제1장

다양한 소스의 보조정보를 이용한 소지역 추정 기법 연구

김서영 · 김재광 · 이정희 · 권순필

제1절 서 론

우리나라 통계청과 같은 국가통계 기관에서 실시하는 대부분의 표본조사는 전국 수준 또는 보다 큰 공간단위에서 통계를 정확하게 생산하는 것을 목표로 하고 있다. 최근 이 사회는 점점 더 빠르고 복잡하게 변하고 있고, 이러한 변화의 다양성을 반영할 수 있는 더 많은 정보가 필요하게 되었다. 통계 작성 기관은 이러한 사회변화의 정보를 가능한 자세하게 파악할 수 있도록 하기 위해 보다 작은 공간(시도 또는 시군구) 또는 영역에서 통계를 생산하는데 관심을 갖게 되었다. 이들 작은 지역들은 “소지역(small area)"이라 부르고, 이때 ”소(小, small)"의 의미는 조사가 실시되는 지역 또는 영역 내에서 표본크기가 작다는 것을 의미한다.

소지역 통계를 제시하고 있는 많은 국가들 중 우리가 흔히 통계 선진국가라고 부르는 미국, 캐나다, 영국 등은 소지역 통계를 소지역 추정기법에 의해 제공하는 방법을 사용하고 있다. 가장 잘 알려진 소지역 추정으로는 SAIPE (Small Area Income and Poverty use of Estimation) 프로그램을 들 수 있을 것이다. 이는 Fay 와 Herriot (1979)에 처음 제안된 것으로서 통계적 모형을 이용하였고, 이 추정 결과는 미국 교육부에서 카운티에 대해 매년 교육정책자금을 배당하는데 사용되고 있다. SAIPE 방법은 소지역에서의 빈곤율을 추정하기 위해 회귀모형을 사용하였으며, 회귀모형은 설명변수로서 세금자료 (tax record)와 식권 (food stamp)과 같은 보조정보를 포함하고 있다. 카운티 단위의 빈곤율 추정은 회귀모형으로부터 추정된 예측치와 미국 CPS (Current Population Survey)의 직접추정치(direct estimate)가 결합되어 얻어진 결과이다. 이 방법에서 등록자

료는 오차가 없다고 간주되었고, 또한 다른 표본조사 자료는 보조정보로 사용되지 않았다.

소지역추정에는 소지역 내에서 우리가 관심 있는 값과 관련된 변수들을 이용한 혼합모형(mixed model)이 자주 사용되었다. 일반적으로 소지역은 표본크기가 작기 때문에 해당 지역이 가지고 있는 정보만으로는 좋은 추정량을 만들기 어렵다. 혼합모형은 보조정보를 이용한 지역 랜덤효과를 통해 다른 소지역들로부터 정보를 빌려(borrow strength) 사용함으로써 우리가 관심 있는 모수를 추정할 수 있도록 한다. Fay 와 Herriot (1979)는 처음 보조변수에 대한 평균벡터를 사용하여 소지역에서의 모수 추정을 향상시켰다. 이후에도 Dempster 등 (1981), Fuller와 Harter (1987) 등 많은 연구가 있었다. Ghosh 와 Rao (1994), Marker (1999), Rao (1999)는 소지역 추정에 관한 많은 연구문헌들을 작성하였다. Datta 등 (1999)은 다변량 소지역 추정에 관한 이론을 유도하였고, Prasad 와 Rao (1999)는 소지역 추정시 설계가중치를 사용함으로써 추정치의 정도를 더욱 좋게 하였다.

많은 연구들에서 소지역 추정량들의 편향 (bias)과 평균제곱오차 (mean squared error)와 같은 특성들은 보조정보에 의해 유도되고, 이들 보조정보들은 모든 지역에 대해서 이용가능하고 오차 없이 측정되었다고 가정되었다. 즉, 지금까지의 많은 소지역 추정 연구들에서 보조정보는 오차가 없는 변수들만이 사용되었다. 그렇지만 실제 상황에서는 어떤 종류에서든 오차가 없는 변수들은 거의 없다고 볼 수 있다. 전수조사와 등록자료의 경우는 표본조사에서 발생하는 표본 오차는 없을 수 있다. 그렇지만 센서스와 같은 전수조사의 경우에도 측정오차는 존재하고, 경우에 따라서는 측정오차가 매우 심각할 수 있다. 또한 등록자료는 커버리지 오차 또는 잘못된 기록에 따른 오차가 발생할 수 있다.

소지역 추정 모형은 이처럼 측정오차가 없는 보조정보 외에도 소지역에 대해 다른 정보에서 기인하는 표본조사 자료도 포함할 수 있다. 이는 어떤 목표변수에 대해서 2개 이상 별도의 조사가 존재하는 경우가 해당된다. 예를 들면, 우리나라 소지역에서의 실업자에 대한 정보는 표본조사자료인 경제활동인구조사 (경활조사), 지역별고용조사 (고용조사), 전수자료인 센서스 그리고 등록자료인 실업급여보험청구자료 (실업급여등록자자료)로부터 얻을 수 있다. 뿐만 아니라 가계소득에 대한 정보는 가계동향조사와 가계금융조사로부터 얻어질 수 있고, 이들 모두는 표본조사자료에 해당한다.

이처럼 표본오차 또는 측정오차가 포함된 정보도 소지역 추정의 보조정보로 활용할 수 있다. 다양한 보조정보 활용에 관한 연구는 1990년대 후반부터 진행되었고 (Zieschang, 1990; Renssen and Nieuwenbroek, 1997), 이는 최근에 이르러 더 많은 주목을 받고 있는 것 같다. Ybarra 와 Lohr (2008)은 측정오차가 있는 보조정보를 이용한 소지역 추정방법으로 다변량 Fay-Herriot 모형을 이용하였다. Merkouris (2010)은 다양한 조사 자

료를 이용한 작은 영역 (small domain) 단위에서의 회귀 추정 방법을 제시하였다. 또한 Lohr 와 Prasad는 보조정보로서 조사 자료를 이용한 소지역 추정방법을 논의하였다.

본 연구에서는 다양한 소스의 보조정보가 있을 때, 특히 보조정보가 오차를 포함하고 있을 때의 소지역 추정 방법을 고려하고자 한다. 이때 지역수준 모형 (area level model)을 사용하고, 보조정보는 측정오차가 있는 조사 자료, 커버리지 오차가 있는 행정등록 자료 및 측정오차 또는 갱신되지 않은 정보를 포함하는 센서스 자료를 고려해 보고자 한다. 제안된 방법은 표본오차 모형(sampling error model)과 구조오차 모형(structural error model)을 사용하고, 모수 추정은 GLS (Generalized Least Squares) 또는 GMM (Generalized method of moment) 방법을 사용한다. 게다가 등록자료의 경우 커버리지 오차를 제어할 수 있도록 한다.

본 연구의 구성은 다음과 같다. 2절에서는 본 연구에서 제안하고자 하는 소지역 추정 방법에 대한 기본 개념들을 설명한다. 3절에서는 경활조사자료, 지역별고용조사자료 및 등록자료를 결합한 복합추정량을 제시한다. 이에 필요한 이론과 방법론에 대해 설명한다. 또한 실제 자료에 대해 제안한 방법을 적용하고 그 결과를 요약・설명한다. 4절에서는 경활조사, 지역별고용조사, 센서스 자료를 결합한 복합추정방법을 제안한다. 이때 등록자료는 결합하지 않는다. 모형에 대한 모수추정방법과 GLS 추정방법을 설명하고, MSE 추정 방법을 제시한다. 5절에서는 4절에서 제안한 복합추정량에 대해 실제자료를 적용하고, 그 추정량을 평가한다. 평가는 각 추정량들의 MSE와 CV 기준을 이용하여 비교 설명한다. 이때 목표 추정치는 우리나라 시군구 실업률이 된다. 마지막으로 6절에서는 4절에서 제시한 GLS 복합추정량을 중심으로 결론을 내리고, 본 연구에서 제안한 방법과 관련하여 향후 연구 방향에 대해 언급한다.

제2절 다양한 소스 자료를 이용한 소지역 추정

1. 개념

동일한 속성에 대한 자료가 여러 개 있는 경우 이것을 적절하게 결합하여 보다 나은 추정을 하고자 하는 것은 통계학의 가장 중요한 문제 중의 하나이다. 이러한 것을 어떻게 잘 결합할 것인가는 결국 모형의 선택과 추정량의 선택으로 귀결된다. 추정 모형과 추정량의 좋고 나쁨은 자료에 따라 달라질 수 있고, 이러한 결정은 몇 가지 주요 시나리오에 따라 얻어지는 추정량의 MSE 추정값으로 해야 할 문제이다.

본 연구에서는 두 개 이상의 서로 다른 조사에서 동일한 속성에 대한 값을 측정하는 경우 이를 어떻게 소지역 추정에 반영하는지를 우선적으로 다루고자 한다. 이와 더불어 등록자료가 있는 경우 특히 커버리지 오차가 큰 자료를 추정에 반영하는 방법도 연구하기로 한다. 모형은 크게 관측치 관련 모형과 구조 관련 모형으로 나눌 수 있다. 구조 관련 모형은 두 조사를 연결하는 연결모형 (linking model)으로서 소지역 추정의 핵심이 된다. 구조 모형이 결정되고 나면 모형에 대한 모수를 추정하게 되고, 이 모형으로부터 모형기반 추정과 합성 추정이 가능하게 된다.

2. 모형

동일한 속성 항목에 대해 측정하는 서로 다른 조사가 두 가지 있다고 하자. 두 조사가 모두 전수조사라 하더라도 동일 항목에 대해 다르게 측정되는 것은 마찬가지 현상이다. 왜냐하면 조사과정에서 동일한 조사원과 동일한 설문 문항을 동일한 시점에서 사용하지 않는 이상 두 값이 항상 같을 수는 없기 때문이다. 두 조사 (A, B 라고 하자)에서 얻어진 관측값을 각각 y_a 와 x_b 라고 하고, 이 때 관심값에 대한 참값을 Y 라고 하자. 그러면 각 개인 i 에 대해 얻어질 수 있는 관측값 $y_{a,i}$ 와 $x_{b,i}$ 는 참값 Y_i 를 측정하는 조사값들이다. 여기서 조사 A 는 경제활동인구조사 (경활조사), 조사 B 는 지역별고용조사 (고용조사) 또는 센서스라고 하면 다음과 같은 모형을 생각할 수 있을 것이다.

$$y_{ai} = Y_i + e_{ai} \tag{2.1}$$

$$x_{bi} = Y_i + e_{bi}$$

이고, 여기서 e_{ai} 와 e_{bi} 는 각각 조사 A 와 조사 B 의 측정오차를 나타내는 확률변수이다. 즉, 동일인에 대해 반복하여 그 조사를 실시하였을 때 얻어지는 변동을 나타낸다. 측정 오차는 그 정의상 평균이 0 이고 분산은 각각 σ_a^2 와 σ_b^2 으로 표현된다. 이러한 측정오차는 거의 없다고 해도 무방하나 조사 B에서 훈련되지 않은 조사원을 사용한다면 σ_b^2 값은 σ_a^2 값보다 상당히 크게 될 것이다. 모형 (2.1)은 이상적인 경우로 두 모형 모두 오차가 존재하긴 하지만 오차의 평균이 둘 다 0 이므로 전국 단위에서는 전체 평균이 거의 같아지게 된다. 만약 전국 단위에서 전체 평균이 다르다면, 이는 두 조사 중 적어도 하나는 오차의 평균이 0 이 아니라는 것을 의미한다. 즉, 이러한 경우에는 측정 편향 (measurement bias)이 존재한다고 볼 수 있을 것이다.

대부분의 경우 두 조사 중 하나는 측정 편향 (measurement bias)이 존재한다. 예를 들어 고용조사는 경활조사에 비해서 체계적(systematic)인 오차가 있다고 볼 수 있는 것이다. 이러한 경우 측정오차 모형은 (2.1)이 아니라 다음과 같이 표현될 수 있다.

$$y_{ai} = Y_i + e_{ai} \tag{2.2}$$

$$x_{bi} = X_i + e_{bi}$$

이때 X_i 는 고용조사에서 얻어지게 될 관측값의 기대값으로서 Y_i 와 관련이 있다. 즉, $X_i - Y_i$ 는 조사 B의 측정 편향을 나타낸다. 이러한 관련성은 다음과 같은 모형으로 표현될 수 있다.

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{2.3a}$$

여기서 u_i 는 Y_i 를 X_i 의 선형 모형으로 설명하고자 할 때 발생하는 모형오차로써 평균이 0이고, 분산이 σ_u^2 인 오차라 할 수 있다. (등분산이 아닐 수도 있다.) 모형 (2.3a) 대신에 다음과 같은 모형을 생각할 수도 있다.

$$X_i = \beta_0 + \beta_1 Y_i + u_i \tag{2.3b}$$

모형 (2.3b)는 모형 (2.3a)와 동일한 개념이지만 Y_i 를 설명변수로 사용할 경우 향후 추정이 더 편리하게 될 수 있다는 장점이 있다. 왜냐하면 조사 A 는 측정 오차가 거의 없다고 간주하기 때문이다. 아무튼 모형 (2.3a)와 (2.3b)는 모형 (2.2)와 본질적으로 다른 모형이다. 모형 (2.2)는 어떤 속성을 측정할 때 발생하는 오차를 나타내는 측정오차 모형 (measurement error model)이고, 모형 (2.3)은 두 변수의 구조적 오차를 나타내는 구조오차 모형 (structural error model)이다. 측정오차 모형은 각 조사에서의 표본오차 (sampling error) 와 관련이 있으며 구조오차 모형은 각 조사에서의 비표본오차 (non-sampling error) 와 관련이 있다. 간단하게 측정오차 모형은 측정모형, 구조오차 모형은 구조모형으로 부르기도 한다.

만약 연령대나 성별 등과 같은 인구학적 변수가 있어서 (2.3a)나 (2.3b)의 모형을 더욱 개선시킬 수 있다고 한다면 이러한 변수 X를 모형에 설명변수로 추가시킬 수 있다. 그렇게 개선된 모형은 σ_u^2 값이 줄어들 것이고 이는 모형기반 추정의 오차를 줄여준다.

모형 (2.3)에서 우리는 (Y_i, X_i) 를 관측하지 못한다. 변수 X_i 와 관련된 측정값 x_b 는 조사 B에서 관측하고 변수 Y 와 관련된 측정값 y_a 는 조사 A에서 관측될 뿐이다. 이 두 별개의 조사를 연결하는 모형이 (2.3)이므로 모형 (2.3)을 연결 모형이라고도 한다. 그렇다면 과연 y_a 와 x_b 만을 이용하여 어떻게 모형 (2.3)의 모수들을 추정하고 그것을 바탕으로 소지역의 Y 값들을 예측 (prediction) 할 수 있을까 하는 것이다. 이러한 질문에 대한 대답은 개인 단위 자료에 대해 연결이 가능한 경우와 그렇지 않은 경우로 나누어서 생각해 볼 수 있다.

모형 (2.2)와 모형 (2.3)은 개인 단위 (unit level)로 표현되었다. 이를 지역 단위 (area level) 형태 모형으로 표현할 수 있다. 지역 h에서 얻어지는 두 가지 추정량을 각각 $\widehat{Y_{ha}} = \sum_{i \in A_h} w_{ia} y_{ai}$ 와 $\widehat{X_{hb}} = \sum_{i \in B_h} w_{ib} x_{bi}$이라고 한다면 두 추정량에 대한 측정 모형은 다음과 같다.

$$\widehat{Y_{ha}} = Y_h + e_{ha} \qquad (2.4)$$

$$\widehat{X_{hb}} = X_h + e_{hb}$$

이때, $Y_h = \sum_{i \in U_h} Y_i$, $X_h = \sum_{i \in U_h} X_i$ 이고, 오차항 $e_{ha} = \widehat{Y_{ha}} - Y_h$ 은 표본 오차와 측정 오차를 모두 포함한다. 다행히 $\widehat{Y_{ha}}$ 의 불편(unbiased) 분산 추정량(= $\hat{V}(\widehat{Y_{ha}})$)은 오차항($e_{ha}$)의 MSE를 편향되지 않게 추정한다. 마찬가지로 $\widehat{X_{hb}}$의 불편 분산 추정량 $\hat{V}(\widehat{X_{hb}})$ 은 오차항(e_{hb})의 MSE를 편향되지 않게 추정한다. 궁극적으로 모형 (2.4)는 직접 추정량의 표본 오차 (측정 오차 포함)를 나타낸다.

마찬가지로 모형 (2.3a)를 이용한 지역 단위의 구조 모형은 다음과 같다.

$$Y_h = \beta_0 N_h + \beta_1 X_h + u_h \qquad (2.5a)$$

여기서 $u_h = \sum_{i \in U_h} u_i$ 는 지역 단위 모형 오차로써 평균이 0 이고 분산이 $N_h\sigma_u^2$이다.

참고로, 지역단위 모형을 설정할 때 지역의 정의를 시군구로 할 것인가 아니면 시군구내 성별, 연령별로 할 것인가는 선택의 문제이다. 일반적으로 추정 단위를 세분화할수록 구조오차 모형의 오차는 더 작아지고, 이에 따라 더 정교한 소지역 추정을 할 수 있지만 지역이 너무 세분화되면 해당 소지역의 표본이 매우 적거나 없어서 모형 (2.4)의 직접 추정량이 구현되지 않을 수 있을 것이다.

모형을 (2.3b)로 사용하는 경우, 모형은 다시 다음과 같이 표현될 수 있다.

$$X_h = \beta_0 N_h + \beta_1 Y_h + u_h. \tag{2.5b}$$

이러한 지역 단위 구조 모형은 모수 추정을 가능하게 하고, 향후 모형 기반 추정량을 계산하는데 사용된다.

