1. 서론

1.1 정의 및 동기

소지역(small area) 혹은 국소지역(local area)이란 통상적으로 작은 지리적 지역, 예컨대, 시, 군, 지방자치단위, 센서스구획 등을 의미한다. 때로는 큰 지리적 지역 내에서의 특정 나이-성별-인종의 사람들로 구성된 작은 부모집단(subpopulation)을 일컫는 소영역(small domain)을 의미하기도 한다. 흔히 영역의 표본 크기는 영역의 크기와 더불어 증가하는 경향이 있다. 그러나 이는 항상 성립하는 것은 아니다. 예컨대, 미국의 NHANES III(Third National Health and National Examination Survey) 조사는 인종과 나이에 따라 분류되는 영역에서 신뢰할 만한 추정값을 생산하도록 설계되었다. 이 조사에서 주(state)는 소지역으로 간주될 수 있다. 왜냐하면 많은 주에서 지역관련 표본의 크기가 작기 때문이다.

소지역통계는 역사적으로 이미 오래 전부터 사용되었다. 예를 들면, 11세기 영국과 17세기 캐나다에서는 이미 센서스나 행정자료에 근거한 소지역통계가 존재했다. 인구통계학자들은 센서스 이후의 연도에서 인구 및 다른 특성치의 소지역추정을 위해 다양한 간접적 방법을 오래 전부터 사용해 왔다. 전형적으로 샘플링은 전통적 인구 통계학적 방법에서 사용되지 않았다.

최근 소지역통계에 대한 수요는 공공부문과 민간부문 모두에서 신뢰할 수 있는 소지역통계에 대한 요구와 필요가 증가함으로 인해 세계적으로 크게 증가하고 있는 추세이다. 이는 무엇보다도 정부예산의 분배와 지역적 계획수립에서 정책과 프로그램을 결정하는데 소지역통계의 사용이 증가했기 때문이다. 정부에 의한 조례나 법령의 제정으로 소지역 통계에 대한 필요가 증가하게 되었고 이러한 경향은 계속되는 추세이다. 뿐만 아니라, 민간부문에서도 소지역통계에 대한 수요가 더욱 커지게 되었다. 왜냐하면 기업의 의사결정에서, 특히 작은 기업체와

관련된 의사결정이 지역의 사회 경제적 조건에 크게 의존하기 때문이다.

소지역추정은 중앙 및 동유럽 국가들과 옛 소련 연방 국가들에서의 과도기적 경제상황에서 특별한 관심사이다. 1990년대에 이들 국가들은 중앙 집권화된 의사결정으로부터 벗어나게 되었고, 그 결과 지금은 표본조사를 사용하여 대지역뿐만 아니라 소지역에 대한 추정값을 작성하고 있다.

표본조사 자료는 큰 지역이나 영역에 대한 합이나 평균에 대해서는 신뢰할 만한 추정값를 제공한다. 그러나 소지역추정을 위해서 각 소지역에 해당되는 표본자료만 가지고 직접조사 추정량을 계산하면 소지역의 표본크기가 너무 작기 때문에 표준오차가 매우 커지게 되어 추정값이 신뢰성을 잃게 된다. 소지역에서의 표본크기가 작은 이유는 예컨대 전국 규모의 조사에서 주어진 정밀도에 맞추어 표본의 크기를 정하면 비용의 제한 등으로 인해 소지역에는 작은 수의 표본만이 할당되기 때문이다.

센서스는 통상 매 5년 혹은 10년에 한번 제한된 수의 항목에 대해 자세한 정보를 제공한다. 행정자료는 시간적으로 더 자주 자료를 제공할 수 있으나 범위가 좁은 단점을 가진다. 한편, 표본조사는 센서스에 비해 적은 비용으로 짧은 시간에 넓은 범위의 과제에 대한 정보를 제공한다. 표본조사로부터 얻은 자료를 사용하여 큰 지역에 대한 신뢰할 만한 직접추정량을 구할 수 있으나, 소지역에서는 표본의 크기가 작기 때문에 직접통계량이 소지역에 대해서는 충분한 정밀도를 제공하지 못한다. 이러한 문제점을 개선하기 위해 관련된 지역들로부터 “정보를 빌려오는” (“borrow strength”) 간접추정량을 찾아서 효과적으로 표본크기를 증가시켜 결국 정밀도를 증가시키자는 것이다.

암시적 모형(explicit model) 가정에 근거한 전통적인 간접추정량은 일반적으로 설계 기반(design-based)이며 간접추정량의 설계분산(즉, 표집설계에 의해 유도된 확률분포에 관한 분산)은 보통 직접추정량의 설계분산에 비해 상대적으로 작다. 그러나 간접추정량은 보통 설계편향(design bias)이 내재되어 있고, 이 설

계편향은 전체 표본크기가 증가할 때 감소하지 아니한다. 그럼에도 불구하고 평균제곱오차(MSE)의 감소가 소지역추정에서 간접추정량을 사용하는 주된 이유임은 분명하다.

모형에 포함된 보조변수들에 의해 설명되지 아니하는 지역간 변동(between area variation)에 해당하는 랜덤 지역효과(random area-specific effects)를 포함하는 명시적 연결 모형(explicit linking model)을 “소지역모형”이라 부른다. 소지역모형을 사용하여 구한 간접추정량을 “모형기반 추정량”이라 부른다. 소지역모형은 크게 두 가지 형태로 분류된다. 첫째는 소지역 직접추정량을 지역에 관련된 공변량과 연결시키는 총계(혹은 지역)수준 모형(aggregate (or area) level model)으로 이러한 모형은 단위수준자료(unit (or element) level data)를 얻을 수 없을 때 필요하다. 둘째로 연구변수의 단위 값을 단위 수준의 공변량과 연관시키는 단위수준 모형(unit level model)이다.

간접추정량이 사용될 때는 명시적 소지역모형에 근거해야만 한다는 인식이 이제는 널리 확산되었다. 이러한 모형은 추정과정에 관련된 자료를 통합시키는 방법을 정의한다. 소지역추정에서 모형에 근거한 접근은 다음과 같은 여러 가지 장점을 제공한다.

(1) 가정된 모형 하에서 “최적” 추정량을 구할 수 있다.

(2) 전통적인 간접추정량과 함께 흔히 사용되는 (소지역들에 대해 평균을 취한) 대역적 측도(global measure)와는 달리 각 소지역의 추정량과 관련된 구체적 소지역 변동(variability)의 측도를 구할 수 있다.

(3) 표본자료를 사용하여 모형의 타당성을 조사할 수 있다.

(4) 반응변수의 특성과 자료구조의 복잡성에 따라 여러 종류의 모형을 사용할 수 있다.

1.2 사례

1.2.1 소지역 소득 및 빈곤 추정

1993년 11월 21일 미국 하원을 통과한 법안 Bill H.R. 1645의 결과로서 현재 미국 Census Bureau에서 진행중인 소지역추정 문제를 살펴보자. 이 법안에 의하면 미국 상무장관은 1996년부터 적어도 매 2년마다 미국 내 빈곤발생률과 관련된 자료를 생산하여 발표해야 한다. 구체적으로 이 법안에 의하면 "가능한 범위까지" 상무장관이 주, 군, 지방정부관할구역, 그리고 학군(school districts)에 대한 빈곤의 추정값을 작성하도록 한다는 것이다. 학군에 대해서는 5-17세 연령의 빈곤아동의 수를 추정한다. 또한 65세 이상의 빈곤한 개인의 수에 대한 주별, 군별 추정값을 작성하도록 명시하고 있다. 지금까지는 매 10년마다 실시하는 센서스만이 이러한 작은 지리적 지역에 대한 개인, 가족 그리고 가구에 대한 소득 분포와 빈곤자료에 대한 정보를 제공하였다.

소지역에 대한 이러한 통계는 민간부문뿐만 아니라 주나 지방자치단위에서의 정책 결정자를 포함한 광범위한 고객에 의해 사용된다. 예를 들면 연방과 주 정부의 기금을 분배하는 일이다. 연방정부 기금이 1994 회계연도에 300억 달러이상이었다. 1990년 센서스 자료가 1990년대 후반의 상황을 충분히 반영하지 못하기 때문에 1990년대 후반에서의 경제적 상황과 관련하여 1990년 센서스 자료를 사용하는 것은 바람직하지 못하다.

미국 Census Bureau는 매 10년마다 실시하는 센서스로부터의 추정값을 개선하기 위해서 SAIPE(Small Area Income and Poverty Estimates) 프로그램에 대한 연구와 개발을 1990년대 중반에 시작하였다. 소득연도(income year) 1993년에서 빈곤가정의 5-17세 연령의 아동 수에 대한 주와 군의 첫 SAIPE 추정값이 1997년 초에 발표되었다. 소득연도 1995년에도 빈곤가정의 5-17세 연령의 아동 수에 대한 주와 군의 추정값이 1999년 초에 발표되었다.

주와 군의 SAIPE 추정값은 가구소득의 중위값, 빈곤자의 수, 5세 이하 빈곤 아동의 수(주 단위만), 5-17세 연령의 빈곤아동의 수, 18세 이하의 빈곤자의 수 등을 포함한다. 또한 1999년 초에 1,400개의 학군에 대해 소득연도 1995년의 학령에 달한 빈곤 아동의 수에 대한 추정값을 발표하였다.

Census Bureau 추정값은 각종 조사와 10년 단위의 센서스와 행정기록으로부터의 자료를 결합하는 통계적 모형 기법을 사용하여 개발되었다. 따라서 추정값은 간접추정값이며, 추정값의 질(quality)은 적절한 통계적 모형의 선택에 의존한다. SAIPE 추정값은 하나의 자료출처로부터 직접적으로 얻어질 수 없다. 왜냐하면 현재까지 가능한 어떠한 자료출처도 직접적인 소지역추정값에 대해 충분히 신뢰할 만한 최신의 정보를 제공하지 못했기 때문이다.

1.2.2 연방-주 협동프로그램

연방-주 협동프로그램(Federal-State Cooperative Program; FSCP)은 1967년 미국 Census Bureau에 의해 시작된 지역인구 추정을 위한 연방 정부와 주 정부간 협동 프로그램이다. 이 프로그램의 기본 목표는 지역 간 비교가능한 양질의 일관된 일련의 군의 인구 추정값을 제공하는 것이다. 49개 주(Massachusetts주 제외)가 현재 이 프로그램에 참가하고 있으며, 이 프로그램 하에서 각 주의 지정된 조사기관과 Census Bureau가 함께 일하고 있다. FSCP의 몇몇 회원 주들은 현재 군별 추정값 외에 군보다 작은 단위(subcounty)의 인구 추정값도 생산하고 있다. FSCP는 센서스 이후 인구 추정값을 만드는데 사용될 수 있는 각종 자료를 Census Bureau에 제공하기 때문에 Census Bureau의 센서스 이후 인구 추정 프로그램에서 중요한 역할을 하고 있다. 주의 조사기관에 의해 작성된 예비 군별 추정값에 대한 검토와 비판을 통해 Census Bureau 역시 주의 관련 조사기관에 도움을 주고 있다.

1.2.3 소지역 개인소득 추정

Fay와 Herriot(1979)은 여러 소지역의 개인소득(Per Capita Income; PCI) 추정을 고려하였다. 미국 Census Bureau는 재무부(Treasury Department)의 일반 세입교부 프로그램(General Revenue Sharing Program)하에서 재원을 지원받는 주와 지방 정부에 대한 개인소득 추정값과 기타 통계를 제공한다. 재무부는 이 통계를 사용하여 해당하는 주의 할당을 분배함으로 다른 주 내에 있는 지방정부에 대한 할당을 결정한다.

1970년대 초 Census Bureau는 (20 퍼센트 표본에 근거한) 1969년 PCI의 1970년 센서스 추정값에 현재 년도의 행정적 PCI 추정값과 1969년의 행정적 PCI의 비를 곱하여 현재 년도의 PCI 추정값을 결정했다. 그러나 1970년에 대충 39,000개의 지방정부단위 중에서 약 15,000개가 500명 이하의 주민을 가진 소지역들이라는 문제에 직면했으며, 이러한 소지역들에 대한 PCI 추정값의 표본오차가 높았다. 예컨대, 500명의 주민을 가진 지역에서는 변동계수가 약 13%인 반면 100명의 주민을 가진 지역에서는 변동계수가 약 30%로 증가하였다. 따라서 Census Bureau는 처음에는 이들 소지역에 대한 센서스 추정값을 제쳐두고 대신 해당하는 군별 PCI 추정값을 사용하였다. 그러나 이 해결책은 많은 소지역에 대한 PCI의 센서스 추정값이 표본오차를 고려할 때 해당하는 군별 PCI 추정값과 매우 다르기 때문에 바람직하지 못하다.

Fay와 Herriot(1979)은 경험적 베이즈 방법을 사용한 보다 나은 추정값을 제안하였고, 경험적 베이즈 추정값이 센서스 표본 추정값이나 군 평균보다 더 작은 평균오차를 가진다는 경험적 증거를 제시하였다. 소지역에 대해 제안된 추정값은 센서스 표본 추정값 및 종속변수로 PCI의 표본 추정값과 독립변수로 표본 추정값의 군 평균, 1969년 납세신고 자료, 1970년 센서스 주택자료를 사용하여 선형회귀모형을 적합시켜 얻은 “합성” 추정값의 가중평균이다. Fay-Herriot 방

법은 1974년 소지역에 대한 개선된 추정값을 작성하기 위해 Census Bureau에 의해 채택되었다.

1.2.4 옥수수와 콩의 재배면적 예측

Battese, Harter and Fuller(1988)는 농장-면담 자료와 LANDSAT 위성자료를 연관지어 아이오와 주 중북부 지역의 12개 군에 대한 옥수수와 콩 재배면적을 추정하였다. 각 군을 면적 구획으로 나누고, 표본 구획을 설정하여 미국 농무부(United States Department of Agriculture; USDA) 통계보고 서비스 현장 직원이 농장주와의 면담을 통해 12개 군의 37개 표본 구획(구획은 약 150 헥타르)에서의 옥수수와 콩의 재배면적을 확인하였다. 이 때 각 군에 속한 표본 구획의 수는 1에서 6이다. 보조자료는 1978년 8월과 9월 동안의 LANDSAT 위성자료에 근거하여 USDA 절차를 사용하여 각 군의 표본 구획을 포함한 모든 면적 구획에 대해 얻어진 옥수수와 콩으로 분류된 화소(pixel; 약 0.45 헥타르에 해당하는 사진 요소)의 수 형태이다. Battese, Harter and Fuller(1988)는 랜덤 소지역 효과와 구획 수준의 자료를 포함하는 지분오차회귀모형(nested error regression model)을 고려하였으며, 전통적인 분산성분 접근방법을 사용하여 옥수수와 콩의 군별 재배면적에 대한 경험적 최량선형불편예측(empirical best linear unbiased prediction; EBLUP) 추정값을 구하였다. 분산성분 추정에 관련된 불확성을 고려함으로 추정값의 평균제곱오차(MSE)의 추정값을 구하였다.

1.2.5 주별 4인 가구 소득의 중앙값 추정

전국, 주별, 군별, 소지역별 4인 가구 소득의 중앙값은 흔히 정부의 여러 가지 정책 결정을 위해 필요하다. 미국 보건복지부(Department of Health and Human Services; HHS)는 저소득 가구를 위한 에너지 보조 프로그램을 공식화

할 때 주별(50개 주와 DC) 4인 가구 소득의 중앙값에 대한 추정값이 필요하다. 이 추정값은 매년 Census Bureau에 의해 보건복지부에 제공된다.

1970년대 후반부터 Census Bureau가 사용하고 있는 방법은 회귀분석 방법이다. 이 방법은 3가지 출처의 자료를 사용하는데, 기본 출처는 전년도 4인 가구 소득의 중앙값을 제공하는 경상인구조사(Current Population Survey; CPS)의 3월 표본에 대한 년도별 인구통계학적 보충이다. 두 번째 자료는 매 10년마다 조사되는 센서스 전년도에 대해 센서스로부터 구한 4인 가구 소득의 중앙값이다. 세 번째 자료는 미국 상무부(Department of Commerce)의 BEA(Bureau of Economic Analysis)에 의해 매년 조사되는 1인당 소득(PCI)의 추정값이다.

CPS 추정값의 직접적 사용은 작은 표본크기로 인해 변동계수가 크기 때문에 보통 바람직하지 않다. 반면에 센서스 추정값은 거의 무시할 수 있는 표본오차를 가진다고 믿어지므로 년도별 소득의 중앙값 추정에 CPS 추정값과 연결하여 사용될 수 있다. 센서스 추정값은 다른 추정값에 대한 평가기준이 되는 “gold standard″로 사용될 수 있다. 그러나 이러한 비교는 센서스 전년도에서만 가능하다. CPS 추정값과는 달리 PCI 추정값은 표본 기법을 사용하여 구한 것이 아니기 때문에 관련된 표본오차가 없다.

현재 Census Bureau가 사용하는 방법은 Fay(1987)에 의해 제안된 이변량 회귀모형(bivariate regression model)을 사용하고 있다. 이 방법은 주목적이 4인 가구 소득의 중앙값 추정이지만 4인 가구의 중앙값 소득과 더불어 3인 가구와 5인 가구의 중앙값 소득도 사용한다. 각 주에 대해 CPS 직접추정값에 근거하여 3인, 4인, 5인 가구의 중앙값 소득을 구하는 방법은 $2,500의 구간으로 범주화된 표로 작성한 소득을 선형보간법을 사용하여 보정한다. 각 주에 대한 기본 자료 셋은 이변량 확률 벡터로서, 한 성분은 4인 가구의 CPS 중앙값 소득이고 다른 성분은 가중값 0.75와 0.25를 사용한 3인 가구와 5인 가구의 CPS 중앙값 소득의 가중평균이다.

Census Bureau가 사용하는 회귀 방정식은 절편항 외에 독립변수로 기본 년도 센서스 중앙값(b)과 4인 가구 및 가중값 0.75와 0.25를 사용한 3인 가구와 5인 가구의 가중평균에 대한 조정된 센서스 중앙값(c)을 사용한다. 여기서 센서스 중앙값(b)은 가장 최근에 이루어진 센서스에서의 4인 가구 소득 중앙값을 나타내며, 센서스 중앙값(c)은 다음 공식으로부터 얻어진다.

$$\text{조정된 센서스 중앙값}(c) = \frac{\text{BEA PCI(c)}}{\text{BEA PCI(b)}} \times \text{센서스 중앙값}(b)$$

여기서 BEA PCI(c)와 BEA PCI(b)는 현재 년도 c와 기본 년도 b에 대해 BEA에 의해 생산된 PCI 추정값을 나타낸다. Fay(1987)에 의하면 위의 공식은 현재 년도의 조정된 중앙값을 구하기 위해 BEA PCI에서의 증가비율에 의해 기본 년도 센서스 중앙값을 조정하고자 한 것이다. 기본 년도 센서스 중앙값(b)을 독립변수에 포함시킨 이유는 현재 년도 중앙값 소득을 추정할 때 BEA PCI에서의 변화효과에 의해 과잉설명 가능성을 조정하기 위해서이다.

결국 현재 중앙값 소득의 CPS 표본 추정값과 이에 해당하는 회귀 추정값의 가중평균을 구한다. 이 가중 추정값은 사전분포의 분산을 구하기 위해 다소 특별히 고안된(ad hoc) 추정량을 사용한 경험적 베이즈 추정 절차를 사용한다.

Datta et al.(1996)은 Fay(1987)의 방법을 확장하여 관측값의 주변분포에 근거한 분산성분의 최대우도추정량(MLE)을 구함으로 경험적 베이즈 절차를 개선하였다. 뿐만 아니라 동일한 추정 문제에 대한 계층적 베이즈 절차를 제안하였다. Ghosh, Nangia와 Kim(1996)은 CPS가 매년 반복되는 점에 착안하여 베이지안 시계열 모형을 사용한 계층적 베이지안 절차를 제안함으로 주별 4인 가구 중앙값 소득에 대한 보다 좋은 추정값을 구하였다.

1.2.6 암 사망률 추정

작은 혹은 평균 크기의 도시에서의 폐암에 의한 사망의 연간 빈도는 아주 낮을 뿐만 아니라 한 도시로부터의 정보는 매우 제한적이다. 모형을 사용하여 다른 사망률을 가진 여러 도시로부터의 정보를 결합시키면 개별 도시에 기초한 원비율(raw rates)보다 사망률의 참값에 대해 좋은 추정값을 구할 수 있다.

원비율은 매년 보고되는 인구 10만명당 발생수에 대한 연간 비율이다. 이는 안정적이지 못하므로 특히 작은 도시에서 하나 혹은 둘의 사망 건수 차이가 원비율에 큰 영향을 끼친다.

Tsutakawa(1985)는 Missouri 주의 84개의 대도시에서 45-64세 연령의 남자 폐암사망률을 고려했다. 여기서 Y_i를 1972-1982년 사이 i번째 도시에서의 폐암 사망자 수라 두고, n_i를 도시의 크기라 하자. 이 경우 모형은 다음과 같다.

(i) $Y_i \sim \text{Poisson}(n_i p_i)$

(ii) $\theta_i = \log \frac{p_i}{1-p_i} \sim N(\mu, \sigma^2)$

(iii) $\mu \sim \text{Uniform}(-\infty, \infty)$; $\sigma^2 \sim$ Inverse Gamma

이는 폐암사망률의 참값이 어떤 분포에 따라 도시(소지역)별로 랜덤으로 변하는 랜덤효과모형으로서 소지역추정에서 흔히 사용되는 모형의 한 예이다.

최근 암이나 백혈병같은 특정 질병에 대한 발병률이나 사망률의 지역적 분포를 지도상에 표시한 질병지도(disease mapping) 작성이 보건관련 분야에서 중요한 관심사이다. 다른 지리적 지역들에 걸친 상대위험도(relative risk)에 대한 질병지도 작성이 특정한 질병이 다발적으로 발생한 집락을 찾거나 특정 질병을 유발하는 환경적 결정자를 파악하는데 도움이 되기 때문이다. 질병지도 작성의 근거가 되는 추정값이 결국 소지역추정 방법에 의해 작성된다.

2. 경험적 베이즈 추정

2.1 경험적 BLUP와 경험적 베이즈

m개의 소지역을 $1, \cdots, m$이라 나타내자. 이 때 관심 모수는 $\Theta_1, \cdots, \Theta_m$으로 소지역 총계 혹은 평균을 나타낸다. 직접조사추정량(direct survey estimator) $\hat{\Theta}_1, \cdots, \hat{\Theta}_m$이 주어졌다고 하자. 뿐만 아니라 소지역별 보조자료(auxiliary data) $\boldsymbol{x}_i = (1, x_{i1}, \cdots, x_{ip})'$ $(i = 1, \cdots, m)$가 사용 가능하다고 하자.

다음과 같은 랜덤효과모형을 고려하자. (Fay와 Herriot(1979) 참조).

$$\hat{\Theta}_i = \Theta_i + e_i ;$$
$$\Theta_i = \boldsymbol{x}_i' \boldsymbol{b} + u_i$$

여기서 $\boldsymbol{b} = (b_0, b_1, \cdots, b_p)$는 곱변량 $\boldsymbol{x}_i$에 대한 회귀계수이다. 표본오차 e_i는 서로 독립이고 평균이 0이고 분산이 V_i인 정규분포를 따른다고 가정하며, 보통 V_i는 기지의 값으로 가정한다. 이 가정에 대해서는 흔히 중심극한정리(central limit theorem)로 정당화된다. 모형오차 u_i는 서로 독립이며, 평균이 0이고 분산이 τ^2인 정규분포를 가정한다. 위의 두 식을 결합하면 모형을 다음과 같이 표현할 수 있다.

$$\hat{\Theta}_i = \boldsymbol{x}_i' \boldsymbol{b} + u_i + e_i, \quad i = 1, \cdots, m$$

위의 모형은 고정효과 $\boldsymbol{b}$와 랜덤효과 u_i를 가지는 혼합모형(mixed model)의 일종이다. 이 모형을 흔히 Fay-Herriot 모형이라 한다.

우리의 관심은 $\boldsymbol{b}$를 추정하는데 있지 않고 $\Theta_i = \boldsymbol{x}_i' \boldsymbol{b} + u_i (i = 1, \cdots, m)$를 추정(예측)하는데 있다. 알려진 표본오차 분산 V_i를 대각요소로 가지는 행렬을 $\boldsymbol{V} = Diag(V_1, \cdots, V_m)$로 나타내며, 보조정보를 $\boldsymbol{X}' = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m)$의 설계행렬로 나타내자. $\hat{\Theta} = (\hat{\Theta}_1, \cdots, \hat{\Theta}_m)'$이라 할 때, 모형을 행렬로 나타내면 다음과 같다.

$$\hat{\Theta} = \boldsymbol{Xb} + \boldsymbol{u} + \boldsymbol{e}$$

여기서 $\boldsymbol{u} = (u_1, \cdots, u_m)'$와 $\boldsymbol{e} = (e_1, \cdots, e_m)'$는 서로 독립이며, $\boldsymbol{u}$는 평균벡터가 $\boldsymbol{0}$이고 분산공분산 행렬이 $\tau^2 \boldsymbol{I}$인 다변량 정규분포를 따르며, $\boldsymbol{e}$는 평균벡터가 $\boldsymbol{0}$이고 분산공분산 행렬이 $\boldsymbol{V}$인 다변량 정규분포를 따른다. 여기서 $\mathrm{rank}(\mathbf{X}) = \mathrm{p} < \mathrm{m}$으로 가정하자. $\boldsymbol{Xb} + \boldsymbol{u}$의 최량예측추정량(best predictor)은 $E[\boldsymbol{Xb} + \boldsymbol{u} \mid \hat{\Theta}]$이다. 만약 정규분포 가정이 없다면 이는 최량선형예측추정량(best linear predictor))이다.

정규분포 가정 하에서 다음이 성립한다.

$$\begin{bmatrix} \hat{\Theta} \\ \boldsymbol{Xb} + \boldsymbol{u} \end{bmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{Xb} \\ \boldsymbol{Xb} \end{pmatrix}, \begin{pmatrix} \boldsymbol{V} + \tau^2 \boldsymbol{I} & \tau^2 \boldsymbol{I} \\ \tau^2 \boldsymbol{I} & \tau^2 \boldsymbol{I} \end{pmatrix} \right]$$

따라서 우리가 구하고자 하는 것을 다음과 같이 얻을 수 있다.

$$E[\boldsymbol{Xb} + \boldsymbol{u} \mid \hat{\Theta}] = \boldsymbol{Xb} + \tau^2 \boldsymbol{I}(\boldsymbol{V} + \tau^2 \boldsymbol{I})^{-1}(\hat{\Theta} - \boldsymbol{Xb})$$

$$= \begin{bmatrix} (1 - B_1)\hat{\Theta}_1 + B_1 \boldsymbol{x}_1' \boldsymbol{b} \\ \vdots \\ (1 - B_m)\hat{\Theta}_m + B_m \boldsymbol{x}_m' \boldsymbol{b} \end{bmatrix}$$

여기서 $B_i = \frac{V_i}{\tau^2 + V_i}$ $(i = 1, \cdots, m)$이다. 결국 우리가 구한 최량예측추정량은 직접추정량 $\hat{\theta}_i$와 회귀추정량 $\boldsymbol{x}_i'\boldsymbol{b}$의 가중합으로 표현된다.

위의 모형을 베이지안 형식으로 표현하면 다음과 같다.

i. $\hat{\theta}_i | \theta_i \overset{ind}{\sim} N(\theta_i, V_i)$

ii. $\theta_i \overset{ind}{\sim} N(\boldsymbol{x}_i'\boldsymbol{b}, \tau^2)$

이 모형에서 사후분포(posterior distribution)를 구하면 아래와 같다.

$$\theta_i | \hat{\theta}_i \overset{ind}{\sim} N[(1 - B_i)\hat{\theta}_i + B_i\boldsymbol{x}_i'\boldsymbol{b}, V_i(1 - B_i)]$$

사후분포의 사후평균 $(1 - B_i)\hat{\theta}_i + B_i\boldsymbol{x}_i'\boldsymbol{b}$이 미지의 모수 $\boldsymbol{b}$와 τ^2을 포함하므로 이를 $\hat{\theta}_i$들의 주변분포(marginal distribution)로부터 추정하여 대입하면 경험적 베이즈 추정량을 얻을 수 있다. 이 경우 $\hat{\theta}_i$의 주변분포는 다음과 같다.

$$\hat{\theta}_i \overset{ind}{\sim} N(\boldsymbol{x}_i'\boldsymbol{b}, \tau^2 + V_i)$$

$\boldsymbol{b}$와 τ^2을 추정하는 첫 번째 방법은 다음 두 식의 해를 반복적으로 구하는 것이다. (Fay와 Herriot, 1979 그리고 Morris, 1983 참조).

