제2장

모형 기반 소지역 실업자 수 추정

제1절 서 론

1. 연구배경

소지역 추정(small area estimation)은 공표되고 있는 통계표보다도 더욱 세세한 항목을 대상으로 한 추정을 말한다. 소지역 통계는 정책 결정을 위하여 지역의 경제상황과 발전수준을 파악하거나 사업을 위한 의사결정 자료로 유용하며, 2000년 이전부터 그 수요가 증가해 왔다. 그러나 인구주택총조사와 같은 센서스를 제외하면 표본조사에서는 소지역별로 제공되는 통계는 거의 없다. 왜냐하면 표본조사가 전국 또는 지역(시도)별 결과를 공표할 수 있는 규모로 설계된 경우가 많기 때문에, 이보다 작은 소지역은 결과에 대한 정도(precision)를 확보할 수 있을 정도의 표본을 확보하기가 어렵기 때문이다. 이처럼 표본조사 자료만으로 소지역에 대한 관심값을 추정하면 할당된 표본 수가 부족하기 때문에 표집오차가 커지게 된다. 이 때문에 표본규모를 확대하지 않고, 표집오차의 증가를 줄이면서 소지역 추정에 대한 정도를 높일 수 있는 통계적 방법에 관한 연구가 수없이 행해져 왔고, 지금도 활발히 연구가 계속되고 있다. 소지역 추정 방법에 대한 자세한 이론적 내용은 Rao(2003)를 참고할 수 있다.

고용통계를 포함하여 국가별 소지역 추정사례를 보면 다음과 같다. 먼저, 미국은 오래 전부터 주(state)별 노동시장의 불균형을 파악하기 위해서 소지역 통계를 작성하고 있다. 영국은 지역수준에서 범죄, 실업, 교

육 등의 문제를 해소하기 위해 소지역 통계의 중요성을 인지하고 연구를 진행하고 있다. 호주는 전국 기준의 장애인 통계조사를 이용하여 각 주 단위의 장애인 수를 추정하고 이를 복지 예산 편성의 기초 자료로 이용하고 있다. 호주의 경우 실업통계에 대해서는 2008년 말경에 소지역 추정 결과를 발표할 예정에 있다(담당자 E-mail). 뿐만 아니라 범국가적 학술모임이나 국제회의장에서도 소지역 추정에 대한 논의가 적극적으로 이루어지면서 국제적인 관심도 높아지고 있다.

2. 연구범위 및 내용

소지역 추정방법은 일반적으로 합성추정량(synthetic estimator)과 복합추정량(composite estimation)을 포함하는 설계 기반(design-based) 추정방법과 경험적 또는 계층적 베이즈 추정량을 포함하는 모형 기반 추정(model-based)방법으로 구분할 수 있다. 이 중에서 본 연구는 모형 기반 추정방법에 초점을 둔다. 특히, 한국의 실업통계에 적용할 만하다고 판단되는 모형 기반 방법으로 Fay-Herriot 모형, Time series & Cross-sectional 모형(Rao-Yu 모형), 시계열회귀 모형과 로지스틱회귀 모형(Logistic regression model)에 관한 내용을 개괄적으로 소개한다. 이 모든 방법에 대해서 이론적 측면과 해외 적용 사례를 충분히 검토해 본 결과, 한국의 실업통계에 적용 가능성이 높다고 판단되는 Fay-Herriot 모형과 Rao-Yu 모형을 우선적으로 고려한다. 나머지 모형에 대한 자세한 검토와 적용은 향후 과제로 남긴다.

추정량은 EBLUP(Empirical Best Linear Unbiased Predictor)과 계층적 베이즈(Hierarchical Bayes: HB) 추정량을 사용한다. 선행 연구들에 따르면 이론적 특성에 비추어 볼 때 추정량의 효율성 측면에서는 HB가 EBLUP보다 일반적으로 더 우수한 추정량으로 알려져 있다(You 등, 2003). 그러나 일반적으로 어떤 추정량이 우수하다고 단언할 수는 없다. 왜냐하면 추정량은 적용되는 자료의 특성이나 조건에 따라 수행능력이 다를 수 있기 때문이다. 그러므로 우리나라 실업통계 자료에서의 두 추정량의 효율성을 확인할 필요가 있다. 이를 위해 모형이 보다 간단한 Fay-Herriot 모형에 두 추정량을 적용하고 그 결과를 비교한다.

한편, 고용통계 소지역별 추정의 근본적인 목적은 한국의 전체 230여개 시군구(소지역을 행정구역으로 정의할 경우)에 대해서 추정통계를 만들어내는 것일 것이다. 즉, 본 연구에서 선택한 2개 모형을 경제활동인구조사에 적용해서 모든 시군구별 실업자 수(또는 실업률)를 추정해 보아야 한다. 현재 경제활동인구조사는 전국과 16개 시도단위에 대해 통계를 공표하고 있고, 이 때 지역단위의 벤치마크 인구로는 추계인구를 사용하고 있다. 실제로 시군구 단위의 공식적인 추계인구는 없다. 이와 관련해서 연구의 대상 소지역을 31개 시군구로 제한한 것은 2가지 이유가 있다. 첫째, 31개 대상지역들은 해당 지자체에서 지역통계를 생산하고 있고, 통계청에서도 이들 지역에 대한 추계인구를 제공하고 있기 때문이다. 둘째, 소지역별로 추정한 결과를 비교할 만한 대상(모수)이 없기 때문에 표본규모가 큰 이들 지역 통계조사를 비교 대상으로 삼고자 하는 것이다. 결론적으로, 본 연구의 가장 큰 목적은 다양한 모형 기반 추정방법을 31개 시군구 지역의 경제활동인구조사 자료에 적용해 봄으로써 향후 통계청의 고용통계 시군구별 소지역 추정 방법으로써의 적용 가능성을 모색하는 것이다.

본 연구는 7개의 절로 구성되었다. 먼저 2절에서 모형을 이용한 추정방법에 대해서 간략하게 소개한다. 3절과 4절은 Fay-Herriot 모형과 Rao-Yu 모형의 이론적 측면을 설명하고, 5절은 경제활동인구조사 자료와 모형에서 보조정보로 사용될 고용보험의 수급자 정보에 대해서 설명한다. 6절에서는 다양한 추정방법을 경제활동인구조사에 적용해 보고, 여러 가지 평가방법에 의해 각각의 추정결과를 비교・분석한다. 마지막으로 연구결과와 논의사항을 언급한다.

제2절 다양한 소지역 추정방법 소개

본 절에서는 학술연구를 포함하여 많은 국가에서 소지역 실업자 수(실업률) 추정에 적용되고 있는 방법에 대해서 각 방법의 특성을 위주로 간단히 소개한다.

1. Fay-Herriot 모형을 이용한 EBLUP

EBLUP 추정량은 영국에서 검토한 바 있는 방법이다. 이 방법에서는 지역(공간) 정보가 추정에 이용된다. Fay-Herriot 모형은 단순한 회귀모형과 다르게 소지역별 효과도 포함된 모형으로, 소지역의 지역적 특성이 추정에 이용될 수 있다. EBLUP 추정량은 자료가 내포하고 있는 구조는 회귀직선으로 관측치가 다음과 같은 Fay-Herriot 모형을 따른다고 가정한다. 이 모형에서는 설명변수 항, $x_i\beta+\nu_i$가 회귀직선상에 있지 않고, 지역 효과 ν_i 만큼 직선에서 떨어져 있다는 것을 전제로 한다.

$$y_i = x_i\beta+\nu_i+e_i,\ e_i \sim N(0,\psi_i),\ \nu_i \sim N(0,\sigma_\nu^2)$$

이때, i는 소지역, y_i는 관측치(실업자 수), x_i는 설명변수(보조정보), ν_i는 지역만의 효과, e_i는 표집오차를 나타낸다. 표집오차 e_i의 분산 ψ_i는 알려져 있다고 가정한다. 실제로는 ψ_i 대신에 관측 자료로부터 계산한 추정량 $\widehat{\psi_i}$을 이용한다. 지역효과 ν_i의 분산 σ_ν^2은 각 지역에서 동일하다고 가정한다. 이 모형을 실제 우리나라 경제활동인구조사 자료에 적용해보았다. 이 방법에 대한 자세한 내용은 3절의 추정방법을 참조하기 바란다.

이 방법은 로지스틱회귀 모형과 그 특성이 유사하고, 회귀 적합이 좋지 않으면 효과적이지 않은 것으로 알려져 있다. 그러나 대규모 표본조사, 즉, 센서스 등의 보조정보를 이용할 수 있는 경우에는 적합한 방법으로 생각된다.

2. Time series & Cross-sectional 모형

이 방법은 캐나다 통계국에서 연구된 것으로, Fay-Herriot 모형을 다변량으로 확장한 Rao-Yu 모형을 이용한다(Rao와 Yu, 1994; You 등, 2003). Rao-Yu 모형은 Time series & Cross-sectional 모형 중 하나로, 이 모형에서는 시계열과 공간 정보를 모두 추정에 이용할 수 있다. 참고로 이런 유형의 Time series & Cross-sectional 모형은 다양하다. 미국에서는 Rao-Yu 모형과는 조금 다르게, 시점에 대한 분산추정 시 랜덤워크(random walk)모형을 적용하여 Time series & Cross-sectional 모형으로 추정한 바 있다(Datta 등, 1999). Rao-Yu 모형은 회귀 모형에 ① 지역만의 특색을 반영하는 확률변수 ν_i, ② AR(1) 모형을 따르면서 랜덤 시계열 효과를 나타내는 확률변수 u가 더해진 모형이다. 모형에 대한 구체적인 식은 4절을 참고하기 바란다.

이 모형은 4절에서 설명되는 바와 같이 상당히 복잡한 추정 과정을 거치지만 추정의 효율을 높이기 위해서 계층적 베이즈 모형(Hierarchical Bayes Model)과 마코프 체인 몬테 칼로 방법(Markov Chain Monte Carlo: MCMC)이 적용된다. 이를 위해 모수 추정을 위한 다양한 사전 분포를 가정하고, 이로부터 사후분포를 구한다. 모수 추정은 MCMC방법의 일종인 깁스 표본자(Gibbs sampler)를 이용하고, 이때 L개의 체인에 대해 평균을 취하는 방법을 이용한다. 이렇게 함으로써 초기값에 의존적이지 않은 안정적인 추정을 할 수 있다. 오차의 분산공분행렬 Σ_i는 오차의 자기상관을 기반으로 하여 추정한다. 관심모수 추정에는 라오-블랙웰 추정량을 이용한다. 이에 의해 초기치의 영향을 줄임으로써 시뮬레이션 오차를 줄일 수 있을 것으로 기대할 수 있다.

3. 로지스틱회귀 모형

영국 통계국은 실업률(실업자) 등의 비율자료(이산형자료)를 추정하기 때문에 로지스틱 변환을 이용한 로지스틱회귀 모형을 이용하고 있다. 이때 미국이 지역단위의 시계열 정보를 이용하는 대신에 영국은 공간(지역)단위의 정보로부터 추정하고 있다. 영국의 경우 여러 차례에 걸

쳐 추정방법을 개선하였고, 2006년에 현재의 추정방법에 대한 결과를 공식적으로 발표하여 사용하고 있다(ONS, 2006). 영국은 모형에서의 핵심변수로 매년 각 소지역(LAD/UA)들 내에서 각 연령별/성별 그룹(남성/여성, 16-24세, 25-49세, 50세 이상)의 16세 이상 인구의 비율을 노동력통계(LFS) 자료에 다음과 같이 적용하고 있다.

$$logit(p_{ij}) = \beta_0 + \beta_1 x_j + \beta_2 D_A + U_j(area\ random\ effect\ term\ _j)$$

여기서, p_{ij}는 성별/연령별 그룹 i내의 소지역 j의 실업률, x_j는 j지역 설명변수(j지역의 $logit$(실업보험수급율)), D_A: 더미변수, U_j는 보조 정보에 의해 설명되지 않는 지역 변이를 나타낸다. U_j항을 모형에 포함시키는 것은 보조자료에 의해 설명되지 않는 지역 변이항이 모형에 있을 때와 없을 때 간의 차이를 설명하기 위함이다. 추정량은 표본 크기가 증가함에 따라 직접 추정량에 많은 가중이 더해지는, 직접추정량과 고정효과추정량의 가중 평균 형태를 취한다. 이것은 표본이 커지면 직접추정량으로의 수렴을 보장하므로, 직접추정치가 모형에 의한 추정치보다 더 정도가 높을 때 어떤 추정치를 선택해야 하는지 쉽게 결정하도록 도와준다.

로지스틱회귀 모형식은 분산동일성을 만족하지 않기 때문에 j지역에서 가중치 w_j를 이용한 가중회귀를 수행한다. 이 방법은 작은 지역의 지역정보를 얻을 수 있을 때에는 불완전하게나마 이용될 수 있지만, 정도 높은 추정은 기대하기 어렵다. 영국에서는 보다 작은 지역단위 수준의 자료를 이용해서, 즉, 성별, 지역별로 사후층화를 하고 추가로 이 가중치를 이용해서 추정치를 구하고 있다. 반대로 소지역에 대해 설명변수를 구할 수 없는 경우에는 회귀 적합이 좋지 않고 그다지 유효하지 않은 것으로 알려져 있다. 이에 대해 영국은 앞에서 언급한 EBLUP 추정량에 대해서도 검토한 바 있지만 계산식이 복잡하고 그 효과에도 의문이 제기되어 지금의 로지스틱회귀 모형을 사용하고 있는 것으로 보인다.

호주에서도 2008년 말경에 로지스틱회귀 모형을 이용한 실업통계 개발완료를 예정하였으나, 2009년 1월 현재 연구 결과 검토 및 보고서 작

성 등을 진행 중에 있다(담당자의 E-mail).

4. 시계열 모형

미국과 일본에서 채택하고 있는 방법이 시계열회귀 모형이다. 이것은 회귀항, 추세항, 계절변동항, 불규칙항, 표집오차항으로 구성된 모형을 가정한다. 이 모형에 칼만-필터(Kalman and Filter)를 이용해서 표집오차항을 추정하고 관측치로부터 이 표집오차항을 제거하여 추정치를 구하는 방식이다. 시계열회귀 모형은 매우 복잡한 추정식을 요한다. 여기서는 간략하게 주요 단계에 대해서만 설명하기로 하고, 자세한 내용은 통계개발원의 해외출장결과보고서(2008)를 참고하기 바란다. 전체적으로 시계열 모형은 시그널과 노이즈를 합한 형태로 시그널 부분과 노이즈에 대한 추정이 이루어진다.

가. 시그널(Signal) + 노이즈(Noise) 모형

소지역별 관측 자료는 시그널과 노이즈라는 독립적인 두 확률과정의 합으로 표현된다.

$$y(t) = \theta(t) + e(t), \qquad t = 1, \cdots, T$$

여기서, $y(t)$는 관측치, $\theta(t)$는 참값에 대한 추정치(시그널), $e(t)$는 표집오차에 대한 추정치(노이즈)이다. 결국 $y(t)$에서 $e(t)$를 빼는 것이 이 방법의 목적이다.

참값, $\theta(t)$는 다음과 같이 네 개의 항으로 분해할 수 있다고 가정한다.

$$\theta(t) = M(t) + T(t) + S(t) + I(t)$$

여기서, $M(t)$는 회귀항, $T(t)$는 추세항, $S(t)$는 계절항, $I(t)$는 불규칙 변동항을 나타내고, 이들 각각이 따르는 모형은 다음과 같다. 먼저, 회귀항 $M(t)$는 $\theta(t)$가운데 보조정보에 의해서 설명되는 부분이다.

$$M(t) = X(t)\beta(t), \quad \beta(t) = \beta(t-1) + v_\beta(t)$$

여기서, $X(t)$는 보조정보(설명변수)를 나타내고, $v_\beta(t) \sim N(0, \sigma_\beta^2)$을 가정한다. 추세항 $T(t)$는 $\theta(t)$ 중에 완만한 변동을 찾아내기 위해 삽입되

며, 각 시점에서 수준을 나타내는 부분 $T(t)$와 기울기를 나타내는 부분 $R(t)$를 이용해서 다음과 같이 표현된다.

$$T(t) = T(t-1) + R(t-1) + v_T(t),$$
$$R(t) = R(t-1) + v_R(t)$$

여기서, $v_T(t) \sim N(0, \sigma_T^2)$, $v_R(t) \sim N(0, \sigma_R^2)$이다. $\sigma_R^2 = \sigma_T^2 = 0$인 경우에는 기울기 $R(t)$가 일정한 1차 다항식(직선)을 추세로 가정하는 셈이 된다. $\sigma_R^2 > 0$ 또는 $\sigma_T^2 > 0$ 인 경우에는 기울기 및 수준이 각각 독립적으로 랜덤워크(random walk)에 따라서 변화하는 유연한 모형이 된다. 회귀항에 의해서 추세를 충분히 설명할 수 있을 경우에는 추세항이 필요 없게 되고 자동적으로 모형에서 빠진다. 이 항은 설명변수로 설명되고 남은 잔차 추세(residual trend)를 끄집어내기 위해 도입된 것이다. 계절항 $S(t)$는 시그널 가운데 계절적인 변동을 추출하기 위해 삽입된다. 계절적인 변동으로는 시계열 성분 가운데 주기가 일년인 것을 가리킨다. 계절항은 계절주파수(12개월 주기, 6개월 주기, 4개월 주기, 3개월 주기, 2.4개월 주기, 2개월 주기)에 대응하는 6개 삼각함수의 합으로써 다음과 같이 표현된다.

$$S(t) = \sum_{j=1}^{t} S_j(t)$$
$$S_j(t) = \cos(\omega_j) S_j(t-1) + \sin(\omega_j) S_j^*(t-1) + v_{S_j}(t),$$
$$S_j^*(t) = -\sin(\omega_j) S_j(t-1) + \cos(\omega_j) S_j^*(t-1) + v_{S_j^*}(t),$$
$$v_{S_j}(t) \sim N(0, \sigma_S^2),\ v_{S_j^*}(t) \sim N(0, \sigma_S^2),\ \omega_j = \frac{2\pi j}{12}$$

$\nu_{S_j}(t)$, $\nu_{S_j^*}(t)$는 모두 상호 독립으로 분산이 동일한(σ_S^2) 정규분포를 따른다. 회귀항에 의해서 계절성을 충분히 설명할 수 있는 경우에는 이 항이 필요 없게 되고 자동적으로 모형에서 빠진다. 이 항은 설명변수로 설명되고 남은 잔차 계절성(residual seasonality)을 끄집어내기 위해 도입된다. 불규칙 변동항 $I(t)$는 위의 각 항에 포함되지 못하고 남은 변동을 나타낸다. 여기서는 경기의 단기적인 변동 등이 들어갈 가능성이 있다.

$$I(t) = v_I(t), \quad v_I(t) \sim NID(0, \sigma_I^2)$$

이때 $\theta(t)$가 $M(t), T(t), S(t)$에 의해서 거의 설명될 경우에는 이 항은

필요 없게 되고 자동적으로 모형에서 빠진다.

마지막으로 표집오차항 $e(t)$모형은 미국 소지역 추정 모형 가운데 가장 중요한 부분이다. 표본에 연동구조가 있는 경우 표집오차에 시계열 상관이 있을 것으로 예상된다. 또 표집오차 분산은 표본크기, 표본설계 변경, 참값 $\theta(t)$의 영향을 받아서 시간에 따라 변한다고 볼 수 있다.

$$e(t) = \gamma(t)e^{*}(t)$$

여기서, $\gamma(t)$는 표집오차에 대한 분산 변화를 나타내는 스칼라 값으로 $e(t)$ 변동의 진폭을 나타낸다. $e^{*}(t)$는 연동구조에 의한 시계열 상관을 나타내는, 즉, 표집오차항 $e(t)$의 변동 패턴을 나타내는 AR(13)모형이다. 이들 움직임이 결합해서 표집오차 $e(t)$가 된다. $e^{*}(t)$와 $\gamma(t)$ 추정 방법에 대해서 다음과 같이 설명한다.

나. 상태공간모형

앞에서 설명한 시그널 + 노이즈 모형은 다음과 같이 상태공간모형으로 표현할 수 있다.

$$y(t) = \boldsymbol{H}(t)\boldsymbol{\alpha}(t) + w(t),\ w(t) \sim N(0,\sigma_I^2)\ :\ \text{관측방정식}$$
$$\boldsymbol{\alpha}(t) = \boldsymbol{F}\boldsymbol{\alpha}(t-1) + \boldsymbol{V}(t),\ \boldsymbol{V}(t) \sim N(0,\boldsymbol{Q})\ :\ \text{전이방정식}$$

여기서, $\boldsymbol{F}$는 전이행렬, $\boldsymbol{\alpha}(t)$는 상태변수 벡터, $\boldsymbol{H}(t)$는 관측행렬, $w(t)$는 관측 노이즈, $V(t)$는 시스템 노이즈를 나타낸다.

초기상태 $\boldsymbol{\alpha}(0)$는 정규분포를 따르고, $\boldsymbol{F}$, $\boldsymbol{H}(t)$는 모두 알려져 있다고 가정한다. $\boldsymbol{Q}$와 σ_I^2에 대해서는 초기상태에서는 알려져 있다고 가정하고, 나중에 추정한다. 상태변수 $\boldsymbol{\alpha}(t)$는 직접 관측할 수 없고 위의 관측방정식을 통해서 $w(t)$가 부가되어 실제 관측치로 관측된다. 또 $\boldsymbol{\alpha}(t)$는 전이방정식에 의해서 시간과 함께 변한다. 여기에 칼만-필터를 적용해서 각 확률과정으로 분해하고 표집오차항을 추정한다. 관측치로부터 추정한 표집오차항을 제거한 값이 구하고자 하는 추정치가 된다.

일본은 미국의 모형을 자국의 현실에 맞게 약간 수정하여 사용하고 있다. 시계열 방법은 지역단위로 모형을 설정하기 때문에 각각의 지역 특성을 반영시킬 수 있고, 시계열자료를 추정하는 데에는 유용한 방법이라 할 수 있다. 그러나 시계열 정보 외에 이웃지역 간의 정보를 모형

설정에 전혀 사용할 수 없다는 단점이 있다. 미국에서도 이런 BLS의 방법의 문제점을 지적하고 Datta 등(1999)이 Time series & Cross-sectional 모형을 제안한 바 있다.

제3절 Fay-Herriot 모형과 EBLUP 추정량

본 절에서는 EBLUP 추정량의 이론에 대해서 Rao(2003)의 5장부터 7장을 적절하게 인용해서 설명한다. 추정식의 도출 등 이론적으로 상세한 내용은 생략하기로 한다. 상세한 내용은 Rao(2003)의 해당 부분을 참고하기 바란다. 본 연구에서는 Fay-Herriot 모형에 EBLUP과 HB 추정량을 적용하고자 한다. Fay-Herriot 모형을 이용한 HB에 대한 이론적 설명은 김달호(2005)와 통계청 용역보고서(2004) 등에 자세히 설명되어 있다. 그리고 HB에 대한 이론적 설명은 4절에서 Times series & Cross-sectional 모형과 함께 설명하기로 한다.

소지역별로 조금이라도 특성이 있는 자료를 다루는 경우를 생각해보면, 단순한 선형회귀 모형에서는 보통 더미변수(dummy variable)를 삽입하는 것으로 이와 같은 문제에 대처한다. 그러나 이런 경우 각 소지역별의 특성을 모형에 반영하기가 어렵고, 게다가 자유도가 낮아지는 문제도 발생한다. 선형혼합효과 모형(Linear mixed effect model)을 기본으로 한 BLUP(Best Linear Unbiased Predictor)추정량은 이와 같은 문제점에 대처할 수 있고 추정의 정도에 대한 향상을 기대할 수 있다.

BLUP 추정량은 원래 축산 관련 분야에서 주로 이용되어진 방법이다. 실제 응용 예로서는 미국의 1세대당 수입에 대한 중위값(median) 분석이나 캐나다 센서스의 조사 누락율 추정 등이 있다. 이런 예들에 대해서는 Rao(2003)에 자세하게 설명되어 있다. EBLUP 추정량에 대해서는 현재도 활발한 연구가 진행되고 있다.

1. Fay-Herriot 모형

전체 m개의 소지역이 있고 지역수준(area level)의 보조정보를 이용할 수 있다고 가정하자. 이때 자료구조는 회귀모형을 고려할 수 있고, 관측치는 다음과 같은 모형을 따른다고 가정한다.

$$y_i = \mathbf{x}_i^t\boldsymbol{\beta} + b_i\nu_i + e_i,\ i = 1,2,\cdots,m \tag{2.1}$$

여기서 y_i : 소지역 i의 관측치

$\mathbf{x_i}$: 소지역 i의 보조정보 벡터(상수항을 포함)

b_i : 미지의 상수, $e_i \sim N(0,\psi_i)$, $\nu_i \sim N(0,\sigma_\nu^2)$이다.

e_i 및 ν_i는 모두 독립이고 σ_ν^2 및 ψ_i는 각각 알려져 있다고 가정한다. 각 소지역 i에 대해서 y_i는 독립이라고 가정한다. 식(2.1)은 이 모형을 처음으로 개발한 사람의 이름을 붙여서 Fay-Herriot 모형이라 부른다(Fay와 Herriot, 1979). 식(2.1)은 다음과 같은 두개의 모형을 조합한 것이다.

$$\theta_i = \mathbf{x}_i^t\boldsymbol{\beta} + b_i\nu_i\ ,\ i = 1,2,\cdots,m \tag{2.2}$$

$$y_i = \theta_i + e_i\ ,\ i = 1,2,\cdots,m \tag{2.3}$$

이때 θ_i는 각 소지역의 참값을 나타내고, 식(2.2)는 θ_i가 선형회귀모형을 따른다는 것을 나타내고 있다. 식(2.3)은 참값 θ_i에 표집오차 e_i가 더해져서 y_i로 관측된다는 것을 나타낸다.

일반적인 선형회귀 모형에는 식(2.2)의 $b_i\nu_i$항이 없고, 참값 θ_i가 회귀직선 $\mathbf{x}_i^t\boldsymbol{\beta}$에 완전히 겹친다고 가정한다. 그러나 이와 같은 가정은 실제 자료에서 현실적이지 못한 경우가 종종 발생한다. Fay-Herriot 모형은 이처럼 선형회귀모형과 다르고, 참값 θ_i에 대해서 회귀직선으로부터 $b_i\nu_i$ 만큼의 폭을 갖는 모형을 고려한다. 이는 이렇게 함으로써 지역의 특성을 반영한 보다 현실적인 추정이 가능할 것이라는 생각에서 비롯된 것이다. 확률변수가 여러 개인 이와 같은 선형모형을 일반적으로 선형혼합효과모형이라 부른다. Fay-Herriot 모형은 이런 선형혼합효과모형의 특수한 형태이다.

e_i와 ν_i는 모두 평균이 0인 확률변수이다. 그러나 이들 두 변수의 분산에 관한 가정에는 다음과 같이 큰 차이가 있다.

◦ ν_i는 지역 i에 의존하지 않는 일정한 분산 σ_ν^2을 갖는다고 가정한다. 이로부터 소지역 간의 효과에 대한 산포(dispersion)는 ν_i보다는

알려진 상수 b_i라는 외적요인에 의해 반영되고 있다. 본 연구에서는 $b_i = 1$로 가정한다. 그러나 분석자가 지역별 변동 ν_i에 관해서 미리 정보를 가지고 있다면 그것을 기초로 b_i값을 정할 수도 있다.

◦ e_i의 분산 ψ_i에는 첨자 i가 붙어 있다. 이는 σ_v^2와 달리 표집오차 e_i의 분산이 소지역에 따라 다르다는 것을 허용한다. 일반적으로 표집오차 분산 ψ_i는 알려져 있다고 가정하지만 실제로는 모르기 때문에 ψ_i 대신 자료로부터 추정한 $\hat{\psi}_i$을 이용한다. 본 연구에서 $\hat{\psi}_i$의 추정은 잭나이프 방법(jackknife method)을 사용하였다.

2. BLUP 추정량

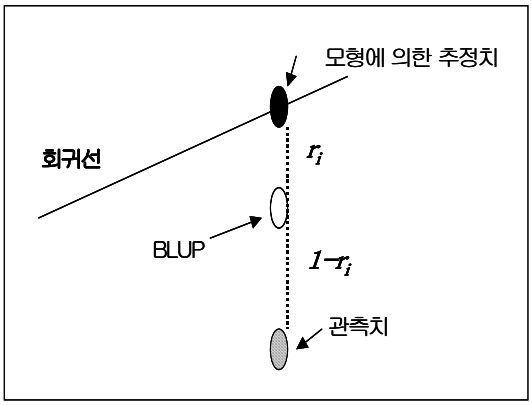
BLUP이란 θ_i의 예측량 $\hat{\theta}_i = \mathbf{x}_i^t\hat{\boldsymbol{\beta}} + \hat{\nu}_i$을 선형불편이라고 할 때 평균제곱오차가 최소가 되는 값을 말한다. Henderson(1950)은 보다 일반적인 선형혼합모형에 대한 BLUP을 도출하였다. Henderson의 이론을 식(2.1)의 모형에 대입하면 다음과 같은 Fay-Herriot 모형의 BLUP 추정량 θ_i^{BLUP}을 구할 수 있다.

$$\theta_i^{BLUP} = \gamma_i\hat{y}_i + (1-\gamma_i)\mathbf{x}_i^t\hat{\boldsymbol{\beta}} \ , \tag{2.4}$$

$$\gamma_i = \frac{b_i^2\sigma_\nu^2}{\psi_i + b_i^2\sigma_\nu^2},$$

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{m}\frac{\mathbf{x}_i^t\mathbf{x}_i}{\psi_i + b_i^2\sigma_\nu^2}\right]^{-1}\left[\sum_{i=1}^{m}\frac{\mathbf{x}_i^t y_i}{\psi_i + b_i^2\sigma_\nu^2}\right].$$

이때 회귀계수 $\boldsymbol{\beta}$의 추정량 $\hat{\boldsymbol{\beta}}$은 확률변수의 분산을 고려한 경우의 일반화최소제곱추정량(GLS)이 된다. θ_i^{BLUP}은 관측치 y_i와 단순한 회귀모형의 예측치 $\mathbf{x}_i^t\boldsymbol{\beta}$를 가중치 γ_i로 가중평균한 형태이다. 즉 BLUP 추정량은 회귀직선과 관측치의 $\gamma_i : (1-\gamma_i)$ 내분점에 해당한다. 따라서 앞에서도 언급한 바와 같이 참값은 회귀선 위에 완전히 겹친다고 보지 않고, 회귀선과 관측치 사이에 있다고 가정하게 된 것이다([그림 2-1]).



