중간보고서

소지역 추정법에 의한 시·군·구 실업통계 개발

2002. 9.

이 계 오 교수

목 차

제1장 서 론

'95년에 시작된 지방자치제도의 활성화로 시군구 단위의 경제활동인구 관련 통계의 요구가 빈번할 뿐 아니라 지식기반 정보화 사회로 발전하면서 통계의 활용성이 증대되면서 경제활동인구의 정확성과 시의성이 요구되고 있다.

현재 통계청에서 생산하는 경제활동인구에 대한 통계는 광역시와 도 단위에 대해서는 기본적인 사항인 취업자 수, 실업자 수와 비경제활동인구 수 등만을 생산하고 있으나 다른 세부항목인 연령계층별, 교육정도별 등에 대한 통계를 생산하지 않음으로써 광역자치단체별로 환경에 적합하고 효율적인 고용증진정책을 수립 추진하는데 애로 사항이 되고 있으므로 특·광역시와 도 단위의 세부항목에 대한 통계생산을 위한 소지역 추정법의 연구가 필요하다.

2000년 「인구주택 총 조사」 자료를 이용하여 가구조사 표본을 전면 개편하여 '03년부터 신규표본에 의해서 경제활동인구 관련 통계를 생산하게 되므로 소지역 추정법을 적용한 시군구단위의 통계기법을 연구하여 실제로 도입하지 않는다면 앞으로 또 다른 5년이 지나서야 시군구 실업통계의 생산이 가능할 것으로 사료되므로 소지역 추정법을 적용한 시군구단위의 경제활동인구 통계와 특·광역시와 도 단위의 세부항목에 대한 통계작성의 기법 뿐 아니라 이에 대한 알고리즘의 연구는 필수적일 것이다.

본 연구의 성공적인 추진을 위해서 새로 개편되는 경제활동인구 통계 관련 사항을 정리 요약하여 새로운 연구의 기본으로 활용할 것이다. 다음으로는 미국, 카나다 와 영국 등의 통계 선진국에서 고용통계분야의 소영역 통계생산에서 적용하는 소지역 추정법을 정리 분석하여 연구방향의 설정과 알고리즘을 연구 개발하는데 참고가 되도록 하겠다. 소지역 추정법의 일반적인 이론과 시군구 단위 경제활동인구 통계 작성에 필요한 절차를 연구 정리하였다. 다음에는 새로운 표본설계에서 적용될 수 있는 시군구 단위의 경제활동인구 통계작성 절차와 이에 대한 알고리즘을 연구하였으며, 알고리즘은 SAS언어로 구현하고 실제 계산결과를 통해서 타당성

과 실용성을 검토하였다. 마지막으로 특·광역시와 도 단위의 세부항목에 대한 통계생산도 소지역 추정법의 적용절차와 계산 알고리즘을 개발하여 이를SAS로 구현하고 수치적인 계산 결과를 통해서 타당성과 신뢰성을 입증하였다.

제2장 경제활동 통계작성 현황

2.1 표본설계 및 추정

2002년에 개편하는 경제활동인구 조사의 표본설계 내용을 요약 분석하여 표본설계와 추정에 관한 내용을 기술하겠음

2.2 표본관리 및 오차관리

경제활동인구조사의 신규 표본설계에서 표본가구 및 오차관리에 관해서 요약 정리할 것임

2.3 소지역(시군구) 추정 필요성

o 시군구가 지방자치행정의 기초단위이므로 지역적인 특성을 반영한 지역행정을 펴기 위해서 경제활동인구 관련 통계는 필수적임

o 지역간의 균형적인 발전을 위한 정책을 펴기 위해서 시군구의 노동력 통계는 참고자료이며 필수적인 정보임.

o 지식기반 정보화 사회에서 정확하고 시의성을 갖춘 통계의 활용성을 증대시키기 위해서 기초행정단위인 시군구의 노동력 관련 통계는 일정 수준의 정확성을 갖도록 생산해야함

* 신규 경제활동인구 조사의 표본설계 내용을 요약정리 후에 연계하여 시군구 단위의 노동력 통계의 필요성에 관해서 기술할 것임

제3장 소지역 추정법에 의한 노동통계 작성사례

3.1 미 국

3.1.1 서 론

1950년대에 미 노동성 고용훈련 행정국에서는 각 주(state)들에 대한 실업률 추정값들을 비교할 수 있도록 실업관련 통계 추정기법을 개발하여 소책자로 발간하였다. 이 소책자를 근거로 하여 1950년대 후반에는 소규모 표본조사를 통한 연속적인 주 단위 실업통계가 작성되었다. 1950년대의 실업통계는 주로 실업보험(UI:unemployment insurance) 자료를 이용하여 작성되었다.

1972년 미 노동 통계국에서 주 단위에 대해서 이용 가능한 노동력, 취업자 및 실업자 추정에 대한 방법과 개념을 연구하기 시작하였고, 1973년에는 통계국이 주관이 되어, 경상인구조사(CPS:Current Population Survey)의 개념, 정의, 추정과 이전의 소책자방법을 결합하여 주 단위와 주의 세부단위까지 노동력 관련 통계를 추정할 수 있는 기법을 개발하였다.

1976년 이후에는 모든 주 단위 실업 관련 통계의 추정값에 대한 신뢰도를 높이기 위해서 각 주별로 표본 가구 수를 몇 배씩 증가시켰으며, 이후부터 CPS 자료를 이용한 노동력 관련 추정값을 공식적으로 발표하기 위해 실업 관련 추정값에 대한 변동계수의 최대 허용기준을 설정하였다(실업률을 6%라고 가정했을 때, 실업 관련 통계에 대한 변동계수의 최대 허용값을 10%로 함). 1978년부터 규모가 큰 10개주(캘리포니아, 플로리다, 일리노이스, 메사츄셑, 미시간, 뉴저지, 뉴욕, 오하이오, 펜실베니아, 텍사스)와 2개 지역(로스엔젤레스, 뉴욕시)의 노동력 관련 통계는 CPS자료만으로 추정된 결과를 공식 통계로 사용토록 하였다.

UI 신청자료의 데이터베이스를 지속적으로 유지하고 개선하였으며, 1976년과 1978년 사이에 걸쳐 UI 신청자료를 모든 주에 대해서 표준화하고, CPS조사의 조사 주 간(매월 12일을 포함한 주)에 실업자로 인정받는 UI 신청자는 자동으로 데이터베이스에 등록되도록 하는 데이터베이스 관

리체계를 개발하여 CPS자료와 연계한 추정법을 개발하였다.

1985년에는 1980년 센서스 자료를 근거로 하여 주 단위들에 대한 CPS 표본설계를 완성하였으며, 이 때 추가적으로 노스캐롤라이나 주를 CPS 자료에서 노동력 관련 통계를 직접 추정하여 공표 하는 주로 포함시켰고, 표본크기가 충분한 총 11개 대규모 주의 실업 관련 통계에 대한 목표변동계수를 8%로 낮추었다. 또한 나머지 39개 주의 연평균 실업통계에 대한 목표변동계수를 8%로 정하였으며, 이 때 실업률의 참값은 6%인 것으로 가정하였다.

표본크기가 충분한 11개 주 이외의 39개 주와 주 내의 부차관심영역들에 대한 공식적인 월별 추정값들은 1989년 이전까지는 노동성에서 개발한 소책자 추정방법에 의해 계산하였으나, 1989년부터 시계열 모형을 이용한 추정기법을 표준 추정법으로 채택하였다. 1992년에는 추정값들의 시계열들에 대해 계절요인 보정을 적용하였고 1994년에는 1990년 센서스 자료를 이용하여 CPS를 재 설계하였으며, 또한 모든 조사에서 컴퓨터 보조면접을 실시할 수 있도록 설문지를 재구성하였다. 새로 개편된 CPS의 표본설계는 1995년 중반까지 단계적으로 도입하였다.

CPS는 매월 미국 내의 노동력의 동향과 이와 관련된 세부항목들(연령대별, 성별, 인종별 등)에 대한 특성을 파악하기 위해 실시되는 경상인구조사로써 층화 다단확률추출에 의해 추출된 가구단위들에 대해 조사가 이루어진다. CPS의 최종추출단위는 가구조사 단위들로써 약 50,000 조사가구들에 대해 조사가 실시되며, 조사가구들은 총 8개의 패널로 구성되어 표본으로 관리된다. CPS의 표본교체방식은 4-8-4체계를 따른다. 표본으로 추출된 최종 추출단위(가구단위)들은 4개월 간 연속조사가 진행되며, 이후 8개월 간 조사에서 제외되었다가 다시 4개월 간 연속조사가 진행된 후 표본목록에서 영구히 삭제된다.

매달 8개 패널의 표본가구에 거주하는 만 15세 이상의 사람들을 대상으로 노동력조사가 이루어지며, 노동력 관련 최종 추정값들은 여러 단계의 추정과정을 통해 작성된다. CPS 자료로부터 전국 단위와 주 단위에

대한 노동력 관련 추정값들을 작성하기 위한 개략적인 추정과정은 다음과 같다.

(a) CPS 가중치를 이용하여 노동력 관련 추정값들의 불편성조정

(b) 무응답에 대한 보정

(c) 1차 추출단위들(PSUs)의 분산을 줄이기 위한 일단계 비보정 (First-stage ratio adjustment)

(d) CPS 추정값들의 분산을 줄이기 위한 이단계 비보정(Second-stage ratio adjustment)

(e) 추정값들의 분산을 줄이기 위해 전 달의 조사자료를 이용하는 복합 추정

(f) 주요한 노동력 통계에 대한 계절요인 조정(Seasonal adjustment)

3.1.2 불편성 조정

대부분의 표본조사에서와 마찬가지로 CPS에서도 무응답이 발생한다. 미국 CPS에서의 단위 가구들에 대한 월별 무응답은 평균적으로 6~7%정도이다. 이러한 무응답이 편향의 잠재적인 요인으로 작용하게 된다. 이 밖에 표본선택 과정에서 발생하는 단위 가구들에 대한 누락이라든가 면접자에 의해 가구 단위 또는 가구 내의 개인 조사단위가 누락될 경우 등이 편향의 요인으로 작용될 수 있다. 이러한 편향의 잠재적인 가능성을 최소화하기 위해 추가적으로 무응답에 대한 보정이 추정과정에 포함된다.

표본 내에 있는 모든 단위들이 동일한 추출확률을 갖는다면 이러한 표본은 자체가중치(Self-weighting)를 갖는 표본으로 분류되며(동일한 추출확률을 갖지 않을 경우는 비자체가중치(Nonself-weighting)를 가짐), 불편추정량은 표본총계에 추출확률의 역수를 곱하여 계산된다. CPS에서 대부분의 주(state) 지역 표본들은 자체가중치를 갖는 지역들에 해당된다.

(1) 기본 가중치

CPS는 총 792개 PSU(Primary sampling units) 지역(층)들로 구성되며, 이들은 432개의 자체가중치를 갖는 SR(Self-representing) 지역(층)들과 360개의 비자체가중치를 갖는 NSR(Nonself-representing) 지역(층)들로 구분된다. CPS에서 주별 SR 및 NSR 지역(층)들의 수 및 추정인구는 다음 <표3.1>에 주어졌다. 여기에서 추정인구는 1990년 센서스에 근거하여 추정된 결과이다.

<표3.1> 각 주별 층의 수 및 추정인구

State	SR 지역		NSR 지역		전체 추출간격
	층의 수	추정인구	층의 수	추정인구	
Alabama	4	1,441,118	10	1,604,901	2,298
Alaska	5	287,192	5	91,906	336
Arizona	3	2,188,880	4	544,402	2,016
Arkansas	4	612,369	13	1,151,315	1,316
California	21	20,866,697	5	1,405,500	2,700
Los Angeles	1	6,672,035	0	0	1,691
Remainder	20	14,194,662	5	1,405,500	3,132
Colorado	5	1,839,521	5	629,640	1,992
Connecticut	20	2,561,010	0	0	2,307
Delaware	3	509,330	0	0	505
District of Columbia	1	486,083	0	0	356
Florida	20	9,096,732	8	1,073,818	2,176
Georgia	9	2,783,587	9	2,038,631	3,077
Hawaii	3	778,364	1	49,720	769
Idaho	8	416,284	8	302,256	590
Illinois	11	6,806,874	12	1,821,548	1,810
Indiana	9	2,215,745	8	1,949,996	3,132
Iowa	5	710,910	11	1,372,871	1,582
Kansas	3	920,819	9	906,626	1,423
Kentucky	9	1,230,062	9	1,546,770	2,089
Louisiana	7	1,638,220	9	1,403,969	2,143
Maine	8	685,937	4	247,682	838
Maryland	4	3,310,546	2	352,766	3,061
Massachusetts	31	4,719,188	0	0	954
Michigan	12	5,643,817	11	1,345,074	1,396
Minnesota	4	2,137,810	7	1,119,797	2,437
Mississippi	6	540,351	14	1,334,898	1,433

Missouri	7	2,394,137	6	1,457,616	3,132
Montana	8	366,320	8	221,287	431
Nebraska	2	557,203	9	608,513	910
Nevada	4	819,424	2	98,933	961
New Hampshire	13	846,029	0	0	857
New Jersey	11	6,023,359	0	0	1,221
New Mexico	6	685,543	8	410,929	867
New York	13	12,485,029	7	1,414,548	1,709
New York City	1	5,721,495	0	0	1,159
Remainder	12	6,763,534	7	1,414,548	2,093
North Carolina	14	3,089,176	23	1,973,074	1,095
North Dakota	4	232,865	9	235,364	363
Ohio	13	6,238,600	13	1,958,138	1,653
Oklahoma	2	1,244,920	11	1,092,513	1,548
Oregon	4	1,396,261	6	761,794	1,904
Pennsylvania	14	7,704,963	11	1,507,946	1,757
Rhode Island	5	784,090	0	0	687
South Carolina	7	1,434,621	7	1,161,298	2,291
South Dakota	5	211,146	11	291,546	376
Tennessee	8	2,502,671	6	1,219,915	3,016
Texas	22	8,779,997	20	3,613,170	2,658
Utah	2	888,524	3	251,746	958
Vermont	11	428,263	0	0	410
Virginia	12	3,093,947	6	1,612,667	3,084
Washington	5	2,437,454	6	1,216,557	2,999
West Virginia	10	803,797	9	581,544	896
Wisconsin	7	1,773,109	9	1,889,141	2,638
Wyoming	8	227,401	6	98,321	253
Total	432	141,876,295	360	45,970,646	2,060

<표3.1>에 주어진 해당 주 별 표본 추출간격을 CPS 기본가중치로 이용한다. 같은 주 내에 있는 대부분의 표본조사 단위들은 동일한 추출확률을 가지나 New York과 California 내에서는 예외적으로 서로 다른 추출확률이 적용된다. 추정과정의 첫 번째 단계에서는 단순 집계된 가구 조사단위들에 대해서 위의 기본가중치가 곱해진다. 인구의 증가에 따라 가구단위들이 증가할 경우 원래의 표본추출간격은 조사 비용과 정확도의 측면을 고려하여 정기적으로 보정된다.

(2) 특별 가중치

CPS에서 최종 추출단위들은 보통 4가구로 구성된다. 경우에 따라서는 최종 추출단위 내에 4가구 이상의 조사가구들이 포함되어 있는 경우가 있을 수 있다. 예를 들면 다음과 같은 예이다. CPS에서 하나의 최종 추출단위 지역에는 4개의 조사가구가 배정되는 것이 일반적이지만 조사자가 면접 당시 하나의 최종 추출단위 지역에 36개 가구단위가 존재하였다고 가정해 보자. 조사자의 업무 부담을 줄이고자 3가구 당 하나의 가구를 부차적으로 추출하여 총 12개 가구를 조사했을 경우 CPS에서는 특별가중치를 3으로 산정하여 해당 지역의 기본가중치에 곱하여 추정값들이 불편성을 만족하도록 보정해 준다. CPS에서는 추정값의 편향과 분산을 고려하여 특별가중치가 4를 넘지 않도록 명시하고 있다.

3.1.3 무응답에 대한 보정

CPS에서는 무응답을 크게 항목무응답(Item nonresponse)과 단위무응답(Unit nonresponse)의 두 가지로 구분한다. 여기에서 단위무응답은 해당 표본단위로부터 조사자료를 수집하지 못한 경우를 의미한다. 예를 들면, 피면접자의 응답 거부 또는 피면접자의 부재 등의 사유로 조사자료를 수집하지 못한 경우 등이 여기에 해당된다. CPS에서는 이러한 단위무응답을 Type A 무응답이라 부르며 매월 약 4~5%의 Type A 무응답이 발생한다. 최근에 들어서는 Type A 무응답의 비율이 6~7%정도로 다소 증가하는 경향을 보인다. CPS에서는 단위무응답 요인들을 고려하여 가중치에 대한 보정을 실시한다.

단위무응답에 대한 보정은 다음과 같이 이루어진다. 우선 해당 주에서 광역통계구역(metropolitan statistical area: MSA)과 크기가 비슷한 1차 추출단위(PSU)들을 한데 묶어 MSA 집락으로 구분하고, 그렇지 못한 PSU들은 non MSA 집락으로 구분한 후, 각 집락 내에서는 두 개의 무응답보정 셀을 형성한다. MSA 집락은 "도시 중심지역"과 "기타 지역"으로, non-MSA 집락은 "도시 지역"과 "농촌 지역"으로 무응답보정 셀이 형성

된다. CPS에서는 총 254개의 무응답보정 셀에 대해서 단위무응답에 대한 보정이 실시되며, 각 무응답보정 셀에 대한 무응답보정값 F_{ij}는 다음과 같이 산정된다.

$$F_{ij} = \frac{Z_{ij} + N_{ij}}{Z_{ij}},$$

여기에서 "$Z_{ij} = i$집락의 j셀에서 조사된 가구단위들의 가중총계", "$N_{ij} = i$집락의 j셀에서 조사된 가구단위들의 Type A 무응답 가중총계"를 나타낸다.