3. 모수 추정과 모형 기반 추정

가. 개인 단위 모형

구조 모형에서 모수 추정은 두 조사의 자료가 서로 연계 가능한 경우와 그렇지 않은 경우로 나누어 생각할 수 있다. 먼저 두 조사 자료 간에 레코별로 연계가 가능한 경우에는 한 조사 자료에 다른 조사에서 얻어진 관측치를 추가할 수 있기 때문에 이를 바탕으로 모형을 세울 수 있다. 이러한 경우 모형 (2.2)와 모형 (2.3a)를 결합하게 되면 다음과 같이 표현된다.

$$y_{ai} = Y_i + e_{ai} = \beta_0 + \beta_1 (x_{bi} - e_{bi}) + u_i + e_{ai} \tag{2.6a}$$

따라서 단순하게 y_{ai} 를 x_{bi} 회귀식에 적합하면 편향된 추정을 초래하게 된다. 이를 해결하기 위해서는 측정오차 모형의 추정 공식을 이용하여 추정해야 한다. 다른 방법으로는 모형 (2.2)와 모형 (2.3b)를 결합하는 것을 생각해 볼 수 있다. 이러한 경우 모형은 다음과 같다.

$$x_{bi} = X_i + e_{bi} = \beta_0 + \beta_1 (y_{ai} - e_{ai}) + u_i + e_{bi} \tag{2.6b}$$

식 (2.6b)는 식 (2.6a)와 비슷해 보이지만 실제로 조사 A 의 측정 오차는 매우 작을 것이므로 이를 무시할 수 있다고 한다면 (2.6b)는 다음의 모형으로 표현될 수 있다.

$$x_{bi} = \beta_0 + \beta_1 y_{ai} + \eta_i \tag{2.6c}$$

여기서 $\eta_i = u_i + e_{bi}$ 는 평균이 0 이고, 분산이 σ_η^2 인 오차항이다. 이 경우 회귀 모수 추정은 간단한 최소자승법을 사용할 수 있게 된다. 실제로는 조사 A 의 가중치를 이용하여 추정해야 한다.

이렇게 해서 얻어진 모수 추정치 $\hat{\beta} = (\hat{\beta_0}, \hat{\beta_1})$을 이용하여 모형 기반 추정량을 구현할 수 있는데 기본적인 원리는 예측에 의한 추정이다. 조사 B 는 조사 A 보다 표본 크기가 훨씬 크기 때문에, 조사 B 의 표본으로부터 y_i의 예측치를 추정하여 소지역 모수에 해당하는 $Y_h = \sum_{i \in U_h} Y_i$ 를 $\widehat{Y_h} = \sum_{i \in B_h} w_{bi} \widehat{Y_i}$ 으로 예측하고자 하는 것이다. 모형 (2.6b)를 이용하면 Y_i 의 예측치는

$$\widehat{Y_i} = \widehat{\beta_1^{-1}} (x_{bi} - \hat{\beta_0}) = \hat{\beta_0^*} + \hat{\beta_1^*} x_{bi} \tag{2.7}$$

으로 계산될 수 있고, 따라서 Y_h 의 모형 기반 추정량은

$$\widehat{Y_{hb}} = \sum_{i \in B_h} w_{ib} (\hat{\beta_0^*} + \hat{\beta_1^*} x_{ib}) = \hat{\beta_0^*} \widehat{N_{hb}} + \hat{\beta_1^*} \widehat{X_{hb}} \tag{2.8}$$

으로 계산될 수 있다. 이 때 회귀계수의 변동성은 근사적으로 무시할 수 있기 때문에, 모형 기반 추정량의 MSE 는 $\alpha' \widehat{V_{hb}} \alpha + N_h \sigma_\eta^2$ 이 된다. 여기서 $\alpha = (\alpha_0, \alpha_1)$ 을 나타낸다. $\widehat{V_{hb}}$ 는 $(\widehat{N_{hb}}, \widehat{X_{hb}})$ 의 분산공분산 행렬을 나타낸다. 분산 모수 σ_η^2 을 추정하기 위해서는

$\hat{\eta}_i = x_{ib} - \hat{\beta}_0 - \hat{\beta}_1 y_{ia}$ 를 계산한 후, 이들의 표본 분산값을 이용하여 계산할 수 있다. 모형 기반 추정량은 (2.7)의 합성값 (synthetic value)을 이용하기 때문에 이를 합성 추정량이라고 부르기도 한다.

나. 지역 단위 모형

두 자료의 레코드 연결이 불가능한 경우에는 개인 단위 모형을 이용하여 모수 추정을 할 수가 없다. 대안적으로 가장 많이 사용되는 방법으로는 지역 단위 모형을 이용하여 추정하는 방법이 있다. 방법적으로는 모형 (2.4)와 모형 (2.5a)를 결합하는 방법과 모형 (2.4)와 모형 (2.5b)를 결합하는 방법이 있는데, 추정하는 과정에서 있어서 후자가 더 편리한 면이 있다는 점에서 그 경우만 고려하기로 한다. 이 경우 모형 기반 추정을 위한 결합 모형은 다음과 같다.

$$\widehat{X_{hb}} = \beta_0 N_h + \beta_1 Y_h + u_h + e_{hb} = \beta_0 N_h + \beta_1 \widehat{Y_{ha}} + u_h - \beta_1 e_{ha} + e_{hb}$$

이를 기반으로 β_0, β_1를 추정할 수 있다면, 모형 기반 추정량 (또는 합성추정량)은 다음과 같이 계산된다.

$$\widehat{Y_{hb}} = \widehat{\beta_1^{-1}}\left(\widehat{X_{hb}} - \hat{\beta}_0 \widehat{N_{hb}}\right)$$

본 연구에서는 위의 지역 기반 모형을 이용한 추정량을 제시하고자 한다. 따라서 이에 대한 모수추정과 추정량에 관한 내용은 다음 절에서 자세하게 논의하기로 한다.

제3절 조사와 등록자료를 보조정보로 이용한 복합추정

1. 기본 이론

본 절에서는 2절에서 설명한 모형에 대한 기본적 이해를 바탕으로 본 연구 목적에 부합된 조사자료와 등록자료의 활용을 전제로 한, 소지역 추정에 필요한 기본적 이론을 전개하고자 한다. 어떤 부분들은 2절의 내용과 중복되기도 하겠지만, 본 절은 실제 자료

의 적용을 전제로 방법론을 전개한다는 점에 집중할 필요가 있다. 우선은 보조정보가 하나만 이용할 수 있는 경우에 대해 설명한다. 즉, 경활조사 자료를 종속변수, 고용조사자료를 보조변수로 간주한다. 다음 절에서는 이러한 개념을 등록자료 등으로 확장하고자 한다.

두 조사 A, B가 있다고 하자. 이 두 조사는 각각 다른 확률표본설계를 따르고 있다. 조사 A 에 의한 설계기반 불편 추정량은 $\hat{Y}_{ha} = \sum_{i \in A_h} w_{ia} y_i$ 이고, 이 추정량의 분산 추정량은 $\hat{V}(\hat{Y}_{ha})$이다. 그리고 조사 B 의 합성추정량 $\hat{Y}_{hb} = \sum_{i \in B_h} w_{ib} \tilde{y}_i$ 은 $Y_h = \sum_{i \in U_h} y_i$의 설계기반 불편 추정량이고, 여기서 $E(\tilde{y}_i - y_i) = 0$, 추정량에 대한 MSE 추정량은 $\widehat{MSE}(\widehat{Y}_{hb})$. 이때 회귀 계수 β_0, β_1가 알려져 있다면, $\tilde{y}_i = (x_{1bi} - \beta_0)/\beta_1$가 된다.

이제 다음의 함수 Q를 최소화하는 Y_h의 추정량을 찾음으로써 두 조사를 연결할 수 있다.

$$Q(Y_1, \cdots, Y_H) = \sum_{h=1}^{H} \begin{pmatrix} \widehat{Y}_{ha} - Y_h \\ \widehat{Y}_{hb} - Y_h \end{pmatrix}' \begin{pmatrix} \hat{V}(\widehat{Y}_{ha}) & 0 \\ 0 & \widehat{MSE}(\widehat{Y}_{hb}) \end{pmatrix}^{-1} \begin{pmatrix} \widehat{Y}_{ha} - Y_h \\ \widehat{Y}_{hb} - Y_h \end{pmatrix}, \tag{3.1}$$

여기서 $\hat{V}(\widehat{Y}_{ha})$는 $\widehat{Y}_{ha}$의 분산추정량이고 $\widehat{MSE}(\widehat{Y}_{hb})$ $\widehat{Y}_{hb}$의 MSE에 대한 불편추정량이다. 이를 최소로 하는 해는 다음과 같다.

$$\widehat{Y}_h^* = \frac{\widehat{Y}_{ha} / \hat{V}(\widehat{Y}_{ha}) + \widehat{Y}_{hb} / \widehat{MSE}(\widehat{Y}_{hb})}{1/\hat{V}(\widehat{Y}_{ha}) + 1/\widehat{MSE}(\widehat{Y}_{hb})} \tag{3.2}$$

식 (3.1)에서 두 조사가 서로 독립이라고 하면, $\widehat{Y}_{hb} = \sum_{i \in B_h} w_{ib} \tilde{y}_i$의 MSE (평균제곱오차)는

$$MSE(\widehat{Y}_{hb}) = E\left\{\left(\sum_{i\in B_h} w_{ib}\widetilde{y}_i - \sum_{i\in U_h} y_i\right)^2\right\}$$

$$E\left\{\left(\sum_{i\in B_h} w_{ib}\widetilde{y}_i - \sum_{i\in U_h}\widetilde{y}_i + \sum_{i\in U_h}\widetilde{y}_i - \sum_{i\in U_h} y_i\right)^2\right\}$$

$$V\left(\sum_{i\in B_h} w_{ib}\widetilde{y}_i\right) + E\left\{\left(\sum_{i\in U_h}\widetilde{y}_i - \sum_{i\in U_h} y_i\right)^2\right\}$$

이다. 그러면, $\widetilde{y}_i = (x_{1bi} - \beta_0)/\beta_1$ 이고 $y_i = (x_{1bi} - \beta_0 - e_{1i})/\beta_1$,

$$V\left(\sum_{i\in B_h} w_{ib}\widetilde{y}_i\right) = \beta_1^{-2} V\left(\sum_{i\in B_h} w_{ib}x_{ib}\right),$$

$$E\left\{\left(\sum_{i\in U_h}\widetilde{y}_i - \sum_{i\in U_h} y_i\right)^2\right\} = \beta_1^{-2} E\left\{\left(\sum_{i\in U_h} e_i\right)^2\right\} = \beta_1^{-2} K_h\sigma_{e1}^2$$

이다.

위의 식에서 K_h는 N_h(지역 h의 모집단 크기)의 함수로서, 지역 내 동질성을 측정한다. 만약 e_i가 독립이면 $K_h = N_h$이다. e_{1i}가 관측치 x_{ib}와 관련된 측정오차를 나타내기 때문에 측정오차가 독립이고 $K_h = N_h$를 가정하는 것은 타당하다. 따라서 만약 $\widehat{\sigma_{e1}^2}$을 이용할 수 있다면, $\widehat{Y}_{hb}$의 MSE는 다음과 같이 계산될 수 있다.

$$\widehat{MSE}(\widehat{Y}_{hb}) = \widehat{\beta_1^{-2}}\,\hat{V}(\widehat{X_{1hb}}) + \widehat{\beta_1^{-2}} K_h\widehat{\sigma_{e1}^2}.$$

이때 $\hat{V}(\hat{X}_{1hb})$는 $\hat{X}_{1hb} = \sum_{i\in B_h} w_{ib}x_{1i}$ 의 분산에 대한 디자인 일치 추정량 (design consistent estimator)이다.

이제 σ_{e1}^2을 추정해 보자. 두 조사의 독립성에 의해 다음과 같은 식을 이용할 수 있다.

$$E\left\{(\widehat{Y_{ha}}-\widehat{Y_{hb}})^2\right\}= V(\widehat{Y_{ha}})+MSE(\widehat{Y_{ha}})$$
$$= V(\widehat{Y_{ha}})+\beta_1^{-2}V(\widehat{X_{1hb}})+\beta_1^{-2}K_h\sigma_{e1}^2$$

적률방법에 의하면

$$\widehat{\sigma_{e1}^2}=\sum_{h=1}^{H}k_hK_h^{-1}\left\{\beta_1^2\left(\widehat{Y_{ha}}-\widehat{Y_{hb}}\right)^2-\beta_2^2\hat{V}(\widehat{Y_{ha}})-\hat{V}(\widehat{X_{hb}})\right\} \tag{3.3}$$

이고, 여기서 k_h는 부모집단 h에 의해 할당된 가중치를 나타낸다. 간단하게는 $k_h=n_{ha}/n_a$를 이용할 수 있고, n_{ha}는 조사 A 의 모집단에 포함된 표본크기 $n_a=\sum_{h=1}^{H}n_{ha}$이다. Rao (2003)은 k_h에 대해서 보다 복잡한 형태를 논의하였다.

이제 모형에서의 모수 β_0와 β_1을 추정해 보자. 이를 위해 y_i는 조사 A로부터 관측된 값이고 x_{1bi}는 조사 B로부터 관측된 값이다. 두 조사가 서로 연계되지 않는다면 2절에서 설명한 모형에 대해 다음과 같이 다시 쓸 수 있다.

$$X_{1h}=N_h\beta_0+Y_h\beta_1+\widetilde{e_{1h}} \tag{3.4}$$

여기서 $(N_h,X_{1h},Y_h,\tilde{e}_{1h})=\sum_{i\in U_h}(1,x_{1i},y_i,e_{1i})$, $\tilde{e}_{1h}\sim(0,N_h\sigma_e^2)$. 모형 (3.4)를 구조모형이라 하고, 이는 두 잠재변수 Y_h와 X_{1h}에 대한 구조적 관계를 나타낸다. 모수를 추정하기 위해 또 다른 모형을 세워보자. 즉 표본오차 모형은 식 (3.5)와 같이 설명될 수 있다.

$$\begin{pmatrix}\widehat{Y_{ha}}\\ \widehat{X_{1hb}}\end{pmatrix}=\begin{pmatrix}Y_{ha}\\ X_{1hb}\end{pmatrix}+\begin{pmatrix}N_ha_h\\ N_hb_h\end{pmatrix} \tag{3.5}$$

이고, a_h와 b_h는 각각 $\hat{Y}_{ha}/N_h$ 와 $\hat{X}_{1hb}/N_h$에 관련된 표본오차를 나타낸다.

식 (3.4)를 모평균의 개념으로 다시 쓸 수 있다.

$$\overline{X_{1h}} = \beta_0 + \overline{y_h}\beta_1 + \overline{e_{1h}} \tag{3.6}$$

여기서 ,

$$(\overline{X_{1h}}, \overline{Y_h}, \overline{e_{1h}}) = N_h^{-1}\sum_{i \in U_h}(x_{1i}, y_i, e_{1i}),\ \overline{e_{1h}} \sim (0, \sigma_{e1}^2/N_h),\ (\overline{y_{ha}}, \overline{x_{1hb}}) = N_h^{-1}(\widehat{Y_{ha}}, \widehat{X_{1hb}})$$

이라고 하면,

$$\begin{pmatrix} \widehat{y_{ha}} \\ \widehat{x_{1hb}} \end{pmatrix} \sim \left[\begin{pmatrix} \overline{Y_h} \\ \overline{X_{1hb}} \end{pmatrix}, \begin{pmatrix} \hat{V}(a_h) & 0 \\ 0 & \hat{V}(b_h) \end{pmatrix} \right].$$

따라서 식 (3.5)와 (3.6)을 결합하면, 다음과 같은 식을 얻을 수 있다.

$$\overline{x_{1hb}} = \beta_0 + (\overline{y_{ha}} - a_h)\beta_1 + \overline{e_{1h}} + b_h.$$

모수 (β_0, β_1)의 일치추정량은 다음의 함수 Q를 최소화하여 구할 수 있다.

$$Q(\beta_0, \beta_1) = \sum_{h=1}^{H} \frac{(\overline{x_{1hb}} - \beta_0 - \beta_1 \overline{y_{ha}})^2}{\beta_1^2 \hat{V}(a_h) + V(\overline{e_{1h}}) + V(b_h)} \tag{3.7}$$

이때 $V(\overline{e}_{1h}) = N_h^{-1}\sigma_{e1}^2$은 매우 작게 될 것이고, 그러면 식 (3.7)에서 $V(\overline{e}_{1h})$을 무시할 수 있다. 따라서 식 (3.7)은 식 (3.8)과 같이 표현될 수 있다.

$$Q(\beta_0,\beta_1)=\sum_{h=1}^{H}\frac{(\overline{x}_{1hb}-\beta_0-\beta_1\overline{y}_{ha})^2}{\beta_1^2\widehat{V}(a_h)+V(a_h)}=\sum_{H=1}^{h}\omega_h\frac{(\overline{x}_{1hb}-\beta_0-\overline{y}_{ha}\beta_1)^2}{\beta_1^2} \tag{3.8}$$

여기서 $\omega_h=1/\left\{V(a_h)+\beta_1^{-2}V(a_h)\right\}\cong 1/V(a_h)$. 따라서 $\beta_1^*=1/\beta_1,\ \beta_0^*=-\beta_0/\beta_1$이라고 하면, 식 (3.8)은 다음과 같이 표현될 수 있다.

$$Q(\beta_0^*,\beta_1^*)=\sum_{h=1}^{H}\omega_h(\overline{y_{ha}}-\beta_0^*-\overline{x_{1hb}}\beta_1^*)^2,$$

이것을 풀면 그 추정치는,

$$\beta_0^*=\left\{\sum_{h=1}^{H}\omega_h\right\}^{-1}\sum_{h=1}^{H}\omega_h(\overline{y_{ha}}-\beta_1^*\overline{x_{1h}}),$$

$$\beta_1^*=\left\{\sum_{h=1}^{H}\omega_h(\overline{x_{1hb}}-\overline{x_{1bw}})^2\right\}^{-1}\sum_{h=1}^{H}\omega_h(\overline{x_{1hb}}-\overline{x_{1bw}})(\overline{y_{ha}}-\overline{y_{aw}}),$$

이고, 여기서 $(\overline{x}_{1bw},\overline{y}_{aw})=\sum_{h=1}^{H}\omega_h(\overline{x}_{1hb},\overline{y}_{ha})/\sum_{h=1}^{H}\omega_h$. 그러면, 식 (3.8)을 최소로 하는 해는 $\widehat{\beta_1}=1/\widehat{\beta_1^*},\ \widehat{\beta_0}=-\hat{\beta}_0^*/\hat{\beta}_1^*$이 된다. 이때 $\omega_h=1/V(a_h)$을 사용하는 대신에 $\omega_h=n_{ha}$를 직접적으로 사용할 수 있다.