[그림 2-1] BLUP 추정량의 의미

이때 가중평균의 가중치 γ_i는 e_i와 ν_i의 분산비에 의해 결정된다. γ_i의 형태에서 다음과 같은 사실을 알 수 있다.

- 관측치의 표집오차 ψ_i가 커짐에 따라서($\psi_i \to \infty$) γ_i는 0에 가까워지고 모형으로부터 계산된 예측치에 걸리는 가중치가 커지게 된다.
- 반대로 지역 효과의 변동 σ_ν^2이 커짐에 따라서($\sigma_\nu^2 \to \infty$) γ_i는 1에 가까워지고 관측치에 걸리는 가중치가 커지게 된다.
- $\psi_i = 0$일 때 가중평균의 가중치 γ_i는 σ_ν^2 값에 상관없이 1이 되고, θ_i^{BLUP}은 관측치 y_i에 거의 일치한다.

이에 의해 θ_i^{BLUP}은 관측치를 신뢰할 수 있는 경우에는 관측치의 가중치를 크게 하고, 그렇지 않은 경우는 그 정도에 따라서 적당한 비율로 모형 정보를 활용한다. 그럼으로써 관측치의 오차에 대응한 상당히 유연한 추정치가 된다. 그리고 모집단 전부를 조사한 경우에는($\psi_i = 0$), θ_i^{BLUP}은 모집단 값에 일치한다.

3. EBLUP 추정량

3.2절에서는 σ_ν^2이 알려져 있다고 가정하고 선형불편 추정량 가운데 최량(최소 분산)인 예측량 θ_i^{BLUP}을 구했다. 그러나 실제로는 σ_ν^2의 값은 알려져 있지 않으므로, 그 값을 관측치와 자료로부터 추정할 필요가 있

다. 이처럼 BLUP 추정량 θ_i^{BLUP}에서 σ_ν^2을 관측치 자료에 의한 추정치 $\widehat{\sigma_\nu^2}$으로 바꾸면, 이것을 EBLUP 추정량이라 부른다. Fay와 Herriot(1979)은 단순 적률법과 간략화한 뉴턴법(newton method)을 조합해서 $\widehat{\sigma_\nu^2}$를 추정했다. 적률법은 사용 방법이 간단하지만 추정효과는 다소 떨어지는 것으로 알려져 있다. 이와 함께 자주 사용되는 방법 중 하나가 최우추정법이다. 최우추정(maximum likelihood estimation)법은 표본크기가 충분한 경우 효과적인 방법으로 널리 사용되고 있지만, 표본이 작을 경우나 모형적합이 좋지 않을 경우 최대값으로의 수렴에 대한 보장이 어렵다는 단점이 있다.

본 연구에서는 최우추정법을 이용하여 $\widehat{\sigma_\nu^2}$을 추정하였다. 최우추정법에서 σ_ν^2의 추정치는 대수우도함수 $l(\widehat{\sigma_\nu^2})$가 최대값이 될 때의 $\widehat{\sigma_\nu^2}$이 된다. 대수우도함수 $l(\widehat{\sigma_\nu^2})$의 구체적인 형태는 다음과 같다.

$$
\begin{aligned}
l(\widehat{\sigma_\nu^2}) &= \log\left\{ \Pi_{i=1}^{m} \frac{1}{\sqrt{2\pi(b_i^2\widehat{\sigma_\nu^2}+\psi_i)}} exp\left[-\frac{1}{2}\frac{(y_i-\mathbf{x}_i^t\boldsymbol{\beta}(\widehat{\sigma_\nu^2}))^2}{b_i^2\widehat{\sigma_\nu^2}+\psi_i} \right] \right\} \\
&= -\frac{m}{2}log2\pi - \frac{1}{2}\sum_{i=1}^{m}\log(b_i^2\widehat{\sigma_\nu^2}+\psi_i) - \frac{1}{2}\sum_{i=1}^{m}\frac{(y_i-\mathbf{x}_i^t\boldsymbol{\beta}(\widehat{\sigma_\nu^2}))^2}{b_i^2\widehat{\sigma_\nu^2}+\psi_i} \qquad (2.5) \\
&\left[\boldsymbol{\beta}(\widehat{\sigma_\nu^2}) = \left[\sum_{i=1}^{m}\frac{\mathbf{x}_i^t\mathbf{x}_i}{\psi_i+b_i^2\widehat{\sigma_\nu^2}} \right]^{-1} \left[\sum_{i=1}^{m}\frac{\mathbf{x}_i^t y_i}{\psi_i+b_i^2\widehat{\sigma_\nu^2}} \right] \right]
\end{aligned}
$$

식(2.5)는 복잡한 비선형 방정식이기 때문에 최소값을 해석적으로 계산하는 것은 불가능하다. 이 경우 수치계산에 의한 비선형최적화 방법을 이용해서 $\widehat{\sigma_\nu^2}$를 추정하면 계산이 편리하다.

함수의 최소화를 위한 수치계산방법은 다양하다. 그 중에서 비선형최적화 방법으로서 스코어(score)법을 소개한다. 이것은 뉴톤-랩슨 방법에서 기울기를 표현하는 Hessian 행렬을 그 기댓값으로 변환한 방법이다. 이 방법은 기댓값을 얻는데 피미분함수가 단순한 형태가 되어 계산이 간단해진다는 장점이 있다. 이 경우, Hessian 행렬의 기댓값은 Fisher의 정보행렬과 같아진다. 이 방법은 Fisher에 의해 처음 사용되었기 때문에 Fisher의 스코어법이라고도 한다. 식(2.5)를 최소화하기 위한 스코어법의 알고리즘은 다음과 같다. 자세한 내용은 Rao(2003)를 참고하기 바란다.

$$\sigma_\nu^{2(a+1)} = \sigma_\nu^{2(a)} + [I(\sigma_\nu^{2(a)})]^{-1} s(\boldsymbol{\beta}^{(a)}, \sigma_\nu^{2(a)}) \tag{2.6}$$

$$\begin{pmatrix} I(\sigma_\nu^{2(a)}) = \frac{1}{2}\sum_{i=1}^{m} \frac{1}{(\psi_i + \sigma_\nu^2)^2} \quad : \quad Fisher\text{의 정보량} \\ s(\boldsymbol{\beta}^{(a)}, \sigma_\nu^{2(a)}) = -\frac{1}{2}\sum_{i=1}^{m} \frac{1}{\psi_i + \sigma_\nu^2} + \frac{1}{2}\sum_{i=1}^{m} \frac{(\hat{\theta}_i - \mathbf{x}_i^t \tilde{\beta})^2}{(\psi_i + \sigma_\nu^2)^2} \\ \boldsymbol{\beta}^{(a)}(\sigma_\nu^{2(a)}) = \left[\sum_{i=1}^{m} \frac{\mathbf{x}_i^t \mathbf{x}_i}{\psi_i + b_i^2 \sigma_\nu^{2(a)}}\right]^{-1} \left[\sum_{i=1}^{m} \frac{\mathbf{x}_i^t y_i}{\psi_i + b_i^2 \sigma_\nu^{2(a)}}\right] \end{pmatrix} \tag{2.7}$$

위의 식 (2.6)과 (2.7)에서 (a)부분은 반복계산에 필요한 반복 횟수를 나타낸다. 계산을 간단하게 하기 위해 우선 β를 고정하고 σ_ν^2에 대해서 대수우도를 최대화한 다음, 그렇게 해서 구한 σ_ν^2로부터 β의 값을 구하는 과정을 밟는다. 구체적으로 보면 다음과 같은 순차적인 계산과정을 거친다. 이때 $\sigma_\nu^{2(a)}$의 초기치는 0으로 한다.

σ_ν^2에 특별한 조건을 두지 않으면 수치계산 과정에서 분산 추정치가 음의 값을 취하거나 최종적인 추정치도 음의 값에 수렴할 가능성이 있다. 이런 경우를 위해 $\widehat{\sigma_\nu^2} = \exp(\tau)$로 변경해서 τ를 추정하는 방법을 생각해 볼 수 있다. 이 경우 식(2.5)의 우도함수는 더욱 복잡해지고 정확한 미분을 구하기가 쉽지 않기 때문에 수치미분을 이용할 수 있다. R의 비선형 최적화 패키지 'optim'에서는 수치미분에 의한 준뉴톤법(Quasi newton method) 등 다양한 최적화 방법을 이용할 수 있다. 본 연구에서는 이들 두 방법을 모두 적용해 본 결과, 추정결과가 거의 동일하다는 것을 확인할 수 있었다. 따라서 최종적으로 준뉴톤법에 의해 추정한 결과를 사용하였다.

스코어법의 알고리즘

step1: 초기치를 설정한다. : $\sigma_\nu^{2(a)} = 0$

step2: 식 (7)에 의해 $\boldsymbol{\beta^{(a+1)}}\left(\sigma_\nu^{2(a+1)}\right)$을 계산한다.

step3: 식 (6)에 의해 $\sigma_\nu^{2(a+1)}$을 계산한다.

step4: a ← a+1로 하고 step2로 돌아감

이하, $\sigma_\nu^{2(a)}$가 수렴할 때까지 반복

EBLUP추정량은 적당한 사전분포를 설정했을 때의 경험적 베이즈 추정량으로도 해석할 수 있다.

4. MSE 추정량

MSE 추정량은 추정 결과의 안정성에 대한 평가기준으로 널리 사용되고 있다. MSE 추정량이 소지역 추정 결과에 대한 비교척도로 사용되는 만큼 이들 추정법들간의 비교 연구도 수없이 행해지고 있다. 이제 EBLUP 추정량을 θ_i^{EBLUP}이라 하자. $\widehat{\sigma_\nu^2}$을 어떤 방법으로 추정하는가에 따라 MSE 추정방법도 다양하다. 식(2.1)에서처럼 $b_i = 1$로 고정하고 $\widehat{\sigma_\nu^2}$을 최우추정법으로 추정한 경우에는 MSE 추정량을 식(2.8)과 같이 구할 수 있다(Rao, 2003).

이 추정량의 성능을 평가하기 위해서 식(2.8)의 MSE 제곱근을 EBLUP 추정치로 나눈, $\sqrt{MSE(\theta_i^{EBLUP})}\,/\theta_i^{EBLUP}$을 EBLUP 추정치의 오차율(error rate of estimate)이라 하고, 추정량 비교를 위한 지표로 사용한다.

$$MSE(\theta_i^{EBLUP}) = g_{1i}(\widehat{\sigma_\nu^2}) + g_{2i}(\widehat{\sigma_\nu^2}) \tag{2.8}$$

$$\begin{pmatrix} g_{1i}(\widehat{\sigma_\nu^2}) = \dfrac{\widehat{\sigma_\nu^2}\psi_i}{\psi_i + \widehat{\sigma_\nu^2}} (= \gamma_i \psi_i) \\ g_{2i}(\widehat{\sigma_\nu^2}) = (1-\gamma_i)^2 \mathbf{x}_{\mathrm{i}}^{\mathrm{t}} \left[\sum_{\mathrm{i}=1}^{\mathrm{m}} \dfrac{\mathbf{x}_{\mathrm{i}}^{\mathrm{t}} \mathbf{x}_{\mathrm{i}}}{\psi_{\mathrm{i}} + \widehat{\sigma_\nu^2}} \right]^{-1} \mathbf{x}_{\mathrm{i}} \\ \overline{\mathrm{V}}(\widehat{\sigma_\nu^2}) = [\mathrm{I}(\widehat{\sigma_\nu^2})]^{-1} = \left[\sum_{\mathrm{i}=1}^{\mathrm{m}} \dfrac{1}{(\psi_{\mathrm{i}} + \widehat{\sigma_\nu^2})^2} \right]^{-1} \end{pmatrix}$$

제4절 Time Series & Cross-sectional 모형과 HB 추정량

본 절에서는 Rao-Yu 모형을 이용한 소지역 실업통계 추정방법에 대해서 살펴본다. Rao-Yu 모형은 Time series & Cross-sectional 모형 중 하나이다. Time series & Cross-sectional 모형이란 3절에서 설명한 Fay-Herriot 모형을 시계열 방향의 다변량 모형으로 확장한 것으로, 공간정보와 시계열정보 모두를 추정에 이용할 수 있다.

경제활동인구조사에 적용 시에는 추정을 위한 계산상의 효율을 높이기 위해 모형을 계층적 베이즈 모형으로 나타내고 깁스 샘플에 의한 MCMC방법을 적용하였다(Rao, 2003). 본 절에서는 추정에 필요한 MCMC 및 HB 모형에 의한 이론을 간단히 언급한다. 이때 3절에서 설명한 Fay-Herriot 모형에 의한 HB 추정 방법도 함께 언급한다.

1. 베이즈 통계학 및 계층적 베이즈 모형

주로 Takabe(2004)와 김달호(2005)를 인용해서 본 연구에서 사용할 베이즈 통계 방법 및 계층적 베이즈 모형에 대해서 설명한다. 이는 독자들의 베이즈 개념에 대한 이해를 돕기 위함이다. 고전적인 통계 분석과 달리 베이즈 통계학에서는 모수(parameter)의 참값이 확률변수라고 가정한다. 그래서 관측치 자료를 적용했을 때 각종 모수에 대한 조건부 확률분포를 이용해서 추정한다. 사후분포는 다음의 '베이즈 정리'를 이용해서 구한다.

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} (p(\theta|y) \propto p(y|\theta)p(\theta)) \tag{2.9}$$

$p(y|\theta)$를 우도(likelihood)라 부른다. 우도는 모수 θ의 함수이고 θ의 그럴듯함 정도를 나타낸다. $p(\theta)$는 사전분포라 한다. 관측치 y가 관측되기 전에 분석자가 가지고 있는 모수에 관한 정보를 나타낸다고 보면 좋다. 사전분포의 모수는 초모수(hyper parameter)라 부른다. $p(\theta|y)$는 사후분포라 한다. 관측치 자료가 적용되었을 경우 모수 θ에 관한 확률분포이다. 베이즈 통계 분석에서는 이 사후분포를 기초로 모든 추정이 이루어진다. 이와 같은 추론 구조를 베이지안 패러다임이라 한다. 베이지안 패러다임의 구성은 다음과 같다. 첫째, 모수의 사전분포 결정, 둘째, 자료의 확률모형과 사전분포를 이용한 사후분포의 계산, 셋째, 사후분포를 이용한 모수의 추론이다(김달호, 2005).

베이즈 통계학에서는 적당히 정한 손실함수 $L(\theta,\overline{\theta})$를 이용해서 모수 추정치와 참값과의 거리를 측정한다. 이때 손실함수로는 제곱오차손실, 즉, $L(\theta,\overline{\theta}) = (\theta - \overline{\theta})^2$이 자주 이용된다. 식(2.9)와 같은 θ의 사후분포에 의한 손실함수 $L(\Theta,\overline{\Theta})$의 평균을 위험함수(risk function)라고 부른다.

$$R(\theta|y) = \int L(\theta,\overline{\theta})P(\theta|y)d\theta \tag{2.10}$$

베이즈 통계 분석에서는 식(2.10)과 같은 위험함수를 최소화하는 $\hat{\theta}$을 점추정치로 사용한다. 손실함수가 제곱오차손실인 경우는 사후분포의 평균(사후평균)이 위험함수를 최소화하는 추정치가 된다(김달호, 2005).

베이즈 통계분석을 위해서는 사전분포를 설정할 필요가 있다. 이때 사전분포로써 공액사전분포(conjugate prior)가 사용되는 경우가 많다. 공액사전분포는 사전분포와 사후분포가 같은 분포족(distribution family)에 속하고 결과를 이해하기 쉬우며 수학적 계산을 매우 간편하게 해준다. 뿐만 아니라 공액사전분포를 사용하면 계산속도 및 추정치의 정도가 향

상된다. 모수에 관한 사전 정보에 확신이 없거나 모수에 관한 사전정보가 미흡할 경우에는 무정보적 사전분포(noninformative prior)가 이용되는 경우가 많다. 이에 의한 추론은 현재 자료가 가지고 있는 정보 이외의 것에 의해서는 거의 영향을 받지 않게 된다. 무정보 사전분포로써는 Jeffreys가 만든 방법이 자주 사용된다(김달호, 2005).

모수에 대해서 계층적인 구조를 가정한 모형은 계층적 베이즈 모형이라 불린다. 일반적으로 모형을 가정할 때 계층적 구조를 가정하면 모형을 간결하게 기술할 수 있는 경우가 많다. 계층적 베이즈 모형에서 사전분포가 다음과 같이 분해된다고 가정한다.

$$P(\theta) = \int \cdots \int P(\theta|\theta_1)P(\theta_1|\theta_2)\cdots P(\theta_{k-1}|\theta_k)P(\theta_k)d\theta_1 \cdots d\theta_k \quad (2.11)$$

식(2.11)에서 $P(\theta_{k-1}|\theta_k)$는 k번째 단계에서의 사전분포이다. 계층적 베이즈 모형은 모수가 아주 많은 모형이라도 단계적 계산과정을 통해 어느 정도는 대처할 수 있다.

그러나 다음에서 설명할 Rao-Yu 모형의 경우는 사후분포가 복잡해지고 평균치를 구할 때 적분의 차원이 높아지는 상황이 발생된다. 이와 같은 복잡한 다중적분에 대한 평가를 해석적으로 하는 것은 거의 불가능하고, 수치적 계산에 의한 근사적 방법을 사용하더라도 계산에 걸리는 부담이 커진다. 만약 사후분포로부터 상호 독립적인 대량의 난수 $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_D)$를 생성할 수 있으면 그들 난수들의 표본평균과 표본분산을 계산함으로써 대수법칙으로부터 사후분포의 평균 및 분산을 추정할 수 있다. 다음 절에서 이와 같은 난수 생성 방법을 소개한다.

2. 마코프 체인 몬테 칼로 방법

확률변수의 시계열 $(x_1, x_2, \cdots) = \{x_t\}_{t=1}^{\infty}$을 생각하자. 마코프 체인은 다음 식(2.12)와 같이, 현 시점(t)에서의 확률변수의 실현값 $\{x_t\}_{t=1}^{\infty}$이 바로 전 시점$(t-1)$에서의 조건부 분포에만 영향을 받을 때, $\{x_t\}_{t=1}^{\infty}$를 마코프 체인이라고 부른다.

$$P(x_t|x_{t-1},\cdots,x_1) = P(x_t|x_{t-1}) \tag{2.12}$$

각 시점에서의 조건부 분포 및 초기분포가 알려져 있는 경우, 마코프 체인에 의한 난수 생성은 다음과 같이 쉽게 할 수 있다.

step1 : $P(x_1) \rightarrow \widetilde{x_1} \quad (t=0)$
step2 : $P(x_t|\widetilde{x}_{t-1}) \rightarrow \widetilde{x_t}$

t를 $t+1$로 바꾸고 step2로 돌아간다.

여기서 $t \rightarrow \infty$일 때 x_t의 분포가 특정한 확률밀도로 수렴할 경우, 마코프 체인은 불변밀도(invariant density)를 갖는다고 하고, 이때의 불변밀도를 확률밀도함수를 갖는 분포로 나타내어 불변분포(invariant distribution)라 부른다. $\{x_t\}_{t=1}^{\infty}$에 불변밀도가 있을 경우에는 위의 알고리즘에서 step1과 step2를 반복함으로써 불변밀도와 유사한 분포로부터 난수를 생성할 수 있다.

만약 확률변수 간에 마코프 체인 관계를 만들 수 있고, 사후분포가 불변밀도를 갖도록 할 수 있다면 위의 알고리즘을 충분히 많이 반복함으로써 사후분포에 근사한 분포로부터 난수를 생성할 수 있다. 이와 같은 난수 생성 방법을 마코프 체인 몬테 칼로(Markov Chain Monte Carlo: MCMC) 방법이라 한다. 이 방법에 의해 생성된 난수는 몬테 칼로 표본(monte carlo sample)이라고도 부른다. MCMC 방법에는 다양한 종류가 있다. 그 중에서도 비교적 간단하게 프로그램을 작성할 수 있는 깁스 샘플링에 대해서 다음과 같이 설명한다.

k개의 확률변수에 대한 결합확률밀도함수 $P(\theta) = P(\theta_1,\theta_2,\cdots,\theta_k)$가 있다고 하자. 특정한 확률변수 외의 나머지 모든 확률변수에 대한 조건부확률밀도를 그 확률변수에 대한 조건부확률분포라 부른다. 깁스 샘플링 알고리즘은 전체 조건부 사후분포에 의해 다음과 같이 구성된다.

$\theta_n^{(t)}$는 t번째 생성된 θ_n의 난수를 나타낸다.

적당한 초기치 $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \cdots, \theta_k^{(0)})$를 세팅한다.

step 1 : $P(\theta_1^{(t)} | (\theta_2^{(t)}, \cdots, \theta_k^{(t)}) \rightarrow \theta_1^{(t+1)}$

step 2 : $P(\theta_2^{(t)} | (\theta_1^{(t+1)}, \cdots, \theta_k^{(t)}) \rightarrow \theta_2^{(t+1)}$

$\vdots$

step k-1 : $P(\theta_{k-1}^{(t)} | (\theta_1^{(t+1)}, \cdots, \theta_{k-2}^{(t+1)}, \theta_k^{(t)}) \rightarrow \theta_{k-1}^{(t+1)}$

step k : $P(\theta_k^{(t)} | (\theta_1^{(t+1)}, \cdots, \theta_{k-1}^{(t+1)}) \rightarrow \theta_k^{(t+1)}$

t를 $t+1$로 하고 step 1로 돌아가서 반복 수행한다.

깁스 샘플링 알고리즘에 의해 생성된 난수는 사후분포가 수렴성을 갖는 마코프 체인이 된다고 알려져 있다(김달호, 2005; Rao, 2003). 결합확률밀도함수, $P(\theta) = P(\theta_1, \theta_2, \cdots, \theta_k)$가 복잡한 경우라도 그 조건부 확률분포는 저차원의 단순한 분포(정규분포, 감마분포 등)가 되는 경우가 많고 그 분포로부터 난수는 상용 통계 패키지에 의해 쉽게 생성될 수 있다.

마코프 체인이 정상분포에 수렴할 때까지 계속해서 난수를 발생시키는 것을 'burn in'이라 부른다. burn in까지 생성된 난수는 초기치의 영향을 강하게 받는다고 볼 수 있기 때문에 표본으로 사용하기가 어렵다. burn in 이후의 난수는 'burn out'이라 부른다. 단일 마코프 체인으로부터 난수를 생성시키는 것을 'single run'이라 하고, 다중 마코프 체인에 의해 난수를 생성해서 그 체인들의 평균을 추정치로 사용하는 방법을 'multi run'이라 한다. multi run의 경우, 모수의 초기 상태에 대한 영향을 줄일 수 있기 때문에 불변분포로의 수렴이 single run에 비해 빠르다 (中妻照雄, 2003). 또한 multi run을 적용한 경우에는 마코프 체인이 정상분포로

수렴하는지 여부를 판단할 수 있다는 점에서 유용하다.

3. Time series & Cross-sectional 모형

가. Rao-Yu 모형

Fay-Herriot 모형을 시계열 방향으로 확장한 것이 Time series & Cross-sectional 모형으로, 이는 시간영역과 공간영역 양 방향의 정보를 측정하는 데 이용할 수 있다. Time series & Cross-sectional 모형의 일종인 Rao-Yu 모형은 지역 효과 ν_i 외에 랜덤한 시계열 효과를 나타내는 확률변수 u를 포함하고 있으며, 다음과 같은 형태로 나타낼 수 있다.

$$y = \theta + e \tag{2.13}$$
$$\theta = \mathbf{x}\boldsymbol{\beta} + \nu_{\mathrm{i}} + \mathrm{u}, \quad \mathrm{i} = 1,\cdots,\mathrm{m}, \mathrm{t} = 1,\cdots,\mathrm{T}$$
$$u = \rho u_{t,t-1} + \epsilon$$

여기서, y: i번째 소지역과 t시점에서의 완전실업률 관측치 자료
$\mathbf{x}$: i번째 소지역과 t시점에서의 보조정보(상수항을 포함)
$\theta_{i,t}$: i번째 소지역과 t시점에서의 완전 실업률 참값
e : i번째 소지역과 t시점에서의 표집오차
$\nu_i \sim N(0,\sigma_\nu^2)$: i번째 소지역의 지역 효과
u : i번째 소지역과 t시점의 시계열 효과
$\epsilon \sim N(0,\sigma_e^2)$.

여기서 y는 각 소지역 i에 대해서 독립이라고 가정한다. 그러나 표집오차는 어느 정도의 시계열 상관이 있을 것으로 예상되기 때문에 시간 t에 대해서는 독립이 아니다. 5절에 의하면 우리나라의 경제활동인구조사 자료에서 표집오차의 시계열 상관은 논리상 35기이다. 이것을 고려하면 36기의 변수를 사용하는 것이 타당하다. 이는 표본설계가 36개 그룹을 갖는 연동표본구조이기 때문이다. 그러나 실제로 경제활동인구조사 자료의 자기상관성을 검토한 결과에 따르면 대체로 24개월 이후 시점에서 그 상관성이 약 0.1 근처로 매우 약해지기 때문에(부그림 2-1) 본

연구에서는 24기($T=24$)의 자료를 사용하였다. 따라서 변수 벡터 차원은 24가 된다.

나. 계층적 베이즈 모형을 이용한 Rao-Yu 모형

식(2.13)의 y에 관한 식은 다음과 같은 분포형태로 바꿔 나타낼 수 있다.

$$\mathbf{y_i} \sim \mathrm{N}(\boldsymbol{\theta_i},\ \Sigma_\mathrm{i}),\ i=1,2,\cdots,m \tag{2.14}$$

$$\begin{cases}\Sigma_i : i\text{번째 소지역(시군구)에 의한 } \boldsymbol{e_i}\text{의 분산공분산행렬} \\ \mathbf{y_i} = (\mathrm{y_{i1}},\cdots,\mathrm{y_{iT}})',\ \boldsymbol{\theta_i} = (\theta_{\mathrm{i1}},\cdots,\theta_{\mathrm{iT}})', \mathbf{e_i} = (\mathrm{e_{i1}},\cdots,\mathrm{e_{iT}})'\end{cases}$$

$\boldsymbol{y_i}$는 $\boldsymbol{\theta_i}$에 대해 불편성을 갖는다. 표집오차의 식(2.14)에서 분산공분산행렬(variance-covariance matrix) Σ_i는 일반적으로 알려져 있다고 가정한다. 그러나 실제로는 모르는 경우가 대부분이기 때문에 표본자료로부터 추정한 추정치 $\widehat{\Sigma_i}$로 대치한다. $\widehat{\Sigma_i}$는 다음과 같이 추정한다. 또 표집오차의 자기상관계수는 일반적으로 사용되는 계산방법을 사용하였다(조신섭, 손영숙, 1999).

$$\widehat{\Sigma_i} = (\hat{\sigma}^2_{its})_{T\times T},\quad i=1,\cdots,m,\quad t,s=1,\cdots,T \tag{2.15}$$

$$\begin{pmatrix}\hat{\sigma}^2_{itt} = \overline{C} y^2_{it},\quad \hat{\sigma}_{its} = \overline{\rho}_{|t-s|}\hat{\sigma}_{itt}\hat{\sigma}_{iss} \\ \overline{C_i} = \dfrac{1}{T}\sum_{t=1}^{T} C_{it},\quad \overline{\rho}_i = \dfrac{1}{m}\sum_{i=1}^{m}\rho_{ik}\end{pmatrix}$$

여기서, C_{it} : i번째 소지역, 시점 t에서의 실업자수의 오차율,
ρ_{ik} : i번째 소지역에 의한 표집오차의 k차 자기상관.

오차율 및 오차의 자기상관을 그대로 이용하는 것보다 식(2.15)와 같이 시간 또는 공간에 대해서 평균을 구해서 사용하는 편이 오차율 개선 측면에서 좋은 결과를 제시한다는 보고가 있다(You 등, 2003).

식(2.13)~식(2.15)에 의해 Rao-Yu 모형은 다음과 같은 계층적 베이즈 모형으로 표현될 수 있다.

$$\begin{aligned}
\mathbf{y_i}|\boldsymbol{\theta_i} &\sim \mathrm{N}(\boldsymbol{\theta_i}, \boldsymbol{\Sigma_i}) \\
\theta|\boldsymbol{\beta}, u, \sigma_\nu^2 &\sim N(\mathbf{x}\boldsymbol{\beta} + \rho \mathrm{u}, \sigma_\nu^2) \\
u|u_{i,t-1}, \sigma_\epsilon^2 &\sim N(\rho u_{i,t-1}, \sigma_\epsilon^2) \\
\sigma_\nu^2 \sim IG(a_1, b_1), \quad \sigma_\epsilon^2 &\sim IG(a_2, b_2) \\
f(\beta) &\propto 1 \text{ (무정보적 사전분포)}
\end{aligned} \tag{2.16}$$

일반적으로 회귀계수 $\boldsymbol{\beta}$에 관한 정보가 거의 없기 때문에 이를 모형에 반영하기 위해서 $\boldsymbol{\beta}$의 사전분포는 무정보적 사전분포(flat prior)를 사용한다. 또 역감마분포는 σ_ν^2의 공액사전분포이다(Rao, 2003; 中妻照雄, 2003; 김달호, 2005). a_1, a_2, b_1, b_2는 양의 값으로 알려져 있고, σ_ν^2와 σ_ϵ^2에 관한 사전 정보가 없다는 것을 반영하기 위해서 보통 충분히 작은 값을 취한다. You 등(2003)은 $a_1 = a_2 = b_1 = b_2 = 0.001$로 설정하였고 본 연구에서도 이 값을 사용하였다. 0.001이란 값은 충분히 작고 다른 수치에는 거의 영향을 미치지 않는다고 가정하는 셈이다.