조사된 개인단위들은 위에서 산정된 무응답보정값 F_{ij}가 곱해져서 단위무응답에 대한 보정이 이루어진다. 참고적으로 CPS에서는 F_{ij}가 2이상인 경우 또는 해당 셀에서 조사된 무응답 조사가구가 50보다 작을 경우에는 예외적인 보정절차가 시행된다. 무응답보정이 이루어진 후 각 개인 조사단위들에 대한 가중치는 "기본가중치×특별가중치×무응답보정값"이 적용된다.

3.1.4 비 보정(Ratio Adjustment)

(1) 일단계 비보정

일차추출단위(PSU)에서 발생하는 주 수준(state level) 추정값들의 분산을 줄이기 위한 목적으로 일단계 비보정이 실시된다. 일단계 비보정값들은 1990년 센서스 자료에 근거하여 산정되며, 비자체가중치를 갖는 NSR 지역(층)의 PSU들에 대해서 적용된다. NSR 지역(층)의 PSU들을 포함하고 있는 각 주별로 두 가지의 인종별 범주(Black, non-Black)에 대해서 일단계 비보정값들이 계산되며, 산정공식은 다음과 같이 주어진다.

$$FS_{sj} = \frac{\sum_{i=1}^{n} C_{sij}}{\sum_{k=1}^{m} \left(\frac{1}{\pi_{sk}}\right) C_{skj}},$$

여기에서

FS_{sj} = s주의 j인종 셀에 대한 일단계 비보정값(j=Black, non-Black),

C_{sij} = s주의 j인종 셀에서 NSR 지역(층)의 전체 PSU 중 i PSU에 대한 만 15세 이상의 1990년 센서스 인구,

C_{skj} = s주의 j인종 셀에서 NSR 지역(층)의 표본 PSU 중 k PSU에 대한 만 15세 이상의 1990년 센서스 인구,

π_{sk} = s주에서 k PSU에 대한 1990년 추출확률,

n = s주에서 NSR 지역(층)에 있는 PSU들의 전체 개수,

m = s주에서 NSR 지역(층)의 표본 PSU들의 개수.

만약, s주에서 일단계 비보정값이 1.3보다 크거나 1/1.3보다 작은 경우, NSR지역(층)의 표본 PSU들이 4개보다 작은 경우에 대해서는 인종 셀의 범주를 Black 또는 non-Black으로 구분하지 않고 일단계 비보정값을 공통적으로 1을 부여한다. 자체가중치를 갖는 SR 지역(층)에 대해서도 마찬가지로 일단계 비보정값은 1이 부여된다. CPS의 일단계 비보정값들은 다음 <표3.2>에 주어졌다. 일단계 비보정이 이루어진 후 각 개인 조사단위들에 대한 가중치는 "기본가중치×특별가중치×무응답보정값×일단계 비보정값"이 적용된다.

<표3.2> 각 주별 일단계 비보정값(1999년 CPS 기준)

State	Black	non-Black
Alabama	0.92986976	1.02360321
Alaska	**1.00000000	**1.00000000
Arizona	**1.00000000	**1.00000000
Arkansas	1.03854268	0.99625041
California	0.92824011	1.01550280
Colorado	**1.00000000	**1.00000000
Connecticut	*1.00000000	*1.00000000
Delaware	*1.00000000	*1.00000000
District of Columbia	*1.00000000	*1.00000000
Florida	1.07779736	1.00025192
Georgia	1.06981965	0.98237807

Hawaii	**1.00000000	**1.00000000
Idaho	**1.00000000	**1.00000000
Illinois	1.01947743	1.00254363
Indiana	1.16715920	0.99747055
Iowa	**1.00000000	**1.00000000
Kansas	**1.00000000	**1.00000000
Kentucky	1.09352656	0.99897341
Louisiana	1.04956759	0.98344772
Maine	**1.00000000	**1.00000000
Maryland	**1.00000000	**1.00000000
Massachusetts	*1.00000000	*1.00000000
Michigan	0.92441097	0.99798724
Minnesota	**1.00000000	**1.00000000
Mississippi	0.98243024	1.00997154
Missouri	**1.00000000	**1.00000000
Montana	**1.00000000	**1.00000000
Nebraska	**1.00000000	**1.00000000
Nevada	**1.00000000	**1.00000000
New Hampshire	*1.00000000	*1.00000000
New Jersey	*1.00000000	*1.00000000
New Mexico	**1.00000000	**1.00000000
New York	0.83647167	1.00163779
North Carolina	1.07378643	0.98057928
North Dakota	**1.00000000	**1.00000000
Ohio	0.97362223	1.00000085
Oklahoma	1.01775196	1.00400963
Oregon	**1.00000000	**1.00000000
Pennsylvania	1.17560284	0.99367587
Rhode Island	*1.00000000	*1.00000000
South Carolina	0.93971915	1.05454366
South Dakota	**1.00000000	**1.00000000
Tennessee	1.08638935	0.98680199
Texas	1.23277658	0.97475648
Utah	**1.00000000	**1.00000000
Vermont	*1.00000000	*1.00000000
Virginia	**1.00000000	**1.00000000
Washington	**1.00000000	**1.00000000
West Virginia	1.22587622	0.99959222
Wisconsin	**1.00000000	**1.00000000
Wyoming	**1.00000000	**1.00000000

* 자체가중치를 갖는 SR 지역(층) (총 8개 주)

** 인종 셀의 범주(Black, non-Black)가 합쳐진 지역 (총 23개 주)

(2) 이단계 비보정

이단계 비보정은 노동력과 관련된 세부항목과 관심영역 및 부차관심영역들에 대한 표본 추정값들의 추정오차를 줄이기 위할 목적으로 실시되며, 각 패널 내의 표본 가중치들에 대해서 보정이 이루어진다. CPS에서는 이단계 비보정 시 매월 서로 독립적으로 기록 유지되는 다음과 같은 세 가지 대조그룹들에 대한 정보들을 이용한다.

- o 50개 주와 콜롬비아 구에 대한 만 15세 이상의 추계인구
- o 라틴 아메리카 계의 14개 성별-연령대별 추계인구, 비 라틴 아메리카 계의 5개 성별-연령대별 추계인구(세부범주는 <표3.3>을참조)
- o 백인에 대한 66개 성별-연령대별 추계인구, 흑인에 대한 42개 성별-연령대별 추계인구, 기타 10개 성별-연령대별 추계인구(세부범주는 <표3.4>를 참조)

<표3.3> 라틴아메리카계(Hispanic)/비 라틴아메리카계 (non-Hispanic)에 대한 성별-연령대별 범주

Ages	Hispanic	
	Male	Female
0-5		
6-13		
*14		
*15		
16-19		
20-29		
30-49		
50+		

Ages	non-Hispanic
*0-5	
*6-13	
*14	
*15	
*16+	

* 표시는 성별 구분이 없는 범주를 나타냄

<표3.4> 백인(White)/흑인(Black)/기타에 대한 성별-연령대별 범주

Ages	White male	White female
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		
10-11		
12-13		
14		
15		
16		
17		
18		
19		
20-24		
25-26		
27-29		
30-34		
35-39		
40-44		
45-49		
50-54		
55-59		
60-62		
63-64		
65-67		
68-69		
70-74		
75+		

Ages	Black male	Black female
0-1		
2-3		
4-5		
6-7		
8-9		
10-11		
12-13		
14		
15		
16-17		
18-19		
20-24		
25-29		
30-34		
35-39		
40-44		
45-49		
50-54		
60-64		
65+		

Ages	Other male	Other female
0-5		
6-13		
*14		
*15		
16-44		
45+		

*는 성별 구분이 없는 범주를 나타내며, 기타 인종은 아메리카 인디언계, 에스키모 계, 알류트 계, 아시아 계, 태평양 인근 섬 주민들이 포함됨.

이단계 비보정은 월 단위 표본을 구성하는 8개의 패널들 각각에 대해서 개별적으로 이루어진다. 한편, 각 대조그룹에서 패널별 추계인구와 표본 추정값 간의 비보정(ratio adjustment)값들은 대조그룹들에 따라서 약간의 차이가 발생한다. CPS에서는 세 가지 대조그룹들에 대한 정보를 동

시에 비보정값 산정에 포함시킬 수 있는 레이킹 비 추정량(raking ratio estimator)을 개발하여 이단계 비보정값들을 산정하고 있다. 레이킹 비추정량에 대해서는 Ireland and Kullback(1968)에 보다 상세히 소개되어 있다.

이단계 비보정이 이루어진 후 각 개인 조사단위들에 대한 가중치는 "기본가중치×특별가중치×무응답보정값×일단계 비보정값×이단계 비보정값"이 적용된다. CPS에서는 위의 가중치가 적용된 추정값을 FSC(first- and second-stage combined)추정값이라 부른다.

3.1.5 복합 가중치(Composite Weights)

CPS에서 복합가중치 산정은 두 단계에 걸쳐 이루어진다. 우선 인구통계적 특성에 의해 분류된 주요 노동력 범주들에 대한 복합추정값(composite estimates)들이 첫 번째 단계에서 추정된다. 다음으로 각 범주별 개인조사단위에 대한 가중합이 추정된 복합추정값과 일치하도록 일련의 비보정 과정을 통하여 보정된다.

먼저 복합추정값 산정 단계에 대해서 살펴보도록 하자. 일반적으로 복합추정값들은 여러 추정값들의 가중평균의 형식을 취한다. CPS에서의 복합추정값은 다음 두 개의 추정값들을 결합하여 작성된다. 두 추정값들 중 하나는 앞 절에서 언급한 FSC추정값이다. 나머지 하나의 추정값은 바로 전 달의 복합추정값과 전 달과 금번 달의 변화량에 대한 추정값의 항을 결합하여 표현한다. 여기에서 월 변화량에 대한 추정값은 금번 달과 바로 전 달의 CPS 조사표본 중 동일한 표본가구들에 대해 연속적으로 조사되는 약 75%의 조사표본 자료에 근거하여 작성된다(CPS 조사표본의 1개월 간격 중복율은 약 75%임).

t번째 달의 실업자 총계에 대한 복합추정값 Y_t'의 추정공식은 다음과 같이 주어진다.

$$Y_t' = (1-K)\,\hat{Y}_t + K\,(Y_{t-1}' + \triangle_t) + A\,\hat{\beta}_t\ .$$

여기에서

$$\hat{Y}_t = \sum_{i=1}^{8} x_{t,i} \ ,$$

$$\triangle_t = \frac{4}{3} \sum_{i \in S} (x_{t,i} - x_{t-1,i-1}) \ ,$$

$$\hat{\beta}_t = \sum_{i \in S} x_{t,i} - \frac{1}{3} \sum_{i \in S} x_{t,i} \ ,$$

$i = 1, 2, \cdots, 8,$

$x_{t,i} = t$번째 달의 i표본에서 응답자들에 대한 이단계 비보정 후의 가중합,

$S = \{2, 3, 4, 6, 7, 8\}$: 전 달과 공통으로 조사되는 6개의 패널들의 집합,

$$K = \begin{cases} 0.4 \ , & \text{실업자총계 추정 시} \\ 0.7 \ , & \text{취업자총계 추정 시} \end{cases},$$

$$A = \begin{cases} 0.3 \ , & \text{실업자총계 추정 시} \\ 0.4 \ , & \text{취업자총계 추정 시} \end{cases}.$$

위에서 명시된 상수 A와 K는 실업자 및 취업자 총계에 대한 월간 변화량의 추정치의 분산을 최소화하는 최적가중치들이다. 상수 K는 금번 달의 FSC추정치 및 전 달의 복합추정치 Y'_{t-1}와 금번 달과 전 달의 변화량에 대한 추정치 $\triangle_t$의 합에 대한 가중치를 결정한다. 여기에서 변화량에 대한 추정치 $\triangle_t$는 금번 달과 전 달의 공통 조사표본인 6개의 패널의 조사자료에 근거하여 추정된다. 상수 A는 $\hat{\beta}_t$의 가중치를 결정하며, 여기에서 $\hat{\beta}_t$는 복합추정량의 분산과 표본 편향을 줄이기 위하여 산정되는 보정항이다(자세한 내용은 Breau and Ernst(1983)과 Bailar(1975)를 참조).

복합가중치는 주요 노동력 범주인 실업자(UE:unemployed), 취업자(E:employed), 비노동인구(NILF:not in labour force)의 세가지 범주에 대해서 각각 작성되며 실업자에 대한 복합가중치 산정절차를 설명하면 다음과 같다.

(i) j번째 주(50개 주와 1개의 콜롬비아 구로 구성된 총 51개의 셀)에 대해서 실업자 총계에 대한 복합추정값 $comp(UE_j)$를 계산한다. 같은 방법으로 전국 9개의 연령대별-성별-인종별 셀(라틴 아메리카 계에 대한 8개의 연령대별-성별 셀, 비 라틴 아메리카 계에 대한 1개의 연령대별-성별 셀들로 구성)들과 66개 연령대별-성별-인종별 셀(백인에 대한 38개의 연령대별-성별 셀, 흑인에 대한 24개의 연령대별-성별 셀, 기타 4개의 연령대별-성별 셀들로 구성)들에 대한 복합추정값들을 계산한다(연령대별-성별-인종별 셀에 대해서는 <표3.3>과 <표3.4>를 참조). 상수K와 A는 실업자 총계 추정에서는 모든 범주들에 대해서 $K=0.4$, $A=0.3$이 적용된다.

(ii) 조사된 표본자료를 각 주별로 분류하고, j번째 주 내에서 실업자 총계에 대한 단순추정값 $simp(UE_j)$을 계산한다. 여기에서 $simp(UE_j)$는 이단계 비보정 후의 가중치가 적용된 추정값이다.

(iii) j번째 주 내에서 실업자에 해당하는 각 개인 조사단위가 갖는 가중치에 다음과 같은 비보정 가중치를 곱한다.

$$\frac{comp(UE_j)}{simp(UE_j)}$$

(iv) 표본조사자료를 연령대별-성별-인종별로 교차분류한 후, 각 셀 내에서 실업자 총계에 대한 단순추정값을 계산한다.

(v) 각각의 연령대별-성별-인종별 셀 내에서 실업자에 해당하는 개인 조사단위들은 절차(iii)과 같은 방법으로 가중치 보정이 이루어진다.

(vi) 절차(iv)와 절차(v)의 과정이 연령대별-성별-인종별 셀들에 대해서 반복된다.

(vii) 절차(ii)에서 절차(vi)까지의 과정이 5회 더 반복된다. 따라서의 총 6회의 반복이 이루어진다.

실업자 및 비노동력 인구에 대해서도 위와 같은 방법이 동일하게 적용

되며, 각 개인조사단위들은 위에서 언급된 복합가중치 적용절차가 끝난 후의 가중치를 최종가중치로 적용받게 된다.

3.1.6 계절요인 보정(seasonal adjustment)

시계열은 시간에 따라 측정 가능한 관측치 또는 추정값들의 계열을 말한다. CPS 자료에 기반을 두고 있는 많은 월별 노동력 통계 자료는 시계열 자료 분석방법에 의해 계절요인이 보정되어 노동력 통계의 시의성 및 정확도를 배가시키고 있다.

계절요인의 보정은 날씨, 휴일, 학기 일정 등과 같은 일상적인 계절 요인 효과를 시계열로부터 측정하여 노동력 지표들을 조정하기 위한 목적으로 실시한다. 실업 및 취업 등과 같은 다양한 추정치들의 월별 변화 양상도 계절요인으로 보다 쉽게 설명될 수 있으며, 주요 경제지표들에 대한 장기순환 및 추이성에 대한 조정도 계절요인의 분석에 의해 이루어질 수 있다. 예를 들어 5월과 6월의 실업 및 취업자 총계에 대한 추정값들을 고려해 보자. 이 경우 6월의 실업 및 취업자 총계는 5월보다는 현격한 증가가 발생할 수 있으며, 학기가 끝나는 5월말을 기준으로 6월에는 노동력인구에 학생들이 유입되기 때문에 경제지표에 대한 장기적인 순환과 추이성에는 영향을 미쳤다고 보기는 어렵다. 따라서 이러한 시의성이 반영되어 6월의 시계열은 5월의 자료에 근거하여 유사하게 보정된다.

계절요인 보정에 대한 연구는 오랜 기간 동안 꾸준히 진행되어오고 있다. 노동력 관련 시계열들의 계절요인 보정을 위해 1980년 이후 지금까지 $X-11$ ARIMA 프로그램이 이용되고 있다. 이 프로그램은 1960년대에 미 센서스 국에 의해 개발된 $X-11$ 프로그램을 기반으로 하여 1970년대 말 캐나다 통계청에 의해 보완된 시계열 보정 프로그램이다. 공식적으로 일년에 두 번(6월과 12월) 실업자, 취업자 및 실업률 등과 같은 노동력 관련 시계열들의 계절요인 보정 추정값들이 $X-11$ ARIMA를 통해 얻어진다.

3.1.7 주(state) 및 주 내의 부차관심영역에 대한 추정

대영역인 51개 주와 주 내의 부차관심영역인 소지역들에 대한 실업관련 추정값들은 각 지역의 경제상황을 나타내는 중요한 지표로 활용된다. 미 노동 통계국에서는 주별 협력 프로그램 하에서 매월 약 6,950개 지역들에 대해서 노동력 및 실업관련 추정값들을 작성하고 있다. 미국 내의 모든 주, 광역시, 구, 인구 25,000이상의 시지역, New England 내의 모든 시지역들과 지방 중심지들이 이러한 추정대상 지역들에 해당된다. 주 및 주 내의 소지역들에 대한 표본의 할당 및 추정결과의 공표뿐만 아니라 이들 지역들의 예산 보조에 대한 적격성 여부는 연방 프로그램에 근거하여 결정된다.