2. 실업급여등록자료 결합

이제 또 다른 보조정보가 있는 경우를 고려해 보자. 게다가 이 정보는 두 조사와 독립적인 것으로 행정등록에 의한 등록자료에 해당한다. 등록자료의 경우, 예를 들어 $g=1,\cdots,G$는 등록자료에서 이용될 수 있는 성별*연령별 그룹이라고 하자. 따라서 $\overline{x_{2hg}}$는 등록자료에서 얻어진 정보로서 부모집단 h에서 그룹 g에서 y에 대한 관측치의 평균값을 나타낸다. 즉, $\overline{x_{2hg}}$는 h지역 내 성*연령별 그룹에서의 실업급여등록율 (=실업급여등록자수/총피보험자수)로 정의된다.

$\overline{Y}_{hg}$를 부모집단 h에서 그룹 g에 대한 y의 평균이라고 하고, $\overline{y_{hg,a}}$는 조사 A 로부터 얻어진 $\overline{Y}_{hg}$의 추정치라고 하자. 그러면 다음과 같은 구조오차 모형을 세울 수 있다.

$$\overline{x_{2hg}} = \alpha_0 + \alpha_1 \overline{Y_{hg}} + u_{hg} \tag{3.9}$$

여기서, $u_{hg} \sim (0, \sigma_u^2)$이고, u_{hg}는 $\overline{x}_{2hg}$에 대해 불완전한 커버리지와 관련된 오차를 나타낸다. 이제 p_{hg}를 부차모집단 h내의 그룹 g의 모집단 비율이라고 하자. 조사 A에 대한 $\overline{Y}_h$의 직접추정량은 $\overline{y}_{ha} = N^{-1} \sum_{I \in A_h} w_{ia} y_i$이다. 행정자료에 대한 $\overline{Y}_h$의 합성추정량은

$$\overline{y_{hc}} = \sum_{g=1}^{G} p_{hg} \overline{y_{hg,c}} = \sum_{g=1}^{G} p_{hg} \alpha_1^{-1} (\overline{x_{2hg}} - \alpha_0).$$

2.1절에서와 같은 개념들을 이용하면 $\overline{y_{hc}}$의 MSE는 다음과 같이 구할 수 있다.

$$MSE(\overline{y_{hc}}) = \sum_{g=1}^{G} p_{hg}^2 \alpha_1^{-2} \sigma_u^2.$$

또한 $\hat{Y}_{hc} = N_h \overline{y}_{hc}$과 $\widehat{MSE}(\hat{Y}_{hc}) = N_h^2 \widehat{MSE}(\overline{y}_{hc})$. 만약 $\widehat{MSE}(\hat{Y}_{hc})$를 이용할 수 있다면 위의 세 개의 정보를 결합한 복합추정량은 다음과 같이 구할 수 있다.

$$\widehat{Y_h^*} = \frac{\widehat{Y_{ha}} / \hat{V}(\widehat{Y_{ha}}) + \widehat{Y_{hb}} / \widehat{MSE}(\widehat{Y_{hb}}) + \widehat{Y_{hc}} / \widehat{MSE}(\widehat{Y_{hc}})}{1/\hat{V}(\widehat{Y_{ha}}) + 1/\widehat{MSE}(\widehat{Y_{hb}}) + 1/\widehat{MSE}(\widehat{Y_{hc}})} \tag{3.10}$$

식 (3.10)으로부터 계산된 복합추정량의 MSE는

$$MSE(\widehat{Y^*}) = (1/\hat{V}(\widehat{Y_{ha}}) + 1/\widehat{MSE}(\widehat{Y_{hb}}) + 1/\widehat{MSE}(\widehat{Y_{hc}}))^{-1},$$

이고, 식 (3.10)을 구하기 위해 우리는 σ_u^2를 추정할 필요가 있다. 식 (3.3)과 유사한 방법

을 사용하면 σ_u^2의 추정치는

$$\widehat{\sigma_u^2} = \sum_{h=1}^{H} k_h C_h^{-1}\left\{(\overline{y_{ha}} - \overline{y_{hc}})^2 - \hat{V}(\overline{y_{ha}})\right\} , \tag{3.11}$$

여기서 $C_h = \sum_{g=1}^{G} p_{hg}^2 \alpha_1^{-2}$이다.

이제 계수 α_0와 α_1을 추정하기 위해 식 (3.9)와 $\bar{y}_{ha} = \hat{Y}_{ha} + a_h$을 결합하여 다음의 식 (3.12)를 얻을 수 있다.

$$\overline{x_{2h}} = \alpha_0 + \alpha_1 \overline{y_{ha}} - \alpha_1 a_h + u_h. \tag{3.12}$$

여기서 $\bar{x}_h = \sum_{g=1}^{G} p_{hg} \bar{x}_{2hg}$, $a_h \sim (0, \hat{V}(a_h))$이고 $u_h = \sum_{g=1}^{G} p_{hg} u_{hg} \sim (0, C_h \alpha_1^2 \sigma_u^2)$이다. 따라서 $(\alpha_{0,}\ \alpha_1)$의 일치추정량은 다음의 함수 Q를 최소화하는 값으로 얻을 수 있다.

$$Q(\alpha_0, \alpha_1) = \sum_{h=1}^{H} \frac{(\overline{x_{2h}} - \alpha_0 - \alpha_1 \overline{y_{ha}})^2}{\alpha_1^2 (\hat{V}(a_h) + C_h \sigma_u^2)}. \tag{3.13}$$

$\alpha_1^* = 1/\alpha_1$, $\alpha_0^* = -\alpha_0/\alpha_1$, $\omega_h = 1/(\hat{V}(a_h) + C_h \hat{\sigma}_u^2)$라고 하면, (3.13)은 다음과 같이 다시 표현될 수 있다.

$$Q(\alpha_0, \alpha_1) = \sum_{h=1}^{H} \omega_h (\overline{y_{ha}} - \alpha_0^* - \alpha_1^* \overline{x_{2h}})^2. \tag{3.14}$$

따라서 이를 풀면 추정치는 각각

$$\widehat{\alpha_0^*} = \left\{\sum_{h=1}^{H} \omega_h\right\}^{-1} \sum_{h=1}^{H} \omega_h (\overline{y_{ha}} - \alpha_1^* \overline{x_{2h}}),$$

$$\alpha_1^* = \left\{ \sum_{h=1}^{H} \omega_h (\overline{x_{2h}} - \overline{x_{2w}})^2 \right\}^{-1} \sum_{h=1}^{H} \omega_h (\overline{x_{2h}} - \overline{x_{2w}})(\overline{y_{ha}} - \overline{y_w})$$

여기서 $(\overline{x}_{2w}, \overline{y}_w) = \sum_{h=1}^{H} \omega_h (\overline{x}_{2h}, \overline{y}_{ha}) / \sum_{h=1}^{H} \omega_h$. 식 (15)를 최소하기 위해 $\widehat{\alpha_1} = 1/\widehat{\alpha}^*$, $\widehat{\alpha_0} = -\widehat{\alpha_0^*}/\alpha_1^*$를 사용할 수 있다.

그렇지만 실제 활용상에 있어서 식 (3.11)로부터 σ_u^2의 해를 찾기가 쉽지 않을 수 있을 것이다. 이런 경우에는 이분법 (bisection method)와 같은 Grid search algorithm 을 이용하면 쉽게 구할 수 있다. 대부분의 상용패키지에서 쉽게 사용할 수 있고, R프로그램의 경우, “bisection.method” 와 같은 함수를 이용할 수 있다. 게다가 우리가 추정해야 할 모수 α_0, α_1, C_h, σ_u^2, ω_h가 서로의 함수관계에 있기 때문에, 우선 $\omega_h = 1/\hat{V}(a_h)$로 초기값으로 사용하여 2~3번 반복한 값을 이용할 수 있다.

3. 시험적 추정결과

가. 지역수준 모형에 기반한 조사자료와 실업급여등록자료 복합추정

① 지역수준(area level) 에서의 분석자료 세팅

- 경활조사, 고용조사의 직접 통계량값과 그 분산을 입력한 지역수준 자료 생성
- 등록자료는 성(남/여)·연령(5세 단위 10개 범주) 그룹별 인구수, 실업급여등록자 생성

② 각각의 모수 추정 $(\alpha,\ \beta, \sigma^2)$

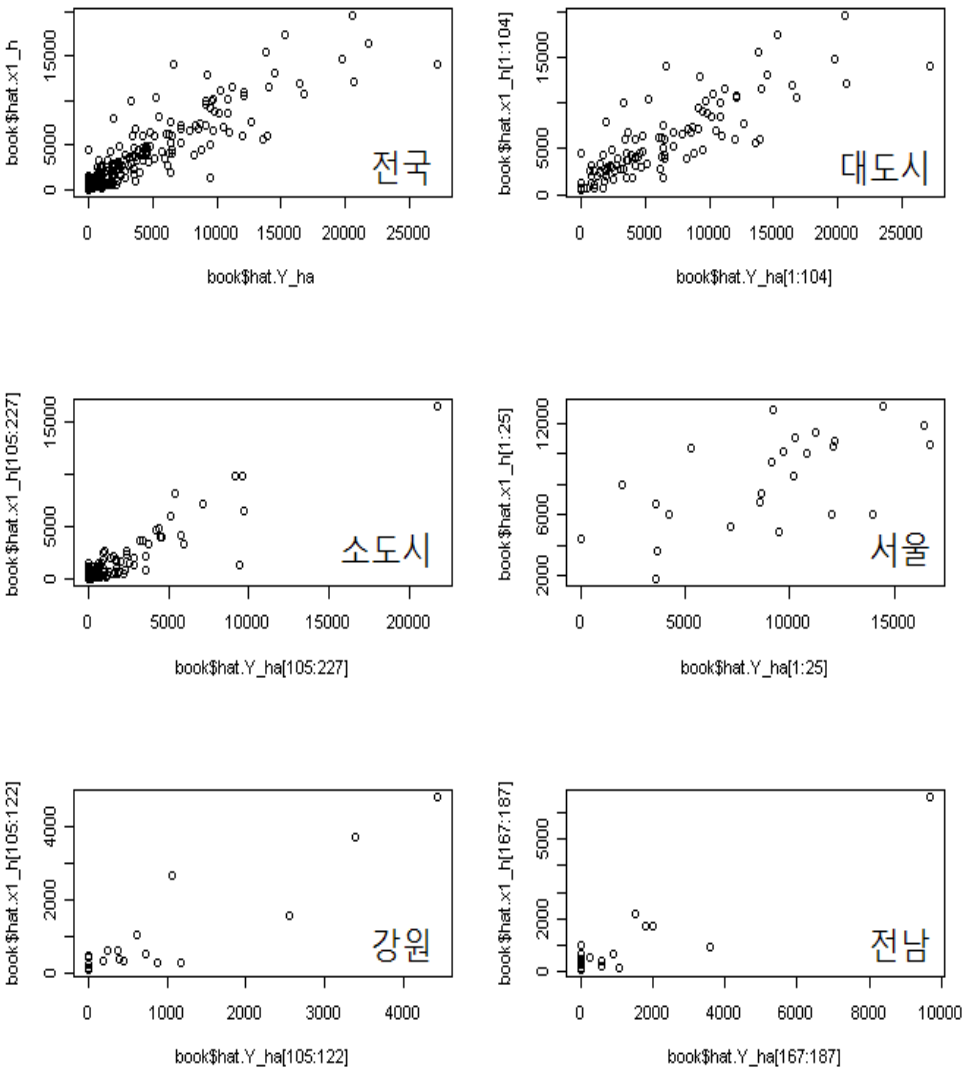
③ 합성 추정량 구현 - 지역별고용조사에 대한 합성추정량($\widehat{Y_{hb}}$), 실업급여등록자료에 대한 합성추정량($\widehat{Y_{hc}}$)

$$\hat{Y}_{hb} = \hat{\beta}_1^{-1}(\widehat{X_{hb}} - N_h\hat{\beta}_0) = N_h\beta_0^* + \beta_1^*\widehat{X_{hb}} = N_h * 0.005 + 0.773 * \widehat{X_{hb}}$$

$$\hat{Y}_{hc} = N_h \sum_g^G p_{hg}\alpha_1^{-1}(\overline{x}_{2hg} - \alpha_0) = \sum_g^G p_{hg} N_h (0.019 - 0.103 * \widehat{X_{hb}})$$

④ 복합 추정량 구현 ($\widehat{Y^*}$)

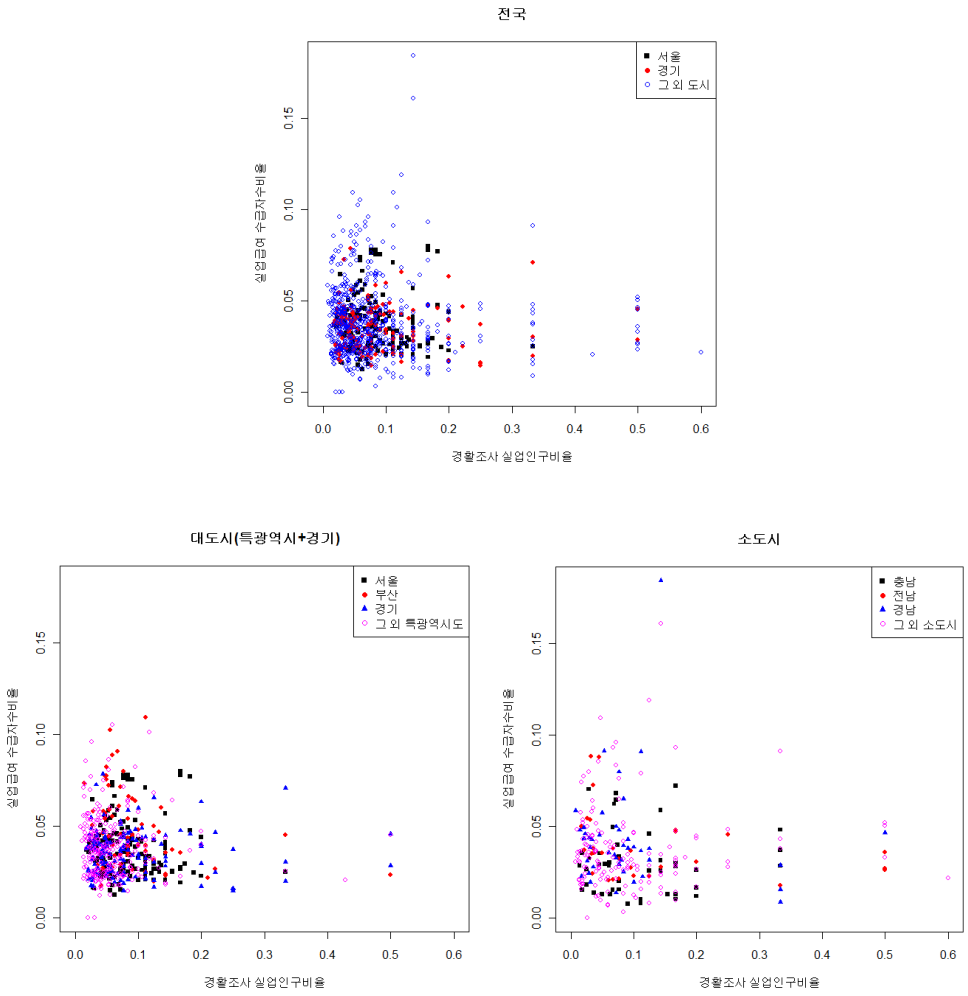
나. 경활조사자료 VS. 지역고용조사, 실업급여등록의 상관성



[그림 1-1] 경활조사와 고용조사의 실업자수에 대한 산점도

[그림 1-1]에서 y축은 경활조사 실업자수, x축은 지역별고용조사로부터 얻어진 실업자수 추정치를 각각 나타낸다. 다음 그림에서 전국 수준과 지역별 수준에서의 구조적 관계가 유사하고, 양의 상관관계를 갖는 것을 알 수 있다. 그림에서 보면 규모가 작은 지역

으로 갈수록 즉, 대도시에 비해 소도시에서 그 관계성이 상대적으로 약해지는 경향이 있다.



