치:

$$S_t = \alpha X_t + \alpha(1-\alpha)X_{t-1} + \alpha(1-\alpha)^2 X_{t-2} + \cdots$$

$\alpha : weight(0 < \alpha < 1)$ → 가까운 시점의 데이터에 더 큰 가중치 부여

→ 평활상수가 작을수록 평활효과가 큼(매끄러워짐)

* 이때, 어떤 가중치 값 α를 주더라도, 가중치의 합은 근사적으로 1이 될 것임(등비급수의 합)

- 이중지수평활법: 시계열 데이터가 추세 패턴을 따르는 경우 사용

$$X_t = c + bt + a_t$$

- 선형 추세의 기댓값은 c+bt, 단순지수평활법의 기댓값은 c+bt+{(1-α)/α}*b로써, {(1-α)/α}*b만큼의 차이가 생김(a_t는 ~iid(0, σ^2)로 생각) → 즉 biased임 → 이를 보정하기 위해서 이중지수평활 사용

[기댓값 증명 참고]

선형 추세모형을 따를 때의 단순지수 평활예측량

- $Z_t = \beta_0 + \beta_1 t + \varepsilon_t$
- (review) $S_n^{(1)} = w \sum_j^{n-1} (1-w)^j Z_{n-j}$

$$\Rightarrow E[S_n^{(1)}] = w \sum_j (1-w)^j E[Z_{n-j}] \quad (Z_t = \beta_0 + \beta_1 t + \varepsilon_t)$$

$$\fallingdotseq w \sum_j^{\infty} (1-w)^j [\beta_0 + \beta_1 (n-j)]$$

$$\fallingdotseq w \sum_j^{\infty} (1-w)^j [\beta_0 + \beta_1 (n-j+1-1)]$$

$$\beta_0 + \beta_1 (n+1) - \beta_1 (j+1)$$

$$= \underbrace{w \sum_j^{\infty} (1-w)^j [\beta_0 + \beta_1 (n+1)]}_{①} - \underbrace{w \beta_1 \sum_j^{\infty} (1-w)^j (j+1)}_{②}$$

$$\frac{1}{1-(1-w)} = \frac{1}{w}$$

① $w\sum_{j=0}^{\infty}(1-w)^j[\beta_0+\beta_1(n+1)]$

$= w\cdot\frac{1}{w}[\beta_0+\beta_1(n+1)]$

$= \beta_0+\beta_1(n+1)$

② $A \overset{let}{=} \sum_{j=0}^{\infty}(1-w)^j(j+1)$

$= 1 + 2(1-w) + 3(1-w)^2 + \cdots$

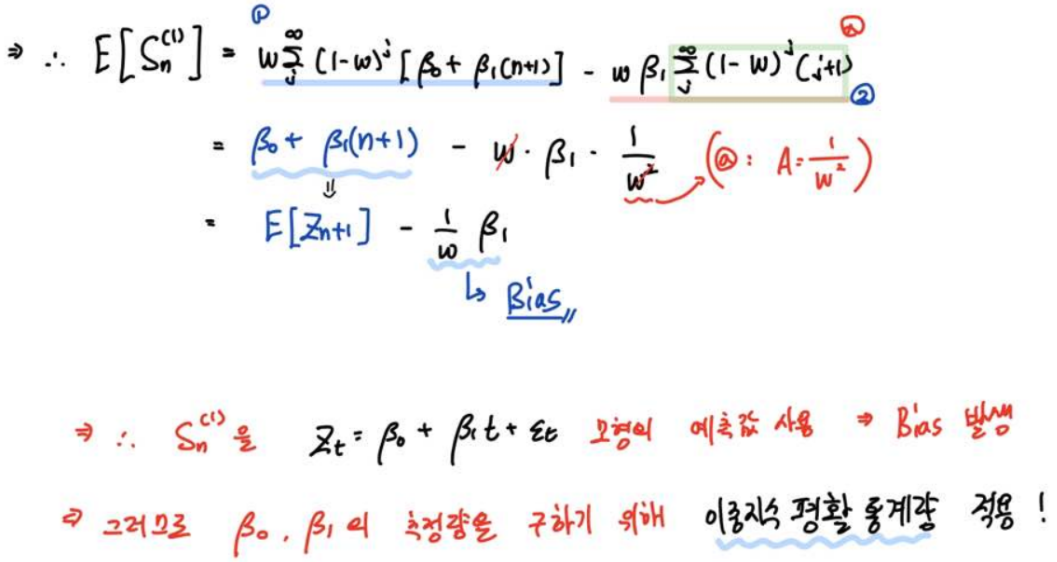
$-\quad (1-w)A = \qquad (1-w) + 2(1-w)^2 + \cdots$

