Kalman-Filter기법을 사용한 소지역 추정

이상은[1] 신민웅[2]

1 서론·

최근 시계열 모형을 이용한 소지역 추정방법이 많은 관심을 끌고 있다. 특히, 관측값에서 발생 할 수 있는 조사(survey) 오차가 시간상에서 상관관계를 갖고,모형화가 합리적으로 만들어 질수 있을 때, Kalman-Filter(KF) 기법이 소지역 총계의 시계열 EBLUP을 구하기 위하여 사용된다. Kalman-Filter기법이 이용되는 예로 소지역의 실업자수 추정에서 설명변수로 경제활동 인구수를 생각할 때,센서스-이후의 관찰된 실업자수를 보정할 때 사용된다 그 이유는 센서스를 실시할 때 여러 가지 이유로 조사오차가 발생하기 때문이다.

2. Kalman Filter 모형

Y_t가 시간 $t(=1,2,\cdots)$에서 변수 y의 관찰값이면. KF모형에서 관측방정식은

$$Y_t = F_t' \theta_t + v_t$$

이다.

여기서 F_t는 알려진 보조변수들의 값들의 벡터이고, v_t는 정규분포 $N(0, V_t)$인 확률변수로 V_t는 기지이며

$$\theta_t = G_t \theta_{t-1} + w_t$$

이라고 가정하고 G_t는 기지의 행렬이고,

$$w_t \sim N(0, W_t)$$

로 W_t는 기지이며

$$w_t \sim iid \quad v_t$$

라 하자.

1 경기대학교 응용정보통계학과 조교수 sanglee@stat.kyonggi.ac.kr
2 한국외국어대학교 정보통계학과 교수 mwshin@stat.hufs.ac.kr

여기서 θ_t 는 다음과 같이 추정한다.

Y_t를 관찰하기 전에 θ_t의 사전 분포로 보면 사후분포는 다음과 같다.

$$\theta_t | Y_t \propto P(\theta_t | Y_{t-1}) P(Y_t | Y_{t-1}, \theta_t)$$

Y_t를 관찰하기 바로 전에 Y_t를 예측하고, $\hat{Y}_t$은 Y_t의 예측값이면. 예측오차는 다음과 같다.

$$(e_t | \theta_t, Y_{t-1}) \sim N(F'_t(\theta_t - G_t \hat{\theta}_{t-1}), V_t)$$

여기서 F_t, G_t, $\hat{\theta}_{t-1}$은 기지이므로, Y_t를 관찰하는 것은 e_t를 관찰하는 것과 동치이다.

그러므로 사후분포는

$$P(\theta_t | Y_t) \propto P(\theta_t | Y_{t-1}) P(e_t | Y_{t-1}, \theta_t)$$

이며 ,여기서

$$(e_t | \theta_t, Y_{t-1}) \sim N(F'_t(\theta_t - G_t \hat{\theta}_{t-1}), V_t)$$

$$(\theta_t | e_t, Y_{t-1}) \sim N(G_t \hat{\theta}_{t-1} + R_t F_t (F'_t R_t F_t + V_t)^{-1} e_t ,$$

$$R_t - R_t F_t (F'_t R_t F_t + V_t)^{-1} F'_t R_t)$$

이 되며, $R_t = G_t \sum_{t-1} G'_t + W_t$ 이다.

3 소지역에서의 응용

KF-기법을 소지역의 실업자수 추정에 활용 해보록 하자.

Y_{ta} = 시간 t에서 소지역 a의 실업자수

X_{ta} = 시간 t에서 소지역 a의 경제활동인구, $a = 1, \cdots, A$

KF-모형은 다음과 같다.

$$Y_{0a} = X_{0a} \beta_0 + v_0 \qquad (3.1)$$

$$v_0 \sim N(0, V_0)$$

여기서 β_0와 V_0는 지역에 독립인 모수들 β_0는

$$\beta_0 \sim N(b_0, W_0)$$

를 갖고, b_0는

$$\hat{b}_0 = \frac{1}{A} \sum_{a=1}^{A} \frac{Y_{0a}}{X_{0a}}$$

으로 W_0는

$$\hat{W}_0 = \frac{1}{A-1} \sum_{a=1}^{A} \left(\frac{Y_{0a}}{X_{0a}} - \hat{b}_0 \right)^2$$

으로 추정된다.

그로므로, 소지역 a에 대하여, Y_{0a} 를 관찰한 후에 β_0 의 베이즈추정치는

$$\hat{\beta}_{0a} = \frac{Y_{0a} X_{0a} W_0 + b_0 V_0}{X_{0a}^2 W_0 + V_0}, \quad a = 1, \cdots, A$$

$$\Sigma_{0a} = \frac{V_0 W_0}{V_0 + X_{0a}^2 W_0}$$

이며, (3.1) 로부터

$$\hat{V}_0 = \frac{1}{A-1} \sum_{a=1}^{A} (Y_{0a} - X_{0a} \overline{\beta}_0)^2$$

$$\hat{\beta}_0 = \sum_a X_{0a} Y_{0a} / \sum_a X_{0a}^2$$

이 된다.

그러므로, $\beta_1 \mid Y_{0a}$ 의 사후 분포는

$$N(\hat{\beta}_{0a}, \Sigma_{0a} + \hat{W}_0 \fallingdotseq \hat{R}_{0a})$$

이므로,

$$\hat{Y}_{1a} = \hat{\beta}_{0a} X_{1a}$$

$$V(Y_{1a} \mid Y_{0a}) = X_{1a}^2 R_{1a} + V_1$$

이다.

이때, $V(Y_{1a} \mid Y_{0a})$ 를 계산하기 위하여, R_{ta} 와 V_t 의 최근에 가능한 값, 즉 $\hat{R}_{0a}$ 와 $\hat{V}_0$ 을 사용한다.

4. 토의

KF-기법을 과거의 자료를 사용하여 t 시점에서 응답 변수의 값을 예측하는데 이용될 수 있다. 또한, KF-기법을 여러 소지역들에 대한 t시점의 인구를 관측한 후에 사후-센서스 인구(t시점의 인구)를 t시점 시점의 인구를 관측한 후에 보정(correction)하는 데 응용된다. 소지역 추정문제를 다루는 데 있어서, 총계등의 합성추정량을 추정하기 위해서는 유사한 지역으로 층화하므로서 더 효율적이 된다. 특히, 많은 변수가 영향을 주는 경우에는 추정하고자 하는 변수에 영향을 주는 여러 설명변수를 찾아 내어서 KF추정치를 사용하는 것이 잡음을 제거하는 데 유용하다.

참고문헌

1. 시계열 분석의 원리(2001). 이상열, 자유아카데미
2. 표본설계(2001) 신민웅, 이상은, 교우사
3. 캐나다 노동력 조사 방법론(2001) 통계기획국,조사관리과
4. Small area estimation in survey sampling(1998) Parimal Mukhopadhyay
5. Introduction to small area estimation(2001) J.N.K.Rao. ISI(2001,Korea)
6. Singh,M.P.,Gambino.J. and Mantel.H.J.(1994). Issues and strategies for small area data.Survey Methodology.20(1).3-22