주 및 지방정부는 고용정책 입안과 예산 책정 시 주 및 부차관심영역들에 대한 노동력 관련 추정값들을 활용하며, 이러한 추정값들은 대부분 CPS자료에 근거하여 작성되나 CPS 표본이 충분하지 않은 관심영역들에 대해서는 경상고용통계(CES:current employment statistics) 자료와 실업보험(UI:unemployment insurance) 자료가 추가적으로 추정과정에 활용된다.

주 지역들에 대한 노동력관련 연 평균 추정값들은 CPS자료로부터 직접적으로 산출된다. 또한 충분한 표본을 갖고있는 캘리포니아, 플로리다, 일리노이스, 메사츄셑, 미시간, 뉴저지, 뉴욕, 노스캐롤라이나, 오하이오, 펜실베니아, 텍사스의 11개 주 지역들과 뉴욕시와 로스엔젤레스-롱비치의 2개 지역들에 대해서는 CPS 월 추정값들이 활용된다.

나머지 39개 주들과 콜롬비아 구에 대해서는 CPS 표본이 충분하지 않기 때문에 CES 자료와 UI 자료를 결합시킨 시계열모형으로부터 월 추정값들이 작성된다. 이러한 지역들에 대한 시계열 추정방법을 소개하면 다음과 같이 요약된다.

시점 t에서 CPS 노동력관련 추정값 $y(t)$는 다음과 같은 두개의 독립적인 확률과정들의 합으로 표현된다.

$$y(t) = \theta(t) + e(t). \tag{3.1}$$

여기에서 $\theta(t)$는 노동력 참값이고 $e(t)$는 표본 오차를 나타낸다. 이 모형을 통해 CPS 복합추정량보다 추정오차가 작은 $\theta(t)$를 추정하는 것이 기본적인 목적이다.

(3.1)식에서 시그널 항인 $\theta(t)$는 다음과 같은 모형을 통해 추정된다.

$$\theta(t) = \theta'(t) + \eta(t)$$

여기에서

$$\theta'(t) = X(t)\beta(t) + T(t) + S(t) \ ,$$

$X(t) =$ 기지인 설명변수들의 벡터,

$\beta(t) =$ 랜덤 계수 벡터,

$T(t) =$ 추이 항(trend component),

$S(t) =$ 계절요인 항(seasonal component),

$\eta(t) =$ 노이즈 항(noise componet).

이러한 항들의 각각에 대한 확률과정적인 특성들은 서로 독립적인 정규분포를 따르는 하나이상의 백색 잡음 항(white noise component)들에 의해 결정된다. 각 항들을 j로 표현할 때 백색 잡음 항들의 분포는 다음과 같이 가정된다.

$$v_j(t) \sim NID(0\ ,\ \sigma_{v_j}^2\ )$$

여러 개의 항들 중 우선 회귀 항을 살펴보자. 회귀 항을 $M(t)$라 놓을 때 이 항은 시간에 따라 변화하는 계수 $\beta(t)$와 설명변수들 $X(t)$로부터

$$M(t) = X(t)\beta(t)$$

와 같이 주어진다. 위의 회귀 항에 포함된 독립변수들은 관측가능한 변수들이며 CPS 표본 오차와 서로 독립인 UI 자료, CES 자료, 표본 센서스 자료들로부터 선택된다. 회귀계수 $\beta(t)$는 다음과 같은 랜덤워크과정으로 모형화된다.

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

여기에서

$$v_\beta(t) \sim NID(0\,,\,\sigma_\beta^2\,)$$

로 주어진다. 한편, 선택된 설명변수들은 CPS 추정값들의 중요한 변동을 모형에 반영하고 있지만 비표본 오차에 대한 변화량은 포함하고 있지 않기 때문에 추가적으로 추이 항과 계절요인 항들이 모형에 포함된다.

추이 항 $T(t)$는 다음과 같이 랜덤인 $T(t)$와 기울기 $R(t)$의 일차근사로 주어진다.

$$T(t) = T(t-1) + R(t-1) + v_T^*(t)\ ,$$

$$R(t) = R(t-1) + v_R(t).$$

여기에서

$$v_T^*(t) = \sum_{k=1}^{m} \delta_k \xi_k(t) + v_T(t)\ ,$$

$$\xi_k(t) = \begin{cases} 1 & \text{if } t = t_k \\ 0 & \text{if } t \neq t_k \end{cases}$$

로 주어지며, 계수 δ_k는 시점 t_k에서의 외부충격을 나타낸다. 교란 항 (disturbance term) $v_T(t)$와 $v_R(t)$는

$$v_T(t) \sim NID\,(0\,,\,\sigma_T^2\,),$$

$$v_R(t) \sim NID\,(0\,,\,\sigma_R^2\,),$$

$$E[\,v_T'(t)\,v_T^*(t)\,] = 0$$

로 가정된다. 분산이 클수록 추이성은 크게 나타나며, 특히 $v_T(t)$의 효과는 상하의 폭을, $v_R(t)$는 기울기의 변화를 나타낸다.

계절요인 항 $S(t)$는 12달의 도수와 관계가 있는 6개의 삼각함수들의 합으로 주어진다.

$$S(t) = \sum_{j=1}^{6} S_j(t).$$

각 항들은

$$S_j(t) = cos(w_j)S_j(t-1) + sin(w_j)S_j^*(t-1) + v_{S_j}(t) \ ,$$

$$S_j^*(t) = -sin(w_j)S_j(t-1) + cos(w_j)S_j^*(t-1) + v_{S_j}^*(t)$$

와 같은 확률변수들의 쌍으로 표현되며, w_j는 $w_j = 2\pi j/12$로 주어진다. 여기에서 v_{S_j}와 $v_{S_j}^*$는 서로 상관관계가 없는 백색 잡음 확률과정이며 평균 0, 공통분산 σ_s^2을 갖는 확률과정으로 가정된다.

노이즈 항 $\eta(t)$는 위에서 언급된 항들로는 설명될 수 없는 불규칙적인 변동과 정상적인 과정에서 벗어난 일시적인 대규모 변동(또는 이상치)의 합으로 주어진다.

$$\eta(t) = I(t) + O(t)$$

여기에서 $I(t)$는 보통 불규칙변동 항(irregular component)으로 불리우며, 보통 시그널과 표본오차 항들을 추정한 후에도 여전히 시그널 항에서 불규칙적인 변동이 탐지될 때 고려되는 항이다. 불규칙변동 항 $I(t)$는

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

로 주어진다. 만약, $I(t)$ 항의 분산이 0일 경우에는 $I(t)$ 항은 0으로 처리되어 모형에서 제외될 수 있다. 이상치의 항 $O(t)$는 관측된 계열에서 탐지되는 일시적인 대규모 변동을 나타내며

$$O(t) = \sum_j \lambda_j \zeta_j(t)$$

로 주어진다. 여기에서

$$\zeta_j(t) = \begin{cases} 1 & \text{if } t = j \\ 0 & otherwise \end{cases}$$

계수 λ_j는 시점j에서 계열들의 변화량을 나타낸다. $O(t)$는 주로 CPS 관측값에 영향을 주는 외부적인 요인에 의해 발생한다. 예를 들면 잘못된 표본가구들의 선정에 의해 관측된 비정상적인 CPS 추정값들은 외부적인 요인에 의해 발생된 이상치로 볼 수 있으며, 이로 인해 발생되는 일시적인 변동은 시계열에 그대로 반영되어 이상치에 대한 인식을 용이하게 해

주며 시그널 항 추정에서 이상치의 양을 줄여주는 보정과정이 이루어진다.

표본오차 $e(t)$는 모집단의 값 $\theta(t)$와 추정값 $y(t)$ 간의 차

$$e(t) = y(t) - \theta(t)$$

로 정의된다. 여기에서 표본오차 $e(t)$는 다음과 같은 특성을 갖는다.

$$E[e(t)] = 0 \ ,$$

$$Var[e(t)] = \sigma^2_{e(t)} \ ,$$

$$\rho_e(l) = \frac{E\{e(t)e(t-l)\}}{\sigma^2_{e(t)}} .$$

CPS 표본오차의 자기상관적 구조는 CPS 패널 설계와 모집단의 특성에 의해 결정되며, 표본오차의 절대 크기는 표본 재설계, 표본크기의 변화, 노동력 수준들의 변동에 의해 고정되어 있는 값이 아니라 항상 변화하는 값이다. CPS 분산은 시간에 따른 변화가 심하고, 자기상관구조는 이에 비해 매우 안정적이라는 데에 착안하여

$$e(t) = \gamma(t)\, e^*(t)$$

와 같은 표본오차의 의 형태를 고려할 수 있다. 위 표현식은 $e(t)$의 자기상관적인 구조와 이분산적인 구조(heteroscedastic structure)를 파악할 수 있게 해 준다. 분산 팽창계수 $\gamma(t)$는 CPS에서 이분산성을 설명하며

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

로 정의된다. 여기에서 $\sigma_{e(t)}$는 CPS 추정값에 대한 일반화분산함수에 의한 추정값이며, σ_{e^*}는 자기회귀이동평균(ARMA:autoregressive moving average) 확률과정 $e^*(t)$의 추정오차이다.

CPS 자기상관 구조는 ARMA 확률과정 모형

$$e^*(t) = \phi^{-1}(L)\theta(L)v_{e^*}(t)$$

를 통해 얻는다. 여기에서 L은 시차 연산자로써 $L^k(X_t) = X_{t-k}$로 정의되

며, $\phi(L)$과 $\theta(L)$은 각각

$$\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p,$$

$$\theta(L) = 1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_q L^q$$

와 같이 주어진다. 모수 $\phi_1, \phi_2, \cdots, \phi_p$와 $\theta_1, \theta_2, \cdots, \theta_q$는 자기상관함수

$$\rho_{e^*}(l) = \frac{\theta(L)\theta(L^{-1})}{\phi(L)\phi(L^{-1})}$$

를 통해 표본오차 시차 상관계수(lag correlation)들로부터 추정된다 (Dempster and Hwang, 1990). 여기에서 $\rho_{e^*}(-l) = \rho_{e^*}(l)$가 가정된다. 계수 θ와 ϕ는 충격응답 가중치(impulse response weights) $\{g_k\}$를 계산하는데 이용되며, 가중치 $\{g_k\}$는 생성함수

$$g(L) = \phi^{-1}(L)\theta(L)$$

를 통해 계산된다. ARMA 확률과정 $e^*(t)$의 분산은

$$\sigma_{e^*}^2 = \sum_{k=0}^{\infty} g_k^2$$

로 계산된다.

한편, 주 내의 모든 구 지역들과 인구 25,000이상의 시지역들 및 주에 부속되어 있으면서 노동시장에 크게 영향을 미치는 부차관심영역들에 대한 추정값들도 CPS 자료에 CES 자료, UI 자료와 센서스 자료가 추가되어 추정된다.

주 지역들에 대한 추정값들은 $X-11$ ARIMA 계절요인 보정절차에 따라 일년에 두 번 계절요인에 대한 보정이 이루어진다.

미 센서스 국은 일년에 한번 추계인구에 대한 전체적인 조정을 실시한다. 따라서 모든 주지역들과 콜롬비아 구에 대해서 CPS 노동력 관련 추정값들은 새로운 추계인구를 기준으로 벤치마킹이 이루어진다. Denton 방법(Denton, 1971)으로 불리우는 벤치마킹 절차를 적용하여 모형기반 노동력 관련 추정값들의 연간 평균을 해당되는 CPS 추정값들의 연간 평균과 일치시키는 작업이 이루어진다.

3.1.8 적용 사례

(1) 회귀 모형 추정법

회귀모형을 이용한 추정법은 인근 유사지역의 정보를 이용하여 추정된 합성 추정값이 지역 변동을 충분히 설명하지 못한 점을 보완하고자 실직보험 자료와 구직등록 자료를 이용하여 1970년대 후반부터 시군구의 실업통계 작성에 이용되었던 기법이다. CPS의 1차 추출단위(PSU:primary sampling unit)에 대한 추정치를 종속변수로 사용하고, 다음과 같은 몇 개의 적절한 독립변수를 선택한 회귀모형을 고려한다.

① 다음의 범주를 고려한 합성추정값

o 직업별-성별-인종별 범주

o 결혼여부-연령대별-성별-인종별 범주

② 센서스에서 추정한 실업자 총계 대비 3-4월의 실직보험 가입자에 대한 가입비율(%)

③ 연말 자료에서 공표되는 "70-단계"의 실업률 추정치

5개월 간의 월별 CPS 추정값들의 평균을 종속변수로 사용하기 위해 CPS PSU와 SMSA를 대응시켰을 때 150개의 SMSA 중 122개의 SMSA가 완전 대응되며 이러한 지역을 회귀모형 추정을 위한 자료로 활용하였다. 실업률 추정을 위해 적합된 회귀모형은 다음과 같다.

$$\hat{Y} = 0.008 - 0.201X_1 + 0.680X_2 + 0.404X_3 \ , \qquad (3.2)$$

$$Residual\ Mean\ Square = 0.868 \times 10^{-4},$$

추정치의 표준오차$= 0.932 \times 10^{-2}$,

$$R^2 = 0.546$$

여기에서

Y=5개월 간의 월별 CPS 실업률 추정값의 평균,

X_1=전체 실업자 대비 보험가입 실업자의 비율(%),

X_2=최종 공표된 연말 실업률(70-단계 추정값),

X_3 ="결혼여부-연령대별-성별-인종별"범주에서 계산된
실업률의 합성추정값

위 회귀모형의 타당성 검증을 위하여 센서스 자료를 이용한 회귀모형을 적합한 결과는 다음과 같이 주어지며, 실업률 추정에서 회귀모형 적용에 대한 타당성을 보여준다.

$$\hat{U} = 0.009 + 0.012X_1 + 0.586X_2 + 0.540X_3 \ , \qquad (3.3)$$

추정치의 표준오차$= 0.213 \times 10^{-4}$,

$R^2 = 0.859$

여기에서 U는 센서스 자료에서 계산된 실업률 추정값이다.

5개월 간의 월별 CPS 실업률 추정값들의 평균을 종속변수로 사용했을 때, 변동의 55%가 회귀모형으로 설명될 수 있고, 센서스의 실업률 추정값들을 종속변수로 사용한 경우에는 변동의 약 86%가 회귀모형으로 설명될 수 있다. CPS 표본 추정값들의 변동이 커질수록 설명변수들의 기여도가 낮아진다는 것을 확인할 수 있다.

다음은 독립변수들을 변화시켜 가면서 추가로 2종의 회귀모형을 추정해 본 결과이다. 먼저 독립변수 X_2 대신 벤치마킹이 이루어지기 전의 실업률에 대한 CPS 추정값 X_4를 모형에 적합시킨 결과이다.

$$\hat{Y} = 0.012 - 0.252X_1 + 0.299X_3 + 0.078X_4 \ , \qquad (3.4)$$

Residual Mean Square $= 0.833 \times 10^{-4}$,

추정값의 표준오차$= 0.912 \times 10^{-2}$,

$R^2 = 0.565$

센서스 자료에 의해 위와 대응되는 회귀모형은 다음과 같다.

$$\hat{U}' = -0.006 + 0.005X_1 + 0.477X_3 + 0.564X_4 \ , \qquad (3.5)$$

Residual Mean Square $= 0.228 \times 10^{-4}$,

추정값의 표준오차$= 0.478 \times 10^{-2}$,

$$R^2 = 0.849$$

회귀모형 (3.4)는 총변동의 약 57%를 설명하며, 회귀모형 (3.2)보다는 어느 정도 개선된 효과가 나타났다.

두 번째 경우에는 최종 공표된 연말 실업률 X_2대신 11개월 간의 CPS 월별 추정값들의 평균 X_5와 "직업별-성별-인종별" 범주의 유사정보를 추정과정에 도입하여 구한 합성추정값 X_6를 독립변수로 선택하였을 때 계산된 회귀모형의 적합 결과이다.

$$\hat{Y}'' = 0.009 - 0.210X_1 + 0.640X_5 + 0.444X_6 \ , \qquad (3.6)$$

$$Residual\ Mean\ Square = 0.883 \times 10^{-4},$$

추정값의 표준오차$= 0.939 \times 10^{-2}$,

$$R^2 = 0.539$$

센서스 자료에 의한 적합 결과는 다음과 같이 주어진다.

$$\hat{U}'' = -0.008 - 0.011X_1 + 0.532X_5 + 0.617X_6 \ , \qquad (3.7)$$

$$Residual\ Mean\ Square = 0.194 \times 10^{-4},$$

추정값의 표준오차$= 0.440 \times 10^{-2}$,

$$R^2 = 0.872$$

위의 적합결과를 살펴보면, 센서스 자료에 의한 추정값을 종속변수로 이용했을 때 회귀식은 약 85 ~ 87%의 설명력을 보이나, CPS자료 추정치를 사용했을 경우에 회귀식의 설명력은 약 54 ~ 56% 정도임을 확인할 수 있다.

다음 <표3.5>는 변수들 간의 상관관계를 나타낸 표이다. "70-단계" 추정값 X_2는 센서스 추정값 U와 CPS 추정값 Y와 높은 상관관계를 보인다. 또한, 합성추정값 X_3와 실직보험가입비율 추정값 X_1과는 낮은 상관성을 나타내며, "70-단계" 추정값 X_2와도 상관계수의 값이 낮다. 따라서 X_2이외에 추가적인 독립변수로써 합성추정값 X_3와 실직보험가입비율 추

정값 X_1을 고려한다면 회귀식의 예측력은 증가할 것임을 알 수 있다.

