K-F기법으로 실업자 수의 소지역추정
- 경제활동인구조사를 중심으로 -

양영춘 이상은[1] 신민웅[2]

요 약

소지역에서 직접(direct) 시계열추정을 할 수 있다면, 소지역들 추정에서 최적선형 불편예측량(BLUP)을 일반화 시킬 수 있다. 특히 조사에서 얻어지는 관측 값의 오차가 시간상으로 상관관계가 있다면 Kalman-Filter(K-F)기법이 사용 될 수 있다. 이 연구는 소지역의 실업자 수 추정에서 K-F기법으로 경제활동인구수를 이용하여 현 시점의 소지역 실업자 수를 예측함수(BLUP)를 통해 추정하였다. 그리고 단순 회귀분석 추정치와 비교하였다.

keywords : Kalman-Filtering, 소지역 추정, BLUP

1. 서론

직접추정치 혹은 간접 혹은 복합 추정치 , 즉 자료를 기반으로 하는 소지역의 추정방법에서 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.

최근 소지역통계에서 시계열모형의 적용으로 향상된 추정치를 얻고 있다. 시계열 모형의 접근은 소지역에서 직접 시계열의 추정이 가능하면 BLUP을 일반화 할 수 있다.

우리나라의 경우 실업자 수의 추정은 경제활동인구 조사에서 얻어지며 이때 조사의 관측 값의 오차가 시간과 관계가 있음을 착안 K-F기법을 이용하여 소지역의 BLUP를 적용하기로 했다.

이 연구는 실업자 수를 얻는 조사에서 회귀모형보다 시계열 모형에 더욱 합리적임을 예측하고 시간상에 나타날 수 있는 오차를 보정해 주는 K-F기법을 활용하기로 하겠다.

이 연구에서의 베이즈 추정에 관한 시뮬레이션은 Winbug에 의해 구하였다.

1) 경기대학교 응용정보통계학과 대학원 tender1@kyonggi.ac.kr
1) 경기대학교 응용정보통계학과 조교수 sanglee@stat.kyonggi.ac.kr
2) 한국외국어대학교 정보통계학과 교수 mwshin@stat.hufs.ac.kr

2 K-F모형과 단순회귀모형

K-F는 과가의 자료를 이용하여 특정 시점, t ,에서의 종속변수의 값을 예측하는 기법이다. 이를 소지역통계에 적용하기로 하자.

KF-모형은 다음과 같다.

$$Y_{t,a} = X_{t,a}\beta_{t,a} + v_t, \quad v_t \sim N(0, V_t) \tag{2.1}$$

$$\beta_{t,a} = \beta_{t-1,a} + w_t, \quad w_t \sim N(0, W_t) \tag{2.2}$$

여기서

Y_{jta} = 시간 t에서 소지역 a의 j 번째 가구의 실업자수,

X_{jta} = 시간 t에서 소지역 a의 j번째 가구의 경제활동인구,

$j=1,\cdots,n_a$, $a=1,\cdots,A$

$Y_{ta}=\sum_{j=1}^{n_a} y_{jta}$, $X_{ta}=\sum_{j=1}^{n_a} x_{jta}$, V_t와 W_t는 알려진 값으로, $w_t \sim^{iid} v_t$임을 가정한다.

이때 $\beta_{t,a}$의 사후분포는 다음과 같다.

$$\beta_{t,a}|Y_{t,a} \sim N(\mathcal{B}_{t,a}, \lambda_{t,a}) \tag{2.3}$$

여기서

$$\mathcal{B}_{t,a} = \mathcal{B}_{t-1,a} + R_{t,a}(X_{t,a}^2 + V_t)^{-1}e_{t,a}$$
$$\lambda_{t,a} = R_{t,a} - R_{t,a}(X_{t,a}^2 + V_t)^{-1}X_{t,a}R_{t,a}$$
$$e_{t,a} = Y_{t,a} - X_{t,a}\mathcal{B}_{t-1,a}$$
$$\mathcal{B}_{t-1,a} = E(\beta_{t-1,a}|Y_{t-1,a})$$
$$R_{ta} = Var(\beta_{t,a}|Y_{t-1,a}) = \Sigma_{t-1,a} + W_{t,a}$$

여기서 $\mathcal{B}_{t-1,a} = E(\beta_{t-1,a}|Y_{t-1a})$과 $\Sigma_{t-1,a} = Var(\theta_{t-1,a}|Y_{t-1,a})$은 t-1시점에서의 회귀모형에서의 베이즈 추정치이며 으로 winbug에의해 계산 되어졌다.