미지의 모수는 $(\boldsymbol{\beta}, \sigma_\nu^2, \sigma_e^2, \boldsymbol{u_1}, \cdots, \boldsymbol{u_m}, \boldsymbol{\theta_1}, \cdots, \boldsymbol{\theta_m})$이고, 이 가운데 우리의 관심 모수는 현 시점 T에서의 m개의 소지역에 대한 $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)$가 된다. 계층적 베이즈 분석에서 θ_{iT}는 이것의 사후평균 $E(\theta_{iT}|y)$에 의해 추정되고, 이 추정값의 불확실성은 사후분산 $Var(\theta_{iT}|y)$에 의해 측정된다. 우리는 θ_{iT}의 사후평균과 사후분산을 구하기 위해 뒤에서 설명할 깁스 샘플링(Gibbs sampling, Gelfand와 Smith, 1991; Gelman과 Rubin, 1992)을 사용한다. 각 모수에 대한 조건부사후분포는 각각 다음과 같이 나타낼 수 있다(You 등, 2003).

$$\begin{aligned}
&\boldsymbol{\beta}|\mathbf{y}, \sigma_\nu^2, \sigma_\mathrm{e}^2, \mathbf{u}, \boldsymbol{\theta} \sim N((X^tX)^{-1}X^t(\boldsymbol{\theta} - \boldsymbol{u}), \ \sigma_\nu^2(X^tX)^{-1}), \\
&\sigma_\nu^2|\mathbf{y}, \sigma_\epsilon^2, \mathbf{u}, \boldsymbol{\theta} \sim IG\left(a_1 \times \frac{m \times T}{2}, \ b_1 + \sum_{i=1}^{m}\sum_{t=1}^{T} \frac{(\theta - \mathbf{x}\boldsymbol{\beta} - \mathrm{u})^2}{2}\right), \\
&\sigma_\epsilon^2|\mathbf{y}, \sigma_\nu^2, \mathbf{u} \sim IG\left(a_2 \times \frac{m \times (T-1)}{2}, \ b_2 + \sum_{i=1}^{m}\sum_{t=1}^{T} \frac{(u - \rho u_{i,i-1})}{2}\right), \\
&u_{i1}|\mathbf{y}, \boldsymbol{\beta}, \sigma_\nu^2, \sigma_\mathrm{e}^2, \mathrm{u_{i2}}, \boldsymbol{\theta} \quad [T=1]
\end{aligned} \tag{2.17}$$

$$\sim N\left(\left(\frac{1}{\sigma_\nu^2}+\frac{\rho^2}{\sigma_e^2}\right)^{-1}\left(\frac{\theta_{i1}-\mathbf{x_{i1}}\boldsymbol{\beta}}{\sigma_\nu^2}+\frac{\rho u_{i2}}{\sigma_e^2}\right),\ \left(\frac{1}{\sigma_\nu^2}+\frac{\rho}{\sigma_e^2}\right)^{-1}\right),$$

$$u|\mathbf{y},\boldsymbol{\beta},\sigma_\nu^2,\sigma_\mathrm{e}^2,\mathrm{u_{i,t-1}},\boldsymbol{\theta}\quad [2\le t\le T-1]$$

$$\sim N\left(\left(\frac{1}{\sigma_\nu^2}+\frac{1+\rho}{\sigma_e^2}\right)^{-1}\left(\frac{\theta-\mathbf{x}\boldsymbol{\beta}}{\sigma_\nu^2}+\frac{\rho u_{i,t-1}+\rho u_{i,t+1}}{\sigma_e^2}\right),\ \left(\frac{1}{\sigma_\nu^2}+\frac{1+\rho}{\sigma_e^2}\right)^{-1}\right),$$

$$u_iT|\mathbf{y},\boldsymbol{\beta},\sigma_\nu^2,\sigma_\mathrm{e}^2,\mathrm{u_{iT-1,}}\boldsymbol{\theta}\quad [t=T]$$

$$\sim N\left(\left(\frac{1}{\sigma_\nu^2}+\frac{1}{\sigma_e^2}\right)^{-1}\left(\frac{\theta_{i,T-1}-\mathbf{x_{i,T-1}}\boldsymbol{\beta}}{\sigma_\nu^2}+\frac{\rho u_{i,T-1}}{\sigma_e^2}\right),\ \left(\frac{1}{\sigma_\nu^2}+\frac{1}{\sigma_e^2}\right)^{-1}\right),$$

$$\boldsymbol{\theta_i}|\mathbf{y},\boldsymbol{\beta},\sigma_\nu^2,\sigma_\mathrm{e}^2,\mathbf{u}$$

$$\sim N\left((\sigma_\nu^{-2}\boldsymbol{I_T}+\Sigma_i^{-1})^{-1}(\Sigma_i^{-1}\mathbf{y_i}+\sigma_\nu^{-2}(\mathrm{X_i}\boldsymbol{\beta}+\mathbf{u_i})),\ (\sigma_\nu^{-2}\mathbf{I_T}+\Sigma_\mathrm{i}^{-1})^{-1}\right).$$

여기서, 각 자료의 형태는 다음과 같다.

$\mathbf{y}=(\mathbf{y_1^t},\cdots,\mathbf{y_m^t})$: $Tm\times 1$ 관측치 벡터,
$X=(X_1^t,\cdots,X_m^t)$: $Tm\times 1$ 보조정보 벡터,
$\boldsymbol{\theta}=(\boldsymbol{\theta_1^t},\cdots,\boldsymbol{\theta_m^t})$: $Tm\times 1$ 모수 벡터,
$\boldsymbol{u}=(\boldsymbol{u_1^t},\cdots,\boldsymbol{u_m^t})$: $Tm\times 1$ 지역내 시계열 효과 벡터,
$\boldsymbol{I_T}$: $T\times T$ 단위행렬.

유사한 방법으로 식(2.2)와 (2.3)의 Fay-Herriot 모형은 다음과 같이 계층적 모형으로 표현될 수 있다. 자세한 내용은 You 등(2003)과 김달호(2005)를 참고할 수 있다. 이 모형에서는 현재 시점(T)만의 추정값을 구한다.

◦ θ에 대한 조건부 $y|\theta\sim N(\theta,\ \sigma^2)$
◦ $\beta_t,\ \sigma_v^2$에 대한 조건부 $\theta|\beta_t,\sigma_{\nu t}^2\sim N(x^t\beta_t,\ \sigma_{\nu t}^2)$
◦ $\sigma_{\nu t}^2\sim IG(a_t,\ b_t),\quad \beta_t\propto 1$ (무정보적 사전분포)

Fay-Herriot 모형의 임의의 시점 t에서의 조건부분포는 다음과 같다. 이 때, $y_t=(y_{1t},\cdots,y_{mt})^t$, $X_t^t=(x_{1t},x_{2t},\cdots,x_{mt})$, $\theta_t^t=(\theta_{1t},\theta_{2t},\cdots,\theta_{mt})^t$, $t=1,2,\cdots,T$ 라 하자. 그러면, $i=1,2,\cdots,m$에 대해서,

$$\beta_t|y_t,\sigma_{\nu t}^2,\theta_t \sim N\big((X_t^tX_t)^{-1}X_t^t\theta_t,\quad \sigma_{\nu t}^2(X_t^tX_t)^{-1}\big),$$
$$\theta|y_t,\beta_t,\sigma_{\nu t} \sim N((1-r)y+rx^t\beta_t,\quad \sigma(1-r)),$$
$$r=\sigma^2/(\sigma^2+\sigma_{\nu t}^2),$$
$$\sigma_{\nu t}^2|y_t,\beta_t,\sigma_e^2,u,\theta \sim IG(a_1+m/2,\quad b_1+\sum_{i=1}^{m}(\theta-x^t\beta_t)^2/2).$$

위에서 설명한 두 모형 각각의 조건부사후분포에 MCMC 방법을 적용하고 소지역별 실업자 수를 추정한다. 경제활동인구조사자료를 Rao-Yu 모형에 적용하여 추정할 경우, 연동표본 설계원칙에 의해 원래 36개월분의 자료가 모두 사용되는 것이 좋을 것이다. 그러나 앞에서 이미 설명한 바와 같이 시점 간 자기상관성이 24개월 이후에는 대체로 상관계수가 '0' 근처로 수렴하는 경향이 있고, 자료처리상의 편리 또는 자료이용상의 제한 등을 고려하여, 현시점의 추정을 위해 24개월 자료를 이용하기로 한다. 그리고 24개월분의 추정치 $\boldsymbol{\theta}$를 산출한다. 이 가운데 가장 최근 월($T=24$)의 수치만을 최종 추정치로 이용한다.

각 확률변수의 조건부사후분포의 평균을 해석적으로 계산할 수 있는 경우에는 그들의 평균과 분산 계산을 통해 보다 효율적인 추정을 할 수 있고, 몬테 칼로 방법에 의한 시뮬레이션 오차를 구할 수 있다. 이와 같은 추정량을 라오-블랙웰 추정량(Rao-Blackwellized estimator)이라 부른다. 라오-블랙웰 추정량 계산에 필요한 조건부사후분포의 평균은 깁스 샘플러 계산 과정에서 구할 수 있기 때문에 계산에 수반되는 추가적인 부담은 많지 않다. multi run으로 추정한 경우, Rao-Yu 모형에 의한 라오-블랙웰 추정량은 다음과 같다(You 등, 2003).

$$E[\theta_i|y_i] = \frac{1}{LD}\sum_{l=1}^{L}\sum_{k=d+1}^{d+D}[(\sigma_\nu^{-2(lk)}\boldsymbol{I_{24}}+\Sigma_i^{-1})\times(\Sigma_i^{-1}\mathbf{y}_i+\sigma_\nu^{-2(lk)}(X_i\boldsymbol{\beta^{(lk)}}+\boldsymbol{u_i^{(lk)}}))] \quad (2.18)$$

$$V[\theta_i|y_i] = \frac{1}{LD}\sum_{l=1}^{L}\sum_{k=d+1}^{d+D}(\sigma_\nu^{-2(lk)}\boldsymbol{I_{24}}+\Sigma_i^{-1}) \quad (2.19)$$
$$+\frac{1}{(LD)^2}\sum_{l=1}^{L}\sum_{k=d+1}^{d+D}[(\sigma_\nu^{-2(lk)}\boldsymbol{I_{24}}+\Sigma_i^{-1})\times(\Sigma_i^{-1}\mathbf{y}_i+\sigma_\nu^{-2(lk)}(X_i\boldsymbol{\beta^{(lk)}}+\boldsymbol{u_i^{(lk)}}))]$$

$$\times[(\sigma_\nu^{-2(lk)}\boldsymbol{I_{24}}+\Sigma_i^{-1})\times(\Sigma_i^{-1}\mathbf{y}_i+\sigma_\nu^{-2(lk)}(X_i\boldsymbol{\beta^{(lk)}}+\boldsymbol{u_i^{(lk)}}))]^t$$

$$-\frac{1}{(LD)^2}[\sum_{l=1}^{L}\sum_{k=d+1}^{d+D}[(\sigma_\nu^{-2(lk)}\boldsymbol{I_{24}}+\Sigma_i^{-1})\times(\Sigma_i^{-1}\mathbf{y}_i+\sigma_\nu^{-2(lk)}(X_i\boldsymbol{\beta^{(lk)}}+\boldsymbol{u_i^{(lk)}}))]]$$

$$\times[\sum_{l=1}^{L}\sum_{k=d+1}^{d+D}[(\sigma_\nu^{-2(lk)}\boldsymbol{I_{24}}+\boldsymbol{\Sigma_i^{-1}})\times(\boldsymbol{\Sigma_i^{-1}}\mathbf{y}_i+\sigma_\nu^{-2(lk)}(X_i\boldsymbol{\beta^{(lk)}}+\boldsymbol{u_i^{(lk)}}))]]^t$$

여기서,

L : 동시에 생성할 마코프 체인의 수

d : brun in까지 각 마코프 체인 안에서 생성된 난수의 총 수

D : burn out 후에 생성된 난수의 총 수

본 연구는 24개월분의 추정치 $\boldsymbol{\theta}$가운데 $T=24$의 수치만을 추정치로 이용하기 때문에 실제로는 식(2.18)은 24행째의 성분에, 식(2.19)는 24행과 24열째의 성분에 관해서 계산하면 된다. 난수들의 안정성 및 깁스 샘플러 계산에 대한 결과는 5절 경제활동인구조사에 적용한 결과 분석에서 자세하게 설명하기로 한다.

다. 불변분포로의 수렴판단 및 모형 선택

1) 수렴판단

앞에서 살펴보았듯이 MCMC 방법에 의해 multi run을 적용한 경우에는 마코프 체인이 정상분포에 수렴하는지를 진단할 수 있다. 진단에는 다음과 같은 통계량을 이용한다(Rao, 2003; 김달호, 2005).

$$\hat{R}_{i,T}=\hat{V}_{i,T}/\hat{W}_{i,T} \tag{2.20}$$

$$\hat{V}_{i,T}=\frac{D-1}{D}\hat{W}_{i,T}+\frac{1}{D}\hat{B}_{i,T},$$

$$\frac{1}{D}\hat{B}_{i,T}=\frac{1}{L-1}\sum_{l=1}^{L}(\overline{\theta}_{i,T}^{(l)}-\overline{\theta}_{i,T})^2 \ : \text{ between run variance}$$

$$\hat{W}_{i,T}=\frac{1}{L(D-1)}\sum_{l=1}^{L}\sum_{k=d+1}^{d+D}(\overline{\theta}_{i,T}^{(lk)}-\overline{\theta}_{i,T}^{(l)})^2 \ : \text{ within run variance}$$

$$\overline{\theta}_{i,T}^{(l)} = \frac{1}{D}\sum_{k=d+1}^{d+D}\theta_{i,T}^{(lk)}, \quad \overline{\theta}_{i,T} = \frac{1}{L}\sum_{l=1}^{L}\overline{\theta}_{i,T}^{(l)}$$

$\theta_{i,T}^{(lk)}$: i 소지역의 시점 $t=T$에서 l번째 체인 내의 k번째 난수

통계량 $\hat{R}$은 현 시점($t=T$)의 결과에 대해서 계산한다. 각 소지역에 대해서 $\hat{R}_{i,T}$ 이 1에 가까워지면 마코프 체인이 정상분포에 수렴한다고 판단된다. 그러나 어느 정도 1에 가까워지면 좋은지에 대한 판단의 문제가 남아 있다. 본 분석에서는 1.5이상 값을 갖는 소지역이 없으면 마코프 체인이 정상분포에 거의 수렴한다고 판단하기로 했다.

2) 모형 선택

계층적 베이즈 모형에 의해 분석할 경우 설명변수 및 모형 내의 상수 선택에 따라 다양한 모형을 고려할 수 있다. 이런 경우 모형 사이의 우월성을 비교해서 적당한 모형을 선택할 필요가 있다. 최적 모형 선택은 예측분포로부터 추출한 몬테 칼로 표본을 이용할 수 있다. 예측분포란 사후평균에 의해 추정한 모수 $\boldsymbol{\theta}$를 우도 함수 $f(\boldsymbol{y}|\boldsymbol{\theta})$에 대입한 것이다. 예측분포에서 몬테 칼로 표본 $\boldsymbol{y}^{*(lk)}$ 추출은 우도 $f(\boldsymbol{y}|\boldsymbol{\theta})$에 $\boldsymbol{\theta}^{(lk)}$를 대입한 분포 $f(\boldsymbol{y}|\boldsymbol{\theta}^{(lk)})$ $(=N(\boldsymbol{y}|\boldsymbol{\theta}^{(lk)}))$를 이용한다. 깁스 샘플에서는 식 (2.14)을 이용해서 비교적 간단하게 $\boldsymbol{y}^{*(lk)}$을 생성할 수 있기 때문에 계산 시에 새롭게 걸리는 부하는 없다.

만약 $\boldsymbol{y}_{obs}$가 관측자료라 할 경우, 모형이 잘 설정되었고, $\boldsymbol{\theta}^{(lk)}$의 정도가 높게 추정되었다면, $\boldsymbol{y}^{*(lk)}$는 사후예측분포 $f(\boldsymbol{y}^{*}|\boldsymbol{y}_{obs})$의 근사 분포로부터 추출된 표본이라고 간주할 수 있다. 이 경우에는 $\boldsymbol{y}^{*(lk)}$와 $\boldsymbol{y}_{obs}$는 비교적 유사한 값이 될 것으로 예상된다. 만약 모형이 적절하지 않다면, 이 두 값들 간에 서로 차이가 생긴다. 여기서 어떻게든 관측치와 $\boldsymbol{y}^{*(lk)}$와의 거리를 측정할 수 있다면 그것을 모형 선택 기준으로 이용할 수 있다(Takabe, 2004). 그래서 이와 같은 거리 개념을 이용하는 두 개의 통계량을 다음과 같이 고려할 수 있다.

가) d값

모형들 간 우월성 비교에는 사후예측분포를 이용한 다음의 통계량을 이용한다.

$$d(\mathbf{y}^*, \mathbf{y}_{obs}) = \mathrm{E}(\|\mathbf{y}^{*(ld)} - \mathbf{y}_{obs}\|/(31\times 24)|\mathbf{y}_{obs}) \\ = \sum_{l=1}^{L}\sum_{k=d+1}^{d+D} \|\mathbf{y}^{*(ld)} - \mathbf{y}_{obs}\|/(31\times 24) \tag{2.21}$$

$\|\cdot\|$은 보통 유클리드 놈(euclid norm: 제곱합의 제곱근)을 나타낸다. 기대치는 LD개의 표본을 이용해서 구한다. 통계량 d값이 작을수록 모형 적합이 잘 되었다고 평가한다. 이 통계량은 모형 간 상대적인 비교 시에 이용될 수 있다.

나) 사후예측 p값

모형 적합은 다음과 같은 사후예측 p값에 해당하는 추정량을 이용하여 진단할 수 있다.

$$\hat{p} = \frac{1}{L}\sum_{l=1}^{L}\hat{p_i}^{(l)} \tag{2.22}$$

$$\hat{p_i}^{(l)} = \frac{1}{D}\sum_{k=d+1}^{d+D} I\{T(\mathbf{y}_i^{*(lk)}, \boldsymbol{\theta}_i^{(lk)}) \geqq T(\mathrm{y}_i^{obs}, \boldsymbol{\theta}_i^{(lk)})\}$$

$$T(\mathbf{y}^*, \boldsymbol{\theta}) = \sum_{\mathrm{i}=1}^{\mathrm{m}}(\mathbf{y}_{\mathrm{i}}^* - \boldsymbol{\theta}_{\mathbf{i}})\boldsymbol{\Sigma}_{\mathbf{i}}^{-1}(\mathbf{y}_{\mathrm{i}}^* - \boldsymbol{\theta}_{\mathbf{i}})^{\mathrm{t}}$$

$$\mathbf{y}^* = (\mathbf{y}_1^*, \cdots, \mathbf{y}_m^*), \quad \mathbf{y}_i^* = (\mathbf{y}_{i,1}^*, \cdots, \mathbf{y}_{i,T}^*)$$

식(2.22)에서 $I\{\cdot\}$는 지시함수로 { }이 참일 때 1이 되고, 그렇지 않으면 0이 된다. 그리고 모형이 자료에 잘 적합되었다면 $T(\boldsymbol{y}_i^{*(lk)}, \boldsymbol{\theta}_i^{(lk)})$와 $T(\boldsymbol{y}_i^{obs}, \boldsymbol{\theta}_i^{(lk)})$가 유사한 값을 갖게 될 것이다. 그러면 통계량 $\hat{p}$은 0.5 근처의 값을 갖게 된다. $\hat{p}$이 극단적으로 0 또는 1에 근사할 경우에는 모형 적합이 좋지 않은 것으로 생각된다. 이 통계량은 모형 간 상대적 비교와 함께 $\hat{p}$이 0.5 근처에 있을지에 대한 절대적인 적합성을 확인하는데 사용될 수 있다.

제5절 경제활동인구조사에 적용

1. 자료설명 및 검토

본 장에서는 분석에 사용된 자료에 대해서 설명하기로 한다. 실업자 수 추정에 필요한 핵심정보(종속변수)는 경제활동인구조사(경활조사) 자료이고, 보조정보(설명변수)로는 고용보험 자료의 실업급여 수급자 수를 사용하였다.

가. 경제활동인구조사 자료

경활조사는 월간조사로서 통계청에서는 전국 및 광역시·도에 대한 실업 및 취업 등의 고용 동향을 월별, 분기별, 연별 자료의 형태로 공표하고 있다. 본 연구의 목표는 소지역별로 월별 실업자 수를 추정하는 것이고, 따라서 모형에 사용될 종속변수는 경활조사자료의 각 소지역에서의 실업자 수가 된다. 분석에 사용된 자료는 2005년 1월부터 2007년 12월까지의 경활조사 자료이다. 추정의 효율성을 높이기 위해서는 더 광범위한 시계열 자료를 사용하는 것이 바람직하겠지만, 보조자료로 사용될 고용보험 자료(노동부, 2008)의 활용범위에 다소 제약이 있기 때문에 추정 연구 범위는 3년으로 제한하였다. 왜냐하면 노동부에 따르면 2005년 이후에 고용보험 자료의 범위가 현재와 같이 확대되고 자료체계가 정비되었기 때문에 그 이후 자료의 활용가치가 높기 때문이다.

본 연구의 궁극적인 목적은 우리나라 모든 시군구에 대해서 실업자를 추정하는 것이고, 2006년 현재 우리나라 행정구역상 시군구 수는 232개이다. 그러나 2008년 현 시점에서 실제로 모든 시군구에 대해서 실업자를 추정하는 것은 두 가지 측면에서 제약이 따른다.

첫째, 2008년 현재 시군구 단위로 추계인구를 활용할 수 있는 지역은 31개 지역에 불과하다. 추계인구는 경활조사 자료에서 시군구별 직접추정치(비추정)를 계산하는 과정에서 각 시군구 단위의 벤치마크로 사용되는 기준값이다. 물론 대안으로 주민등록인구를 사용할 수 있을 것이고, 실제로 31개 지역에 대해서 추계인구와 주민등록인구의 15세 이상

인구를 비교한 결과 몇 개 지역을 제외하면 큰 차이는 없었다. 참고로 경기도에서 수행하는 지역통계조사에서는 벤치마크로써 추계인구 대신에 주민등록인구를 사용하고 있고, 그 밖의 지역에서는 통계청에서 작성한 추계인구를 기본으로 하여 실업자 수를 추정하고 있다. 그러므로 지역마다 사용하는 벤치마크가 다르다는 현실적인 문제를 감안해서 결과를 비교할 필요가 있다. 소지역별 추계인구는 2008년 10월경에 작성될 예정에 있다(인구동향과, 통계청).

둘째, 추정치의 결과를 비교할 “gold standard"가 없는 상태에서 이들 31개 지역들의 대규모 표본조사 결과가 모형추정치와 비교대상이 될 만하다고 판단되었다. 여기서 "gold standard"는 참값에 유사한 값이어야 하고 소지역 추정 결과의 정확성 검증을 위한 벤치마크로 사용된다.

나. 보조정보

추정의 효율성을 높이기 위한 보조정보로는 고용보험 자료의 실업급여 수급자 수와 경활조사 자료내의 소지역이 포함된 대지역(광역시·도)의 실업자 수를 고려한다. 회귀분석을 포함한 모형 기반 추정에서 설명변수의 선택은 추정 결과에 중요한 영향을 끼친다. 소지역 추정 모형에서 이러한 설명변수는 보조정보(auxiliary information) 또는 공변량(covariate)이라 부르기도 한다. 보조정보는 전수조사 자료를 이용하거나 행정자료를 이용하는 것이 일반적이다. 왜냐하면 이들 자료는 표집오차의 발생가능성이 적기 때문이다. 우리나라의 실업자 정보는 몇 개의 행정자료를 통해 이용될 수 있다. 대표적으로 고용보험, 국민건강보험, 국민연금 자료 등을 꼽을 수 있다. 그러나 이들 모두 가입 대상이 다르고, 경활조사에서 사용하는 실업자 정의와도 다르다. 본 연구에서는 사전검토 결과, 보조자료로서 가장 가치가 있을 것으로 판단되는 고용보험 자료의 실업급여 수급자 수를 우선적으로 검토·분석하여 추정에 이용하였다. 그러나 추후 다른 행정자료의 활용가치에 대해서 충분히 검토할 필요는 있다. 고용보험 자료에 대한 독자들의 이해를 위해서 다음에서 보조정보 위주로 간략하게 소개하였다.

1) 고용보험의 실업급여 수급자 수

우리나라의 고용보험 제도는 실업의 예방, 고용의 촉진 및 근로자의 직업능력 개발과 향상은 물론 근로자의 생활에 필요한 급여를 지급하여 실직근로자의 생활안정 및 재취업을 지원하는 사회보험 제도의 하나로 1995년 7월 1일부터 시행되었다. 노동부가 주무 부처로써 노동부 지방노동관서, 고용지원센터와 근로복지공단에서 관련 업무를 수행한다.

고용보험의 적용대상은 ① 상시근로 1인 이상 모든 사업장('98.10.01부터), ② 2천만 원 이상 건설공사, ③ 농업・임업・어업 및 수렵업 중 상시 5인 이상(법인인 경우 1인 이상) 근로자를 고용하는 사업(장)으로 상시 근로자는 고용형태를 불문하고 사실상 고용된 모든 근로자(일용직 제외)를 포함한다. 또한 적용 연령은 65세 미만(고용안정・직업능력개발사업은 65세 이상자도 해당)이다(4대사회보험정보연계센터, 2006).

고용보험의 피보험자란 고용보험 적용사업에 고용되는 근로자로서 고용보험에 가입된 근로자(일용직 제외)를 말한다. 이는 본 연구를 위해 사용되는 고용보험 자료가 경활조사와는 달리 자영업, 무급가족종사자, 일용근로자 등을 제외한 임금근로자만 포함하는 한계를 갖는다는 것을 의미한다. 또한 법인의 대표이사 및 개인사업체의 대표자는 사업주로 근로자가 아니므로 피보험자에 해당하지 않고, ① 65세 이상인 자(다만, 고용안정・직업능력개발사업 제외), ② 소정근로시간이 1월간 60시간(1주간 15시간) 미만인 자(다만, 일용근로자이거나 생업을 목적으로 근로를 제공하는 자 중 3개월 이상 계속하여 근로를 제공하는 자는 제외), ③ 국가공무원법 및 지방공무원법에 의한 공무원, ④ 사립학교교직원연금법의 적용을 받는 자 등도 제외된다(4대사회보험정보연계센터, 2006). 참고로 2005년 고용보험 적용자는 전체 대상근로자의 72.1%로 2006년 임금노동자를 기준으로 고용보험 적용비율은 전체 58.1%였다(권혜자와 노현국, 2007).

고용보험 자료를 사전 검토한 결과, 연구에 사용할 수 있는 자료는 크게 사업장과 피보험자 자료 형태로 대별된다. 이 중 피보험자 정보가 본 연구의 관심 사항이므로 노동부에 요청한 자료는 <표 2-1>과 같이 2005년 1월 ~ 2007년 12월(3년)간 고용보험의 시군구별 피보험자 및 실

업급여 수급자 자료이다.

〈표 2-1〉 고용보험 피보험자 및 실업급여 수급자 자료

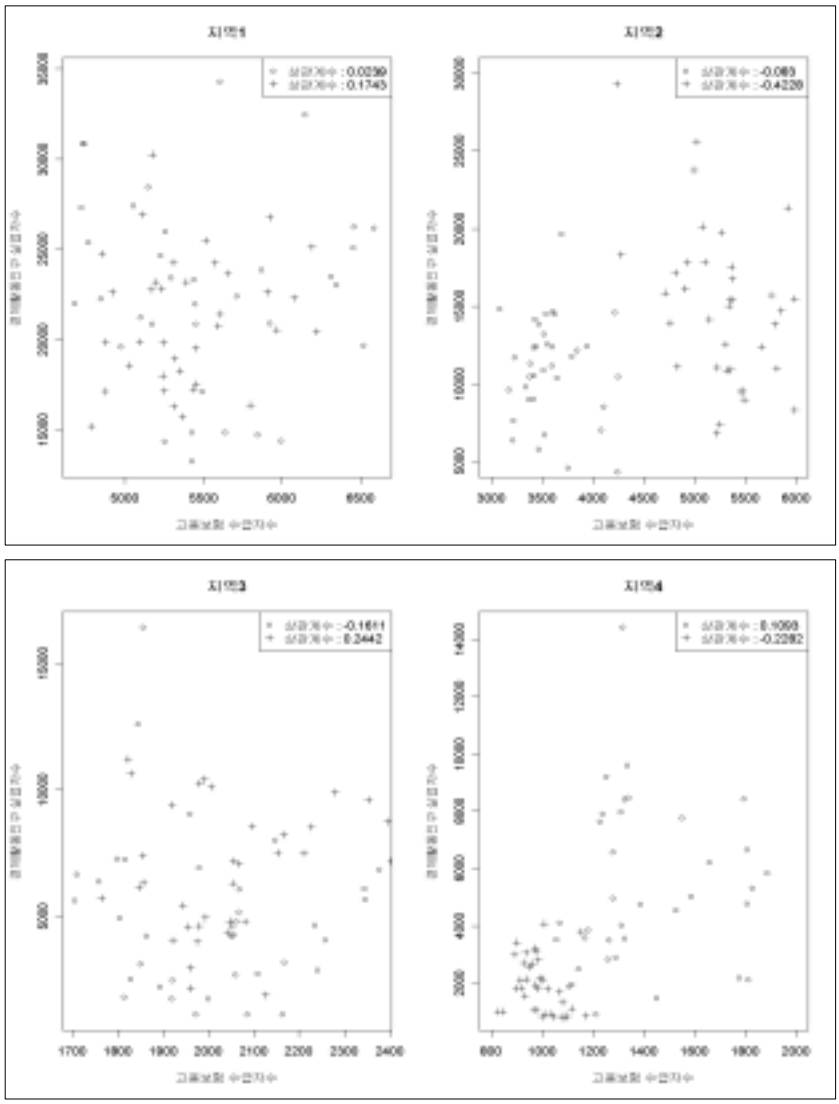
피보험자	실업급여 수급자
· 총피보험자 수 · 취득피보험자 수 · 상실피보험자 수 · 신규취득피보험자 수 · 경력취득피보험자 수 · 일용직피보험자 수 · 기타피보험자 수	· 수급자 수

주 : 시군구별, 성별, 연령별 자료

자료 검토 결과 고용보험의 사업장 및 피보험자 관련 자료는 주소가 사업장 소재지 중심으로 작성되어 있어, 지역별 경활조사 실업률의 보조지표로 사용되기 어려웠다. 따라서 수급자의 주소로 집계되어 있는 실업급여 수급자 자료를 보조정보로 이용하기로 하였다. 실업급여 수급자의 조건은 고용보험 사업장에서 이직일 이전 18개월 동안에 6개월 이상 근무하고, 회사의 경영사정과 관련하여 비자발적으로 이직한 경우로 한정된다. 이는 경활조사는 가구 표본조사에 의한 실업 현황을 보여주는 반면, 고용보험의 실업은 '고용보험 통계에서 포착된 비자발적 실업' 현황임을 보여준다. 즉, 고용보험의 실업급여 수급자 정보는 다양한 실업자들과 실업 유형들 중 임금근로자의 일부분만 설명할 수 있다는 것이다(권혜자와 노현국, 2007).