[그림 1-2] 경활조사실업인구비율과 실업급여등록자 비율과의 산점도

[그림 1-2]는 경활조사에 의한 실업인구비율과 실업급여등록자비율과의 산점도를 나타낸 것이다. 실제로 실업급여등록자료에 대한 모형을 세울 때 이 등록자료가 갖는 모집단 포괄 문제를 극복하기 위하여 전국 인구비율로 조정한 실업급여등록자비율자료를 이

용하였다. 전국단위의 경활조사실업률과 실업급여등록률 간에 강한 상관관계가 존재하지 않고, 특히 소도시로 갈수록 두 변수간의 구조적 관계는 거의 존재하지 않는다. 따라서 이러한 관계 현상은 모형추정을 어렵게 만드는 요인이 될 수 있다. 비율자료가 아닌 총수 자료를 이용할 경우에는 두 변수간의 전국 단위에서 강한 관계가 존재하지만, 이것을 모집단 인구 비율로 표준화하게 되면 그 관계정도가 약해지는 경향이 나타나게 된다. 대도시와 소도시로 나누어 추정을 시도하였다.

다. 추정결과

경활조사 직접추정량 :Y_ha,

지역별고용조사 직접추정량 : x1_h,

지역별고용조사 합성추정량 : synth.x1_h,

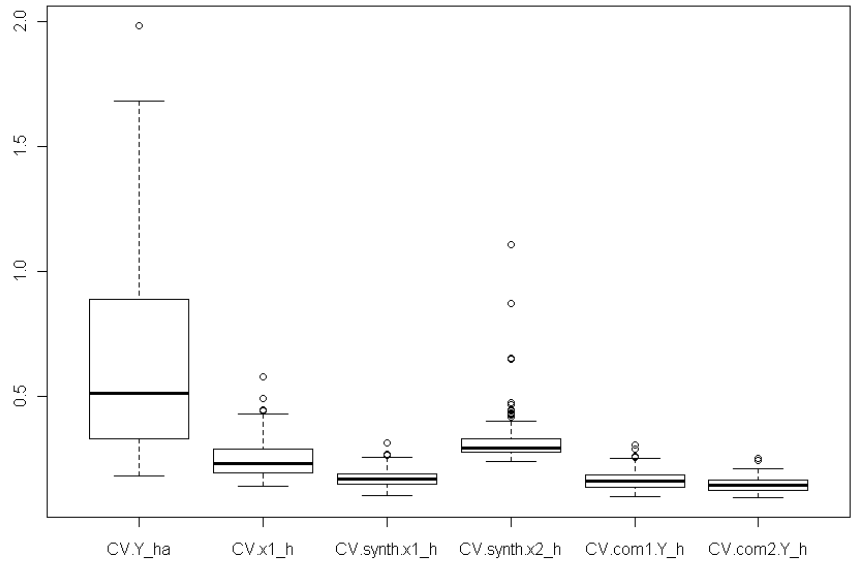
실업급여등록자료 합성추정량 : synth.x2_h,

com1.Y_h: 경활조사직접추정+지역별고용조사고용조사 합성추정,

com2.Y_h:경활조사직접추정+지역별고용조사합성추정+실업급여등록자료 합성추정

[그림 1-3]은 각 추정량들에 대한 상자그림을 나타낸 것이다. [그림 1-3]에서 보면, 모든 추정량은 경활직접추정량의 CV보다 작고, 직접추정량을 제외한 다른 추정량들에서는 등록자료의 합성추정량(synth.x2_h)의 CV가 약간 큰 것을 알 수 있다. 복합추정량의 경우, com1.Y_h 의 CV가 경활직접추정량의 CV를 크게 개선시키고 있다. 이는 com1의 경우 이 복합추정량은 대체로 고용조사의 영향을 상대적으로 많이 받는 것으로 파악되었으며, 실제로 추정치가 x1_h에 상당한 유사한 것으로 나타났다.

여기에 등록자료를 더 연결하여 추정한 결과가 com2에 해당되고, 이 추정량 com2는 com1의 CV를 더욱 향상시키는 결과를 보여준다. com2의 CV는 대부분 지역에서 25% 보다 작게 나타났다. 따라서 본 연구에서 제안한 방법으로 우리나라의 시군구 실업자수를 추정할 경우, 조사자료와 등록자료를 이용한 복합추정량이 직접추정량에 대해 CV측면에서 효과가 있다고 볼 수 있다.



[그림 1-3] 추정량들의 CV

그러나 우리가 제안한 방법인 com2는 com1에 비해 과소 추정하는 경향이 있다. 이는 실업급여등록자료로부터 구한 합성추정량이 과소 추정되면서 발생하는 효과로 판단된다. 이러한 효과는 대도시일수록 더 크게 나타난다. 이와 같은 현상은 실업급여등록자료와 경활조사의 실업자수 자료 간의 관계가 대도시와 소도시에 서로 다른 경향을 보이고, 그 관계성이 소도시에서 더 약하다는 점에서 비롯된다고 볼 수 있다. 따라서 소도시와 대도시의 특성을 구분하여 추정을 시도해 보았다. 두 그룹을 구분하는 기준은 간단하게 특광역시와 도단위로 나누었다. 이에 대한 구분 방법은 시구/군 또는 도시/농촌 등으로 나누어 볼 수도 있다.

<표 1-1>은 추정방법에 따른 시도단위의 실업자 총수를 나타낸다. <표 1-1>의 “그룹” 칼럼은 전국을 두 개 그룹으로 나누어 추정한 결과이다. 이처럼 두 개 그룹으로 나누어 추정하면, 대도시 지역의 추정치는 그렇지 않은 경우에 비해 경활조사에서 더 근사해지는 것을 알 수 있다. 그렇지만 소도시 지역에서는 오히려 추정치간의 차이가 더 커지는 것을 알 수 있다. 결국 소도시에서는 전국 자료를 이용해서 추정하는 것이 경활조사 추정치에 더 근사해진다고 볼 수 있다.

따라서 실업급여등록자료를 결합하는 과정에서 발생하는 복합추정량의 과소 추정문제를 해결하기 위해 다양한 방법을 시도하였다. 그렇지만 제3절에서 제안한 방법으로는 모형의 모수 추정이 여전히 불안정할 뿐만 아니라 그 추정치들 간의 차이를 줄이기는 쉽지 않았다. 따라서 본 연구의 제3절에서 제시한 실업급여등록자료를 연결하는 복합추정량에 관한 내용은 향후 개선과제로 남기기로 하고, 제4절에서는 실업급여등록자료를 제외하고 경활조사와, 지역별고용조사, 센서스 자료를 연결한 복합추정량을 제안하기로 하였다.

〈표 1-1〉 추정방법별 시도단위의 실업자 총수

region	경활(Y_ha)	지역고용(x1)	com1.Y_h	com2.Y_h	그룹
서울	225,000	208,226	199,434	166,842	192,941
부산	58,000	54,183	54,049	49,541	57,136
대구	46,000	50,027	47,823	40,783	47,027
인천	68,000	64,265	61,705	49,806	57,503
광주	20,000	22,601	22,074	19,993	22,894
대전	18,000	21,078	20,232	19,748	22,133
울산	19,000	17,943	18,237	16,898	19,048
경기	219,000	166,772	174,245	168,539	187,982
강원	17,000	18,580	19,349	18,839	18,839
충북	15,000	18,410	18,502	18,600	18,600
충남	33,000	28,549	29,482	28,568	28,568
전북	18,000	22,046	23,330	22,967	22,967
전남	16,000	18,032	20,121	19,778	19,778
경북	35,000	31,115	34,188	33,422	33,422
경남	48,000	38,739	42,340	41,546	41,546
제주	5,000	5,427	5,926	6,068	6,068
총합	857,000	785,993	791,037	721,938	796,482

제4절 조사자료와 센서스 자료를 이용한 복합추정

3절에서는 등록자료를 보조변수로 활용하여 소지역 추정하는 방법을 시도하였다. 시군구 단위의 실업자 추정에 있어서 시군구별 실업급여등록자료는 현재 우리나라의 고용상황을 파악할 수 있는 좋은 등록 정보인 것은 사실이다. 그러나 측정오차 모형을 사용함에 있어서 이 자료에 포함되어 있는 커버리지 오차 또는 등록오차 등은 극복하기 어려운 문제라 여겨지고, 실제 연구에서 제시한 방법으로는 모형 적합에 있어서도 그다지 좋은 결과를 주지 못했다. 따라서 본 절에서는 등록자료를 보조정보로 사용하지 않는 방법을 재설계하고자 한다. 전체적으로 3절에서 정리한 내용과 유사하나 모형을 세팅하는 과정에서 가정사항 등을 약간 다르게 사용하였다.

3절과 달라진 점은 크게 두 가지이다. 3절에서는 ① 측정오차 모형을 세팅할 때 경활조사를 y로, 지역별고용조사를 x로 정의하였다. 이제는 이를 바꾸어서 측정오차모형에서 종속변수를 지역별고용조사 y, 설명변수를 경활조사값 x로 정의하였다. 이렇게 한 이유는 센서스 조사값도 경활조사값 x에 쉽게 연결할 수 있도록 하기 위해서이다. ② 또한 3절에서와는 달리 경활조사와 지역고용조사가 독립적이지 않고 표본 오차에 대한 공분산이 존재한다고 가정하였다. 이는 경활조사 표본이 지역고용조사 표본의 부분이라는 점에 근거하고 있다. ③ 여기서는 실업자수 추정이 아닌 실업률 추정을 위한 모형으로 접근하였다. 총수 추정에 대한 민감성을 고려할 때 실업률을 직접 추정함으로써 실업률 추정에 대한 정도를 더 높일 수 있다고 판단하였다. ④ 이렇게 함으로써 3절에 비해 예측 및 추정 공식이 더 쉽고 확장이 편리하도록 하였다.

1. 기본이론

두 조사, A와 B가 있다고 하자. 조사 A로부터 x_i 를 관측하고, 조사 B로부터 y_{1i} 를 관측한다고 하자. 이때 y_{1i} 은 상당한 수준의 측정오차 등을 포함하고, x_i 는 적은 수준의 측정오차를 포함한다고 할 때, 다음의 모형을 가정할 수 있다.

$$y_{1i} = \beta_0 + \beta_1 x_i + e_{1i} \tag{4.1}$$

여기서 e_{1i} 는 평균 0, 분산 σ_{e1}^2 인 확률변수이다. 선형회귀모형에 필요한 가정 또는 등분산가정들에 대해서는 이후에 완화될 수 있다. 만약 회귀계수에 해당하는 모수

$(\beta_0,\beta_1)=(0,1)$이면 모형 (4.1)은 측정오차가 없다는 것을 의미한다.

이때 이 두 조사는 별개의 확률표본설계에 의한 것이다. 조사 A로부터 설계기반 불편추정량에 해당하는 $\widehat{X}_h=\sum_{i\in A_h} w_{ia}x_i$ 와 분산추정량 $\hat{V}(\widehat{X}_h)$ 을 얻을 수 있고, 조사 B로부터 $Y_{1h}=\sum_{i\in U_h} y_{1i}$ 에 대한 설계-모형기반 불편추정량 $\widehat{Y_{1h}}=\sum_{i\in B_h} w_{ib}y_{1i}$ 을 얻을 수 있다. 두 추정량의 $(\widehat{X}_h,\widehat{Y}_h)$의 표본오차는 다음의 표본오차 모형으로 표현될 수 있다.

$$\begin{pmatrix}\widehat{X}_h\\ \widehat{Y_{1h}}\end{pmatrix}=\begin{pmatrix}X_h\\ Y_{1h}\end{pmatrix}+\begin{pmatrix}N_h a_h\\ N_h b_h\end{pmatrix} \tag{4.2}$$

이때 a_h, b_h은 $\widehat{X}_h/N_h$, $\widehat{Y}_h/N_h$에 관련된 표본오차를 나타낸다.

두 조사로부터 (x_{1i},y_i) 를 동시에 얻을 수 없다면, 즉 두 조사의 레코드 매칭이 어렵다면, 우리는 식 (2.3b)의 측정오차 모형으로부터 다음과 같은 지역수준 모형을 유도할 수 있다.

$$Y_{1h}=N_h\beta_0+\beta_1 X_h+\widetilde{e_{1h}} \tag{4.3}$$

여기서 $(N_h,X_{1h},Y_h,\tilde{e}_{1h})=\sum_{i\in U_h}(1,x_{1i},y_i,e_{1i})$이고, $\tilde{e}_{1h}\sim(0,N_h\sigma_e^2)$을 따른다. 모형 (4.3)을 구조오차 모형이라 부른다. 이는 모형이 두 잠재변수 Y_h와 X_h에 대한 구조적 관계를 설명하기 때문이다. 모형 (4.3)을 모집단 평균 개념으로 다시 표현하면

$$\overline{Y_{1h}}=\beta_0+\widehat{X}_h\beta_1+\overline{e_{1h}}, \tag{4.4}$$

이고 여기서 $(\overline{X}_h,\overline{Y}_{1h},\overline{e}_{1h})=\sum_{i\in U_h}(x_i,y_{1i},e_{1i})/N_h$이다. 만약 우리가 내포오차 모형 (nested error model)

$$e_{1hi} = \epsilon_h + u_{hi}, \tag{4.5}$$

을 사용하면, $\bar{e}_{1h} \sim (0, \sigma_e^2 + \sigma_u / N_h)$이다. 여기서 $\epsilon \sim (0, \sigma_e^2)$이고, $u_{hi} \sim (0, \sigma_u^2)$을 따른다. 내포오차 모형은 소지역 추정에서 보편화된 모형 (Battese 등, 1988)이고 $Cov(e_{1hi}, e_{1hj}) = \sigma_e^2,\ i \neq j$. 대체로 N_h는 상당히 크기 때문에 $\bar{e}_{1h} \sim (0, \sigma_e^2)$을 가정할 수 있게 된다.

이제 $(\bar{y}_{1h}, \bar{x}_h) = N_h^{-1}(\hat{Y}_{1h}, \hat{X}_h)$라 하고 표본오차 모형 (4.2)와 구조오차 모형 (4.4)를 결합하면 다음과 같은 형식을 얻을 수 있고,

$$\begin{pmatrix} \overline{y_{1h}} \\ \overline{x_h} \end{pmatrix} = \begin{pmatrix} \beta_0 & \beta_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \overline{X_h} \end{pmatrix} + \begin{pmatrix} b_h + \overline{e_{1h}} \\ a_h \end{pmatrix},$$

이는 다시 모형 (4.6)과 같이 표현될 수 있다.

$$\begin{pmatrix} \overline{y_{1h}} - \beta_0 \\ \overline{x_h} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ 1 \end{pmatrix} \overline{X_h} + \begin{pmatrix} b_h + \overline{e_{1h}} \\ a_h \end{pmatrix} \tag{4.6}$$

따라서 모형 (4.6)에 포함된 모든 모수들이 알려져 있다면, $\overline{X}_h$의 최량 추정량 (best estimator)은 (4.7)과 같이 계산될 수 있다.

$$\widehat{\overline{X}}_h = \left\{ (\beta_1, 1) V_h^{-1} (\beta_1, 1)' \right\}^{-1} (\beta_1, 1) V_h^{-1} (\overline{y_{1h}} - \beta_0, \overline{x_h})' \tag{4.7}$$

여기서 V_h는 $(b_h + \bar{e}_{1h}, a_h)$의 분산-공분산 행렬을 나타낸다. 식 (4.7)은 선형모형이론의 일반화최소제곱법(generalized least sqaures method)을 사용한다는 점에서 이 추정량은 GLS 추정량이라 할 수 있다. GLS 방법은 다양한 정보를 쉽게 이용할 수 있다는 점에서 유용하다. 예를 들면 $\overline{Y}_{2h}$에 대한 또 다른 추정량 $\bar{y}_{2h}$를 사용할 수 있다면 이 추정량은 다음을 만족한다.

$$\overline{Y_{2h}} = \gamma_0 + \gamma_1 \overline{X_h} + \overline{e_{2h}}$$

이것을 GLS 모형에 삽입하면 식 (4.8)과 같고

$$\begin{pmatrix} \overline{y_{2h}} - \gamma_0 \\ \overline{y_{1h}} - \beta_0 \\ \overline{x_h} \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} \overline{X_h} + \begin{pmatrix} c_h + \overline{e_{2h}} \\ b_h + \overline{e_{1h}} \\ a_h \end{pmatrix} \tag{4.8}$$

이때 GLS 추정량은 식 (4.7)에서와 유사하게 구할 수 있다.

Remark 1. 모형 (4.6)은 (4.9)와 같이 표현할 수 있다.

$$\begin{pmatrix} \beta_1^{-1}(\overline{y_{1h}} - \beta_0) \\ \overline{x_h} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \overline{X_h} + \begin{pmatrix} (b_h + \overline{e_{1h}})/\beta_1 \\ a_h \end{pmatrix} \tag{4.9}$$

(4.9)로부터 계산된 GLS 추정량은 (4.6)에 의해 계산된 GLS 추정량과 같아지게 되고, 그 때의 GLS 추정량은 (4.10)과 같다.

$$\widehat{\overline{X}_h} = \alpha_h \overline{x_h} + (1 - \alpha_h) \widetilde{x_h} \tag{4.10}$$

여기서 $\tilde{x}_h = \beta_1^{-1}(\overline{y}_{1h} - \beta_0)$이고 α_h는 다음과 같다.

$$\alpha_h = \frac{V(\widetilde{x_h}) - Cov(\overline{x_h}, \widetilde{x_h})}{V(\overline{x_h}) + V(\widetilde{x_h}) - 2Cov(\overline{x_h}, \widetilde{x_h})}$$

$$= \frac{\sigma_{e,h}^2 + V(b_h) - \beta_1 C(a_h, b_h)}{\sigma_{e,h}^2 + V(b_h) + \beta_1^2 V(a_h) - 2\beta_1 C(a_h, b_h)}$$

회귀계수의 추정치 $\hat{\beta}=(\hat{\beta}_0,\hat{\beta}_1)$가 계산되면 이 추정량 $\widetilde{x_h}$을 합성추정량(synthetic estimator)이라 하고, (4.10)의 최적 추정량은 복합추정량(composite estimator)이라 부르기도 한다. 여기서 모수 β를 추정하는데, 만약 β를 추정하는 효과를 무시할 수 있다면 복합추정량의 분산은 (4.11)과 같아진다.

$$V(\hat{\overline{X}}_h-\overline{X}_h)=\alpha_h V(\overline{x}_h)+(1-\alpha_h)Cov(\overline{x}_h,\tilde{x}_h) \tag{4.11}$$

(4.11)에서 $V(\overline{x}_h)$는 직접추정량의 분산, $Cov(\overline{x}_h,\tilde{x}_h)$는 직접추정량과 합성추정량의 공분산을 나타낸다. $\alpha_h<1$면, 복합추정량은 직접추정량보다 더 효과적이게 된다.