<표3.5> 변수들의 가중상관계수

	X_1	X_2	X_3	X_4	X_5	X_6	Y	Z	U
X_1	1.000								
X_2	0.676	1.000							
X_3	0.285	0.512	1.000						
X_4	0.682	0.961	0.574	1.000					
X_5	0.666	0.995	0.477	0.959	1.000				
X_6	0.369	0.584	0.974	0.633	0.548	1.000			
Y	0.372	0.692	0.577	0.720	0.682	0.599	1.000		
Z	0.259	0.525	0.340	0.543	0.521	0.339	0.700	1.000	
U	0.554	0.851	0.756	0.868	0.815	0.810	0.741	0.512	1.000

* Z=1970년 5월 CPS 실업률 추정값

센서스 중간 해에는 회귀모형의 종속변수는 CPS 1차 추출단위에 대한 추정값이 되며, 독립변수로 센서스에서 추정한 실업률U를 선택할 수 있으며, 적합된 회귀식은 다음과 같다.

$$\hat{Y} = 0.010 + 0.450U + 0.326X_4 + 0.089X_6 , \qquad (3.8)$$

추정값의 표준오차$= 0.914 \times 10^{-4}$,

$R^2 = 0.563$

만약 종속변수로 1970년 5월 CPS 실업률 추정값 Z를 사용한다면 적합된 회귀모형은 다음과 같다.

$$\hat{Z} = 0.019 + 0.422U + 0.430X_4 - 0.246X_6 , \qquad (3.9)$$

Residual Mean Square $= 2.040 \times 10^{-4}$,

추정값의 표준오차$= 1.428 \times 10^{-2}$,

$R^2 = 0.291$

회귀적합모형 (3.8)은 독립변수로 센서스 추정값을 이용한 결과이다. 회귀모형은 약 56%의 설명력을 가지며, 또한 Y를 종속변수로 갖는 다른 회귀적합모형과 비교했을 때 결코 떨어지지 않는 설명력을 갖는다. 1970년 5월 CPS 실업 추정치를 종속변수로 선택한 회귀적합모형 (3.9)는 약 29%의 설명력을 갖는다. 이러한 낮은 설명력은 종속변수의 큰 분산(변동)에 기인한다고 볼 수 있다.

노동력의 규모가 서로 상이한 지역들에 대한 잔차분석 결과가 <표 3.6>에 주어졌다. 독립변수는 각각 (X_1, X_3, X_4), (X_1, X_5, X_6)이고, 종속변수로써 각각 1970년 센서스 추정값, 5개월간의 월별 CPS 추정값, 1970년 5월의 CPS 추정값을 이용한 회귀모형에 대한 잔차의 평균과 잔차의 표준오차이다.

<표3.6> 잔차의 평균과 표준오차 (단위:%점)

1970센서스 노동력크기	SMSA의 수	센서스실업률		CPS 5개월 평균실업률		CPS 5월 실업률(1970년)	
		잔차 평균	잔차 표준오차	잔차 평균	잔차 표준오차	잔차 평균	잔차 표준오차
(X_1, X_3, X_4)							
1,000,000 이상	9	-0.18	0.31	-0.01	0.39	-0.00	0.54
500,000-999,999	17	-0.02	0.54	-0.06	0.66	0.24	0.88
250,000-499,999	21	0.15	0.36	0.15	1.17	-0.21	1.23
100,000-249,999	51	0.14	0.66	0.03	1.55	0.10	2.77
100,000 미만	24	0.37	0.78	0.41	1.83	-0.46	2.83
(X_1, X_5, X_6)							
1,000,000 이상	9	-0.07	0.34	0.07	0.38	0.06	0.55
500,000-999,999	17	-0.10	0.49	-0.16	0.75	0.10	0.90
250,000-499,999	21	0.06	0.39	0.06	1.23	-0.28	1.37
100,000-249,999	51	0.10	0.61	0.03	1.57	0.14	2.80
100,000 미만	24	0.11	0.71	0.19	1.75	-0.63	2.86

전반적으로 노동력의 규모가 작은 지역들보다는 노동력의 규모가 큰

지역들에서 추정값의 효율이 좋다. 잔차의 표준오차는 노동력의 규모가 작을수록 증가하는 경향을 보인다. 센서스 추정값을 종속변수로 취한 회귀모형에서 노동력의 규모가 100,000이하인 지역을 살펴보면 잔차의 표준오차는 노동력의 규모가 백만 이상인 지역보다 2배 이상 큰 표준오차를 갖는다. 5개월 간의 월별 CPS 평균에 대한 추정값을 이용한 회귀모형이 1970년 5월 CPS 추정값을 이용한 회귀모형 보다 잔차의 표준오차가 훨씬 작게 나타난다.

1970년 센서스 실업률과 회귀모형 추정값 간의 차이에 대한 분포가 다음 <표3.7>에 주어졌다. 독립변수는 (X_1, X_5, X_6)를 이용하였다. CPS 추정값을 이용한 경우보다 센서스 실업률 추정값을 이용했을 경우가 훨씬 대칭적인 분포를 보인다.

<표3.7> 센서스실업률과 회귀추정값의 차이에 대한 분포
(122개의 SMSA 이용)

차이의 범주 (%점)	센서스 실업률		5개월 간의 CPS 평균 실업률		1970년 5월의 CPS 실업률	
	SMSA의 갯수	백분율 (%)	SMSA의 갯수	백분율 (%)	SMSA의 갯수	백분율 (%)
3.00 이상	0	0	0	0	0	0
2.00-3.00	1	0.8	1	0.8	1	0.8
1.50-2.00	0	0	0	0	0	0
1.00-1.50	6	4.9	0	0	2	1.6
0.50-1.00	14	11.5	4	3.3	5	4.1
0.25-0.50	15	12.3	5	4.1	8	6.6
0.10-0.25	19	15.6	1	0.8	3	2.5
-0.10-0.10	23	18.9	9	7.4	6	4.9
-0.25～-0.10	10	8.2	11	9.0	8	6.6
-0.50～-0.25	15	12.3	15	12.3	16	13.1
-1.00～-0.50	16	13.1	47	38.5	43	35.2
-1.50～-1.00	3	2.5	25	20.5	23	18.9
-2.00～-1.50	0	0	2	1.6	5	4.1
-3.00～-2.00	0	0	2	1.6	2	1.6
-3.00 이하	0	0	0	0	0	0

-1.0에서 1.0까지의 구간을 살펴보면, 센서스 실업률을 종속변수로 취한 회귀모형에서는 약 91.9%의 SMSA지역들이 이 구간 내에 포함되며, 5개월 간의 CPS 추정값의 평균을 종속변수로 취한 회귀모형에서는 약 75.3%, 1970년 5월 CPS 추정값을 종속변수로 취한 회귀모형에서는 약 73.0%의 SMSA지역들이 포함되어 있다. -0.50 이하의 범주에서는 5개월 간의 CPS 실업률 추정값의 평균을 종속변수로 취한 회귀모형이 약 62.2%, 1970년 5월 CPS 실업률 추정값을 종속변수로 취한 회귀모형이 약 59.8%의 SMSA지역들을 포함하고 있으며, 이 범주에서는 센서스 실업률을 종속변수로 취한 회귀모형보다는 CPS 추정값을 종속변수로 취한 회귀모형이 더 많은 SMSA지역들을 포함하고 있다.

5개월 간의 월별 CPS 추정값의 평균을 종속변수로 취한 회귀모형에서

$$MSE = E\{\frac{(Y_0 - \hat{Y})^t(Y_0 - \hat{Y})}{n}\} - \frac{(n-2p-2)\sigma_w^2}{n}$$

을 이용하여 계산된 *MSE* 추정결과는

독립변수	*MSE*
X_1, X_2, X_3	0.405×10^{-4}
X_1, X_3, X_4	0.369×10^{-4}
X_1, X_5, X_6	0.419×10^{-4}

와 같이 주어진다. 여기에서 "Y_0=관측값 Y", "n=관측값의 개수", "p=독립변수의 개수", "σ_w^2=PSU 내의 오차"를 나타낸다.

상기한 자료들은 다음과 같은 내용을 시사한다. 미 노동통계국에서 연말 공표하는 실업률 추정값은 CPS의 1차 추출단위 추정값을 독립변수로 취한 회귀모형을 이용하여 개선될 수 있음을 시사한다. 추가적인 독립변수들로 70-단계 추정값 외에 *UI*자료, 센서스자료, CPS 자료를 이용한 합성추정값, 센서스 중간년도의 표본 자료 추정값 등을 이용할 수 있다. 회귀모형에 의한 소지역 실업 통계의 추정은 적절한 독립변수와 종속변수의 선택이 중요한 과제이지만, 센서스와 CPS자료를 이용한 전반적인

검토 및 실업에 관련된 사회경제적 요인에 대한 파악도 중요한 과제이다.

(2) 시계열 회귀모형

시계열 회귀모형은 현재 미국에서 표본크기가 충분하지 못한 주 단위 및 주 내의 부차관심영역들의 실업통계를 작성하는데 적용하고 있는 방법으로 표본조사 추정량을 개선하기위해 모집단의 값을 확률과정으로 생각하여 시계열 분석의 시그널 적출 기법을 적용한 것이다.

1989년 1월 노동통계국(BLS)에서는 39개 주와 콜롬비아 구의 총 40개 지역에 대해 월별 취업 및 실업자 추정을 위한 새로운 방법을 소개하였다. 이 방법은 CPS의 월별 표본자료에 시계열 모형을 적합시키는 방법이다.

모집단의 특성을 추정하기 위한 직접적인 방법은 표본 설계에 근거하여 대규모 표본조사를 시행하는 것이다. 대부분 추정값들은 대 영역에 대해서는 신뢰할 만한 수준이나 대 영역 내의 부차관심영역들에 대해서는 반드시 그러하지는 않다. 주기적인 조사의 경우 소규모의 주 지역 및 주 내의 부차관심영역들에 대해서 시계열 기법이 관심을 받고 있으며, CPS 추정값은 이러한 기법을 적용시키기에 적합한 자료이다. 매달 약 50,000 표본 가구가 조사되어 모집단의 노동력 상태를 추정하는데, 전국 수준의 추정값 또는 인구수가 많은 11개 주에 대한 월별 추정값는 비교적 신뢰할 만 하다. 인구수가 적은 나머지 40개 지역에 대해서는 월별 CPS 추정값을 그대로 사용하는 것은 바람직하지 못하다.

새로운 추정은 CPS 추정값을 시간에 따라 확률적으로 변화하는 시계열로 간주하여 시그널 성분과 노이즈 성분의 합으로 표현되는 시계열 모형 추정 방법에 근거한다. 월별 CPS 추정값은 시계열 모형에서 실업보험(UI :Unemployment Insurance)자료와 경상고용통계CES:Current Employment Statistics)자료를 결합하여 계산된다. 즉, 이전에 추정한 방법과는 달리 좀더 체계적인 방법으로써 이용 가능한 보조 자료와 과거와 현재의 표본자료를 모두 이용하여 표본크기가 작은데에서 발생하는 CPS 추정값의 추

정오차를 줄이고자 하는 것이 기본적인 생각이다.

비 관측된 모집단 값의 동적인 변화와 표본오차의 자기공분산을 나타내는 모형이 주어진다면, KF(Kalman Filter)는 참값을 추정하기 위한 방법으로 적용될 수 있으며, 다음과 같은 몇가지 유용성을 갖는다. 첫째, KF는 시그널 성분과 노이즈 성분으로 표현된 모형들에 대한 다양한 추정을 허용한다. 둘째, KF의 반복적인 추정 방법은 주 및 주 내의 부차관심 영역들에 대한 월별 노동력 관련 추정값을 산출하는 데에 있어서 매우 효율이 좋은 알고리즘을 제공한다. 셋째, KF는 동적인 모형들에 대해서 미지인 모수들을 효율적으로 추정하기 위한 매우 유용한 도구이다.

① CPS 자료를 모형화하기 위한 시계열 방법

CPS 노동력 추정값 $y(t)$는 시그널 성분 $\theta(t)$와 노이즈 성분 $e(t)$의 합인

$$y(t) = \theta(t) + e(t)$$

로 주어진다. 시그널 성분(노동력의 참값)은 시간에 따라 변화하는 $\mu_X(t)$와 오차항 $u(t)$로 표현할 수 있다.

$$\begin{aligned} \theta(t) &= \mu_X(t) + u(t) \\ &= X(t)\beta(t) + u(t). \end{aligned} \tag{3.10}$$

여기에서 $X(t)$는 $1 \times k$벡터로써 관측된 값이며, $\beta(t)$는 $k \times 1$ 확률계수 벡터이다. 회귀계수 $\beta(t)$는 1차 자기상관회귀 과정에 따라 확률적으로 변화하는 형태를 수식으로 표현하면

$$\beta(t) = T_\beta \beta(t-1) + v_\beta(t) \tag{3.11}$$

로 주어진다. 여기에서 T_β는 $k \times k$행렬로 고정모수들의 행렬을 나타내며, $v_\beta(t)$는 $k \times 1$ 백색잡음오차 벡터를 나타낸다. 식(3.10)에서 오차항 $u(t)$는 자기상관성을 고려하여 $u(t) \sim ARMA(p_u, q_u)$로 가정되며

$$u(t) = \phi_u(L)\psi_u(L)^{-1} v_u(t) \tag{3.12}$$

와 같이 표현된다.

여기에서

$v_u(t) = u(t)$에 대한 백색잡음오차,

$\psi_u(L) = 1 - \sum_{i=1}^{p_u} \psi_u(i) L^i$, ($u(t)$에 대한 자기회귀연산자)

$\phi_u(L) = 1 + \sum_{i=1}^{q_u} \phi_u(i) L^i$, ($u(t)$에 대한 MA연산자)

이고 L은 시차연산자로써 $L^i y(t) = y(t-i)$를 만족한다. 랜덤오차 $v_\beta(t)$와 $v_u(t)$는 평균이 0이고 서로 독립이라고 가정한다. 즉,

$$\begin{pmatrix} v_\beta(t) \\ v_u(t) \end{pmatrix} \sim ID\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} Q & 0 \\ 0 & \sigma_{v_u v_u} \end{pmatrix} \right) \tag{3.13}$$

단, $Q = Cov(v_\beta(t)) = Diag(\sigma_{\beta_1\beta_1}, \cdots, \sigma_{\beta_k\beta_k})$.

노이즈 성분은 전체 모집단에서 표본 일부를 추출하는 과정에서 발생하는 표본오차로 취급된다. CPS는 모집단으로부터 추출된 다 단계 확률표본이다. 일차적으로 모집단으로부터 일차추출단위(PSU)들이 뽑히고, 다음으로 PSU 내에서 조사가구 단위들이 추출된다.

$y(t)$가 노동력 특성을 나타내는 $\theta(t)$에 대한 CPS 추정치로 주어지면, 표본오차 $e(t) = y(t) - \theta(t)$의 분산과 공분산 함수는 다음과 같다.

$$\sigma_{e(t)e(t)} = D_y S_y^2 \tag{3.14}$$

$$\gamma_{ts} = Cov(e(t), e(s))$$

여기에서

D_y = 단순임의추출표본 추정치의 분산에 대한 CPS 추정치의 분산의 비(Design Effect),

$S_y^2 = \frac{N(t)}{n(t)} \theta(t)(1 - P(t))$,

$N(t)$ = 모집단의 크기,

$n(t)$ = 표본의 크기,

$$P(t) = \frac{\theta(t)}{N(t)}.$$

위의 식에서 설명된 것과 같이 CPS 표본오차는 이분산적 구조와 자기상관적 구조를 동시에 갖는다. 식(3.14)는 표본 재설계에서 설계효과 D_y, 표본구간 $N(t)/n(t)$, 참값 $\theta(t)$와 $P(t)$의 변화와 같은 이분산성의 세가지 주요한 요인이 반영된 식이다. 부연하여 설명하면, CPS는 10년 간의 센서스 자료를 사용하여 매 십년 마다 표본 추출 구조 및 추정 절차를 갱신하기 위하여 재 설계된다. 표본조사의 재 설계보다는 오히려 주 단위에 대한 표본 크기의 조정이 자주 있었고, 이것은 주 수준의 분산에 중요한 영향을 끼치게 되었다. 일종의 고정된 설계와 고정된 표본 크기일지라도 오차 분산은 참 노동력 크기의 함수이므로 변하게 된다. 노동력은 매우 순환적이면서 동시에 계절적이므로 분산도 이와 비슷한 형태를 가진다는 것을 예측해야만 한다.

$e(t)$의 자기공분산 구조는 다음의 세 가지 사실에 기인한다. 첫째, 월별 표본은 8개의 서로 독립적인 연동교체그룹들인 부 표본들로 구성되어 있다. 연동교체그룹들은 4개월 간 연속 조사되고, 8개월 동안 조사되지 않다가 다시 4개월 간 연속 조사 후 표본에서 영구히 제거된다. 이 과정에서 동일한 조사가구 단위들이 시간에 따라 중복되어 조사되므로 당연히 상관성이 존재한다. 4-8-4 연동교체표본 체계는 월별 추정값의 신뢰성을 높이기 위해 15개월 주기 동안 첫 달은 전월에 조사된 표본의 약 75%를 중복하여 월 표본으로 사용하고, 나머지는 교체표본 추출방식에 의해 새로운 표본을 조사한다. 2개월 간의 중복율은 약 50%정도이다. 둘째, 연동교체표본 체계의 사용은 표본들에 대한 주기적 선택을 요구한다. 한 집락의 조사가구 단위들이 연동교체표본 그룹으로부터 영구히 제외될 때, 근처에 있는 단위들로 대체된다. 따라서 새로운 단위들도 대체 단위들과 비슷한 특성을 가질 것이므로 같은 연동교체표본 그룹에서 다른 가구들과도 높은 상관성을 갖게 된다. 셋째, 표본오차는 복합추정량에 의해 영향을 받게 된다.