한편, 실업급여 수급자 수가 소지역 추정에서 보조정보로서 가치가 있으려면 경활조사의 실업자 수와 어느 정도의 상관성이 보장되어야 한다. 이는 우리가 사용하는 모형이 회귀모형 형태로 종속변수와 설명변수와의 상관성이 추정의 정확성을 높이는 중요한 관건이 되기 때문이다. 실업급여 수급자 수와 경활조사의 실업자 수 사이의 상관관계를 분석해 보았다. 그 결과 전국 수준에서는 두 변수 간의 상관관계가 상당히 높은(상관계수=0.966) 반면, 각 지역(광역시·도 또는 시군구)의 하부단

위에서의 두 변수 간 상관성은 약해지고 그 경향도 양과 음의 방향으로 일관성이 없었다.



[그림 2-2] 지역별 경제활동인구 실업자 수와 실업급여 수급자 수

[그림 2-2]는 31개 시군구 중 일부(4개) 지역의 경활조사의 실업자 수와 실업급여 수급자 수 간의 산점도를 나타낸다. 이 그림에서 지역1~4는 지역 번호가 클수록 인구 규모는 작다. 이때 상관계수의 범위는 0.5에서 0.05로 그 값이 작고, 이는 경활조사의 실업자 수와 고용보험 자료의 실업급여 수급자 수 사이의 상관관계가 약하다는 것을 의미한다. 이런 의미에서 본다면 실업급여 수급자 수는 소지역 추정을 위한 보조정보로서 큰 의미가 없어 보인다. 이는 분석에 사용된 자료가 3개년 자료에 불과하고, 지역별 실업급여 수급자들의 특성에 따라 다른 양태를 보일 수 있다는 것도 원인이 될 수 있다. 또한 2005년 이후부터 고용보험 적용 범위가 현재의 범위로 확대되고 자료 체계가 정비되었기 때문이다. 이런 여러 가지 한계에도 불구하고 고용보험 자료가 계속해서 보완되고 있다는 측면에서 소지역 추정을 위한 보조정보로서 지속적인 검토가 있어야 하며 더 나은 행정자료를 확보하기 위한 노력도 추가되어야 할 것이다.

2) 공간정보

또 하나 고려된 보조 정보는 공간정보(대지역 실업 정보)이다. 이는 해당 시군구를 포함한 상위 지역, 즉, 광역시·도의 실업자 수를 이용하는 것이다. 이 경우에 사용 가능한 보조정보는 다음과 같이 계산될 수 있다.

$$\text{보조정보} = \frac{\text{소지역의 추계인구}}{\text{대지역의 추계인구}} \times \text{대지역의 추정 실업자수}$$

이는 대지역의 고용동향과 소지역의 고용동향이 강한 상관관계를 가질 것이라는 가정 하에 이용한 정보이다. 이 경우는 소지역의 고용특성이 대지역의 특성과 동일하다고 가정하기 때문에 그렇지 못할 경우에는 추정치에 큰 편향이 발생할 수 있다는 점에도 주의해야 한다. 이 경우에도 소지역이 포함된 대지역의 정보를 빌려다 쓴다는 점에서 자기상관이 높아진다는 점과 대지역의 추정 실업자 수 정보가 표본조사 자료라는 측면에서 분산이 과소하게 계산될 수 있다는 한계가 있다.

그럼에도 불구하고, 모형 연구를 위한 보조정보 탐색이라는 측면을 강조하여 이 두 변수를 모두 연구에 적용하기로 하였고 두 결과를 비교해 보고자 하였다.

2. 추정방법 및 결과 평가방법

여기서는 Fay-Herriot 모형과 Rao-Yu 모형에 대한 추정과 평가방법이 조금씩 다르고 다소 복잡하기 때문에 각각 설명하기로 한다. 본 연구는 2개의 추정량(EBLUP, HB)을 포함한 두 개 모형의 적용 결과에 대해서 직접추정치와 지역통계와의 상대적 비교를 시도하였다. 이 때문에 모든 추정결과들의 동시 비교는 이해하기가 상당히 어려울 것으로 예상된다. 따라서 복잡한 비교 결과를 보다 효과적으로 설명하기 위해 단계적으로 비교하기로 한다.

STEP1. 우선 Fay-Herriot 모형에 의해 EBLUP과 HB 추정량을 적용하고, 이 결과를 직접추정치와 지역통계자료에 비교해서 각 추정량의 우수성을 평가한다. 이때 평가측도로는 오차율(CV)과 상대편향(ARB) 등을 이용한다.

STEP2. STEP1에서 선택된 추정량을 Rao-Yu 모형에도 적용한다. 본 연구에서는 STEP1에서 HB가 선택되었다. 또한 많은 선행연구들도 HB가 EBLUP(또는 EB)보다 더 우수한 추정량이라는 이론적, 수치적 증명을 하였기 때문에 Rao-Yu 모형에서 EBLUP을 비교하지 않아도 크게 무리가 없을 것으로 생각된다.

STEP3. 모형 간 비교를 시도한다. 즉, STEP1의 Fay-Herriot 모형에 의한 HB(HB1)와 STEP2의 Rao-Yu 모형에 의한 HB(HB2)를 비교 분석하고, 최종적인 결론을 얻는다.

가. Fay-Herriot 모형에 의한 EBUP과 HB

제3절에서 소개한 EBLUP 추정량(θ_i^{EBLUP})과 HB 추정량(θ_i^{HB1})을 경활조사 자료에 적용해 보았다. 추정에 이용한 관측 자료는 2005년 1월부터 2007년 12월까지 3년간의 31개 시군구 월별 실업자 수이다. 추정과

평가 체계는 다음과 같다.

1) 추정방법

◦ 식(2.1)의 Fay-Herriot 모형에서 $b_i = 1$로 세팅한다.
◦ 시군구별 표집오차 ψ_i는 잭나이프방법으로 추정한다.
◦ EBLUP 추정 : 지역분산 σ_ν^2을 최우추정법으로 추정할 때 비선형 최적화를 위해 $\widehat{\sigma_\nu^2} = \exp(\tau)$로 변환해서 준뉴튼법(BFGS 공식)을 사용한다(3.3절 참조).
◦ HB1 추정 : 조건부확률분포로부터 깁스 샘플러에 의해 몬테 칼로 표본을 생성한다. 이때, $L = 10,\ d + D = 2d = 2000$으로 하고, 각 마코프 체인 내에서 생성된 2000개의 난수 가운데 처음 1000개(d)는 버리고 추정치 계산에는 이후 1000개(D)의 난수만을 이용한다.
◦ MCMC 방법에서 미지의 모수에 대한 초기치는 각각 $\beta, \sigma_\nu^2 \sim N(0,1),\ \theta_i \sim G(0.01, 0.01)$로부터 얻는다.
◦ 회귀 부분에 대한 보조정보로는 고용보험의 시군구별 월별 실업급여 수급자 수를 이용한다(추가 분석에서는 공간 정보도 이용).

2) 평가방법

비추정치(직접추정치), EBLUP과 HB1 각 추정량의 추정치에 대해서 다음과 같은 방법으로 비교하고 결과를 고찰한다.

① 벤치마크[1](gold standard로서 지역통계를 의미)와의 비교(ARB)
② 오차율 비교(CV 사용)
③ 특히, HB 추정량의 경우 마코프 체인의 수렴성(R통계량)과 모형 적합성(p값)을 ①, ②에 우선하여 확인

벤치마크가 참값에 근사하다고 가정하고 벤치마크와의 거리는

1) 지역통계 결과는 표본조사 결과로 엄밀한 의미에서 “Gold Standar”는 아니다. 센서스와 같은 전수형태의 비교 자료가 필요하지만, 센서스는 조사시점과 조사방법 및 조사원 등의 경활조사와는 다른 측면이 있기 때문에 지역통계 자료를 이용한 상대적 비교를 시도하였다.

ARB(Absolute Relative Bias: 상대편향)로 측정한다. ARB는 다음과 같이 정의된다.

$$ARB = \frac{|\theta_i^E - \theta_i^T|}{\theta_i^T} \tag{2.23}$$

식(2.23)에서 θ_i^T는 벤치마크 값, θ_i^E는 소지역에서의 EBLUP과 HB1추정치를 각각 의미한다. 이 식에 의해 벤치마크가 되는 2007년 9월과 12월 실업자에 대한 EBLUP, HB1추정량과 직접추정량 중 어느 쪽이 벤치마크에 가까운가를 비교한다.

나. Time series & Cross-sectional 모형에 의한 HB

1) 추정방법

Rao-Yu 모형과 MCMC 방법을 적용해서 시군구 월별 실업자를 추정한다. 추정절차는 다음과 같다.

① 식(2.17)의 조건부사후분포로부터 깁스 샘플러에 의해 몬테 칼로 표본을 생성한다. 이때 $L = 10,\ 2d = 2000$으로 설정하고, 각 마코프 체인 내에서 생성된 2000개의 난수 가운데 처음 1000개는 burn in(d) 상태로 처리하고 이후 1000개(D)만을 추정치 계산에 사용한다.

② 시군구별 표집오차 ψ_i 및 표집오차의 자기상관 ρ_{ik}를 추정한다. 표집오차 추정은 잭나이프방법, 자기상관은 R의 패키지 'acf'에서 제공하는 결과를 이용한다. 36개월 자료에 대해 자기상관을 검토한 결과, 자기상관 그래프([부그림 2-1])을 보면 대부분의 지역에서 24개월 이후에는 상관계수가 거의 0 근처로 수렴하는 경향을 보인다. 따라서 이 모형에서는 24개월의 자료를 적용해서 추정하기로 하였다.

③ 회귀부분의 보조정보는 고용보험 자료의 시군구내 월별 실업급여 수급자 수를 이용한다.

④ MCMC 방법에 의한 미지의 모수에 대한 초기치는 다음과 같이 정한다. 이때 초기치가 광범위하게 흩어지도록 잡아 추정치가 초기

치에 의존하지 않도록 했다.

◦ $\boldsymbol{\beta}, \sigma_\nu^2, \sigma_e^2, \cdots\ U(0,1000)$: 균일분포에서 초기치 생성
◦ $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_{31}) \cdots\ U(0,1000)$: 균일분포에서 초기치 생성
◦ $\boldsymbol{u} = (\boldsymbol{u_1}, \cdots, \boldsymbol{u_{31}}) \cdots\ N(0,1000)$: 정규분포에서 초기치 생성.

2) 평가방법

AR(1) 과정의 u의 계수로, $\rho = 1, 0.75, 0.5$를 고려하여 이들 3개의 모형을 비교하였다. $\rho = 1$인 경우에는 u가 랜덤워크(random work)를 따른다고 가정하는 것과 같아진다. 모든 모형을 동시에 추정하게 되면 분석량이 많아지기 때문에 다음과 같은 순서로 추정에 이용할 모형을 하나씩 시험해 보기로 하였다.

STEP1. 식(2.20)의 추정량 $\widehat{R_{i,24}}$에 의해 마코프 체인의 정상 분포로의 수렴 상태를 우선적으로 평가한다. 본 분석에서는 $\widehat{R_{i,24}}$을 소수 넷째 자리에서 반올림해서 해당 시군구에서 통계량값이 1로 근사하는 경우에 정상분포에 수렴하는 것으로 판단한다(1 근사의 정도는 1.5까지 인정).

STEP2. STEP1에서 선택된 모형에 대해서 식(2.21)의 통계량 $d(\mathbf{y}^*, \mathbf{y_{obs}})$과 식(2.22)의 통계량 $\hat{p}$를 사용하여 각 추정량의 우월성을 비교한다. 이렇게 해서 선정된 Rao-Yu 모형의 HB2와 직접추정치를 벤치마크(지역통계 조사의 실업자 수)인 2007년 9월과 12월의 지역통계 조사의 실업자 수와 비교해 벤치마크에 근사한가를 비교한다. 이를 위해 상대편향 및 오차율을 계산한다. 아래 식에서 θ_i^{HB2}는 HB2 추정치를 의미한다.

$$ARB = \frac{|\theta_i^{HB2} - \theta_i^T|}{\theta_i^T}$$

MCMC 방법의 프로그램 작성은 R을 이용했다. 그러나 R은 interpreter 언어이고 MCMC처럼 루프가 많을 경우 계산이 느려지는 경향이 있다. Rao-Yu 모형과 같은 복잡한 모형의 경우, 추정에 걸리는 시

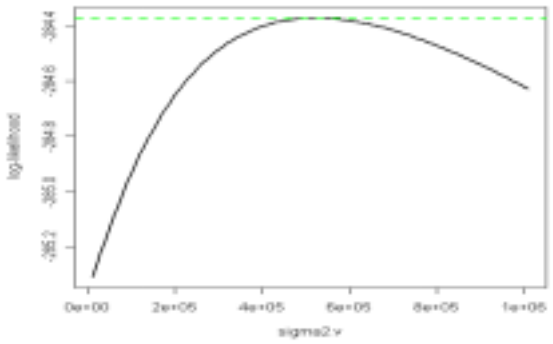
간이 더 길어진다. 본 연구에서처럼 $L=10$, $2d=2000$으로 할 경우, 1회 추정에 약 40분 정도 시간이 소요되었다.

제6절 최종 결과 비교 분석

1. Fay-Herriot 모형의 EBLUP과 HB 비교

가. HB에 대한 정상분포 수렴성 및 모형적합성 검정

① EBLUP의 경우, [그림 2-3]은 2007년 1월의 σ_ν^2에 대한 우도함수의 잠정 형태를 나타낸 것이다. 회귀모형의 설명변수로는 같은 시기의 시군구별 실업급여 수급자 수가 각각 사용되었다. 우도함수는 상당히 급한 곡선이고 σ_ν^2은 512,455 근처에서 최대치가 됨을 알 수 있다. 스코어법이나 준뉴튼법 모두 유사한 값을 찾았다.



[그림 2-3] 우도함수 그래프

② HB의 경우, 마코프 체인의 수렴성 여부를 판단한다. <표 2-2>와 <표 2-3>의 마지막 열인 모형수렴부분에서 HB1의 R값을 보면, 분석에 사용된 9월과 12월 자료의 경우 31개 모든 시군구에서 R값이 거의 1에 근사하게 되므로, 마코프 체인이 비교적 정상분포에 잘 수렴한다고 말할 수 있다.

③ 그런데 각 표의 마지막 행의 모형적합 p값을 보면, 9월 자료에서는 0.229, 12월에서는 0.301로서 절대적 기준을 적용할 경우에는 모형적합이 그다지 좋다고 볼 수 없다. 이와 같은 원인으로는 관측자료와 보조자료 간의 상관관계가 낮거나 추정에 사용되는 실업자 수의 범위가 100~10000단위 값으로 상당히 넓게 분포하기 때문일 수 있다. 자료의 범위가 너무 광범위할 경우, 전체적인 경향(회귀추정선)에서 벗어나는 값이 발생할 가능성이 높기 때문에 모형추정에 영향을 미칠 수 있다. 이와 같은 문제는 자료의 변환이나 비율 자료를 활용함으로써 어느 정도는 극복할 수 있다.

나. 각 추정치의 비교

<표 2-2>와 <표 2-3>과 [그림 2-4]~[그림 2-6]를 이용하여 추정치를 비교한다. <표 2-2>와 <표 2-3>은 2007년 9월과 12월의 실업자 수에 대한 각각의 추정결과를 정리한 것이다. [그림 2-4]~[그림 2-6]은 추정치, ARB, 오차율을 각각 나타낸다. 이 결과로부터 다음의 사실들을 알 수 있다.

① 직접추정치와 Fay-Herriot 모형에 의한 EBLUP 및 HB1 추정치를 직접 비교한다. 모형추정치 EBLUP과 HB1은 모두 직접추정치와는 차이가 있고, EBLUP과 HB1 간에는 유사한 추정치를 갖는 것으로 보인다. [그림 2-4]를 보면 해당 월에 상관없이 전체적으로 EBLUP과 HB1은 직접추정치보다 낮게 추정되는 경향이 있다. 또한 대부분의 경기도 지역에서의 지역통계 값은 직접추정치나 모형추정치에 비해 상당히 높은 실업자 수를 나타내고 있다. 이는 경기도 지역이 벤치마크 인구로 추계인구가 아닌 주민등록인구를 사용했다는 점에서 다른 지역들과 보이는 차이일 수 있다.

② ARB를 비교해보자. 직접추정치보다 EBLUP 또는 HB1의 ARB가 큰 셀에는 <표 2-2>와 <표 2-3>에서 진하게 표시해 두었다. 9월 자료의 경우는 직접추정치와 모형추정치들 모두 벤치마크와 유사하게 나타났고, 12월 자료에서는 모형추정치들이 직접추정치보다 벤치마크(지역통계 값)에 유사한 지역이 약간 많았다(31개 중 18개 지역). [그림 2-5]를

보면, 특히 12월에는 경활조사의 표본조사구 수가 4개 이하인 지역(파주, 김포, 여주, 태백, 진해, 통영, 사천, 밀양) 가운데 4개 지역(파주, 김포, 태백, 진해)에서 직접추정치의 ARB가 세 추정치 중 가장 크게 나타났고, 나머지 3개 지역에서 모형추정치가 더 낮거나 비슷한 값을 보였다.

그리고 표본조사구수가 15개 이상으로 상대적으로 많은 지역(수원, 성남, 전주, 창원)에서도 대체로 직접추정치의 ARB가 크게 나타났다. 9월 자료에서는 전주, 창원에서조차 직접추정치의 ARB가 상대적으로 크고, 12월의 경우는 표본 수가 많은 4개 지역 가운데 수원, 성남, 창원 등 세 지역에서 월등히 크게 나타났다.

다시, EBLUP과 HB1만을 비교하면 HB1의 ARB가 시기에 상관없이 대체로 낮은 지역들이 많고, 표본 수가 많은 지역뿐만 아니라 적은 지역에서조차도 HB1의 ARB가 EBLUP보다 더 작은 값을 갖는다.

③ 오차율(CV)을 비교해 보자. 전체적으로 직접추정치의 오차율이 모형추정치에 비해 매우 크다. 그러나 표본 수가 매우 적은 지역들에서는 모형추정치의 오차율이 큰 경우도 있다. 예를 들면 12월에는 통영, 사천, 밀양은 직접추정치보다 오차율이 크게 나타나고, 9월에도 사천에서는 HB1이 크게 나타난다. 한편, 지역통계 값과 비교해볼 때, 모형추정치의 오차율은 표본 수가 6개 이하로 적은 지역을 제외하고는 모두 30% 이내로 표본 수가 안정적인 지역통계 값에 견줄만하다고 할 수 있다.

다시 EBLUP과 HB1만의 오차율을 비교한다. 전체적으로 EBLUP의 오차율이 가장 낮게 추정된다. 특이한 사실은 오차율 추정에 있어서 여주를 제외한 모든 경기도 지역들은 EBLUP의 오차율에 거의 변화가 없다. 그러나 이론적 결과에 따르면 EBLUP은 HB1과 추정치 면에서는 유사하지만 분산이 과소하게 추정되는 경향이 있고, 이런 면에서 볼 때 우리의 결과 또한 그리 놀라운 사실은 아닌 듯하다.

〈표 2-2〉 2007년 9월 자료에 대한 각 추정량 비교, $\rho = 0.5$: 추정치, CV, ARB, R, P값

	조사구수(개)		실업자 수(명)					CV(%)					ARB(%)				모형수렴R	
	지역	직접	지역	직접	FH		TSCS	지역	직접	FH		TSCS	직접	FH		TSCS	FH	TSCS
	통계	추정량	통계	추정량	EBLUP	HB1	HB2	통계	추정량	EBLUP	HB1	HB2	추정량	EBLUP	HB1	HB2	HB1	HB2
수원	40	16	17,000	22,383	11,668	12,603	13,740	18.00	21.97	11.03	14.45	5.79	31.66	31.37	25.87	19.17	1.049	1.341
성남	49	15	20,498	18,561	10,208	10,898	12,066	13.40	36.15	10.90	14.68	7.46	9.45	**50.20**	**46.83**	**41.14**	1.041	1.282
의정부	38	6	8,530	1,098	3,722	3,049	2,442	23.10	101.06	9.83	27.59	22.24	87.13	56.36	64.26	71.38	1.009	1.186
안양	50	10	11,460	10,450	6,716	7,188	7,987	15.20	44.22	10.42	16.48	10.73	8.82	**41.40**	**37.28**	**30.31**	1.035	1.356
부천	50	13	17,179	7,434	10,684	10,970	11,783	18.20	40.26	10.94	13.28	6.86	56.73	37.81	36.14	31.41	1.044	1.241
광명	31	6	7,632	4,038	3,616	3,816	4,177	19.00	69.79	9.83	23.29	15.96	47.09	**52.62**	**50.00**	45.27	1.012	1.091
평택	40	8	13,261	2,687	3,327	3,366	3,625	14.60	71.35	9.82	23.80	18.92	79.74	74.91	74.61	72.67	1.008	1.282
안산	33	9	18,135	12,634	8,581	9,182	10,090	17.10	33.58	10.71	14.94	8.90	30.33	**52.68**	**49.37**	**44.36**	1.040	1.480
고양	40	13	13,682	18,744	7,714	8,319	9,008	21.30	34.86	10.59	16.12	10.28	37.00	43.62	39.20	34.16	1.032	1.365
남양주	40	11	-	2,635	4,192	4,025	4,015	-	57.30	9.90	19.82	17.22	-	-	-	-	1.014	1.416
시흥	40	5	7,095	3,107	4,205	4,108	3,416	17.90	45.78	9.91	18.75	24.15	56.21	40.74	42.09	51.86	1.014	1.267
용인	32	10	16,505	23,201	7,534	8,288	9,141	16.40	24.49	10.56	16.84	8.77	40.57	**54.36**	**49.79**	**44.62**	1.037	1.161
파주	31	4	6,690	7,756	2,826	3,424	3,373	21.50	32.43	9.93	29.19	23.84	15.93	**57.75**	**48.82**	**49.58**	1.017	1.178
이천	30	5	-	2,084	982	1,191	1,428	-	86.08	20.25	66.37	50.42	-	-	-	-	1.003	1.031
안성	30	5	2,915	933	1,001	998	962	18.90	101.82	19.83	59.96	58.42	68.00	65.65	65.76	67.00	1.004	1.208
김포	31	4	4,536	856	1,655	1,401	1,411	21.70	106.76	12.42	43.50	37.11	81.12	63.50	69.11	68.89	1.001	1.111
화성	30	6	7,956	4,203	3,085	3,299	3,530	17.30	78.23	9.85	27.22	24.21	47.18	**61.22**	**58.54**	**55.63**	1.011	1.202

<표 2-2>의 계속

	조사구수(개)		실업자 수(명)					CV(%)					상대편향(%)				모형수렴R	
	지역 통계	직접 추정량	지역 통계	직접 추정량	FH		TSCS	지역 통계	직접 추정량	FH		TSCS	직접 추정량	FH		TSCS	FH	TSCS
					EBLUP	HB1	HB2			EBLUP	HB1	HB2		EBLUP	HB1	HB2	HB1	HB2
여주	30	4	1,487	1,432	118	240	707	25.10	174.03	200.28	361.11	87.42	3.72	**92.08**	**83.85**	**52.42**	1.004	1.048
태백	25	3	321	2,480	90	353	475	68.70	91.80	263.57	249.82	137.33	150.00	71.92	9.87	48.07	1.005	1.023
전주	46	22	5,791	4,202	5,577	5,588	5,818	22.90	43.63	10.19	16.05	10.87	27.44	3.69	3.51	0.47	1.020	1.396
광양	46	6	1,100	950	600	769	982	30.10	74.56	34.92	67.42	57.82	13.63	**45.45**	**30.11**	10.74	1.004	1.123
창원	45	16	4,493	7,030	5,363	5,791	6,302	21.60	31.80	10.15	16.86	11.01	56.46	19.36	28.88	40.26	1.032	1.655
마산	60	13	5,750	5,697	4,823	5,130	5,259	19.10	41.54	10.03	18.11	15.06	0.93	**16.12**	**10.79**	**8.54**	1.025	1.364
진주	55	10	4,561	3,316	2,652	2,873	2,825	18.10	48.06	10.04	26.45	28.53	27.30	**41.85**	**37.00**	**38.07**	1.008	1.189
진해	39	4	1,639	1,790	1,229	1,363	1,641	25.90	97.89	16.07	56.48	43.76	9.22	**25.03**	**16.84**	0.10	1.002	1.173
통영	50	4	1,756	413	438	426	506	20.40	95.61	49.42	80.51	72.96	76.47	75.03	75.76	71.21	1.004	1.152
사천	50	4	1,776	2,010	368	579	621	27.60	106.07	59.88	145.89	134.81	13.19	**79.26**	**67.40**	**65.03**	1.005	1.170
김해	55	10	5,293	4,044	5,668	5,710	6,029	20.40	52.75	10.21	16.39	10.94	23.59	7.09	7.88	13.91	1.021	1.499
거제	50	6	1,565	733	1,082	970	1,050	25.20	119.10	18.28	59.52	50.19	53.17	30.86	38.01	32.89	1.001	1.032
밀양	50	4	2,122	360	332	355	483	27.40	95.51	66.88	86.56	66.03	83.05	**84.34**	**83.27**	77.22	1.004	1.048
양산	50	6	3,704	6,806	2,858	3,233	3,845	18.20	48.36	9.92	29	21.82	83.75	22.84	12.71	3.80	1.011	1.079
															모형적합		0.229	0.92

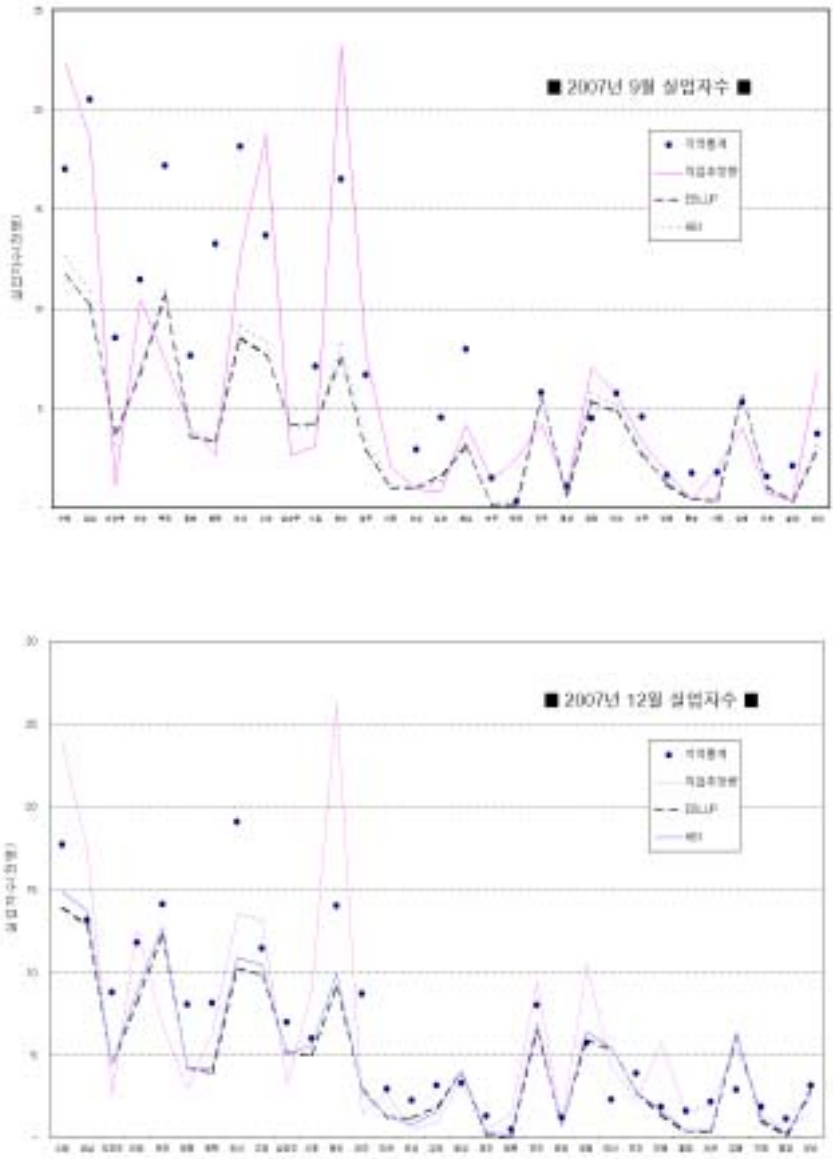
주: FH : Fay-Herriot 모형, TSCS : Time series & Cross-sectional 모형