(i) $\tilde{\boldsymbol{b}} = (\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}^{-1}\hat{\theta}$

(ii) $\sum_{i=1}^{m}(\hat{\theta}_i - \boldsymbol{x}_i'\tilde{\boldsymbol{b}})^2 / (\tau^2 + V_i) = m - p$

여기서 $\boldsymbol{D} = \mathrm{Diag}(V_1 + \tau^2, \cdots, V_m + \tau^2)$이다.

위 식의 아이디어는 다음과 같다. 먼저 τ^2을 안다고 할 때 $\boldsymbol{b}$의 최량비편향

추정량(best unbiased estimator)은 가중최소제곱추정량(weighted least squared estimator) $\tilde{\boldsymbol{b}}$ 이다. 여기서 다음 식을 계산할 수 있다.

$$
\begin{aligned}
& E(\hat{\Theta}_i - \boldsymbol{x}_i'\tilde{\boldsymbol{b}})^2 \\
&= E[(\hat{\Theta}_i - \boldsymbol{x}_i'\tilde{\boldsymbol{b}}) - \boldsymbol{x}_i'(\tilde{\boldsymbol{b}} - \boldsymbol{b})]^2 \\
&= V(\hat{\Theta}_i) + \boldsymbol{x}_i' V(\tilde{\boldsymbol{b}})\boldsymbol{x}_i - 2\,Cov(\boldsymbol{x}_i'\tilde{\boldsymbol{b}}, \hat{\Theta}_i) \\
&= \tau^2 + V_i + \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_i - 2\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_i
\end{aligned}
$$

즉, $E[(\hat{\Theta}_i - \boldsymbol{x}_i'\tilde{\boldsymbol{b}})^2/(\tau^2 + V_i)] = 1 - \dfrac{\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_i}{\tau^2 + V_i}$ 이다. 따라서

$$
\begin{aligned}
& E[\sum_{i=1}^{m}(\hat{\Theta}_i - \boldsymbol{x}_i'\tilde{\boldsymbol{b}})^2/(\tau^2 + V_i)] \\
&= m - tr[(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\sum_{i=1}^{m}\boldsymbol{x}_i(\tau^2 + V_i)^{-1}\boldsymbol{x}_i'] \\
&= m - tr[(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})] = m - p
\end{aligned}
$$

이다. 여기서 적률법(method of moments)을 사용하여 식(ii)를 구한다.

이러한 반복방법(iterative method)을 피하기 위해 Prasad와 Rao(1990)는 τ^2 을 추정하기 위해 비가중최소제곱방법(unweighted least squares approach)을 제안하였다. $\boldsymbol{b}$ 의 최소제곱추정량(LSE)은

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\Theta}$$

이다. 여기서 다음 식을 계산할 수 있다.

$$
\begin{aligned}
& E||\hat{\Theta} - \boldsymbol{X}\hat{\boldsymbol{b}}||^2 \\
&= E||\hat{\Theta} - \boldsymbol{X}\boldsymbol{b} - \boldsymbol{X}(\hat{\boldsymbol{b}} - \boldsymbol{b})||^2
\end{aligned}
$$

$$= tr[V(\hat{\Theta})] + tr[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}]$$
$$- 2\sum_{i=1}^{m}\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i(\tau^2 + V_i)$$
$$= \sum_{i=1}^{m}(\tau^2 + V_i) + \sum_{i=1}^{m}\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i(\tau^2 + V_i)$$
$$- 2\sum_{i=1}^{m}\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i(\tau^2 + V_i)$$
$$= (m-p)\tau^2 + \sum_{i=1}^{m}V_i\{1-\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\}$$

따라서 τ^2의 적률법 추정량을 다음과 같이 구할 수 있다.

$$\hat{\tau^2} = \max\left(0, \frac{||\hat{\Theta} - \boldsymbol{X}\hat{\boldsymbol{b}}||^2 - \sum_{i=1}^{m}V_i(1 - r_i)}{m - p}\right)$$

여기서 $r_i = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$, $i = 1, \ldots, m$이다. $\hat{B}_i = \frac{V_i}{\hat{\tau^2} + V_i}$라 두면, $\Theta = (\Theta_1, \cdots, \Theta_m)'$의 경험적 베이즈(EB) 추정량은 다음과 같이 주어진다.

$$\hat{\Theta}^{EB} = (\hat{\Theta}_1^{EB}, \cdots, \hat{\Theta}_m^{EB})'$$

여기서 $\hat{\Theta}_i^{EB} = (1 - \hat{B}_i)\hat{\Theta}_i + \hat{B}_i\boldsymbol{x}_i'\tilde{\boldsymbol{b}}(\hat{\tau^2})$ 이며,

$\tilde{\boldsymbol{b}}(\tau^2) = (\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}^{-1}\hat{\Theta}$이다.

경험적 베이즈 추정량을 EBLUP으로 해석할 수 있다. 먼저 τ^2을 알 때 Θ의 BLUP은 $\hat{\Theta}^{BLUP} = (\hat{\Theta}_1^{BLUP}, \cdots, \hat{\Theta}_m^{BLUP})'$으로 주어진다. 여기서

$$\hat{\Theta}_i^{BLUP} = (1 - B_i)\hat{\Theta}_i + B_i\boldsymbol{x}_i'\tilde{\boldsymbol{b}}(\tau^2)$$

이다. τ^2에 $\widehat{\tau^2}$을 대입하면 θ의 EBLUP가 된다. 여기서 BLUP 결과는 정규분포 가정이 필요없다.

2.2 평균제곱오차(MSE) 근사

Prasad와 Rao(1990)는 제곱오차손실(squared error loss)과 주관적 사전분포 $\theta_i \overset{iid}{\sim} N(\boldsymbol{x}_i'\boldsymbol{b},\ V_i)$ 하에서 θ의 추정량으로서 $\widehat{\theta}^{EB}$의 베이즈 위험(Bayes risk) (Prasad와 Rao(1990)는 이를 평균제곱오차(mean squared error; MSE)라 부름)에 대한 수학적 근사를 계산하였다.

베이즈 위험을 구하기 위해 다음과 같이 표현하면 편리하다.

$$\widehat{\theta}_i^{B} = (1 - B_i)\widehat{\theta}_i + B_i \boldsymbol{x}_i'\boldsymbol{b};$$

$$\widehat{\theta}_i^{EB}(\tau^2) = (1 - B_i)\widehat{\theta}_i + B_i \boldsymbol{x}_i'\tilde{\boldsymbol{b}}(\tau^2)$$

$$\widehat{\theta}_i^{EB}(\widehat{\tau^2}) \equiv \widehat{\theta}_i^{EB} = (1 - \widehat{B}_i)\widehat{\theta}_i + \widehat{B}_i \boldsymbol{x}_i'\tilde{\boldsymbol{b}}(\widehat{\tau^2})$$

직교성(orthogonality)을 이용하여 다음과 같이 계산할 수 있다.

$$\begin{aligned} & E\| \widehat{\theta}^{EB}(\hat{\tau}^2) - \theta \|^2 \\ &= \sum_{i=1}^{m} E[\theta_i - \widehat{\theta}_i^{B} + \widehat{\theta}_i^{B} - \widehat{\theta}_i^{EB}(\tau^2) + \widehat{\theta}_i^{EB}(\tau^2) - \widehat{\theta}_i^{EB}(\widehat{\tau^2})]^2 \\ &= \sum_{i=1}^{m} E(\theta_i - \widehat{\theta}_i^{B})^2 + \sum_{i=1}^{m} E(\widehat{\theta}_i^{B} - \widehat{\theta}_i^{EB}(\tau^2))^2 \\ &\quad + \sum_{i=1}^{m} E(\widehat{\theta}_i^{EB}(\tau^2) - \widehat{\theta}_i^{EB}(\widehat{\tau^2}))^2 \end{aligned}$$

여기서 E는 $\widehat{\theta}$과 θ의 결합분포에 관한 기대값을 나타낸다.

위 공식에 대해 다음과 같은 해석이 가능하다. 첫째 항은 주관적 사전분포로 인한 평균제곱오차를 나타내고, 둘째 항은 미지의 $\boldsymbol{b}$로 인한 평균제곱오차를 나타내며, 셋째 항은 미지의 τ^2으로 인한 평균제곱오차를 나타낸다.

위 공식에서 둘째 항과 셋째 항의 직교성은 Kackar과 Harville(1984)의 일반적 결과로부터 얻어진 것이다.

여기서 다음 식을 계산할 수 있다.

$$\begin{aligned} E[\theta_i - \hat{\theta}_i^B]^2 &= E[E(\theta_i - \hat{\theta}_i^B)^2 | \hat{\theta}_i] \\ &= E[V_i(1 - B_i) | \hat{\theta}_i] \\ &= V_i(1 - B_i); \end{aligned}$$

$$\begin{aligned} E(\hat{\theta}_i^B - \hat{\theta}_i^{EB}(\tau^2))^2 &= B_i^2 \boldsymbol{x}_i' V(\tilde{\boldsymbol{b}}(\tau^2)) \boldsymbol{x}_i \\ &= B_i^2 \boldsymbol{x}_i' (\boldsymbol{X}' \boldsymbol{D}^{-1} \boldsymbol{X})^{-1} \boldsymbol{x}_i \end{aligned}$$

따라서 다음의 결과를 얻는다.

$$\begin{aligned} & E[\theta_i - \hat{\theta}_i^{EB}(\hat{\tau}^2)]^2 \\ &= V_i(1 - B_i) + B_i^2 \boldsymbol{x}_i' (\boldsymbol{X}' \boldsymbol{D}^{-1}(\tau^2) \boldsymbol{X})^{-1} \boldsymbol{x}_i \\ &\quad + E(\hat{\theta}_i^{EB}(\hat{\tau}^2) - \hat{\theta}_i^{EB}(\tau^2))^2 \end{aligned}$$

여기서 Prasad와 Rao(1990)는 먼저 일차 테일러 급수 근사(1st order Taylor series approximation)를 이용하여 다음의 근사식을 구하였다.

$$\begin{aligned} & E(\hat{\theta}_i^{EB}(\hat{\tau}^2) - \hat{\theta}_i^{EB}(\tau^2))^2 \\ &\doteq E[(\hat{\tau}^2 - \tau^2) \frac{\partial}{\partial \tau^2} \hat{\theta}_i^{EB}(\tau^2)]^2 \\ &\doteq B_i^2 (\tau^2 + V_i)^{-1} V(\hat{\tau}^2) \end{aligned}$$

결국 Prasad와 Rao가 구한 MSE의 수학적 근사는 다음과 같다.

$$MSE(\hat{\Theta}_i^{EB}) \approx V_i(1-\hat{B}_i) + \hat{B}_i^2 \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{D}^{-1}(\hat{\tau}^2)\boldsymbol{X})^{-1}\boldsymbol{x}_i + 2\hat{B}_i^2(V_i+\hat{\tau}^2)^{-1}(\hat{\tau}^2)^2 m^{-2}\sum_{i=1}^{m}(1-\hat{B}_i)^{-2}$$

최근 Jing, Lahiri와 Wan(2002)는 경험적 베이즈 추정량의 MSE 추정에 대한 잭나이프 방법을 제안하였다. 이 방법은 이산형 반응(binary 혹은 count 자료)을 가지는 소지역모형에도 쉽게 적용이 가능한 일반적인 방법이다.

먼저 $MSE(\hat{\Theta}_i^{EB})$를 다음과 같이 분해하자.

$$\begin{aligned} MSE(\hat{\Theta}_i^{EB}) &= E(\hat{\Theta}_i^{EB}-\hat{\Theta}_i^{B})^2 + E(\hat{\Theta}_i^{B}-\Theta_i)^2 \\ &= E(\hat{\Theta}_i^{EB}-\hat{\Theta}_i^{B})^2 + g_{1i}(\tau^2) \\ &= M_{2i}+M_{1i} \end{aligned}$$

여기서 기대는 $(\hat{\Theta}_i, \Theta_i)$ $(i=1,\cdots,m)$의 결합분포에 대한 것이다. $\hat{\Theta}_i^{EB}$는 추정량 $\hat{\boldsymbol{b}}$와 $\hat{\tau}^2$을 통해 모든 자료 $\hat{\boldsymbol{\Theta}} = (\hat{\Theta}_1, \cdots, \hat{\Theta}_m)'$에 의존한다.

경험적 베이즈 추정량을 직접추정량 $\hat{\Theta}_i$와 모수 추정량 $\hat{\boldsymbol{b}}$, $\hat{\tau}^2$의 함수로 표현하여 $\hat{\Theta}_i^{EB} = k_i(\hat{\Theta}_i, \hat{\boldsymbol{b}}, \hat{\tau}^2)$라 할 때 $MSE(\hat{\Theta}_i^{EB})$를 추정하는 잭나이프 절차는 다음과 같다.

단계 1) l번째 지역의 자료 $(\hat{\Theta}_l, \boldsymbol{x}_l)$을 제외한 자료로부터 $(\boldsymbol{b}, \tau^2)$의 l-제거 추정량 $(\hat{\boldsymbol{b}}_{-l}, \hat{\tau}^2_{-l})$, $l=1,\cdots,m$을 계산한다. 이로부터 다음에 주어지는 m개의 경험적 베이즈 추정량을 계산한다.

$$\hat{\Theta}_{i,-l}^{EB} = k_i(\hat{\Theta}_i, \hat{\boldsymbol{b}}_{-l}, \hat{\tau}_{-l}^2);\ i=1,\cdots,m$$

단계 2) M_{2i}의 추정량을 다음과 같이 계산한다.

$$\hat{M}_{2i} = \frac{m-1}{m}\sum_{l=1}^{m}(\hat{\Theta}_{i,-l}^{EB} - \hat{\Theta}_i^{EB})^2$$

단계 3) M_{1i}의 추정량을 다음과 같이 계산한다.

$$\hat{M}_{1i} = g_{1i}(\hat{\tau}^2) - \frac{m-1}{m}\sum_{l=1}^{m}[g_{1i}(\hat{\tau}_{-l}^2) - g_{1i}(\hat{\tau}^2)]$$

여기서 추정량 $\hat{M}_{1i}$은 $g_{1i}(\hat{\tau}^2)$의 편향을 수정한다.

단계 4) $MSE(\hat{\Theta}_i^{EB})$의 잭나이프 추정량을 다음과 같이 계산한다.

$$MSE(\hat{\Theta}_i^{EB}) = \hat{M}_{1i} + \hat{M}_{2i}$$

$\hat{M}_{1i}$은 모형의 모수를 안다고 가정할 때 MSE에 대한 추정량이며, $\hat{M}_{2i}$는 모형 모수의 추정에 기인하는 추가적인 변동을 추정한다. MSE의 잭나이프 추정량은 근사적으로 불편이다.

2.3 실업자 총계의 경험적 베이즈 추정

이제 2003년 9월의 경제활동인구조사 자료를 사용하여 소지역 i에 대한 실업자 총계 추정문제를 생각하자. 소지역 i의 실업자 총계에 대한 직접추정값은 소지역에 배정된 표본만을 이용하여 추정되며, 경제활동인구조사 체계에서의 실

업자 총계에 대한 직접추정공식은 다음과 같이 주어진다.

$$\hat{Y}_{i\cdot} = \sum_{s=1}^{2}\sum_{k}^{11} {}_{sk}\hat{Y}_{i\cdot} \quad (i=1,\cdots,m)$$
$$= \sum_{s=1}^{2}\sum_{k}^{11}\sum_{h=1}^{n_i} {}_{sk}\hat{Y}_{ih} = \sum_{s=1}^{2}\sum_{k}^{11}\sum_{h=1}^{n_i} {}_{sk}M_{i}\ {}_{sk}Y_{ih}$$

여기서 s는 성별(남,여), k는 연령층(15-19, 20-24, ..., 60-64, 65+)을 나타내며 n_i는 경제활동인구조사에서 소지역 i에 할당된 표본조사구 수, ${}_{sk}y_{ih}$는 각 성별 및 연령층에 대하여 소지역 i의 h번째 표본조사구에서 조사된 실업자 수를 나타낸다. 승수 ${}_{sk}M_{i} = {}_{sk}\hat{X}_{i} / {}_{sk}x_{i\cdot}$는 직접추정량 $\hat{Y}_{i\cdot}$이 불편추정량이 되도록 산정된다. 승수 표현식에서 ${}_{sk}\hat{X}_{i}$는 소지역 i에 대한 15세 이상의 상주추계인구를 나타내며, ${}_{sk}x_{i\cdot}$는 경제활동인구조사에서 집계된 15세 이상의 상주조사인구를 나타낸다. 소지역 i에 대한 직접추정량 $\hat{Y}_{i\cdot}$의 분산은 선형화에 기초한 분산추정량을 사용하여 추정된다.

합성추정량은 다음과 같이 구한다. 먼저 대영역 내에 m개의 소지역이 있다고 가정하자. 대영역을 특성이 유사한 시단위, 군단위 및 구단위들의 3개의 그룹으로 분할하고, 각 그룹들을 4개의 성별(남,여)-연령대별(15-29세, 30세이상) 범주로 구분한다. 여기서 $m = \sum_{k=1}^{4} m_k$이며, 시, 군 및 구 그룹들은 각각 m_1, m_2, m_3, m_4개의 동질적인 소지역 단위들로 구성된다.

합성추정량을 정의하기 위해 다음과 같은 기호들을 사용하자.

N_i= 표본추출틀에서 소지역 i의 조사구 수,

n_i= 경제활동인구조사에서 소지역 i에 할당된 표본조사구 수,

${}_{j}P^{C}_{i,2000}$= 2000년 센서스로부터 추계된 j범주에 대한 소지역 i의 상주인구,

${}_jP^R_{i,2000}$= j범주에 대한 소지역 i의 2000년 주민등록인구,

${}_jP^R_{i,month}$= j범주에 대한 소지역 i의 경제활동인구조사 달의 주민등록인구,

${}_j\widehat{X}_i$= j범주에 대한 소지역 i의 상주추정인구,

${}_jY_{ih}$= j범주에 대한 소지역 i의 h번째 표본조사구의 실업자 수.

m_k개의 소지역들을 포함하고 있는 부차관심영역인 각 시, 군, 구 그룹 내에서 소지역 i의 실업자 총계에 대한 합성추정량 $\widehat{Y}^S_{i.}$는 다음과 같이 주어진다.

$$\widehat{Y}^S_{i.} = \sum_{j=1}^{4} \frac{{}_j\widehat{P}_i}{{}_j\widehat{X}_{.}} \, {}_j\widehat{Y}_{dir} \quad (i=1,\cdots,m_k)$$

여기서

$${}_j\widehat{P}_i = \frac{{}_jP^C_{i,2000}}{{}_jP^R_{i,2000}} \, {}_jP^R_{i,month},$$

$${}_j\widehat{X}_{.} = \sum_{i=1}^{m_i} {}_j\widehat{X}_i,$$

$${}_j\widehat{Y}_{dir} = \sum_{i=1}^{m_k}\sum_{h=1}^{n_i} {}_jM_i \; {}_jY_{ih}$$

이다. ${}_j\widehat{P}_i$는 행정보고 자료로부터 산정된 j범주에 대한 소지역 i의 상주추정인구를 나타내며, ${}_j\widehat{X}_{.}$는 경제활동인구조사 자료로부터 산정되는 상주추정인구를 나타낸다. 또한, ${}_j\widehat{Y}_{dir}$는 j번째 성별-연령대별 범주의 실업자 총계에 대한 직접추정량을 의미하며, 경제활동인구조사 자료로부터 산정된다. 소지역 i의 j범주에 대한 승수는 ${}_jM_i = {}_j\widehat{X}_i / {}_jX_i$. 로 주어진다.

2003년 9월 경제활동인구조사 자료로부터 구한 직접추정량 및 합성추정량을

구성하는 성별-연령대별 범주에 대한 실업자 총계가 Fay-Herriot 모형에서의 변수로 사용된다. 구체적으로 다음의 기호를 정의하자.

$\hat{\Theta}_i =$ 소지역 i에서 실업자 총수에 대한 직접추정량

$V_i =$ 소지역 i에 대한 직접추정량의 분산

보조자료로는 경제활동인구조사와 연관이 있는 다른 조사로부터 자료를 구하기가 여의치 않으므로 Chung, Lee와 Kim(2003)에서와 같이 합성추정량을 구성하는 4개의 성별(남,여)-연령대별(15-29세, 30세이상) 범주에 해당하는 자료를 공변량으로 사용한다.

$$x_{1i} = \frac{{}_1\hat{P}_i}{{}_1\hat{X}_{\cdot}} \; {}_1\hat{Y}_{dir}, \quad x_{2i} = \frac{{}_2\hat{P}_i}{{}_2\hat{X}_{\cdot}} \; {}_2\hat{Y}_{dir},$$

$$x_{3i} = \frac{{}_3\hat{P}_i}{{}_3\hat{X}_{\cdot}} \; {}_3\hat{Y}_{dir}, \quad x_{4i} = \frac{{}_4\hat{P}_i}{{}_4\hat{X}_{\cdot}} \; {}_4\hat{Y}_{dir}$$

위에서 정의한 반응변수와 보조변수를 사용한 소지역모형은 다음과 같다.

i. $\hat{\Theta}_i | \Theta_i \overset{ind}{\sim} N(\Theta_i, \; V_i)$

ii $\Theta_i \overset{iid}{\sim} N(\boldsymbol{x}_i' \boldsymbol{b}, \; \tau^2)$

여기서 $\boldsymbol{b} = (b_0, b_1, b_2, b_3, b_4)'$ 그리고 $\boldsymbol{x}_i = (1, x_{1i}, x_{2i}, x_{3i}, x_{4i})'$ 이다.

Θ_i에 대한 경험적 베이즈 추정량을 계산하기 위해 다음의 행렬 기호를 사용한다.

$$X = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_m' \end{pmatrix} \text{ 혹은 } \boldsymbol{X}' = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m)$$

$$\hat{\Theta} = (\hat{\Theta}_1, \cdots, \hat{\Theta}_m)'$$

$$\boldsymbol{D} = \text{Diag}(V_1 + \tau^2, \cdots, V_m + \tau^2)$$

Θ_i의 사후평균이 포함하고 있는 미지의 모수 $\boldsymbol{b}$와 τ^2를 $\hat{\Theta}_i$들의 주변분포로부터 추정하는 두 가지 방법은 다음과 같다.

방법 I (Prasad와 Rao, 1990): 먼저 다음의 추정량을 구한다.

$$\hat{\boldsymbol{b}} = \left[\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i' \right]^{-1} \left[\sum_{i=1}^{m} \boldsymbol{x}_i \hat{\Theta}_i \right]$$

$$\hat{\tau}^2 = \max \left(0, \frac{\sum_{i=1}^{m} (\hat{\Theta}_i - \boldsymbol{x}_i' \hat{\boldsymbol{b}})^2 - \sum_{i=1}^{m} V_i(1 - r_i)}{m - p} \right)$$

단, $r_i = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$, $i = 1, \ldots, m$이다. 따라서 Θ_i에 대한 경험적 베이즈 추정량은 다음과 같다.

$$\hat{\Theta}_i^{EB} = (1 - \hat{B}_i)\hat{\Theta}_i + \hat{B}_i \boldsymbol{x}_i' \hat{\boldsymbol{b}}$$

여기서 $\hat{B}_i = \dfrac{V_i}{\hat{\tau}^2 + V_i}$이다.

방법 II (Fay와 Herriot, 1979): 먼저 다음의 두 식을 사용하여 $\boldsymbol{b}$와 τ^2의 추정량을 반복적으로 구한다.

(i) $\tilde{\boldsymbol{b}}(\tau^2) = \left[\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i' / (V_i + \tau^2) \right]^{-1} \left[\sum_{i=1}^{m} \boldsymbol{x}_i \hat{\Theta}_i / (V_i + \tau^2) \right]$

(ii) $h(\tau^2) = m - p$

여기서 $h(\tau^2) = \sum_{i=1}^{m} (\hat{\Theta}_i - \boldsymbol{x}_i' \tilde{\boldsymbol{b}})^2 / (\tau^2 + V_i)$ 이다. (ii)식에서 τ^2를 수치적으로 계산하는 알고리즘은 기본적으로 Newton-Raphson 방법으로서 다음과 같다.

초기값은 $\tau_0^2 = 0$이며 수치적 계산식은

$$\tau_{k+1}^2 = \max(0,\ \tau_k^2 - \frac{h(\tau_k^2) - (m-p)}{h'(\tau_k^2)})$$

이다. 여기서 $h'(\tau^2) = -\sum_{i=1}^{m} (\hat{\Theta}_i - \boldsymbol{x}_i' \tilde{\boldsymbol{b}})^2 / (\tau^2 + V_i)^2$ 이며 이는 $h(\tau^2)$의 미분에 대한 근사이다. 이 수치적 근사식은 수렴속도가 빨라서 일반적으로 10회 이하의 반복으로 충분히 수렴한다.

따라서 Θ_i에 대한 경험적 베이즈 추정량은 다음과 같다.

$$\hat{\Theta}_i^{EB} = (1 - \hat{B}_i)\hat{\Theta}_i + \hat{B}_i \boldsymbol{x}_i' \tilde{\boldsymbol{b}}(\hat{\tau}^2)$$

여기서 $\tilde{\boldsymbol{b}}(\hat{\tau}^2) = (\boldsymbol{X}' \hat{\boldsymbol{D}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}' \hat{\boldsymbol{D}}^{-1} \hat{\Theta}$ 이며

$\hat{\boldsymbol{D}} = \mathrm{Diag}(\mathrm{V}_1 + \hat{\tau}^2, \cdots, \mathrm{V}_m + \hat{\tau}^2)$이다.

위의 방법 I과 II에 대해 Prasad와 Rao(1990)가 제안한 MSE 근사를 다음과 같이 구한다.

$$MSE(\hat{\Theta}_i^{EB}) \approx V_i(1 - \hat{B}_i) + \hat{B}_i^2 \boldsymbol{x}_i' (\boldsymbol{X}^T \hat{\boldsymbol{D}}^{-1} X)^{-1} \boldsymbol{x}_i$$
$$+ 2\, \hat{B}_i^2 (V_i + \hat{\tau}^2)^{-1} (\hat{\tau}^2)^2 m^{-2} \sum_{i=1}^{m} (1 - \hat{B}_i)^{-2}$$

여기서 $\widehat{B}_i = \frac{V_i}{\widehat{\tau}^2 + V_i}$ 이다.

2003년 9월 경제활동인구조사 자료를 사용하여 위의 방법 I과 II에 근거하여 경험적 베이즈 추정값을 계산하였다. 여기서는 방법 II에 의한 τ^2의 추정값이 보다 안정적이므로 이에 관한 경험적 베이즈 추정값을 계층적 베이즈 추정결과와 함께 보고한다. 방법 II의 경우 Prasad와 Rao(1990)가 제안한 MSE 근사를 계산하였으나, MSE의 잭나이프 추정은 시도되지 않았다.

3. 계층적 베이즈 추정

3.1 경험적 베이즈와 계층적 베이즈의 차이

경험적 베이즈(EB) 분석의 특징은 모수에 대한 어떤 사전분포를 생각할 때, 이 사전분포가 가지는 모수인 초모수(hyperparameter)를 자료로부터 이용하여 추정하는 것이다. 즉, 관측값의 주변분포(marginal distribution)로부터 초모수를 추정한다.

이와는 달리 계층적 베이즈(HB) 분석은 사전분포를 단계적으로 모형화하는데, 첫 번째 단계에서는 먼저 초모수가 주어졌을 때 모수에 대한 사전분포를 생각하고, 두 번째 단계에서는 초모수에 대한 초사전분포(hyperprior)를 생각한다. 흔히 초사전분포로는 모수공간 상에 균일한 분포를 사용한다.

경험적 베이즈와 계층적 베이즈 절차 모두 사전정보에서의 불확실성을 인식하여 사전분포를 생각한다. 그러나 두 방법의 차이점으로 계층적 베이즈 절차는 사전분포가 가지는 초모수에 분포를 고려함으로 사전정보에서의 불확실성을 모형화하는 반면 경험적 베이즈 절차는 미지의 초모수를 기존의 추정방법인 최우추정법, 적률법 혹은 UMVUE 등을 사용하여 추정함으로 결국 추정된 사전분포

를 사용한다.

점추정 면에서 두 방법은 흔히 비교적 비슷한 결과를 제공한다. 그러나 추정량과 관련된 표준오차 계산 문제에서는 계층적 베이즈 방법이 단순(naive) 경험적 베이즈 절차보다 확실한 장점이 있다. 그 이유는 단순 경험적 베이즈 절차는 미지의 초모수를 추정함으로 생기는 불확실성을 무시하기 때문이다. 따라서 표준오차를 작게 추정(underestimate)하게 된다. 이와는 달리 계층적 베이즈 방법은 사후평균과 관련된 오차로서 사후표준편차를 사용한다.