CPS 추정량의 이분산적이며 자기상관적인 구조는 $e(t)$를

$$e(t) = \gamma(t)\, e^*(t)$$

와 같은 승법적 구조로 모형화 함으로써 설명될 수 있다(Bell and Hillmer(1989)). 여기에서 $e^*(t)$는 $ARMA$과정을 따르며 상수인 분산을 갖는다.

$$e^*(t) = \phi_e(L)\psi_e(L)^{-1} v_e(t) \tag{3.15}$$

여기에서

$$v_e(t) \sim NID(0, \sigma_{v_e}),$$

$$\sigma_{e^*(t)e^*(t)} = \sigma_{v_e} \sum_{k=0}^{\infty} g_k^2$$

이며, 가중치 $\{g_k\}$는 생성함수 $g(L) = \phi_e(L)\psi_e(L)^{-1}$로 계산된다. $e(t)$의 이분산성 성분 $\gamma(t) = \sqrt{\frac{\sigma_{e(t)e(t)}}{\sigma_{e^*(t)e^*(t)}}}$ 는 분산비의 제곱근이다. $e(t)$의 자기상관구조는 표본 설계에 의해 영향을 받게 된다. 예를 들면, 표본 재설계에서 복합추정량에서의 가중치들이 모두 교체되는 것과 같은 경우이다.

② 상태공간형식(State Space Form)과 KF 알고리즘

노동력 모형의 시그널 성분과 노이즈 성분을 상태공간형식에 삽입한다. 먼저 추정에는 적절치 못할지라도 융통성을 보여 주기에는 유용할 것 같은 매우 일반적인 형식을 설명하고 실질적인 적용에 있어서는 고려되어야 할 것 같은 제한사항들에 대해서 설명한다. 상태공간 작성에서는 비관측된 시그널과 노이즈가 상태 변수들이고, 이 들의 시간에 따른 전개는 변이함수들의 집합에 의해 설명된다. 관측점 함수는 상태 변수들을 관측 표본시계열들로 변환시킨다. 상태공간 체계에서 변이함수들은 1계의 VAR 형태를 취한다.

비 관측변수는 $\beta(t)$, $u(t)$, $e^*(t)$이다. 계수벡터인 $\beta(t)$는 (3.11)식과

같이 이미 적절한 형식이 있다. $u(t)$와 $e^*(t)$는 (3.12)식과 (3.15)식에서와 같이 $ARMA$과정으로 명시되나, 각각 1계 VAR형태인 $S_u(t)$와 $S_e(t)$벡터들로 변환된다. 임의의 $ARMA(p, q)$과정은 일종의 $r\times 1$ 1계 VAR형태로 변환될 수 있다는 것이 기본적인 규칙이며, 여기에서 $r = max(p, q+1)$이다(Harvey(1981)). 변이함수들은 다음과 같이 주어지며, 여기에서 $S(t)$는 $\beta(t)$, $S_u(t)$와 $S_e(t)$로 구성되는 상태벡터이다.

$$S(t) = T(t)S(t-1)+\Gamma(t)v(t),$$

$$(m\times 1)(m\times m)\ (m\times l)(l\times 1)$$

$$\begin{pmatrix}\beta(t)\\ S_u(t)\\ S_e(t)\end{pmatrix}=\begin{pmatrix} T_\beta & 0 & 0\\ 0 & T_u & 0\\ 0 & 0 & T_e(t)\end{pmatrix}\begin{pmatrix}\beta(t-1)\\ S_u(t-1)\\ S_e(t-1)\end{pmatrix}+\begin{pmatrix} I_k & 0 & 0\\ 0 & \Gamma_u & 0\\ 0 & 0 & \Gamma_e(t)\end{pmatrix}\begin{pmatrix}v_\beta(t)\\ v_u(t)\\ v_e(t)\end{pmatrix},$$

$$E(v(t)v(t)^t) = block\ diagonal(Q,\ \sigma_{v_u v_u},\ \sigma_{v_{u(t)}v_{e(t)}}).$$

여기에서

$$T_u=\begin{pmatrix}\psi_u(1) & : & I_{r_u-1}\\ \psi_u(2) & :\\ \cdot & :\\ \cdot & :\\ \cdot & :\\ \cdots\cdots & \cdots & \cdots\\ \psi_u(r_u) & : & 0\end{pmatrix},\ T_e(t)=\begin{pmatrix}\psi_e(1) & : & I_{r_e-1}\\ \psi_e(2) & :\\ \cdot & :\\ \cdot & :\\ \cdot & :\\ \cdots\cdots & \cdots & \cdots\\ \psi_e(r_e) & : & 0\end{pmatrix}$$

$$\Gamma_u=\begin{pmatrix}1\\ \phi_u(1)\\ \phi_u(2)\\ \cdot\\ \cdot\\ \cdot\\ \phi_u(r_u-1)\end{pmatrix},\quad \Gamma_e=\begin{pmatrix}1\\ \phi_e(1)\\ \phi_e(2)\\ \cdot\\ \cdot\\ \cdot\\ \phi_e(r_e-1)\end{pmatrix},$$

$m = k + r_u + r_e$,

k = 회귀변수의 수,

$l = k + 2$,

$r_u = max(p_u, q_u+1)$, p_u와 q_u는 $u(t)$의 $ARMA$모수,

$$r_e = max(p_e,\ q_e+1),\quad p_e와\ q_e는\ e(t)의\ ARMA모수.$$

관측점 함수는 벡터 $H(t)$를 선택하여, 시그널 성분과 노이즈 성분을 만들기 위한 상태변수들의 일차결합을 취한다.

$$Y(t) = H(t)S(t) = \theta(t) + e(t), \tag{3.16}$$

여기에서

$$H(t) = (X(t) \mid 1 \mid 0_{r_\theta - 1} \mid \gamma(t) \mid 0_{m-k-r_e-2}),$$
$$\theta(t) = H_\theta(t)S(t),$$
$$e(t) = H_e(t)S(t),$$
$$H_\theta(t) = (X(t) \mid 1 \mid 0_{m-k-1}),$$
$$H_e(t) = (0_{k+r_\theta} \mid \gamma(t) \mid 0_{r_e-1}).$$

비 관측된 시그널과 노이즈 성분들의 상태공간형태가 주어진다면, KF는 시그널과 노이즈를 추정하기 위한 방법을 제공한다. 이러한 알고리즘을 설명하기 위해 시간 $t-j$까지 관측된 자료에 대한 $S(t)$의 조건부 평균 및 공분산 행렬을 다음과 같이 표현하자.

$$S(t \mid t-j) = E(S(t) \mid Y_{t-j}, \cdots, Y_1),$$
$$P(t \mid t-j) = E\{(S(t) - S(t \mid t-j))(S(t) - S(t \mid t-j))' \mid Y_{t-j}, \cdots, Y_1)\}.$$

또한 바로 전까지의 값들이 주어진 경우의 표본 추정값 $Y(t)$의 예측값을

$$Y(t \mid t-1) = H(t)S(t \mid t-1),$$

$Y(t)$의 분산을

$$E(Y(t) - Y(t \mid t-1))^2 = H(t)P(t \mid t-1)H(t)'$$
$$= f(t \mid t-1)$$

로 표현하자.

t번째 관측점까지의(t번째 관측점은 제외) 자료에 근거한 $S(t)$의 추정값이 주어진다면, 최근의 자료를 고려한 $S(t)$의 추정량은 $S(t \mid t-1)$과 최근의 표본 추정값 $Y(t)$의 가중평균으로

$$S(t \mid t) = (I - K(t)H(t))S(t \mid t-1) + K(t)Y(t)$$

$$= S(t|\,t-1) + K(t)(Y(t) - Y(t|\,t-1))$$

와 같이 표현되며, 공분산 행렬은

$$P(t|\,t) = (I - K(t)H(t)^t)P(t|\,t-1)$$

로 주어지며,

$$S(t|\,t-1) = TS(t-1|\,t-1),$$

$$P(t|\,t-1) = TP(t-1|\,t-1)T^t + \Gamma E(v(t)v(t)^t)\Gamma^t$$

로부터 반복적으로 추정된다. 여기에서 가중벡터 $K(t)$(gain of filter)는

$$K(t) = \frac{P(t|\,t-1)H(t)^t}{f(t|\,t-1)}$$

로 표현되며, $K(t)$의 원소들은 $P(t|\,t)$의 대각원소들의 합을 최소화하여 결정한다(Gelb(1974)).

KF 방정식을 이용하여 시간 t에서 관측된 표본 추정값은 시그널 성분과 노이즈 성분으로 분해된다.

$$Y(t) = \theta(t|\,t) + e(t|\,t).$$

여기에서

$$\theta(t|\,t) = \theta(t|\,t-1) + h_\theta(t)\tilde{Y}(t),$$

$$e(t|\,t) = e(t|\,t-1) + (1 - h_\theta(t))\tilde{Y}(t),$$

$$\tilde{Y}(t) = Y(t) - Y(t|\,t-1),$$

$$h_\theta(t) = H_\theta(t)K(t)$$

$$= \frac{Var\left(\frac{\theta(t)}{\theta(t-1)}\right) + H_\theta(t)TP(t-1|\,t-1)T^tH(t)}{f(t|\,t-1)},$$

$$1 - h_\theta(t) = H_e(t)K(t)$$

$$= \frac{\sigma_{e(t)e(t)} + H_e(t)TP(t-1|\,t-1)T^tH(t)}{f(t|\,t-1)},$$

$$Var\left(\frac{\theta(t)}{\theta(t|\,t-1)}\right) = \sum_{i=1}^{k} X_i(t)^2\sigma_{\beta_i\beta_i} + \sigma_{v_tv_t},$$

$$\sigma_{e(t)e(t)} = \gamma(t)^2 \sigma_{v_e v_e} .$$

가중치 $h_\theta(t)$는 예측오차 $\tilde{Y}(t)$를 시그널과 노이즈 성분으로 분해하며, 이것은 KF가 시그널 성분의 최소평균제곱오차 추정값을 만들기 위해 시계열 추정값 $\theta(t|t-1)$과 최근의 표본 추정값 $Y(t)$를 결합하는 구체적인 방법을 설명한다. $\theta(t|t-1)$의 $Y(t)$에 대한 보정 양은 이분산성의 표본오차 분산 $\sigma_{e(t)e(t)}$와 상대적으로 비교하여 시계열 분산성분 $Var\left(\frac{\theta(t)}{\theta(t-1)}\right)$의 크기의 함수로 표현된다. $\sigma_{e(t)e(t)}$의 값이 크면 $h_\theta(t)$의 값이 작게 되며, 따라서 $\theta(t|t)$를 이끌어 내는데 있어서 시계열 예측값 $\theta(t|t-1)$에 대한 작은 양의 보정만이 일어난다. 역으로 만약 표본 분산이 작다면 $\theta(t|t)$는 최근의 표본 추정값 $Y(t)$와는 아주 다른 값이 될 것이다.

KF는 반복적인 방법으로 상태벡터 $S(t)$의 최소평균제곱오차를 제공하며, 또한 KF는 새로운 자료가 각 주기에서 이용될 수 있는 실시간 상황에 적합하다. 그러나 자료가 시간 t 이후에 이용될 수 있을 경우에는 추정값 $S(t|t)$는 이러한 새로운 정보를 반영시키지는 못한다. 왜냐하면 KF는 시간 t의 앞쪽으로만 이동하기 때문이다. 이 전 주기의 추정값들에 대한 부분적인 최적화는 평활을 통해 쉽게 조정될 수 있다.

평활의 방법은 두 가지 형태의 KF 추정량을 결합하는 방법이다. 첫 번째 형태의 KF 추정량은 이 전에 설명했던 것과 같이 전방 Filter 추정값으로써 시간 t에서 모든 과거 표본자료와 현재 표본자료에 의해 추정되는 $S(t|t)$이다. 두 번째 형태의 KF 추정량은 후방 Filter 추정치로써, 표본 주기의 끝에서 출발하여(가령 $t=n$에서 출발) 처음까지 진행하며, 미래의 자료에만 근거하여, 각 시간 t에서의 예측값들을 이끌어 내는데, 이러한 추정량을 $S(t|t+1)$로 나타낸다. 이 때 최적인 평활 추정량은 두

추정량의 평균제곱오차 값들의 비율을 결합하여 작성되며 구체적인 형태는 다음과 같이 주어진다.

$$S(t|n) = P(t|n)\{\frac{S(t|t)}{P(t|t)} + \frac{S(t|t+1)}{P(t|t+1)}\}.$$

여기에서

$$P(t|n) = 1/\{\frac{1}{P(t|t)} + \frac{1}{P(t|t+1)}\}.$$

한편, $S(t|n)$의 공분산 표현식으로부터

$$\frac{1}{P(t|n)} = \frac{1}{P(t|t)} + \frac{1}{P(t|t+1)}$$

을 얻을 수 있고 $P(t|n) - P(t|t)$는 음 반정치(negative semidefinite)임을 알 수 있다. 따라서 $S(t)$의 평활 추정량은 전방 Filter추정량보다는 훨씬 좋은 추정량이 되며, 이러한 이유 때문에 노동력 추정값들은 평활 알고리즘을 이용하여 만들어진다.

③ 보충 설명

상태공간형식은 시그널 성분을 세분화하는데 상당한 유연성을 허용하며, 특별한 경우들로써 표본조사 방법에 근거한 다음의 두 가지 형태의 모형을 포함한다.

만약 $Q = Cov(v_\beta(t)) = 0$이고 $e(t)$와 $u(t)$가 백색잡음오차 변수이면, 시스템은 Ericksen(1974)의 표본회귀모형이 된다. 이 경우 시그널 적출 문제는 관측된 표본자료에 대해 가중 최소제곱방정식을 적합시켜 해결된다. $\beta(t) = 0$, $Q = 0$일 경우 Wiener-Kolmogorov의 시그널 적출 이론에 근거한 모형을 얻는다. 회귀식의 평균은 없어지고 시그널은 공분산 정상과정이 된다. 추가적으로 $e(t)$과정의 분산과 $ARMA$ 모수들이 상수인 경우에는 $e(t)$가 공분산 정상과정이 된다.

Scott와 Smith(1974)는 공분산이 추정되어야 하는 조사 자료에 대해 전통적인 시그널 적출 방법을 채택하였다. Bell과 Hillmer(1987a)는 시그

널 과정에서 비 정상성을 다루는 방법들을 토의하였다. 앞서 소개되었던 내용에서는 시그널의 비 정상성 문제를 회귀변수들과 회귀계수들의 확률적인 변화로 다루었다. 회귀계수들의 움직임을 통제하는 (3.11)식의 추이 방정식은 다양한 형태를 수용할 수 있다(Los(1985)).

CPS 자료를 이용한 연구는 많이 진행되고 있지는 않다. Hausman과 Watson(1985)은 CPS 4-8-4 연동표본교체와 합성추정절차를 통합시킨 전 지역의 십대의 실업률 시계열에 대한 오차과정의 $ARMA(1,15)$모형을 개발했다. Bell과 Hillmer(1987b)는 Train(1978)등에 의해 추정된 design에 근거한 자기공분산의 근사로써 $ARMA(1,1)$모형을 개발했다.

CPS와 같은 복잡한 조사에 대해서 공분산 추정값을 산출하는 것은 대규모의 자료가 사용되어 비용이 많이 든다. 최근에는 필요로 하는 모든 자료를 이용하지 않고 추정값을 산출하는 방법이 모색되고 있다. 설계 기반 표본오차 공분산 계산의 어려움 때문에 방정식에서의 오차들의 효과와 표본오차의 효과를 추정하지 않고 회귀방정식을 적합시키기도 한다. 만약 두 성분 오차들이 $ARMA$과정이면, 이 때 합도 $ARMA$과정이 된다 (Granger and Morris(1975)). 즉,

$$u(t) \sim ARMA(p_u, q_u),\ e(t) \sim ARMA(p_e, q_e)\text{이면}$$

$$w(t) = u(t) + e(t) \sim ARMA(p, q).$$

여기에서 $p \leq p_u + p_e,\ q \leq max(p_u + q_e, p_u + q_e)$를 만족한다. KF는 회귀성분과 총 오차를 적출하기 위해 사용될 수 있다.

④ 실업률 추정에 적용

CPS 자료 외의 미 연방-주의 실업보험 체계로부터 생산된 실업보험(UI) 자료와 비농인구에 대한 경상고용통계조사(CES) 자료가 있다. 이러한 자료들은 1960년대 초 이래로 주의 특정 추정치를 생산하기 위하여 미 노동통계국에서 작성한 소책자 추정방법에서 이용되었던 자료들이다. 이러한 자료들을 이용하여 순환적이며 계절적인 노동력의 움직임을 통제하

고 표본오차의 영향을 줄이기 위한 시계열 추정방법이 적용되었다. 이러한 방법은 구성된 변수들의 오차 문제를 다루기 때문에, 즉 계수들보다는 오히려 종속변수의 참값을 추정하는 데에 초점이 맞추어져 있기 때문에 기존의 방법과는 다르다.

보험가입 실직자 수에 대한 월별 주(state) 단위 자료는 CPS 자료와는 독립적으로 얻을 수 있는 실직자에 대한 최신정보이다. 이러한 자료가 실업률 모형 개발을 위한 출발점이 된다. *UI* 자료는 *UI* 혜택을 신청하고 있는 노동자들의 수를 완전히 집계한다. *UI* 보상은 해고되어 주의 특정한 재정상의 적격 기준을 충족시킨 노동자들에게만 혜택이 돌아간다. 대조적으로 CPS에서 이용되는 실직의 개념은 조사기간 동안 직업을 갖고 있지 않은 모든 사람들, 해고되어 직업을 찾고 있는 사람 또는 일시 해고되어 대기발령 중인 사람들을 모두 포함한다. *UI* 자료에 집계되지 않은 실직자 범주를 살펴보면 <표3.8>와 같다.

