응용통계연구 제 16권 2호, 2003년, pp. 239-246

KALMAN FILTER기법을 이용한 실업자 수의 소지역 추정

양영춘 [1] 이상은 [2] 신민웅 [3]

요 약

소지역에서 직접(direct) 시계열추정을 할 수 있다면, 소지역들 추정에서 최적선형 불편 예측량(BLUP)을 일반화 시킬 수 있다. 특히 조사에서 얻어지는 관측 값의 오차가 시간상으로 상관관계가 있다면 Kalman Filter(KF)기법이 사용 될 수 있다. 이 연구는 예측 값을 활용한 소지역의 실업자 수 추정에서 표본으로 추출되지 않은, 즉 관측되지 않은 값의 예측모형에 KF기법을 적용하였다. 이는 경제활동인구수를 이용하여 현 시점의 소지역 실업자 수를 예측함수(BLUP)를 통해 추정하게 된다. 그리고 이를 단순 회귀분석 추정치와 비교하였다.

주요용어: Kalman-Filtering, 소지역 추정, BLUP

1. 서론

직접추정 혹은 간접 혹은 복합 추정, 즉 자료를 기반으로 하는 소지역의 추정은 물론 모델을 기반으로 하는 추정방법의 이론적 연구는 매우 활발하게 진행되고 있다. 또한 모형기반의 추정치가 자료기반의 추정치보다 안정적임은 이미 우리가 알고 있다. 그러므로 추정하고자하는 변수의 가장 적절한 모델을 찾는 것이 더 좋은 소지역 통계치를 얻는 관건이 됐다.

우리나라의 경우 실업자 수의 추정은 경제활동인구 조사에서 얻어지며 이때 조사의 관측 값의 오차가 시간과 관계가 있음을 착안하여 시계열 모형을 시도하기 전에 필터링 기법을 활용하고자한다. 일반적으로 관측오차에 필터링의 기법을 적용한다. 그러나 이 연구에서는 자료의 특성으로 관측오차대신 모수추정의 오차에 필터링 기법으로 KF기법을 이용하여 소지역의 BLUP를 적용하였다.

이 연구에서는 예측 값을 이용한 소지역 통계(이상은, 2001)의 모형에서 표본으로 추출되지 않은, 즉 관측되지 않은 값을 예측하는 모형에 KF기법을 적용하였다. 모의실험 자료로는 실업자 수를 추정 할 수 있는 1999년 경기도 경제활동인구조사 자료를 이용하였다.

이 연구에서의 베이즈 추정에 관한 시뮬레이션은 Winbugs에 의해 구하였다.

1) (450-701) (442-760) 경기도 수원시 팔달구 이의동 산 94-6, 경기대학교 경제학부 응용정보통계전공, 대학원
E-mail : tender1@kyonggi.ac.kr
2) (442-760) 경기도 수원시 팔달구 이의동 산 94-6, 경기대학교 경제학부 응용정보통계전공, 조교수
E-mail : sanglee@stat.kyonggi.ac.kr
3) (449-791) 경기도 용인시 한국 외국어 대학 정보통계학과, 교수
E-mail : mwshin@stat.hufs.ac.kr

Winbugs란 Windows Bayesian Using Gibbs Sampling의 약자이며 마코프 연쇄 몬테카를로(Markov Chain Monte Carlo : MCMC)기법을 활용하여 복잡한 통계적 모형에서의 베이지안 분석을 위한 벅스(BUGS)의 윈도우 버전 프로그램이다.

www.mrc – bsu.cam.ac.uk/bugs 에서 다운받아 활용할 수 있다.