〈표 2-3〉 2007년 12월 자료에 대한 각 추정량 비교, $\rho = 0.5$: 추정치, CV, ARB, R, P값

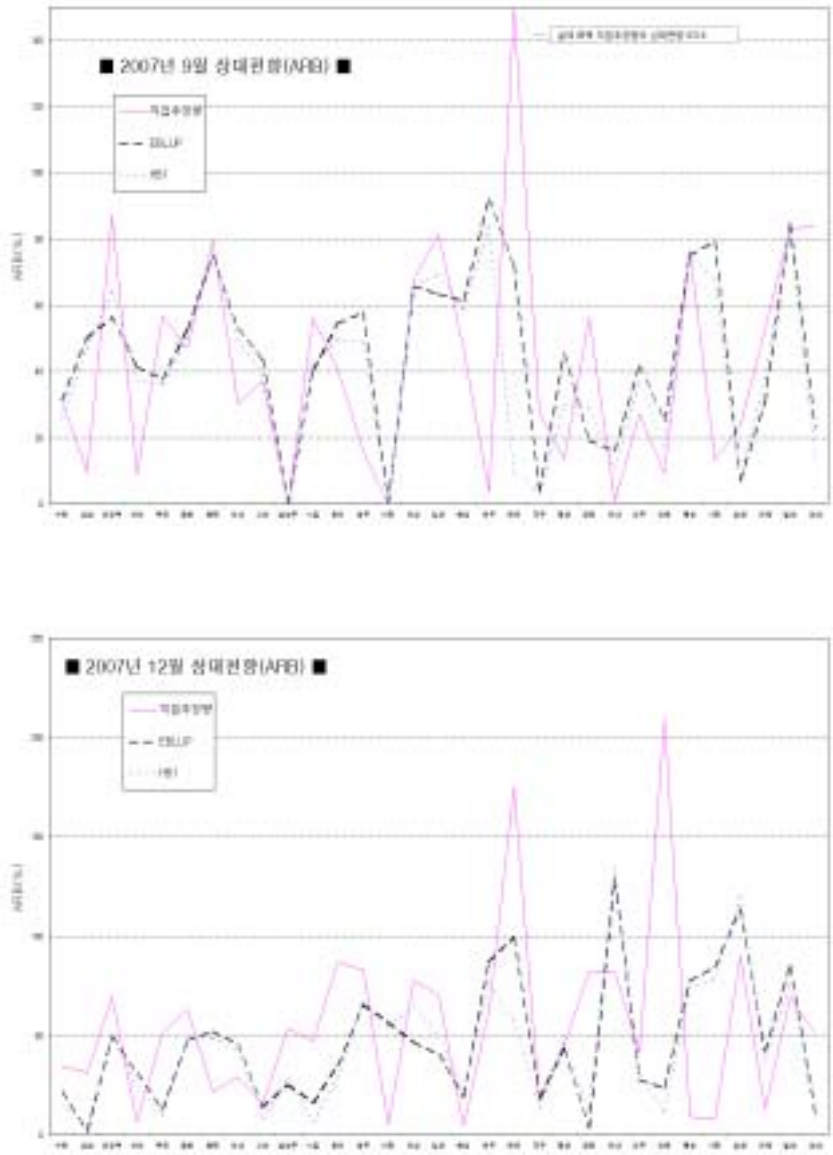
	조사구수(개)		실업자 수(명)					CV(%)					상대편향(%)				모형수렴R	
	지역 통계	직접 추정량	지역 통계	직접 추정량	FH		TSCS	지역 통계	직접 추정량	FH		TSCS	직접 추정량	FH		TSCS	FH	TSCS
					EBLUP	HB1	HB2			EBLUP	HB1	HB2		EBLUP	HB1	HB2	HB1	HB2
수원	40	16	17,713	23,867	13,929	14,881	13,673	21.40	29.90	10.35	12.44	5.81	34.74	21.37	15.99	22.81	1.084	1.354
성남	49	15	13,161	17,279	12,862	13,705	12,920	16.90	30.58	10.30	12.36	5.37	31.29	2.27	4.13	1.83	1.081	1.243
의정부	38	6	8,807	2,677	4,478	4,406	3,716	16.10	71.22	9.39	18.33	18.69	69.60	49.16	49.97	57.81	1.021	1.393
안양	50	10	11,783	12,550	8,175	8,876	8,180	17.30	23.15	9.90	13.57	9.30	6.51	**30.62**	**24.67**	**30.58**	1.065	1.358
부천	50	13	14,119	6,833	12,313	12,711	11,231	15.20	49.76	10.26	11.87	6.10	51.61	12.79	9.97	20.45	1.076	1.171
광명	31	6	8,061	2,958	4,232	4,221	3,960	18.70	65.34	9.36	18.76	14.19	63.31	47.50	47.63	50.88	1.025	1.013
평택	40	8	8,145	6,356	3,912	4,190	3,767	15.30	54.86	9.34	21.09	21.68	21.96	**51.97**	**48.56**	**53.75**	1.021	1.153
안산	33	9	19,052	13,489	10,242	10,897	10,111	17.00	36.42	10.11	12.83	8.03	29.20	**46.24**	**42.80**	**46.93**	1.071	1.346
고양	40	13	11,441	13,220	9,840	10,453	9,608	23.90	45.08	10.07	13.06	9.20	15.55	14.00	8.63	**16.02**	1.073	1.263
남양주	40	11	7,006	3,260	5,192	5,057	4,606	14.30	49.51	9.48	16.16	14.12	53.47	25.89	27.82	34.25	1.028	1.359
시흥	40	5	6,000	8,834	5,009	5,590	4,889	16.80	26.19	9.46	17.33	16.83	47.24	16.52	6.83	18.52	1.041	1.395
용인	32	10	14,003	26,194	9,110	9,963	9,506	19.10	21.22	10.00	14.20	10.43	87.06	34.94	28.85	32.11	1.071	1.343
파주	31	4	8,684	1,486	2,967	2,789	2,452	17.60	106.50	9.48	26.38	19.11	82.89	65.84	67.89	71.77	1.015	1.201
이천	30	5	2,937	2,795	1,255	1,328	1,675	26.40	121.65	15.48	62.54	37.26	4.85	**57.27**	**54.79**	**42.97**	1.001	1.052
안성	30	5	2,267	505	1,210	811	672	26.50	101.95	16.02	54.00	64.14	77.74	46.63	64.25	70.37	1.011	1.044
김포	31	4	3,179	945	1,892	1,600	1,378	34.60	100.67	11.16	38.00	36.27	70.27	40.47	49.67	56.65	1.010	1.094
화성	30	6	3,317	3,460	3,952	4,069	3,960	26.20	78.87	9.34	20.45	15.76	4.31	**19.14**	**22.68**	**19.38**	1.017	1.176

<표 2-3>의 계속

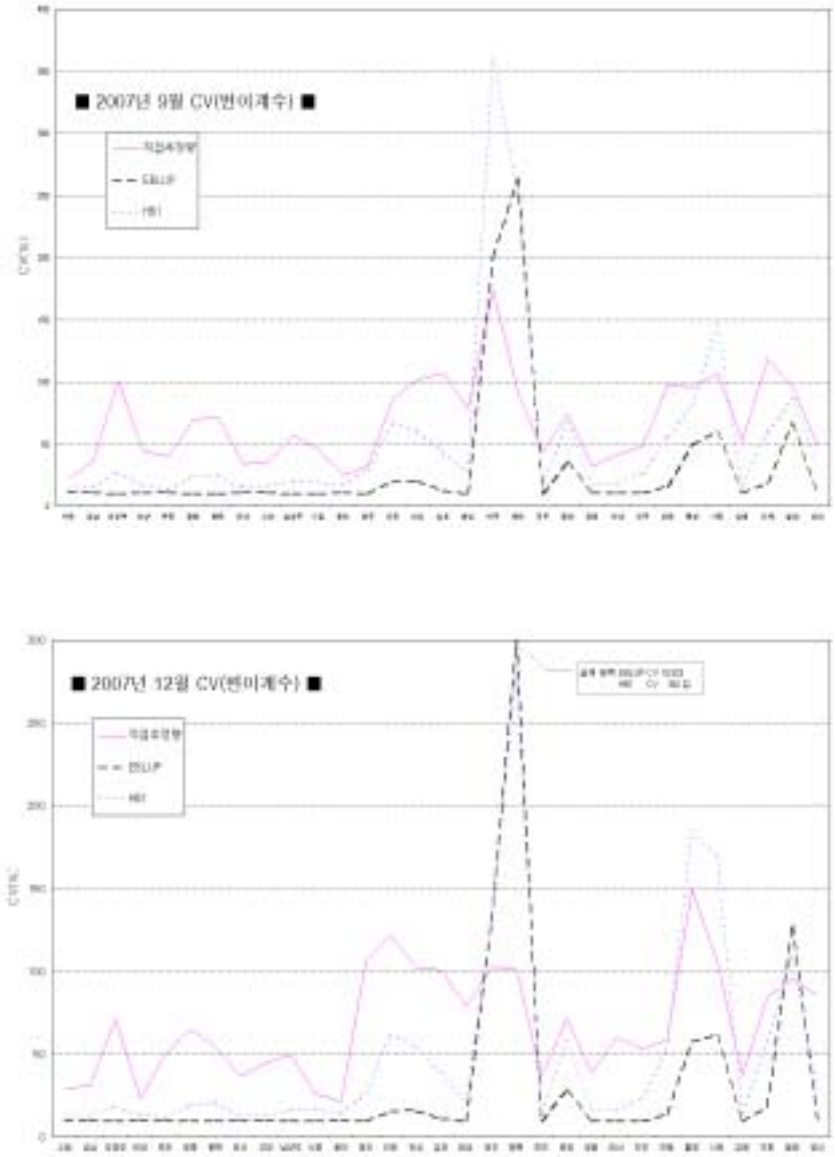
	조사구수(개)		실업자 수(명)					CV(%)					상대편향(%)				모형수렴R	
	지역 통계	직접 추정량	지역 통계	직접 추정량	FH		TSCS	지역 통계	직접 추정량	FH		TSCS	직접 추정량	FH		TSCS	FH	TSCS
					EBLUP	HB1	HB2			EBLUP	HB1	HB2		EBLUP	HB1	HB2	HB1	HB2
여주	30	4	1,342	504	171	326	501	30.00	102.96	129.76	130.50	92.20	62.45	**87.26**	**75.67**	**62.65**	1.004	1.173
태백	25	3	516	1,421	2	217	408	57.90	101.28	300.00	280.00	152.48	175.37	99.57	58.00	20.89	1.005	1.121
전주	46	22	8,035	9,394	6,473	6,935	6,235	23.50	36.58	9.68	14.99	10.83	16.92	**19.44**	13.70	**22.40**	1.047	1.263
광양	46	6	1,254	679	712	676	751	32.70	71.50	28.03	58.12	61.29	45.88	43.25	**46.08**	40.14	1.000	1.146
창원	45	16	5,763	10,500	5,964	6,425	5,933	26.10	38.38	9.60	16.02	13.88	82.19	3.49	11.48	2.94	1.038	1.196
마산	60	13	2,321	4,234	5,304	5,445	4,826	32.30	60.09	9.50	16.19	15.27	82.42	**128.53**	**134.59**	**107.95**	1.041	1.244
진주	55	10	3,919	2,296	2,833	2,764	2,553	17.50	53.36	9.56	23.73	23.84	41.42	27.71	29.47	34.86	1.004	1.170
진해	39	4	1,866	5,770	1,429	1,656	1,841	30.10	58.70	13.75	52.08	41.16	209.22	23.43	11.28	1.32	1.004	1.394
통영	50	4	1,634	1,479	369	430	759	20.00	149.62	57.44	184.56	93.76	9.46	**77.40**	**73.69**	**53.53**	1.004	1.059
사천	50	4	2,178	2,010	345	473	800	21.50	106.07	61.90	168.85	83.69	7.70	**84.18**	**78.27**	**63.26**	1.004	1.241
김해	55	10	2,914	5,536	6,220	6,407	5,921	26.20	37.24	9.64	14.04	12.80	89.99	**113.44**	**119.85**	**103.20**	1.044	1.349
거제	50	6	1,873	1,635	1,084	1,185	1,343	22.10	84.33	17.89	57.76	49.04	12.71	**42.13**	**36.75**	**28.28**	1.002	1.225
밀양	50	4	1,166	360	174	285	395	23.20	95.51	127.65	109.61	74.67	69.16	**85.11**	**75.57**	66.14	1.004	1.004
양산	50	6	3,177	1,552	2,855	2,633	2,330	19.90	85.20	9.54	27	20.37	51.14	10.13	17.11	26.66	1.011	1.247
															모형적합		0.301	0.90



[그림 2-4] FH 모형에 기반한 추정량 비교: 추정치



[그림 2-5] FH 모형에 기반한 추정량 비교: ARB



[그림 2-6] FH 모형에 기반한 추정량 비교: CV값

한 가지 주목할 만한 사실은 태백시의 분산추정이 모든 추정량에 있어서 매우 크다는 것이다. 그 이유로는 표본 수가 3개로 가장 작기 때문에 잭나이프에 의한 표본분산 추정이 다소 불안정할 수 있다는 점을 들 수 있다. 표본 수가 4개인 지역들에서도 표본 수가 많은 지역들에 비해 분산이 다소 크게 추정된다는 점도 이 지역의 분산추정의 문제를 뒷받침해 준다.

다. 요약

Fay-Herriot 모형에 의한 EBLUP과 HB1 비교 결과를 정리하면 다음과 같다. 우선 추정량의 모형 선택 면에서 보면, EBLUP 추정 시 MLE을 이용한 지역분산 σ_v^2을 추정하는 과정에서 월별 자료에 따라 MLE의 최대값 수렴이 다소 불안정한 경향이 있다. 어떤 월의 자료에서는 거의 초기값 부근에서 최대값을 찾는 현상이 발생하기도 한다. MLE의 수렴이 잘 안되는 이유는 표본 수가 적다거나 가정한 분포가 적절하지 않다거나 등의 이론적 근거가 있을 수 있다. 이 문제에 대해서는 추정방법을 MLE가 아닌 적률법과 같은 보다 간단한 방법을 적용해 보거나 현재 31개로 국한된 지역수를 더 추가해서 분석함으로써 해결될 수 있을 것으로 보인다. HB1 추정에서 MCMC 방법에 대해서는 R값이 거의 1 근처로 정상분포에 수렴하였다. 단 p값은 0.5보다 약간 작은 값으로 모형적합은 그렇게 좋은 편이 아님을 시사했다.

추정 결과 면에서 보면, 모형추정량이 직접추정량보다 벤치마크에 근사한 지역들이 많다. 대부분의 지역에서 HB1의 추정결과가 좋다고 말할 수 있다. 오차율도 직접추정치와 비교했을 때 HB1이 대부분의 지역에 작게 나타난다. 또한 지역통계와의 비교에서도 표본 수가 6개 이하로 적은 지역을 제외하고는 오차율이 30% 이내로 안정된 값을 보이는 것으로 나타났다. HB1이 직접추정치보다 좋지 않은 경우가 일부 표본 수가 작은 지역에서 나타나기도 하지만, HB1은 직접추정량보다는 오차율을 상당히 줄일 수 있고, EBLUP의 분산 과소 추정에도 대처할 수 있을 것으로 보인다.

2. Time series & Cross-sectional 모형의 HB 비교

가. 몬테 칼로 표본의 적정성 검정

MCMC 방법을 적용할 때 표본이 계획한 대로 잘 뽑혔는지를 확인하는 것이 추정과정에서 첫 번째 해야 할 일이다. 표본이 불안정하면 추정결과의 신뢰도가 떨어질 수 있고, 특히 시뮬레이션 연구인 경우 이들의 특성을 확인하는 절차는 반드시 필요하다.

이를 확인하기 위해 조건부사후분포에 MCMC 방법을 적용한 몬테 칼로 표본의 표본 경로를 살펴볼 수 있다. 관측치 자료는 2005년 1월부터 2007년 12월까지의 31개 시군구별 월별 실업자 수를 이용했다. 보조정보는 같은 시기의 실업급여 수급자 수를 사용했다. 또 시계열 효과 u의 계수는 0.5로 했다. multi run에 의한 몬테 칼로 표본을 생성하기 위해, $L=5,\ 2d=2000$으로 설정했다. 다음의 각 모수에 대해서 샘플 경로를 나타내고 현황을 파악하였다. MCMC 방법에 의한 알려지지 않은 모수의 초기치는 (5.2.나) 절에서 설명한 방법으로 설정했다. 최종 추정시점은 2007년 12월이다.

◦ 회귀계수 β

◦ 분산모수 σ_ν^2 및 σ_e^2

◦ 수원 $t=24$에 의한 u_{it} 및 $\theta(u_{1,24},\ \theta_{1,24})$

[부그림 2-2]는 이들 모수들의 샘플 경로이고, [부그림 2-3]은 히스토그램이다. 각 그래프에서 제목과 모수의 관계는 다음과 같다.

Beta(intercept) : 회귀모형의 상수(절편), Beta(coefficient): 계수(기울기)

sigma2.v: σ_ν^2, sigma2.e : σ_e^2, u : $u_{1,24}$, theta.it : $\theta_{1,24}$.

이 그래프에 의해 파악된 모수들의 특성은 다음과 같다.

◦ 회귀계수 β의 샘플경로는 회귀계수 및 상수 모두 평균 근처에서 분산되어 있는 것처럼 보인다. 5개의 체인은 적당히 서로 섞여 있어 각각의 마코프 체인이 전체 평균에 근사해 있음을 알 수 있다. 각각 샘플 경로의 자기상관은 높다고 볼 수 있다. 히스토그램을

보면 회귀계수의 분포는 약간 오른쪽으로 기울어져 있고, 상수의 분포는 봉우리 부분이 약간 완만한 패턴을 보이고 있다. 그러나 거의 정규분포에 가까운 형태를 하고 있다.

◦ 분산 모수 σ_ν^2 및 σ_e^2에 대해서는 처음에 매우 큰 값을 택했지만, 그 후 빠르게 안정 상태가 되었다. 단, σ_ν^2의 한 체인에서 반복횟수가 1000을 약간 넘는 지점에서 급격한 변동을 갖는다. 이에 대한 정확한 이유는 알 수 없지만 체인 수를 늘리면 다소 해결될 수 있을 것으로 본다. 히스토그램으로 보면 역감마분포의 형태를 잘 반영하고 있어 보인다. 각 샘플 경로에는 일부 주기적인 움직임이 보이지만, 자기상관이 약간 높은 것으로 생각된다.

◦ u 및 θ에 대해서는 5개의 체인이 매우 잘 섞여 있고, 초기치의 영향은 거의 받지 않고 있다. 자기상관도 꽤 작아 보인다. 안정된 정상분포에 거의 수렴한다고 볼 수 있다.

나. HB에 대한 정상분포 수렴성 및 모형적합성 검정

① 우선 마코프 체인의 수렴성 여부를 보자. <표 2-4>의 R값을 보면 $\rho=1$일 때는 어떤 시점에서도 R값이 크다. 많은 지역에서 2보다 넘는 값이 발생한다. 따라서 마코프 체인이 수렴한다고 볼 수 없다. 한편, $\rho=0.75$와 $\rho=0.5$인 경우 몇 개 지역을 제외하면 전체적으로 최대 1.5 근처의 값을 갖는다.

② d값을 비교하면 9월, 12월 모두에서 $\rho=0.5$ 일대가 가장 작다. 따라서 3개의 모형 중 $\rho=0.5$ 일 때가 가장 적합이 좋다고 말할 수 있다.

③ 단 p값을 보면 어떤 자료에서도 0.9에 가까운 값이 되므로 절대적 기준에서 본다면 모형적합은 그리 좋다고 볼 수 없다. p값이 이렇게 큰 값을 갖는 이유에 대해서는 Fay-Herriot 모형의 경우와 같이 관측치가 보조정보와 상관성이 낮다거나 특정지역에서 매우 크거나 작은 관측값이 존재하기 때문이라 볼 수 있다.

이상의 결과에 따라 연구자들은 $\rho=0.5$의 모형을 비교에 이용하기로 하였다.

〈표 2-4〉 2007년 12월, 9월 R값 비교

	2007.12			2007. 9		
	ρ=0.5	ρ=0.75	ρ=1	ρ=0.5	ρ=0.75	ρ=1
수원	1.341	1.680	1.964	1.354	1.980	1.732
성남	1.282	1.484	2.896	1.243	2.984	2.826
의정부	1.186	1.635	1.906	1.393	1.635	1.690
안양	1.356	1.789	2.093	1.358	1.989	3.944
부천	1.241	1.662	3.855	1.171	1.962	1.722
광명	1.091	1.509	2.879	1.013	1.609	3.214
평택	1.282	1.575	2.924	1.153	1.575	1.559
안산	1.480	1.498	1.610	1.346	1.798	2.547
고양	1.365	1.487	2.244	1.263	1.987	1.424
남양주	1.416	1.176	1.767	1.359	1.176	2.003
시흥	1.267	1.609	2.385	1.395	1.609	1.381
용인	1.161	1.261	2.125	1.343	1.261	1.263
파주	1.178	1.198	7.222	1.201	1.198	1.131
이천	1.031	1.833	6.667	1.052	1.833	3.139
안성	1.208	1.099	1.596	1.044	1.099	2.407
김포	1.111	1.244	1.528	1.094	1.244	1.225
화성	1.202	1.367	3.607	1.176	1.367	1.532
여주	1.048	1.156	2.086	1.173	1.156	1.005
태백	1.023	1.231	1.434	1.121	1.231	1.933
전주	1.396	1.586	2.604	1.263	1.586	1.274
광양	1.126	1.062	1.142	1.146	1.062	1.502
창원	**1.655**	1.262	6.530	1.196	1.262	1.625
마산	1.364	1.835	1.912	1.244	2.835	2.456
진주	1.189	1.127	1.741	1.170	1.127	2.495
진해	1.173	1.575	2.263	1.394	1.575	1.363
통영	1.152	1.397	1.846	1.059	1.397	1.208
사천	1.170	1.216	1.801	1.241	1.216	1.495
김해	**1.499**	1.677	3.368	1.349	1.677	2.884
거제	1.032	1.998	1.837	1.225	1.998	1.019
밀양	1.048	1.033	1.450	1.004	1.033	1.024
양산	1.079	1.396	2.136	1.247	1.396	2.452

	2007년 12월			2007년 9월		
ρ	0.5	0.75	1	0.5	0.75	1
p	0.90	0.97	0.99	0.92	1.0	0.99
d	27459795	27535174	31380092	2706606	27917576	28131039

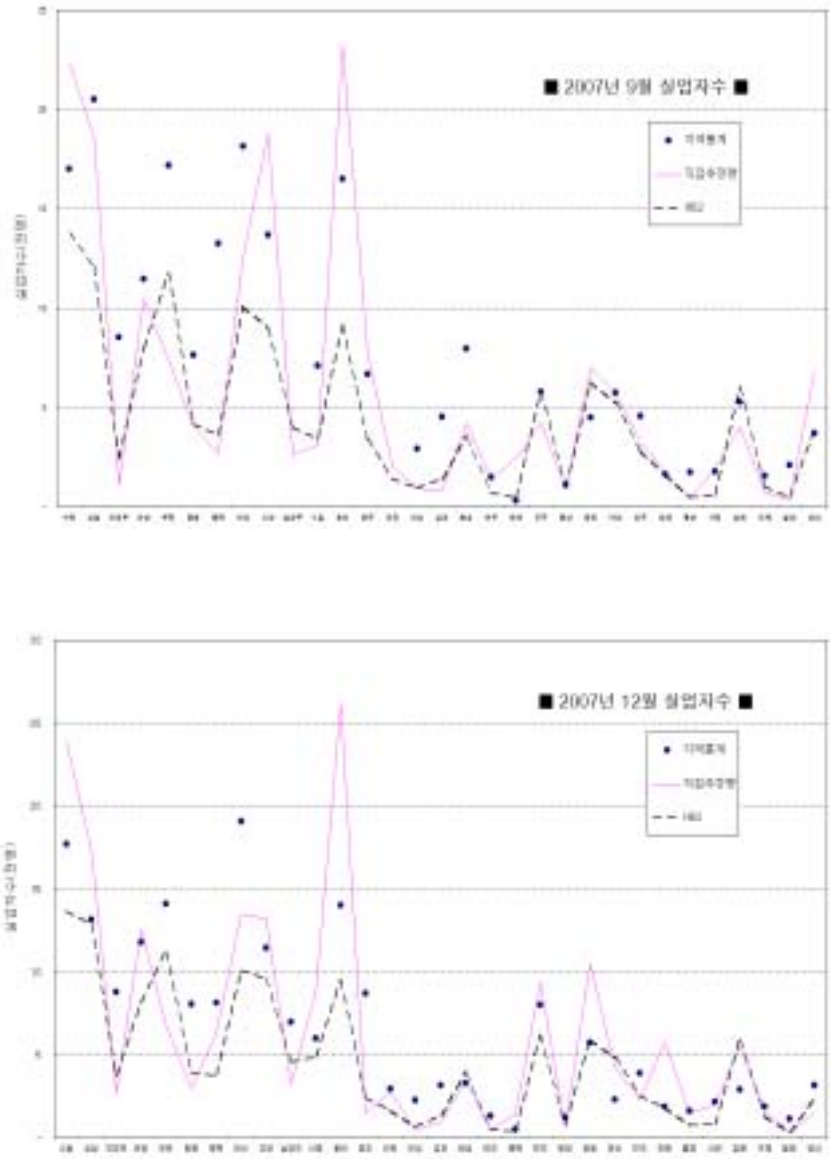
다. 각 추정치의 비교

<표 2-2>와 <표 2-3>에서 'TSCS'부분이 Rao-Yu 모형에 의한 HB 추정량 결과이다. [그림 2-7] ~ [그림 2-9]는 각 추정량들의 결과를 그림으로 표현한 것이다. 이들 결과로부터 다음의 사실들을 알 수 있다.

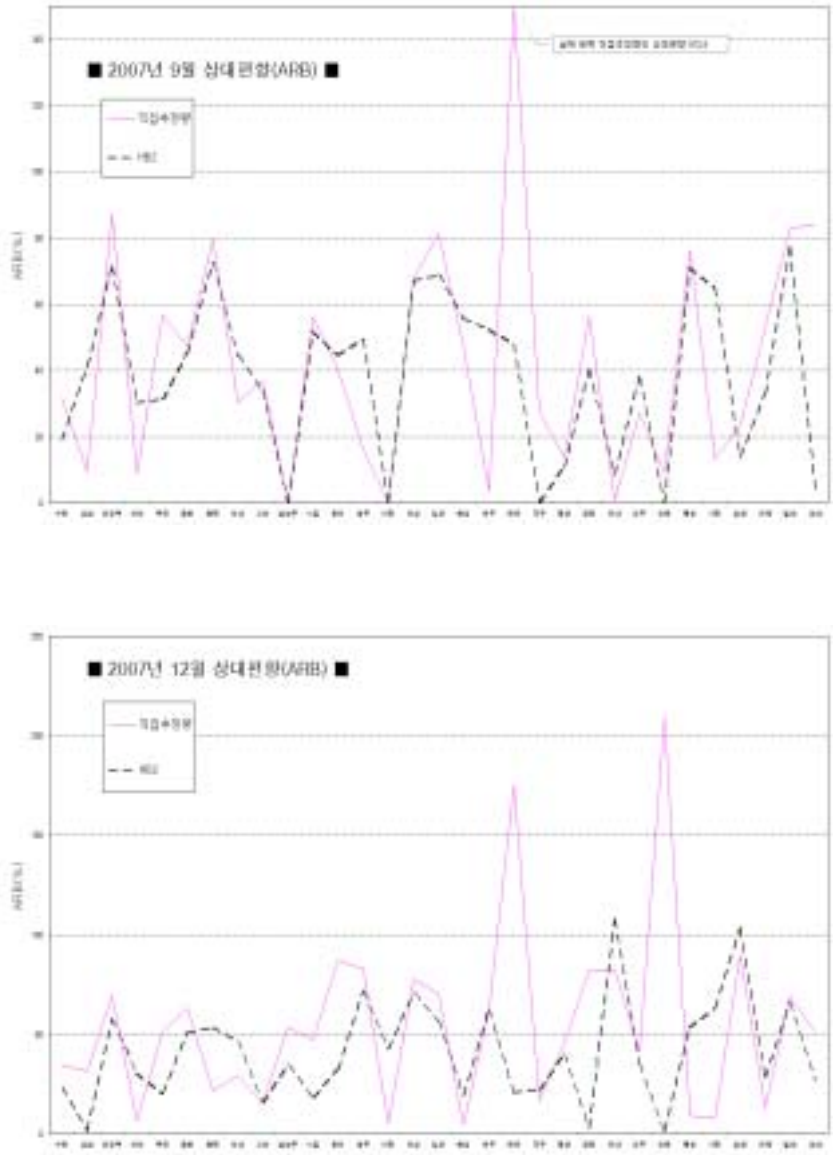
① 직접추정치와 Rao-Yu 모형에 의한 HB 추정치(HB2)를 비교해 보자. 지역에 따라 두 추정치 간에는 큰 차이가 발생한다. 또한 전체적으로 보면 9월 12월 모두에서 대체로 HB2가 직접추정치보다 낮게 추정된 것처럼 보인다.

② ARB를 비교해 보자. <표 2-2>와 <표 2-3>에서 직접추정치보다 HB2 추정치가 낮은 셀을 진하게 표시하였다. HB2가 직접추정치보다도 벤치마크에 가까운 지역이 어느 시점에서도 많다. 그래프를 보면 9월의 파주, 여주, 사천, 12월의 통영과 사천에 대해서는 HB2가 상당히 좋지 않다. 그러나 그 외의 지역에서는 HB2의 ARB가 직접추정치에 비해 개선되었다고 볼 수 있다.