<표3.8> *UI*자료에 집계되지 않은 실직자 범주

1. 다음의 범주에 해당되는 실직자
   a. Exhaustees : 수혜 권리를 다 써버린 노동자들
   b. 재정적 부적격자 : 주의 적격 요구기준을 충족시키지 못하는 우선 고용자 또는 소득을 갖고 있는 노동자들
   c. 유예된 비 신청자 : 실직시기의 처음에 혜택을 신청하지 않은 적격한 노동자들
2. 노동력 신규 유입자 : 최근의 실직기간 전에 노동인구에 편입되지 않았던 노동자들
3. 이직자 : 이 전의 직업을 그만두고 다른 직업을 찾고 있는 노동자

UI 자료에 집계되지 않은 실업자 중 신규 유입자가 차지하는 비율이 가장 크며, 다니는 직장을 그만두고 다른 직업을 찾는 이직 희망자들은 적어도 그 기간만큼은 보험 상의 혜택을 받지는 못한다. 만약 *UI* 보상을 받지 못하는 실직자의 상대적인 규모가 시간에 따라 안정적이라면, 보험

지불요구율은 전체 실업률 추정에 대용으로 사용될 수 있다. 실제로 노동시장에서 특히 실직자와 신규 유입자 간의 실업률 분포는 순환적이며 계절적인 특성변화를 나타낸다.

먼저 실업률 분포에 있어서 계절적 변화를 고려해 보자. 가장 중요한 현상은 실직자와 유입자는 매우 다른 계절적 형태를 띤다는 점이다. 청년과 여성이 유입자의 대부분을 차지하고, 청년층의 실업은 휘발성의 계절적 형태를 띠는데, 이는 학기가 끝나면서 신규 노동력이 유입되고 학기 시작과 함께 빠져나가는 청년층의 노동력의 변화가 반영되었기 때문이다. 이와는 대조적으로 성인 남성의 실직에 있어서 가장 일반적인 원인은 자동차 산업과 건설 산업과 같이 산업의 연간 생산 순환주기에 영향을 받아 계절적인 해고와 재 고용이 발생한다. <표3.9>는 CPS 신규유입 및 실직률의 계절적 형태(40개 주에 대한 평균값)와 전체 비율에 대한 이러한 항들의 순수효과 간의 차이를 나타내며, 또한 *UI* 가입자에 대한 계절적인 형태를 보여준다.

<표3.9> 40개 주에 대한 실업의 계절적인 요인들에 대한 평균 (1979-1985년)

월	CPS 실업률				*UI* 지불요구율
	전체	신규유입	실직	이직	
1	110.6	99.1	120.9	104.3	131.2
2	110.9	98.1	126.1	100.7	133.8
3	105.3	96.2	115.7	95.5	120.8
4	97.0	90.7	103.5	91.3	104.0
5	93.2	96.4	91.0	92.8	91.0
6	106.0	127.2	89.6	93.6	85.8
7	98.7	106.0	90.8	101.0	95.9
8	98.5	102.3	92.1	110.7	90.2
9	95.5	102.7	84.7	112.8	77.6
10	93.6	98.3	88.0	107.0	79.3
11	94.9	94.5	93.5	101.3	87.7
12	95.9	88.4	104.1	89.5	102.4

* 계절 요인들은 *ARIMA X−11*로 계산됨. CPS 실업률의 분모는 각각의 범주에 대해서 CPS 고용과 실직인원의 합이고, *UI* 지불요구율에서 분모는 CES 고용 총계임.

여기에서 실업의 계절적인 요인이 100보다 큰 값은 평균실업률보다 큰 달을 말한다. 신규 유입률은 겨울에는 평균실업률보다 낮고 여름에는 평균실업률보다 높다. 반면, 실직률은 반대의 형태를 보인다. 각 그룹에서의 큰 계절 실업률은 전체 실업률에 강한 영향을 미친다. 실직자와 신규 유입자는 경기순환과 다른 형태를 띤다는 점에서 수치적으로 중요하며, 전체 실업률의 약 반 정도를 설명한다. 이직자는 전체의 약 15%를 설명하므로 수치적으로 덜 중요하다. 늦여름과 가을에 구별되는 계절적 형태를 나타낸다. *UI* 지불요구율은 실직자의 계절적 형태를 따르나 신규 유입 또는 이직자의 계절적 형태를 따르지는 않는다.

다음의 <표3.10>은 경기순환과 관련된 실업률 분포의 변화를 나타낸다. 경기후퇴 기간 중에는 노동력 요구는 떨어지고 실직자는 증가한다. 따라서 *UI* 지불요구율은 충분히 순환적인 지표가 된다. 그러나 *UI* 지불요구는 실직자의 순환적인 변화를 전적으로 반영하지는 못한다. 경기후퇴의 후반부 쪽에서는 실직기간이 길어져서 *UI* 보상혜택을 모두 써버린 노동자들이 증가한다. 또한 일단 재 고용되면 이러한 노동자들이 차후의 실직기간 동안 *UI* 보상 자격을 얻기 위한 고용신뢰도와 충분한 임금을 얻기까지는 얼마간의 시간이 걸린다. 중요한 사실은 실직자들에 대한 *UI* 범위를 살펴보면 오랜 기간에 걸쳐 지속적으로 줄어든 적이 있었다는 점이며, <표3.10>에서 이러한 사실을 확인할 수 있다.

Burtless et al.(1984)와 Corson et al.(1988)의 연구에 의하면 이러한 현상은 경제적인 변화나 인구 통계적인 변화와는 무관하며 *UI* 대상자들이 혜택을 적용 받지 못한 이유 때문으로 설명한다. 정책의 변화가 주요한 원인으로 주목된다.

이상의 토의는 대표적인 주의 움직임을 설명한 반면, 모형화 과정에서 설명되어야만 하는 중요한 주 간의 차이점들이 존재한다. *UI* 자료에서 보면 주의 적격요구 기준, 수혜기간, *UI* 적용범위에 대한 행정관례에서의 변동사항들이다. 순환적이며 계절적인 움직임에 있어서 실제적인 차이들은 회귀계수와 모형의 모수에 영향을 끼친다.

<표3.10> 실직범주들의 상대적인 크기(40개 주 평균)

년도	CPS 실업(%)					이직
	U.S. 실업률	CPS 실직자 *UI*지불요구	지불요구	실직	신규유입	
76	7.7	93.1	36.9	43.1	41.9	14.8
77	7.1	86.5	34.5	41.3	43.6	14.9
78	6.1	87.6	31.4	37.7	46.1	16.2
79	5.8	87.5	32.8	39.4	44.4	16.1
80	7.1	76.7	35.8	47.6	38.3	14.0
81	7.6	68.1	31.9	48.8	38.4	12.7
82	9.7	61.8	34.3	56.3	34.5	9.1
83	9.6	50.7	27.7	55.1	35.5	9.2
84	7.5	52.2	25.7	50.1	38.9	10.9
85	7.2	56.6	26.9	48.3	40.1	11.5
86	7.0	60.3	27.9	47.7	38.9	13.3
87	6.2	57.0	25.3	45.6	40.2	14.0
88	5.5	62.3	26.4	43.9	40.4	15.5

UI 적용에 있어서 변화량은 <표3.11>에 주어졌다. <표3.11>은 주별 전체 CPS실업자에 대한 *UI* 지불요구율의 연간 평균값들의 전체평균(%), 주 내에서의 *CV*값, 연간 평균값들 중 최소값과 최대값을 나타냈다.

비율모형에서 회귀성분의 일반적인 형태는 다음과 같이 주어진다.

$$\text{실업률} = \beta_0(t) + \beta_1(t)(\text{지불요구율}) + \beta_2(t)(\text{인구 대 취업자비}) + \beta_3(t)(\text{신규유입률}) \tag{3.17}$$

여기에서

$$\text{지불요구율} = \frac{continued\ claims\ w/o\ earnings}{CES\ employment} \times 100,$$

$$\text{인구 대 취업자비} = \frac{CPS\ employment}{CPS\ population\ (16+)} \times 100,$$

$$\text{신규유입률} = \frac{CPS\ \text{취업 신규유입자}}{CPS\ \text{신규유입자} + CPS\ employment} \times 100.$$

<표3.11> 주별 전체 CPS실업자에 대한 *UI* 지불요구자(1976-1987)

주	*UI*지불 요구	*CV*	최소	최대	주	*UI*지불 요구	*CV*	최소	최대
AL	25.5	23.7	17.5	37.5	MT	32.2	21.6	23.1	44.2
AK	57.1	21.2	43.1	83.0	NE	30.0	18.7	20.7	43.9
AZ	24.3	13.8	19.8	29.5	NV	34.7	19.7	25.3	48.2
AR	28.1	21.6	20.2	38.5	NH	26.4	17.4	19.2	34.7
CO	23.6	10.9	20.2	28.1	NM	24.2	10.5	20.4	28.1
CT	37.5	17.0	29.3	48.9	ND	31.1	14.7	24.3	36.7
DE	29.9	17.1	22.8	38.0	OK	27.5	21.7	19.4	39.0
DC	26.6	10.7	21.9	30.3	OR	37.6	16.7	28.8	48.4
GA	25.8	15.0	21.2	33.4	RI	51.2	12.7	39.5	61.4
HI	32.9	9.5	29.0	39.8	SC	26.3	18.1	19.8	33.9
ID	31.6	15.4	24.7	39.5	SD	22.5	30.6	14.7	35.0
IN	24.4	17.5	19.4	33.5	TN	27.2	26.2	18.1	41.0
LA	30.3	20.9	22.6	42.1	UT	31.4	25.7	19.9	41.1
KS	35.5	13.6	29.5	45.9	VT	40.0	9.7	33.9	46.3
KY	29.2	29.7	17.2	40.0	VA	17.6	20.1	13.3	24.7
LA	27.9	20.6	20.0	37.6	WA	36.1	15.7	30.1	49.1
ME	35.8	11.2	28.7	41.1	WV	32.8	24.7	21.1	42.7
MD	28.5	12.2	23.6	34.0	WI	34.2	24.1	24.8	47.1
MN	33.7	18.7	24.0	44.4	WY	28.7	28.7	20.7	49.7
MS	26.1	19.7	20.1	35.3	전체	31.0	23.2	17.6	57.1
MO	32.8	19.5	24.3	44.2					

지불요구율은 *UI* 혜택을 받고 있는 실직자들의 상대적인 규모를 나타내는 측도이다. 인구 대 취업자비는 실직보험 지불 요구인원에 포함되지 않은 실직자들을 나타내는 측도로 이용된다. 이직자의 수에 영향을 미치는 노동요구에 있어서의 계절적 변동을 인구 대 취업자비에서 찾을 수 있다. 즉 여름에 정점을 취하고 가을에 떨어지는 계절적 변동을 나타낸다. 이러한 주기 동안 수많은 계절적 직업들이 발생하고, 이직에 기인한 실직이 증가한다.

대부분의 주에서는 인구 대 취업자비에서 분모의 값을 CPS 취업자료를 이용하여 계산한다. 주의 CPS 취업자료는 표본오차의 영향을 받을 가능성이 있지만, *CV*값은 주의 CPS 추정값보다는 5~6배 정도 작다. 소수의 주에서 인구 대 취업자비의 계산에서 CES 자료를 이용한다. 대부분의

주들에서는 CES 자료는 CPS 취업측도와는 어느 정도 다른 계절적 형태를 갖고 있으며 실업률의 참값과는 상관성이 높지 않은 것으로 나타난다. 표본오차의 효과를 줄이기 위하여 CPS 신규 유입률 변수는 주보다는 훨씬 큰 지역(전국 또는 센서스 지역)에 대해서 계산된다. 몇몇 경우에 있어서는 주의 CPS 신규 유입률의 3-month 이동평균이 사용된다.

40개 주의 각각에 대한 모형들이 1976-1987년 동안의 월별 CPS 실업률의 시계열 자료에 적합되었다. 이 전에 언급한 것과 같이 CPS 자료는 다음과 같이 시그널 성분과 노이즈 성분으로 표현된다.

$$Y(t) = \theta(t) + e(t),$$

여기에서

$$\theta(t) = X(t)\beta(t) + u(t)$$

로 주어지고 계수들은

$$\beta(t) = \beta(t-1) + v_\beta(t)$$

와 같은 랜덤워크 과정을 따른다. $u(t)$와 $e(t)$의 효과는 분리하여 추정하지 않기 때문에 관측된 값들은

$$Y(t) = X(t)\beta(t) + w(t),$$
$$w(t) = u(t) + e(t) \sim ARMA(p, q)$$

와 같이 표현할 수 있다.

$S_w(t)$를 $max(p, q+1)$과 같은 계(order)를 갖는 $w(t)$의 상태벡터라 하면, 상태공간 모형은 다음과 같은 전이방정식(transition equation)을 갖는다.

$$S(t) = \begin{pmatrix} \beta(t) \\ S_w(t) \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & T_w \end{pmatrix} \begin{pmatrix} \beta(t-1) \\ S_w(t-1) \end{pmatrix} + \begin{pmatrix} I & 0 \\ 0 & \Gamma_w \end{pmatrix} \begin{pmatrix} v_\beta(t) \\ v_w(t) \end{pmatrix},$$

$$Y(t) = (X(t)\ 1\ 0\ \cdots\ 0)S(t).$$

이 시스템의 모수들은 $Cov(v_\beta(t)) = Diag(\sigma_{\beta_1\beta_1}, \cdots, \sigma_{\beta_k\beta_k})$, $Var(v_w(t))$이고, T_w는 p AR모수들을 갖고, Γ_w는 q AR모수들을 갖는다. 이러한 모수들은 우도함수의 새로운 형태를 이용하여 추정된다(Schweppe(1965)).

만약 백색잡음오차 $v_\beta(t)$와 $v_w(t)$가 정규분포를 따른다면, 1-step 앞의 예측오차들인 $\tilde{Y}(t)$는 독립인 $N(0, f(t|t-1))$을 따르는 확률변수가 된다.

관측된 표본오차의 결합확률은 각각의 밀도함수들의 곱으로 표현된다. 만약 상태벡터가 l개의 비 정상원소를 갖고 있다면, 결합밀도함수는 처음 l개의 관측점에 대한 조건부 함수로 표현되며, 미지인 모수들인 Ω의 함수로써 로그우도함수

$$L(\Omega) = -\frac{1}{2}\left(\sum_{t=l}^{n}\log f(t|t-1) + \frac{Y(t)^2}{f(t)}\right)$$

는 상수의 범위 내에 있게된다. 만약 Ω가 주어진다면, 초기값 $S(l+1|l)$, $P(l+1|l)$을 계산하기 위한 처음의 l개의 관측값을 이용하여 $L(\Omega)$를 결정하는데, 이때 KF반복법이 사용된다. 일반적으로 이러한 방법은 어려운 비선형 최적화 문제에 해당된다. 초기에 우리는 $Cov(v_\beta(t)) = qD$로 나타냈다. 여기에서 q는 상수이고, D는 사전에 명시된 상수들의 대각행렬이다. 또한 $w(t)$에 대한 1계인 AR모형으로 출발하여, 두 모수들의 추정문제를 다루었다. 계수들과 AR 모수 값의 변동의 정도에 대한 개략적인 추정값들을 계산하였고, 몇몇 경우에서 이러한 추정값들은 Watson과 Engle(1983)에 의해 개발된 $EM-scoring$ 알고리즘의 초기값들로 이용되어 좀 더 다듬어지게 되었다. 이러한 알고리즘은 계수의 변화와 관측오차를 동시에 고려하는 일반적인 자기회귀 구조에도 이용된다.

모수들에 대한 초기값들은 합리적인 반복횟수 내에서 수렴값을 얻기 위해 매우 중요하다. 계수들의 표준편차 $\sqrt{\sigma_{\beta,\beta}}$는 관측오차에 대한 표준편차의 약 0.6%정도였다. 초기값들로 이러한 사실들을 이용한다면 EM 알고리즘은 일반적으로 3~6회의 반복 후 수렴하게 된다. Harvey와 Phillips(1979), Ansley와 Kohn(1985), Bell과 Hillmer(1987a)에 의하면 KF를 초기화하는 많은 방법들을 소개하고 있다.

3.2 캐나다

3.2.1 서 론

(1) 배경 및 목적

캐나다 노동력조사(LFS:Labour Force Survey)는 대규모 노동시장의 변화 양상 및 시의성 있는 노동시장의 정보를 파악하기 위해 2차 세계대전 이후 도입되었고, 주로 주(Province) 지역 및 국가 단위의 고용 및 실업통계를 생산할 목적으로 설계되었다. LFS는 1945년 분기별 조사로 시작하여 1952년 월별 조사로 변경되었고, 1960년부터 캐나다 실업통계를 생산하기 위한 공식조사로 승인되었다. 그 후 LFS를 통해 노동시장의 다양한 통계를 작성할 수 있도록 표본개편 및 조사방법에 관한 연구가 지속적으로 진행되었고, 현재는 캐나다 노동시장의 세부적인 변화에 관한 정보를 제공할 수 있을 정도로 발전을 거듭하였다. 매월 고용인구와 실업인구 총계 및 실업률에 관한 추정치, 노동인구의 특성(연령, 결혼여부, 교육정도, 가족현황) 등에 관한 공식통계는 LFS를 통해 작성된다.