3.2 소지역 추정을 위한 계층적 베이지안 모형

소지역 추정을 위해 다음과 같은 계층적 베이지안 모형을 생각하자.

i. $\hat{\Theta}_i | \Theta, \boldsymbol{b}, \tau^2 \overset{ind}{\sim} N(\Theta_i, V_i)$, $i = 1, \cdots, m$;

ii. $\Theta_i | \boldsymbol{b}, \tau^2 \overset{iid}{\sim} N(\boldsymbol{x}_i' \boldsymbol{b}, \tau^2)$, $i = 1, \cdots, m$;

iii. $\boldsymbol{b}$와 τ^2은 독립이며, $\boldsymbol{b} \sim \text{Uniform}(R^p)$이고, τ^2의 분포는 $g(\tau^2) \propto 1$이다.

여기서 목표는 $\hat{\Theta}$이 주어졌을 때 Θ의 사후분포를 구하는 것이다. 이를 위해 먼저 $\hat{\Theta}$이 주어졌을 때 Θ, $\boldsymbol{b}$와 τ^2의 사후결합확률분포를 구하면 다음과 같다.

$$\begin{aligned} &\pi(\Theta, \boldsymbol{b}, \tau^2 | \hat{\Theta}) \\ &\propto \exp\left[-\frac{1}{2}\sum_1^m (\hat{\Theta}_i - \Theta_i)^2 / V_i\right] \\ &\quad \times (\tau^2)^{-\frac{m}{2}} \exp\left[-\frac{1}{2\tau^2}\sum_1^m (\Theta_i - \boldsymbol{x}_i' \boldsymbol{b})^2\right] \end{aligned}$$

여기서 $\boldsymbol{X}' = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m)$라 두고, $\text{rank}(\mathbf{X}) = p$를 가정하자. 사후결합확률

분포를 $\boldsymbol{b}$에 관해 적분하면 다음과 같은 결과를 얻는다.

$$\pi(\Theta, \tau^2 | \widehat{\Theta})$$

$$\propto (\tau^2)^{-\frac{1}{2}(m-p)} \exp\left[-\frac{1}{2}(\widehat{\Theta} - \Theta)' \boldsymbol{V}^{-1}(\widehat{\Theta} - \Theta) - \frac{1}{2\tau^2}\Theta'(\boldsymbol{I}_m - \boldsymbol{P}_{\boldsymbol{X}})\Theta \right]$$

여기서 $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ 이며, $\boldsymbol{V} = \mathrm{Diag}(\mathrm{V}_1, \cdots, \mathrm{V}_m)$ 이다.

이제 $\boldsymbol{E}^{-1} = \boldsymbol{V}^{-1} + \tau^{-2}(\boldsymbol{I}_m - \boldsymbol{P}_{\boldsymbol{X}})$라 두면 다음의 결과를 얻는다.

$$\begin{aligned}
&(\widehat{\Theta} - \Theta)' \boldsymbol{V}^{-1}(\widehat{\Theta} - \Theta) + \tau^{-2}\Theta'(\boldsymbol{I}_m - \boldsymbol{P}_{\boldsymbol{X}})\Theta \\
&= \Theta' \boldsymbol{E}^{-1}\Theta - 2\Theta' \boldsymbol{V}^{-1}\widehat{\Theta} + \widehat{\Theta}' \boldsymbol{V}^{-1}\widehat{\Theta} \\
&= (\Theta - \boldsymbol{E}\boldsymbol{V}^{-1}\widehat{\Theta})' \boldsymbol{E}^{-1}(\Theta - \boldsymbol{E}\boldsymbol{V}^{-1}\widehat{\Theta}) \\
&\quad + \widehat{\Theta}'(\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{E}\boldsymbol{V}^{-1})\widehat{\Theta}
\end{aligned}$$

따라서 $\Theta | \tau^2, \widehat{\Theta} \sim N_m(\boldsymbol{E}\boldsymbol{V}^{-1}\widehat{\Theta}, \boldsymbol{E})$이다. 여기서 평균백터와 분산공분산 행렬을 보다 구체적으로 표현하기 위해 다음 계산을 수행한다.

$$\boldsymbol{E} = \tau^2[(\boldsymbol{I}_{\boldsymbol{m}} - \boldsymbol{D}_0) + (\boldsymbol{I}_m - \boldsymbol{D}_0)\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{I}_m - \boldsymbol{D}_0)] ;$$

$$\boldsymbol{E}\boldsymbol{V}^{-1}\widehat{\Theta} = [(1 - B_1)\widehat{\Theta_1} + B_1 \boldsymbol{x}_1' \widetilde{\boldsymbol{b}}, \cdots, (1 - B_m)\widehat{\Theta_m} + B_m \boldsymbol{x}_m' \widetilde{\boldsymbol{b}}]^T$$

여기서

$$\begin{aligned}
B_i &= V_i/(\tau^2 + V_i), \\
\boldsymbol{D}_0 &= \mathrm{Diag}(1 - \mathrm{B}_1, \cdots, 1 - \mathrm{B}_m), \\
\widetilde{\boldsymbol{b}} &= (\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}^{-1}\widehat{\Theta}
\end{aligned}$$

이다.

한편, $\boldsymbol{b}$와 τ^2의 사후결합확률분포에서 Θ에 관해 적분하여 τ^2에 관한 주변 사후분포를 다음과 같이 구할 수 있다.

$$\pi(\tau^2|\hat{\Theta}) \propto (\tau^2)^{\frac{1}{2}p}\{\prod_1^m(\tau^2+V_i)\}^{-\frac{1}{2}}|\sum_1^m(1-B_i)\boldsymbol{x}_i\boldsymbol{x}_i'|^{-\frac{1}{2}}$$

$$\times \exp[-\frac{1}{2\tau^2}\{\sum_1^m(1-B_i)\hat{\Theta}_i^2 - (\sum_1^m(1-B_i)\boldsymbol{x}_i\hat{\Theta}_i)'(\sum_1^m(1-B_i)\boldsymbol{x}_i\boldsymbol{x}_i')^{-1} \times (\sum_1^m(1-B_i)\boldsymbol{x}_i\hat{\Theta}_i)\}]$$

따라서 다음과 같이 사후평균과 사후분산을 구할 수 있다.

$$E(\Theta_i|\hat{\Theta}) = [1-E(B_i|\hat{\Theta})]\hat{\Theta}_i + E[B_i\boldsymbol{x}_i'\tilde{\boldsymbol{b}}|\hat{\Theta}];$$
$$V(\Theta_i|\hat{\Theta}) = E[V(\Theta_i|\tau^2,\hat{\Theta})|\hat{\Theta}] + V[E(\Theta_i|\tau^2,\hat{\Theta})|\hat{\Theta}]$$

여기서 $V(\Theta_i|\hat{\Theta})$를 좀더 간단히 정리하면

$$E[V_i(1-B_i)+B_i^2\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_i|\hat{\Theta}] + V[B_i(\hat{\Theta}_i-\boldsymbol{x}_i'\tilde{\boldsymbol{b}})|\hat{\Theta}]$$

이다. 단, $\boldsymbol{D} = \mathrm{Diag}(\tau^2+\mathrm{V}_1,\cdots,\tau^2+\mathrm{V}_m)$이다.

Morris(1983)는 $E(\Theta_i|\hat{\Theta})$를 다음과 같이 근사시켰는데

$$(1-\hat{B}_i)\hat{\Theta}_i + \hat{B}_i\boldsymbol{x}_i'\tilde{\boldsymbol{b}}(\hat{\tau^2})$$

이는 Prasad-Rao 추정량과 동일하다. 뿐만 아니라, Morris(1983)는 $V(\Theta_i|\hat{\Theta})$도 다음과 같이 근사시켰다.

$$V_i(1 - \widehat{B}_i) + \widehat{B}_i^2 \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{D}^{-1}(\widehat{\tau^2})\boldsymbol{X})^{-1}\boldsymbol{x}_i$$
$$+ \frac{2}{m-p-2}\widehat{B}_i^2(\widehat{\tau^2} + \widetilde{V})/(\widehat{\tau^2} + V_i)$$

여기서 $\widetilde{V} = \sum_1^m V_i(\widehat{\tau^2} + V_i)^{-1} \sum_1^m (\widehat{\tau^2} + V_i)^{-1}$ 이다. Morris의 사후분산 근사식에서 첫째항과 둘째항은 Prasad와 Rao(1990)의 결과와 일치하지만 셋째항은 다르다.

그러나 계층적 베이지안 분석은 깁스 표본자(Gibbs sampler)를 사용하여 훨씬 쉽게 수행할 수 있다. 깁스 표본자는 모형으로부터 모든 확률변수의 결합확률분포가 주어졌을 때 저차원의 모든 조건부 확률분포로부터 표본을 반복하여 충분히 생성하면 결국 생성된 표본이 결합확률분포와 주변확률분포에 대한 표본이 된다는 이론에 근거한다. 즉, 모든 조건부 확률분포로부터 생성된 표본을 사용하여 결합확률분포와 주변확률분포에 대한 통계적 추론을 수행하는 방법이다. 따라서 깁스 표본자를 사용하기 위해서는 먼저 $\widehat{\Theta}$이 주어졌을 때 Θ, $\boldsymbol{b}$와 τ^2의 사후결합확률분포로부터 모든 모수에 대한 조건부 확률분포를 구해야만 한다. 이들 조건부 확률분포를 구하면 다음과 같다.

$$\boldsymbol{b}|\Theta, \tau^2, \widehat{\Theta} \sim N_p[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\Theta,\ \tau^2(\boldsymbol{X}'\boldsymbol{X})^{-1}];$$

$$\tau^2|\Theta, \boldsymbol{b}, \widehat{\Theta} \sim IG\left(\frac{1}{2}\sum_1^m(\Theta_i - \boldsymbol{x}_i'\boldsymbol{b})^2,\ \frac{m-2}{2}\right);$$

$$\Theta_i|\boldsymbol{b}, \tau^2, \widehat{\Theta} \overset{ind}{\sim} N[(V_i^{-1} + \tau^{-2})^{-1}(V_i^{-1}\widehat{\Theta}_i + \tau^{-2}\boldsymbol{x}_i'\boldsymbol{b}),\ (V_i^{-1} + \tau^{-2})^{-1}]$$

여기서 $Z \sim IG(a, b)$는 확률밀도함수가 $f(z) \propto \exp(-a/z)z^{-(b+1)}$ 형태를 가진다는 것을 의미한다.

주어진 조건부 확률분포에 근거하여 깁스 표본자를 사용한 베이지안 계산을 수행할 때, Gelman과 Rubin(1992)에 따라 길이가 $2d$인 L개의 병렬체인을 고

려한다. 왜냐하면 병렬체인을 사용하면 깁스 표본자의 수렴 여부를 쉽게 검증할 수 있기 때문이다. 각 체인에서 Θ, $\boldsymbol{b}$와 τ^2에 대해 $2d$개의 난수를 반복하여 생성한 후, 초기값의 영향을 제거하기 위해 반복의 앞 부분 d개를 버리고 나머지 d개를 사용하여 사후평균과 사후분산을 계산한다. 이 때 Gelfand와 Smith(1991)에서 제안된 바와 같이, 사후평균과 사후분산으로 Rao-Blackwellized 추정량을 계산한다. 따라서 사후평균 $E(\Theta_i|\widehat{\Theta})$는 다음과 같이 추정된다.

$$(Ld)^{-1}\sum_{l=1}^{L}\sum_{k=d+1}^{2d}(V_i^{-1}+\tau^{-2(lk)})^{-1}(V_i^{-1}\widehat{\Theta}_i+\tau^{-2(lk)}\boldsymbol{x}_i'\boldsymbol{b}^{(lk)})$$

한편 사후분산 $V(\Theta_i|\widehat{\Theta})$도 다음과 같이 추정된다.

$$\begin{aligned}&(Ld)^{-1}\sum_{l=1}^{L}\sum_{k=d+1}^{2d}(V_i^{-1}+\tau^{-2(lk)})^{-1}\\&+(Ld)^{-1}\sum_{l=1}^{L}\sum_{k=d+1}^{2d}(V_i^{-1}+\tau^{-2(lk)})^{-2}(V_i^{-1}\widehat{\Theta}_i+\tau^{-2(lk)}\boldsymbol{x}_i'\boldsymbol{b}^{(lk)})^2\\&-[(Ld)^{-1}\sum_{l=1}^{L}\sum_{k=d+1}^{2d}(V_i^{-1}+\tau^{-2(lk)})^{-1}(V_i^{-1}\widehat{\Theta}_i+\tau^{-2(lk)}\boldsymbol{x}_i'\boldsymbol{b}^{(lk)})]^2\end{aligned}$$

위에서 $\tau^{2(lk)}$와 $\boldsymbol{b}^{(lk)}$는 l번째 체인의 k번째 반복에서 생성된 τ^2과 $\boldsymbol{b}$의 값을 나타낸다.

3.3 깁스 표본자의 수렴과 모형적합

3.3.1 깁스 표본자의 수렴

깁스 샘플링을 사용한 사후평균과 사후분산에 대한 계산이 정당성을 가지기 위해서는 관심모수 $\Theta_i\ (i=1,\ldots,m)$에 대한 깁스 표본자의 수렴성을 조사해야

한다. 이를 위해 Gelman과 Rubin(1992)의 방법대로 다음과 같은 절차를 따른다.

단계 1: 전체평균

$$\overline{\Theta}_i = \sum_{l=1}^{L} \sum_{k=d+1}^{2d} \Theta_i^{(lk)}/(Ld)$$

그리고 수열내 평균

$$\overline{\Theta}_i^{(l)} = \sum_{k=d+1}^{2d} \Theta_i^{(lk)}/d, l=1,\ldots,L$$

를 계산한다. 그리고 L 수열 평균간의 분산인 B_i/d를 구한다. 여기서

$B_i/d = \sum_{l=1}^{L} (\overline{\Theta}_i - \overline{\Theta}_i^{(l)})^2/(L-1)$ 이다.

단계 2: 각 자유도가 $(d-1)$인 L개의 수열내 분산 s_{il}^2의 평균인 W_i를 계산한다. 즉, $W_i = \sum_{l=1}^{L} s_{il}^2/L$ 이다.

단계 3: $s_i^2 = (d-1)\, W_i/d + B_i/d$ 와 $V_i = s_i^2 + B_i/(Ld)$ 를 계산한다.

단계 4: $\widehat{R}_i = V_i/W_i$ $(i=1,\ldots m)$를 계산한다. 여기서 $\widehat{R}_i$는 잠재적 척도축소인자(potential scale reduction factor)이다. 만약 모든 관심모수 Θ_i에 대한 $\widehat{R}_i$가 1에 가까우면 이는 깁스 샘플링이 수렴한다는 것을 암시한다.

3.3.2 모형적합

사용된 Fay-Herriot 모형의 적합을 검증하기 위해 사후예측 p값을 계산한다. 이 방법에서는 적절한 불일치측도(discrepancy measure)의 모의실험 값이

사후예측분포로부터 생성되며 이를 관측된 자료에 대한 불일치측도 값과 비교한다. (Sinha와 Dey (1997) 참조). Fay-Herriot 모형에 대해 자료 $\boldsymbol{y}$와 모수 Θ에 의존하는 불일치측도로 다음의 통계량을 고려한다.

$$d(\boldsymbol{y}, \Theta) = \sum_{i=1}^{m} (y_i - \Theta_i)^2 / V_i$$

주어진 모형이 현재의 자료에 잘 적합한지를 조사하기 위해 주어진 모형에 근거한 새로운 자료를 생성할 수 있다. $\mathbf{y}_{obs}$와 $\mathbf{y}_{new}$를 관측된 자료와 모의실험으로 생성된 자료라 하자. 만약 모형이 관측된 자료 $\mathbf{y}_{obs}$를 충분히 잘 적합하면 생성된 새로운 자료 $\mathbf{y}_{new}$도 관측된 자료와 비슷해야 한다.

따라서 불일치측도는 자료에 따라 $d(\mathbf{y}_{obs}, \Theta)$와 $d(\mathbf{y}_{new}, \Theta)$가 있다. 이 때 사후예측 p값은 다음과 같이 정의된다.

$$p = \mathrm{P}\left(d(\boldsymbol{y}_{new}, \Theta) > d(\boldsymbol{y}_{obs}, \Theta) \mid \boldsymbol{y}_{obs}\right)$$

여기서 확률은 관측된 자료가 주어졌을 때 Θ의 사후분포에 관한 것이다.

깁스 샘플링 결과를 사용하여 사후분포 $f(\Theta \mid \mathbf{y}_{obs})$로부터 $\Theta^{(l)}$을 생성하고, 이를 사용하여 $f(\mathbf{y} \mid \Theta^{(l)})$로부터 $\mathbf{y}^{(l)}$을 생성한 후, $d(\mathbf{y}_{obs}, \Theta^{(l)})$과 $d(\mathbf{y}^{(l)}, \Theta^{(l)})$ ($l = 1, \cdots, B$)을 계산한다. 여기서 B는 Θ값의 깁스 반복의 총 수이다.

생성된 표본을 사용하여 $p = P\{d(\mathbf{y}_{new}, \Theta) \geq d(\mathbf{y}_{obs}, \Theta) \mid \mathbf{y}_{obs}\}$에 대한 근사값을 다음과 같이 구할 수 있다.

$$B^{-1} \sum_{l=1}^{B} I\{d(\mathbf{y}^{(l)}, \Theta^{(l)}) \geq d(\mathbf{y}_{obs}, \Theta^{(l)})\}$$

여기서 $I(\cdot)$은 지시함수이다. 이 p값을 계산했을 때 극단적인(0이나 1에 가까운) 값으로 주어지면 모형의 적합결여를 나타낸다고 판단한다. 반면에 모형이 자료를 잘 적합하면 이 p값이 0.5에 가깝다.

3.4 실업자 총계의 계층적 베이즈 추정

3.4.1 계층적 베이즈 추정량 계산

이제 2003년 9월의 경제활동인구조사 자료를 사용하여 소지역 i에 대한 실업자 추정문제에 대한 계층적 베이즈 방법을 고려하자. 경험적 베이즈 추정에서와 동일한 직접추정량을 사용한다.

$\hat{\Theta}_i =$ 소지역 i에 대한 직접추정량

$V_i =$ 소지역 i에 대한 직접추정량의 분산

이 때 앞에서와 동일한 보조정보를 고려한다. 합성추정량을 구성하는 4개의 성별(남,여)-연령대별(15-29세, 30세이상) 범주에 해당하는 자료를 보조정보로 사용하여 4개의 공변량을 고려하자.

소지역추정을 위한 계층적 베이지안 모형은 다음과 같다.

i. $\hat{\Theta}_i | \Theta_i \overset{ind}{\sim} N(\Theta_i,\ V_i)$

ii. $\Theta_i \overset{ind}{\sim} N(\boldsymbol{x}_i'\boldsymbol{b},\ \tau^2)$

iii. $\boldsymbol{b}$와 τ^2은 독립이며, $\boldsymbol{b} \sim \text{Uniform}(R^2)$이고, τ^2의 분포는 $g(\tau^2) \propto 1$이다.

여기서 $\boldsymbol{b} = (b_0, b_1, b_2, b_3, b_4)'$ 이며 $\boldsymbol{x}_i = (1, x_{1i}, x_{2i}, x_{3i}, x_{4i})'$ 이다.

따라서 깁스 표본자를 생성하기 위한 조건부 확률분포는 다음과 같다.

$$\boldsymbol{b}|\,\Theta,\tau^2,\widehat{\Theta}\sim N_p[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\Theta,\ \tau^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]\ ;$$

$$\tau^2|\,\Theta,\boldsymbol{b},\widehat{\Theta}\sim IG\left(\frac{1}{2}\sum_1^m(\Theta_i-\boldsymbol{x}_i'\boldsymbol{b})^2,\ \frac{m-2}{2}\right);$$

$$\Theta_i|\,\boldsymbol{b},\tau^2,\widehat{\Theta}\ \overset{ind}{\sim}\ N[(V_i+\tau^2)^{-1}(\tau^2\widehat{\Theta}_i+V_i\boldsymbol{x}_i'\boldsymbol{b}),\ \tau^2V_i(V_i+\tau^2)^{-1}]$$

여기서 $Z\sim IG(a,b)$는 확률밀도함수가 $f(z)\propto\exp(-a/z)z^{-(b+1)}$ 형태를 가진다는 것을 의미한다.

위의 조건부 확률분포를 사용하여 Gelman과 Rubin(1992)에 따라 깁스 샘플링 계산을 수행한다. 구체적으로 $L=10$개의 병렬체인을 고려하고 각 체인에서 $2d=1000$개의 반복을 사용하여 Θ, $\boldsymbol{b}$와 τ^2의 난수를 각각 1000개 생성한 후, 초기값의 영향을 제거하기 위해 반복의 앞 부분 500개를 버리고 나머지 500개를 사용하여 사후평균과 사후분산을 계산한다. 이 때 Rao-Blackwellization을 사용한 사후평균과 사후분산은 아래와 같이 주어진다.

$$E(\Theta_i|\,\widehat{\Theta})\approx(Ld)^{-1}\sum_{l=1}^{L}\sum_{k=d+1}^{2d}(V_i+\tau^{2(lk)})^{-1}(\tau^{2(lk)}\widehat{\Theta}_i+V_i\boldsymbol{x}_i'\boldsymbol{b}^{(lk)})$$

$$\begin{aligned}V(\Theta_i|\,\widehat{\Theta})\approx\ &(Ld)^{-1}\sum_{l=1}^{L}\sum_{k=d+1}^{2d}\tau^{2(lk)}V_i(V_i+\tau^{2(lk)})^{-1}\\&+(Ld)^{-1}\sum_{l=1}^{L}\sum_{k=d+1}^{2d}(V_i+\tau^{2(lk)})^{-2}(\tau^{2(lk)}\widehat{\Theta}_i+V_i\boldsymbol{x}_i'\boldsymbol{b}^{(lk)})^2\\&-[(Ld)^{-1}\sum_{l=1}^{L}\sum_{k=d+1}^{2d}(V_i+\tau^{2(lk)})^{-1}(\tau^{2(lk)}\widehat{\Theta}_i+V_i\boldsymbol{x}_i'\boldsymbol{b}^{(lk)})]^2\end{aligned}$$

여기서 $\tau^{2(lk)}$와 $\boldsymbol{b}^{(lk)}$는 l번째 체인의 k번째 반복에서 생성된 τ^2과 $\boldsymbol{b}$의 값을 나타낸다.

위의 공식을 사용한 계층적 베이즈 추정값 계산결과는 다음 절에서 경험적 베이즈 추정값과 함께 표로 정리했다.

주어진 실업률 추정 문제에서 계층적 베이지안 Fay-Herriot 모형에 대한 깁

스 표본자의 수렴 여부를 검증하기 위해 10개의 병렬체인에서 각 1000개의 반복을 통해 생성된 난수를 사용하여 $\Theta_i\ (i=1,\ldots,191)$에 대한 잠재적 척도축소인자 $\widehat{R}_i$값을 계산한다. 이 때 191개의 Θ_i에 대해 모두 1에 매우 가깝기 때문에 깁스 표본자가 매우 잘 수렴한다고 말할 수 있다.

깁스 표본자로부터 생성된 $\Theta^{(l)}$을 사용하여 5000개의 $\mathbf{y}_{new}$를 생성한다. 이 때 불일치측도 $d(\mathbf{y}_{obs}, \Theta^{(l)})$와 $d(\mathbf{y}_{new}, \Theta^{(l)})$이 5000개 만들어지고, 이를 사용하여 사후예측 p값을 구한 결과 0.5062 이었다. 따라서 사용된 Fay-Herriot 모형의 총체적 모형적합이 좋지 않다는 징후는 없다.

3.3.4 경험적 베이즈 및 계층적 베이즈 추정결과

2003년 9월의 경제활동인구조사 자료를 사용하여 전체 191개의 소지역에 대한 실업자 총계 추정을 다음의 2가지 경우로 나누어서 고려해 보았다. 보조정보로 공변량을 4개를 사용하여 모든 소지역을 하나의 모형 혹은 특별시, 광역시, 도에 대해 각각 개별 모형에 적합시켜 보았다. 이를 다음과 같이 구분하여 표를 작성하였다.

(1) 4개의 공변량을 사용하여 모든 소지역을 동일 모형에서 동시에 추정했을 때

(2) 4개의 공변량을 사용하여 특별시, 광역시, 도에 대해 개별 모형을 사용하여 추정했을 때

(2)의 경우 계산결과가 (1)의 경우와 크게 차이가 나지 않으므로 여기서는 서울특별시, 부산광역시, 경기도, 충청북도의 경우만 예시를 위해 표로 작성하였다.

모든 소지역에 대한 자료를 병합한 (1)의 경우 랜덤지역효과와 사전분포를

무시하고 고전적 가중회귀모형을 적합시켰을 때 결정계수가 약 40%정도이나, 특별시, 광역시, 도에 대해 개별 자료로 분할하여 각각 고전적 가중회귀모형을 적합시키면 지역에 따라 결정계수값의 범위가 10-90%정도로 상당한 차이가 난다. 그 중 서울특별시는 결정계수값이 약 10%정도 밖에 되지 않으나 충청북도의 경우 결정계수값이 88%정도까지 커진다. 한편, 각 지역별 회귀모형에서 단계별 변수선택을 실시하여도 선택되는 변수에 상당한 차이가 있으므로 어느 변수가 중요한 변수인지 파악하기가 쉽지 않다.

보조정보로 합성추정량을 구성하는 4개의 성별(남,여)-연령대별(15-29세, 30세이상) 범주에 해당하는 4개의 공변량 대신 4개의 변수를 합한 합성추정량을 모형의 공변량으로 사용하여 소지역추정을 시도해 보았으나 계산결과가 거의 비슷하므로 여기서는 보고를 생략하였다.

(1)에서 모든 특별시, 광역시, 도에 대해 계층적 베이즈 추정값과 경험적 베이즈 추정값이 크게 차이가 없으며 대부분 직접추정값보다 CV값이 상당히 줄어든다. 그러나 계층적 베이즈 추정값이 경험적 베이즈 추정값보다 전체적으로 작은 CV값을 가진다. (2)에서 예시한 4가지 경우에도 동일한 결과를 얻었다.

경험적 베이즈 추정값에 대해 잭나이프 추정량을 사용하여 MSE 계산을 시도해보지 않았지만 아마도 계층적 베이즈 추정값에 대한 사후표준편차와 비슷한 결과가 나오리라고 예상되며, 잭나이프 추정으로 경험적 베이즈 추정량의 MSE 값을 개선할 여지가 있으리라 사료된다.