$wA = 1 + (1-w) + (1-w)^2 + \cdots$

$r = (1-w)$

$= \sum_{j=0}^{\infty}(1-w)^j = \frac{1}{1-(1-w)} = \frac{1}{w}$

$\Rightarrow \quad wA = \frac{1}{w} \rightarrow \therefore A = \frac{1}{w^2}$



따라서, 이중지수평활통계량으로 b와 c를 추정하고자 함

$$S_t = \alpha X_t + (1-\alpha) S_{t-1}$$
$$S_t^{(2)} = \alpha S_t + (1-\alpha) S_{t-1}^{(2)}$$
$$E[S_t^{(2)}] = E[S_t] - \frac{1-\alpha}{\alpha} b$$
$$E[S_t] - E[S_t^{(2)}] = \frac{1-\alpha}{\alpha} b$$

$$\hat{b} = \frac{\alpha}{1-\alpha}(S_T - S_T^{(2)}) \qquad \hat{C} = 2 \cdot S_T - S_T^{(2)} - n \cdot \hat{b}$$

이를 통해, T시점에서 다음 시점인 T+1 시점에서의 예측값을 얻을 수 있음

$$\hat{f}_{T,1} = \hat{c} + \hat{b}(T+1) = 2S_T - S_T^{(2)} + \hat{b}$$

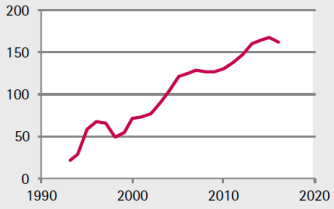
(예시 계산)

이중 지수평활법

$\checkmark\ 39.07 = 0.2 \cdot 52.38 + 0.8 \cdot 35.75$

$\therefore\ \hat{b} = \frac{0.2}{0.8}(52.38 - 39.07) = 3.33$

- 예 (특허건수) 아래는 우리나라 연도별 (1993-2016) 특허건수 (천건)를 나타낸 것이다.
- 이중지수평활을 적용하여 시간에 따라 한단계이후를 예측해 보자. (α=0.2 사용)



년도	건수	s_t	$s_t^{(2)}$	$\hat{b}$	예측치
1999	56.0	47.27	35.75	2.88	
2000	72.8	52.38	39.07	3.33	61.68
2001	73.7	56.64	42.59	3.51	69.01
2002	76.6	60.64	46.20	3.61	74.21
2015	167.3	144.00	120.22	5.95	168.07
2016	163.4	147.88	125.75	5.53	173.74

#2

[이중 지수평활]

$\checkmark\ 52.38 = 0.2 \cdot 72.8 + 0.8 \cdot 47.27$

POSTECH

- 홀트 모형: 추세 패턴이 있는 시계열 자료에 사용
  - 수평 수준과 추세를 각각 갱신함

$$\text{Forecast equation} \qquad \hat{y}_{t+h|t} = \ell_t + hb_t$$

$$\text{Level equation} \qquad \ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})$$
$$\text{Trend equation} \qquad b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1},$$

→ l_t는 시간 t에서의 시계열의 수평 수준 추정값, b_t는 시간 t에서의 시계열의 추세 (기울기)추정값, α는 수준에 대한 매개변수, β^*는 추세에 대한 매개변수
→ 위 식을 해석해보면, 수평 수준 식은 l_t가 관측 y_t에 대한 가중평균이라는 것, $l_{t-1} + b_{t-1}$로 주어지는 시간 t에 대한 '한 단계 이전 학습의 예측'을 의미
→ 추세 식은 b_t가 추세의 이전 추정값인 $l_t - l_{t-1}, b_{t-1}$에 기초한, 시간 t에서의 추정된 추세의 이동 평균이라는 것을 나타냄

$$f_{T,k} = L_T + kb_T$$

k-단계 앞을 예측하는 예측값

위 예측값은 마지막으로 추정된 수평 수준에 마지막으로 추정된 추세의 값에 k배를 한 것을 더한 것임 → 예측값은 k에 대한 선형 함수임

3. **계절성 고려 모형 - 윈터스 모형 & 분해법**
   - 윈터스 모형: 홀트 모형 + 계절성 반영
   - 분해법: 추세와 계절성을 분해한 후 다시 결합(추세를 제외하고 구한 계절성 & 계절성을 제외하고 구한 추세)

- 윈터스 모형
  - 수평 수준, 추세, 계절성을 각각 갱신하는 모형

$$\text{수평수준: } L_t = \alpha \frac{X_t}{s_{t-m}} + (1-\alpha)(L_{t-1} + b_{t-1}), \quad (0\langle\alpha\langle 1)$$
$$\text{추세: } b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}, \quad (0\langle\beta\langle 1)$$
$$\text{계절성: } s_t = \gamma \frac{X_t}{L_t} + (1-\gamma)s_{t-m}, \quad (0\langle\gamma\langle 1)$$

  - 각각을 해석하면,
    수평 수준: α*(계절성 제거한 시계열) + (1-α)*(이전 수평 수준 + 이전 추세)
    추세: β*(수평 수준 변화량) + (1-β)*(이전 추세)
    계절성: γ*(수평 수준 제거한 시계열) + (1-γ) * (이전 주기의 계절성)
  - 시점 T에서 T+k의 값 예측

$$f_{T,k} = (L_T + kb_T))s_{T-m+k}$$

(수평수준 + 추세 k단계) * (계절성)