2. 모수 추정

이제 모형 (4.4)의 모수추정에 대해 살펴보기로 한다. $\overline{X}_1,\cdots,\overline{X}_H$가 알려져 있다면 모수 (β_0,β_1)의 GLS 추정량은 다음의 함수 Q를 최소화하는 β를 말한다.

$$Q(\beta_0,\beta_1)=\sum_{h=1}^{H}\begin{pmatrix}\overline{y}_{1h}-\beta_0-\beta_1\overline{X}_h\\ \overline{x}_h-\overline{X}_h\end{pmatrix}'\begin{pmatrix}\sigma_{e,h}^2+V(b_h) & C(a_h,b_h)\\ C(a_h,b_h) & V(a_h)\end{pmatrix}^{-1}\begin{pmatrix}\overline{y}_{1h}-\beta_0-\beta_1\overline{X}_h\\ \overline{x}_h-\overline{X}_h\end{pmatrix} \tag{4.12}$$

실제로 $\overline{X}_1,\cdots,\overline{X}_H$ 을 알지 못하기 때문에 우리는 이에 대한 추정량, $\hat{\overline{X}}_h=\hat{\overline{X}}_h(\beta_0,\beta_1)$을 대신 이용한다. 즉, Q대신에 Q^*를 이용하여 β를 추정할 수 있다.

$$Q^*(\beta_0,\beta_1)=\sum_{h=1}^{H}\begin{pmatrix}\overline{y}_{1h}-\beta_0-\beta_1\hat{\overline{X}}_h\\ \beta_1(\overline{x}_h-\hat{\overline{X}}_h)\end{pmatrix}\left\{V\begin{pmatrix}\overline{y}_{1h}-\beta_0-\beta_1\hat{\overline{X}}_h\\ \beta_1(\overline{x}_h-\hat{\overline{X}}_h)\end{pmatrix}\right\}^{-1}\begin{pmatrix}\overline{y}_{1h}-\beta_0-\beta_1\hat{\overline{X}}_h\\ \beta_1(\overline{x}_h-\hat{\overline{X}}_h)\end{pmatrix} \tag{4.13}$$

이를 계산하면 (4.13)은 (4.14)와 같다.

$$Q^*(\beta_0,\beta_1)=\sum_{h=1}^{H}\frac{(\bar{y}_{1h}-\beta_0-\beta_1\bar{x}_h)^2}{V(\bar{y}_{1h}-\beta_0-\beta_1\bar{x}_h)} \tag{4.14}$$

식 (4.2), (4.4), (4.6)을 이용하여 다음과 같이 표현할 수 있고

$$\bar{y}_{1h}-\beta_0-\bar{x}_h\beta_1 = -a_h\beta_1+b_h+\bar{e}_{1h},$$

이것을 분산에 대해 표현하면 다음의 식 (4.15)와 같다.

$$V(\bar{y}_{1h}-\beta_0-\bar{x}_h\beta_1)=\sigma_{e,h}^2+(-\beta_1,1)\sum_h(-\beta_1,1)' \tag{4.15}$$

여기서 $\sigma_{e,h}^2=V(\bar{e}_{1h})$이고, $\Sigma_h=V\{(a_h,b_h)'\}$이다. (a_h,b_h)의 분산-공분산행렬에 대한 일치추정량(consistent estimator)을 구할 수 있기 때문에, 만약 $\sigma_{e,h}^2$을 알고 있다면 식 (4.14)의 $Q^*(\beta_0,\beta_1)$을 최소화하는 β를 구할 수 있다.

식 (4.14)와 식 (4.15)를 이용하여 $Q^*(\beta_0,\beta_1)$을 다시 쓰면

$$Q^*(\beta_0,\beta_1)=\sum_{h=1}^{H}w_h(\beta_1)(\overline{y_{1h}}-\beta_0-\beta_1\overline{x_h})^2 \tag{4.16}$$

이고, 여기서 $w_h(\beta_1)=\{\sigma_{e,h}^2+(-\beta_1,1)\Sigma_h(-\beta_1,1)'\}^{-1}$이다.

$$\frac{\partial}{\partial\beta_0}Q^*=0 \quad\Leftrightarrow\quad \sum_{h=1}^{H}w_h(\beta_1)(\overline{y_{1h}}-\beta_0-\beta_1\overline{x_h})=0,$$

이고,

$$\hat{\beta}_0=\overline{y_w}-\hat{\beta}_1\overline{x_w} \tag{4.17}$$

여기서 $(\bar{x}_w, \bar{y}_w) = \{\Sigma_{h=1}^{H} w_h(\hat{\beta}_1)\}^{-1} \Sigma_{h=1}^{H} w_h(\hat{\beta}_1)(\bar{x}_h, \bar{y}_h)$이다. 식 (4.17)에서 추정한 추정치 $\hat{\beta}_0$을 식 (4.16)에 대입하면

$$Q_1^*(\beta_1) = \sum_{h=1}^{H} w_h(\beta_1)\{\overline{y_{1h}} - \overline{y_w} - \beta_1(\overline{x_h} - \overline{x_w})\}^2 \tag{4.18}$$

이 된다. 따라서 식 (4.18)을 β_1에 대해 미분 $\partial Q_1^*/\partial\beta_1 = 0$ 하면

$$\frac{\partial}{\partial\beta_1} Q_1^* = \sum_{h=1}^{H} \left\{ \frac{\partial}{\partial\beta_1} w_h(\beta_1) \right\} \{\overline{y_{1h}} - \overline{y_w} - \beta_1(\overline{x_h} - \overline{x_w})\}^2$$
$$-2\sum_{h=1}^{H} w_h(\beta_1)(\overline{x_h} - \overline{x_w})\{\overline{y_{1h}} - \overline{y_w} - \beta_1(\overline{x_h} - \overline{x_w})\}$$

이고,

$$\frac{\partial}{\partial\beta_1} w_h(\beta_1) = -2\{w_h(\beta_1)\}^2\{\beta_1 V(a_h) - C(a_h, b_h)\}$$

과

$$\{\overline{y_{1h}} - \overline{y_w} - \beta_1(\overline{x_h} - \overline{x_w})\}^2 \to_p \sigma_{e,h}^2 + (-\beta_1, 1)\sum_h (-\beta_1, 1)' = 1/w_h(\beta_1)$$

을 사용하면

$\partial Q_1^*/\partial\beta_1 = 0$ 대한 최종 해는 (4.19)와 같다.

$$\hat{\beta}_1 = \frac{\sum_{h=1}^{H} w_h(\hat{\beta}_1)\{(\overline{x_h} - \overline{x_w})(\overline{y_{1h}} - \overline{y_{1w}}) - C(a_h, b_h)\}}{\sum_{h=1}^{H} w_h(\hat{\beta}_1)\{(\overline{x_h} - \overline{x_w})^2 - V(a_h)\}} \tag{4.19}$$

가중치 $w_h(\beta_1)$은 β_1에 의존하게 된다. 따라서 (4.19)는 반복 알고리즘을 사용하여 구할 수 있다. $\hat{\beta}_1$이 계산되면 $\hat{\beta}_0$은 식 (4.17)에 의해 구해진다.

이제 모형분산 $\widehat{\sigma_{e,h}^2}$을 추정하는 방법을 살펴보자. 가장 간단한 방법은 적률법(MOM, method of moment)을 이용하는 것이다. 즉, $e_{e,h}^2$의 불편추정량을 얻기 위해서 식 (4.20)을 이용할 수 있다.

$$E\{(\overline{y_{1h}} - \beta_0 - \overline{x_h}\beta_1)^2 - \beta_1^2 V(a_h) + 2\beta_1 C(a_h, b_h) - V(b_h)\} = \sigma_{e,h}^2 \tag{4.20}$$

내포오차 모형 (4.5)하에서 $\sigma_{e,h}^2 = \sigma_e^2$ 이고,

$$E\{(\overline{y_{1h}} - \beta_0 - \overline{x_h}\beta_1)^2 - \beta_1^2 V(a_h) + 2\beta_1 C(a_h, b_h) - V(b_h)\} = \sigma_e^2 \tag{4.21}$$

따라서 Fuller (2009)와 유사한 방법을 사용하면 σ_e^2의 MOM 추정량은 (4.22)와 같이 계산될 수 있다.

$$\hat{\sigma_e^2} = \sum_{h=1}^{H} \kappa_h \left\{ (\overline{y_{1h}} - \hat{\beta}_0 - \overline{x_h}\hat{\beta}_1)^2 - (-\hat{\beta}_1, 1) \sum_h (-\hat{\beta}_1, 1) \right\} \tag{4.22}$$

여기서

$$\kappa_h \propto (\hat{\sigma_e^2} + (-\hat{\beta}_1, 1)\Sigma_h(-\hat{\beta}_1, 1)')^{-1},$$

$\Sigma_{h=1}^{H}\kappa_h = 1$. 이때 κ_h는 $\widehat{\sigma_e^2}$에 의존하기 때문에 반복 계산에 의해 구할 수 있고, 초기값으로 $\widehat{\sigma_e^2}=0$을 사용한다. Fay 와 Herriot (1979)은 비선형방정식의 근을 구하기 위해 반복수행 알고리즘을 사용하였다.

$$\sum_{h=1}^{H}\frac{(\overline{y_{1h}}-\widehat{\beta_0}-\widehat{\beta_1}\overline{x_h})^2}{\sigma_e^2+(-\widehat{\beta_1},1)\sum(-\widehat{\beta_1},1)'}=H-2$$

위의 방정식을 $g(\sigma_e^2)=H-2$이라 하고, 여기서 H는 소지역 개수라 하면, $\theta=\sigma_e^2$ 인 $g(\theta)=0$에 대한 뉴튼 타입 방법은 다음과 같이 구할 수 있다.

$$\theta^{(t+1)}=\theta^{(t)}+\frac{1}{g'(\theta^{(t)})}(H-2-g(\theta^{(t)})) \tag{4.23}$$

여기서

$$g'(\theta)=-\sum_{h=1}^{H}\frac{(\overline{y_{1h}}-\widehat{\beta_0}-\widehat{\beta_1}\overline{x_h})^2}{\left\{\theta+(-\widehat{\beta_{1,}}1)\sum(-\widehat{\beta_{1,}}1)'\right\}^2}$$

이때 $\sigma_{e,h}^2\equiv\sigma_e^2$ 이라고 하면, 전체 추정 절차는 다음과 같다.

<모수 추정 절차>

step1. (4.17), (4.19)에서 $\hat{\sigma}_e^2=0$때 초기값 $(\beta_{0,}\beta_1)$계산

step2. step1에서 계산된 $(\beta_{0,}\beta_1)$에 대해 $\widehat{\sigma^2}$을 (4.23)의 반복알고리즘에 의해 계산

step3. step2에서 구한 $\widehat{\sigma^2}$에 대해 $(\beta_{0,}\beta_1)$ 업데이트

step4. step2-step3을 추정치가 수렴할 때까지 반복수행

Remark2. 만약 $\sigma_{e,h}^2 = \sigma_e^2$이 만족되지 않으면, 대안으로 (4.24)를 생각할 수 있다.

$$\overline{e_h} \sim (0 \ , \ \overline{X_h}\sigma_e^2) \tag{4.24}$$

모형 (4.24)가 유효한지를 확인하기 위해 (4.25)의 ν_h,

$$\nu_h = (\overline{y_{1h}} - \hat{\beta}_0 - \overline{x_h}\hat{\beta}_1)^2 - \hat{\beta}_1^2 V(a_h) + 2\hat{\beta}_1 \hat{C}(a_h, b_h) - V(b_h) \tag{4.25}$$

을 계산한 후 ν_h와 $\overline{x_h}$에 대해 선형그림을 그려본다. 만약 그림에서 두 변수 간에 선향 관계를 보이면 모형 (4.24)는 타당한 모형으로 간주될 수 있다. 모형 (4.24)하에서 비 방법(ratio method)에 의해 σ_e^2을 구할 수 있다.

$$\hat{\sigma}_e^2 = \frac{\Sigma_{h=1}^H \kappa_h \nu_h}{\Sigma_{h=1}^H \kappa_h \widehat{\overline{X}}_h} \tag{4.26}$$

여기서

$$\kappa_h \propto \left\{ \widehat{\overline{X}}_h \hat{\sigma}_e^2 + (-\hat{\beta}_1, 1) \Sigma_h (-\hat{\beta}_1, 1)' \right\}^{-1}$$

$\Sigma_{h=1}^H \kappa = 1$, κ_h가 σ_e^2에 의존하기 때문에 (4.26)의 해는 반복 수행에 의해 구해진다.

Remark 3. 정규분포 근사를 위해 $\overline{x_h^*} = T(\overline{x_h})$, $\overline{y_{1h}^*} = T(\overline{y_{1h}})$ 와 같은 변수변환을 고려할 수 있다. 정규성을 확인하기 위해서 우선 $n_{ha}\overline{V}(\overline{x_h})$ 를 $\overline{x_h}$에 대해 산점도를 그린다. 만약 산점도가 조금이라도 구조적 관계를 보이면 정규성 가정은 위배된다고 본다. 다음의 로그변환을 생각해 보자.

$$T(x) = \log(x) \tag{4.27}$$

$\overline{x_h^*} = T(\overline{x_h})$의 근사분산은 다음과 같다.

$$V(\overline{x_h^*}) \doteq \frac{1}{(\overline{x_h})^2} V(\overline{x_h}).$$

이런 변수변환은 일종의 분산안정화변환에 해당하고, 정규분포 가정 하에서 모수추정하고자 할 때 유용하다.

일단 $\overline{X}_h^*$ 에 대한 추정량 $\widehat{\overline{X}}_h^*$ 가 구해지면, 이를 역변환(back transformation)하여 최량추정량을 구한다. 즉, $\overline{X}_h = T^{-1}(\overline{X}_h^*) := Q(\overline{X}_h^*)$. 테일러 전개에 의해

$$Q(\widehat{\overline{X}_h^*}) \doteq Q(\overline{X_h^*}) + Q'(\overline{X_h^*})(\widehat{\overline{X}_h^*} - \overline{X}_h) + \frac{1}{2} Q'(\overline{X_h^*})(\widehat{\overline{X}_h^*} - \overline{X}_h)^2$$

이고, $\overline{X} = Q(\overline{X}_h^*)$의 추정량을 $Q(\widehat{\overline{X}}_h^*)$이라 하면, 낮은 차수의 항을 무시하여 다음의 식을 얻을 수 있다.

$$E\left\{Q(\widehat{\overline{X}_h^*})\right\} = \overline{X}_h + \frac{1}{2} Q'(\overline{X_h^*})\, V(\widehat{\overline{X}_h^*}) = \overline{X}_h + \frac{1}{2} \overline{X}_h\, V(\widehat{\overline{X}_h^*})$$

(4.27)의 로그변환에 대해서 $Q(\widehat{\overline{X}}_h^*) = \exp(\overline{X}_h^*)$, $Q''(\overline{X}_h^*) = \overline{X}_h$. 따라서 $\widehat{\overline{X}}_h = Q(\overline{X}_h^*)$,

$$E(\widehat{\overline{X}_h}) \cong \overline{X}_h + \frac{1}{2} \overline{X}_h\, V(\widehat{\overline{X}_h^*}),$$

이고, $\overline{X}_h$의 편향을 보정한 추정량(bias-corrected estimator)은

$$\widehat{\overline{X}_{h,bc}} = \frac{\widehat{\overline{X}}_h}{1+0.5\,V(\widehat{\overline{X}}_h^*)} \tag{4.28}$$

이고, 분산 $V(\widehat{\overline{X}}_h^*)$은 4.3절의 MSE 추정방법에 의해 계산될 수 있다.

3. MSE 추정

이 절에서는 식 (4.10)에 주어진 GLS 추정량 $\hat{\overline{X}}_h$에 대한 MSE 추정에 대해서 설명한다. GLS 추정량은 모수 (β_0,β_1)과 σ_e^2의 함수적 관계에 있다. 만약 모형에서의 모수들을 알고 있다면, $\widehat{\overline{X}}_h$의 MSE는 remark 1에서 논의한 바와 같이 $M_{h1}=\alpha_h V(\overline{x}_h)$이다. 즉, $\theta=(\beta_0,\beta_1,\sigma_e^2)$이고, $\widehat{\overline{X}}_h=\widehat{\overline{X}}_h(\theta)$이라고 하면, $\overline{X}_h$에 대한 실제 예측은 $\hat{\overline{X}}_{eh}=\hat{\overline{X}}_h(\hat{\theta})$에 의해 계산된다. 우선 $MSE(\widehat{\overline{X}}_h^*)$의 다음과 같은 분해를 생각하자.

$$\begin{aligned} MSE(\widehat{\overline{X}}_{eh}) &= MSE(\widehat{\overline{X}}_h) + E\left\{(\widehat{\overline{X}}_{eh}-\widehat{\overline{X}}_h)^2\right\} \\ &=: M_{h1}+M_{h2} \end{aligned}$$

MSE를 추정하기 위해 잭나이프 방법을 사용한다. 잭나이프 계산은 다음과 같은 과정을 거쳐 계산한다.