(1) 보조정보로 4개의 공변량을 사용하여 모든 소지역을 동일 모형에서 동시에 추정했을 때

<표 3.1> 서울특별시의 구단위 실업자 총계 추정결과

지구	시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
		추정값	추정오차	CV (%)	추정값	추정오차	CV (%)	추정값	추정오차	CV (%)
강남동	서초구	6029	1373	22.8	6872	750	10.9	6179	1305	21.1
	강남구	11863	2485	20.9	10209	922	9.0	11453	2155	18.8
	송파구	13698	3100	22.6	11747	943	8.0	13105	2516	19.2
	강동구	6952	2208	31.8	8428	842	10.0	7439	1954	26.3
강남서	양천구	13505	4235	31.4	8652	883	10.2	11150	2974	26.7
	강서구	5167	2011	38.9	8961	850	9.5	6109	1816	29.7
	구로구	4231	1625	38.4	6759	778	11.5	4731	1513	32.0
	금천구	12704	5212	41.0	4853	844	17.4	7994	3235	40.5
	영등포구	5847	1156	19.8	6815	699	10.3	5991	1114	18.6
	동작구	11048	2123	19.2	7889	827	10.5	10370	1892	18.2
	관악구	10127	2772	27.4	10026	918	9.2	10269	2329	22.7
강북동	성동구	6965	2938	42.2	6095	829	13.6	6794	2396	35.3
	광진구	11768	1920	16.3	7792	821	10.5	10994	1744	15.9
	중랑구	7708	2959	38.4	7782	850	10.9	7892	2412	30.6
	성북구	10243	2986	29.2	8131	855	10.5	9620	2427	25.2
	강북구	6004	3740	62.3	6613	849	12.8	6488	2777	42.8
	도봉구	7731	2335	30.2	6590	816	12.4	7518	2035	27.1
	노원구	11002	4164	37.8	10859	929	8.6	11166	2971	26.6
강북서	종로구	4187	1687	40.3	3398	755	22.2	4095	1561	38.1
	중구	7685	950	12.4	4674	693	14.8	7432	926	12.5
	용산구	5131	1540	30.0	4303	745	17.3	5042	1443	28.6
	은평구	8423	1817	21.6	8325	804	9.7	8478	1666	19.6
	서대문구	5102	1992	39.0	6197	796	12.8	5420	1796	33.1
	마포구	4298	1028	23.9	5763	672	11.7	4463	998	22.4
	동대문구	7693	3360	43.7	7083	850	12.0	7626	2612	34.3
계		205,111			184,816			197,818		

<표 3.2> 부산광역시의 구/군단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
중구/기장군	3663	1131	30.9	2418	684	28.3	3555	1091	30.7
서구	3018	1659	55.0	2244	755	33.6	2938	1540	52.4
영도구	1591	762	47.9	1976	566	28.7	1633	750	45.9
동래구	2069	886	42.8	3161	628	19.9	2182	867	39.7
남구	4239	1468	34.6	4354	756	17.4	4319	1385	32.1
사하구	5296	1843	34.8	5236	816	15.6	5384	1689	31.4
금정구	2057	691	33.6	2822	547	19.4	2122	682	32.1
연제구	4127	1887	45.7	3389	785	23.2	4053	1718	42.4
수영구	2096	1352	64.5	2386	718	30.1	2174	1285	59.1
동구	2123	802	37.8	1906	580	30.4	2120	788	37.1
부산진구	7510	2379	31.7	6121	869	14.2	7283	2076	28.5
북구	7325	1874	25.6	4853	809	16.7	6909	1710	24.8
해운대구	10063	2532	25.2	5747	866	15.1	8914	2172	24.4
강서구	0	0	NA	0	0	NA	0	0	NA
사상구	4749	1602	33.7	4197	772	18.4	4726	1496	31.7
계	59,926			50,810			58,312		

<표 3.3> 인천시의 구/군단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
중구	2566	610	23.8	2107	498	23.6	2546	604	23.7
동구	803	567	70.5	931	468	50.3	817	561	68.7
남구	6935	2677	38.6	7272	940	12.9	7296	2283	31.3
연수구	7720	1566	20.3	4703	789	16.8	7301	1467	20.1
남동구	8197	1863	22.7	6629	866	13.1	8005	1711	21.4
부평구	9324	1725	18.5	8995	946	10.5	9406	1615	17.2
계양구	3235	892	27.6	4202	652	15.5	3347	874	26.1
서구	8529	2620	30.7	5566	899	16.1	7799	2236	28.7
강화군	441	456	103.2	566	400	70.6	452	453	100.1
계	47,750			40,971			46,969		

<표 3.4> 대구광역시의 구/군단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
중구	2422	1140	47.1	1751	679	38.8	2365	1099	46.5
동구	5330	1708	32.0	5283	791	15.0	5365	1581	29.5
서구	6220	1732	27.8	4746	776	16.4	5992	1598	26.7
남구	5324	1512	28.4	3609	750	20.8	5085	1420	27.9
북구	5260	1530	29.1	6401	796	12.4	5478	1439	26.3
수성구	6328	1742	27.5	6809	826	12.1	6455	1611	25.0
달서구	10146	2663	26.2	9347	938	10.0	9969	2267	22.7
달성군	4335	1702	39.3	2609	767	29.4	4062	1574	38.7
계	45,365			40,555			44,771		

<표 3.5> 광주광역시의 구단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
동구	2085	588	28.2	2002	485	24.2	2087	582	27.9
서구	4815	1128	23.4	4583	722	15.7	4828	1091	22.6
남구	4172	1098	26.3	3696	707	19.1	4153	1063	25.6
북구	7948	1581	19.9	7207	937	13.0	7920	1498	18.9
광산구	2166	691	31.9	2742	544	19.8	2219	681	30.7
계	21,186			20,230			21,207		

<표 3.6> 대전광역시의 구단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
동구	4510	1319	29.2	3524	721	20.4	4400	1257	28.6
중구	5861	1517	25.9	3952	758	19.2	5582	1425	25.5
서구	7107	1224	17.2	6495	745	11.5	7039	1176	16.7
유성구	604	395	65.3	909	360	39.7	621	393	63.3
대덕구	1909	835	43.7	2323	595	25.6	1949	818	42.0
계	19,991			17,203			19,591		

<표 3.7> 울산광역시의 구/군단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
중구	2878	709	24.6	2815	543	19.3	2886	699	24.2
남구	5872	1139	19.4	4610	701	15.2	5769	1099	19.0
동구	1478	430	29.1	1635	383	23.4	1490	427	28.7
북구	1137	412	36.3	1142	369	32.4	1141	410	36.0
울주군	2314	1037	44.8	1987	655	33.0	2306	1006	43.6
계	13,679			12,189			13,592		

<표 3.8> 경기도의 시/군단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
평택시	3745	1280	34.2	4026	718	17.8	3820	1223	32.0
용인시	6195	1885	30.4	6572	844	12.8	6370	1723	27.0
파주시	1972	1335	67.7	2553	721	28.2	2085	1271	60.9
이천시	5025	1655	32.9	2759	761	27.6	4676	1536	32.8
안성시	484	379	78.2	717	346	48.3	498	377	75.7
광주시	2864	414	14.5	2740	372	13.6	2862	412	14.4
포천군	0	0	NA	0	0	NA	0	0	NA
가평/양평군	0	0	NA	0	0	NA	0	0	NA
수원시	7419	3033	40.9	11993	981	8.2	9420	2490	26.4
성남시	8294	2256	27.2	11266	938	8.3	9245	2003	21.7
안양시	7614	2443	32.1	7268	853	11.7	7652	2111	27.6
부천시	17212	4212	24.5	10533	954	9.1	14005	2999	21.4
동두천시	2139	1081	50.5	1315	667	50.7	2072	1046	50.5
고양시	12326	3586	29.1	9505	1078	11.3	11308	2812	24.9
과천/구리시	2163	1703	78.7	2873	763	26.5	2345	1575	67.2
군포시	5342	1982	37.1	3452	785	22.8	4992	1787	35.8
하남시	5277	1060	20.1	3000	688	22.9	5068	1026	20.3
김포시	0	0	NA	0	0	NA	0	0	NA
화성시	1949	622	31.9	2253	501	22.3	1977	615	31.1
여주/연천군	2089	1467	70.2	1365	732	53.6	2019	1382	68.5
의정부시	5071	2128	42.0	4563	810	17.8	5037	1894	37.6
광명시	7366	2690	36.5	4423	821	18.6	6518	2254	34.6
안산시	12976	3433	26.5	7898	891	11.3	11026	2658	24.1
남양주시	3828	2207	57.6	4488	827	18.4	4102	1952	47.6
오산시	1079	835	77.3	1182	590	49.9	1099	818	74.4
시흥시	2025	1240	61.2	3386	711	21.0	2231	1188	53.3
의왕시	6672	685	10.3	4612	596	12.9	6544	675	10.3
양주군	2955	2006	67.9	1322	780	59.0	2651	1804	68.1
계	134,081			116,064			129,622		

<표 3.9> 강원도의 시/군단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
춘천시	2240	1355	60.5	1683	720	42.8	2185	1288	59.0
원주시	2593	865	33.3	2072	608	29.3	2559	846	33.1
강릉시	679	495	72.9	856	426	49.8	692	491	71.0
동해시	780	781	100.1	683	571	83.6	781	768	98.3
태백시	409	389	95.1	388	352	90.8	410	387	94.4
속초시	777	547	70.4	683	457	66.9	775	542	69.9
삼척시	370	334	90.4	377	310	82.1	372	333	89.7
홍천군	276	220	79.8	270	213	78.8	276	220	79.5
횡성/영월 /고성군	318	163	51.4	319	160	50.3	318	163	51.3
평창/정선 /양양군	0	0	NA	0	0	NA	0	0	NA
철원/화천군	1021	829	81.2	604	591	97.9	997	813	81.6
양구/인제군	0	0	NA	0	0	NA	0	0	NA
계	9,463			7,935			9,365		

<표 3.10> 충청북도의 시/군단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
청주시	6691	1238	18.5	6373	750	11.8	6655	1189	17.9
충주시	2392	1276	53.3	2188	705	32.2	2383	1219	51.2
제천시	1332	529	39.7	1358	446	32.9	1336	525	39.3
청원시	1929	479	24.8	1753	416	23.7	1922	475	24.7
보은/영동군	278	205	73.9	290	199	68.6	279	205	73.5
옥천/진천군	900	801	89.0	765	578	75.6	900	786	87.3
괴산군	296	319	107.5	305	297	97.3	298	318	106.4
음성군	635	464	73.0	585	405	69.2	636	461	72.4
단양군	605	450	74.4	503	396	78.8	602	447	74.2
계	15,058			14,120			15,011		

<표 3.11> 충청남도의 시/군단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
천안시	6836	1842	26.9	4464	789	17.7	6397	1684	26.3
공주시	1340	688	51.4	1295	531	41.0	1343	679	50.5
보령시	457	272	59.6	514	259	50.4	461	272	59.0
아산시	986	523	53.1	1216	444	36.5	1002	519	51.8
서산시	1804	1012	56.1	1500	647	43.1	1789	983	55.0
논산시	1466	990	67.6	1503	640	42.6	1483	963	65.0
군산/서천 /청양군	1006	1259	125.2	1420	702	49.4	1082	1204	111.3
연기/홍성군	4104	2124	51.8	2012	797	39.6	3654	1891	51.7
부여군	0	0	NA	0	0	NA	0	0	NA
예산/태안군	1567	1231	78.6	1560	698	44.7	1589	1180	74.2
당진군	573	612	106.9	754	494	65.5	590	605	102.7
계	20,139			16,238			19,390		

<표 3.12> 전라북도의 시/군단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
전주시	5906	1998	33.8	6472	1073	16.6	5972	1854	31.0
군산시	4486	2925	65.2	2878	854	29.7	3916	2401	61.3
익산시	1837	974	53.0	2711	678	25.0	1916	950	49.6
정읍시	2539	722	28.4	2008	552	27.5	2506	711	28.4
남원시	1429	661	46.2	1242	519	41.8	1421	652	45.9
김제시	1048	631	60.2	1060	504	47.5	1052	624	59.3
완주군	1590	1037	65.2	832	662	79.6	1541	1006	65.3
진안/장수 임실군	401	306	76.4	398	288	72.4	404	306	75.7
무주/순창 /고창군	389	374	96.1	414	343	82.9	395	373	94.4
부안군	0	0	NA	0	0	NA	0	0	NA
계	19,625			18,015			19,123		

<표 3.13> 전라남도의 시/군단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
목포시	5284	1368	25.9	3678	729	19.8	5091	1298	25.5
여수시	6120	1876	30.7	4308	797	18.5	5802	1711	29.5
순천시	2629	714	27.2	2906	549	18.9	2656	704	26.5
나주시	1355	483	35.7	1333	419	31.4	1357	480	35.4
광양시	486	475	97.6	757	414	54.7	504	472	93.5
담양군	492	409	83.0	428	367	85.8	491	407	82.8
곡성/구례 /보성군	484	410	84.9	475	369	77.6	485	408	84.2
고흥군	528	447	84.6	478	394	82.5	528	444	84.1
화순/장흥군	970	1150	118.6	647	683	105.6	948	1108	116.9
강진군	480	493	102.7	386	425	109.9	478	489	102.3
해남군	982	685	69.7	709	532	75.0	970	676	69.7
영암군	0	0	NA	0	0	NA	0	0	NA
무안군	0	0	NA	0	0	NA	0	0	NA
함평군	0	0	NA	0	0	NA	0	0	NA
영광/완도 /진도군	0	0	NA	0	0	NA	0	0	NA
장성군	0	0	NA	0	0	NA	0	0	NA
계	19,810			16,105			19,310		

<표 3.14> 경상북도의 시/군단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
포항시	11599	2029	17.5	6921	950	13.7	10702	1848	17.3
경주시	3376	1614	47.8	3299	777	23.5	3440	1506	43.8
김천시	1700	952	56.0	1700	633	37.2	1721	927	53.9
안동시	1812	1184	65.3	2036	694	34.1	1873	1139	60.8
구미시	4416	1044	23.6	4128	689	16.7	4427	1014	22.9
영주시	378	344	91.0	542	319	58.8	388	342	88.2
영천시	536	517	96.3	757	440	58.2	554	513	92.5
상주시	0	0	NA	0	0	NA	0	0	NA
문경시	482	475	98.6	600	413	68.9	493	472	95.8
경산시	3381	1745	51.6	2660	777	29.2	3322	1610	48.5
군위/영양 /성주군	0	0	NA	0	0	NA	0	0	NA
의성/청송군	564	415	73.6	520	372	71.4	564	413	73.3
영덕/청도군	1018	793	77.9	716	578	80.7	1001	779	77.8
고령/봉화군	468	561	119.7	400	466	116.4	468	556	118.7
칠곡/예천 /울진군	480	635	132.5	627	510	81.3	491	628	127.9
계	30,210			24,906			29,444		

<표 3.15> 경상남도의 시/군단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
창원시	8991	1811	20.1	5760	807	14.0	8448	1662	19.7
마산시	2938	1386	47.2	4304	744	17.3	3183	1316	41.3
진주시	2938	991	33.7	3351	647	19.3	3001	964	32.1
진해시	1208	1429	118.3	1416	724	51.1	1275	1351	105.9
통영시	3965	1811	45.7	1880	770	41.0	3610	1659	45.9
사천시	492	458	93.1	654	401	61.3	504	455	90.3
김해시	4088	1815	44.4	3982	772	19.4	4151	1662	40.0
거제시	1640	578	35.2	1506	475	31.6	1639	572	34.9
밀양시	2993	2001	66.9	2022	779	38.5	2848	1801	63.2
양산시	533	403	75.6	818	365	44.6	551	401	72.8
의령/합천군	0	0	NA	0	0	NA	0	0	NA
함안/하동 /거창군	1018	831	81.6	1025	593	57.8	1033	815	78.8
창녕/고성군	1622	971	59.9	1096	639	58.3	1591	946	59.4
남해/산청 /함양군	1089	408	37.5	1023	367	35.9	1089	406	37.3
계	33,515			28,837			32,923		

<표 3.16> 제주도의 시/군단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
제주시	2535	740	29.2	2548	588	23.1	2532	730	28.8
서귀포시	1598	600	37.6	1297	491	37.8	1583	594	37.5
북제주군	667	230	34.4	641	222	34.6	667	230	34.4
남제주군	167	174	103.9	169	170	100.5	168	174	103.5
계	4,967			4,655			4,950		

(2) 보조정보로 4개의 공변량을 사용하여 특별시, 광역시, 도에 대해 개별 모형을 사용하여 추정했을 때

<표3.1-1> 서울특별시의 구단위 실업자 총계 추정결과

지구	시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
		추정값	추정오차	CV(%)	추정값	추정오차	CV(%)	추정값	추정오차	CV(%)
강남동	서초구	6029	1373	22.8	6796	1247	18.4	6092	1360	22.3
	강남구	11863	2485	20.9	11132	2081	18.7	11783	2442	20.7
	송파구	13698	3100	22.6	11224	2211	19.7	13375	2971	22.2
	강동구	6952	2208	31.8	7950	1786	22.5	7099	2161	30.4
강남서	양천구	13505	4235	31.4	9226	2319	25.1	12603	3862	30.6
	강서구	5167	2011	38.9	6944	1606	23.1	5386	1966	36.5
	구로구	4231	1625	38.4	5441	1411	25.9	4358	1601	36.7
	금천구	12704	5212	41.0	7270	2367	32.6	11098	4569	41.2
	영등포구	5847	1156	19.8	6046	1057	17.5	5875	1149	19.6
	동작구	11048	2123	19.2	9499	1632	17.2	10902	2067	19.0
	관악구	10127	2772	27.4	10161	2291	22.5	10179	2712	26.6
강북동	성동구	6965	2938	42.2	6569	1891	28.8	6958	2809	40.4
	광진구	11768	1920	16.3	10020	1659	16.6	11608	1884	16.2
	중랑구	7708	2959	38.4	7497	1860	24.8	7744	2819	36.4
	성북구	10243	2986	29.2	8437	2014	23.9	10033	2848	28.4
	강북구	6004	3740	62.3	6738	2035	30.2	6239	3465	55.5
	도봉구	7731	2335	30.2	7200	1697	23.6	7704	2266	29.4
	노원구	11002	4164	37.8	8728	2905	33.3	10681	3952	37.0
강북서	종로구	4187	1687	40.3	4869	1464	30.1	4249	1665	39.2
	중구	7685	950	12.4	7259	923	12.7	7658	946	12.4
	용산구	5131	1540	30.0	5351	1301	24.3	5157	1521	29.5
	은평구	8423	1817	21.6	7860	1512	19.2	8406	1785	21.2
	서대문구	5102	1992	39.0	5826	1531	26.3	5203	1949	37.5
	마포구	4298	1028	23.9	4835	980	20.3	4339	1023	23.6
	동대문구	7693	3360	43.7	7518	2005	26.7	7731	3163	40.9
계		205,111	37	0.6	190,396			202,460		

<표3.2-1> 부산광역시의 구/군단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
중구/기장군	3663	1131	30.9	3609	850	23.6	3667	1114	30.4
서구	3018	1659	55.0	2457	833	33.9	2973	1597	53.7
영도구	1591	762	47.9	2205	618	28.0	1616	756	46.8
동래구	2069	886	42.8	2063	715	34.7	2071	879	42.5
남구	4239	1468	34.6	3747	1133	30.2	4208	1440	34.2
사하구	5296	1843	34.8	6077	1141	18.8	5406	1775	32.8
금정구	2057	691	33.6	2185	583	26.7	2062	687	33.3
연제구	4127	1887	45.7	3665	999	27.3	4079	1803	44.2
수영구	2096	1352	64.5	2022	1010	50.0	2094	1328	63.4
동구	2123	802	37.8	1741	625	35.9	2113	796	37.7
부산진구	7510	2379	31.7	7448	1333	17.9	7521	2251	29.9
북구	7325	1874	25.6	7056	1204	17.1	7305	1805	24.7
해운대구	10063	2532	25.2	9775	1691	17.3	10021	2449	24.4
강서구	0	0	NA	0	0	NA	0	0	NA
사상구	4749	1602	33.7	3854	1182	30.7	4674	1566	33.5
계	59,926			57,904			59,810		

<표3.8-1> 경기도의 시/군단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
평택시	3745	1280	34.2	3795	1094	28.8	3769	1250	33.2
용인시	6195	1885	30.4	5927	1513	25.5	6197	1807	29.2
파주시	1972	1335	67.7	2244	1123	50.0	2041	1300	63.7
이천시	5025	1655	32.9	4214	1333	31.6	4887	1594	32.6
안성시	484	379	78.2	571	374	65.5	495	378	76.4
광주시	2864	414	14.5	2847	406	14.3	2865	413	14.4
포천군	0	0	NA	0	0	NA	0	0	NA
가평/양평군	0	0	NA	0	0	NA	0	0	NA
수원시	7419	3033	40.9	10693	2280	21.3	8673	2818	32.5
성남시	8294	2256	27.2	9963	1830	18.4	8751	2146	24.5
안양시	7614	2443	32.1	7646	1763	23.1	7586	2285	30.1
부천시	17212	4212	24.5	11239	2343	20.8	14641	3540	24.2
동두천시	2139	1081	50.5	2124	963	45.4	2137	1063	49.8
고양시	12326	3586	29.1	9264	2770	29.9	11372	3372	29.7
과천/구리시	2163	1703	78.7	2976	1351	45.4	2321	1635	70.4
군포시	5342	1982	37.1	4721	1454	30.8	5210	1876	36.0
하남시	5277	1060	20.1	4616	966	20.9	5174	1042	20.1
김포시	0	0	NA	0	0	NA	0	0	NA
화성군	1949	622	31.9	1972	597	30.3	1962	619	31.5
여주/연천군	2089	1467	70.2	1811	1224	67.6	2049	1425	69.6
의정부시	5071	2128	42.0	5048	1519	30.1	5067	1998	39.4
광명시	7366	2690	36.5	6021	1726	28.7	6964	2451	35.2
안산시	12976	3433	26.5	7946	2463	31.0	11492	3136	27.3
남양주시	3828	2207	57.6	3909	1570	40.2	3945	2067	52.4
오산시	1079	835	77.3	1308	788	60.3	1116	827	74.1
시흥시	2025	1240	61.2	2224	1118	50.3	2118	1219	57.5
의왕시	6672	685	10.3	6212	673	10.8	6610	680	10.3
양주군	2955	2006	67.9	2077	1486	71.5	2790	1899	68.1
계	134,081			121,368			130,232		

<표3.10-1> 충청북도의 시/군단위 실업자 총계 추정결과

시군구	직접추정량			계층적 베이즈 추정량			경험적 베이즈 추정량		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
청주시	6691	1238	18.5	6746	1257	18.6	6694	1239	18.5
충주시	2392	1276	53.3	2139	980	45.8	2372	1254	52.8
제천시	1332	529	39.7	1431	476	33.2	1338	525	39.3
청원시	1929	479	24.8	1825	433	23.7	1924	477	24.8
보은/영동군	278	205	73.9	311	201	64.6	279	205	73.5
옥천/진천군	900	801	89.0	620	655	105.6	886	792	89.4
괴산군	296	319	107.5	290	309	106.8	297	318	107.3
음성군	635	464	73.0	612	402	65.7	636	461	72.5
단양군	605	450	74.4	597	423	70.9	603	449	74.4
계	15,058			14,571			15,029		

4. 미국의 시계열모형을 사용한 실업률 추정

4.1 배경

실업률은 미국 BLS(Bureau of Labor Statistics)에 의해 발표되는 다섯 개의 중요한 경제변수 중 하나이며 실업자 수를 노동력의 백분율로 나타낸다. BLS는 미국 전체뿐만 아니라 여러 지리적 그리고 인구통계적 부영역(subdomain)에 대한 월별 실업률을 발표한다. 예를 들면, 50개 주와 컬럼비아 특별구(D.C.), 모든 도시통계지역(Metroplitan Statistical Area; MSA), 모든 군(New England에 있는 시와 읍), 인구 25,000명 이상의 도시에 대한 실업률 추정값이 발표된다. 지역 실업률은 지역적 계획 수립이나 여러 연방 보조 프로그램 하에서의 기금 분배에 사용된다. 예를 들면, 1991 회계연도에 90억 달러 이상의 정부예산이 소지역 실업통계에 완전히 혹은 부분적으로 근거하여 주와 소지역에서 집행되었다.

전국적인 실업률은 CPS 자료를 사용하여 소위 A-K 추정량에 의해 추정된다. CPS 월별 표본은 72,000개의 주택 단위로 구성되어 있으며, 모든 주와 컬럼비아 특별구에 속하는 1,000개 이상의 군으로 구성된 729개 지역에서 수집된다. CPS는 4-8-4 연동교체설계(rotating panel design)를 사용하며 월별로 미국 Census Bureau에서 조사한다. 주어진 달에서 CPS 표본은 8개의 교체패널에 해당하는 8개의 부표본(subsample)으로 집단화 된다. 특별한 교체패널에 속하는 모든 단위들은 동시에 표본에 들어오고 나간다. 주어진 교체패널은 연속적인 4개월 동안 표본에 머물고, 이어서 8개월간 표본을 떠났다가, 다시 연속되는 4개월 동안 표본에 되돌아 온다. 그 후 표본에서 완전히 탈락하고 가까운 가구 그룹에 의해 대체된다. 매달 표본이 되는 두 개의 새로운 교체 그룹 중에서 하나는 완전히 새로운 그룹(처음 등장하는 패널)이고 다른 하나는 되돌아오는 그룹(8개월 동안 표본에서 나가 있다가 다섯 번째 등장하는 패널)이다. 따라서 CPS 설계에서 8개의 연동교체 그룹 중에서 6개는 두 연속적인 달에서 공통이고(즉, 75% 중복), 4개는 연속적인 연도에서 같은 달에 공통이다(즉, 50% 중복).

미국의 각 주는 CPS 표본에서 주에 대한 가능한 표본의 크기에 따라 direct-use state 혹은 indirect-use state로 분류된다. 만약 주로부터 가능한 표본의 크기가 크면 direct-use state로 분류되고, 그렇지 않으면 indirect-use state로 분류된다. California, Florida, Illinois, Massachusetts, Michigan, New Jersey, New York, North Carolina, Ohio, Pennsylvania 그리고 Texas는 CPS에서 11개의 direct-use state이다. 나머지 주와 컬럼비아 특별구는 indirect-use state로 고려된다. indirect-use state에 대한 직접 CPS 추정값은 신뢰할 수 없다. 왜냐하면 CPS는 indirect-use state에 대해서 충분한 표본을 제공하지 않기 때문이다. 따라서 indirect-use state에 대해서 CPS 추정값을 개선할 필요가 있다.

CPS 자료의 반복적 특징과 설명변수로 사용가능한 다른 행정자료의 가능성은 신뢰할 만한 모형기반 주별 실업률 추정량의 개발에 대한 좋은 연구영역를 제공한다. 사실 현재 BLS 방법은 본질적으로 CPS 시계열 자료로부터의 정보와 실업보험(unemployment insurance; UI) 시스템으로부터의 관련된 보조 자료를 결합시키기 위해 칼만 필터(Kalman filter)를 사용하는 추정된 최량선형비편향예측(best linear unbiased prediction) 방법이다. 이 방법은 Tiller(1992)에 의해 개발되었고 이 과제에 대한 그의 초기 연구에 근거하고 있다. (Tiller, 1989 참조). Scott과 Smith(1974)의 독창적 최초 아이디어에 따라 Tiller는 CPS 표본 추정값이 확률적으로 변하는 실제 실업 급수(신호)와 샘플링에 의해 생성되는 오차(잡음)의 합으로 표현될 수 있다고 가정하였다. 시간에 따른 CPS 표본의 상당한 중복과 탈락된 가구의 가까운 가구로의 대체 정책으로 말미암아 표본오차 급수에는 강한 자기상관이 존재한다. 잡음급수에 대한 표본 자기상관을 반영하기 위해 자기회귀이동평균(autoregressive moving average; ARMA) 모형(Bell과 Hillmer, 1990)과 보조변수로서 UI 청구률을 사용한 구조적 시계열모형을 사용함으로 Tiller(1992)는 주별 실업률의 추정량을 제안하였다. UI 청구률은 전체 비농업 노동력 중에서 UI 수혜를 요구한 실업노동자의 백분율로 정의된다.

BLS 방법의 중요한 단점은 주별 실업률의 추정값에 관련된 불확실성의 측도를 제공하지 못한다는 것이다. 따라서 변동의 모든 출처를 반영하는 계층적 베이즈 방법을 사용함으로 이 문제를 극복하고자 한다. 제안된 방법은 Ghosh, Nangia와 Kim(1996), Pfeffermann과 Burck(1990), Rao와 Yu(1992,1994), 그리고 Singh, Mantel과 Thomas(1991) 등에 의해 이미 고려된 모형과는 다른 계층적 횡단면 및 시계열 모형을 사용한다. CPS의 가까운 가구 교체정책으로 인해 상당히 긴 시차까지의 표본오차상관을 포함하게 된다. Dempster와 Hwang(1993) 그리고 Rao와 Yu(1992, 1994)를 제외한 다른 연구들은 연동구조가 자동적으로 표본오차에 대해 비교적 간단한 이동평균(moving average)모형을 유도한다고 상당히 간단하게 가정한다. 제안한 모형은 Rao와 Yu(1994)의 모형과도 다르다. 중요한 차이점은 제안한 모형이 시간성분에 대해 확률보행모형(random walk model)을 사용하는 반면 Rao와 Yu는 정상자기회귀모형(stationary autoregre-ssive model)을 가정한다.

제안한 모형외에 두 대안적 모형을 고려한다. 첫 번째 대안적 모형은 잘 알려진 Fay-Herriot 모형을 계층적 베이지안 방법으로 표현한 것이다. Fay-Herriot 모형은 과거 조사로부터의 정보는 결합하지 아니하고 단지 다른 지역(주)으로부터의 정보만을 결합하는 모형이다. Tiller(1989, 1992)의 아이디어에 따른 두 번째 대안적 모형은 다른 주로부터의 횡단면 자료는 무시하고 단지 시간으로부터의 정보만을 결합하는 모형이다. 이는 단지 한 개의 주에 대해 제안한 모형의 특별한 경우이다. 이 대안적 시계열모형이 제한된 수의 시점을 사용하여 모형에 있는 많은 수의 모수를 추정하기 때문에 상당히 불안정하고 의미없는 추정값을 생산한다. 이러한 이유로 이 시계열모형을 더 이상 고려하지 아니한다. 이 문제는 Neyman-Scott 문제와 비슷하다.