2. 예측 값을 이용한 소지역통계 모형

예측 값을 이용한 소지역통계 모형은 소지역에서의 관심 있는 변수(반응변수) Y , 를 표본으로 추출되어 관측된 값 $Y^{(1)}$ (이 후의 논문에서는 관측된 값으로 표기함)과 표본으로 추출되지 않아 관측되지 않은 값 $Y^{(2)}$ (이 후의 논문에서는 관측되지 않은 값으로 표기함)으로 나눈다. 이 때 Y 의 추정을 $Y^{(1)}$ 과 $Y^{(2)}$ 의 가중평균으로 구하게 되며 모형설정은 다음과 같다.

a 지역의 반응변수의 Y_a 값 를 두 부분 $Y_a^{(1)}$, $Y_a^{(2)}$로 나눈다. 이는 a 지역에서 관측된 값과 관측되지 않은 값으로 각각 표기한다. $a = 1, \cdots, A$

$$Y_a = \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \end{pmatrix}$$

여기서

$Y_a^{(1)} = (Y_1^{(1)}a, \cdots, Y_{n_a a}^{(1)})'$ 지역에서관측되어진 값(sampled unit)

$Y_a^{(2)} = (Y_{n_a+1a}^{(2)}, \cdots, Y'(2)_{N_a a})'$ 지역에서 관측 되지 않은 값(unsampled unit)

n_a a 지역의 표본수 $(n = \sum_{a=1}^{A} n_a)$

N_a a 지역의 모집단 수

이 때의 Y_a 추정을 다음과 같이 한다.

a 소지역의 총계 추정량 $\hat{Y}_{a(tot)}$ 은

$$\hat{Y}_{a(tot)} = N_a[f_a\hat{y}_a + (1 - f_a)\hat{Y}_a] \tag{2.1}$$

이며 여기서

$f_a = \frac{n_a}{N_a}$

$\hat{y}_a$ 는 표본평균

$\hat{Y}_a$ 는 모형을 이용한 평균 예측 값이 된다.

이 연구에서는 $\hat{Y}_a$의 값을 얻기 위한 모형에 KF를 적용하였다. 예측 값을 이용한 소지역통계의 기본 모델을 이상은(2001)에 참고하기 바란다.

3. Kalman Filter(KF) 모형

KF의 목적은 오차를 포함하고 있는 관측 값으로부터 관측방정식을 추정하는 것이기도 하다. 여기서는 과거의 자료를 사용하여 시점 t 에서의 반응 변수를 예측하는 모델로 모수

의 상태방정식을 이용한다.

Y_t는 시간 $t(=1,2,\cdots)$에서 변수 Y의 관찰 값이면 KF 모형에서 관측방정식은 다음과 같다.

$$Y_t = F_t'\theta_t + \nu_t \tag{3.1}$$

여기서 F_t 는 알려진 보조변수들의 값들의 벡터이고, ν_t 는 정규분포 $N(0,\nu_t)$인 확률변수로 ν_t는 기지이며, 모수의 상태방정식은

$$\theta_t = G_t\theta_t + \omega_t \tag{3.2}$$

이며 여기서 G_t는 기지의 행렬이며 $\omega_t \sim N(0,\Omega_t)$로 Ω_t 는 기지이다.

$$\omega_t \sim^{iid} \nu_t \tag{3.3}$$

θ_t의 추정을 다음 같이 한다.

우선 Y_{t-1}이 관측된 후 θ_{t-1}의 베이즈 추정량은 사후분포 $\theta_t|Y_{t-1}$, 의 평균을 $\hat{\theta}_{t-1}$, 그의 분산을 Σ_{t-1} 이라 하자. 이때 식(3.2)를 이용하면

$$\theta_t|Y_{t-1} \sim N(G_t\hat{\theta}_{t-1}, G_t\sigma_{t-1}G_t' + \Omega_t = R_t) \tag{3.4}$$

이 된다.