Step1. 전체 자료셋 $\left\{(\overline{x}_h,\overline{y}_{1h})\,;\,h=1,2,\cdots,H\right\}$ 에서 k 번째 지역의 자료 $(\overline{x}_k,\overline{y}_{1k})$를 제거해 가면서 $\hat{\theta}$에 대한 $\theta^{(-k)}$를 계산한다. 이 계산은 각 k 에서 $\theta:\left\{\hat{\theta}^{(-k)}\,;k=1,\cdots,H\right\}$를 구하고 이를 H개 반복해서 구한다. 다음으로 H개의 $\hat{\overline{X}}_h:\left\{\hat{\overline{X}}_h^{(-k)}\,;k=1,2,\cdots,H\right\}$, $\hat{\overline{X}}_h^{(-k)}=\hat{\overline{X}}_h(\hat{\theta}^{(-k)})$을 구한다.

Step2. M_{h2}를 구한다.

$$\widehat{M_{2h}} = \frac{H-1}{H}\sum_{k=1}^{H}(\widehat{\overline{X}_h^{(-k)}} - \widehat{\overline{X}_h})^2. \tag{4.29}$$

Step 3. M_{h1}을 구한다.

$$\widehat{M_{1h}} = \widehat{\alpha_h^{(JK)}}\,\hat{V}(\overline{x_h}) + (1-\widehat{\alpha_h^{(JK)}})\hat{C}(\overline{x_h}, \overline{y_{1h}}) \tag{4.30}$$

여기서 $\hat{\alpha}_h^{(JK)}$ 은 α_h에 대한 편향이 보정된 추정량이다.

$$\widehat{\alpha_h^{(JK)}} = \widehat{\alpha_h} - \frac{H-1}{H}\sum_{k=1}^{H}(\widehat{\alpha_h^{(-k)}} - \widehat{\alpha_h}),$$

$$\widehat{\alpha_h} = \frac{\widehat{\sigma_e^2} + V(b_h) - \widehat{\beta_1}C(a_h, b_h)}{\widehat{\sigma_e^2} + V(b_h) + \widehat{\beta_1^2}V(a_h) - 2\widehat{\beta_1}C(a_h, b_h)},$$

그리고,

$$\widehat{\alpha_h^{(-k)}} = \frac{\widehat{\alpha_e^{(-k)2}} + V(b_h) - \widehat{\beta_1^{(-k)}}C(a_h, b_h)}{\widehat{\sigma_e^{(-k)2}} + V(b_h) + (\widehat{\beta_1^{(-k)}})^2 V(a_h) - 2\widehat{\beta_1^{(-k)}}C(a_h, b_h)}.$$

Remark 4. 변수변환 식 (4.27)을 적용하면 편향이 보정된 추정량 (4.28)을 사용하게 된다. 그리고 이에 따른 MSE 추정방법도 바꿀 필요가 있다. 이를 위해 편향이 보정된 추정량, $\widehat{\overline{X}}_{eh,bc}$에 대해 정리하면 다음과 구할 수 있다.

$$
\begin{aligned}
MSE(\widehat{\overline{X}_{eh,bc}}) &= MSE(\widehat{\overline{X}_{eh}}) \\
&= MSE\left\{Q(\widehat{\overline{X}_{eh}^{*}})\right\} \\
&\cong \left\{Q'(\widehat{\overline{X}_{h}^{*}})\right\}^2 \cdot MSE(\widehat{\overline{X}_{eh}^{*}}) \\
&= \overline{X_h^2} \cdot MSE(\widehat{\overline{X}_{eh}^{*}})
\end{aligned}
$$

여기서 처음 등식은 $\hat{\overline{X}}_{h,bc} - \hat{\overline{X}}_{h}$이 $O_p(n_h^{-1})$을 따른다. $\hat{\overline{X}}_{h}^{*}$, 즉 변수변환 후의 $\overline{X}_{h}^{*}$의 GLS 추정량의 MSE는 식 (4.29)와 식 (4.30)에 의해 계산된다. $MSE(\hat{\overline{X}}_{eh}^{*})$가 추정되면 여기에 $\hat{\overline{X}}_{eh,bc}^{*}$ 를 곱하면 역변환 된 GLS 추정량 $\hat{\overline{X}}_{eh,bc}$의 MSE를 구할 수 있다.

제5절 경활조사에 적용

1. 자료

경활조사는 우리나라 16개 시도 단위의 고용통계를 공표할 목적으로 설계된 것으로 전국 약 32,000가구 (72,000여명)에 대해 매월 조사를 실시한다. 지역별고용조사는 특·광역시와 9개 도내에 포함된 158개 시군에서의 고용통계 공표를 목표로 전국 약 17만 5천여 가구 (약 33만 명)를 대상으로 하는 조사로서, 2008년과 2009년은 연 1회 실시하였고, 2010년 3분기부터는 분기조사를 실시하고 있다. 한편 3절에서 제시한 모형에서의 등록자료는 한국고용정보원에서 제공하는 실업급여등록자자료로서 지역, 성, 연령별 피보험자수, 실업급여등록자 정보를 사용하고 있다. 이 자료는 우리나라 임금근로자로 한정되어 있기 때문에 전체 실업자 정보를 대표하기에는 상당한 커버리지 오차가 포함되어 있다.

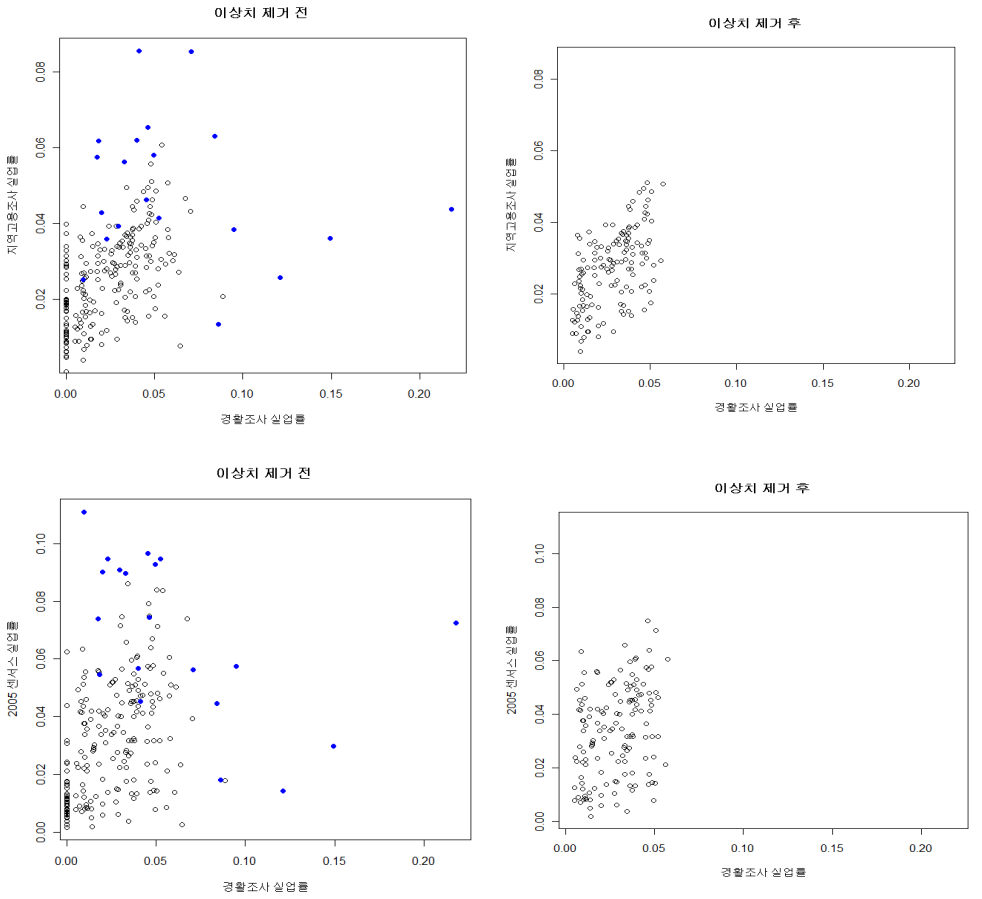
소지역 추정에 이용할 수 있는 보조정보는 <표 1-2>와 같다. <표 1-2>에 제시된 기호는 4절에서 사용된 기호를 중심으로 작성되었다.

〈표 1-2〉 자료의 특성

자료	관측정보	지역수준 추정	특성
조사 A	직접 관측 x_i	$\widehat{X_h}$, $\widehat{V}(\widehat{X_h})$	· 표본오차 (large)
조사 B	보조 관측 y_{1i}	$\widehat{Y_{1h}}$, $\widehat{V}(\widehat{Y_h})$	· 측정오차 · 표본오차 (small)
센서스	보조 관측 y_{2i}	$\widehat{Y_{2h}}$	· 측정오차 · 갱신되지 않은 정보

조사 A는 경활조사자료로서 회귀모형에서 설명변수에 해당한다. 이 자료는 소규모 표본조사자료로서 특히 소지역에서는 큰 표본오차를 수반한다. 조사 B는 지역별고용조사자료로서 회귀모형에서 종속변수에 해당한다. 지역별고용조사 자료는 대규모 조사 자료로서 경활조사에 비해 상대적으로 작은 표본오차를 포함하고 있으나, 조사원 효과 또는 조사기간 등에 따른 측정오차가 더 발생할 수 있다고 간주한다. 센서스는 전수자료로서 표본오차는 포함하지 않는다. 그렇지만 대규모 조사인 만큼 측정오차가 존재하며 우리나라의 경우, 5년 주기 센서스로 그 기간 동안에는 매월 또는 연단위로 정보가 갱신되지 않는다는 특성이 있다. 참고로 <표 1-2>에는 제시되지 않았지만 등록자료는 실업급여등록자 자료로서 표본오차는 존재하지 않지만, 경우에 따라서 커버리지 오차가 크게 발생할 수 있다.포함될 수 있다.

다음 [그림 1-4]는 경활조사 실업률 vs. 지역별고용조사 실업률, 경활조사 실업률 vs. 센서스 실업률 간의 산점도를 그린 것이다. 이상치 제거 전 자료, 즉 원자료는 실업률이 0인 지역을 포함하고 있고, 진한색 점으로 표시된 지역들은 이상치로 간주될 있다. 따라서 이 지역들을 제외하면, 이상치 제거 후 자료와 같이 어느 정도 두 변수들 간의 이상치의 효과가 없어질 것으로 예상된다. 실제로 모형을 세울 때 이상치를 포함한 경우와 그렇지 않은 경우를 모두 고려하여 이상치 효과를 확인하였고, 최종 분석은 이상치 효과를 제외한 결과를 사용하였다.



[그림 1-4] 이상치 제거전후 경활조사, 지역별고조사, 센서자료 간의 산점도

2. 방법

본 연구에서 제안한 방법을 경활조사에 적용해 보자. 이미 앞에서 설명한바와 같이 고용관련 두 개 조사가 있다. 경활조사는 소지역에서 심한 표본오차가 존재하지만, 측정오차는 없다고 가정한다. $\overline{X}_h$가 지역 h에서의 진짜 실업률이라고 하자. 경활조사에서 $\overline{x_h}$ 를 관측하고 지역별고용조사에서 $\overline{y_{1h}}$ 를 관측한다. 먼저 모집단을 두 개의 그룹, 즉, 도시 지역과 농촌 지역으로 나눈다. 이때 나누는 기준은 농업에 종사의 가구의 비율을 사용 한다 (20%). 각 그룹 내에서 모형을 각각 세우고 각 모형으로부터 모수를 각각 추정한다.

구조모형은 다음의 식 (5.1)과 같다.

$$\overline{Y_h} = \beta_1 \overline{X_h} + e_h \tag{5.1}$$

이때 $e_h \sim (0, \sigma_e^2)$를 따른다. 여기서 $\overline{X}_h$의 GLS 추정량이 음수가 되지 않도록 하기 위해 $\beta_0 = 0$을 세팅한다. 표본오차 모형은 앞 절에서 세운 모형과 같다. 이 경우에 모수 β_1은 다음의 식 (5.2)와 같이 추정할 수 있다

$$\hat{\beta}_1 = \frac{\Sigma_{h=1}^{H} w_h(\hat{\beta}_1)\{\overline{x_h}\,\overline{y_{1h}} - C(a_h, b_h)\}}{\Sigma_{h=1}^{H} w_h(\hat{\beta}_1)\{\overline{x_h^2} - V(a_h)\}}. \tag{5.2}$$

모형분산은 (4.22)의 적률방법에 의해 추정된다. 이때 $\hat{\beta}_0 = 0$이다. 이 추정량은 (4.10)에 의해 구해지고, 이때 $\tilde{x}_h = \hat{\beta}_1^{-1}\,\overline{y}_{1h}$이다. 여기에 센서스 정보를 이용할 수 있다. 이때 세 개의 정보를 이용한 GLS 추정량은 다음과 같이 표현될 수 있다.

$$\begin{pmatrix} \overline{Y_{2h}} \\ \overline{y_{1h}} \\ \overline{x_h} \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} \overline{X_h} + \begin{pmatrix} \overline{e_{2h}} \\ b_h + \overline{e_{1h}} \\ a_h \end{pmatrix}$$

여기서 $\overline{Y}_{2h}$는 소지역에서의 센서스 조사값을 나타낸다. 센서스값은 표본오차에 영향을 받지 않기 때문에, 모형을 $E(\overline{Y}_{2h}) = \gamma_1 \overline{X}_h$ 라 하면 모형오차 e_{2h}가 유일한 오차를 대표한다. 모형에서 모수는 4.2절에서 소개한 방법을 이용하여 추정된다. 이때 분산행렬은 $\Sigma = diag(0, V(a_h))$이다. $\overline{X}_h$의 GLS 추정량은 쉽게 구할 수 있다. MSE 추정은 M_{h1} 부분은 다음과 같이 구하고,

$$V(\widehat{\overline{X}_h} - \overline{X_h}) = \left[\begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix}' \left\{ V \begin{pmatrix} \overline{e_{2h}} \\ b_h + \overline{e_{1h}} \\ a_h \end{pmatrix} \right\} \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} RIGHT \right]^{-1} := M_{h1}$$

편향보정에 대해서는 잭나이프 방법을 적용하여 계산될 수 있다.

MSE 추정치에 대한 CV는 다음과 같이 계산된다.

$$CV_{MSE(}\hat{\theta}) = \frac{\sqrt{mse(\hat{\theta})}}{\hat{\theta}}$$

제안한 방법에 의한 소지역 추정결과는 5.3절에서 제시한다. 전체적으로 직접추정량의 CV_{MSE} 를 개선하는 효과가 크지만, 소지역 추정치는 CV_{MSE} 에 대해 30% 기준을 사용한다고 할 때, CV가 30% 이상인 지역이 기대했던 것보다 많게 나타난다.

3. 결과

가. 모형세팅 방법

○ 지역별고용조사 직접추정량의 분산이 다소 불안정한 점을 고려하여, 경활조사 대비 지역별고용조사 표본의 비를 이용하여 분산추정량을 스무딩한 보정을 취했다. 이렇게 한 이유는 경활조사의 표본이 지역별고용조사 내에 포함되어 있다는 점을 고려한 것이다.

○ 회귀절편이 있는 회귀모형을 사용할 경우, 기울기가 음수값을 갖는 경우가 발생할 수 있기 때문에 이를 고려하여 회귀절편이 없는 모형($\beta_0 = 0$)을 사용하였다.

○ 우리나라 230개 시군구 지역을 크게 도시 지역과 농촌 지역 두 개 그룹으로 나누어 분석하였다. 두 개 그룹에 대한 분리 기준은 농업가구 비율이 20% 미만인 지역을 도시, 그 이상인 지역은 농촌 지역으로 간주하였다. 이렇게 한 이유는 설명변수와 종속변수간의 관계가 지역에 따라 다른 양상을 보이고, 특히 농촌과 도시 지역 개념에서 그 특성이 다르다는 점에 착안하였다.

○ 회귀모형을 통한 모수 추정은 자료에 민감하게 반응하기 때문에 우선 경활조사의 실업률 추정치가 0인 지역은 모수추정을 위한 자료에서 제외하고, 이 자료에서 이상치(outlier)로 판단되는 자료 또한 모수추정 자료에서 제외하였다. 즉, 모수추정에 사용된 최종자료는 경활조사자료의 실업률이 0인 지역과 이상치로 판단되는 지역의 자료가 제외된 총 180개 지역의 자료가 사용되었다.

○ 위에서 추정된 모수를 이용하여 섬 지역(경활조사 표본제외지역)을 제외한 227개 소지역에 대해 추정치를 계산하였다.

나. 모형세팅 확인 과정

위에서 정의한 최종 모형을 세팅하기 위해 다양한 분석을 시도하였다. 모형세팅은 이상치 제거 유무와 도시/비도시 구분 여부를 확인하기 위한 것이다. 이를 위해 모형은 단순하게 설정하였다. 즉, 경활조사 자료와 지역별고용조사 자료만을 이용하여 회귀계수와 모형오차 분산 (β_1, σ_e^2)을 구하고, 이에 대한 MSE를 계산하였다.