제안한 계층적 베이즈 모형의 계산을 위해 Gelfand와 Smith(1990)의 깁스 샘플링 알고리즘을 사용하고 깁스 표본자(Gibbs sampler)의 수렴성을 조사하기 위해 Gelman과 Rubin(1992)의 알고리즘을 사용한다. 1985년 1월부터 1988년 12월

까지 CPS 시계열자료를 사용하여 자료분석을 수행한다. New York 주의 UI 청구율 자료는 고려 중인 기간동안에 매우 신뢰할 수 없기 때문에 (Tiller, 1992 참조), 연구에서 New York 주를 제외한다. 따라서 나머지 49개 주와 컬럼비아 특별구에 대한 48개월 자료를 사용한다. 계층적 모형에 대해 최근에 개발된 진단 도구를 사용하여 앞에서 언급한 두 대안적 모형을 비교하고자 한다.

4.2 CPS 추정량

주어진 월(month)에 대한 CPS의 연동교체설계는 CPS 표본에서 가구로 구성된 8개의 다른 연동교체 그룹을 만든다. 특정 월 t에서 만약 어느 가구가 (고려중인 월을 포함하여) r번째 시점의 표본에 포함된다면 ($r = 1, \cdots, 8$), 그 가구는 r번째 연동교체 그룹에 속한다. r번째 연동교체 그룹만의 자료를 사용하여 t번째 월의 실업률 추정값 m_{rt}를 구할 수 있다. 이 추정값을 month-in-sample 추정값이라 부른다. 따라서 주어진 달에는 각각 동일한 실업률의 추정값으로 8개의 month-in-sample 추정값이 있다.

현재월과 과거월에서 공통인 여섯 개의 연동교체 그룹에 대한 month-in-sample 추정값을 사용하여 과거월로부터 실업수준의 변화를 추정하는 것이 가능하다. 과거월의 실업률 추정값에 이 변화의 추정값을 합하면 현재월의 실업률의 교대 추정값(alternate estimates)이 만들어 진다. Rao와 Graham(1964)이 이 교대 추정값과 8개의 month-in-sample 추정값의 단순 평균의 볼록결합(convex combination)으로 주어지는 추정값의 집합을 제안하였다.

동일시점과 관련된 많은 특성에 대해 8개의 연동교체 그룹으로부터의 추정값은 같은 기대값을 가지지 않는다. 가장 큰 차이점은 첫 번째 month-in-sample 추정값(전혀 새로운 연동교체 그룹에 근거한 추정값)을 모든 8개의 연동교체 그룹의 평균 추정값과 비교할 때 일어난다. (Bailar, 1975; Huang와 Ernst, 1981 참

조). 첫 번째와 다섯 번째 연동교체 그룹에 더 큰 가중치를 부여하기 위해 Gurney와 Daly(1965)는 Rao와 Graham(1964)에 의해 제안된 추정량의 집합을 확장하였다. Gurney와 Daly(1965)의 복합추정량이 A-K 추정량으로 알려졌다. 시점 t에 대한 전국적 실업률 추정값을 생산하기 위해 BLS는 다음과 같이 주어지는 A-K 추정량을 사용한다.

$$Y_t = \sum_{r=1}^{8} a_r m_{rt} - K \sum_{r=1}^{8} b_r m_{r,t-1} + Y_{t-1}$$

여기서 m_{rt}는 시점 t에서 r번 째 month-in-sample 추정량이고,

$a_1 = a_5 = 1/8(1-(K-A))$,

$a_2 = a_3 = a_4 = a_6 = a_7 = a_8 = 1/8(1 + 1/3(K-A))$,

$b_1 = b_2 = b_3 = b_5 = b_6 = b_7 = 1/6$, $b_4 = b_8 = 0$, $K = .4$, and $A = .2$ 이다. (CPS에서 많은 특성에 대한 A와 K의 최적값을 구하기 위해서는 Huang과 Ernst(1981)의 표 2를 참고하기 바람). 시점 t에서 i번째 주에 대한 CPS 추정량을 $Y_{it}(i = 1, \cdots, m; t = 1, \cdots, T)$로 나타내자. CPS 주별 실업률 추정값은 특히 indirect-use state에서 매우 작은 표본을 개별 주로부터 얻을 수 있기 때문에 신뢰할 수 없다.

4.3 Fay-Herriot 모형의 시계열 확장

y_{it} $(i = 1, \cdots, m;\ t = 1, \cdots, T)$를 시점 t에서 i번 째 주의 CPS 실업률 추정값이라 하자. 가능한 자료는 49개 주와 컬럼비아 특별구에 대한 1985년 1월에서 1988년 12월까지의 자료이다.

먼저, y_{it}를 $y_{it} = \Theta_{it} + e_{it}$로 분해하자. 여기서 Θ_{it}는 시점 t에서 i번째

주의 실업률의 참값이며, e_{it}는 평균이 0이고 분산을 σ_{itt}라고 가정하는 표본 오차이다. $(i = 1, \cdots, 50\,;\, t = 1, \cdots, 48)$. CPS 설계가 시간에 따른 표본 단위들의 상당한 중복을 일으키기 때문에 e_{it}와 $e_{it'}$ 간의 상관을 설명하는 것이 중요하다. 이에 관한 공분산을 $\sigma_{itt'}$ $(i = 1, \cdots, 50\,;\, t, t' = 1, \cdots, 48,\ t \neq t')$으로 나타내자.

따라서 CPS 설계로부터 동기를 부여받은 대로 Rao와 Yu(1994)에서와 같이 다음을 가정한다.

$$\boldsymbol{Y}_i | \boldsymbol{\Theta}_i \overset{ind}{\sim} N_{48}(\boldsymbol{\Theta}_i, \boldsymbol{\Sigma}_i),\ i = 1, \cdots, 50$$

여기서 $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{i48})'$, $\boldsymbol{\Theta}_i = (\Theta_{i1}, \cdots, \Theta_{i48})'$, $\boldsymbol{\Sigma}_i = ((\sigma_{itt'}))$ 48×48 양정치행렬이다. Ghosh, Nangia와 Kim(1996), Pfeffermann과 Burck(1990), Singh, Mantel과 Thomas(1991)과는 달리, 본 연구에서는 표본분산-공분산행렬 $\boldsymbol{\Sigma}_i$에 특별한 상관구조를 가정하지 아니한다. 표본분산(즉, $\boldsymbol{\Sigma}_i$의 대각 원소)은 표본 시차상관의 추정값을 제공한 Tiller(1989, 1992)에 의한 일반화분산함수(GVF) 방법을 사용하여 계산된다. 이 표본시차상관과 분산 추정값을 사용하여 49개 주와 컬럼비아 특별구에 대한 $\boldsymbol{\Sigma}_i$의 추정값을 구한다. Datta와 Lahiri(1992)는 Alabama(AL)와 Massachusetts(MA)에 대한 47차까지의 시차상관의 추정값을 구하였다.

다른 출처로부터 정보를 빌려오기 위해 Θ_{it}에 대한 모형을 개발하였다. CPS 추정값 y_{it}는 계절적으로 조정되지 않았기 때문에 년과 월 효과를 가지고 계절성을 설명한다. f_{itu}를 u번째 월에 대한 지시변수로 다음과 같이 정의하자. 만약 $t = u \pmod{12}$이면 $f_{itu} = 1$이고, 그렇지 않으면 $f_{itu} = 0$이다. 또한 만약 $t =$ 12, 24, 36, 48이면 $f_{it12} = 1$이고, 그렇지 않으면 $f_{it12} = 0$이다. 비슷하게

g_{itw}를 w번째 년에 대한 지시변수로 다음과 같이 정의하자. 만약 $12(w-1) < t \leq 12w$이면 $g_{itw} = 1$이고, 그렇지 않으면 $g_{itw} = 0$이다.

앞에서 설명한 표기를 사용하여 Θ_{it}를 다음과 같이 분해하자.

$$\Theta_{it} = x_{it}\beta_i + v_i + \sum_{u=1}^{12} f_{itu}\gamma_{iu} + \sum_{w=1}^{4} g_{itw}\zeta_{iw} + \alpha_{it}$$

여기서 v_i와 β_i는 주에 관련된 절편과 기울기이며 α_{it}는 위의 모형에 포함되지 않은 변동을 설명하기 위해 필요한 오차항이다. 완전계수(full-rank) 선형모형을 가지기 위한 통상적 제약조건으로 $\gamma_{i12} = -\Sigma_{u=1}^{11}\gamma_{iu}$과 $\zeta_{i4} = -\Sigma_{w=1}^{3}\zeta_{iw}$을 부과한다.

49개 주와 컬럼비아 특별구에 대해 $16 \times 50 = 800$ 개의 모형 모수 $v_i, \beta_i, \boldsymbol{\gamma}_i = (\gamma_{i1}, \cdots, r_{i11})'$, $\boldsymbol{\zeta}_i = (\zeta_{i1}, \zeta_{i2}, \zeta_{i3})'$가 있으며 단지 $48 \times 50 = 2{,}400$ 개의 관측값 y_{it} $(i = 1, \cdots, 50,\ t = 1, \cdots, 48)$을 가지고 이들 모수들을 추정한다. 이는 심각한 추정문제를 일으킨다. Ghosh, Nangia와 Kim(1996)에서도 비슷한 문제에 직면했으나 그들은 초모수가 주에 관련되지 않는다고 가정함으로서 문제를 해결한 바 있다. 그러나 이러한 가정은 주별 실업률의 추정값이 지나치게 평활될 가능성이 있다고 보여진다. 절충으로 초모수가 다음에 주어지는 공통 확률분포로부터의 50개 독립 구현이라고 가정한다.

$$v_i \overset{ind}{\sim} N(v, r_v^{-1}),\ \beta_i \overset{ind}{\sim} N(\beta, r_\beta^{-1}),\ \boldsymbol{\gamma}_i \overset{ind}{\sim} N_{11}(\boldsymbol{\gamma}, \boldsymbol{W}_1^{-1})$$

$$\boldsymbol{\zeta}_i \overset{ind}{\sim} N_3(\boldsymbol{\zeta}, \boldsymbol{W}_2^{-1})$$

여기서 $v_i, \beta_i, \boldsymbol{\gamma}_i$와 $\boldsymbol{\zeta}_i$는 서로 독립이다$(i = 1, \cdots, 50)$. 이와같이 모형을 설정하면 다른 주에 대한 y_{it}사이의 상관이 허용된다. 한편, 오차항 α_{it}는 다음과

같은 확률보행모형을 따른다고 가정한다.

$$\alpha_{it} | \alpha_{i,t-1} \sim N(\alpha_{i,t-1}, r_\alpha^{-1})$$

여기서 $t = 2, \cdots, 48$ 이며 $i = 1, \cdots, 50$ 과 독립이다. 이는 α_{it}에 정상모형을 가정한 Rao와 Yu(1994)의 연구와 대조된다. Ghosh, Nangia와 Kim(1996)의 시간성분에 대한 확률보행모형은 주에 관련되어 있지 않다.

마지막으로 초모수에 다음과 같은 부적절 사전분포(improper prior)를 가정한다.

$$\begin{aligned} & f(v, \beta, \gamma, \zeta, r_v, r_\beta, \boldsymbol{W}_1, \boldsymbol{W}_2, r_\alpha) \\ & \propto r_v^{1/2b-1} e^{-1/2a\, r_v} \\ & \quad \times r_\beta^{1/2d-1} e^{-1/2c\, r_\beta} \times |\boldsymbol{W}_1|^{(k-11-1)/2} e^{-1/2\, tr(\boldsymbol{S}_1 \boldsymbol{W}_1)} \\ & \quad \times |\boldsymbol{W}_2|^{(l-3-1)/2} e^{-1/2\, tr(\boldsymbol{S}_2 \boldsymbol{W}_2)} \times r_\alpha^{1/2f-1} e^{-1/2e\, r_\alpha} \end{aligned}$$

여기서 $\boldsymbol{S}_1 = \Delta_1 \boldsymbol{I}_{11}$, $\boldsymbol{S}_2 = \Delta_2 \boldsymbol{I}_3$, $k = 12$, $l = 4$이며 Δ_1과 Δ_2는 큰 양수이고, $b = d = f = 2$, 그리고 a, c, e는 작은 양수이다. a, c, e의 여러 가지 작은 값과 Δ_1과 Δ_2의 여러 가지 큰 값을 시도하였으나 Θ_{iT}의 사후평균과 사후분산은 거의 변하지 않았다. 이와같은 모형설정으로 모든 $48 \times 50 = 2{,}400$개의 관측값이 초모수를 추정하는데 사용되었다.

제안한 모형에 대한 두가지 대안이 가능하다. 첫 번째는 잘 알려진 Fay-Herriot 모형을 변형시켜 과거 시계열 자료는 결합하지 아니하고 단지 다른 주로부터의 횡단면 정보만을 결합하는 모형이다. 정해진 시간 t에 대해 다음과 같은 모형을 가정하자.

$$y_{it} | \Theta_{it} \overset{ind}{\sim} N(\Theta_{it}, \sigma_{itt}), \quad i = 1, \cdots, 50;$$

$$\Theta_{it} \overset{ind}{\sim} N(v_t + x_{it}\beta_t,\ \Psi_t),\ \ i = 1, \cdots, 50;$$

$$v_t,\ \beta_t \sim uniform(R^2),\ \ \Psi_t \sim uniform(R^+)$$

여기서 v_t와 β_t는 Ψ_t와 서로 독립이다.

두번 째 모형은 Tiller(1989, 1992)의 관점에서 제안된 모형이다. 이 모형은 다른 주로부터의 횡단면 정보는 결합하지 아니하는 시계열모형이다. 이 모형은 다음과 같이 표현된다.

$$\boldsymbol{Y}_i | \Theta_i \overset{ind}{\sim} N_{48}(\Theta_i, \Sigma_i),\ \ i = 1, \cdots, 50;$$

$$\Theta_{it} = x_{it}\beta_i + v_i + \sum_{u=1}^{12} f_{itu}\gamma_{iu} + \sum_{w=1}^{4} g_{itw}\zeta_{iw} + \alpha_{it};$$

$$\alpha_{it} | \alpha_{i,t-1} \sim N(\alpha_{i,t-1},\ r_{i\alpha}^{-1})$$

이 때 초모수에 대한 사전분포는 다음과 같이 주어진다.

$$f(v_i, \beta_i, \gamma_i, \zeta_i, r_{i\alpha}) \propto r_{i\alpha}^{1/2f-1} e^{-1/2e\, r_{i\alpha}}$$

여기서 $e \approx 0$ 이고 $f = 2$ 이다.

마지막 모형은 근본적으로 모든 17개의 초모수(즉, v_i, β_i, γ_i, ζ_i, $r_{i\alpha}$)가 단지 48개의 관측값을 사용하여 추정되어야만 한다는 것을 의미한다. 실업률 추정 문제의 응용에서 초모수 추정값이 매우 불안정하기 때문에 이는 제안한 모형에 대한 실행 가능한 대안은 아니다.

4.4 깁스 표본자와 제안한 계층적 베이즈 모형의 실행

우리의 주된 목표는 현재시점 T에서 i번째 주의 실업률의 참값인 소지역 평균 Θ_{iT}를 추정하는 것이다. 계층적 베이즈 접근에서 Θ_{iT}는 사후평균으로 추정한다. 계층적 베이즈 추정값에서의 불확실성은 Θ_{iT}의 사후표준편차에 의해 측정된다. 깁스 표본자를 사용하여 Θ_{iT}의 사후평균과 사후분산을 계산한다. (예컨대, Gelman과 Rubin, 1992; Gelfand와 Smith, 1990 참조).

깁스 표본자는 확률변수 $U_1, \cdots, U_n$의 주변, 조건부, 그리고 결합분포를 제공하는 몬테칼로 마코브 갱신 절차이다. $U_1, \cdots, U_n$의 결합밀도는 $\boldsymbol{U}_{-r} = (U_1, \cdots, U_{r-1}, U_{r+1}, \cdots, U_n)'$이 주어졌을 때 모든 U_r의 조건부밀도를 모은 것, 즉 $f(U_r | \boldsymbol{U}_{-r})$, $r = 1, \cdots, n$에 의해 유일하게 결정된다. 깁스 표본자는 조건부밀도 $f(U_r | \boldsymbol{U}_{-r})$, $r = 1, \cdots, n$로부터 표본의 생성을 요구한다. 초기값 $U_1^{(0)}, \cdots, U_n^{(0)}$에서 시작하여 난수 $U_1^{(1)}$을 $f(U_1 | \boldsymbol{U}_{-1}^{(0)})$으로부터 처음 생성한다. 비슷하게 $U_2^{(1)}, \cdots, U_n^{(1)}$가 $f(U_2 | U_1^{(1)}, U_3^{(0)}, \cdots, U_n^{(0)})$, …, $f(U_n | \boldsymbol{U}_{-n}^{(1)})$으로부터 계속하여 생성된다. 따라서 첫 반복 $U_1^{(1)}, \cdots, U_n^{(1)}$을 얻는다. 반복수 R이 무한대로 증가할 때 $(U_1^{(R)}, \cdots, U_n^{(R)})$이 어떤 정칙조건 하에서 $(U_1, \cdots, U_n)$로 분포수렴한다.

깁스 표본자에 의한 마코브 연쇄 몬테칼로 방법을 사용하여 사후분포로부터 표본을 생성하기 위해 모든 조건부 분포로부터 표본을 추출하는 것이 필요하다. 우리의 응용에 필요한 모든 조건부 분포는 다음과 같이 주어진다. 주어진 조건부분포에서 "rest"는 데이터셋과 조건부분포가 고려 중인 모수를 제외한 모형에 있는 모든 다른 모수를 지칭한다. 아래에 주어지는 조건부 분포는 모두 쉽게 표본이 추출되는 표준분포들이다.

(1) (현재시점 T에서) Fay-Herriot 모형의 모든 조건부분포

아래에서 $\beta_T = (\beta_{0T}, \beta_{1T})'$, $\boldsymbol{A}_T = col_{1\le i\le m}(1, x_{iT})$, $\boldsymbol{d}_T = col_{1\le i\le m} y_{iT}$,

$\boldsymbol{p}_T = col_{1\le i\le m}\theta_{iT}$, $\boldsymbol{S}_T = \mathrm{diag}(\sigma_{1TT}, \cdots, \sigma_{mTT})$ 이라 하자. t개의 성분을 가진 열벡터 $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m$에 대해 $col_{1\le i\le m}\boldsymbol{a}_i' = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m)'$ 은 $m\times t$ 행렬이다.

(i) $\boldsymbol{p}_T \mid \text{rest} \sim$

$$N_m\{(\boldsymbol{S}_T^{-1} + \phi_T \boldsymbol{I}_T)^{-1}(\boldsymbol{S}_T^{-1}\boldsymbol{d}_T + \phi_T \boldsymbol{A}_T \beta_T),\ (\boldsymbol{S}_T^{-1} + \phi_T \boldsymbol{I}_T)^{-1}\};$$

(ii) $\beta_T \mid \text{rest} \sim N_2\{(\boldsymbol{A}_T'\boldsymbol{A}_T)^{-1}\boldsymbol{A}_T'\boldsymbol{p}_T,\ \phi_T(\boldsymbol{A}_T'\boldsymbol{A}_T)^{-1}\}$;

(iii) $\phi_T \mid \text{rest} \sim \text{gamma}\{[(\boldsymbol{p}_T - \boldsymbol{A}_T\beta_T)'(\boldsymbol{p}_T - \boldsymbol{A}_T\beta_T)]/2, (m-2)/2\}$;

여기서 $\phi_T = \psi_T^{-1}$이고 gamma $(a/2, b/2)$ 는 형태모수 $b/2$와 평균 b/a를 가지는 감마분포이다.

(2) 제안한 모형에 대한 모든 조건부분포

다음과 같은 $col_{1\le t\le T} x_{it} = \boldsymbol{X}_i$, $((f_{itu} - f_{it12}))_{t=1,\cdots,T,\, u=1,\cdots,11} = \boldsymbol{F}_i$,

$((g_{itw} - g_{it4}))_{t=1,\cdots T,\, w=1,2,3} = \boldsymbol{G}_i$, $\boldsymbol{K}_i = (1 \mid \boldsymbol{X}_i \mid \boldsymbol{F}_i \mid \boldsymbol{G}_i)$,

$(v_i, \beta_i, \gamma_i', \zeta_i')' = \delta_i$, $col_{1\le t\le T}\alpha_{it} = \alpha_i$를 정의한다.

(i) $\theta_i \mid \text{rest}$ 는 $\theta_i = \boldsymbol{K}_i\delta_i + \alpha_i$로 주어진다.

여기서 δ_i와 α_i는 조건부셋에서 결정된다.

(ii) $\delta_i \mid \text{rest} \sim$

$$N_{16}\{(\boldsymbol{K}_i'\Sigma_i^{-1}\boldsymbol{K}_i+\Omega)^{-1}(\boldsymbol{K}_i'\Sigma_i^{-1}(\mathbf{y}_i-\alpha_i)+\Omega\delta),(\boldsymbol{K}_i'\Sigma_i^{-1}\boldsymbol{K}_i+\Omega)^{-1}\}$$

여기서 $\Omega = \text{block diag}\ (r_v, r_\beta, \boldsymbol{W}_1, \boldsymbol{W}_2)$이고 $\delta=(v,\beta,\gamma',\zeta')'$이다.

(iii) $t=T$에 대해,

$\alpha_{iT} \mid \text{rest} \sim$

$$N\left(\frac{\sigma_i^{TT}\left\{\alpha_{iT}^{(l-1)}+\sum_{j=1}^{T}\frac{\sigma_i^{Tj}}{\sigma_i^{TT}}(y_{ij}-\Theta_{ij}^{(l-1)})\right\}+r_\alpha\alpha_{i,T-1}^{(l-1)}}{\sigma_i^{TT}+r_\alpha},(\sigma_i^{TT}+r_\alpha)^{-1}\right);$$

$2\le t\le T-1$에 대해,

$\alpha_{it} \mid \text{rest} \sim$

$$N\left(\frac{\sigma_i^{tt}\left\{\alpha_{it}^{(l-1)}+\sum_{j=1}^{T}\frac{\sigma_i^{tj}}{\sigma_i^{tt}}(y_{ij}-\Theta_{ij}^{(l-1)})\right\}+r_\alpha(\alpha_{i,t-1}^{(l-1)}+\alpha_{i,t+1}^{(l-1)})}{\sigma_i^{tt}+2r_\alpha},(\sigma_i^{tt}+2r_\alpha)^{-1}\right);$$

$t=1$에 대해,

$\alpha_{i1} \mid \text{rest} \sim$

$$N\left(\frac{\sigma_i^{11}\left\{\alpha_{i1}^{(l-1)}+\sum_{j=1}^{T}\frac{\sigma_i^{1j}}{\sigma_i^{11}}(y_{ij}-\Theta_{ij}^{(l-1)})\right\}+r_\alpha\alpha_{i2}^{(l-1)}}{\sigma_i^{11}+r_\alpha},(\sigma_i^{11}+r_\alpha)^{-1}\right);$$

여기서 l은 현재의 반복을 나타내고, $\alpha_{it}^{(l-1)}$은 마지막 반복으로부터 α_{it}의 값

이며, σ_i^{hk}는 Σ_i^{-1}의 (h,k)번째 요소이다.

(iv) $r_\beta \mid \text{rest} \sim \text{gamma}\left[\frac{1}{2}\left\{\sum_{i=1}^{m}(\beta_i-\beta)^2+c\right\},\ \frac{1}{2}(d+m)\right]$;

(v) $r_v \mid \text{rest} \sim \text{gamma}\left[\frac{1}{2}\left\{\sum_{i=1}^{m}(v_i-v)^2+a\right\},\ \frac{1}{2}(b+m)\right]$;

(vi) $r_\alpha \mid \text{rest} \sim \text{gamma}\left[\frac{1}{2}\left\{\sum_{i=1}^{m}\sum_{t=1}^{T}(\alpha_{it}-\alpha_{i,t-1})^2+e\right\},\ \frac{1}{2}(f+mT)\right]$;

(vii) $f(\boldsymbol{W}_1 \mid \text{rest}) \propto |\boldsymbol{W}_1|^{1/2(k-12+m)} e^{-1/2tr\left[\left\{\boldsymbol{S}_1+\sum_{i=1}^{m}(\boldsymbol{\gamma}_i-\boldsymbol{\gamma})(\boldsymbol{\gamma}_i-\boldsymbol{\gamma})^T\right\}\boldsymbol{W}_1\right]}$.

(viii) $f(\boldsymbol{W}_2 \mid \text{rest}) \propto |\boldsymbol{W}_2|^{1/2(l-4+m)} e^{-1/2tr\left[\left\{\boldsymbol{S}_2+\sum_{i=1}^{m}(\boldsymbol{\zeta}_i-\boldsymbol{\zeta})(\boldsymbol{\zeta}_i-\boldsymbol{\zeta})^T\right\}\boldsymbol{W}_2\right]}$.

(ix) $\delta \mid \text{rest} \sim N_{16}(\bar{\delta}, (1/m)\Omega^{-1})$

여기서 $\bar{\delta} = (1/m)\sum_{i=1}^{m}\delta_i$이다.

(3) 두 번째 대안으로 주어진 시계열모형에 대한 모든 조건부 분포

(i) 제안된 모형에서의 (i)과 동일하다.

(ii) $\delta_i \mid \text{rest} \sim$

$$N_{16}\left[(\boldsymbol{K}_i'\Sigma_i^{-1}\boldsymbol{K}_i)^{-1}\boldsymbol{K}_i'\Sigma_i^{-1}(\boldsymbol{y}_i-\alpha_i),(\boldsymbol{K}_i'\Sigma_i^{-1}\boldsymbol{K}_i)^{-1}\right];$$

(iii) 특정 i에 대해 그리고 r_α을 $r_{i\alpha}$로 교체했을 때 제안된 모형에서의 (iii)과 동일하다.

(iv) $r_{i\alpha} \mid \text{rest} \sim \text{gamma}\left[1/2\left\{\sum_{t=1}^{T}(\alpha_{it}-\alpha_{i,t-1})^2+e\right\}, 1/2(f+T)\right]$.

깁스 샘플링을 수행하기 위해 Gelman과 Rubin(1992)의 알고리즘을 따른다. 각 경로에 대해 과대산포를 가지는 분포로부터 생성된 초기값을 사용하는 $s=10$개의 경로(병렬체인)와 $2n=1,000$의 반복을 고려한다. Gelman과 Rubin (1992)에서 제안된 것과 같이 최종 결과에 대한 초기값의 영향을 감소시키기 위해 각 체인에서의 첫 500개 반복을 burn-in 표본으로 간주한다. 각 체인에서의 나머지 500개 반복을 사용하여 사후분포를 근사시키고 기술통계량(예컨대, 사후평균과 분산)을 계산하고 깁스 표본자의 수렴성을 검사한다. 수렴의 모니터링을 설명하기 위해 주된 관심모수인 Θ_{iT}만 고려하자. Gelman과 Rubin (1992)에 따라 각 500개 반복에 근거한 10개의 수열 평균 $\overline{\Theta}_{igT}$간의 분산인 $B_{iT}/500=\sum_{g=1}^{10}(\overline{\Theta}_{igT}-\overline{\Theta}_{iT})^2/(10-1)$를 계산한다. 여기서 $\overline{\Theta}_{igT}$는 g번째 수열 평균이며 $\overline{\Theta}_{iT}=\sum_{g=1}^{10}\overline{\Theta}_{igT}/10$ 이다. 한편, W_{iT}는 각각 (500-1) 자유도에 근거한 10개의 수열내 분산 s_{igT}^2의 평균을 나타낸다. 즉, $W_{iT}=\sum_{g=1}^{100}s_{igT}^2/10$ 이다. 따라서 다음 값을 구한다.

$$s_{iT}^2=\frac{500-1}{500}W_{iT}+\frac{1}{500}B_{iT}$$

$$V_{iT}=s_{iT}^2+((10)(500))^{-1}B_{iT}$$

마지막으로 잠재적 척도감소인자 $\hat{R}=V_{iT}/W_{iT}\,(i=1,\cdots,50)$를 구한다. 만약 이 잠재적 척도감소인자가 모든 Θ_{iT}에 관해서 1에 가까우면 이는 깁스 표본자에서 수렴이 성취되었음을 암시한다. 잠재적 측도감소인자는 수렴을 검사하기 위해 모형의 다른 모수에 대해서도 비슷하게 정의된다. 모든 Θ_{iT}에 대해 잠재

적 측도감소인자 관점에서 깁스 표본자가 수렴함을 알수있다.