이 (3.3)식으로 부터 θ_t 의 사전분포함수를 이용하여 Y_t 가 관측된 후 θ_t 의 사후분포 $\theta_t|Y_t$ 는 다음과 같이 얻어진다.

$$\theta_t|Y_{t-1} \propto P(\theta_t|Y_{t-1})P(Y_t|Y_{t-1},\theta_t) \tag{3.5}$$

여기서 Y_t 를 관찰하기 바로 전에 Y_t를 예측하고, Y_t의 예측 값을 $\hat{Y}_t$라 할 때 예측오차 $\epsilon_t = Y_t - \hat{Y}_t$ 의 분포는 다음과 같다.

$$(\epsilon_t|\theta_t, Y_{t-1}) \sim N(F_t'(\theta_t - G_t\hat{\theta}_{t-1}), \nu_t) \tag{3.6}$$

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

그러므로 사후분포는 다음과 같이 쓸 수 있다.

$$P(\theta_t|Y_t) \propto P(\theta_t|Y_t - 1)(\epsilon_t|Y_t - 1, \theta_t) \tag{3.7}$$

여기서

$$(\epsilon_t|\theta_t, Y_{t-1}) \sim N(F_t'(\theta_t - G_t\hat{\theta}_{t-1}), \nu_t)$$

$$(\theta_t|\epsilon_t, Y_{t-1}) \sim N(G_t\hat{\theta}_{t-1} + R_tF_t(F_t'R_tF_t + \nu_t)^{-1}\epsilon_t, R_t - R_tF_t(F_t'R_tF_t + \nu_t)^{-1}F_t'R_t)$$

이 되며 $R_t = G_t\sum_{t-1}G_t' + \Omega_t$이다.

4. 소지역에 적용하는 KF 모형과 단순회귀 모형

KF 모형의 관측(3.1), 상태(3.2) 방정식을 다음과 같이 설정한다.

$$Y_{t,a_j} = X_{t,a}\beta_{t,a} + \nu_t, \nu_t \sim N(0, \nu_t) \tag{4.1}$$

$$\beta_{t,a} = \beta_{t-1,a} + \omega_t, \omega_t \sim N(0, \Omega_t) \tag{4.2}$$

여기서Y_{t,a_j} = 시간 t 에서 소지역 a의 j번째 가구의 실업자수,
X_{t,a_j} = 시간 t 에서 소지역 a의 j번째 가구의 경제활동인구, $j = 1, \cdots, n_a, a = 1, \cdots, A$
$Y_{t,a} = \sum_{j=1}^{N_a} y_{t,a_j}, X_{t,a} = \sum_{j=1}^{n_a} x_{t,a_j}$
ν_t 와 Ω_t 는 알려진 값으로 ω_t와 ν_t는 서로 독립임을 가정한다.
이때 $\beta_{t,a}$ 의 사후분포는 다음과 같다.

$$\beta_{t,a}|Y_{t,a} \sim N(\hat{\beta}_{t,a}, \Lambda_{t,a}) \tag{4.3}$$

여기서
$\hat{\beta}_{t,a} = \hat{\beta}_{t-1,a} + R_{t,a})(X_{t,a}^2 + \nu_t)^{-1}\epsilon_{t,a}$
$\Lambda_{t,a} = R_{t,a} - R_{t,a}(X_{t,a}^2 + \nu_t)^{-1}X_{t,a}R_{t,a}$
$\epsilon_{t,a} = Y_{t,a} - X_{t,a}\hat{\beta}_{t-1,a}$
$\hat{\beta}_{t-1,a} = E(\beta_{t-1,a}|Y_{t-1,a})$
$R_{t,a} = Var(\beta_{t,a}|Y_{t-1,a}) = \sum_{t-1,a} + \Omega_{t,a}$
여기서 $\hat{\beta}_{t-1,a} = E(\beta_{t-1,a}|Y_{t-1,a})$ 과 $R_{t,a} = Var(\beta_{t,a}|Y_{t-1,a})$은 t-1시점에서의 회귀모형에서의 베이즈 추정치를 winbug에 의해 계산하였다.