LFS는 주 지역 및 전국 단위의 고용 및 실업통계 작성 외에 고용보험 경제구역(EIER:Employment Insurance Economic Regions), 센서스 도시지역(CMA:Census Metropolitan Areas) 등과 같은 주 내의 특정 행정구에 대한 통계작성도 가능하도록 표본이 설계 되어있다. 최근에 들어서는 주 지역 내의 센서스 조사구(CD:Census Divisions)와 같은 소지역들에 대해서도 소지역 추정기법을 이용하여 관련 통계지표를 작성하고 있으며, 지방정부의 소지역에 대한 예산 배분 또는 정책 결정 등의 사안에 이러한 소지역 통계들이 이용되고 있다.

LFS 추정치들은 매월 "Labour Force Information"라는 책자를 통해 공표된다. 또한 노동시장의 좀 더 다양한 정보들은 캐나다 통계국의 전자정보 데이터베이스의 일종인 "CANSIM"을 통해 획득할 수 있으며, LFS의 결과로부터 매월 9000 항목 이상의 시계열 자료들이 정기적으로 수정보완된다. 이외에 노동시장의 중심지표가 되는 다양한 주제에 대한 세부적인 고찰을 다룬 "Labour Force Update"가 1997년부터 계간지로써 출간

되고 있으며, 1976년 이래로 최근까지의 방대한 시계열 자료(time series data) 및 횡단면 자료(cross-sectional data)를 포함하고 있는 "Labour Force Historical Review on CD-ROM"이 매년 제작되고 있다.
캐나다 노동력 조사에 의해서 매월 발표되는 통계수치는 자영업, 부업과 전업을 포함한 취업자 총수와 실업자 총수이다. 매월 발표하는 노동시장의 표준지표는 실업률, 취업률과 경제활동 참가율이고 노동력 조사의 주요 정보 요소로서 15세 이상 인구의 개인적 특성은 나이, 성별, 혼인상태, 교육정도와 가족사항이다.

취업통계의 추정값들에는 인구학적 특성, 산업과 업종, 정규직과 통상적인 근로시간 등이 포함되어 있으며 설문내용에는 비자발적 부업적 취업, 복수 직업 여부와 휴직 등에 대해서 분석할 수 있는 주제들이 포함되어 있다. 특히 1997년 이후에는 근로자들의 노조가입 여부와 임금수준에 대한 정보와 작업장의 근로자 수 및 직업의 정규직 또는 임시직 여부에 대한 정보를 제공하고 있다.

실업통계의 추정값은 인구학적 범주별, 실업기간, 구직활동 전의 활동 및 바로 이전 직장에서 이직한 이유 등에 대한 정보를 제공하고 있다. 노동력 조사에 의해서 발표되는 통계는 국가 단위와 주 단위 추정값이 핵심적인 내용이지만 경제구역(ER : Economic Region)과 센서스 도시지역(CMA ; Census Metropolitan Area)과 같은 소지역 단위에 대한 노동력 상태의 추정값을 제공하고 있다.

(2) 노동력 상태 결정과정

취업과 실업의 개념은 생산의 요소로서 노동력 공급이론을 근거로 정의하였으며 생산은 국민계정 체계(SNA : the System of National Accounts)에서 언급한 것과 같이 상품과 서비스로 정의되는 개념이므로 작업의 목적이나 성질에서는 보수를 받는 근로활동과 조금도 차이가 없는 무보수의 가사노동이나 자원봉사활동은 근로시간으로 계산하지 않고 있다.

노동력 공급의 측정단위는 개별적인 근로시간이지만 조사에서 모집단의 개별적인 구성원들의 구분은 취업, 실업, 비경제활동 인구로 분류되어야 한다. 조사기간 중에 보수 근로 중인 사람은 근로시간에 상관없이 취업자로 구분하고 근로시간과 무관하게 노동시장에서 구직 행위가 있을 경우에는 실업자로 구분한다. 나머지 인구는 현재 일을 하고 있지 않거나 또는 노동시장에서 구직 활동하지 않는 경우로 비경제활동 인구로 정의한다.

통계조사에서 적용하는 취업자와 실업자의 정의와 개념은 국제노동기구(ILO : International Labor Organization)의 기준에 준거하고 있다.

(가) 취업자(Employment)

조사대상 기간 중에 직업이 있거나 개인 사업에서 일을 하는 사람을 말하며 자기에게 직접적인 소득이 없을 지라도 가구단위로 운영되는 농장, 사업체 또는 전문적인 기관에서 일하는 경우도 포함된다. 또한 직업이나 사업체가 있을지라도 일시적인 병이나, 휴가, 노동쟁의 등의 이유로 일을 하지 못하는 일시적인 휴직자도 포함한다.

(나) 실업자(Unemployment)

이용되지 못하는 공급된 노동력으로 실업자를 정의하고 있으며 실업자의 구분은 구직행위와 근로행위의 준비여부를 기준으로 하고 있으나 구직행위는 가구조사에서 구체적이고 지속적인 의사 표명을 전제로 하고 있다. 실업자는 조사 대상 기간 중 다음 3가지 항목 중 하나에 해당하는 사람으로 정의될 수 있다.

① 현재 휴직 중이지만 근로 활동이 가능하고 복직을 기대할 수 있는 경우

② 일을 하지 않고 있으나 구직 활동을 하고 있으며 일이 주어지면 바로 시작할 수 있는 경우

③ 조사대상 기간으로부터 4주 이내에 새로운 일을 할 직업이 주어졌으며 바로 일을 할 수 있는 경우

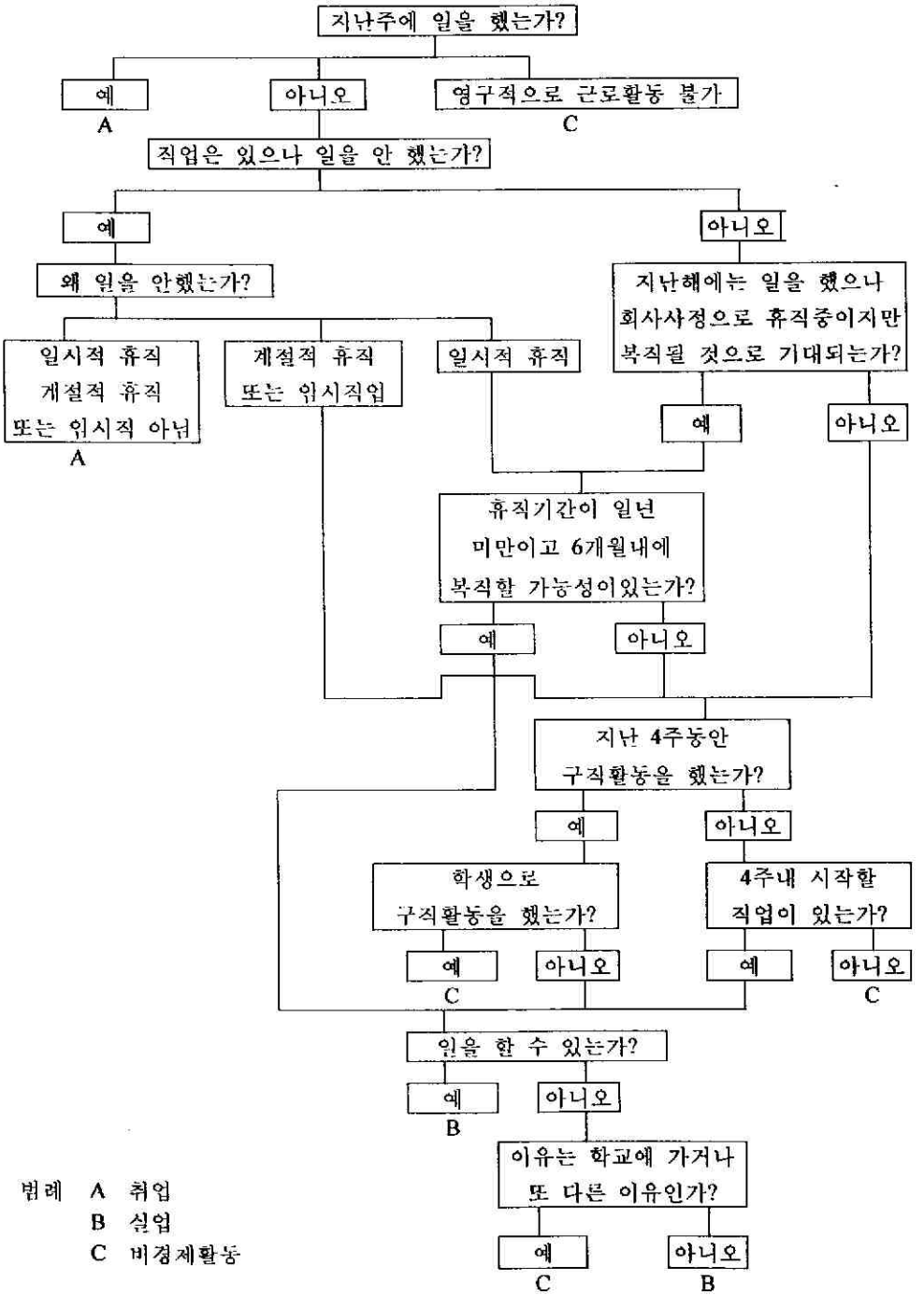
현재 학교에 다니는 학생이 전업적인 일을 찾고 있을 지라도 구직활동

으로 간주하지 않으며 여름방학동안의 근로 활동이나 구직 활동 역시 실업자의 분류로 고려하지 않고 있다.

(다) 비경제활동인구

노동시장에서 노동력을 제공하거나 참여할 의사가 없거나 가능성이 없는 사람을 말하며 이들은 실업자도 아니고 취업자도 아닌 상태의 사람들이다. 여기에는 순수하게 가사만 돌보는 사람, 학생, 투자배당금이나 연금 등을 받아서 생활하는 사람들이 포함된다.

■ 노동력 상태결정


지난주에 일을 했는가?
예
A
아니오
영구적으로 근로활동 불가
C
직업은 있으나 일을 안 했는가?
예
아니오
왜 일을 안했는가?
지난해에는 일을 했으나 회사사정으로 휴직중이지만 복직될 것으로 기대되는가?
일시적 휴직
계절적 휴직
또는 임시직 아님
A
계절적 휴직
또는 임시직업
일시적 휴직
예
아니오
휴직기간이 일년 미만이고 6개월내에 복직할 가능성이있는가?
예
아니오
지난 4주동안 구직활동을 했는가?
예
아니오
학생으로 구직활동을 했는가?
4주내 시작할 직업이 있는가?
예
C
아니오
예
아니오
C
일을 할 수 있는가?
예
B
아니오
이유는 학교에 가거나 또 다른 이유인가?
예
C
아니오
B
범례 A 취업
B 실업
C 비경제활동

LFS는 캐나다 인구의 약 98%에 해당하는 인구를 목표모집단(target population)으로 설정하였다. 캐나다 북서부지역, 인디언 보호구역, 왕실 소유지, 수감자 및 직업군인에 해당하는 약 2%는 LFS의 조사대상에서 제외된다. 1998년 12월 현재 LFS의 표본크기는 52,350가구이다. 1970년 표본개편을 통해 LFS의 표본크기는 이전의 35,000가구에서 50,000가구로 증편된 후, 1980년 47,000가구로 감소되었다가 1989년 주 지역 내의 행정구역인 EIER 지역의 통계작성에 신뢰성을 확보하기 위해 63,000조사가구로 표본설계가 대폭 개편되었으며, 1993년에 약 59,000 조사가구로 다시 감소된 후 1995년부터 52,350 조사가구(EIER 지역에 16,500 조사가구, 나머지 지역에 35,850 조사가구)로 조정되어 현재에 이르고 있다.

LFS의 표본교체(sample rotation) 체계는 일종의 연동교체표본설계(rotating panel sample design)를 따른다. 표본가구는 6개의 부차표본(sub-sample)로 분리되어 6개월간 관리되며, 매월 1/6의 표본이 새로운 표본으로 대체되는 형식이다. 캐나다의 LFS는 매월 15일이 포함되는 주중에 실시되며, 80명의 조사전문인력을 포함하여 약 850명의 조사인력이 투입되어 조사가 이루어지고, 조사결과는 5개의 지방사무소(RO: Regional Office)에서 각각 취합되어 중앙으로 이관된다. 첫 달의 조사는 방문면접 형식을 취하며, 이 후 연속되는 다섯달은 전화조사를 통해 조사가 이루어진다. 조사자는 휴대용 컴퓨터를 이용하여 직접 설문 항목의 결과를 입력하는 일종의 컴퓨터 보조 면접(CAI: Computer Assisted Interviewing) 방식을 이용한다. 조사결과는 조사완료 시점에서 정확히 13일 후에 공표된다.

3.2.2 층화 및 추출단위 구성

캐나다의 주(Province) 지역들은 지리적인 경계에 의해 여러 개의 경제구역(ER:Economic Regions)들로 분할되며, 현재 캐나다 ER 지역은 총 72개가 존재한다. LFS에서는 1960년대 이후로 이러한 ER 지역들을 캐나다 노동력 조사의 1차 층(primary strata)으로 이용해오고 있다. 주 지역

및 전국 단위의 통계 작성에 ER 지역들이 이용된다.

초기의 LFS에서는 ER 지역이 표본설계 시 반영되었던 주 지역 내의 유일한 행정구역이었고 노동력 조사의 관심은 이러한 ER 지역에 집중되었으나, 1989년부터 HRDC(Human Resources Development Canada)의 자금지원에 의해 16,500개의 표본조사가구가 LFS에 추가되면서 EIER 지역 에 대한 노동력 조사도 추가적으로 가능하게 되었다. 따라서 현재의 노동력조사에서는 ER 지역과 EIER 지역 모두가 표본설계 시 층화에 반영되며, 추가표본은 주로 EIER 지역의 추정치의 신뢰도를 확보하기 위해 할당된다. 여기에서 ER 지역과 EIER 지역은 서로 조사 목적이 상이한 지역들이며 133개 지역이 서로 중복되어 조사된다. 한편 LFS 층화 시 반영되는 행정구역으로 CMA 지역을 들 수 있다. CMA 지역은 인구 100,000명 이상인 지역들로써 현재의 CMA 지역은 EIER 지역과 정확히 일치한다.

대영역 내에서의 세부적인 층화는 지리적인 구분에 관계없이 집락화 알고리즘에 의해 이루어진다. 그룹들 간의 가중 제곱합을 최소화하는 층화변수들을 이용하여 가능한한 동질적인 층으로 분할되며, 세부적인 알고리즘은 Drew et al.(1985)과 Singh et al.(1990)에서 참조할 수 있다. 층화 알고리즘에 이용되는 층화변수들은 다음과 같다. 이 층화변수들은 1991년 센서스 자료를 이용하여 선정되었으며, 각각의 층화변수들은 전체 인구의 2%이상을 설명할 수 변수들로 선정되었다.

- o 농업부문 종사자 수
- o 임업, 어업부문 종사자 수
- o 광업부문 종사자 수
- o 제조업(소비재분야) 종사자 수
- o 제조업(고무, 플라스틱, 가죽분야) 종사자 수
- o 제조업(섬유, 의류 분야) 종사자 수
- o 제조업(가구, 펄프, 제지, 인쇄, 목재분야) 종사자 수
- o 제조업(금속, 광업분야) 종사자 수

o 제조업(석유화학, 화학분야) 종사자 수
o 운수업 부문 종사자 수
o 건설업 부문 종사자 수
o 서비스업(상업분야) 종사자 수
o 서비스업(금융분야) 종사자 수
o 서비스업(개인/사업분야) 종사자 수
o 서비스업(정부분야) 종사자 수
o 종사인원 총계
o 총 소득
o 15세 이상 인구
o 15-24세 인구
o 55세 이상 인구
o 1인 거주 가구 수
o 2인 거주 가구 수
o 개인 소유 가구 수
o 총 임차료
o 고졸학력 인구
o 영어를 모국어로 하는 인구
o 프랑스어를 모국어로 하는 인구
o 영어/프랑스어 이외의 언어를 모국어로 하는 인구

LFS 추출틀은 농촌 지역, 인구 50,000명 이상의 대도시 지역과 소도시 지역의 3가지 유형의 지역들로 구분된다. 각 지역에 대한 층화는 다음과 같은 방법으로 이루어진다.

농촌지역에서 층화는 EI 지역과 EIER 지역의 교집합 내에 있는 2~3개의 CD(Census Division) 지역들을 함께 묶어 지리적 층으로 구성하였다.

캐나다 내의 대도시 지역인 17개의 CMA 지역들에 대해서는 충분한 수의 아파트들이 있기 때문에 각각의 CMA 지역들에 대해 독립적인 아파

트 추출틀을 작성하며, 아파트 추출틀을 제외한 나머지 지역에 대해서는 일종의 지역 추출틀(area frame)을 형성하였다. 또한 $100,000 이상의 평균소득을 갖는 고소득 지역들은 독립된 층으로 구성하여 고소득 가구들에 대한 대표성을 제고하였고, 부수적으로 소득관련 조사 및 소득관련 정보 수집이 용이하도록 하였다. 이러한 부수적인 정보는 고소득층에 대한 무응답의 경향을 분석하고자 하는 경우에도 유용하게 이용될 수 있다. 아파트 추출틀과 고소득 층을 제외한 나머지 지역들은 SNF(Street Network File) 지역으로 구분하였다. 최종적으로 마지막 층에는 적어도 48가구의 표본이 배정되도록 하였다.

LFS 표본설계에서 대규모 CMA 지역들에 대해서는 아파트 추출틀을 사용해오고 있다. 현재 18개 CMA 지역에서 표본 추출틀로써 이 목록이 이용되며, 각 CMA 지역에서 새로운 아파트가 건설되면 바로 표본목록에 추가된다. 또한 7개의 CMA 지역에 대해 평균소득이 $20,000 미만의 저소득 아파트 단지를 파악하여 저소득 아파트 층을 구성하여, 고소득 아파트 층과 더불어 소득관련 정보를 획득한다. 캐나다의 각 CMA 지역에 대한 아파트 추출틀 층화 현황은 다음 <표3.12>에 주어졌다.