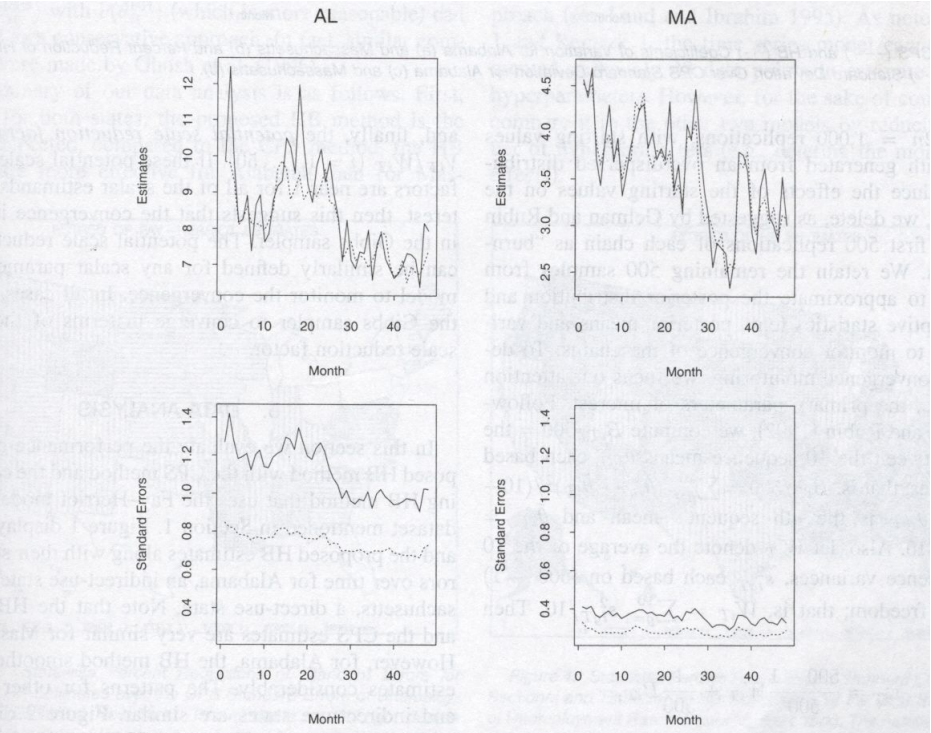

Figure 1. CPS (– – –) and HB (· · ·) Estimates and Standard Errors for Alabama [(a), (c)] and Massachusetts [(b), (d)].

4.5 자료분석

CPS 방법 및 단지 다른 지역(주)으로부터의 횡단면 데이터셋 만을 사용하는 Fay-Herriot 모형에 근거한 계층적 베이즈 방법과 제안한 계층적 베이즈 방법의 수행을 평가한다. Figure 1은 indirect-use-state인 Alabama 주와 direct-use-state인 Massachusetts 주에 대해 시간에 따라 CPS와 제안한 계층적 베이즈 추정값을 표준오차와 함께 나타낸 것이다. HB 추정값과 CPS 추정값은 Massachusetts 주에 대해서는 매우 비슷하다. 그러나 Alabama 주에 대해서는 HB 방법이 CPS 추정값을 상당히 평활시킨다. 이러한 패턴은 다른 direct-use와

indirect-use 주에 대해서 비슷하다. Figure 2는 시간에 따라 CPS와 제안한 HB 추정값에 대한 변동계수를 Fay-Herriot과 제안한 모형에 대한 표준오차의 CPS 표준오차에 대한 퍼센트 감소와 함께 나타낸 것이다. 퍼센트 감소는 PCTRED로 나타내며, 퍼센트 감소의 정의는 CPS 표준오차에 대한 사후표준편차와 CPS 표준오차 차이를 백분율로 표현한 것이다. 따라서 PCTRED = 100 × {(CPS 표준오차 - HB 표준오차)/CPS 표준오차}이다.

다른 모형기반 추정값과 설계기반 CPS 추정값을 비교하기 위한 PCTRED 측도는 주의를 필요로 한다. 일반적으로 주어진 계층적 모형 하에서 사후분산과 CPS 분산 추정값(즉, σ_{itt})은 동일한 불확실성 측도를 추정하지 아니므로 엄밀히 말해 비교 대상이 아니다. 다른 계층적 모형 하에서의 사후분산 비교에 관해서도 동일한 원리가 적용된다.

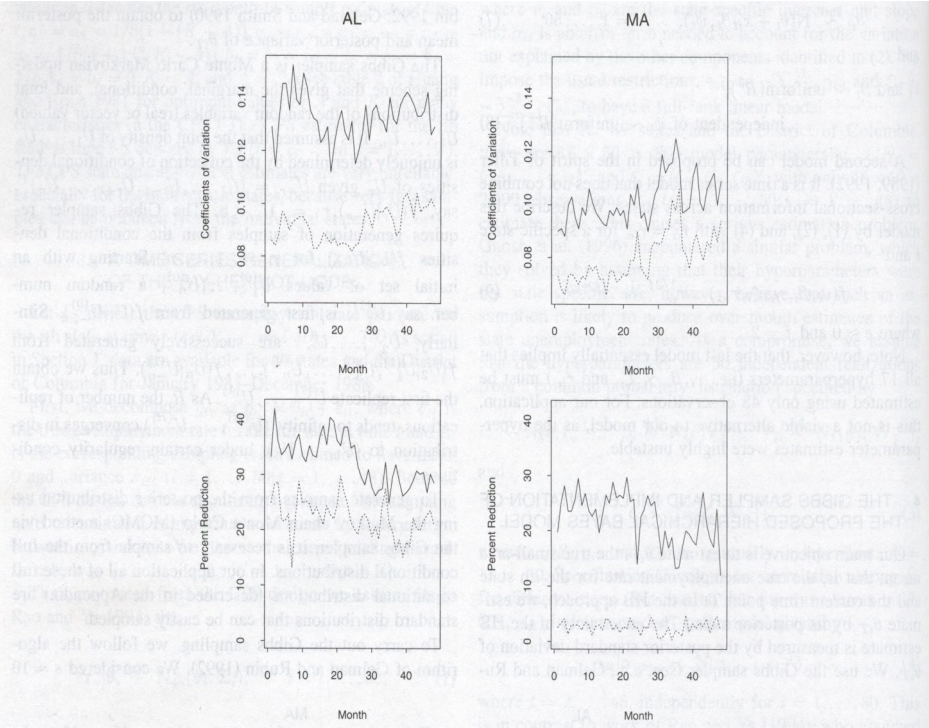

Figure 2. CPS (---) and HB (···) Coefficients of Variation for Alabama (a) and Massachusetts (b) and Percent Reduction of HB (---) and Fay-Herriot (···) Standard Deviation Over CPS Standard Deviation for Alabama (c) and Massachusetts (d).

다음 절에서의 모형적합과 모형비교에 의하면 제안된 모형이 고려된 세 모형 중에서 최적의 모형인 것으로 판명되었다. 따라서 Θ_{it}의 어떤 특정 추정값 d_{it}의 불확실성의 합당한 측도는 $E\{(d_{it}-\Theta_{it})^2|\boldsymbol{y};P\}=v(d_{it})$이다. 여기서 조건부 기대는 제안된 모형 P에 관한 것이다. 다른 추정값 d_{it}에 대한 $v(d_{it})$를 비교함으로 다른 추정 절차를 비교할 수 있다. 그러나 우리의 경우 통상의 불확실성의 측도를 사용한다. 예를 들면, Fay-Herriot 모형에 대해 점추정값 $\hat{\Theta}_{it}^{FH}$의 불확실성 측도로 사후분산 $E\left[(\hat{\Theta}_{it}^{FH}-\Theta_{it})^2|\boldsymbol{y};FH\right]=V(\Theta_{it}|\boldsymbol{y};FH)$을 보고한다. 여기서 조건부 기대는 Fay-Herriot 모형에 관한 것이다. $v(d_{it})$는 제안된 HB 추정값 $\hat{\Theta}_{it}^{HB}$에 대해서 가장 작다. 따라서 $v(\hat{\Theta}_{it}^{HB})$를 $v(\hat{\Theta}_{it}^{FH})$와 비교하는 것(더 합리적임) 보다 차라리 $v(\hat{\Theta}_{it}^{FH})$를 $V(\Theta_{it}|\boldsymbol{y};FH)$와 비교하는 것이 보수적 접근으로 간주된다. 사실 비슷한 비교가 Ghosh, Nangia와 Kim(1996)에서 이루어졌다.

우리의 자료분석 요약은 다음과 같다. 첫 째, Alabama 주와 Massachusetts 주 모두에서 제안된 HB 방법이 최적이다. 기대한 바와 같이 CPS 방법과 비교하여 HB방법이 Massachusetts 주에서 보다 Alabama 주에서 더 효과적이다. 왜냐하면 Massachusetts 주에서 보다 Alabama 주에서 더 적은 수의 CPS 표본 가구들이 가능하였기 때문이다. 제안한 HB 추정값의 사후표준편차는 CPS 추정값의 표준오차보다 모든 시간과 모든 주에 걸쳐서 상당히 작았다. 그러나 Fay-Herriot HB 추정값은 CPS 추정값보다 항상 월등하게 좋은 것은 아니었다. 사실 Fay-Herriot 모형 추정값의 감소는 모든 direct-use state에 대해서는 10% 보다 작았으며 어떤 월(예컨대, North Carolina 주 1988년 3월)에서는 0.1%만큼 작았다. 대부분 indirect-use state에 대해서 Fay-Herriot 모형에 대한 상당한 크

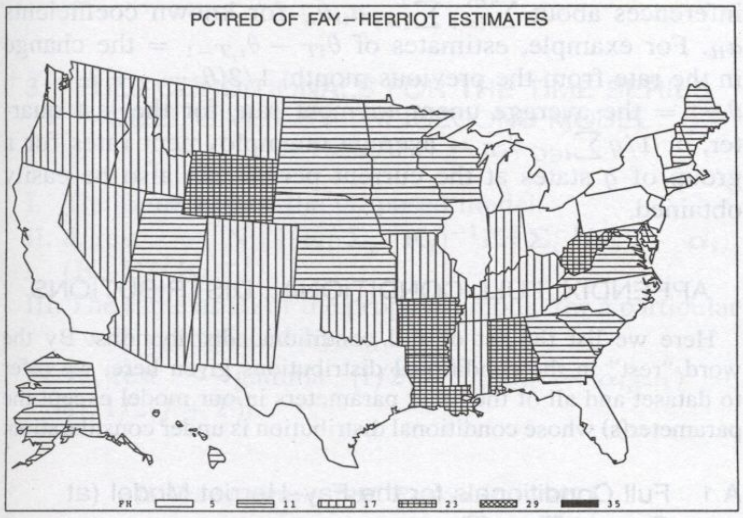

Figure 3. Statewise Percent Reductions of Standard Errors for Fay–Herriot HB Estimates over the CPS Standard Errors of Unemployment Rates (Excluding New York). The numbers on the right denote midpoints of intervals of length 6; for example, 5 means 2–8%, and so on.

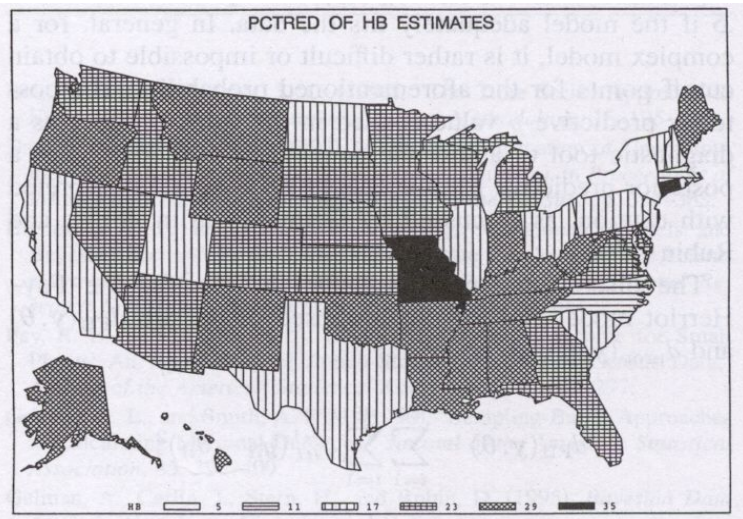

Figure 4. Statewise Percent Reductions of Standard Errors for Cross-Sectional and Time Series HB Estimates over the CPS Standard Errors of Unemployment Rates (Excluding New York). The numbers on the right denote midpoints of intervals of length 6; for example, 5 means 2–8%, and so on.

기의 감소가 있지만 감소는 2%(예컨대, New Hampshire 주 1986년 6월)만큼 작았다. 전반적으로 제안된 HB 방법이 CPS 추정값을 상당히 개선한다.

Figure 3과 Figure 4는 1988년 12월에 New York 주를 제외한 49개 주와 컬럼비아 특별구에 대해 두 HB 방법의 표준오차와 CPS를 비교하기 위해 만든 지도이다. 이 지도는 CPS 추정값에 대한 HB 추정값의 주별 퍼센트 감소를 보여준다. 그림에서 모든 주에 대해 HB 방법이 CPS 보다 훨씬 좋은 것이 분명하다. 제안한 HB 모형이 모든 50개 주에 대해 Fay-Herriot 모형보다 좋다.

4.6 모형적합과 모형비교

사후예측평가(posterior predictive assessment) 접근(Gelman, Meng과 Stern, 1996; Sinha와 Dey, 1997)과 사후예측발산(posterior predictive divergence) 접근(Laud와 Ibrahim, 1995)에 근거하여 제안한 모형과 Fay-Herriot 모형을 비교한다. 앞에서 언급한 지역에 관한 횡단면 정보를 결합하지 아니하는 시계열모형은 18개의 초모수를 추정하기 위해 48개의 관측값만을 가지고 계산을 수행할 수 없다. 그러나 단지 비교를 위하여 초모수(예컨대, 월과 년 효과 제거)의 수를 줄여서 시계열모형과 다른 두 모형을 비교한다.

4.6.1 사후예측평가 접근

이 접근방법에서 적절한 불일치측도의 모의실험 값이 사후예측분포로부터 생성되며 이를 관측된 자료에 대한 불일치측도 값과 비교한다. Sinha와 Dey (1997)에 따라 $\mathbf{y}_{obs}$와 $\mathbf{y}_{new}$를 관측된 자료와 모의실험으로 생성된 자료라 하자. $f(\Theta \mid \mathbf{y}_{obs})$, $f(d(\mathbf{y}_{obs}, \Theta) \mid \mathbf{y}_{obs})$, 그리고 $f(d(\mathbf{y}_{new}, \Theta) \mid \mathbf{y}_{obs})$을 Θ의 사후(예측)분포라 하고, $d(\mathbf{y}, \Theta)$를 우리의 모형에 대한 적당한 불일치측도

(나중에 정의함)라 하자. 불일치측도는 자료에 따라 $d(\mathbf{y}_{obs}, \Theta)$와 $d(\mathbf{y}_{new}, \Theta)$가 있다.

사후예측평가는 깁스 샘플링을 사용하여 근사적 방법으로 계산한다. 깁스 출력을 사용하여 $f(\Theta \mid \mathbf{y}_{obs})$로부터 $\Theta^{(l)}$을 생성하고 $f(\mathbf{y} \mid \Theta^{(l)})$로부터 $\mathbf{y}^{(l)}$을 생성한 후, $d(\mathbf{y}_{obs}, \Theta^{(l)})$과 $d(\mathbf{y}^{(l)}, \Theta^{(l)})$ ($l = 1, \cdots, B$)을 계산한다. 여기서 B는 Θ값의 깁스 반복의 총수이다.

따라서 $\{d(\mathbf{y}_{obs}, \Theta^{(l)}), 1 \le l \le B\}$와 $\{d(\mathbf{y}^{(l)}, \Theta^{(l)}), 1 \le l \le B\}$는 $f(d(\mathbf{y}_{obs}, \Theta) \mid \mathbf{y}_{obs})$와 $f(d(\mathbf{y}_{new}, \Theta) \mid \mathbf{y}_{obs})$으로부터의 표본을 나타낸다. 이 생성된 표본을 사용하여 $P\{d(\mathbf{y}_{new}, \Theta) \ge d(\mathbf{y}_{obs}, \Theta) \mid \mathbf{y}_{obs}\}$에 대한 근사값을 다음과 같이 구할 수 있다.

$$B^{-1} \sum_{l=1}^{B} I\{d(\mathbf{y}^{(l)}, \Theta^{(l)}) \ge d(\mathbf{y}_{obs}, \Theta^{(l)})\}$$

여기서 $I(\cdot)$은 지시함수이다. 확률 $P\{d(\mathbf{y}_{new}, \Theta) \ge d(\mathbf{y}_{obs}, \Theta) \mid \mathbf{y}_{obs}\}$의 극단적인(0이나 1에 가까운) 값은 주어진 모형의 적합성결여를 나타낸다. 반면에 모형이 적합하면 이 확률값이 0.5에 가깝다.

주어진 모형이 현재의 자료에 잘 적합하는지를 조사하기 위해서는 주어진 모형에 근거한 새로운 자료를 생성했을 때, 만약 모형이 관측된 자료 $\mathbf{y}_{obs}$를 충분히 잘 적합하면 생성된 새로운 자료 $\mathbf{y}_{new}$도 관측된 자료와 비슷해야 한다. 이 유사성을 계량화하기 위해 불일치측도에 의한 $\mathbf{y}_{obs}$와 $\mathbf{y}_{new}$를 비교하기 위해서는 적절한 불일치측도 $d(\mathbf{y}, \Theta)$가 사용된다.

만약 모형이 관측된 자료를 적합하면 불일치측도의 두 값이 비슷하다. 즉, 만

약 주어진 모형이 관측된 자료를 충분히 잘 적합하면, $d(\mathbf{y}_{obs}, \Theta)$는 $\mathbf{y}_{new}$가 사후예측분포에 근거하여 반복적으로 생성되었을 때 만들어지는 $d(\mathbf{y}_{new}, \Theta)$의 히스토그램의 중심부에 가까워야 한다. 따라서 만약 주어진 모형이 관측된 자료를 충분히 잘 적합하면 사후예측 p값, $P\{d(\mathbf{y}_{new}, \Theta) \geq d(\mathbf{y}_{obs}, \Theta) \mid \mathbf{y}_{obs}\}$이 거의 0.5에 가까운 값이 예상된다. 일반적으로, 복잡한 모형에 대해서 사후예측 p값에 대한 경계값을 구하기가 다소 어렵거나 불가능할 수 있다. 따라서 주어진 모형의 적합을 평가하는 진단도구로서 탐색적 자료분석(EDA)의 관점에서 사후예측 p값이 사용된다. 사후예측 p값은 Gelman, Carlin, Stern과 Rubin(1995)에서 지적한 대로 주의해서 해석해야 한다.

Fay-Herriot 모형과 제안한 모형에 대해 사용하는 불일치측도는 각각 다음과 같다.

$$d_{FH}(\mathbf{y}, \Theta) = \sum_{i=1}^{50} \sum_{t=1}^{48} \sigma_{itt}^{-1}(y_{it} - \Theta_{it})^2$$

$$d_{prop}(\mathbf{y}, \Theta) = \sum_{i=1}^{50} (\mathrm{y}_i - \Theta_i)' \Sigma_i^{-1} (\mathrm{y}_i - \Theta_i)$$

Fay-Herriot 모형 하에서 사후예측 p값의 추정된 값 0.999는 모형의 적합결여를 나타내는 반면 제안한 모형에 대한 값 0.614는 모형이 적합함을 강하게 암시한다.

월과 년 효과를 제거할 때 시계열모형과 제안한 모형에 대한 사후예측 p값은 0.802와 0.758이다. 변형된 두 모형이 모두 불만족스럽지만 시계열모형이 더 불만족스럽다고 말할 수 있다.

4.6.2 사후예측발산 접근에 의한 모형비교

사후예측분포 $f(\boldsymbol{y}_{new}, \Theta \mid \boldsymbol{y}_{obs})$으로부터 생성된 자료 벡터 $\boldsymbol{y}_{new}$를 사용하여

$$d(\boldsymbol{y}_{new}, \boldsymbol{y}_{obs}) = E\{n^{-1}\|\boldsymbol{y}_{new} - \boldsymbol{y}_{obs}\|^2 \mid \boldsymbol{y}_{obs}\}$$

(Laud와 Ibrahim(1995)의 발산측도)를 $(nB)^{-1}\sum_{l=1}^{B}\|\boldsymbol{y}^{(l)} - \boldsymbol{y}_{obs}\|^2$의해 근사한다. 여기서 n은 $\boldsymbol{y}_{obs}$의 차원이다. 두 개의 모형 중에서 이 발산측도의 값이 작은 모형을 선택한다. 우리의 자료분석에서 발산측도 값이 Fay-Herriot 모형에서는 1.6256이고 제안한 모형에서는 0.0007이므로 제안한 모형이 훨씬 좋다. 월과 년의 효과를 제거했을 때 제안한 모형과 시계열모형에 대한 발산측도 값은 각각 0.0008과 1.4243이다.

4.7 맺는말

CPS에 대한 HB 방법의 유효성을 평가하기 위해 세 모형을 고려하였다. 이 중 횡단면과 시계열자료를 결합하는 모형이 최적이다. 우리의 분석에 의하면 개별 주에 대한 시계열 모형은 자료에 비해 모수가 너무 많이 존재하기 때문에 적합되지 않는다. 따라서 주별 실업률의 개선된 추정값을 개발하기 위해 횡단면과 시계열자료를 결합하는 것이 중요하다. 현재시점 T에서 i번째 주의 실업률의 참값인 Θ_{iT}의 추정을 고려한다. 그러나 방법은 계수 a_{it}를 알 때 $\sum_{i=1}^{m}\sum_{t=1}^{T} a_{it}\Theta_{it}$에 대한 추론 문제에도 적용이 가능하다. 예를 들면, $\Theta_{iT} - \Theta_{i,T-1}$ = 전월로부터의 실업률 변화, $1/3(\Theta_{i,T-2} + \Theta_{i,T-1} + \Theta_{iT})$ = 마지막 분기에 대한 평균실업률, 혹은 $1/g\sum_{i\in G}\Theta_{iT}$ = 현재시점에서 g개의 주 그룹에 대한 평균실업률 등을 쉽게 구할 수 있다.

5 캐나다의 시계열모형을 사용한 실업률 추정

5.1 배경

캐나다에서 주(Province)별 실업률과 국가 전체 실업률은 대부분 언론의 주목을 받지만 주보다 작은 단위에 대한 실업률 통계도 매우 중요하다. 이러한 실업률 통계는 EI(Employment Insurance)에 의해 사용되어 프로그램을 운영하는 규칙을 결정한다. 뿐만 아니라, CMA(Census Metropolitan Area; 인구 100,000명 이상의 도시)와 CA(Census Agglomeration; 다른 도시 중심지)에 대한 실업률은 지역 수준에서의 관심의 대상이다. 그러나 많은 CA들이 적절한 직접추정량을 생산하기 위한 충분한 수의 표본을 가지지 못한다. 목표는 주어진 달에 주어진 CMA나 CA에서의 표본만을 사용하여 구한 직접추정량을 개선하기 위해 모형에 근거한 추정량을 구하는 것이다. 편의상 CMA와 CA를 구별하지 아니하고 모두 CA라 하자.

캐나다에서 실업률은 캐나다 노동력조사(Labor Force Survey; LFS)를 통해 작성된다. LFS는 층화 다단계 설계를 사용하여 추출된 53,000 조사가구에 대한 월별 조사이다. 매월 1/6의 표본이 새로운 표본으로 교체되며, 두 연속 월에서는 표본의 5/6이 공통이다. 이 중복된 표본으로 인해 상관이 생기게 되므로 지역뿐만 아니라 시간으로부터도 정보를 빌려오는(borrow strength) 분석방법을 사용하면 소지역에서 바람직한 성질을 가지는 개선된 추정값을 생산할 수 있다. 따라서 공간과 시간에서 동시에 정보를 빌려오는 소지역추정 방법을 고려하고자 한다. 지금까지 이러한 접근 방법에 의해 소지역 추정값을 계산한 연구로는 Rao와 Yu(1994), Ghosh, Nangia와 Kim(1996), Datta, Maiti와 Lu(1999) 등이 있다.

여기서는 캐나다 노동력조사 자료에 베이지안 시계열 모형을 생각한다. Datta, Maiti와 Lu(1999)와는 달리 소지역에 교차된 짧은 시계열 자료를 사용한다. 따라서 모형에서 계절적 모수는 포함하지 아니한다. 이로 인해 추정해야할

모수의 수는 상당히 감소된다. 이러한 단순화에도 불구하고 충분히 좋은 모형 적합을 얻으며, 실업률의 소지역 추정값의 변동계수(CV)에 상당한 감소를 가져온다. 변동계수의 감소는 직접 LFS 추정량의 표본 공분산 행렬의 평활추정값(smoothed estimates)을 구하기 위해 평활 변동계수(smoothed CV)와 시차상관(lag correlation)을 사용하는 공분산 행렬의 계산 방법에 부분적으로 기인한다.

5.2 횡단면 및 시계열 모형

y_{it}를 시점 t에서 i번째 CA(소지역)의 실업률의 참값 Θ_{it}의 직접 LFS 추정값이라 하자. ($i = 1, \cdots, m$, $t = 1, \cdots, T$). 여기서 m은 CA의 총 개수이고, T는 관심있는 현 시점이다.

$$y_{it} = \Theta_{it} + e_{it}, \qquad i = 1, \cdots, m \, , \; t = 1, \cdots, T$$

여기서 e_{it}는 표본 오차이다. $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iT})'$, $\boldsymbol{\Theta}_i = (\Theta_{i1}, \cdots, \Theta_{iT})'$, 그리고 $\boldsymbol{e}_i = (e_{i1}, \cdots, e_{iT})'$로 표현하면 $\boldsymbol{e}_i$는 i번째 CA에 대한 표본 오차 벡터이다. LFS 설계에서 CA는 층(strata)으로 간주된다. 따라서 표본 벡터 $\boldsymbol{e}_i$는 CA간에는 상관이 없다. 그러나 LFS의 표본연동 형태로 인해 소지역 내 짧은 시간에 걸친 상당한 표본의 중복이 있으므로 e_{it}와 $e_{is}(t \neq s)$ 사이의 상관은 반드시 고려되어야 한다. $\boldsymbol{e}_i$는 평균벡터가 $\boldsymbol{0}$이고 공분산행렬이 $\boldsymbol{\Sigma}_i$인 다변량 정규분포를 따른다고 가정하자. 즉, $\boldsymbol{e}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$ 이다. 위의 모형을 벡터를 사용하여 표현하면 다음과 같다.

$$\boldsymbol{y}_i \sim N_T(\boldsymbol{\Theta}_i, \boldsymbol{\Sigma}_i), \qquad i = 1, \cdots, m$$

따라서 $\boldsymbol{y}_i$는 $\boldsymbol{\Theta}_i$에 대해 설계불편(design-unbiased)이다. 분산공분산행렬 $\boldsymbol{\Sigma}_i$

는 안다고 가정하자. 정규성과 Σ_i를 안다는 가정은 모형에 근거한 소지역추정에서의 일종의 관습이다. (Fay와 Herriot, 1979; Ghosh와 Rao, 1994; Datta et al., 1999; Rao, 1999 참조).

그러나 실제에 있어서 Σ_i를 정하는 것이 쉽지 않다. 모형에서 Σ_i의 평활추정량을 사용하고, 이를 Σ_i의 참값으로 간주한다. LFS에서 Σ_i의 평활추정량을 구성하는 자세한 절차는 (4)에서 설명한다. 캐나다 노동력조사 같은 연동교체설계(rotating-panel design)에서 표본오차의 자기상관을 추정하는 간단한 방법은 Pfeffermann, Feder와 Signorelli(1998)에 소개되어 있다. 주어진 상황에서 이 방법의 사용 가능성을 조사하는 것도 유용할 것이다.

지역과 시간으로부터 동시에 정보를 빌려오기 위해 보조변수 $\boldsymbol{x}_{it}$와 랜덤 지역효과를 가지는 선형회귀모형을 사용하여 실업률의 참값인 Θ_{it}를 모형화 한다.

$$\Theta_{it} = \boldsymbol{x}_{it}'\beta + \nu_i + u_{it}, \qquad i = 1, \cdots, m,\ t = 1, \cdots, T$$

여기서 $\boldsymbol{x}_{it} = (x_{it1}, \cdots, x_{itp})'$는 시간 t에서 i번째 CA에 대한 지역 수준의 보조자료의 벡터이고, β는 p개의 회귀계수 벡터이고, ν_i는 랜덤 지역효과로서 $\nu_i \overset{iid}{\sim} N(0, \sigma_\nu^2)$ 이며, u_{it}는 랜덤 시간 성분이다.

