제8장 이상값 분석

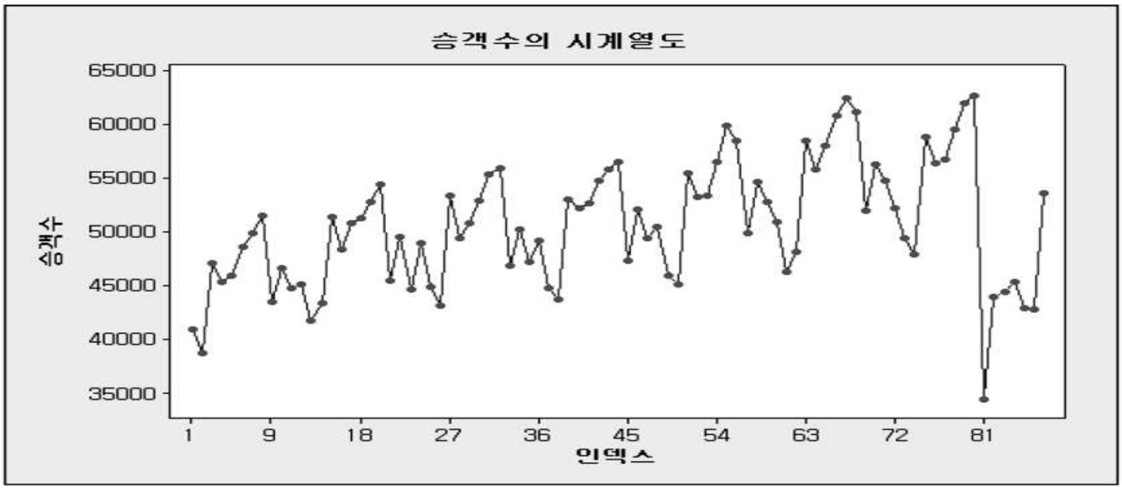

그림 8.1 월별 미국 항공기 승객수(1995년 1월 – 2002년 3월)

■ 구조변화를 포함하는 시계열 모델링

- 알려진 사건 시점 이후의 데이터를 사용할 수 있다.
- 그러나 사건 시점 이후의 관측값이 많지 않을 경우 모델링이 어려울 수 있다.

→ 이러한 사건이 시계열에 미치는 영향의 형태를 몇 가지로 가정하고
전체 데이터를 활용하여 모형을 구축하는 방안을 고려해야 한다.

■ 이상값 탐지(outlier detection) 문제

- 시계열에서 이상값이 존재하면 모형을 잘못 식별할 수 있다.
- 이상값은 존재하는 시점이 알려지지 않은 경우가 대부분이므로
시점을 추정하고 이의 영향을 최소화하는 시계열 모형을 추정해야 한다.

8.1 시점이 알려진 구조변화모형

■ 함수 정의(Box and Tiao, 1975)

① 펄스 함수(pulse function)

$$P_t^{(T)} = \begin{cases} 1, & t = T \\ 0, & t \neq T \end{cases}$$

② 스텝 함수(step function)

$$S_t^{(T)} = \begin{cases} 1, & t \geq T \\ 0, & t < T \end{cases} \quad \text{또는} \quad P_t^{(T)} = S_t^{(T)} - S_{t-1}^{(T)} = (1-B)S_t^{(T)}$$

■ 원시계열이 다음과 같이 수평적 패턴을 가지고 있다: $Z_t = \theta_0 + a_t$

[Case 1]

시점 T에서 사건이 발생하고, 그 직후인 시점 $T+1$에만 시계열에 영향을 일시적으로 준 뒤 시점 $T+2$부터는 다시 제자리로 돌아온다.