이제 KF에 의한 추정량 $\beta_{t,a}$,과 소지역의 경제활동인구수 $X_{t,a}$를 이용하여 소지역에서 관측되지 않은 실업자 수를 예측하고 이를 추정된 소지역의 평균 실업자수$\hat{Y}_{t,a}$,라고 하자. 그러면 위의 결과를 이용하여 소지역의 실업자수의 추정량 $\hat{Y}_{a(tot)}$는 다음과 같다.

$$\hat{Y}_{a(tot)} = N_a[f_a\bar{Y}_{t,a} + (1 - f_a)\hat{Y}_{t,a}], f_a = \frac{n_a}{N_a} \tag{4.4}$$

여기서
$\bar{Y}_{t,a}$: 관측된 평균 실업자 수
$\hat{Y}_{t,a}$: 추정된 소지역의 평균 실업자 수로 다음과 같이 얻어진다.
N_a : a 지역의 전체 가구 수
n_a : a 지역의 표본 가구 수
마찬가지로 단순회귀식을 이용하면 다음과 같다. 소지역의 전체 실업자 수의 추정량 $\hat{Y}_{a(tot_R)}$은

$$\hat{Y}_{a(tot_R)} = N_a[f_a\bar{Y}_a + (1 - f_a)\hat{Y}_a], f_a = \frac{n_a}{N_a} \tag{4.5}$$

이며, $\bar{Y}_a$: 관측된 평균 실업자 수 $\hat{Y}_a$: 추정된 소지역의 평균 실업자 수는 다음과 같이 얻는다.

$$\hat{Y}_a = \bar{X}_a\hat{\beta}, \hat{\beta} = (X'X)^{-1}XY$$

마지막으로 위에서 언급한 추정량을 비교하기위해 편의(Bias)와 평균제곱오차(MSE)를 다음과 같이 계산하였다.

$$Bias = \frac{1}{R}\sum_{r=1}^{R}\hat{Y}_{a(tot)} - Y_a$$

$$MSE = \frac{1}{R}\sum_{r=1}^{R}(\hat{Y}_{a(tot)} - Y_a)^2$$

여기서
R 은 반복실험횟수
$\hat{Y}_{a(tot)}$는 소지역 a 의 추정된 총계 값
Y_a는 소지역 a의 실제 총계

5. 모의실험

5.1. 자료소개

경제활동인구조사는 취업, 실업, 노동력 등과 같은 인구의 경제적 특성을 조사하여 노동공급, 노동투입, 고용구조, 가용노동시간 및 인력자원의 활용정도 파악, 고용 창출 등을 위한 정부정책 입안 및 평가자료 제공을 목적의 조사이다.

이 연구에서는 1999년도 5월 경기도의 경제활동인구조사에서 조사된 자료를 모집단으로 하였다.

이 때 경기도 23개시 각각의 전체 모집단의 실업자 수의 값(true value)은 다음과 같다.

시	31001	31002	31003	31004	31005	31006	31007	31008
실업자수	28	34	21	32	41	6	18	3
시	31009	31010	31011	31012	31013	31014	31015	31016
실업자수	21	15	1	2	6	2	12	3
시	31017	31018	31019	31020	31021	31022	31023	
실업자수	3	6	4	7	4	2	3	

모의실험의 표본은 모집단을 행정구역 (23개의 시)순으로 나열한 후 모집단의 약 20 percert 인 500개의 가구를 계통추출 하였으며 이를 1000번 반복실험 하였다.

5.2. 결과 요약

3장에서 제시한 2가지 방법(KF, 회귀모형)에 의한 결과는 다음과 같다.