① 모수추정(그룹화/또는 이상치 제거 유무 고려)

〈표 1-3〉 모형 세팅을 위한 모수 추정치

구분	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\widehat{\sigma_e^2}$
GLS1(이상치 제거 전)	0.000	0.6582	0.0003
GLS2(이상치 제거 후)	0.000	0.8814	0.0001
GLS3(도시)	0.000	0.7131	0.0004
GLS4(농촌)	0.000	0.4162	0.0001
GLS3(도시-이상치 제거 후)	0.000	0.9026	0.0001
GLS4(농촌-이상치 제거 후)	0.000	0.6900	0.0000

먼저 이상치 제거 유무에 따른 결과를 살펴보자. 이상치를 제거하면 기울기($\widehat{\beta_1}$)가 증가하고, 모형분산은 더 감소하는 것을 알 수 있다. 도시와 농촌 그룹 여부에 따라서는 도시 (GLS3)는 그룹하기 전(GLS1)보다 그룹한 후에 기울기가 약간 증가하는 것에 대해 농촌은 오히려 기울기가 그룹하기 전 기울기가 0.6582(GLS1)보다 그룹한 후에 기울기가 0.4162(GLS4)로 크게 감소하는 것을 알 수 있다. 그리고 지역 그룹화한 후에 모형분산은 약간 커지거나 같은 것을 알 수 있다. 이 결과로부터 모형 세팅을 위해 이상치를 제거하는 것이 좋은 추정 결과를 준다고 볼 수 있다. 그룹 여부에 대한 효과를 판단하기 위해 그룹하기 전과 후의 MSE를 비교해 보자.

② MSE 추정 분포

〈표 1-4〉 각 추정량의 MSE에 대한 기초통계량

180개 지역	최소값	1사분위수	중앙값	평균	3사분위수	최대값
경활조사	0.00000000	0.00006307	0.00012100	0.00024760	0.00023950	0.00205800
지역별고용조사	0.00010560	0.00012460	0.00013680	0.00015080	0.00016010	0.00047000
GLS1(이상치제거전)	0.00000969	0.00003631	0.00005242	0.00007343	0.00009315	0.00035540
GLS2(이상치제거후)	0.00000928	0.00003202	0.00004640	0.00005839	0.00007932	0.00022970
GLS3(그룹-이상치제거전)	0.00000972	0.00003698	0.00005367	0.00007815	0.00009888	0.00039720
GLS4(그룹-이상치제거후	0.00000924	0.00003154	0.00004491	0.00005771	0.00007844	0.00022060

<표 1-4>는 실업률이 '0'이 아닌 지역에 대한 MSE 분포를 나타낸다. 지역별고용조사의 MSE를 계산할 때 모형분산은 $\sigma_e^2 = 0.0001$을 가정하고 이것을 표본 분산에 더해서 계산하였다.

①에서 설명한 <표 1-2>에서와 같이 이상치를 제거하면 MSE가 감소하는 효과가 있지만, 그룹 여부에 대해서는 MSE에 대해 큰 변화가 없는 것으로 나타났다. 그리고 180개 시군 지역에 대한 GLS 추정치는 경활조사의 MSE를 크게 감소시키는 효과가 있지만 지역별고용조사의 MSE 만큼은 줄어들지 않은 것으로 나타났다. 이러한 결과를 바탕으로 섬지역을 제외한 227개 지역에 대해 추정결과를 해석해 볼 수 있다.

다. 최종 모형

섬지역을 제외한 전국 227개 시군구 지역에 대한 최종 모형은 이상치를 제거한 자료로부터 모수를 추정하고, 이 모수 추정치를 이용하여 소지역 복합추정량과 이에 대한 MSE 를 계산하였다. 최종 모형을 다시 쓰면 다음과 같고,

$$\begin{pmatrix} \overline{Y_{2h}} \\ \overline{y_{1h}} \\ \overline{x_h} \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} \overline{X_h} + \begin{pmatrix} \overline{e_{2h}} \\ b_h + \overline{e_{1h}} \\ a_h \end{pmatrix}$$

이 모형에 대한 모수추정치와 모형 분산은 각각 다음과 같다.

$\beta_1 = 1.0586(\sigma^2_{e1h} = 0.00004285),\ \gamma_1 = 1.3713(\sigma^2_{e2h} = 0.0001987),$

$V(a_h)$: 경활조사 실업률 분산,

$V(b_h)$: 지역별고용조사 실업률 분산, $Co\,V(a_h, b_h) = V_h(b_h)$

GLS 복합추정량을 구하기 위해 자료, 회귀계수 추정치, 분산공분산 추정치 행렬을 다음과 같이 표현하면,

$$Y = \begin{pmatrix} \overline{Y_{2h}} \\ \overline{y_{1h}} \\ \overline{x_h} \end{pmatrix},\ \theta = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix},\ V_h = \begin{pmatrix} \sigma^2_{e2} & 0 & 0 \\ 0 & \sigma^2_{e1} + V(b_h) & Cov(a_h, b_h) \\ 0 & Cov(a_h, b_h) & V(a_h) \end{pmatrix}$$

과 같고, GLS 복합추정량 $\widehat{\overline{X}}_h$는 다음과 같이 계산된다.

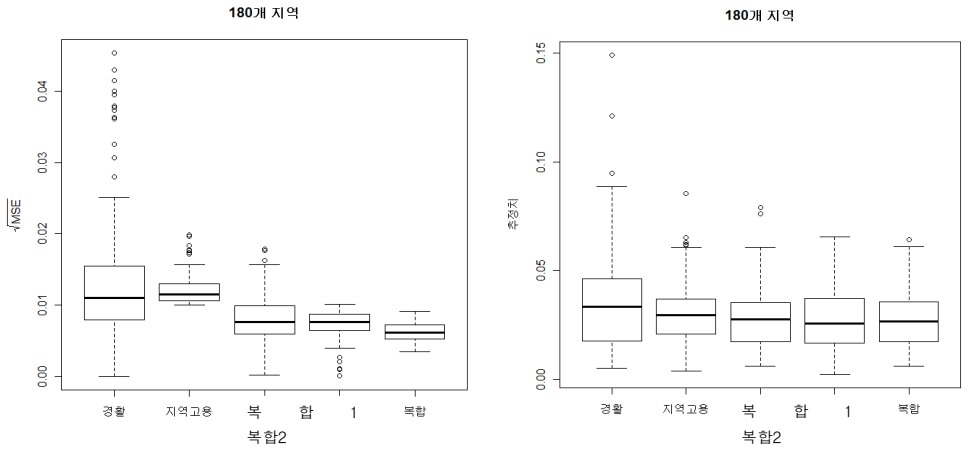
$$\widehat{\overline{X}}_h = (\theta' V_h^{-1} \theta)^{-1} \theta\, V_h^{-1} Y_h'.$$

MSE 분포

최종 227지역에 대한 MSE 분포를 각 추정량별로 나타내면 다음의 <표 1-5>와 같다. 그리고 추정치의 MSE 180개 지역과 227개 지역 각각에 대해 나타낸 것이 [그림 1-5]와 [그림 1-6]이고, 180개 227개 지역들의 CV를 비교한 그림은 [그림 1-7]과 같다.

〈표 1-5〉 직접추정량과 복합추정량의 MSE 분포

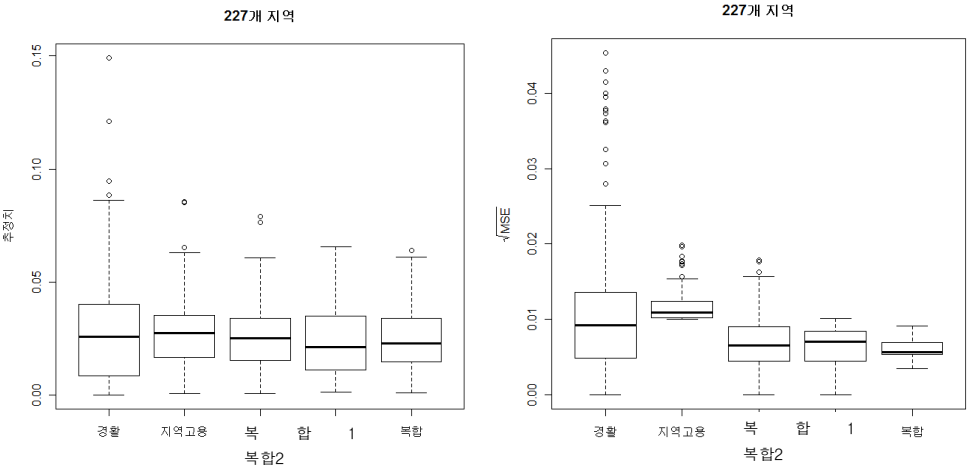
추정량	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
경활	0.0000000	0.00002396	0.00008608	0.00019630	0.00018400	0.00205800
지역고용	0.0001000	0.00010400	0.00011920	0.00013820	0.00015320	0.00039470
복합추정	0.0000125	0.00002868	0.00003219	0.00003836	0.00004893	0.00008400



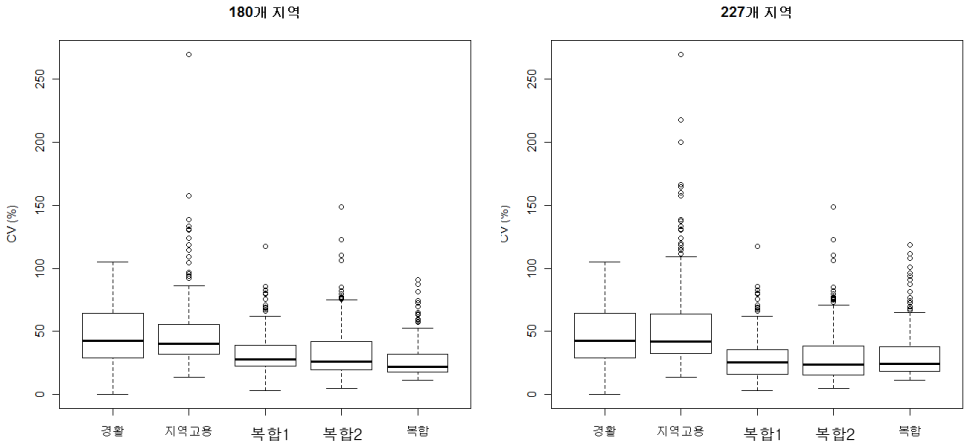
[그림 1-5] 180개 지역에 대한 추정치의 $\sqrt{MSE}$ 추정치와 실업률 추정치 분포

< [그림 1-5] ~ [그림 1-7]에서 사용한 추정량의 이름 >

복합1 : 지역별고용조사 자료만 보조변수로 사용한 경우, 복합추정량
복합2 : 센서스 자료만 보조변수로 사용한 경우, 복합추정량
복합 : 지역별고용조사와 센서스 자료 모두를 사용한 경우, 복합추정량



[그림 1-6] 227개 지역에 대한 추정치의 $\sqrt{MSE}$ 추정치와 실업률 추정치 분포



[그림 1-7] 180개 vs. 227개 지역에 대한 CV 분포

위에서 설명한 바와 같이 [그림 1-5] ~ [그림 1-7]에서 지역은 이상치와 실업률이 0인 지역들을 제외한 180개 지역과 전국 230개 소지역 중 섬지역 만을 제외한 227 지역으로 구분하였다. 모형을 세울 때, 이상치와 실업률 0인 지역들이 모형에 미치는 효과를 제거하기 위해 180개 지역만을 사용하기로 하였다.

180개 지역의 추정결과

- 각 모형추정방법 복합1, 복합2, 복합 추정량 중 지역별고용조사와 센서스자료를 모두 결합한 복합 추정량(복합)의 MSE가 가장 작고, 이 MSE의 변동성도 가장 적게 나타났다.
- 모든 복합 추정량들은 모두 경활 직접추정량의 MSE를 훨씬 줄이는 효과가 있다.
- MSE는 복합 추정량에서 지역별고용조사에서 보다 훨씬 작게 추정되었다.
- 실업률 추정치는 각 추정방법에 따라 대체로 비슷하고, 복합 추정치는 경활직접추정치와 지역별고용조사 직접추정치보다 낮게 나타난다.
- CV 는 복합1, 2가 낮고, 그 중에서 복합추정치가 가장 낮다.
- CV 를 30% 기준으로 볼 때, 복합추정량의 경우가 51개 지역으로 가장 적게 나타났다(<표 1-6>). 경활조사와 지역별고용조사 직접추정량들의 경우는 227개 지역 중 각각 131개, 148개 지역에서 CV가 30% 이상으로 것으로 나타났다

〈표 1-6〉 180개 지역에 대한 CV기준에 따른 지역수 분포

180개 지역	10% 미만	10~20%	20~30%	30% 이상
경활	5	6	38	131
지역고용	-	4	30	146
복합1	5	17	76	82
복합2	5	43	52	80
복합	-	73	56	51

227개 소지역 추정 결과

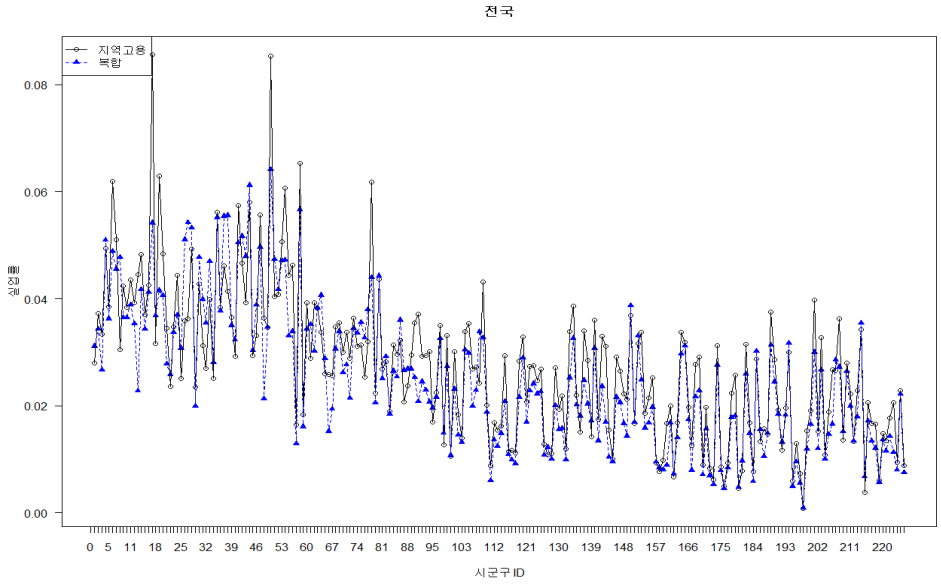
- 227개 지역추정에 있어서 MSE는 180개 지역과 거의 유사한 결과를 보인다.
- CV 는 180개 지역에 비해 모형추정치들 간에 유사한 경향을 보인다.
- 지역별고용조사의 CV 가 30% 이상인 지역의 수는 190개 지역이고, 경활조사의 경우는 180개 지역과 동일한 것처럼 보이지만 실제는 실업률이 0 지역(47개)이 제외된 숫자이다.
- 복합 추정량의 경우 CV 가 30% 이상인 지역수가 227개 중 83개 지역에 해당한다.

〈표 1-7〉 227개 지역에 대한 CV기준에 따른 지역수 분포

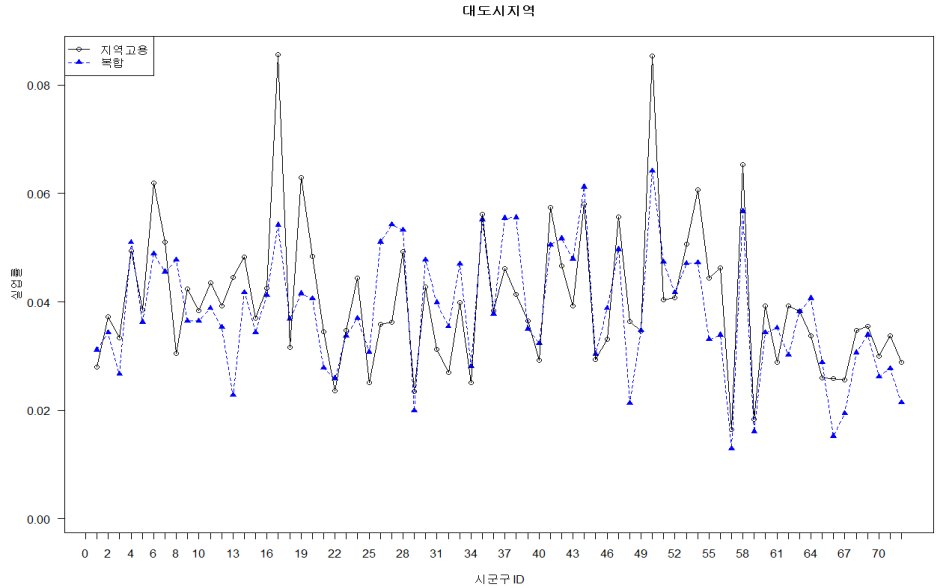
227개 지역	10% 미만	10~20%	20~30%	30% 이상
경활	5	6	38	131(47)
지역고용	-	4	33	190
복합1	52	17	76	82
복합2	52	43	52	80
복합	-	78	66	83

227개 지역의 추정치

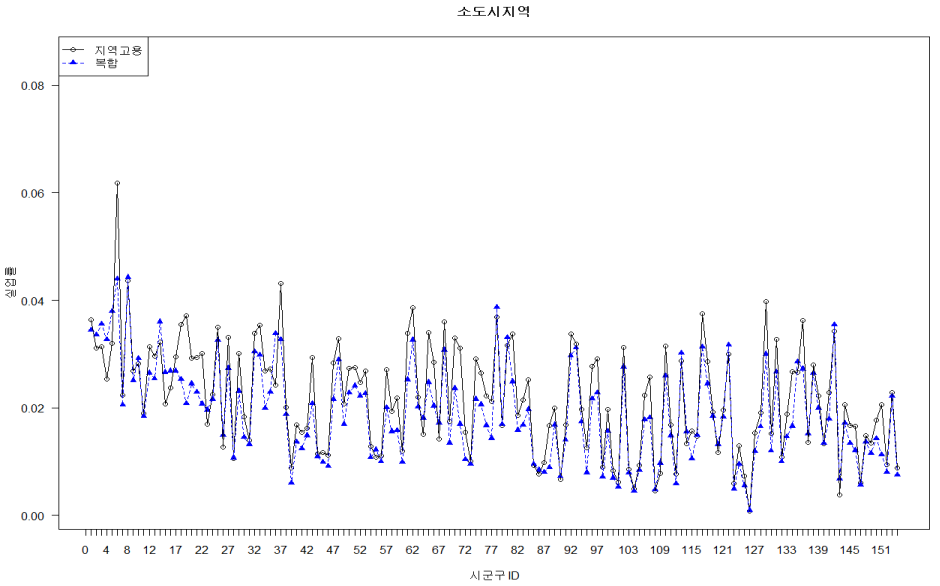
복합 추정치는 지역별고용조사 추정치에 비해 전반적으로 실업률을 낮게 추정하는 것을 알 수 있다. 대도시(특광역시)보다는 소도시(도 지역)에서 복합 추정치 실업률이 지역별고용조사보다 더 낮은 경향이 있다. 다음의 그림 [그림 1-8]은 지역 구분에 따른 시군구 실업률 추정치를 나타낸 것이다. 그림에서 가로축으로 갈수록 표본수가 적거나 소도시에 해당되는 지역들이다.