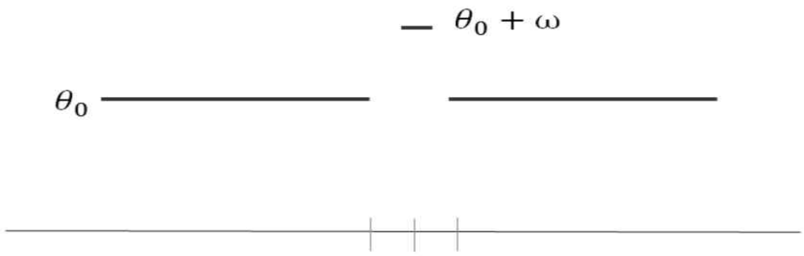

그림 8.2 펄스 형태 영향

$$Z_t = \theta_0 + \omega P_{t-1}^{(T)} + a_t \quad \Rightarrow \quad Z_{T+1} = \theta_0 + \omega + a_{T+1}$$

[일반화]

초기 영향이 $T+b$ 시점에서 일어나고 점진적으로 감소하여 원래의 평균값으로 수렴하는 경우

$$Z_t = \theta_0 + \frac{wB^b}{1-\delta B} P_t^{(T)} + a_t, \ 0 \le \delta \le 1$$

$$Z_t = \theta_0 + wP_{t-b}^{(T)} + w\delta P_{t-b-1}^{(T)} + w\delta^2 P_{t-b-2}^{(T)} + \cdots + a_t$$

$$E[Z_t] = \begin{cases} \theta_0, & t < T+b \\ \theta_0 + w, & t = T+b \\ \theta_0 + w\delta^k, & t = T+b+k, k = 1,2,\cdots \end{cases}$$

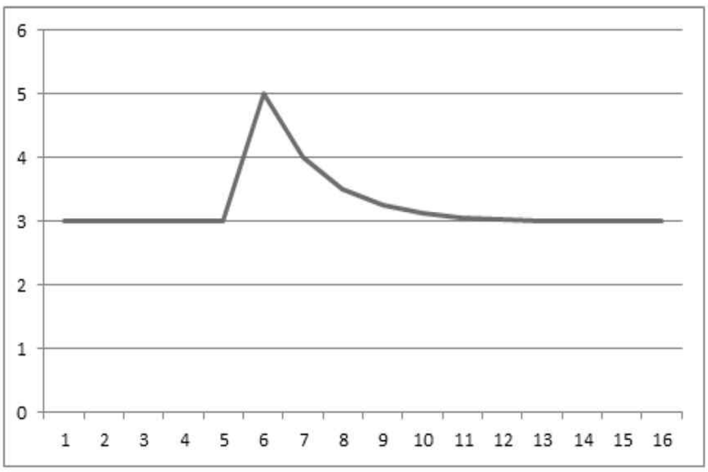

그림 8.5 점진적으로 원래로 복귀하는 형태($T=5,\ \theta_0=3,\ b=1,\ w=2,\ \delta=0.5$)

[Case 2] 사건이 발생한 시점 T 이후부터 시계열 평균이 변화하여 지속되는 경우

그림 8.3 스텝 함수 형태 영향

$$Z_t = \theta_0 + \omega S_{t-1}^{(T)} + a_t \quad \rightarrow \quad Z_t = \theta_0 + \omega B S_t^{(T)} + a_t$$

[일반화]

시점 T에서의 사건의 영향이 1시간 이후가 아닌 b시간 이후에 나타나는 경우

$$Z_t = \theta_0 + \omega B^b S_t^{(T)} + a_t$$

[일반화]

초기 영향이 b시간 이후에 나타나고 점진적으로 증가하여 새로운 평균으로 변하는 경우

$$Z_t = \theta_0 + \frac{wB^b}{1-\delta B} S_t^{(T)} + a_t,\ 0 \le \delta \le 1$$

$$Z_t = \theta_0 + wS_{t-b}^{(T)} + w\delta S_{t-b-1}^{(T)} + w\delta^2 S_{t-b-2}^{(T)} + \cdots + a_t$$

$$E[Z_t] = \begin{cases} \theta_0, & t < T+b \\ \theta_0 + w, & t = T+b \\ \theta_0 + w + w\sum_{j=1}^{k} \delta^j, & t = T+b+k, k = 1, 2, \cdots \end{cases}$$