- 분해법
  - 추세 성분: “지금까지 ~비율로 꾸준히 증가했으니, 앞으로도 ~비율로 증가하겠지”
  - 계절 성분: “A이 주기만큼 증감이 반복되었으니, 앞으로도 A 크기 정도는 움직임이 비슷하겠지”
  —> 계절에 따른 패턴을 제거하는 것이 분해법의 목적 중 하나!
  (ex.) 연말에 항상 증가하는 계절적 특성을 띠는 백화점 매출 → 추후 매출이 어떻게 될지 예측하고자 함 → 계절 조정을 하지 않고 마침 겨울 시즌의 매출을 예측한다면 매출의 순수한 상승인지 계절의 순환에 의한 상승이지 판단할 수 없음

가법적 모형

$$X_t = b_t + s_t + \varepsilon_t;\ s_t = s_{t-m},\ \sum_{i=1}^{m} s_i = 0$$

승법적 모형

$$X_t = b_t \times s_t \times \varepsilon_t;\ s_t = s_{t-m},\ \sum_{i=1}^{m} s_i = m$$

- 가법모형: 계절성분의 진폭이 시계열의 수준에 관계없이 일정한 수준일 때 사용
- 승법모형: 시계열의 수준에 따라 진폭이 달라짐
  → 승법모형에서 각 성분이 '곱'으로 연결되어 있으므로, 양변에 log를 취할 수 있음
  → $log(X_t) = log(b_t) + log(s_t) + log(e_t)$: 로그가법모형
  → 경제 관련 자료들의 경우, 시간의 흐름에 따라 점차 증가하며 이분산성을 줄이고 단위 scaling의 효과도 볼 수 있는 로그가법형태를 분석에서 많이 사용함

- 분해법에 의한 예측 절차
  1. 중심이동평균으로 평활치 산출(특정 시점 전후의 data의 평균 사용)
  2. 추세 제거한 시계열 산출(계절성 계산을 위해)
  3. 계절성 지수 산출
  4. 계절성 제거한 시계열 산출
  5. 회귀모형으로 추세 추정
  6. 예측치 산출

- step 1.

$$CM_t = \begin{cases} \frac{1}{m}\left(0.5X_{t-q} + X_{t-q+1} + \cdots + X_{t+q+1} + 0.5X_{t+q}\right) & m = 2q \text{ (주기 짝수)} \\ \frac{1}{m}\left(X_{t-q} + X_{t-q+1} + \cdots + X_{t+q}\right) & m = 2q+1 \text{ (주기 홀수)} \end{cases}$$

m개의 데이터 사용해서 구한 중심이동평균

- step 2~3. 추세 제거한 시계열 = $DX_t^{(T)} = X_t / CM_t$
  → 계절별 추세제거 시계열값의 평균으로 계절성 지수 s_i산출(s_i의 평균은 1이 되도록)

- step 4. 계절성 제거한 시계열 = $DX_t^{(S)} = X_t / s_t$

- step 5. 추세 추정: 산점도 등을 통해 1차회귀식 or 2차회귀식 적합

1차식 이용 $DX_t^{(S)} = \beta_0 + \beta_1 t + \varepsilon_t \Rightarrow b_t = \hat{\beta}_0 + \hat{\beta}_1 t$

2차식 이용 $DX_t^{(S)} = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t \Rightarrow b_t = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 t^2$

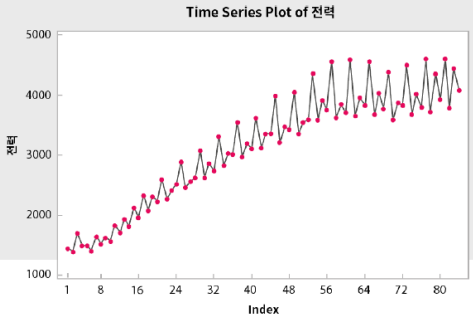
- step 6. 예측치 산출: 추세 X 계절성

$$f_{T,k} = b_{T+k} \times s_{T+k}$$

(예시)

분해법

- 예 (가정상업용 전력사용량) 아래 그림은 1997년-2017년 사이의 분기별 가정 및 상업용 전력 사용량을 나타낸다.
- 분해법 (승법 모형)을 2015년까지 데이터에 적용하여 추세 및 계절성지수를 추정한 후 2016-2017년 분기별 전력사용량을 예측해 보자.



m=4

X_t / CM_t

년도/분기	X_t	CM_t	$DX_t^{(T)}$	s_t	$DX_t^{(S)}$
1997/1	1,461			1.1035	1323.894
/2	1,406			0.9292	1513.185
/3	1,710	1527.75	1.1193	1.0067	1698.607
/4	1,514	1535.25	0.9862	0.9606	1576.157
1998/1	1,501	1530.87	0.9805	1.1035	1360.141
2015/1	4505	3994.62	1.1278	1.1035	4082.234
/2	3684	4009.12	0.9189	0.9292	3964.847
/3	4030			1.0067	4003.149
/4	3801			0.9606	3957.048

#1