(가) 전국 단위



(나) 대도시 단위 (특광역시 의미함)



(다) 소도시 단위(도 단위 의미)

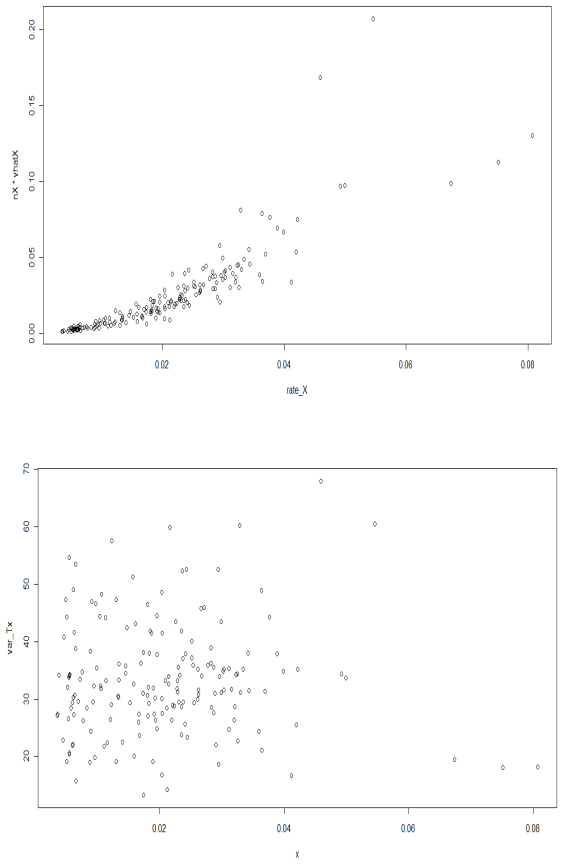
[그림 1-8] 지역 구분에 따라 추정량별 시군구 실업률 추정치 분포:지역별고용조사 vs. 복합

· CV가 30% 이상인 지역들은 대부분이 표본수가 작거나 실업률이 0인 지역들이 포함된다. CV가 30% 이상인 지역은 전체 83개로 이들 지역 중 표본 조사구수가 3개 이하인 지역수는 46개 지역에 해당된다. 그렇지만 표본 조사구수가 일정수준 이상인 지역들에서도 CV가 30% 이상인 경우가 많이 발생한다. 표본 조사구수가 7개 이상인 지역은 가구수가 약 140여 가구가 포함되기 때문에 일반적인 가구 조사 소지역 추정의 경우 표본수가 그렇게 작은 경우는 아니다. 여러 가지 측면에서 그 이유를 찾을 수 있겠지만, 대체로 자료의 비대칭성 또는 직접추정량의 분산추정에서 기인하는 것으로 예상된다.

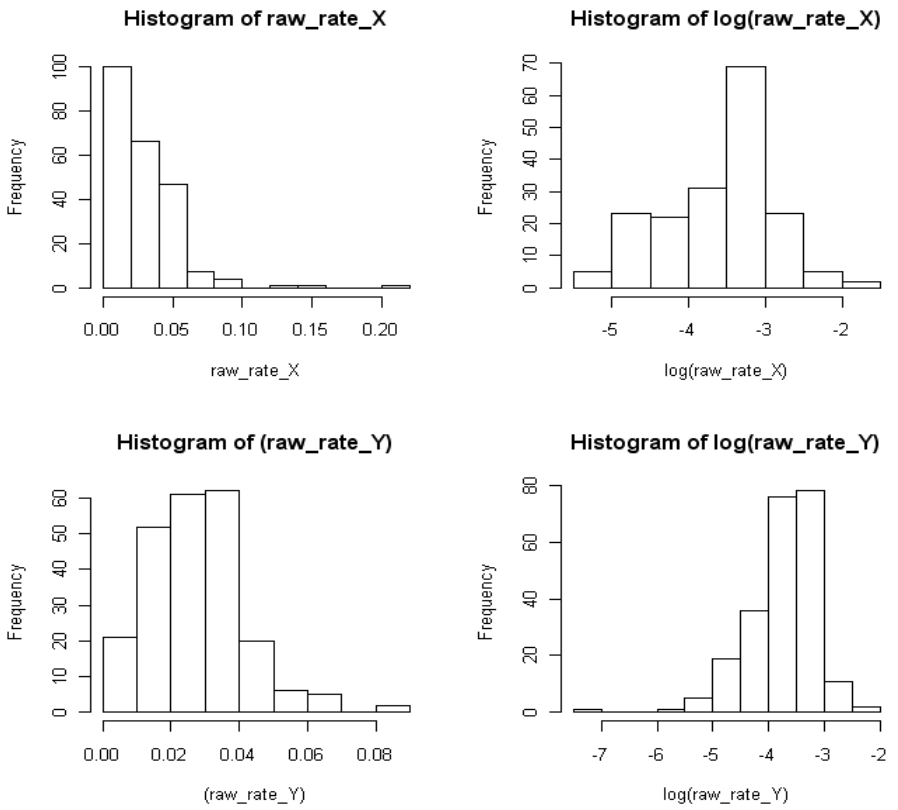
〈표 1-8〉 227개 지역 중 CV>30%인 지역에 대한 표본수 분포

표본규모(조사구)	1개	2개	3개	4개~6개	7개~
지역수	6	25	15	26	17

· 이에 대해 추가적으로 살펴보면, 직접추정량들에 대한 분산안정화를 꾀함으로써 추정치 향상을 기대할 수 있다. 하나의 예로 경활조사의 실업률과 분산과의 산점도를 그려보면, 두 변수 간에 구조적 관계가 있음을 알 수 있다. 즉, 이 변수가 정규성 가정에 위배되었음을 알 수 있다([그림 1-9]의 위쪽 그림). 이를 해결하기 위해 로그 변환을 취하면, 변환 전에 나타났던 구조적 관계가 없어짐을 확인할 수 있다([그림 1-9]의 아래쪽 그림). 이러한 변환을 통해 분산안정화를 기대할 수 있고, 정규성 가정하에서 모수의 추정을 향상시킬 수 있다. 또한 원 자료와 로그 변환자료를 비교해 볼 때, 변환 후에 훨씬 대칭적인 분포 형태를 따르는 자료임을 알 수 있다([그림 1-10]). 이렇게 함으로써 어느 정도의 정규성을 가정한 모수추정을 통해 추정치의 정도를 높일 수 있을 것이다.



[그림 1-9] 자료변환 전·후 실업률과 분산과의 산점도: 전 (위쪽), 후(아래쪽)



[그림 1-10] 자료변환전후 실업률 분포: 전(왼쪽), 후(오른쪽), 위(경활조사 실업률), 아래 (지역별고용조사 실업률)

종합해 보면, 본 연구에서 제안한 GLS 추정량은 경활조사의 MSE와 CV를 크게 개선하는 효과가 있다. 제안된 복합 추정량은 표본수가 적거나 실업률이 '0'인 지역에서 CV가 크게 추정되는 경향이 있고, 조사추정치에 비해 실업률이 낮게 추정되는 경향이 있다. 현재 제안된 방법에 사용된 자료의 분산이 불안정하여 자료의 정규성 가정이 위배된다고 볼 수 있고, 이는 분산안정화 방법을 더 연구·개발함으로써 제안된 추정량의 신뢰성을 개선할 수 있을 것으로 예상된다.

제6절 결론 및 논의

본 연구에서는 다양한 소스의 정보를 이용한 새로운 소지역 추정방법을 제안하였다. 효과적인 추정을 위해서 서로 다른 소스에 의한 정보를 잘 결합하여 사용하는 것은 매우 중요하고 실질적인 문제이다. 이러한 다양한 소스는 조사자료, 행정등록자료 또는 전수조사자료, 표본조사자료가 해당될 수 있다. 지금까지 모형기반 소지역 추정 연구에 있어서 보조변수는 표본오차 또는 측정오차가 없는 자료를 사용하는 것이 일반적이었고, 아주 최근에 이르러서 측정오차가 있는 보조변수를 이용할 수 있는 소지역 추정 방법에 관한 연구가 진행되어 왔다.

본 연구의 출발점은 경활조사, 지역별고용조사, 실업급여등록자료, 센서스 자료를 연결할 수 있는 소지역 추정방법을 개발한다는 것이고, 이때 지역수준 모형(area level model)을 사용하여 실업자수 또는 실업률을 추정하는 방법을 제시하는 것이었다. 이때 두 개의 표본조사인 경활조사와 지역별고용조사는 서로 독립일 수 있거나 두 조사 간의 공분산이 존재한다고 각각 가정할 수 있다. 이러한 시나리오 하에서 크게 두 가지 방법을 시도하였다.

- **경활조사와 지역별고용조사의 독립성을 가정한 실업자수 추정**

  먼저 3절에서 설명한바와 같이 경활조사, 지역별고용조사, 등록자료와의 결합을 통한 방법을 시도하였다. 이때 목표 추정치는 소지역 실업자수이고, 경활조사와 지역별고용조사는 서로 독립이라고 가정하였다. 실제 추정해 본 결과 실업급여등록자료의 결합(3절, com2)이 오히려 추정치의 정도를 약화시키는 것으로 나타났다. 즉, 소지역을 포함한 시도 수준에서 총 실업자수 추정치(3절 com2)가 경활조사 총 실업자수와 격차가 더 커지는 현상이 발생하였다.

  이에 대한 이유는 추정에 사용한 실업급여등록자수비율과 경활조사실업자비율 자료간의 구조적 관계가 매우 미약하기 때문인 것으로 이해될 수 있다. 또한 우리나라 실업급여등록자료의 커버리지가 임금근로자에 제한되어 있고, 등록과정에 발생하는 여러 가지 현실적인 문제들이 자료에 내포되어 있어서 제한된 모형을 통해 이러한 복잡한 현실성을 충분히 반영하지 못한 것으로 이해된다. 실은 이를 해결하기 위해 다양한 방법을 시도하였으나, 이 난점을 극복하지는 못했다. 그렇다하더라도 이 추정방법이 이론적으로 충분히 증명할 수 있다는 점에서 변수간의 설명력이 있는 자료에 대해서는 좋은 추정방법이 될 수 있을 것이다. 이러한 이유에서 첫 번째 제안 방법은 더 이상의 확장을 시도하지 않았다. 결국 실업급여등록자료를 활용하지 않는 새로운 모형 적합을 시도하였다.

- **경활조사와 지역별고용조사의 비독립성을 가정한 실업률 추정**

다음으로 4절에서는 3절과 달리, 경활조사와 지역별고용조사의 공분산성을 가정한 실업률 추정을 시도하였다. 실제로 지역별고용조사 자료 내에는 경활조사자료가 모두 포함되기 때문에 사실상 두 조사 자료가 완전히 독립이라고 보기는 어렵다. 따라서 경활조사, 지역별고용조사, 센서스 자료를 결합한 복합추정량을 제시하였다. 보조정보로 행정자료를 사용하지 않게 됨으로써 모형확장이 보다 쉽도록 모형을 설계하였다. 또한 시군구의 특성을 농촌과 도시로 구분하여 추정을 시도하였지만, 그룹으로 나누어 추정하더라도, 추정치 신뢰성이 크게 향상되지 않은 것으로 나타났다. 따라서 도시와 농촌으로 구분하지 않은 모형을 최종 추정 모형으로 사용하였다. 따라서 본 연구의 최종 추정결과는 4절에서 제시한 경활조사, 지역별고용조사, 센서스 자료를 이용한 GLS 추정량으로 복합추정량을 의미한다.

- **GLS 복합추정량**

본 연구에서는 측정오차 등을 포함하여 다양한 오차를 포함한 보조변수를 이용한 새로운 복합추정량을 제안하였다. 우리나라 230개 시군구의 실업률을 목표 추정치로 하였다. 추정방법은 지역수준 모형으로 구조오차모형을 이용하였다. 모형의 모수 추정은 적률법(MOM)에 의한 반복 추정 기법을 사용하였고, 추정량의 MSE는 잭나이프 방법을 사용하였다. 각 자료의 결합은 먼저 경활조사를 이용한 지역별고용조사의 합성추정량과 경활조사를 이용한 센서스의 합성추정량을 구한 후, 이들 추정량들과 경활조사 직접추정량과 연결하는 방법을 사용하였다. 이때 모형은 회귀절편이 0이 모형을 사용하였고, 지역별고용조사의 경우, MSE는 표본오차만을 그대로 사용하지 않고 이 조사의 편향을 고려하여 추정하였다.

추정결과, 제시한 복합추정량은 직접추정량의 MSE와 CV를 크게 개선하는 효과가 있었다. 또한 대규모 조사인 지역별고용조사의 MSE 또는 CV에 비해서도 상대적으로 낮은 값을 갖는 것을 알 수 있다. 그렇지만 소지역추정치의 신뢰성을 CV=30% 기준으로 설명할 때 여전히 많은 지역들에서 30% 이상 수준으로 나타났다. 이 지역들의 대부분은 표본 조사구수가 아주 작거나 실업률이 0인 지역들, 또는 표본수가 많은 지역들 중에서는 직접 추정량들의 분산이 큰 지역들이 해당된다. 본 연구에서 제안한 방법은 실업률이 '0'인 지역과 분산이 불안정한 지역들에서 잘 추정하지 못할 수 있다는 약점이 있다. 이 문제는 향후 분산안정화 방법을 개발하여 개선할 수 있다.

- **결론 및 향후과제**

  결론적으로 본 연구에서 제안한 방법이 이론적으로 충분이 좋은 방법으로 증명이 되었더라도, 실제 자료를 적용할 때 다양한 문제가 발생하는 것이 사실이다. 특히 직접추정량의 분산추정에 의존하는 경향이 있기 때문에 직접추정량의 분산추정이 매우 정확하게 계산되는 것이 중요하다. 앞으로도 본 연구에서 개발원 소지역 추정이 실제 자료에 잘 적용될 수 있도록 하기 위한 추가 연구가 지속되어야 할 것이다. 특히 CV 가 큰 지역을 중심으로 그 원인을 찾고, 이들 지역들이 갖는 특성, 직접추정량의 분산추정과 분산안정화를 위한 방법론 개발함으로써 소지역 추정치의 정도를 높일 수 있게 된다.

마지막으로, 본 연구는 다양한 소스자료를 이용한 소지역 추정의 개발 연구로서 우리나라 소지역 실업률 추정을 위한 새로운 방법을 제시했다는데 그 의미가 있겠다. 특히 이론적 방법을 개발하는 과정에서 실제 자료의 현실성을 반영하지 못한다는 것을 완전히 무시할 수는 없을 것이다. 실제 자료는 다양한 얘기치 못한 현상들을 포함하고 있고, 이를 고려하지 않은 이론 전개는 그 방법의 활용성 면에서는 그 가치가 덜할 수도 있기 때문이다. 본 연구에서 실제 자료에서 발생하고 있는 또는 발생할 수 있는 문제점들을 최대한 모형에 반영하고자 하였다. 그렇기 때문에 연구 과정에서 여러 가지 얘기치 못한 어려움에 직면하기도 하였다. 우리가 실제를 이론을 전개하는 과정에서 현실적이지 못한 가정을 하게 됨으로써 발생하는 많은 문제들을 경험할 수 있었다.

향후 추가 연구를 통해 현실적인 문제를 개선한다면 우리나라 고용자료에 적합한 고유의 소지역추정방법을 제시할 수 있게 될 것이다. 그리고 본 연구가 단순한 적용 연구가 아닌 새로운 방법을 제안한 이론 연구의 성격이 강하다는 점에서 학술적 차원의 좋은 아이디어도 제공할 수 있을 것이다.

참고문헌

Elliot, M.R. and Davis, W.W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: Combining data from two surveys. *Applied Statistics* 54, 595-609.

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

Fuller, W.A. (1987). *Measurement error models*, Wiley.

Fuller, W.A. (2009). *Sampling Statistics*, Wiley.

Lohr, L.S. and Prasad, N.G.N. (2003). Small area estimation with auxiliary survey data.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation, Journal of the Royal Statistical Society: Series B 68, 509-521.

Pfeffermann, D. (2002). Small area estimation - New developments and directions. International Statistical Review 70, 125-144.

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

Rao, J.N.K. (2003). Small Area Estimation. Wiley Series in Survey Methodology.

Yabarra, L.M.R. and Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error, Biometrika 95, 919-931.