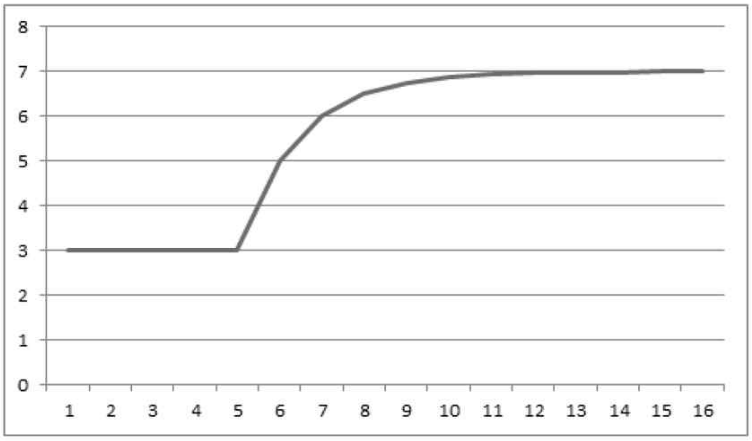

그림 8.4 점진적 증가 형태($T=5,\ \theta_0=3,\ b=1,\ w=2,\ \delta=0.5$)

■ W_t: 정상적 시계열

$$Z_t = W_t + \frac{w(B)B^b}{\delta(B)} I_t^{(T)}$$

$$\text{여기서, } w(B) = w_0 - w_1 B - \cdots - w_s B^s$$

$$\delta(B) = 1 - \delta_1 B - \cdots - \delta_r B^r$$

■ 한 시계열에 여러 사건이 존재하는 경우

$$Z_t = W_t + \sum_{j=1}^{K} \frac{w_j(B)B^{b_j}}{\delta_j(B)} I_{jt}^{(T_j)}$$

여기서, K: 사건의 수, j: 각 사건에 대응하는 것

[예] W_t가 ARMA(p,q)로 모형화되고, $w_j(B) = w_j$, $\delta_j(B) = 1 - \delta_j B$일 때, 시계열은

$$Z_t = \frac{\Theta_q(B)}{\Phi_p(B)} a_t + \sum_{j=1}^{K} \frac{w_j B^{b_j}}{1 - \delta_j B} I_{jt}^{(T_j)}$$

8.2 이상값 분석모형

8.2.1 이상값 형태 및 분석 모형

■ <u>가법적 이상값(additive outlier; AO)</u>

: 시점 T의 값만 변화시키고 그 이후의 시계열에는 영향을 주지 않는 경우

$$Z_t = W_t + wP_t^{(T)} = \frac{\Theta_q(B)}{\Phi_p(B)} a_t + wP_t^{(T)}$$

$$\text{여기서, } P_t^{(T)} = \begin{cases} 1, & t = T \\ 0, & t \neq T \end{cases}$$

■ <u>혁신적 이상값(innovational outlier; IO)</u>

: 시점 T의 값뿐만 아니라 그 이후에도 ARMA 구조를 통하여 모두 영향을 주는 형태의 이상값

$$Z_t = W_t + \frac{\Theta_q(B)}{\Phi_p(B)} wP_t^{(T)} = \frac{\Theta_q(B)}{\Phi_p(B)} \left(a_t + wP_t^{(T)}\right)$$

$$\text{여기서, } P_t^{(T)} = \begin{cases} 1, & t = T \\ 0, & t \neq T \end{cases}$$

■ 한 시계열에 AO 및 IO 형태가 섞인 K개의 이상값이 있는 경우

$$Z_t = W_t + \sum_{j=1}^{K} w_j \nu_j(B) P_t^{(T_j)}$$

여기서, $\nu_j(B) = \begin{cases} 1 & , AO \\ \dfrac{\Theta_q(B)}{\Phi_p(B)}, & IO \end{cases}$, $T_j \ (j=1, \cdots, K)$: 이상값의 존재 시점