이제 K-F에 의해 보정된 추정량, $\mathcal{B}_{t,a}$,과 소지역의 경제활동인구수, $X_{t,a}$,를 이용하여 소지역에서 관측되지 않은 실업자 수를 예측하고 이를 추정된 소지역의 실업자수, $\mathcal{Y}_{t,a}$, 라고 하자.

그러면 위의 결과를 이용하여 소지역의 전체 실업자 수의 추정량, $\mathcal{Y}_{tot}$, 는 다음과 같다.

$$\mathcal{Y}_{tot} = f\,\overline{Y}_{ta} + (1-f)\,\mathcal{Y}_{ta}, \quad f = \frac{n_a}{N_a} \tag{2.4}$$

여기서

$\overline{Y}$: 관측된 실업자수

Y_{ta} : 추정된 소지역의 실업자수

여기서 추정된 소지역의 실업자수, Y_{ta}는 다음과 같이 얻는다.

$$Y_{ta} = X_{ta}\beta_{t,a}$$

N_a : a 지역의 전체 가구수

n_a : a 지역의 표본 가구수

마찬가지로 단순회귀식을 이용하면 다음과 같다.

소지역의 전체 실업자 수의 추정량, Y_{tot_R} , 은

$$Y_{tot_R} = f\,\overline{Y}_a + (1-f)\,Y_a, \quad f = \frac{n_a}{N_a}. \tag{2.5}$$

이며, $\overline{Y}$: 관측된 실업자수

Y_a : 추정된 소지역의 실업자수:

여기서 추정된 소지역의 실업자수, Y_a는 다음과 같이 얻는다.

$$Y_a = X_a\beta_a, \quad \beta_a = (X_a'X_a)^{-1}X_aY_a ,$$

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

$$Bias = \frac{1}{R}\sum_{r=1}^{R}(Y_{ta} - Y_{ta})$$

$$Mse = \frac{1}{R}\sum_{r=1}^{R}(Y_{ta} - Y_{ta})^2 \quad , R\text{은 반복실험횟수}$$

3. 모의실험

경기도의 경제활동인구조사 조사된 자료를 모집단으로 하였으며 경기도 23개시 각각의 전체 실업자 수의 실제자료 값은 다음과 같다.

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

그리고 모집단을 행정구역 (23개의 시)순으로 나열한 후 모집단의 약 10%인 274개의 가구를 계통추출 하였으며 이를 1000번 반복실험 하였다.

2장에서 제시한 2가지 방법(K-F, 회귀모형)에 의한 결과는 다음과 같다.

시	Y_{tot_R}	Y_{tot}	$Bias(Y_{tot_R})$	$Bias(Y_{tot})$	$Mse(Y_{tot_R})$	$Mse(Y_{tot})$
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	2.09
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

* Y_{tot_R} : 단순회귀모델

  Y_{tot} : K-F 모델

위의 표로부터 단순회귀모형에서 K-F모형으로 오차항을 보정해 줌으로써 대부분의 경우 편의(Bias)는 물론 평균제곱오차(Mse)가 크게 줄었음을 볼 수 있다.

4. 토의

소지역 추정을 할 때에 센서스-추정치들은 여러 가지 오류에 영향을 받으므로 KF 추정치는 사후-센서스 추정치로서 이러한 오류를 제거한다. 그러나 센서스 추정치와 KF 추정치 간의 현저한 차이가 있을 때에는 에디팅(editing)을 하거나, 비 표본 오차 등을 조사하여야 한다. 그리고 β_{ta}들도 유사한 소지역으로 묶어서 추정함으로서 소지역 추정의 효율성을 높일 수 있다.

참고문헌

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