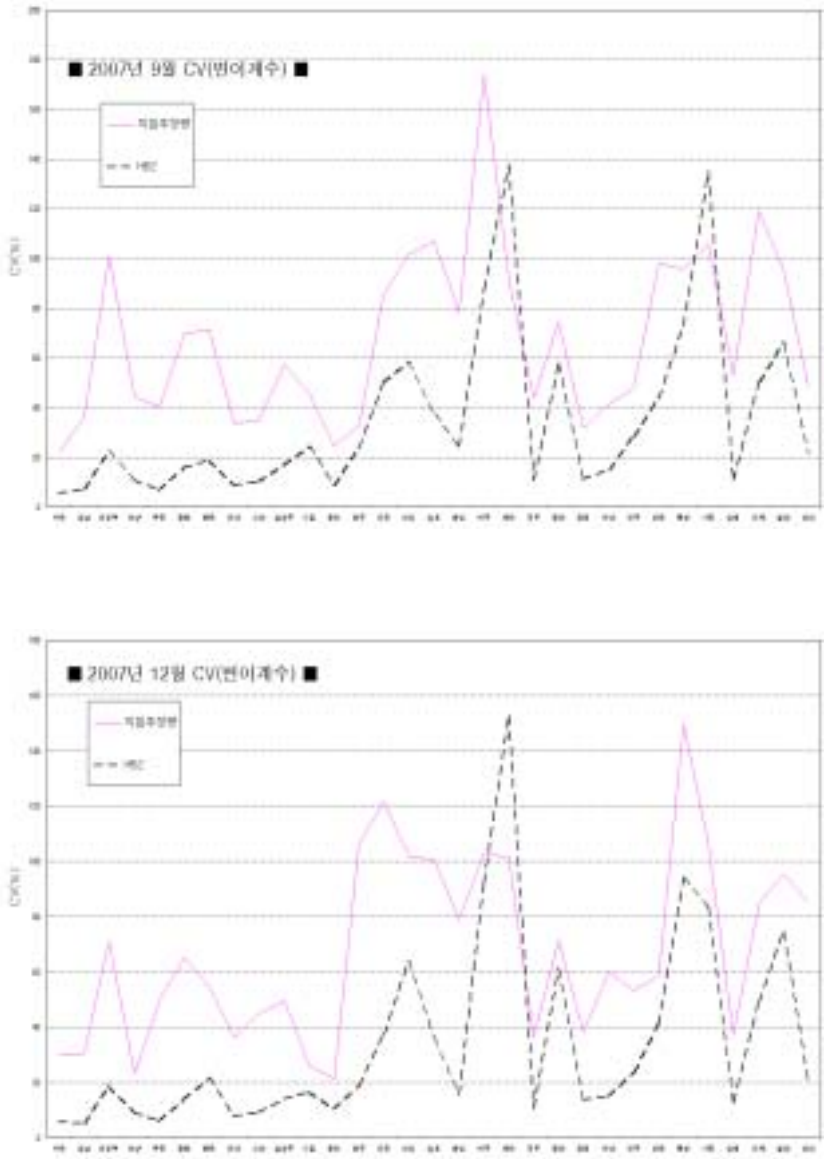
③ 오차율을 비교해 보자. 대부분의 지역에서 HB2가 직접추정치에 비해 전체적으로 낮은 오차율을 나타내고 있다. 그래프를 보면, 태백시를 제외하면 파주, 김포, 통영, 진해, 밀양 등 표본 수가 작은 소지역에서는 HB2의 직접추정치에 비해 오차율이 대폭으로 감소한다는 것을 알 수 있다. 단, 태백, 사천, 여주 등의 소지역에서는 어떤 시기에는 직접추정치보다도 오차가 높게 나타나고 있다. 지역통계와의 비교에서도 HB2의 ARB는 표본 수가 5개 이하로 적은 지역을 제외하고는 오차율 CV가 모두 30% 이내로 지역통계와 상당히 유사한 값을 갖는 것으로 나타났다. 이는 Rao-Yu 모형에 의한 HB2의 오차율은 MSE를 근거로 계산하기 때문에 직접추정치의 오차율과 비교해서 편향만큼 높게 도출된다는 것을 하나의 원인으로 생각할 수 있다. 이에 대해서 마코프 체인의 샘플 경로나 히스토그램([부그림 2-2], [부그림 2-3])을 보면 정규분포 또는 역감마분포에서 약간 어긋나 있다거나 자기상관이 보이는 부분이 있다. 이는 난수를 생성하는 횟수나 한 번에 생성하는 마코프 체인의 난수를 증가시킴으로써 개선될 수도 있다.



[그림 2-7] TSCS 모형에 기반한 추정량 비교: 추정치



[그림 2-8] TSCS 모형에 기반한 추정량 비교: ARB



[그림 2-9] TSCS 모형에 기반한 추정량 비교: CV값

다. 요약

Rao-Yu 모형에 의한 HB2 추정결과를 요약하면 다음과 같다. MCMC 방법에 대해서 R통계량 값에 의한 분석에서는 $\rho = 1$인 경우는 제외하고 완벽하지는 않지만 정상분포에 거의 수렴하고 있다고 볼 수 있다. d 및 p값에 의한 분석에서는 $\rho = 0.5$인 모형이 선택되었다. 단 어떤 모형에 있어서도 p는 높은 값을 보이고, 이는 절대적 기준에서 보면 모형적합이 그다지 좋지 않음을 시사한다. Rao-Yu 모형에 의한 추정이 직접추정치보다도 벤치마크에 가까운 지역 수가 더 많다. 이에 31개 지역 중 거의 모든 소지역에서는 Rao-Yu 모형에 의한 추정이 좋은 결과를 보여준다고 할 수 있다. 오차율은 Rao-Yu 모형 쪽이 직접추정치보다 낮고, 특히 소규모 표본 지역에서 상당히 개선된 결과를 보였다. 표본 수를 6개 이상 유지하면 오차율이 30% 이내로 상당히 안정적이며 오차율면에서 지역통계 값과 유사할 정도의 안정성을 유지한다고 볼 수 있다.

지금까지의 비교분석에 의하면 전체적으로 직접추정치보다 모형추정치가 더 좋은 결과를 보였고, 특히 Fay-Herriot 모형에서는 HB1이 EBLUP보다 나은 결과를 보였다. 마지막으로 Fay-Herriot 모형과 Time series & Cross-sectional 모형을 각각 적용했을 때는 HB 추정량의 결과들이 어떤 경향을 보이는지 직접적으로 비교해 보기로 한다.

3. HB1과 HB2 추정량 비교

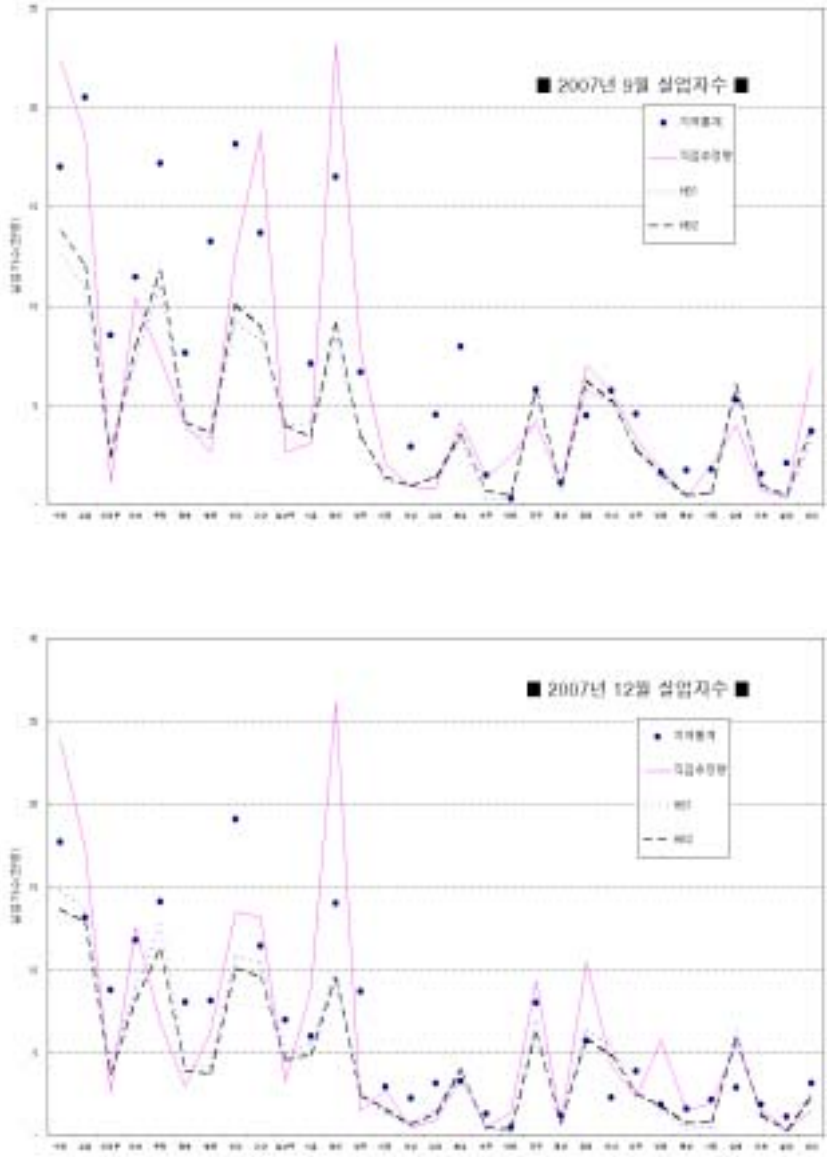
두 모형에 의한 HB추정량에 대해서 직접 비교해 보자 [그림 2-10]~[그림 2-12]는 HB1과 HB2의 결과를 비교하기 위한 것이다. 결과적으로, 직접추정치는 HB추정치들보다 시기에 관계없이 전체적으로 실업자 수가 높게 추정되었다. 전반적으로 HB1과 HB2는 큰 차이가 없지만 실업자 수가 아주 적은 여주, 태백, 사천 등에서는 다소 큰 차이를 보이기도 했다. 앞에서도 설명한 바와 같이 경기도 지역은 지역통계 값이 모형추정치 또는 직접추정치보다 실업자 수가 상당히 높게 나타나고 있다. 이는 주민등록인구를 벤치마크 인구로 사용한 것도 하나의 원인으로 볼 수 있겠다.

① ARB를 보면, 9월의 경우는 HB2가 Fay-Herriot 모형의 HB1보다 더 작은 값을 갖는 지역이 많았고, 12월은 추정치 간의 ARB 수준이 비슷하였다. 그러나 표본 수가 적은 통영, 사천, 여주 등에서는 HB2가 더 좋은 결과를 보였다.

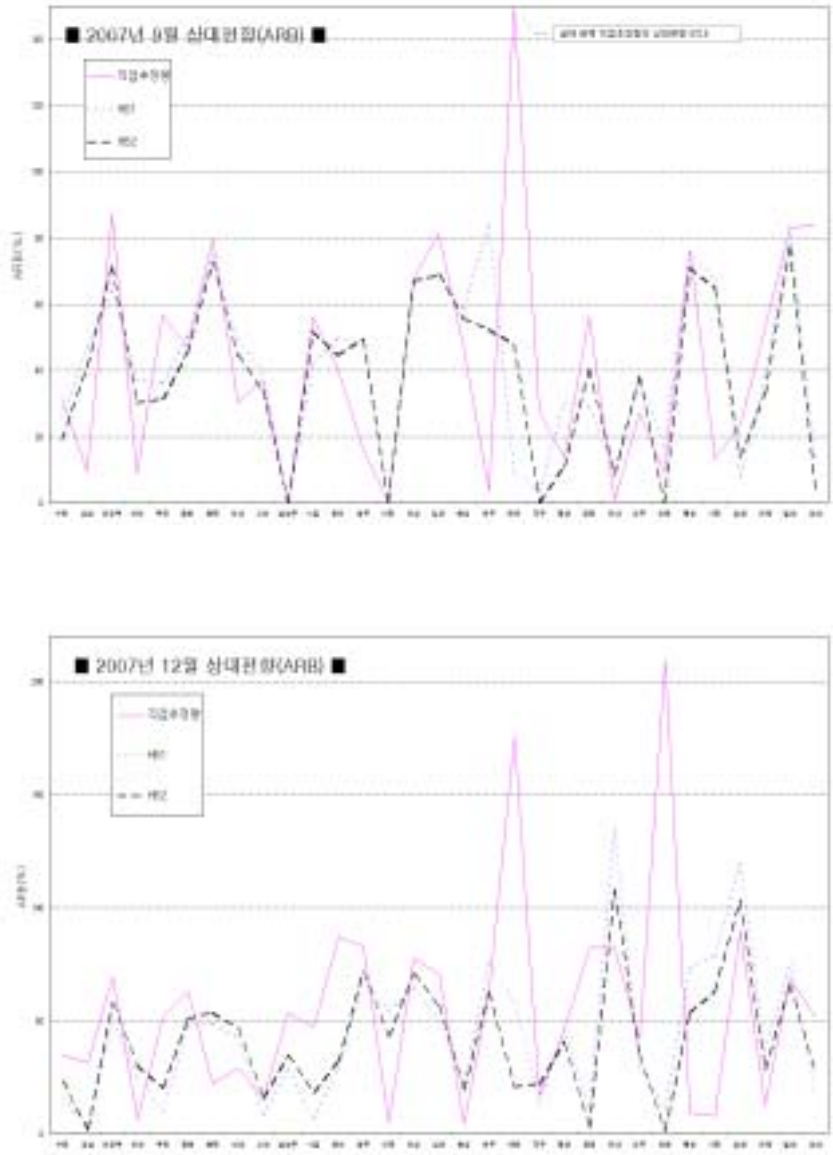
② 오차율을 보면, 시기에 관계없이 직접추정치의 오차율은 HB추정치들보다 전체적으로 높다. HB2가 시기에 상관없이 확실히 오차율이 낮고, 특히 표본 수가 적은 파주, 김포, 여주, 태백, 통영, 진해, 사천, 밀양 지역에서 오차율이 크게 개선되었다.

③ Rao(2003)는 소지역 추정에서 추정치에 대한 오차율(CV)이 약 25%~30% 이내면 사용하는 데 무리가 없다고 설명하고 있다. 이에 따르면 HB 추정량은 31개 지역 중 20개 지역은 30% 이내의 오차율을 유지하면서 지역통계의 오차율과도 큰 차이가 없이 안정적인 것으로 나타났다([그림 2-11]). 이들 20개 지역은 경활조사의 표본 수가 모두 6개 이상인 지역에 해당된다. CV가 30%를 넘는 11개 지역은 모두 표본 수가 6개 이하인 지역으로, 지역통계의 CV와도 상당히 차이를 보이고 있다. 이처럼 안정적인 오차율을 유지하기 위해서는 적정 최소 표본 수를 정하는 것이 중요하다. 이 분석 결과에 의하면 각 소지역은 적어도 7~8개 이상의 표본을 유지하는 것이 필요하다.

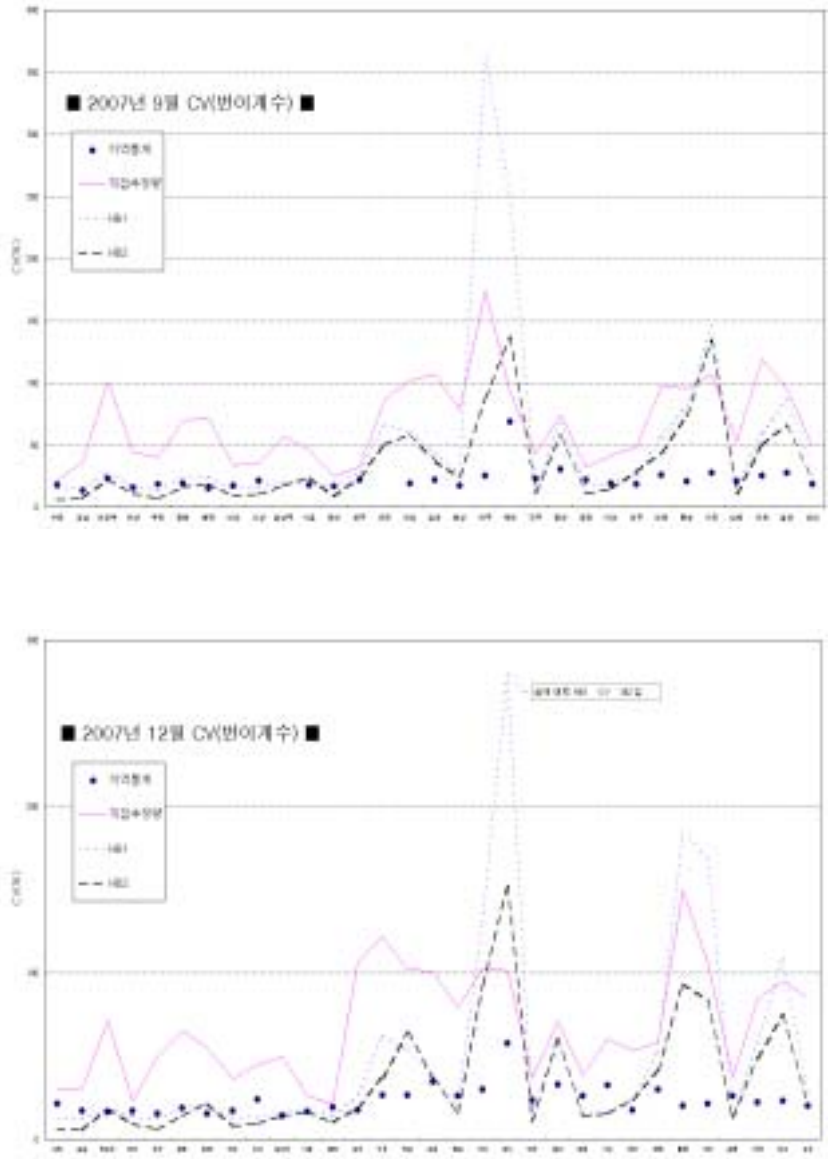
결론적으로 ARB나 오차율 측면에서는 모형추정량들이 직접추정량보다 더 좋았고, 모형추정량 중에서는 HB가 EBLUP보다 더 좋다는 선행 연구 결과들이 확인되었다. 동일한 추정량 HB에 대해서 Fay-Herriot 모형과 Rao-Yu 모형을 각각 적용했을 때, 예상대로 Rao-Yu 모형에 의한 추정량이 가장 좋은 결과를 보이는 것으로 나타났다.



[그림 2-10] 계층적 베이즈 추정량 HB1과 HB2 비교: 추정치



[그림 2-11] 계층적 베이즈 추정량 HB1과 HB2 비교: ARB



[그림 2-12] 계층적 베이즈 추정량 HB1과 HB2 비교: CV값

4. 신뢰구간을 이용한 추정치 평가

가. 추정치의 품질 측정

추정치의 품질측정과 관련하여 핵심 지표로 사용되는 것이 표준오차(standard error)이다. 그렇다면 표준오차는 무엇이고, 언제 발생하며 어떻게 계산할까? 표준오차는 모형 기반 추정의 경우 다음에 의해 생기는 '불확실성'의 정도를 나타낸다.

- 소지역내에서 실업자를 추정하기 위해 모형을 사용한다는 것
- 소지역내의 인구추정을 위해 경활조사를 사용하는 데서 비롯되는 표집오차
- 모형에서 설명되지 않는 지역들의 효과를 허용한다는 것.

각 추정치는 표준오차가 결정되면 신뢰구간이 계산될 수 있다. 95% 신뢰 구간의 상한과 하한은 다음과 같다.

$$\text{실업자수} \pm 1.96 \times \text{표준오차}(\text{또는 } \sqrt{\text{추정치 분산}}) \qquad (2.24)$$

나. 실업자 수 분포와 지역 순위 문제

추정치에 순위를 매기는 것과 관련하여 소지역들의 순위를 해석하는 데 있어서는 매우 신중함을 기해야 한다. 또한 이들 순위를 사용할 때, 추정치의 변동은 반드시 설명되어야 한다. 예를 들면 가장 높은 순위를 갖는 지역들에 해당하는 신뢰구간은 그 추정치가 가장 많은 실업자 수를 갖는 지역이라기보다는 그 지역은 가장 많은 실업자 수를 갖는 그룹에 포함된다고 해석하는 데 도움이 된다. 2개의 특정 지역들 간 추정치들은 오직 이들 추정치들에 대한 신뢰구간이 겹치지 않으면 통계적으로 유의한 차이가 있다고 설명할 수 있다. 신뢰구간은 식(2.24)를 이용한다.

비록 지역의 실업자 수에 대한 모형 기반 추정치들에 순위를 매길 수 있다 하더라도, 이들은 지역들의 실업분포에 대한 어떠한 추론에도 사용될 수 없다. 추정절차는 총 인구에 대해 평균 실업자에 대한 추정치를 축소시키는(shrink) 경향이 있을 것이다. 그래서 모형 기반 추정치들

은 scale의 끝에서 과소 또는 과대 추정되는 경향이 있다. 그럼에도 불구하고 어떤 경우의 추론에는 사용될 수도 있다. 즉, 소지역A의 실업률이 소지역B보다 더 크다(물론 두 신뢰구간들이 겹치지 않는다면)는 식의 추론은 가능하다. 그러나 전체 소지역들 중 X%는 Y보다 더 큰 실업률을 갖는다고 단언하는 것은 의미가 없다.

예를 들어 두 개의 추정치를 비교할 때, 두 지역들의 신뢰구간이 겹치지 않는다면 두 지역들 간의 실업자 수는 통계적으로 유의한 차이를 갖는다고 말할 수 있다. <표 2-5>에서 소지역C는 소지역A보다 유의하게 낮은 추정치를 갖는다. 왜냐하면 95% 신뢰구간들이 서로 겹치지 않기 때문이다. 반면에 소지역B는 소지역A보다 유의적으로 낮은 추정치를 갖는다고 말하기 어렵다. 이는 두 지역 추정치에 대한 신뢰구간이 겹치기 때문이다. 실업률에 대해서도 똑같은 내용을 적용할 수 있다.

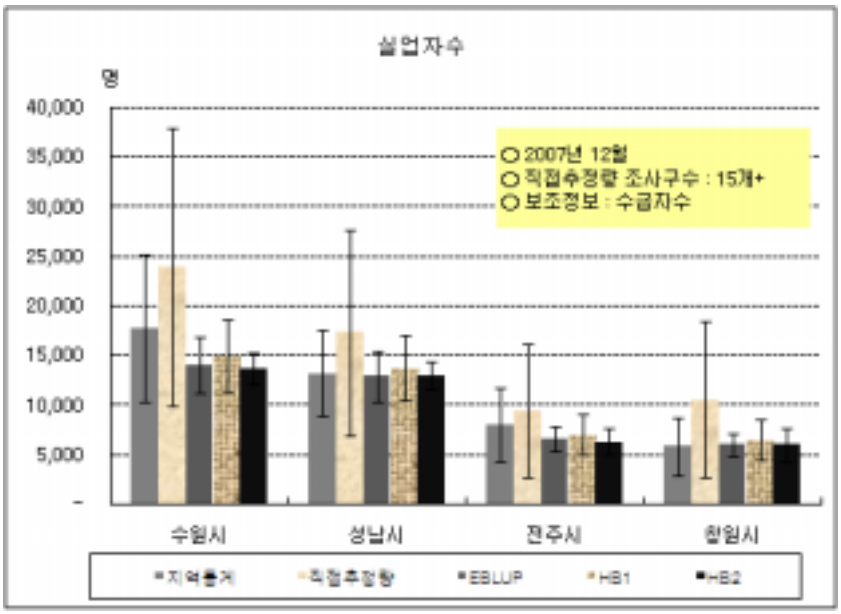
〈표 2-5〉 3개 소지역의 실업자 추정치와 신뢰구간

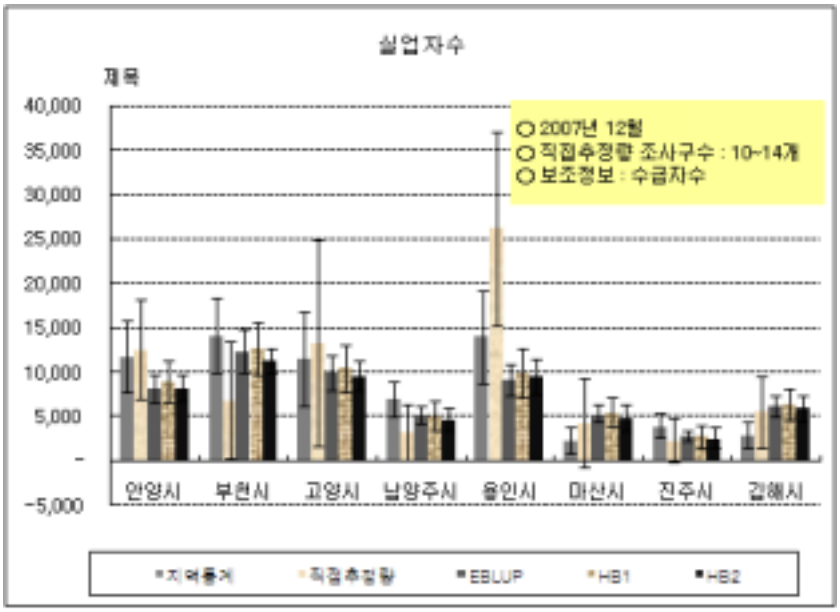
소지역	실업자 추정에 대한 95% 신뢰구간		
	추정치	상한	하한
A	13,740	12,181	15,300
B	12,066	10,301	13,831
C	10,090	8,331	11,850

다. 표본 수에 따른 추정치의 안정성

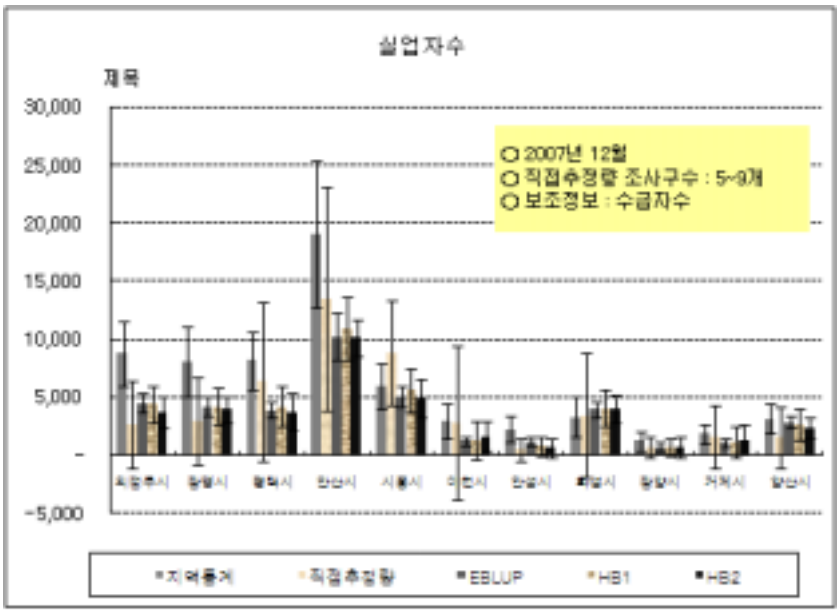
신뢰구간을 이용하여 표본 크기에 따른 추정량들의 안정성이 설명될 수 있다. [그림 2-13]은 2007년 12월의 31개 시군구들을 표본규모에 따라 4개 그룹으로 나누어 각 추정량들의 신뢰구간을 표시한 것이다.

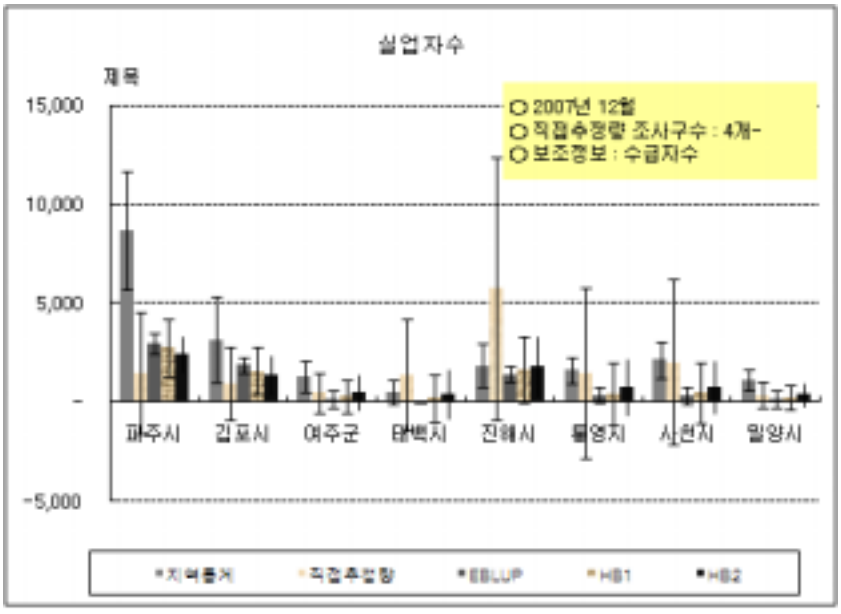
그림에서 보면 표본규모 그룹에 상관없이 모든 지역에서 직접추정치의 신뢰구간이 가장 크다. 이는 그만큼 신뢰오차의 범위가 크다는 것을 의미한다. 특히 표본규모가 10개 미만인 그룹들에서는 직접추정치의 신뢰구간이 음의 값을 갖기도 한다. 실업자는 양수 값을 갖는다고 보면 이 결과는 받아들이기 힘들다. 즉, 이 결과는 직접추정치의 경우 표본 수가 적으면 사용하기 어렵다는 것을 증명해 준다.