주어진 지역 i에 대해 Datta et al.(1999)은 u_{it}가 시간 $t = 1, \cdots, T$에 대해 확률보행과정(random walk process)를 따른다고 가정하였다. 즉,

$$u_{it} = u_{i,t-1} + \varepsilon_{it}, \qquad i = 1, \cdots, m,\ t = 2, \cdots, T$$

여기서 $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$ 이다. 이 때 $cov(u_{it}, u_{is}) = \min(t, s)\,\sigma_\varepsilon^2$ 이다. 또한, $\{\nu_i\}$, $\{\varepsilon_{it}\}$ 그리고 $\{e_i\}$는 상호독립이라고 가정한다. 모형에서 회귀계수 β와

분산성분 σ_v^2와 σ_ε^2는 미지의 모수이다.

Rao와 Yu(1994)는 u_{it}에 대해 정상자기회귀모형(stationary autoregressive model), AR(1)을 사용하였다. 즉, $u_{it} = \rho u_{i,t-1} + \varepsilon_{it}$, ($|\rho| < 1$). Datta et al.(1999)은 분석에서 긴 시계열($T=48$ 개월)을 사용하여 Θ_{it}에 대해 계절적 효과로서 월 및 년 효과를 포함하였다. 그러나, 여기서는 캐나다 LFS 설계의 6개월 연동주기에 근거하여 평활을 위해 6개월의 자료만을 사용하였다. 따라서 Θ_{it}에 관한 모형은 Datta et al.(1999)의 모형보다 간단하다. 이 단순화가 평활 공분산행렬 $\boldsymbol{\Sigma}_i$의 불안정성을 감소시키는 듯 하다.

자료 $\{y_{it}\}$를 $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iT})'$를 사용하여 벡터 $\boldsymbol{y} = (\boldsymbol{y}_1', \cdots, \boldsymbol{y}_m')'$로 표현하여 위의 모형을 행렬형태로 표현하면 다음과 같다.

$$\boldsymbol{y}_i = \boldsymbol{X}_i\beta + \mathbf{1}_T v_i + \boldsymbol{u}_i + \boldsymbol{e}_i, \qquad i = 1, \cdots, m$$

여기서 $\boldsymbol{X}_i' = (\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{iT})$, $\boldsymbol{u}_i' = (u_{i1}, \cdots, u_{iT})$, 그리고 $\mathbf{1}_T$는 1로 이루어진 $T \times 1$ 벡터이다. 이 모형은 일반화선형혼합모형의 특별한 경우이며, 지역과 시간으로부터 동시에 정보를 빌려오는 잘 알려진 Fay-Harriot 모형의 확장이다.

자료분석에서 비교를 위해 시점 $t = 1, \cdots, T$에 대한 Fay-Herriot 모형을 고려하자. 시간 t에서 이 모형은 다음과 같이 주어진다.

$$y_{it} = \Theta_{it} + e_{it}, \qquad i = 1, \cdots, m$$

$$\Theta_{it} = \boldsymbol{x}_{it}'\beta_t + v_{it}, \qquad i = 1, \cdots, m$$

여기서 시점 t에서 표본오차는 $e_{it} \overset{ind}{\sim} N(0, \sigma_{it}^2)$ 이며, 지역 랜덤효과 v_{it}와 $v_{it'}$, $t' \neq t$는 서로 독립이고 $v_{it} \overset{ind}{\sim} N(0, \sigma_{vt}^2)$ 이다. 그리고 표본분산 σ_{it}^2은 안

다고(평활추정값) 가정하나 σ_{vt}^2는 모르는 모수이다. Fay-Herriot 모형은 Θ_{it}를 추정할 때 각 시점 t에서의 횡단면 정보만을 결합하며, 과거 시점으로부터 정보를 빌려오지는 아니한다.

여기서 우리의 관심사는 Θ_{iT}(현재시점 $t=T$에서의 Θ_{it})의 모형기반 추정량을 구하는 것이다. Datta, Lahiri와 Maiti(2002) 그리고 You(1999)는 EBLUP 방법으로 Θ_{iT}의 이 단계 추정량과 추정량의 MSE 근사를 구하였다. 여기서는 깁스 샘플링 을 사용한 계층적 베이즈(HB) 방법으로 u_{it}에 대한 AR(1)과 확률보행과정을 연구한다.

5.3 계층적 베이즈 분석

앞에서 제안한 횡단면 및 시계열 모형과 Fay-Herriot 모형에 대해 계층적 베이즈 방법을 적용하고자 한다. 깁스 샘플링 방법을 사용하여 소지역평균 Θ_{iT}의 사후평균과 사후분산의 추정값을 구한다.

5.3.1 계층적 베이즈 모형

이제 횡단면 및 시계열 모형을 계층적 베이즈 구조로 표현하면 다음과 같다.

- 모수 $\boldsymbol{\Theta}_i = (\Theta_{i1}, \cdots, \Theta_{iT})'$가 주어졌을 때, $[\boldsymbol{y}_i | \boldsymbol{\Theta}_i] \overset{ind}{\sim} N(\boldsymbol{\Theta}_i, \boldsymbol{\Sigma}_i)$;
- 모수 $\boldsymbol{\beta}$, u_{it} 그리고 σ_v^2가 주어졌을 때,

  $[\Theta_{it} | \boldsymbol{\beta}, u_{it}, \sigma_v^2] \overset{ind}{\sim} N(\boldsymbol{x}_{it}'\boldsymbol{\beta} + \rho u_{it}, \sigma_v^2)$;
- 모수 $u_{i,t-1}$와 σ_ε^2가 주어졌을 때, $[u_{it} | u_{i,t-1}, \sigma_\varepsilon^2] \overset{ind}{\sim} N(\rho u_{i,t-1}, \sigma_\varepsilon^2)$;

여기서 주변확률분포의 관점에서 β, σ_v^2, σ_ε^2는 서로 독립이며 사전분포 $\pi(\beta) \propto 1$, $\sigma_v^2 \sim IG(a_1, b_1)$, $\sigma_\varepsilon^2 \sim IG(a_2, b_2)$를 따른다. IG는 역감마분포를 나타내며, a_1, b_1, a_2, b_2는 알려진 양의 상수로서 σ_v^2와 σ_ε^2에 대한 막연한 사전정보를 나타내기 위해 보통 매우 작은 값으로 정한다. 확률보행과정에서는 $\rho = 1$으로 두고, AR(1) 모형에 대해 ρ는 $|\rho| < 1$의 알려진 값으로 가정한다.

우리의 관심사는 현재 실업률 Θ_{iT}를 추정하는 것이다. 계층적 베이지안 분석에서 Θ_{iT}는 사후평균 $E(\Theta_{iT}|\boldsymbol{y})$에 의해 추정되며 추정에 관련된 불확실성은 사후분산 $V(\Theta_{iT}|\boldsymbol{y})$에 의해 측정된다. 깁스 샘플링을 사용하여 사후평균과 사후분산을 구할 수 있다. (Gelfand와 Smith, 1990; Gelman과 Rubin, 1992 참조).

한편, Fay-Herriot 모형을 계층적 베이즈 구조로 표현하면 다음과 같다.

- 모수 Θ_{it}가 주어졌을 때, $[y_{it}|\Theta_{it}] \overset{ind}{\sim} N(\Theta_{it}, \sigma_{it}^2)$;
- 모수 β_t와 σ_{vt}^2가 주어졌을 때, $[\Theta_{it}|\beta_t, \sigma_{vt}^2] \overset{ind}{\sim} N(\boldsymbol{x}_{it}'\beta_t, \sigma_{vt}^2)$;

여기서 주변확률분포의 관점에서 β_t와 σ_{vt}^2는 서로 독립이며 사전분포 $\pi(\beta) \propto 1$, $\sigma_{vt}^2 \sim IG(a_t, b_t)$를 따른다.

5.3.2 깁스 샘플링 방법

깁스 샘플링 방법은 반복적 마코브 연쇄 몬테칼로(Markov chain Monte Carlo) 표본기법으로서 저차원의 밀도함수로부터 표본을 생성함으로 결국 확률변수의 결합확률분포로부터 표본을 생성하여 결합확률분포와 주변확률분포에 대한 통계적 추론을 수행한다. 이 방법의 가장 뛰어난 응용이 베이지안 구조 하에

서의 추론이다.

베이지안 추론에서 관심사는 관심모수의 사후분포이다. $y_i|\Theta$가 조건부밀도함수 $f(y_i|\Theta)$ $(i=1,\cdots,n)$를 가지며, $\Theta=(\Theta_1,\cdots,\Theta_k)'$에 대한 사전정보가 사전분포 $\pi(\Theta)$에 의해 요약된다고 가정하자. $\pi(\Theta|\boldsymbol{y})$를 자료 $\boldsymbol{y}=(y_1,\cdots,y_n)'$가 주어졌을 때 Θ의 사후분포를 나타낸다고 하자. 실제 문제에서는 Θ에 관한 고차원의 적분문제 때문에 $\pi(\Theta|\boldsymbol{y})$에서 직접 표본을 생성하기 어렵다. 그러나 깁스 표본자를 사용하면 극한분포로서 $\pi(\Theta|\boldsymbol{y})$를 가지는 마코브 연쇄 $\{\Theta^{(g)}=(\Theta_1^{(g)},\cdots,\Theta_k^{(g)})'\}$를 구성할 수 있다.

예를 들어, $\Theta=(\Theta_1,\Theta_2)'$를 고려하자. 초기값 $\Theta^{(0)}=(\Theta_1^{(0)},\Theta_2^{(0)})'$를 사용하여 $\pi(\Theta_1|\Theta_2^{(0)},\boldsymbol{y})$로부터 $\Theta_1^{(1)}$를 추출하고 그리고 $\pi(\Theta_2|\Theta_1^{(1)},\boldsymbol{y})$로부터 $\Theta_2^{(1)}$를 추출하여 $\Theta^{(1)}=(\Theta_1^{(1)},\Theta_2^{(1)})'$를 생성한다. $\Theta^{(1)}=(\Theta_1^{(1)},\Theta_2^{(1)})'$를 이용하여 $\Theta^{(2)}$를 생성한다면, 어떤 정칙조건 하에서 $\Theta^{(g)}=(\Theta_1^{(g)},\Theta_2^{(g)})'$는 $g\to\infty$일 때 $\pi(\Theta|\boldsymbol{y})$로 수렴한다. 주변사후분포 $\pi(\Theta_i|\boldsymbol{y})$에 대한 추론은 g가 클 때 주변표본 $\{\Theta_i^{(g+k)};\ k=1,2,\cdots\}$에 기초한다.

앞에서 주어진 계층적 베이즈 모형에서 깁스 표본자를 실행하기 위해서는 모수 β, σ_v^2, σ_ε^2, u_{it} 그리고 Θ_i에 관한 모든 조건부 분포로부터 표본을 생성할 필요가 있다.

먼저 횡단면 및 시계열 모형에 대한 조건부 분포는 다음과 같다. 여기서 행렬 표현을 위해 기호 $\boldsymbol{X}=(X_1',\cdots,X_m')$, $\Theta=(\Theta_1',\cdots,\Theta_m')$, $\boldsymbol{u}=(\boldsymbol{u}_1',\cdots,\boldsymbol{u}_m')'$, $\Theta_i'=(\Theta_{i1},\cdots,\Theta_{iT})$, $\boldsymbol{u}_i'=(u_{i1},\cdots,u_{iT})$를 사용한다.

(i) $\beta | \boldsymbol{y}, \sigma_v^2, \sigma_\varepsilon^2, \boldsymbol{u}, \Theta \sim N[(\boldsymbol{X}'\boldsymbol{X})^{-1}(\Theta - \boldsymbol{u}), \sigma_v^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]$;

(ii) $\sigma_v^2 | \boldsymbol{y}, \beta, \sigma_\varepsilon^2, \boldsymbol{u}, \Theta \sim IG(a_1 + mT/2, b_1 + \sum_{i=1}^{m}\sum_{t=1}^{T}(\Theta_{it} - \boldsymbol{x}_{it}'\beta - u_{it})^2/2)$;

(iii) $\sigma_\varepsilon^2 | \boldsymbol{y}, \beta, \sigma_v^2, \boldsymbol{u}, \Theta \sim IG(a_1 + m(T-1)/2, b_2 + \sum_{i=1}^{m}\sum_{t=2}^{T}(u_{it} - \rho u_{i,t-1})^2/2)$;

(iv) $u_{i1} | \boldsymbol{y}, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i2}, \Theta$ ($i = 1, \cdots, m$)

$$\sim N\left[\left(\frac{1}{\sigma_v^2} + \frac{\rho^2}{\sigma_\varepsilon^2}\right)^{-1}\left(\frac{\Theta_{i1} - \boldsymbol{x}_{i1}'\beta}{\sigma_v^2} + \frac{\rho^2 u_{i2}}{\sigma_\varepsilon^2}\right), \left(\frac{1}{\sigma_v^2} + \frac{\rho^2}{\sigma_\varepsilon^2}\right)^{-1}\right];$$

(v) $u_{i1} | \boldsymbol{y}, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i,t-1} u_{i,t+1}, \Theta$ ($i = 1, \cdots, m;\ 2 \le t \le T-1$)

$$\sim N\left[\left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1}\left(\frac{\Theta_{iT} - \boldsymbol{x}_i'\beta}{\sigma_v^2} + \frac{\rho u_{i,T-1}}{\sigma_\varepsilon^2}\right), \left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1}\right];$$

(vi) $\Theta_i | \boldsymbol{y}, \beta, \sigma_v^2, \sigma_\varepsilon^2, \boldsymbol{u}$ ($i = 1, \cdots, m$)

$$\sim N\left[(\sigma_v^2 \boldsymbol{I}_T + \Sigma_i^{-1})^{-1}(\Sigma_i^{-1}\boldsymbol{y}_i + \sigma_v^2(\boldsymbol{X}_i\beta + \boldsymbol{u}_i)), (\sigma_\varepsilon^2 \boldsymbol{I}_T + \Sigma_i^{-1})^{-1}\right]$$

한편, Fay-Herriot 모형에 대해 시점 t에서의 조건부 분포는 다음과 같다. 여기서 행렬표현을 위해 $t = 1, \cdots, T$에 대해 $\boldsymbol{y}_t = (y_{1t}, \cdots, y_{mt})'$, $\boldsymbol{X}_t = (x_{1t}, \cdots, x_{mt})$, $\Theta_t = (\Theta_{1t}, \cdots, \Theta_{mt})'$의 기호를 사용하자.

(i) $\beta_t | \boldsymbol{y}_t, \sigma_{vt}^2, \Theta_t \sim N[(\boldsymbol{X}_t'\boldsymbol{X}_t)^{-1}\boldsymbol{X}_t'\Theta_t, \sigma_{vt}^2(\boldsymbol{X}_t'\boldsymbol{X}_t)^{-1}]$;

(ii) $\sigma_{vt}^2 | \boldsymbol{y}_t, \beta_t, \sigma_\varepsilon^2, \boldsymbol{u}, \Theta \sim IG(a_1 + m/2, b_1 + \sum_{i=1}^{m}(\Theta_{it} - \boldsymbol{x}_{it}'\beta_t)^2/2)$;

(iii) $\Theta_{it} | \boldsymbol{y}_t, \beta_t, \sigma_{vt}^2 \sim N[(1 - r_{it})y_{it} + r_{it}\boldsymbol{x}_{i1}'\beta_t, \sigma_{it}^2(1 - r_{it})]$ ($i = 1, \cdots, m$)

여기서 $r_{it} = \sigma_{itl}^2/(\sigma_{it}^2 + \sigma_{vt}^2)$ 이다.

위에서 모든 조건부분포는 쉽게 표본을 생성할 수 있는 정규분포와 역감마분포 등 표준분포이다.

5.3.3 사후추정

깁스 샘플링을 수행하기 위해 Gelman과 Rubin(1992)의 제안에 따라 $2d$개의 길이를 가지는 $L(L>2)$개의 독립적인 병렬체인을 고려한다. 각 체인에서 첫 d개의 반복은 초기값의 영향을 제거하기 위해 버린다. 첫 d개의 반복 후 생성된 모든 표본을 사용하여 사후평균과 사후분산을 계산할 뿐만 아니라 깁스 표본자의 수렴성을 확인한다.

이 때 가능하면 Rao-Blackwellization을 사용하여 사후평균과 사후분산의 추정량을 계산한다. 왜냐하면 시뮬레이션으로 생성한 표본으로 단순히 추정값을 계산하는 것 보다 Rao-Blackwellization을 이용하면 시뮬레이션 오차를 상당히 줄일 수 있기 때문이다. (Gelfand와 Smith, 1991; You와 Rao, 2000 참조). 횡단면 및 시계열 모형 하에서 사후평균 $E(\theta_i|\boldsymbol{y})$와 사후분산 $V(\theta_i|\boldsymbol{y})$의 Rao-Blackwellized 추정값을 다음과 같이 구할 수 있다.

$$\widehat{E}(\theta_i|\boldsymbol{y}) = \sum_{l=1}^{L}\sum_{k=d+1}^{2d}[(\sigma_v^{-2(lk)}\boldsymbol{I}_T + \boldsymbol{\Sigma}_i^{-1}) \times(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i + \sigma_v^{-2(lk)}(\boldsymbol{X}_i\beta^{(lk)} + \boldsymbol{u}_i^{(lk)}))]/(Ld)$$

$$\widehat{V}(\theta_i|\boldsymbol{y}) = \sum_{l=1}^{L}\sum_{k=d+1}^{2d}(\sigma_v^{-2(lk)}\boldsymbol{I}_T + \boldsymbol{\Sigma}_i^{-1})/(Ld)$$
$$+ \sum_{l=1}^{L}\sum_{k=d+1}^{2d}[(\sigma_v^{-2(lk)}\boldsymbol{I}_T + \boldsymbol{\Sigma}_i^{-1})^{-1}\times(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i + \sigma_v^{-2(lk)}(\boldsymbol{X}_i\beta^{(lk)} + \boldsymbol{u}_i^{(lk)}))]$$
$$\times[(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i + \sigma_v^{-2(lk)}(\boldsymbol{X}_i\beta^{(lk)} + \boldsymbol{u}_i^{(lk)}))'\times(\sigma_v^{-2(lk)}\boldsymbol{I}_T + \boldsymbol{\Sigma}_i^{-1})^{-1}]'/(Ld)$$

$$+ \sum_{l=1}^{L}\sum_{k=d+1}^{2d}[(\sigma_v^{-2(lk)}\boldsymbol{I}_T + \boldsymbol{\Sigma}_i^{-1})^{-1}\times(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i + \sigma_v^{-2(lk)}(\boldsymbol{X}_i\beta^{(lk)} + \boldsymbol{u}_i^{(lk)}))]$$
$$\times[(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i + \sigma_v^{-2(lk)}(\boldsymbol{X}_i\beta^{(lk)} + \boldsymbol{u}_i^{(lk)}))'\times(\sigma_v^{-2(lk)}\boldsymbol{I}_T + \boldsymbol{\Sigma}_i^{-1})^{-1}]'/(Ld)$$
$$- [\sum_{l=1}^{L}\sum_{k=d+1}^{2d}(\sigma_v^{-2(lk)}\boldsymbol{I}_T + \boldsymbol{\Sigma}_i^{-1})^{-1}\times(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i + \sigma_v^{-2(lk)}(\boldsymbol{X}_i\beta^{(lk)} + \boldsymbol{u}_i^{(lk)}))]$$
$$\times [\sum_{l=1}^{L}\sum_{k=d+1}^{2d}(\sigma_v^{-2(lk)}\boldsymbol{I}_T + \boldsymbol{\Sigma}_i^{-1})^{-1}$$
$$\times (\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i + \sigma_v^{-2(lk)}(\boldsymbol{X}_i\beta^{(lk)} + \boldsymbol{u}_i^{(lk)}))]'/(Ld)^2$$

여기서 $\{\beta^{(lk)}, \sigma_v^{2(lk)}, \boldsymbol{u}_i^{(lk)}; k = d+1, \cdots, 2d,\ l = 1, \cdots, L\}$ 는 깁스 표본자로부터 생성된 표본이고 $\boldsymbol{I}_T$는 T차 항등행렬이다. 따라서 깁스 샘플링을 사용하여 각 지역에 대해 현재 시점의 소지역 평균 Θ_{iT}와 과거시점 $t = 1, \cdots, T-1$에서의 소지역 평균 Θ_{it}를 동시에 추정할 수 있다. 사후공분산행렬의 추정값 $\widehat{V}(\Theta_i|\boldsymbol{y})$을 계산하면 $t \neq s = 1, \cdots, T$에 대한 Θ_{it}와 Θ_{is}의 사후공분산의 추정값을 구할 수 있다.

Fay-Herriot 모형 하에서 $\boldsymbol{y}_T = (y_{1T}, \cdots, y_{mT})'$ 가 현 시점에서의 횡단면 자료를 나타낸다고 하면, $E(\Theta_{iT}|\boldsymbol{y}_T)$와 $V(\Theta_{iT}|\boldsymbol{y}_T)$의 Rao-Blackwellized 추정값을 다음과 같이 구할 수 있다.

$$\widehat{E}(\Theta_{iT}|\boldsymbol{y}_T) = \sum_{l=1}^{L}\sum_{k=d+1}^{2d}[(1 - r_{iT}^{(lk)})y_{iT} + r_{iT}^{(lk)}\boldsymbol{x}_{iT}'\beta_T^{(lk)}]/(Ld)$$

$$\widehat{V}(\Theta_{iT}|\boldsymbol{y}_t) = \sum_{l=1}^{L}\sum_{k=d+1}^{2d}[\sigma_{iT}^2(1 - r_{iT}^{(lk)})]/(Ld)$$
$$+ \sum_{l=1}^{L}\sum_{k=d+1}^{2d}[(1 - r_{iT}^{(lk)})y_{iT} + r_{iT}^{(lk)}\boldsymbol{x}_{iT}'\beta_T^{(lk)}]^2/(Ld)$$
$$- \{\sum_{l=1}^{L}\sum_{k=d+1}^{2d}[(1 - r_{iT}^{(lk)})y_{iT} + r_{iT}^{(lk)}\boldsymbol{x}_{iT}'\beta_T^{(lk)}]\}^2/(Ld)^2$$

여기서 $r_{iT}^{(lk)} = \sigma_{iT}^2/(\sigma_{iT}^2 + \sigma_v^{2(lk)})$ 이다. $E(\Theta_{iT}|\boldsymbol{y}_T)$와 $V(\Theta_{iT}|\boldsymbol{y}_T)$는 시점 $t=T$에서 횡단면 자료만을 사용한다. 그 결과 $E(\Theta_{iT}|\boldsymbol{y}_T)$는 모든 자료를 사용하여 추정한 계층적 베이즈 추정량 $E(\Theta_{iT}|\boldsymbol{y}_T)$보다 효율성이 낮다.

5.4 LFS에의 응용

5.4.1 자료 설명 및 계산 수행

1999년 LFS 실업자수 추정값을 사용하여 계층적 베이즈 분석을 실시하였다. 캐나다 전국에 걸쳐 64개의 CA가 있으며 고용보험 EI 수혜률이 모형에서 보조자료 x_{it}로 사용되었다. 그러나 EI 수혜자 자료는 단지 62개 CA에서만 가능하다. 따라서 $m=62$ CA만 모형에 포함한다. 각 CA 내에서 1999년 1월부터 1999년 6월까지 연속적인 6개월 추정값 y_{it}를 고려하므로 $T=6$ 이다. 관심모수 Θ_{iT}는 1999년 6월에 지역 i에서의 실업률의 참값이다. 6개월 자료 만을 사용하는 이유는 LFS 표본 연동이 6개월 주기로 이루어지기 때문이다. 매달 LFS 표본의 1/6이 교체된다. 따라서 6개월 이후에는 추정값간의 상관이 매우 약하다. 1개월 시차상관이 약 0.48이며, 시차상관계수는 시차가 증가함에 따라 감소한다. Figure 1은 LFS 실업률 추정값에 대한 추정된(평활된) 시차상관계수를 나타내준다. 6개월 후 시차상관은 모두 0.1보다 작은 것이 확실하다.

모형에서 사용되는 표본공분산행렬 $\boldsymbol{\Sigma}_i$의 평활 추정값을 구하기 위해 각 CA에서 시간에 대한 (이 연구에서는 12개월) 평균 변동계수를 구하여 이를 $\overline{\mathrm{CV}}_i$로 나타내고, 시간과 모든 CA에 대한 평균 시차상관계수를 계산한다. 이 평활 CV와 시차상관계수를 사용하여 평활 추정값 $\widehat{\boldsymbol{\Sigma}}_i$를 구한다. 여기서 $\widehat{\boldsymbol{\Sigma}}_i$의

대각요소는 $\hat{\sigma}_{itt}=(\overline{CV}_i)^2 y_{it}^2$ 이고 비대각요소는 $\hat{\sigma}_{its}=\bar{\rho}_{|t-s|}(\hat{\sigma}_{itt}\hat{\sigma}_{iss})^{1/2}$ 이다. 여기서 $\bar{\rho}_{|t-s|}$는 시차 $|t-s|$의 평균 시차상관계수이다. 이렇게 구한 $\hat{\Sigma}_i$를 참 Σ_i로 간주한다. 조사한 바에 따르면 모형에서 Σ_i의 평활 추정값을 사용했을 때 CV 감소의 관점에서 추정량을 상당히 개선하는 것으로 알려졌다.

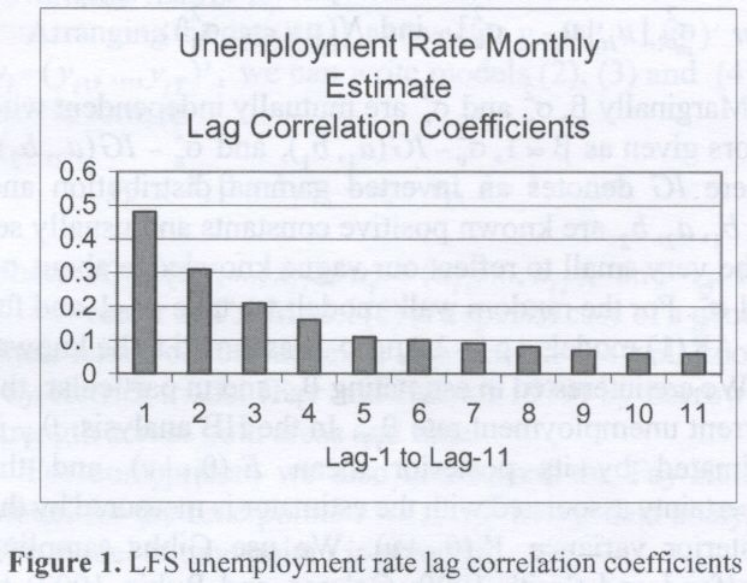

Figure 1. LFS unemployment rate lag correlation coefficients

깁스 샘플링을 수행하기 위해 각각 길이가 $2d=2,000$인 $L=10$개의 병렬 체인을 고려한다. 여기서 첫 $d=1,000$ burn-in 반복을 제거한다. 관심모수 Θ_{iT} $(i=1,\ldots,m)$에 대한 깁스 표본자의 수렴성을 조사하기 위해 Gelman과 Rubin(1992)의 방법대로 다음과 같은 절차를 따른다. 각 Θ_{iT}에 대해 Θ_{iT}^{lk}가 l번째 체인에서 k번째 난수를 나타낸다고 하자. 여기서 $k=d+1,\ldots,2d$이다. ($l=1,\ldots,L$). 첫 번째 단계는 전체평균

$$\overline{\Theta}_{iT}=\sum_{l=1}^{L}\sum_{k=d+1}^{2d}\Theta_{iT}^{(lk)}/(Ld)$$

과 수열내 평균

$$\overline{\Theta}_{iT}^{(l)} = \sum_{k=d+1}^{2d} \Theta_{iT}^{(lk)}/d, l=1,\ldots,L$$

를 계산한다. 그리고 L 수열 평균간의 분산인 B_{iT}/d를 구한다. 여기서

$B_{iT}/d = \sum_{l=1}^{L} (\overline{\Theta}_{iT} - \overline{\Theta}_{iT}^{(l)})^2/(L-1)$ 이다. 두 번째 단계는 각 자유도가 $(d-1)$인 L 수열내 분산 $s^2{}_{iTl}$의 평균인 W_{iT}를 계산한다. 즉, $W_{iT} = \sum_{l=1}^{L} s^2{}_{iTl}/L$ 이다. 세 번째 단계는 $s_{iT}^2 = (d-1)W_{iT}/d + B_{iT}/d$와 $V_{iT} = s_{iT}^2 + B_{iT}/(Ld)$ 를 계산한다. 마지막 단계는 잠재적 척도축소인자인 $\hat{R}_{iT} = V_{iT}/W_{iT}$ $(i=1,\ldots,m)$를 구한다. 만약 모든 관심모수 Θ_{iT}에 대한 잠재적 척도축소인자가 1에 가까우면 이는 깁스 표본자가 수렴한다는 것을 암시한다. 이 연구에서는 $\hat{R}_{iT}$ 값의 관점에서 깁스 표본자가 매우 잘 수렴하였다.