시	$\hat{Y}_{(tot)_R}$	$\hat{Y}_{tot}$	$Bias(\hat{Y}_{(tot)_R})$	$Bisa(\hat{Y}_{tot})$	$MSE(\hat{Y}_{(tot)_R})$	$MSE(\hat{Y}_{tot})$
31001	19.58	26.67	8.42	1.33	71.31	2.81
31002	23.85	32.92	10.15	1.08	103.54	1.82
31003	14.65	15.06	6.35	5.94	40.73	36.70
31004	22.31	30.32	9.69	1.68	94.35	3.43
31005	28.60	32.36	12.40	8.64	154.34	76.21
31006	4.19	2.99	1.81	3.01	3.27	8.79
31007	12.65	13.89	5.35	4.11	28.74	17.78
31008	2.10	3.68	0.90	0.68	0.85	0.39
31009	14.59	15.30	6.41	5.70	41.15	32.42
31010	10.48	16.67	4.52	1.67	20.51	2.36
31011	0.71	0.76	0.29	0.24	0.09	0.14
31012	1.38	2.30	0.62	0.30	0.40	0.07
31013	4.19	4.21	1.81	1.79	3.32	3.41
31014	1.38	0.56	0.62	1.44	0.40	2.0
31015	8.44	10.44	3.56	1.56	12.67	2.56
31016	2.05	2.47	0.95	0.53	0.92	0.39
31017	2.10	4.69	0.90	1.69	0.83	2.56
31018	4.22	3.17	1.78	2.83	3.20	7.88
31019	2.80	3.02	1.20	0.98	1.44	0.96
31020	4.93	3.44	2.07	3.56	4.30	12.48
31021	2.79	1.71	1.21	2.29	1.47	4.75
31022	1.40	0.97	0.60	1.03	0.36	1.06
31023	2.09	5.01	0.91	2.01	0.88	3.47

* $\hat{Y}_{tot_R}$: 단순회귀모델$\hat{Y}_{tot}$: KF 모델

위의 표로부터 단순회귀모형에서 KF모형으로 오차항을 보정해 줌으로써 대부분의 경우 편의(Bias)는 물론 평균제곱오차(MSE)가 크게 줄었음을 볼 수 있다.

6. 토의

소지역 통계에서 센서스-추정치들은 여러 가지 오류에 영향을 받으므로 KF추정치는 사후-센서스 추정치로서 이러한 오류를 제거한다. 그러나 센서스 추정치와 KF추정치 간의 현저한 차이가 있을 때에는 에디팅(editing)을 하거나, 비 표본 오차 등을 조사하여야 한다. 그리고 들도 유사한 소지역으로 묶어서 추정함으로서 소지역 추정의 효율성을 높일 수 있다.

자금까지의 소지역 연구에서는 자료에 KF를 적용하거나 시계열 모형을 적용하여 추정하였다. 그러나 실험에서 얻어지는 자료가 아닌 실업자 수의 경우 모수에 KF기법을 적용함으로서 자료에서 얻을 수 있는 오류에 반해 모수의 추정에서 얻을 수 있는 오류를 수정하였다.

참고문헌

[1] 신민웅, 이상은(2001)*표본설계, 교우사*

[2] 이상은(2001) *Bayes Prediction for Small Area Estimation* ¡한국통계학회논문집¿, 8, 407쪽 416쪽

[3] Parimal Mukhopadhyay(1998)*Small area estimation in survey sampling*

[4] 통계기획국,조사관리과(2001) *캐나다 노동력 조사 방법론*

[5] J.N.K.Rao.(2001) *Introduction to small area estimation* ¡2001년 ISI proceeding¿

[6] Singh,M.P.,Gambino.J. and Mantel.H.J.(1994). *Issues and strategies for small area* ¡Survey Methodology¿ .20

[2003년 1월 접수, 2003년 5월 채택]

Small Area Estimation of Unemplyoment Using Kalman Filter Method

Young Chun Yang [1] Sang Eun Lee[2] Min Woong Shin [3]

ABSTRACT

In small area estimation, Best Linear Unbaised Predictor(BLUP) can be directly implicated ,specially, in use of the time series estimation. If there are correlations between observations and error terms over the time, Kalman Filter method can be used. Therefore, using kalman Filtering technique small area estimation of total of unemployments are estimated by BLUP. And for the example of this study, Economic Active Population Survey data were used.

Keywords: Kalman-Filtering; small area estimation; BLUP

1) Graduate student, Department of Statistics, Kyonggi University.
E-mail : tender1@kyonggi.ac.kr
2) Assistant Professor, Department of Statistics, Kyonggi University.
E-mail : sanglee@stat.kyonggi.ac.kr
3) Professor, Department of Statistics, Hankook University of Foreign Studies.
E-mail : mwshin@stat.hufs.ac.kr