[그림 2-13] 표본규모별 추정치의 신뢰구간


실업자수
제목
30,000
25,000
20,000
15,000
10,000
5,000
-
-5,000
2007년 12월
직접추정량 조사구수 : 5~9개
보조정보 : 수급자수
지역통계
직접추정량
EBLUP
HB1
HB2


실업자수
제목
15,000
10,000
5,000
-
-5,000
2007년 12월
직접추정량 조사구수 : 4개-
보조정보 : 수급자수
파주시
김포시
여주군
태백지
진해시
통영지
사천지
밀양시
지역통계
직접추정량
EBLUP
HB1
HB2

[그림 2-13]의 계속

표본규모에 따라 구체적으로 살펴보자. 표본크기가 15개 이상인 경우에는 직접추정치를 제외한 나머지 추정치들은 지역통계 값의 신뢰구간과 완전히 겹친다. 지역통계를 벤치마크로 했을 때, 95% 신뢰수준 내에서는 두 추정치에 차이가 없다고 볼 수 있다. 표본 수가 10개에서 14개인 그룹에서는 약간 다르다. 마산과 김해에서는 직접추정치는 별도로 하더라도 나머지 추정치들마저 지역통계 값과 신뢰구간이 완벽하게 겹치지 않는다. 이는 두 지역에서는 추정치들이 명백히 다르다는 것을 나타낸다. 이와 같은 양상은 표본 수가 적어질수록 많은 지역에서 나타난다. 특히 표본 수가 4~5개 이하인 지역에서는 추정치의 신뢰구간이 음의 값을 갖거나 지역통계의 신뢰구간과 전혀 겹치지 않는다는 점에서 추정치가 불안정할 수 있다는 점에 주의해야 한다.

이러한 결과를 통해 볼 때 모형을 이용해 실업자 수(율)를 추정하고자 한다면 표본 수는 최소한 일정 수준(약 6~7개 이상의 조사구) 이상은 필요하다고 보인다. 그러나 추정에 필요한 표본 수 예측에 관한 문제는 별도의 연구로서 보다 포괄적이고 자세하게 이루어져야 하는 문제이지 여기서 단언할 문제는 아니다.

5. 추가 분석 결과

가. 공간정보를 활용한 결과

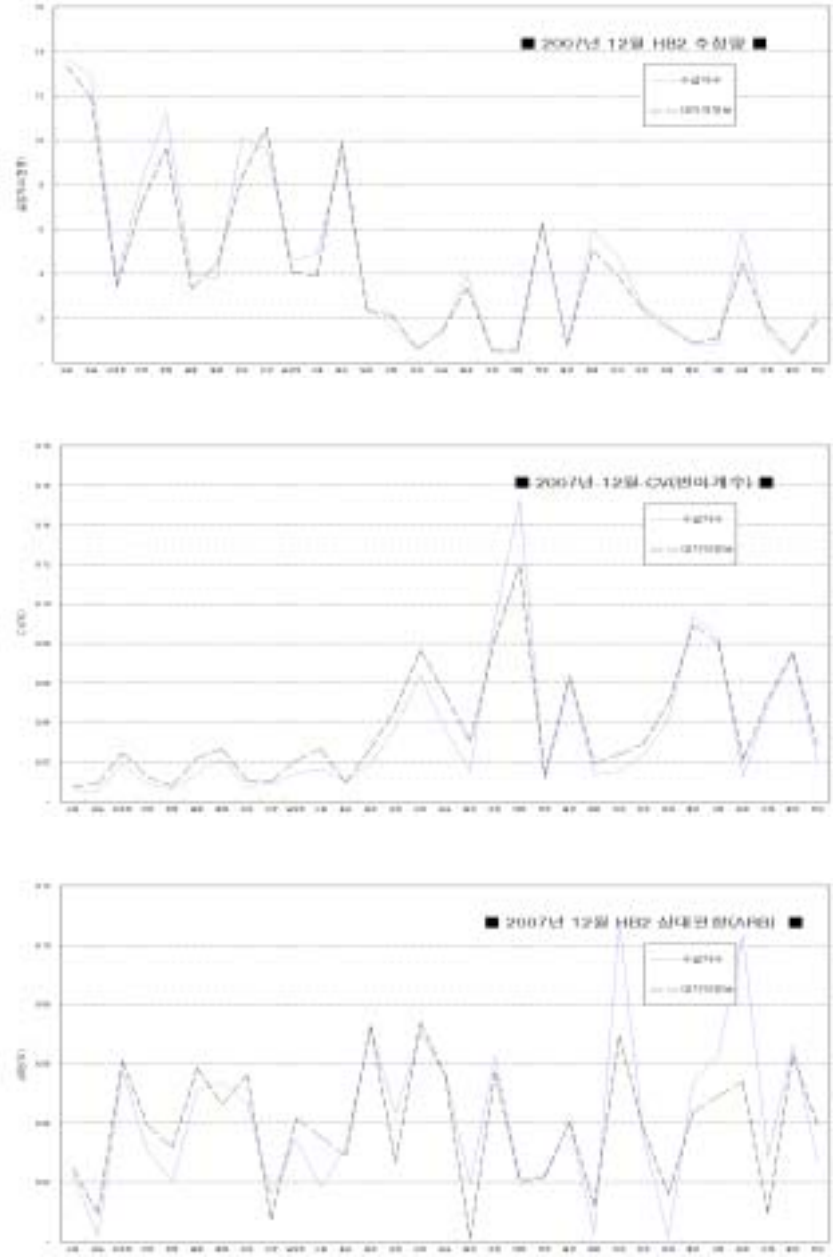
지금까지는 보조정보 탐색을 강조하여, 고용보험 자료의 실업급여 수급자 수를 보조정보로 사용한 분석결과를 위주로 설명하였다. 그러나 (5.1.나)절에서 설명한 바와 같이 실업급여 수급자 수가 보조정보로써 큰 역할을 하지 못했다. 이에 연구자들은 대안으로 공간정보의 보조정보로써의 활용 가능성에 대해 추가로 검토하였다. 대지역의 실업자 수를 소지역의 인구수 대비 대지역의 인구수로 비례할당해서 보조정보로 사용하였다.

분석결과, 시기에 관계없이 전체적으로 보면 수급자를 사용한 지금까지의 추정결과를 크게 개선하지는 못했다. 경우에 따라서는 오히려 수급자 정보를 사용했을 때보다 좋지 않은 결과를 보이기도 했다. [그림

2-14]는 2007년 12월 자료에서 실업급여 수급자 수와 공간정보를 사용했을 때의 HB2 추정량에 대한 추정치, 상대편향, 오차율의 비교를 나타낸다.

수급자 정보에서 실업자 수가 약간 높거나 비슷한 수준에서 추정된다. 지역통계와의 ARB는 수급자 수를 사용했을 때는 표본 수가 많은 지역에서는 대체로 낮은 편이지만, 표본 수가 적은 지역들에서는 오히려 높다. 이는 표본 수가 적은 경우의 모형은 보조정보에 의해서도 크게 좌우될 수 있음을 보여준다. 오차율은 수급자 수를 사용했을 때가 대지역의 경우보다 더 안정적으로 나타났다. 이와 같이 소지역을 포함하는 대지역의 실업자 수를 소지역의 정보로 사용하고자 하는 경우, 소지역은 대지역과 거의 동일하다는 가정을 전제로 하고 있다. 만약 이 가정이 만족되지 않으면 추정 결과에 편향을 초래할 수 있다. 또한 대지역의 표본조사결과를 이용한다는 측면에서도 과소분산추정 결과를 제시할 가능성이 높다. 따라서 대지역 정보를 사용하고자 하는 경우에는 주의할 필요가 있다.

본 보고서에는 대지역정보를 사용했을 때의 추정결과를 자세하게 설명하지 않을 것이다. 다만, 독자들의 이해를 돕기 위해 HB의 추정결과는 [부록]의 <부표 2-1>과 <부표 2-2>에 월별로 제시하였다. 자세한 이해와 해석은 본문의 설명단계를 차근차근 따라가면서 쉽게 할 수 있을 것이다.



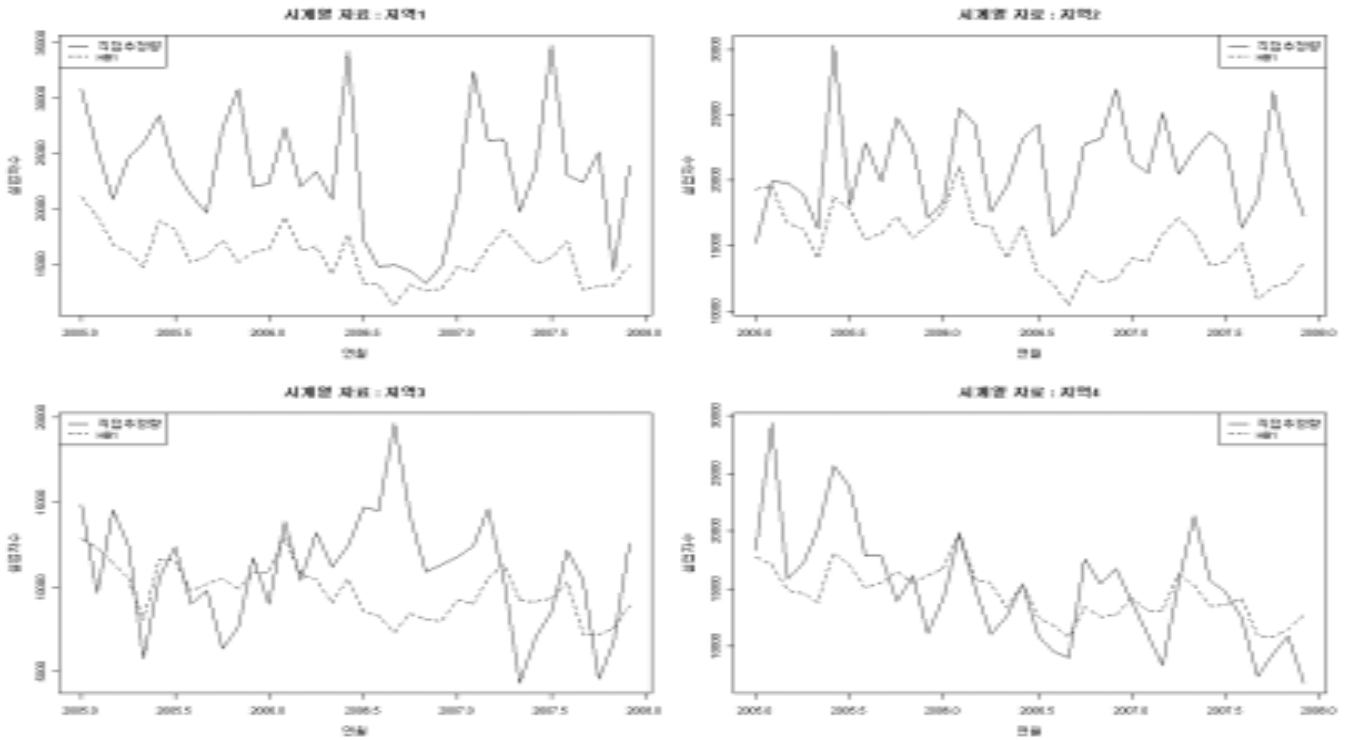
[그림 2-14] HB 추정량을 이용한 보조정보별 추정효과 비교 : 2007년 12월

나. 추정치들의 시계열 비교

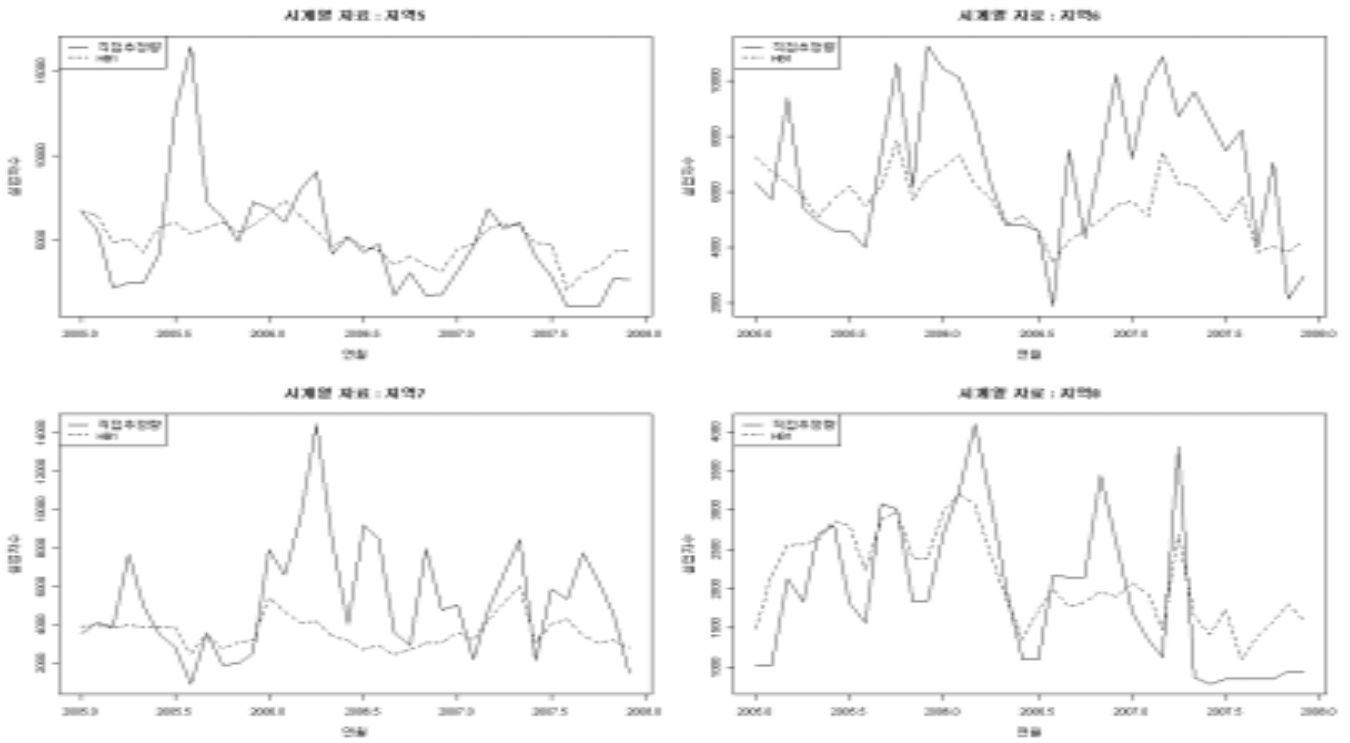
추정된 결과에 대해서 추정치들의 시계열을 이해하는 것은 중요하다. 반복조사의 경우 전월 또는 전년대비를 통해 사회·경제현상의 변화를 파악하는 데 사용하기 때문이다. 그러나 시계열에 따른 추세를 정확하게 파악하기 위해서 충분한 자료가 확보되어야 한다. 월별 자료인 경우, 최소한 48개월에서 60개월 자료는 있어야 한다. 이런 이유로 본문에서 따로 해석하지 않았지만, 여기서는 시계열에 의한 변화 양상을 개괄적으로 이해하는 데 도움이 되기 위해 추가설명을 하고자 한다.

일례로, 직접추정치와 HB1추정량의 시계열 그래프를 그려보았다. 표본규모별로 31개 지역을 4개 그룹으로 나누고 각 규모에 대해 2개 지역을 임의로 선택하여 이들의 직접추정치와 HB1의 시계열을 [그림 2-15]에 나타내었다. [그림 2-15]에서 1번째 행의 지역1과 지역2는 표본 규모가 15개 이상, 2번째 행의 지역3과 4는 10~14개, 3번째 행의 지역5와 6은 5~9개, 4번째 행의 지역7과 8은 4개 이하의 표본을 가지는 지역들을 나타낸다. 이것에 의해 다음과 같은 사실들을 알 수 있다.

- 추정치의 크기는 고려하지 않더라도 전체적인 변화패턴은 직접추정치와 HB1이 유사한 경향을 갖는다고 볼 수 있다.
- 그러나, 표본규모가 가장 많은 지역1과 2에서는 HB1 추정치들이 모두 직접추정치보다 아래쪽에 위치하고 있고, 나머지 지역에서도 어떤 시기에서는 이런 현상이 발생하고 있다. 이와 같은 현상은 Fay-Herriot 모형에서 회귀직선으로부터 산포가 일정하다는 가정이 타당하지 않다고 볼 수 있는 부분이다. HB1 추정량은 이웃지역 정보로 회귀분석을 하기 때문에, 관심값이 전국 평균보다도 크게 나오기 쉬운 소지역은 직접추정치보다 항상 아래쪽에 해당 소지역의 추정량이 위치할 가능성이 높다고 말할 수 있다. 즉, 전국 평균 실업자 쪽으로 추정치를 축소시키는 경향이 있다. 또한 그런 높은 값 근처의 값을 갖는 소지역도 이들 값의 영향을 받아서 결과가 왜곡될 가능성이 있다. 이런 경우에는, 각 시군구에 관한 사전정보를 충분히 얻을 수 있다면 Fay-Herriot 모형의 b_i의 부분을 사전정보로부터 조정하는 등의 대처를 할 수 있다.



[그림 2-15] 추정치의 시계열 그래프: 직접추정량 VS. HB1



[그림 2-15]의 계속

◦ 이외에도 두 시계열의 차이에 대해서 몇 가지 이유를 더 확인해 볼 수 있다. 예를 들어 일정기간 꾸준히 안정적이다가 갑자기 어떤 시점 이후에서부터 추정치의 적합이 좋지 않은 지역에서는 이 시기에 보조정보와의 상관관계에 변화가 발생했을 가능성이 있다. 이를 테면, 이 시점에 전국 실업자 수가 크게 증가했을 수도 있다.

◦ 그리고 특정 시기에 크게 추정치가 감소했다가 직후에 바로 원래의 수준을 회복하는 경우도 있다. 이 달의 실업자 수는 이상치일 가능성이 높다. HB1 추정치는 이 이상치의 영향을 강하게 받을 수 있다.

제7절 결론 및 토의

본 절에서는 지금까지의 연구결과를 종합적으로 요약하고, 연구를 통해 드러난 여러 가지 문제점들에 대한 해결방안을 모색하고자 한다. 특히, 우리나라 고용통계 소지역 추정 연구의 발전을 위해 필요한 보조정보의 활용방안과 최종 추정방법을 선택하기까지 진행되어야 할 연구과제들을 중심으로 논의하고자 한다.

1. 요약

지금까지 모형 기반 추정방법과 그 결과들에 대해서 살펴보았다. 모형은 Fay-Herriot과 Time series & Cross-sectional 모형을 고려하였고, 추정량으로는 EBLUP과 HB를 사용하였다. Fay-Herriot 모형에는 EBLUP과 HB(HB1)를 각각 적용하였고, Time series & Cross-sectional 모형에는 HB(HB2)를 적용하였다. EBLUP추정에서 지역분산은 최우추정법을 사용하였고, HB에서는 MCMC방법에 의한 깁스 샘플러를 사용하여 라오-블랙웰 추정량에 의해 사후평균을 계산했다. 본 연구에서는 모형 추정량들의 특성을 평가하기 위해 여러 가지 평가방법을 사용하여 비교분석하였다.

연구결과를 요약하면 다음과 같다.

◦ Fay-Herriot의 EBLUP추정량은 오차율과 ARB 측면에서 직접추정치보다 낮고, 정도(precision)도 개선되었다고 말할 수 있다. HB1은 몇 개의 지역을 제외하면 EBLUP보다 오차율과 ARB가 낮아서 더 우수한 추정량이라 말할 수 있다.

◦ Time series & Cross-sectional 모형의 HB2추정량은 오차율과 ARB 면에서 직접추정치보다도 낮고 정도도 상당히 개선된 것으로 나타났다. MCMC 방법을 이용함으로써 복잡한 모형 추정도 간단하

게 할 수 있다.

◦ 결과적으로 HB2가 HB1보다 지역통계에 더 근접한 결과를 보였다. 또한 오차율 면에서도 가장 안정적이었으며 특히 직접추정치의 오차율을 가장 많이 개선하는 것으로 나타났다. 그러나 표본규모가 작은 지역에서는 모형추정치가 다소 불안한 지역도 있었다. 직접추정치는 전체적으로 오차범위가 매우 넓고, 더구나 표본규모가 작은 지역에서는 직접추정치의 신뢰구간이 음의 범위를 포함함으로써 추정치로서의 가치를 의심케 하였다.

◦ 또한 31개 소지역 중 20개 지역에서 모형추정치의 오차율은 30% 이내로 안정적이었으며 지역통계의 오차율과도 큰 차이를 보이지 않았다. 그러나 표본 수가 적은(6개 이하) 소지역에서는 모형추정치의 오차율이 모두 30% 이상으로 다소 불안정한 상태를 보였다. 추정치의 안정성 확보를 위해서는 최소 표본 수를 확보할 필요가 있으며, 이 경우 적어도 7~8개 이상은 필요할 것으로 보인다.

◦ 어떤 소지역 추정 방법을 이용해도 모형이나 보조정보를 잘 선택함으로써 원래의 직접추정치보다는 더 좋은 결과를 얻을 수 있다. 이에 보조정보와 실업자 수와의 상관이 높지 않거나 이상치의 문제 등 향후 지속적으로 검토해야 할 과제가 있다는 것을 알 수 있다.

◦ 또한 일부 지역을 대상으로 살펴본 결과에 의하면 모형추정치는 시계열 측면에서 직접추정치와 유사한 추세를 갖는 것으로 나타났다.

본 연구 결과, 추정량들을 편향(ARB)과 효율성(CV) 측면에서 비교해 볼 때 Time series & Cross-sectional 모형에 의한 HB2가 가장 선호될 만하다. 그러나 이 추정량은 계산과정이 복잡하고 시뮬레이션을 통한 추정을 하기 때문에 재현성이 다소 떨어질 수 있다는 점을 주의해야 한다.

Fay-Herriot 모형에 기반한 HB1은 HB2에 비해 효율성 측면에서는 다소 떨어질 수 있지만, CV 기준을 30% 수준으로 놓고 보면 HB2와 거의 유사한 능력을 갖는다. 그러나 표본 수가 아주 적은 몇몇 소지역에서는 HB2보다 더 편향된 경향이 있다. EBLUP은 HB추정량들에 비해 효율성 측면에서는 더 우수하지만 이론적으로는 과소분산 추정 문제를 안고 있다. 이 추정량의 편향 정도는 HB2와 거의 유사한 수준으로 양호하지만 표본이 작은 소지역에 대해서는 상당히 편향된 경향이 있다.

지금까지의 비교는 각 지자체에서 수행하고 있는 지역의 고용통계조사 결과를 기준값으로 사용하였다. 그러나 지역의 고용통계조사 역시 표본조사이고 표본 수를 충분히 확보함으로써 조사 추정치에 대한 불편성은 유지할 수 있다 하더라도 여전히 효율성 문제는 남게 된다. 따라서 이 지역조사 통계를 기준으로 삼은 본 연구의 비교결과는 이러한 측면을 고려해서 이해하기를 바란다.

2. 고용보험 자료 활용 방안

모형을 이용한 추정연구에서 보조정보의 선택은 최종 추정결과에 중요한 영향력을 지닌다. 본 연구는 경활조사의 실업자를 주 정보로 하고, 행정자료 중 고용보험 자료의 실업급여 수급자 수를 보조정보로 활용하였다. 실업급여 수급자 수는 고용보험실업률을 측정하는 지표로 사용되고 있고, 경활조사의 실업자와 밀접한 관계가 있을 것으로 판단되었기 때문이다.

한국고용정보원에서 정의하고 있는 고용보험 실업률은 다음과 같이 계산한다.

고용보험실업률 = (월 실업급여 등록자수 / 월 피보험자 수)×100

여기서, 실업급여 등록자 수는 실업급여 신규신청자와 실업급여 순수수급자의 합으로 구성되고, 이때 중복자는 제외된다. 본 연구에서는 실업급여 등록자수의 일부인 순수 수급자 수를 보조정보로 활용하였다(권혜자와 노현국, 2007). 향후 신규신청자 수를 포함하는 실업급여 등록자 수를 고려함으로써 시군구 실업자 추정의 정도를 더 높일 수 있을 것으로 기대한다.

그러나 고용보험 자료와 같은 행정자료를 활용함에 있어 몇 가지 주의사항이 요구된다.

첫째, 행정자료에 대한 정확한 이해 부분이다. 흔히 사용자들은 자신의 목적에 치중하다가 제공자의 자료 상태를 정확하게 이해하지 못함으로써 여러 가지 오류를 범하게 된다. 일반적으로 행정자료는 마이크로 자료를 DB에 저장하고, 필요(또는 사용자의 요청)에 따라 새로운 자료 형태로 가공하게 된다. 이때 사용자와 제공자 간에 자료에 대한 범위 및 기준설정이 일치하지 않는다면 서로가 엉뚱한 자료를 주고받게 되는 문제가 발생할 수 있게 된다.

둘째, 행정자료를 통계자료로 사용하기 위해서는 통계목적에 부합하도록 다년간의 정비노력이 필요할 것으로 보인다. 우선 행정자료 그대로는 통계자료로 사용하기 곤란하다. 행정자료 DB에 저장되어 있는 자료는 통계용 자료와 기본적인 정의부터 다를 수 있다. 뿐만 아니라 행정자료는 조사통계 자료와 다르게 행정절차에 필요한 시간이 소요되는데, 특히 월별 통계의 경우는 시간적 요소가 크게 작용할 수 있다. 따라서 이런 절차상 문제에 대해서도 통계용 목적과 부합하도록 합의점을 찾아야 할 것이다.

셋째, 지리적 소지역 또는 표본설계 외 영역의 통계 작성을 목적으로 행정자료를 활용할 경우, 지역별로 주요항목들의 특성이 다를 수 있음을 충분히 인지해야 한다. 고용보험의 경우, 고용보험 실업률은 지역별 피보험자 수 또는 실업급여 등록자의 특성 등에 따라 다르다. 그러므로 지역 간 비교 지표로 사용하기 위해서는 이러한 지역별 특성 차이가 충분히 고려되어야 한다. 어떤 지역은 농업 발달 지역인 반면 어떤 지역은 제조업 발달 지역이라면 이 지역 간 피보험자 수 및 실업급여등록자 수는 큰 차이를 보일 수 있기 때문이다.

마지막으로 고용보험 자료와 같은 행정자료는 정부정책에 따라 크게 변동될 수 있기 때문에 그 자료를 사용함에 있어 주의가 요구된다. 이는 통계의 시계열적 해석에 적지 않은 영향을 미칠 수 있다. 이 경우 통계자료로서 안정성을 확보할 수 있는 장치를 마련해야 할 것이다.

고용보험 자료로부터 몇 가지 사례를 통해 행정자료 이용 시에 발생할 수 있는 부분을 예측해보자. 고용보험 자료에서 말하는 피보험자 수

는 고용정보원의 고용보험 DB에 자료가 입력된 날짜를 기준으로 집계된다. 즉, 피험보자가 실제로 피보험자격을 얻거나 잃게 되는 퇴직일 또는 입사일을 기준으로 하지 않는다는 것이다. A라는 피보험자가 2008년 7월 1일이 실제 회사 입사일로서 피보험자격을 얻었지만, 고용정보원 DB에는 2008년 8월 2일에 자료가 입력・완료되었다면 이 사람은 2008년 8월 피보험자격 취득자로 집계되는 셈이다. 그리고 만약 이 사람이 동시에 경활조사의 표본대상이었다면, 경활조사에서 이 사람은 2008년 7월 실업자로 집계되면서 고용보험 실업자와 경활조사의 실업자 수에는 괴리가 발생할 수 있게 된다.

또한 고용보험 자료에서 매월 실업급여 등록자는 중복집계 될 수 있다. 이미 언급한 바와 같이 실업급여 등록자는 신규 신청자와 수급자 수로 집계되고 있다. 이때 실업급여 신청과 지급 절차 사이에 실업인정 절차가 있다. 실업인정은 2주마다 이루어지기 때문에 2008년 7월 1일에 신규 신청한 사람은 한 달에 2번 실업인정을 받게 되어 중복 집계되는 상황이 발생하게 된다.

이처럼 고용보험 자료와 경활조사 자료는 실업통계 자료라는 동일한 목적을 가지는 반면에 실업정의에서부터 실업인정 날짜 등 상이한 부분이 많다. 이처럼 두 자료의 특성이 다르기 때문에 통계목적을 위해 완벽하게 일치시키는 것은 쉽지 않다. 그러므로 그 격차를 줄일 수 있는 부단한 노력이 필요하며 상호보완을 목적으로 꾸준한 검토가 필요할 것으로 보인다.

3. 향후 연구 및 논의

본 연구의 범위는 연구추진 절차상의 문제나 보조정보의 가용성 측면에서 다소 축소된 점이 없지 않다. 그럼에도 불구하고 이 연구가 우리나라 고용통계 소지역 추정 연구에 박차를 가한 것임에는 틀림없다. 우리나라의 상황에 적합한 방법을 최종 결정하기 위해서는 좀 더 포괄적이면서 구체적으로 연구되어야 할 부분들이 남아 있다. 이를 위해 향후 지속되어야 할 소지역 추정 과제들을 언급하고자 한다.

첫째, 본 연구에서는 적용되지 않은 다른 방법들에 대한 검토가 이루

어져야 한다. 방법 간 비교는 그 비교 기준이 되는 절대적 지표가 없기 때문에 실제로 상당히 어려운 문제이다. 그러나 신뢰할 수 있는 벤치마크와 비교하는 방법이라든가 MCMC에 의한 분석 등을 포괄하는 종합적인 판단을 향후에 해야 할 필요가 있을 것으로 생각된다.

둘째, 다양한 추정량들의 안정성 및 우수성을 평가할 필요가 있다. 이는 인구센서스 자료와 같은 전수자료로부터 모집단과 닮은 표본을 반복해서 추출하여 실업자 수(실업률)를 소지역별로 각 추정방법에 따라 추정해 봄으로써 평가할 수 있다.