5.4.2 모형선택

이 절에서는 제안된 모형과 Rao와 You(1994)의 AR(1) 시간성분모형을 비교하고자 한다. 베이지안 구조에서 모형비교 방법이 많이 개발되었으며 그 중 몇 가지는 잘 알려진 BUGS 프로그램에 실려있다. (Spiegelhalter, Thomas, Best 그리고 Gilks, 1996 참조). 실제에 있어서 관심있는 모형이 하나 이상일 때 베이즈 요인(Bayes factor)을 근거로 베이지안 모형선택을 수행하지만 베이즈 요인은 직접 계산하기가 쉽지 않다.

모형선택에 대한 다른 대안으로 예측우도(predictive likelihood)와 예측 로그-우도를 사용한다. 특히 Dempster(1974)는 관측자료의 로그-우도의 사후분포를

조사하였다. 로그-우도의 사후분포의 양은 이탈도(deviance)인 $-2\log f(y|\Theta)$의 사후예측분포로부터 구할 수 있다. 사후 이탈도는 깁스 샘플링 결과를 사용하여 쉽게 추정할 수 있다. 왜냐하면 이는 $-2\log f(y|\Theta)$의 사후분포 $\pi(\Theta|y)$에 대한 기대이기 때문이다. 비계층적 모형에서 $-2\log f(y|\Theta)$의 가능한 최소값은 전통적 이탈도 통계량이다. 계층적 모형에서 이탈도의 최소값은 표본 최소값에 의해 불완전하게 추정될 가능성이 있으므로 이탈도의 평균이 더 합리적인 측도이다. (Karim과 Zeger, 1992; Gilks, Wang, Yvonnet 그리고 Coursagt, 1993 참조).

AR(1) 시간성분모형에서 ρ에 대한 2개의 값 $\rho=0.75$와 $\rho=0.5$를 고려하였다. 깁스 표본자의 각 반복에서 로그-우도를 계산하여 예측 사후이탈도의 평균을 계산한다. 이는 제안된 모형에서는 1311.5이고 $\rho=0.5$를 사용한 AR(1) 모형에서는 1372.8, $\rho=0.75$를 사용한 AR(1) 모형에서는 1358.3이다. 따라서 이탈도 측도에 의하면 u_{it}에 관한 확률보행모형이 AR(1) 모형보다 자료를 더 잘 적합시킨다.

모형비교를 위해 사후예측분포에 근거한 Laud와 Ibrahim(1995)의 발산측도(divergence measure)를 계산하였다. Θ^*를 y가 주어졌을 때 Θ의 사후분포로부터 추출한 값이라 두고, y^*를 사후예측분포 $f(y|y_{obs})$로부터의 표본이라 하자. 여기서 y_{obs}는 관측된 자료이다. Laud와 Ibrahim(1995)의 발산측도 기대값은 다음과 같이 $d(y^*, y_{obs}) = E(k^{-1}|y^* - y_{obs}|^2|y_{obs})$로 주어진다. 여기서 k는 y_{obs}의 차원이다. 이 두 모형 중에서 발산측도 기대값이 작은 모형을 선택한다.

Datta, Day 와 Maiti(1988) 그리고 Datta et al.(1999)와 같이 사후예측분포로부터 생성한 표본을 사용하여 발산측도 $d(y^*, y_{obs})$의 근사값을 계산한다. 깁

스 샘플링 결과를 사용하여 발산측도값을 계산하면 제안된 모형에서는 13.36, $\rho = 0.5$를 사용한 AR(1) 모형에서는 14.62, $\rho = 0.75$를 사용한 AR(1) 모형에서는 14.52이다. 따라서 발산측도에 의하면 확률보행모형이 AR(1) 모형보다 자료를 약간 더 잘 적합시킨다고 할 수 있다.

사후 이탈도와 발산측도는 두 개 이상의 가능한 모형을 비교하는 방법이다. 일단 하나의 모형을 선택한 후에는 선택한 모형이 자료를 잘 적합시키는지를 조사해야 한다.

5.4.3 모형적합 검정

제안한 모형의 총체적 적합을 조사하기 위해 사후예측 p값을 사용한다. (Meng, 1994; Gelman, Carlin, Stern 그리고 Rubin, 1995 참조). 이 방법에서는 적절한 불일치측도(discrepancy measure)의 모의실험 값이 사후예측분포로부터 생성되며 이를 관측된 자료에 대한 불일치측도 값과 비교한다. 구체적으로 $T(y, \Theta)$을 자료 y와 모수 Θ에 의존하는 불일치측도라 하자. 사후예측 p값은 다음과 같이 정의된다.

$$p = \text{prob}\left(T(y^*, \Theta) > T(y_{obs}, \Theta) | y_{obs} \right)$$

여기서 y^*는 사후예측분포 $f(y|y_{obs})$로부터의 표본이다. 확률은 관측된 자료가 주어졌을 때 Θ의 사후분포에 관한 것이다. 이는 보통 사용하는 p값의 베이지안 관점에서의 확장이다.

만약 모형이 관측된 자료를 적합시키면 불일치측도의 두 값이 비슷하다. 즉, 주어진 모형이 관측된 자료를 잘 적합시킨다면 $T(y_{obs}, \Theta)$이 사후예측분포로부터 y^*를 반복해서 생성했을 때 만들어지는 $T(y^*, \Theta)$의 히스토그램의 중심

부에 가까와야 한다. 따라서 만약 모형이 관측된 자료를 잘 적합시킨다면 사후예측 p값이 거의 0.5가 될 것으로 기대한다. 극단적인 사후예측 p값(거의 0이나 1)은 적합이 잘 이루어지지 않았음을 뜻한다.

깁스 표본자로부터 Θ^*의 모의실험으로 생성된 값을 사용하여 p값을 비교적 쉽게 계산할 수 있다. 각각 생성된 값 Θ^*에 대해 모형으로부터 y^*를 생성하고 $T(y^*, \Theta^*)$와 $T(y_{obs}, \Theta^*)$를 계산할 수 있다. 횡단면 및 시계열 모형에 대해 총체적 적합에 대해 사용되는 불일치측도는

$$d(\boldsymbol{y}, \Theta) = \sum_{i=1}^{m} (\boldsymbol{y}_i - \Theta_i)' \Sigma_i^{-1} (\boldsymbol{y}_i - \Theta_i)$$

이다. Datta et al.(1999)도 동일한 불일치측도를 사용하였다. 10개의 병렬체인으로부터 생성한 Θ^*와 y^*를 결합하여 p값을 계산한 결과 0.615이었다. 따라서 확률보행 시계열 및 횡단면 모형의 총체적 모형적합이 좋지 않다는 징후는 없다고 할 수 있다.

현재 시점의 횡단면 자료만 사용하는 Fay-Herriot 모형에 대한 근사적 불일치측도는 다음과 같이 주어진다.

$$d_{FH}(\boldsymbol{y}_T, \Theta_T) = \sum_{i=1}^{m} (y_{iT} - \Theta_{iT})^2 / \sigma^2_{iT}$$

여기서 $\Theta_T = (\Theta_{1T}, \ldots, \Theta_{mT})'$이다. 이 경우 추정된 p값은 0.587이며 이는 현재 시점의 횡단면 자료만 사용한 Fay-Herriot 모형이 자료를 잘 적합한다는 표시이다. 그러나 Fay-Herriot 모형에 의한 계층적 베이즈 추정값은 지역과 시간에 대해 동시에 정보를 빌려오는 제안한 횡단면 및 시계열 모형보다 유효성이 상당히 떨어진다. 이는 Figure 3과 Figure 4에 잘 나타나 있다.

사후예측 p값의 한계는 먼저 $f(y|y_{obs})$로부터 표본을 생성하고 또한 p값

을 계산하기 위해 관측된 자료 y_{obs}을 두 번 사용한다는 것이다. 이러한 자료의 중복 사용은 Bayarri와 Berger(2000)가 지적한 대로 자연스럽지 못한 성질을 유발한다. 자료의 중복 사용을 피하기 위해 Bayarri와 Berger(2000)는 두 가지 다른 대안의 p-측도를 제안하였다. 그러나 이들 측도는 사후예측 p값보다 특히 시계열 및 횡단면 모형같은 복잡한 모형에서 계산하기가 더 어렵다.

5.4.4 추정

확률보행 시계열 및 횡단면 모형 하에서 실업률의 사후 추정값을 구하자. 앞에서 주어진 Rao-Blackwellezed 추정량을 사용하여 Θ_{iT}의 사후평균과 사후분산의 추정값을 구하고 이를 HB1으로 나타내자. 모형에서 표본공분산행렬 Σ_i의 평활 추정값의 사용에 관한 영향을 조사하기 위해, 동일한 모형에 Σ_i의 직접조사 추정값을 사용하여 구한 추정값을 HB2라 하자. 시간으로부터 정보를 빌려오

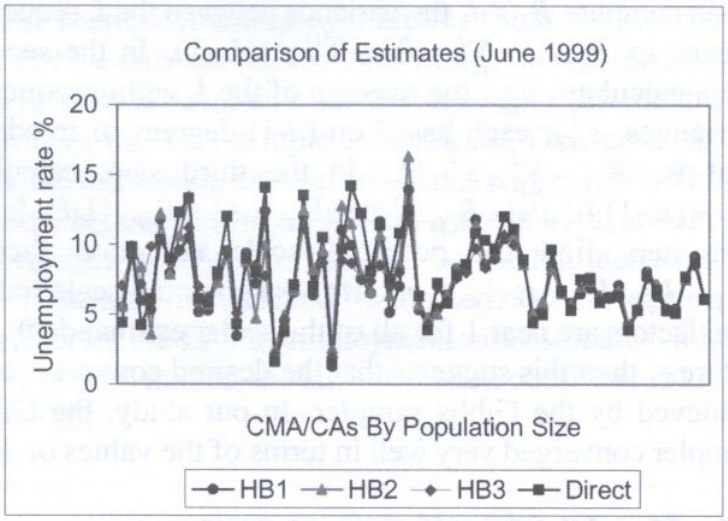

Figure 2. Comparison of direct and HB estimates

는 효과를 조사하기 위해 현재 시점의 횡단면 자료만을 사용한 Fay-Herriot 모형 하에서 Θ_{iT}의 계층적 베이즈 추정값을 구하고 이를 HB3로 나타내자.

Figure 2는 캐나다 전역에 걸친 62개의 CA에 대한 1999년 6월의 LFS 직접 추정값과 세 개의 경험적 베이즈 추정값을 나타낸 것이다. 62개의 CA를 모집단 크기에 따라 왼쪽에서 가장 작은 CA(Dawson Creek, BC, 인구 10,107명)부터 오른쪽의 가장 큰 CA(Totonto, Ont., 인구 3,746,123명)까지 나열하였다. 점추정 관점에서 Fay-Herriot 모형(HB3)이 실업률의 평균을 향해 추정값을 수축하는(shrink) 경향이 있으며, 이는 일반적으로 평활을 너무 많이 한 추정값을 만든다. HB2는 HB1보다 더 많은 변동을 가질 뿐만 아니라 더 극단적인 값을 가지는 경향이 있다. 이는 HB2가 표본오차에 영향을 받는 Σ_i의 직접추정량을 사용하기 때문이다. HB1은 직접 LFS 추정값의 적절한 평활이다. 많은 인구를 가짐으로 큰 표본을 가지는 큰 CA에 대해서는 직접추정값과 계층적 베이즈 추정값이 매우 비슷하나, 작은 CA들에 대해서는 어떤 지역에서 직접추정값과 계층적 추정값이 상당히 다르다.

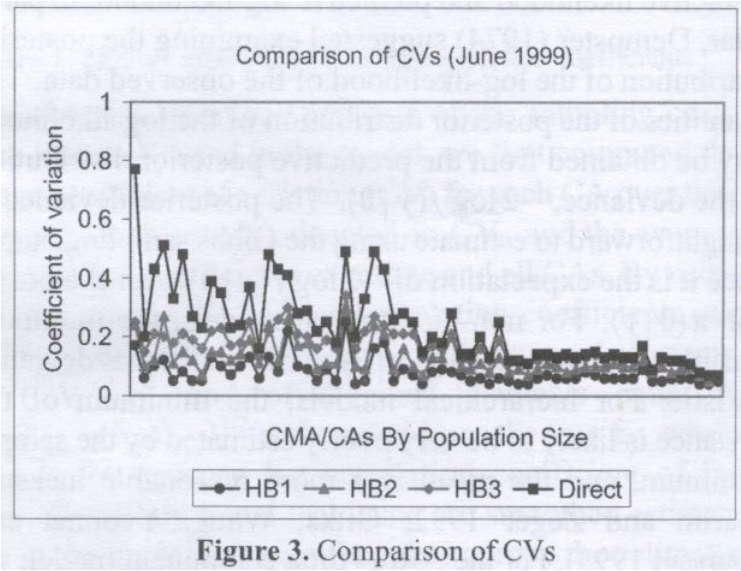

Figure 3. Comparison of CVs

Figure 3은 추정값의 변동계수(CV)를 나타낸다. 계층적 베이즈 추정값의 CV는 사후분산의 제곱근을 사후평균으로 나눈 값이다. Figure 3에서 직접추정값이 가장 큰 CV를 가지며, HB1이 가장 작은 CV를 가지는 것이 확실하다. HB1이 모든 CA에서 HB2보다 작은 CV를 가지며, HB2는 두 개의 비교적 작은 CA를 제외하고는 HB3보다 작은 CV를 가진다. HB 추정값의 효율성 이득이 특히 작은 인구를 가지는 CA에서 명백하다.

Figure 4는 HB1, HB2, HB3에 대한 직접조사 추정값의 퍼센트 CV 감소를 나타낸다. 퍼센트 CV 감소는 LFS CV에 대한 LFS CV와 HB CV의 차이로 정의되며, 백분율로 표현된다. HB1이 가장 큰 CV 감소를 가지며 HB3가 가장 작은 CV 감소를 가진다. Fay-Herriot 모형(HB3)의 직접 LFS 추정값에 대한 CV의 평균 퍼센트 감소는 21%이며, HB2에 대해서는 40%이고 HB3에 대해서는 62%이다. 또한 작은 CA에 대한 CV 감소는 큰 CA에 대한 CV 감소보다 더 현저하다. 모집단 크기가 증가함에 따라 CV 감소는 감소하는 경향이 있다.

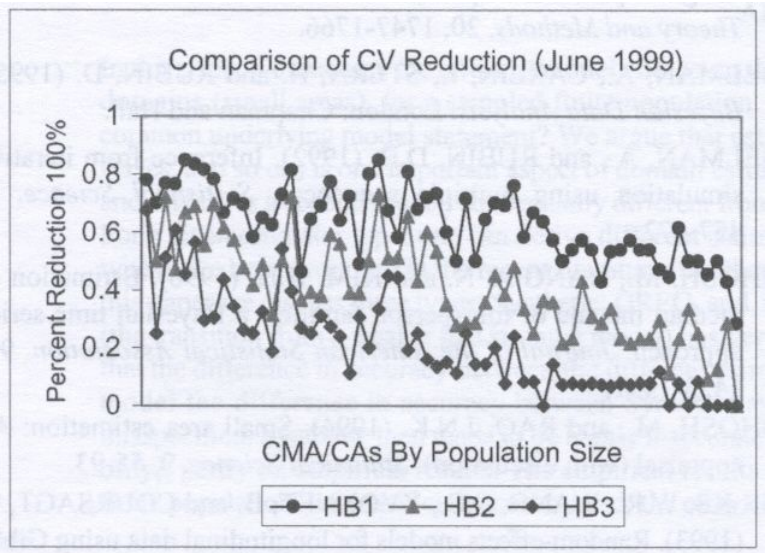

Figure 4. Comparison of CV reduction

이상의 통계적 분석을 요약하면 다음과 같은 결론을 얻는다.

(1) 모형기반 HB 추정값이 직접 LFS 추정값을 개선한다. 특히, 횡단면 및 확률보행 시계열 모형(HB1)이 LFS 추정값을 CV 감소의 관점에서 상당히 개선한다.

(2) 횡단면 및 확률보행 시계열 모형이 Fay-Herriot 모형보다 더 효과적이다.

(3) 표본분산-공분산행렬 Σ_i의 평활 추정값의 사용이 매우 효과적이다.

5.5 맺는말

LFS 자료를 사용하여 캐나다에서의 CA들에 대한 실업률의 모형기반 추정값을 구하기 위해 계층적 베이지안 횡단면 및 시계열 모형을 제안하였다. 이 모형은 지역과 과거시점으로부터 동시에 정보를 빌려오는 것이다. 분석에 의하면 랜덤 시계열성분에 확률보행과정을 가지는 모형이 자료를 잘 적합시킨다. 이 모형에 근거한 계층적 베이즈 추정값이 CV의 관점에서, 특히 작은 인구를 가진 작은 CA에 대해 직접조사 추정값을 상당히 개선한다. 그러나 이 CV는 모형의 표본분산공분산행렬 Σ_i을 안다는 가정에 기초한다. 결국 Σ_i의 추정과 관련된 불확실성을 무시한다.

현재시점 T에서의 자료만 사용하여 횡단면 정보만을 결합하는 잘 알려진 Fay-Herriot 모형도 고려하였다. Fay-Herriot 모형 하에서 CV는 직접추정량과 제안한 모형에 근거한 추정량 사이의 값이다. 횡단면 및 시계열 모형은 CV 감소의 관점에서 Fay-Herriot 모형에 비해 월등히 좋다. 제안된 모형이 공간 뿐만 아니라 시간에 대해서도 정보를 빌려온다는 관점에서 이는 Fay-Herriot 모형의 확장이기 때문에 예상된 결과이다.

LFS에 대한 응용에서 표본분산공분산행렬 Σ_i의 간단한 평활추정값을 $\boldsymbol{\Sigma}_i$의

참값으로 간주한다. Σ_i를 평활하는 다른 방법들에 대한 소지역 모수 Θ_{iT}의 계층적 베이즈 추정값과 관련된 CV의 민감도를 조사할 예정이다. 특히, 사용된 간단한 평활추정값 대신 $\tilde{\sigma}_{itt}=(\overline{CV_i})^2\Theta_{it}^2$와 $\tilde{\sigma}_{its}=\bar{\rho}_{|t-s|}(\tilde{\sigma}_{itt}\tilde{\sigma}_{iss})^{1/2}$ 형태의 평활추정값을 사용하는 것이 보다 현실적인지도 모른다. 그러나 이 경우 $\tilde{\sigma}_{itt}$와 $\tilde{\sigma}_{its}$는 미지의 모수 Θ_{it}에 의존하기 때문에 계층적 베이즈 방법의 계산이 더 어려워진다.

앞의 모형 설정에서 표본모형 $y_{it}=\Theta_{it}+e_{it}$와 대응하도록 모수 Θ_{it}에 선형혼합연결모형 $\Theta_{it}=\boldsymbol{x}_{it}'\beta+v_i+u_{it}$을 사용하였다. 최근 You와 Rao(2002)는 횡단면 자료에 위의 표본모형과는 달리 연결모형으로 비선형혼합모형을 사용하는 비대응 표본 및 연결모형을 개발하였다. You, Chen 그리고 Gambino(2002)는 Θ_{it}에 로그-선형모형을 사용하여 횡단면 및 시계열 자료에서 이 방법을 확장하였다.

6. 토의

본 연구 보고서는 고용통계의 모형기반 추정기법 개발 및 적용에 대한 작은 시도이다. 구체적으로 2003년 9월 경제활동인구조사 자료를 사용하여 191개의 소지역의 실업자 총계에 대한 경험적 베이즈 추정값과 계층적 베이즈 추정값을 계산하고 추정과 관련된 CV를 조사하였다. 기대한 바와 같이 직접추정값과 비교하여 CV값의 상당한 감소가 있었다.

사용한 Fay-Herriot 모형에 대한 계층적 베이즈 분석에서 적합결여 검증과 깁스 표본자에 대한 수렴여부 검증에도 별다른 문제가 없는 것으로 판명되었고, 경험적 베이즈 추정값과 계층적 베이즈 추정값 계산결과에서 모형기반 추정량의

평활 효과도 상당한 듯 하다.

한편, 경제활동인구조사 자료에 대해 Fay-Herriot 모형을 개선하는 노력 혹은 다른 형태의 모형을 동일 자료에 적용해보는 노력 등은 앞으로 모색되어야 할 과제인 듯 하다.

실업자 수의 총계 추정문제에서 가장 중요한 향후과제는 시군구의 소지역과 월과 년의 시계열효과를 모두 고려하는 횡단면 및 시계열 모형을 사용하여 분석하는 문제이다. 경제활동인구조사의 시계열적 특성을 모형화하는 노력의 일환으로 미국과 캐나다에서의 최근 중요한 연구동향을 요약하였다. 사실 이 보고서에서 고려한 횡단면 자료만 사용한 Fay-Herriot 모형을 사용한 분석이 앞으로 이루어질 횡단면 및 시계열 모형의 전단계 분석으로 이해하면 좋을 듯 하다.

또다른 관점에서 지금까지 제안된 합성추정법, 복합추정법, 경험적 베이즈 추정법, 계층적 베이즈 추정법을 비교하는 시뮬레이션 연구가 필요하다. 뿐만 아니라 특별시나 광역시 혹은 도 단위에서의 몇 개 지역을 선정하여 표본의 수를 상당히 많이 늘여서 조사(예컨대 20% 표본)를 실시한 후 그 결과를 평가기준이 되는 “gold standard”로 간주하고 가능한 여러 소지역 추정방법을 실증적으로 비교해 보는 것도 중요한 향후과제인 듯 하다.

참고문헌

[1] Bailar, B. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

[2] Battese, G., Harter, R. and Fuller, W. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of American Statistical Association*, 83, 28-36.

[3] Bayarri, M.J. and Berger, J.O. (2000). P values for composite mull models. *Journal of the American Statistical Association*, 95, 1127-1142.

[4] Bell, W.R. and Hillmer, S.C. (1990). The time series approach to estimation for repeated surveys. *Survey Methodology*, 16, 195-215.

[5] Chung, Y.S., Lee, K.-O and Kim, B.C. (2003). Adjustment of unemployment estimates based on small area estimation in Korea. *Survey Methodology*, 29, 45-52.

[6] Clayton, D. and Kalder, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics, 43, 671-681.

[7] Datta, G.S., Day, B. and Maiti, T. (1998). Multivariate Bayesian small area estimation: An application to survey and satellite data. *Sankhya*, 60, 344-362.

[8] Datta, G.S. and Ghosh, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. Annals of Statistics, 19, 1748-1770.

[9] Datta, G.S., Ghosh, M., Nangia,N. and Natarajan, K. (1996). Estimation of median income of four-person families: A Bayesian approach. In Bayesian

Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zeller (Eds. D.A. Berry, K.M. Chaloner and J.K. Geweke). New York: Wiley, pp. 129-140.

[10] Datta, G.S. and Lahiri, P. (1992). Composite estimation of unemployment rates for the small domains (with discussion). In *Proceedings of the Annual Research Conference, Bureau of the Census*, pp. 353-363.

[11] Datta, G.S., Lahiri, P. and Maiti, T. (2002). Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*, 102, 83-97.

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

[13] Dempster, A.P. (1974). The direct use of likelihood for significance testing (with discussion). In *Proceedings of Conference on Foundational Questions in Statistical Inference* (Eds. O. Barndorff-Nielsen, P. Blaeslid and G. Schou), Dept. of Theretical Statistics, University of Aarhus, Denmark, 335-354.

[14] Dempster, A.P. and Hwang, J.S. (1993). Component models and Bayesian technology for estimation of state employment and unemployment rates (with discussion). In *Proceedings of the Annual Research Conference, Bureau of the Census*, pp. 571-589.

[15] Fay, R.E. (1987). Application of multivariate regression to small domain estimation. In *Small Area Statistics* (Eds. R. Platek, J.N.K. Rao, C.E. Sarndal, and M.P. Singh). New York: Wiley.

[16] 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.

[17] Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. and Lindeyer, J. (1998). *Methodology of the Canadian Labour Force Survey*, Statistics Cananda, Catalogue No. 71-526.

[18] Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

[19] Gelfand, A.E. and Smith, A.F.M. (1991). Gibbs sampling for marginal posterior expections. *Communications In Statistics - Theory and Methods*, 20, 1747-1766.

[20] Gelman, A., Carlin, J., Stern, H. and Rubin, D. (1995). *Bayesian Data Analysis*, New York: Chapman and Hall.

[21] Gelman, A., Meng, X.L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 6, 733-807.

[22] Gelman, A. and Rubin, D.B. (1992), Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457-511

[23] Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*, New York: Chapman & Hall.

[24] Ghosh, M., Nangia, N. and Kim, D. (1996). Estimation of median income of four-person families: A Bayesian time series approach. *Journal of the*

American Statistical Association, 91, 1423-1431.

[25] Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B. (1998). Generalized linear models for small area estimation. *Journal of American Statistical Association*, 93, 273-282.

[26] Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 55-93.

[27] Gilks, W.R., Wang, C.C., Yvonnet, B. and Coursagt, P. (1993). Random-effects models for longitudinal data using Gibbs sampling. *Biometrics*, 49, 441-453.

[28] Gurney, M. and Daly, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. In *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 242-257.

[29] Huang, E.T. and Ernst, L.R. (1981). Comparison of an alternative estimator to the current composite estimator in CPS. In *Proceedings of the Survey Research Section, American Statistical Association*, pp. 303-308.

[30] Jiang, J., Lahiri, P. and Wan, S-M. (2002). A unified jacknife theory. Annals of Statistics, 30, in press.

[31] Kackar, R.N. and Harville, D.A. (1984). Approximations for standard errors of estimators for fixed and random effects in mixed models. *Journal of the American Statistical Association*, 79, 853-862.

[32] Karim, M.R. and Zeger, S.L. (1992). Generalized linear models with random effects: Salamander mating revisited. *Biometrics*, 48, 631-644.

[33] Laud, P. and Ibrahim, J. (1995). Predictive model selection. *Journal of the*

Royal Statistical Society, Ser. B, 57, 247-262.

[34] Meng, X.L. (1994). Posterior predictive p value. *The Annals of Statistics*, 22, 1142-1160.

[35] Morris, C.N. (1983). Parametric empirical Bayes inference: theory and applications (with discussion). *Journal of the American Statistical Association*, 78, 47-65.

[36] Pheffermann, D. and Burck, L. (1990). Robust small-area estimation combining time Series and cross-sectional data. *Survey Methodology*, 16, 217-237.

[37] Pheffermann, D., Feder, M. and Signorelli, D. (1998). Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business and Economic Statistics*, 16, 339-348.

[38] Platek, R., Rao, J.N.K., Sarndal, C.E. and Singh, M.P. (Eds.) (1987). *Small Area Estimation*. New York: Wiley.

[39] Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

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

[41] Rao, J.N.K. (2003). Small Area Estimation. New York: Wiley.

[42] Rao, J.N.K. and Graham, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

[43] Rao, J.N.K. and Yu, M. (1992). Small-area estimation by combining time series and cross-sectional data. In *Proceedings of the Survey Research Section, American Statistical Association*, pp 1-9.

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

[45] Scott, A.J. and Smith, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.

[46] Singh, A.C., Mantel, H., and Thomas, B.W. (1991). Time series generalizations of the Fay-Herriot estimator for small areas. In *Proceedings of the Survey Research Section, American Statistical Association*, pp. 455-460.

[47] Sinha, D, and Dey, D. (1997). Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association*, 92, 1195-1212.

[48] Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996). BUGS 0.6: Bayesing inference Using Gibbs Sampling Manual. Available at http://www.mrc-bsu.cam.ac.uk/bugs.

[49] Tiller, R. (1989). A Kalman filter approach to labor force estimates using survey data. In *Proceedings of the Survey Research Section, American Statistical Association*, pp. 16-25.

[50] Tiller, R. (1992). Time series modeling of sample survey data from the U.S. current population survey. *Journal of Official Statistics*, 8, 149-166.

[51] Tsutakawa, R.K. (1985). Estimation of cancer mortality rates: A

Bayesian analysis of small frequencies. *Biometrics*, 41, 69-79.

[52] You, Y. (1999), *Hierarchical Bayes and Related Methods for Model-Based Small Area Estimation.* Unpublished Ph.D. Thesis, School of Mathematics and Statistics, Carleton University, Ottawa, Canada.

[53] You, Y., Chen, E. and Gambino, J. (2002). Nonlinear mixed effects cross-sectional and time series models for unemployment rate estimation. In *2002 Proceedings of the American Statistical Association, Section on Government Statistics* [CD-ROM], Alexander, VA: American Statistical Association.

[54] You, Y., and Rao, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*, 26, 173-181.

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

모형기반 소지역추정
방법 연구결과 자료

모형기반 소지역추정방법

2004. 1

통 계 기 획 국
조 사 관 리 과

목 차