[예] W_t가 ARMA(1,1)을 따르고, T_1에서 AO가, T_2에서 IO가 존재할 때, 시계열은

$$Z_t = \frac{1-\theta B}{1-\phi B}(a_t + w_2 P_t^{(T_2)}) + w_1 P_t^{(T_1)}$$

$$Z_t = \phi Z_{t-1} + a_t - \theta a_{t-1} + w_2 P_t^{(T_2)} - \theta w_2 P_{t-1}^{(T_2)} + w_1 P_t^{(T_1)} - \phi w_1 P_{t-1}^{(T_1)}$$

8.2.2 시점을 아는 경우 이상값 검정

■ ARMA모형에서 모든 파라미터를 안다고 할 때, 잔차(residual)는

$$e_t = \pi(B) Z_t$$

여기서, $\pi(B) = \dfrac{\Phi_p(B)}{\Theta_q(B)} = 1 - \pi_1 B - \pi_2 B^2 - \cdots$

■ 시점 T에 AO 또는 IO가 있는 경우 잔차는

$$e_t = \begin{cases} w\pi(B) P_t^{(T)} + a_t, & AO \\ w P_t^{(T)} + a_t & , IO \end{cases}$$

■ 시점 T에 AO가 있는 경우 잔차는

$$e_t = w\pi(B)P_t^{(T)} + a_t$$

• 잔차에 대한 n개의 관측값이 있을 때

$$\begin{bmatrix} e_1 \\ \vdots \\ e_{T-1} \\ e_T \\ e_{T+1} \\ \vdots \\ e_n \end{bmatrix} = w \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ -\pi_1 \\ \vdots \\ -\pi_{n-T} \end{bmatrix} + \begin{bmatrix} a_1 \\ \vdots \\ a_{T-1} \\ a_T \\ a_{T+1} \\ \vdots \\ a_n \end{bmatrix}$$

• AO 모형의 경우, 시점 T에서 계수 w의 최소제곱 추정값과 그 분산

$$\hat{w}_T^A = \frac{e_T - \sum_{j=1}^{n-T} \pi_j e_{T+j}}{1 + \sum_{j=1}^{n-T} \pi_j^2}$$

$$Var(\hat{w}_T^A) = \frac{\sigma_a^2}{1 + \sum_{j=1}^{n-T} \pi_j^2}$$

■ 시점 T에 IO가 있는 경우 잔차는

$$e_t = wP_t^{(T)} + a_t$$

• 잔차에 대한 n개의 관측값이 있을 때

$$\begin{bmatrix} e_1 \\ \vdots \\ e_{T-1} \\ e_T \\ e_{T+1} \\ \vdots \\ e_n \end{bmatrix} = w \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} a_1 \\ \vdots \\ a_{T-1} \\ a_T \\ a_{T+1} \\ \vdots \\ a_n \end{bmatrix}$$

• IO 모형의 경우, 시점 T에서 계수 w의 최소제곱 추정값과 그 분산

$$\hat{w}_T^I = e_T \qquad Var(\hat{w}_T^I) = \sigma_a^2$$

■ 이상값 형태에 대한 가설검정

• 가설

H_0: 시점 T의 이상값이 AO 또는 IO 형태가 아니다.

H_1: 시점 T의 이상값이 AO 형태이다.

H_1: 시점 T의 이상값이 IO 형태이다.

• 검정통계량

$H_1 \quad vs \quad H_0: \lambda_{1T} = \dfrac{\hat{w}_T^A \sqrt{1 + \sum_{j=1}^{n-T} \pi_j^2}}{\sigma_a}$	$H_2 \quad vs \quad H_0: \lambda_{2T} = \dfrac{\hat{w}_T^I}{\sigma_a}$

※ H_0이 옳을 때, $\lambda_{1T} \sim N(0, 1^2)$, $\lambda_{2T} \sim N(0, 1^2)$

• 판정

① 표준정규분포 하에서 유의수준에 대응하는 기각역을 구한다.