셋째, 7.1절에서 설명한 바와 같이 모형 기반 연구에서 장기간의 시계열 검토는 여러 가지 모형의 안정성과 해당 지역의 특성을 이해하는 데 도움이 된다. 이는 보조정보를 실업급여 수급자 수보다는 장기적으로 활용가능한 대지역 정보를 활용하여 검토할 수 있는 부분으로 향후 지속적으로 검토・연구되어야 한다.

넷째, 소지역 추정 시 아주 작은 값 또는 결측값(missing value) 처리에 대한 연구가 필요하다. 실업자 수와 같이 어떤 특성치가 아주 작은 값으로 조사되는 경우 소지역 추정은 그 값에 크게 영향을 받게 된다. 예를 들면, 어떤 소지역의 조사결과 표본규모가 아주 작은 경우, 표본에서 실업자가 한 명도 관측되지 않을 수 있다. 그렇게 되면 소지역의 직접추정치는 '0'이 된다. 즉, 실업자가 한 명도 없는 셈이다. 그러나 이는 조사가 잘못된 것이 아니라, 그 소지역의 표본규모가 작고 특성값의 관측이 희박한 경우에 발생하는 문제다. 따라서 소지역 추정에 이런 현상을 어떻게 처리할 것인지 고민해야 할 것이다.

다섯째, 소지역 추정을 위한 최소 표본규모를 찾을 필요가 있다. 추정을 위해 필요한 최소 규모의 표본을 해당 소지역에 할당할 수 있다면 최소의 노력과 예산으로 대규모의 표본조사 결과에 견줄 만한 추정결과를 얻을 수 있을 것이다.

여섯째, 소지역 통계의 사용 목적을 정확하게 설정할 필요가 있다. 이는 소지역의 정의(definition)와 직결되는 문제이기 때문이다. 행정구역 단위로 할 것인지 아니면 특정 목적단위(경제구역단위 등)로 할 것인지가 명백해야 한다. 이번 연구와 같이 모든 행정구역 단위를 소지역으로 정의한다 해도, 향후 소지역 추정을 위해서는 작성된 통계의 활용빈도

나 가치를 고려하여 최적의 단위를 설정해야 할 것이다. 적정 인구규모 또는 유사한 경제권역과 산업권역으로 나누는 것도 하나의 방법이 될 수 있다.

일곱째, 이 외에도 소지역 추정 연구에서 앞으로 해야 할 일들은 곳곳에 산재되어 있다. 본 연구는 실업자 총 수를 분석대상으로 삼았지만, 굳이 실업자 수를 추정하지 않아도 된다면 소지역의 실업률을 대상으로 재추정해 보아야 할 필요가 있다. 실업자 수는 고용통계부문(실업자 수, 실업률, 취업자 수, 취업률 등) 가운데 가장 변동이 크고 불안정한 수치이다. 오히려 실업자 수 추정에서 사용한 동일 모형을 실업률(취업률) 추정에 적용한다면 추정치의 정도 및 오차율을 훨씬 개선할 수 있을 것으로 연구자들은 기대한다. 뿐만 아니라, 전체 시군구를 대상으로 본 연구결과를 적용함으로써 종합적인 비교를 하는 것이 바람직하다 할 것이다.

소지역 추정과 관련하여 이용자들은 추정결과를 해석하는 데 있어서 특히 신중해야 할 부분이 있다. 소지역 추정결과를 사용할 때 가장 주의해야 할 것은 지역별 순위를 매기는 것이다. 일반적으로 사용자들은 추정치에 대해 각 지역별로 순위를 매겨서 지역 간 평가를 하는데 관심이 많다. 그러나 본문에서 설명한 바와 같이 추정치만으로 지역별 순위를 매기는 것은 매우 위험한 행동이다. 추정치는 추정오차를 동반하기 때문에 이것까지 고려해서 해석을 해야 한다. 가장 이해하기 좋은 방법은 신뢰구간을 이용하는 것이다. 두 지역의 추정치들에 대한 신뢰구간이 겹치지 않으면 그 두 지역이 충분히 차이가 있다고 해석을 하지만, 그렇지 않다면 그 지역들의 추정치에는 유의한 차이가 있다고 볼 수 없다. 이런 경우에 두 지역들의 순위가 다르다고 해석하는 것은 잘못된 것이다.

마지막으로 최소한 국가통계에 있어서 소지역 추정은 충분한 연구와 검토를 통해서 이루어져야 한다. 본 연구를 통해서도 우리는 소지역 추정 결과를 신뢰하기 위해 얼마나 많은 작업(보조정보 개선, 시계열비교, 충분한 시계열자료 확보, 벤치마크 등의 기준설정 문제, 소지역정의 문제 등)이 해결되어야 하는지 알 수 있다. 그렇다고 신뢰할 만한 추정결

과를 얻는 것이 그렇게 어려운 것도 아니다. 꾸준히 투자하고 연구를 지속한다면 소지역 추정에 의한 통계는 많은 예산을 투입하는 대규모 표본조사보다도 더 많은 가치가 있을 것으로 확신한다. 연구에 그치지 않고, 지속적으로 작성·활용하고 보완하는 과정을 거침으로써 추정통계의 활용 시기를 단축시킬 수 있을 것으로 기대한다. 추정통계에서 추정오차를 인정하는 것이 조사통계에서 다양한 조사오차를 인정하는 것과 다를 것이 없음을 이용자에게 끊임없이 이해시키는 노력도 통계청의 몫이 되어야 할 것이다.

본 연구에서 소개한 다양한 소지역 추정 방법들이 다른 표본조사에 응용됨으로써 소지역 추정에 관한 유용한 정보를 얻을 수 있기를 기대한다.

참고문헌

권혜자, 노현국(2007), “고용보험 실업률의 추이와 특징”, 한국고용정보원.

김달호(2005), “R과 Winbugs를 이용한 베이지안 통계분석”, 자유아카데미.

4대 사회보험정보연계센터(2006), 『4대 사회보험 업무담당자 전산교육 교재』.

조신섭, 손영숙(1999), “시계열분석”, 율곡출판사.

통계개발원(2008), “고용통계 소지역 추정 및 사후조사와 인구추계”, 일본총무성 통계국 방문 결과 보고.

통계청(2004), “모형 기반 소지역 추정 방법”(내부자료).

허명회(2007), “R을 활용한 탐색적 자료분석”, 자유아카데미.

허문열 외 4인(2007), “R과 통계계산”, 박영사.

中妻照雄(2003), ファイナンスのためのＭＣＭＣ法によるベイズ分析, 三譲經濟研究所.

Ambler, R., Caplan, D., Chambers, R., Kovacevic, M., and Wang, S. (2001), "Combining Unemployment Benefits Data and LFS Data to Estimate ILO Unemployment for Small Areas: an Application of a Modified Fay-Herriot Method", 영국 통계청.

Chatterjee, S., Lahiri, P., and Li, H. (2005), "On Small area prediction interval problems", ASA Section on Survey Research Methods, pp. 2831-2838.

Datta, G. S., Lahiri, P., Maiti, T., and K. L. Lu (1999), "Hierarchical Bayes estimation of unemployment rate for the states of the U.S.", Journal of the American Statistical Association. 94(448), pp. 1074-1082.

Gelfand, A.E., and Smith, A.F.M. (1991), "Gibbs sampling for marginal posterior expectations, Communications In Statistics-Theory and Methods, 20, 1747-1766.

Gelman, A., and Rubin, D.B. (1992), "Inference from iterative sumulation using multiple sequences. Statistical Science, 7, 457-472.

Fay, R.E., and Herriot, R.A. (1979), "Estimates of income for small places: An application of James-Stein procedures to census data", Journal of the American Statistical Association, 74, 269-277.

Hastings, D., Maine, N., Brown, G., and Cruddas, M. (2003), "Development of improved estimation methods for local area unemployment levels and rates", 영국통계청.

Henderson, C.R. (1950), "Estimation of genetic parameters", Annals of Mathematical statistics, vol. 21, 309-310.

Lahiri, P. (2003), "On the Impact of Bootstrap in Survey sampling and Small-area estimation", Statistical Science, 18(2), pp. 199-210.

Lee, S. E. and Shin, K. I. (2003), " Model-data based small area estimation", The Korean Communications in Statistics, 10(3), pp. 637-645.

Mohadjer, L., Rao, J.N.K., Liu, B., and Krenzke, T. (2007), "Hierarchical bayes small area estimates of adult literacy using unmatched sampling and linking models", Joing Statisical mettings, Salt Lake City, Utah, July 29-Audust 2, 2007.

ONS(2006), "Model-Based estimates of ILO Unemployment for LAD/UA in Great Britain Guide for users", July 2006.

Rao, J.N.K. (1999), "Some recent advances in model-based small area estimation", Survey Methodology, 25, 175-186.

Rao, J.N.K. (2003), "Small area estimation", Wiley.

Rao, J.N.K. (2007), "Jackknife and bootstrap methods for small area estimation", Joing Statisical mettings, Salt Lake City, Utah, July 29-Audust 2, 2007.

Rao, J.N.K. and Yu, M. (1994), "Small-area estimation by combining time-series and cross-sectional data", The Canadian Journal of Statistics, 22, 511-528.

Takabe, I. (2004), "Application of the small area estimation methods to the labour force survey: Prefecture level unemployment rate estimation", 日本 統計局 勞動力人口統計室.

Tiller, R.B. (1992), "Time Series Modeling of sample survey data from the U.S.

Current Population Survey", Journal of Official Statistics, 8(2), pp. 149-166.

You, Y., and Rao, J.N.K. (2002), "Small area estimation using unmatched sampling and linking models", The Canadian Journal of Statistics, 30, 3-15.

You, Y., Rao, J.N.K., and Gambino, J. (2003), "Model-Based Unemployment rate estimation for the Canadian Labour Force Survey: A Hierarchical bayes approach", Statistics Canada, 29(1), pp. 25-32.

< 부 록 >

〈부표 2-1〉 2007년 9월 자료에 대한 각 추정량 비교: 공간정보 이용, $\rho = 0.5$: 추정치, CV, ARB, R, P값

	조사구수(개)		실업자 수(명)				CV(%)				상대편향(%)			모형수렴R	
	지역 통계	직접 추정량	지역 통계	직접 추정량	FH HB1	TSCS HB2	지역 통계	직접 추정량	FH HB1	TSCS HB2	직접 추정량	FH HB1	TSCS HB2	FH HB1	TSCS HB2
수원	40	16	17,713	23,867	14,429	14,022	21.40	29.90	17.47	8.01	34.74	15.12	17.52	1.016	1.792
성남	49	15	13,161	17,279	12,851	12,871	16.90	30.58	17.97	8.76	31.29	37.30	37.21	1.015	1.435
의정부	38	6	8,807	2,677	2,945	2,049	16.10	71.22	37.65	28.72	69.60	65.47	75.98	1.021	1.064
안양	50	10	11,783	12,550	8,078	8,321	17.30	23.15	20.95	12.63	6.51	29.51	27.39	1.011	1.535
부천	50	13	14,119	6,833	10,555	10,650	15.20	49.76	16.12	9.97	51.61	38.56	38.00	1.009	1.788
광명	31	6	8,061	2,958	3,639	3,970	18.70	65.34	35.00	21.47	63.31	52.32	47.98	1.000	1.116
평택	40	8	8,145	6,356	4,081	3,819	15.30	54.86	27.62	23.25	21.96	69.22	71.20	1.003	1.202
안산	33	9	19,052	13,489	9,070	9,089	17.00	36.42	19.78	10.52	29.20	49.99	49.88	1.011	1.296
고양	40	13	11,441	13,220	11,828	11,522	23.90	45.08	18.54	9.95	15.55	13.55	15.79	1.012	1.535
남양주	40	11	7,006	3,260	4,443	3,907	14.30	49.51	24.76	23.46	53.47			1.005	1.479
시흥	40	5	6,000	8,834	3,846	2,946	16.80	26.19	25.19	30.40	47.24	45.79	58.48	1.003	1.184
용인	32	10	14,003	26,194	10,640	10,156	19.10	21.22	20.26	10.59	87.06	35.53	38.47	1.010	1.360
파주	31	4	8,684	1,486	4,459	3,948	17.60	106.50	32.07	27.11	82.89	33.36	40.99	1.013	1.888
이천	30	5	2,937	2,795	2,035	2,394	26.40	121.65	52.27	34.85	4.85			1.002	1.150
안성	30	5	2,267	505	1,208	1,272	26.50	101.95	61.82	62.22	77.74	58.56	56.36	1.001	1.338
김포	31	4	3,179	945	1,391	1,121	34.60	100.67	54.58	51.76	70.27	69.34	75.29	1.008	1.094
화성	30	6	3,317	3,460	3,615	3,736	26.20	78.87	36.64	27.37	4.31	54.56	53.04	1.003	1.281

<부표 2-1>의 계속

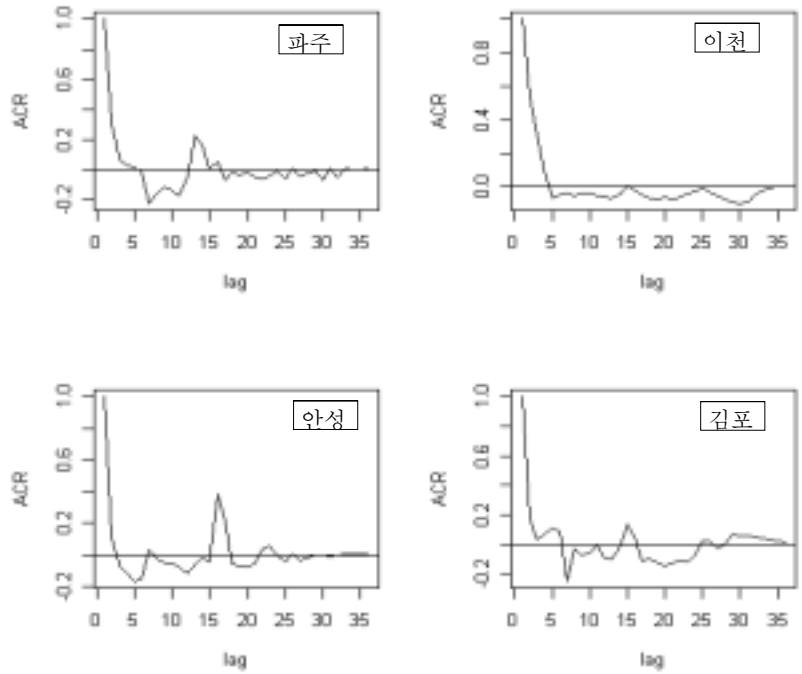
	조사구수(개)		실업자 수(명)				CV(%)				상대편향(%)			모형수렴R	
	지역 통계	직접 추정량	지역 통계	직접 추정량	FH HB1	TSCS HB2	지역 통계	직접 추정량	FH HB1	TSCS HB2	직접 추정량	FH HB1	TSCS HB2	FH HB1	TSCS HB2
여주	30	4	1,342	504	1,081	1,113	30.00	102.96	112.70	82.32	62.45	27.28	25.15	1.000	1.382
태백	25	3	516	1,421	484	740	57.90	101.28	262.91	120.00	175.37	50.63	130.63	1.005	1.223
전주	46	22	8,035	9,394	4,169	4,549	23.50	36.58	26.35	22.42	16.92	28.01	21.44	1.004	1.391
광양	46	6	1,254	679	763	913	32.70	71.50	81.58	68.69	45.88	30.67	16.97	1.003	1.134
창원	45	16	5,763	10,500	4,678	5,154	26.10	38.38	28.79	19.95	82.19	4.12	14.71	1.011	1.480
마산	60	13	2,321	4,234	3,870	3,625	32.30	60.09	33.31	28.29	82.42	32.70	36.96	1.004	1.604
진주	55	10	3,919	2,296	2,752	2,564	17.50	53.36	37.70	32.65	41.42	39.66	43.79	1.002	1.302
진해	39	4	1,866	5,770	1,188	1,581	30.10	58.70	90.88	53.14	209.22	27.50	3.57	1.001	1.244
통영	50	4	1,634	1,479	448	488	20.00	149.62	82.69	75.77	9.46	74.47	72.21	1.001	1.056
사천	50	4	2,178	2,010	892	1,269	21.50	106.07	133.41	73.66	7.70	49.75	28.56	1.002	1.241
김해	55	10	2,914	5,536	3,504	3,902	26.20	37.24	33.43	23.25	89.99	33.81	26.28	1.007	1.360
거제	50	6	1,873	1,635	847	896	22.10	84.33	82.75	58.08	12.71	45.86	42.74	1.001	1.156
밀양	50	4	1,166	360	383	459	23.20	95.51	85.46	70.53	69.16	81.97	78.36	1.002	1.049
양산	50	6	3,177	1,552	2,210	2,787	19.90	85.20	66	32.98	51.14	40.32	24.76	1.005	1.515
												모형적합		0.295	0.928

〈부표 2-2〉 2007년 12월 자료에 대한 각 추정량 비교: 공간정보 이용, $\rho = 0.5$: 추정치, CV, ARB, R, P값

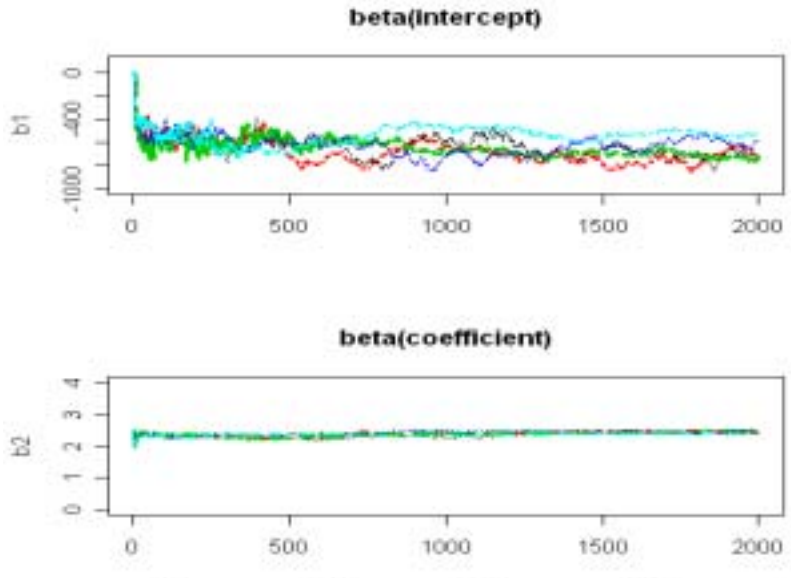
	조사구수(개)		실업자 수(명)				CV(%)				상대편향(%)			모형수렴R	
	지역 통계	직접 추정량	지역 통계	직접 추정량	FH HB1	TSCS HB2	지역 통계	직접 추정량	FH HB1	TSCS HB2	직접 추정량	FH HB1	TSCS HB2	FH HB1	TSCS HB2
수원	40	16	17,713	23,867	15,505	13,354	21.40	29.90	14.12	7.80	34.74	12.47	24.61	1.058	1.779
성남	49	15	13,161	17,279	13,828	11,921	16.90	30.58	14.06	9.24	31.29	5.07	9.42	1.052	1.417
의정부	38	6	8,807	2,677	4,795	3,435	16.10	71.22	19.86	24.66	69.60	45.55	61.00	1.015	1.363
안양	50	10	11,783	12,550	8,870	7,161	17.30	23.15	15.86	12.72	6.51	24.72	39.23	1.045	1.654
부천	50	13	14,119	6,833	11,835	9,652	15.20	49.76	13.42	7.89	51.61	16.18	31.64	1.050	1.226
광명	31	6	8,061	2,958	3,410	3,330	18.70	65.34	25.87	22.03	63.31	57.69	58.69	1.005	1.314
평택	40	8	8,145	6,356	5,132	4,391	15.30	54.86	21.30	26.71	21.96	36.99	46.09	1.018	1.492
안산	33	9	19,052	13,489	9,692	8,326	17.00	36.42	15.55	11.15	29.20	49.13	56.30	1.047	1.420
고양	40	13	11,441	13,220	12,677	10,583	23.90	45.08	14.24	9.91	15.55	10.81	7.50	1.054	1.917
남양주	40	11	7,006	3,260	5,506	4,087	14.30	49.51	17.54	21.09	53.47	21.41	41.67	1.022	1.491
시흥	40	5	6,000	8,834	5,222	3,882	16.80	26.19	22.42	26.68	47.24	12.96	35.30	1.026	1.902
용인	32	10	14,003	26,194	11,713	9,936	19.10	21.22	15.74	9.51	87.06	16.35	29.05	1.049	1.593
파주	31	4	8,684	1,486	3,259	2,359	17.60	106.50	26.88	27.03	82.89	62.47	72.83	1.006	1.178
이천	30	5	2,937	2,795	1,953	2,157	26.40	121.65	50.80	46.87	4.85	33.49	26.56	1.001	1.706
안성	30	5	2,267	505	819	591	26.50	101.95	56.90	76.76	77.74	63.86	73.94	1.009	1.302
김포	31	4	3,179	945	1,587	1,396	34.60	100.67	42.17	54.52	70.27	50.08	56.08	1.004	1.341
화성	30	6	3,317	3,460	3,702	3,349	26.20	78.87	26.29	30.25	4.31	11.60	0.96	1.008	1.538

<부표 2-2>의 계속

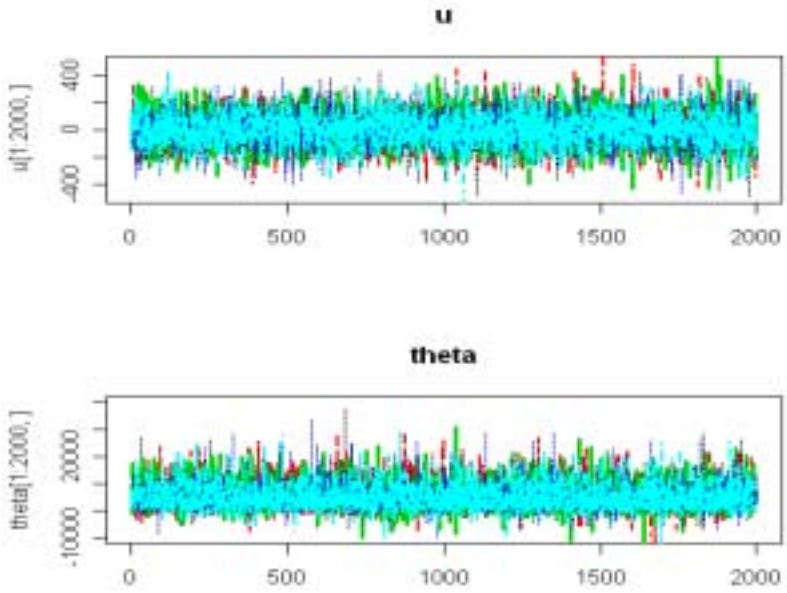
	조사구수(개)		실업자 수(명)				CV(%)				상대편향(%)			모형수렴R	
	지역 통계	직접 추정량	지역 통계	직접 추정량	FH HB1	TSCS HB2	지역 통계	직접 추정량	FH HB1	TSCS HB2	직접 추정량	FH HB1	TSCS HB2	FH HB1	TSCS HB2
여주	30	4	1,342	504	527	575	30.00	102.96	82.11	82.05	62.45	60.73	57.13	1.002	1.107
태백	25	3	516	1,421	-1	619	57.90	101.28	-60,616	119.47	175.37	100.28	19.88	1.006	1.022
전주	46	22	8,035	9,394	7,342	6,287	23.50	36.58	17.01	13.48	16.92	8.63	21.75	1.027	1.362
광양	46	6	1,254	679	532	749	32.70	71.50	80.42	63.90	45.88	57.58	40.27	1.005	1.115
창원	45	16	5,763	10,500	5,327	5,068	26.10	38.38	22.32	19.37	82.19	7.57	12.06	1.024	1.580
마산	60	13	2,321	4,234	4,169	3,937	32.30	60.09	23.45	23.92	82.42	79.64	69.61	1.009	2.027
진주	55	10	3,919	2,296	2,738	2,432	17.50	53.36	26.81	29.29	41.42	30.13	37.96	1.005	1.326
진해	39	4	1,866	5,770	1,255	1,568	30.10	58.70	85.07	50.03	209.22	32.73	15.97	1.003	1.237
통영	50	4	1,634	1,479	728	921	20.00	149.62	126.68	89.45	9.46	55.47	43.66	1.000	1.189
사천	50	4	2,178	2,010	606	1,108	21.50	106.07	154.78	80.53	7.70	72.19	49.12	1.003	1.524
김해	55	10	2,914	5,536	4,442	4,487	26.20	37.24	21.91	21.42	89.99	52.44	53.96	1.013	1.310
거제	50	6	1,873	1,635	1,411	1,695	22.10	84.33	54.81	51.99	12.71	24.65	9.52	1.000	1.404
밀양	50	4	1,166	360	357	432	23.20	95.51	87.44	75.66	69.16	69.35	62.97	1.002	1.133
양산	50	6	3,177	1,552	1,611	1,932	19.90	85.20	47	28.34	51.14	49.30	39.19	1.000	1.132
												모형적합		0.295	0.980



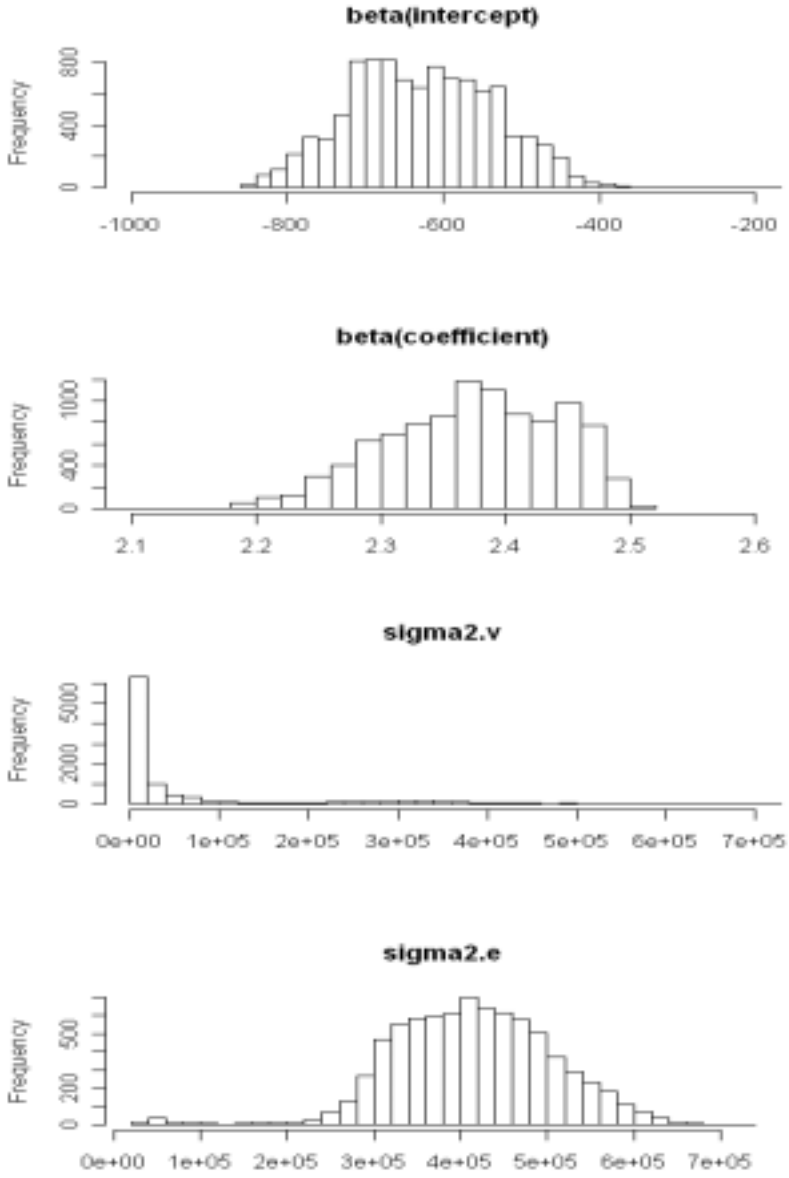
[부그림 2-1] 자기상관 그래프: 임의의 4개 지역



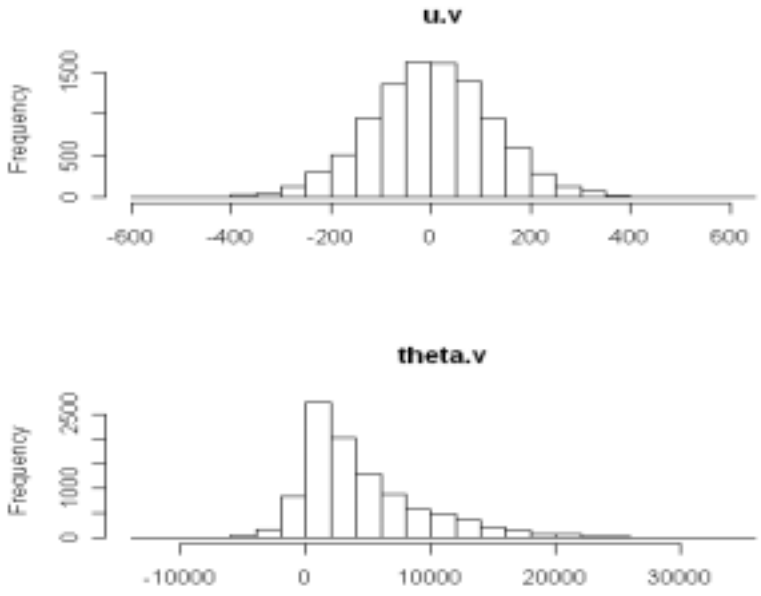
[부그림 2-2] 각 모수들의 샘플 경로


u
u[1:2000,]
-400
0
400
0
500
1000
1500
2000
theta
theta[1:2000,]
-10000
20000
0
500
1000
1500
2000

[부그림 2-2]의 계속



[부그림 2-3] 각 모수들의 히스토그램


u.v
Frequency
0
500
1500
-600
-400
-200
0
200
400
600
theta.v
Frequency
0
1000
2500
-10000
0
10000
20000
30000

[부그림 2-3]의 계속