② λ_{1T}가 기각역에 속하면 AO 형태의 이상값으로 판정한다.

λ_{2T}가 기각역에 속하면 IO 형태의 이상값으로 판정한다.

두 경우 모두 기각하지 못하면 이상값이 아닌 것으로 판정한다.

8.2.3 시점을 모르는 경우 이상값 탐지

■ 많은 경우 이상값 시점들을 모르는 상황에서 시계열 모형 파라미터를 추정해야 하는 어려움에 직면하게 된다. 이상값이 포함된 시계열의 경우 모형 파라미터가 제대로 추정될 수 없다. 이상값은 특히 오차항 분산 추정에 큰 영향을 주는 것으로 알려져 있다.

■ Chang et al.(1988): 시계열에 AO 또는 IO 형태의 이상값이 여러 개 있을 수 있는 상황에서 이상값을 탐지하는 반복적 절차를 제안

① 이상값이 없는 것으로 가정하고 시계열 모형을 추정한다. 잔차 $\hat{e}_t$들을 계산하고 오차항 분산의 초깃값을 다음으로 추정한다.

$$\hat{\sigma}_a^2 = \frac{1}{n} \sum_{t=1}^{n} \hat{e}_t^2$$

② 각 시점에 대한 검정통계량 $\hat{\lambda}_{1t}$와 $\hat{\lambda}_{2t}$ $(t=1,2,\ \cdots,n)$를 계산하면 다음을 만족하는 시점 T를 찾는다.

$$\hat{\lambda}_T = \max_t\left(\left|\hat{\lambda}_{1t}\right|,\ \left|\hat{\lambda}_{2t}\right|\right)$$

- $\hat{\lambda}_T = \left|\hat{\lambda}_{1T}\right| > c$ (c는 미리 정한 3 정도의 상수)

→ 시점 T에 AO 형태의 이상값이 있다고 판정

→ 오차항 분산을 조정된 잔차로 다시 추정

$$\hat{e}_t \leftarrow \hat{e}_t - \hat{w}_{AT}\,\hat{\pi}(B)I_t^{(T)},\ \ t \geq T$$

- $\hat{\lambda}_T = \left|\hat{\lambda}_{2T}\right| > c$ (c는 미리 정한 3 정도의 상수)

→ 시점 T에 IO 형태의 이상값이 있다고 판정

→ 오차항 분산을 조정된 잔차로 다시 추정

$\hat{e}_t \leftarrow \hat{e}_T - \hat{w}$

③ 조정된 잔차와 새로운 오차항 분산 추정치를 사용하여 ②를 반복하여 다른 이상값의 존재 여부를 판단한다. 이때 모형 파라미터 $\hat{\pi}(B)$는 초깃값을 그대로 사용한다.

④ 총 K개의 이상값이 탐지되었다고 하고 이들의 시점을 $T_1,\ \cdots,T_K$라 하자. 이들 시점들이 알려진 것으로 간주하고 이상값 계수들 $w_1,\ \cdots,w_K$를 추정한다. 또한 다음 모형으로부터 파라미터들을 동시에 추정한다.

$$Z_t = \frac{\Theta(B)}{\Phi(B)}a_t + \sum_{j=1}^{K} w_j \nu_j(B) P_t^{(T_j)}$$

그리고 잔차를 아래와 같이 조정하고 오차항 분산을 다시 추정한다.

$$\hat{e}_t = \hat{\pi}(B)\left[Z_t - \sum_{j=1}^{K}\hat{w}_j\hat{\nu}_j(B)P_t^{(T_j)}\right]$$

- Chen and Liu(1993)

- Bui and Jun(2012)