<표 3.12> 아파트 추출틀 층화

CMA	지리적 층	층의 총수	CMA	지리적 층	층의 총수
Halifax	2	2	London	1	2
Quebec	2	2	Windsor	1	2
Montreal*	4	9	Winnipeg*	1	6
Ottawa-Hull*	3	6	Saskatoon	1	1
Oshawa	1	2	Calgary*	1	3
Toronto*	6	16	Edmonton*	1	3
Hamilton	2	4	Vancouver*	4	6
St. Catharines	1	1	Victoria	1	1
Kitchener	2	2	Total	34	68

(주) ① "*" 는 저소득 아파트 층을 갖는 CMA 지역을 표시함

② "층의 총수"는 저소득 아파트 층을 포함한 수임

대부분의 소도시에서는 EA(enumeration area) 지역을 층화단위로 이용하였다. 조사 비용을 절약하기 위해 각 층에서 직접 표본가구를 추출하지 않고, 우선 각 층을 여러 개의 집락으로 구분하고 추출된 집락 내에서 표본가구들을 추출하였다. 통상적으로 농촌지역에서는 EA 지역이 집락으로 사용되었고, 도시지역에서는 좀 더 다양한 형태의 집락들이 이용되었다. 다음 <표3.13>은 LFS 표본설계에 이용된 일 단계 단위(first-stage unit)의 유형들을 요약한 것이다. 여기에서 "추출가구 수"는 LFS에서 조사되는 가구 수를 나타낸다. 집락과 표본가구에 대한 추출방법은 다음 절에서 논의하기로 한다.

<표 3.13>일단계 단위(first-stage unit), 단위당 가구 수와 추출가구 수

지 역	추출단위	단위당 가구수	추출가구수
Toronto, Montreal, Vancouver	cluster	200-250	6
Other cities	cluster	150-200	8
Apartment frame	apartment	varies	5
Most rural areas and non-SNF prarts of cities	EA	300	10

3.2.3 표본배정, 추출, 순환 (Sample allocation, selection and rotation)

LFS의 표본설계는 캐나다 전체, 주(Province) 지역, EIER 지역, CMA 지역, ER 지역에 대해 각각 다음과 같은 실업자 추정값의 목표 CV값이 만족되도록 설계되었다. 캐나다 전체의 실업자 추정치의 목표 CV 는 약 2%이내, 주(Province) 지역의 CV 값은 약 4~7%선에서 관리되도록 하였다. EIER 지역과 CMA 지역은 분기 실업자 추정치의 CV 값이 15% 이내가 되도록 하였고, 여기에서 하나의 EIER 지역에 배당되는 최소 표본크기는 매월 600가구가 배정된다. ER 지역에 대한 분기별 목표 CV 값은 25% 이내가 되기를 기대하고 있다. 캐나다의 각 주 내의 ER 지역, EIER

지역과 CMA 지역의 현황은 다음 <표3.14>와 같다.

LFS의 표본크기는 총 52,350가구로써 HRDC의 자금지원에 의해 추가된 16,500 표본가구를 제외한 35,850 표본가구는 전국 및 주(Province) 지역의 최적 추정을 위해 배정되며, 이 표본을 핵심표본(core sample)이라고 지칭한다. 핵심표본은 각 주의 실업률 추정의 목표정도를 만족하도록 배정된다. 주 내에서 각각의 ER 지역에 대한 핵심표본의 배정은 총 가구수에 비례하여 배정하며, 최소한 200~300 표본가구가 배정되도록 하였다. 또한 16,500 가구의 추가표본은 핵심표본을 보충하여 EIER 지역의 추정의 정확도를 높이기 위해 추가로 배정된다. 우선 핵심표본 만으로 EIER 지역에 대한 CV 값을 계산 한 후, CV 값이 큰 EIER 지역에 대해 추가표본을 배정하는 형식을 취하며, EIER 지역에 대해서는 최소한 600 가구가 배정되도록 하였다.

<표 3.14> ER, EIER, CMA 지역 현황

주(Province)	ERs	EIERs	CMAs
Newfoundland	4	3	1
Prince Edward I	1	1	0
Nova Scotia	5	5	1
New Brunswick	5	4	1
Quebec	16	13	6*
Ontario	11	18	10*
Manitoba	8	3	1
Saskatchewan	6	4	2
Alberta	8	4	2
British Columbia	8	6	2
Canada	72	61	25*

(주) Ottawa-Hull CMA 는 Ontario와 Quebec 양쪽에서 카운트 됨.

LFS에서 표본 추출은 대부분의 지역에서 이 단계 추출(two-stage sampling) 만을 이용한다. 일 단계 추출은 지역 추출(area sample)이고,

이 단계 추출은 일 단계 추출 지역들에 대해 거주가구 목록을 작성하여 이 목록으로부터 표본가구를 추출한다.

농촌지역의 경우 일 단계 추출단위는 EA지역, 이 단계 추출단위는 가구 단위가 된다. 일 단계의 지역 추출은 확률비례계통추출법, 이 단계의 가구 추출은 계통추출법이 이용된다. 아파트 추출틀을 이용하지 않는 주요 도시지역의 경우 일 단계 추출단위는 EA지역, 이 단계 추출단위는 가구 단위가 된다. 일 단계의 지역 추출은 Rao-Hartley-Cochran(RHC)의 랜덤그룹법이 이용되며, 이 단계의 가구 추출은 계통추출법이 이용된다. 아파트 추출틀을 이용하지 않는 기타 도시 지역들도 몇몇 특별한 경우를 제외하고는 주요 도시지역과 유사한 방법을 취하고 있다. 랜덤그룹법은 인구유입 및 유출양상을 LFS의 표본크기에 반영시킨 방법으로써 시기에 따라 유동성이 큰 대도시 지역의 인구변화를 표본크기에 반영시킨 방법이다. 아파트 추출틀을 이용하는 주요 도시지역의 경우 일 단계 추출단위는 아파트 단지, 이 단계 추출단위는 가구 단위가 된다. 일 단계 아파트 단지에 대한 추출은 확률비례계통추출법, 이 단계 가구단위의 추출은 계통추출법이 이용된다.

LFS에서는 표본의 일부가 매달 새로운 표본으로 교체된다. 다 단계 표본설계의 매 단계에서 표본 단위의 교체가 이루어지며, 표본추출의 최종 단위인 가구는 6개월마다 교체된다. 조사자의 업무 부담과 응답가구가 오랜 기간 동안 표본으로 조사되는 동안 발생할 수 있는 무응답에 대한 가능성을 최소화하기 위해 매 달 1/6의 표본이 대체된다. 따라서 하나의 집락에 포함된 표본가구는 연속적으로 6개월 간 조사된 후 표본에서 완전히 삭제되고 새로운 표본으로 대체된다.

3.2.4 특별조사와 보충조사

캐나다 노동력조사 외에 추가적인 많은 조사가 LFS 추출틀 또는 LFS 표본을 이용하여 실시되며, 이러한 조사들은 캐나다 정부 부처의 자금 지원에 의해 시행된다. 다음 <표3.15>은 LFS 연동교체표본 또는 LFS

추출틀을 이용한 특별조사 및 보충조사 현황을 요약한 것이다.

<표3.15> 특별조사 및 보충조사 현황(1998년 현재)

Survey	조사기간	Survey	조사기간
Canadian Travel Survey	1-12월 (매월조사)	Homeowner Repair and Renovation Survey	3월
Employment Insurance Coverage	1월	Survey of Consumer Finances	4월
Survey of Household Spending	1월-3월	Cultural Capital Survey	4월
Survey of Labour and Income Dynamics	1월, 5월	National Population Health Survey	2, 6, 8, 11월
Adult Education and Training Survey	1월	National Longitudinal Survey of Children	11월
Resident Telephon Services Survey	2, 5, 8, 11월	Survey of Work Arrangements	11월
Survey of Household Energy Use	2월		

SCF(Survey of Consumer Finances)는 일년에 한번 시행되는 소비자 재정에 관한 조사로써 보통 4월에 실시된다. 모든 가구들이 4개의 순환 그룹에 배정되어 LFS 조사에 추가된다. 4월의 LFS 조사에 앞서 각 가정에 우편으로 설문지가 배달되고 LFS 조사 기간 동안 회수된다. SCF에서 조사하는 주요 정보는 소비자의 평균 소득과 세금 공제 전과 공제 후의 소득관련 정보들로써, 이러한 결과들은 저소득 기준점 결정 등과 같은 소득 관련 측도들로 이용된다.

SHS(Survey of Household Spending)는 일년에 한번 실시되는 가계비 지출 및 식료품 지출 조사로써 보통 1월~3월에 걸쳐 시행되며, 소비자 가격 지표 산정의 정보로써 이용된다. SHS 조사는 LFS 표본을 포함하는 집락들에서 표본가구를 추출하나 표본가구들은 LFS 조사와는 별도로 조사된다.

SLID(Survey of Labour and Income Dynamics)는 노동력과 소득 변천과정 파악을 조사의 목적으로 일년에 2번, 1월과 5월에 조사가 이루어

지며 LFS와 병행하여 시행된다. 조사 결과는 저소득 층의 유입, 유출 동향, 노동 시장의 변화, 가족변화와 경제적인 복지와의 상관관계 등을 분석하기 위해 이용된다.

NPHS(National Population Health Survey)는 일종의 국민 건강조사로써 분기별로 실시되며 국민 건강의 계절적 요인을 파악하기 위해 NPHS 표본을 2월, 6월, 8월과 11월조사에 각각 1/4씩 배분한다. NPHS 조사는 주 정부의 자금지원에 의해 실시되며, LFS 조사와 병행하여 실시되지는 않는다. 초기에 선택된 가구의 구성원은 2년에 한번 심층 면접을 받으며 20년 동안 지속적으로 관리된다. 기초적인 건강정보는 거주 가구의 모든 구성원들에 대해서 취합되며 이러한 시계열 자료는 횡단면 추정 목적에 이용된다.

NLSCY(National Longitudinal Survey of Children and Youth)는 아동에 대해 유아기부터 성인기까지의 발달 과정을 모니터하는 조사로써 일년에 한번 실시되는 일종의 시계열적 장기조사에 해당한다. LFS 조사가구의 약 30%만이 대상연령에 있는 아동을 포함하는 관계로 NPHS의 추가 표본이 조사에 이용되며, 조사방법은 NPHS 조사와 유사하다.

3.2.5 가중치와 추정

LFS의 가중치는 다음의 세가지 요인들에 기인하여 작성된다. 표본설계를 반영하는 일종의 설계 가중치, 무응답 가구를 보정하는 일종의 무응답 보정 가중치와 모집단 총계에 표본 추정치를 일치시키는 일종의 사후 층화 가중치(g-factor)의 3가지 요인들에 의해 LFS의 최종 가중치가 결정된다. LFS 조사 추정치는 LFS 표본이 확률표본이기 때문에 추정치의 표본오차를 추정하여 신뢰도를 판단할 수 있다. 표본설계에 계획되지 않은 관심지역들에 대한 추정문제는 소지역 추정기법을 도입하여 추정량의 신뢰도를 확보하였다. EIER 지역들이 여기에 해당된다.

LFS에서는 가구단위에 대해 계통추출이 이루어지는 마지막 단계를 제외하고는 모든 단계에서 확률비례추출법으로 표본이 추출되는 층화 다단

계 추출법이 이용된다. 설계가중치는 이러한 복합표본설계에서 가장 기본적인 가중치로써 각 추출단위에 대해서 추출률의 역수로 결정된다. 예를 들어 층에 대한 설계가중치는 $R_h = N_h/n_h$와 같이 표현될 수 있다. 여기에서 n_h는 층 h에 대한 표본가구 수, N_h는 표본설계 시 층 h에 있는 총 가구 수를 나타낸다.

이 단계 추출의 경우는 다음과 같이 설명될 수 있다. n_{hj}^*를 층 h에 있는 j번째의 일 단계 추출단위(FSU: First Stage Unit)라고 하자. 추출해야 할 FSU의 수는 $n_{1h} = n_h/n_{hj}^*$ 로 주어진다. h층에서 j번째 FSU에 있는 가구들의 수를 N_{hj}라 하면 j번째의 FSU에 대한 추출율은 $R_{hj} = N_{hj}/n_{hj}^*$로 주어지며, 이때 j번째 FSU에 대한 일 단계 산입확률(inclusion probability)은

$$\pi_{1hj} = \frac{n_{1h}}{\sum_{j \in h} R_{hj}} R_{hj}$$

로 나타낼 수 있다. j번째 FSU가 주어진 상태에서 k번째 가구가 선택될 조건부 산입확률은

$$\pi_{k/j} = \frac{n_{hj}^*}{N_{hj}} = \frac{1}{R_{hj}}$$

로 주어지며, h번째 층에서 k번째 가구에 대한 산입확률은

$$\begin{aligned} \pi_{hk} &= \pi_{1hj}\pi_{k/j} \\ &= \frac{n_{1h}}{\sum_{j \in h} R_{hj}} R_{hj} \frac{1}{R_{hj}} \\ &= \frac{n_{1h}}{\sum_{j \in h} R_{hj}} \end{aligned}$$

로 계산될 수 있다. 여기에서

$$\sum_{j \in h} R_{hj} = \sum_{j \in h} \frac{N_{hj}}{n^*_{hj}}$$

$$= \frac{n_{1h}}{n_h} \sum_{j \in h} N_{hj}$$

$$= n_{1h} R_h$$

인 관계가 성립하며, 결국 산입확률은 $1/R_h$과 같게 된다. 참고적으로 LFS에서는 기본적으로 각 층 내에서 동일한 설계 가중치 R_h를 이용한다.

LFS와 같은 연속조사에서는 고정된 추출률을 유지할 경우, 인구 증가 및 인구 유입 등으로 시간이 지남에 따라 표본 수도 증가하게 된다. 급속한 인구 증가가 발생하는 특정지역의 표본 수는 급격히 증가되어 조사원의 업무부담을 가중시키고 조사의 질을 떨어뜨릴 가능성이 있다. 이러한 문제점을 해결하기 위해 LFS 표본 설계에서는 이러한 집락들에 대해서 표본을 부차추출(subsampling)하여 전체 표본크기를 일정하게 유지한 후 집락 부차가중치(cluster subweight)를 산출하여 LFS 설계가중치에 추가하여 사용하였다. 구체적으로 SC(Subclustering)방법, SRC(Self-Representing Cluster)방법, CSS(Cluster Subsampling)방법 등을 이용한다.

SC 방법은 인구유입 등으로 인해 원래의 집락의 크기가 약 3배를 초과할 때 이러한 집락들을 여러 개의 작은 집락으로 분할하여 표본 가구를 추출하는 방법이다. 분할된 작은 집락들의 표본 수를 n_{2hj}라 하자. N_{hji}를 새로운 집락의 크기, n_{hji}를 새로운 집락의 표본크기, R_{hji}를 새로운 집락들에 대한 추출률이라 할 때, 새로운 부차집락들로부터 집락추출률

$$R^*_{hj} = \sum_{i \in j} \frac{R_{hji}}{n_{2hj}}$$

를 얻을 수 있다. 이때 집락 부차가중치는 $K = R^*_{hj}/R_{hj}$로 주어지며, 설계가중치는 원래의 가중치에 이러한 부차가중치를 곱하여 계산된다.

SRC 방법은 임의의 층에서 증가하는 거주단위들의 특성이 나머지 단

위들의 특성과 구별되거나 집락의 규모가 층 규모의 20%를 초과할 경우, 해당 집락을 하나의 층으로 재분류하여 가중치를 결정하고, 나머지 집락들에 대해서는 가중치를 재 보정해 주는 방법이다. 이렇게 생성된 새로운 층을 (hj) 라 하자. 이러한 층 내에서 새로운 집락들이 구성되어 표본들이 추출된다. $N^{N}_{(hj)}$을 새로운 층의 크기, $n^{N}_{(hj)}$을 새로운 층의 표본크기라 할 때, 층 추출률은 $R^{N}_{(hj)} = N^{N}_{(hj)}/n^{N}_{(hj)}$으로 주어지며, 이러한 새로운 층으로부터 추출된 가구들은 가중치 $K = R^{N}_{(hj)}/R_h$가 배정된다. 나머지 집락들의 가구들에 대한 가중치는 다음과 같은 방법으로 보정된다. 하나의 층에서 6개의 집락들이 추출되었다고 하자. 층 내에서 1개의 성장집락이 발생하여 새로운 층으로 재 분류 되었다면 나머지 5개의 집락의 가구들에 대한 가중치는 보정되어야만 한다. $N^{R}_{h} = N_h - N^{N}_{(hj)}$를 새로운 층을 제외한 층의 크기, $n^{R}_{h} = n_h - n_{hj}$를 표본의 크기라 할 때, 층에 대한 추출률은 $R^{R}_{h} = N^{R}_{h}/n^{R}_{h}$로 계산되며, 이때 집락 부차가중치는 $K = R^{R}_{h}/R_h$로 결정된다.

CSS 방법은 집락단위가 부차표본으로 추출되는 경우에 이용된다. 임의의 집락이 추출률 R_{hj}로 추출되고, 부차추출이 추출률 $R^{\star}_{hj}$로 추출되었다면, 이때 집락 부차가중치로써 $K = R^{\star}_{hj}/R_{hj}$를 이용하는 방법이다.

표본추출의 마지막 단계는 계통추출이 이용된다. 가구에 대한 추출률은 일관되게 적용되기 때문에 인구 증가로 인한 표본가구수의 증가는 조사규모 및 조사비용의 증가를 초래한다. 이러한 조사비용을 통제하기 위해 표본 수 안정화 작업이 수행된다. 표본 수 안정화 작업은 전체 표본가구 수를 적절한 수준에서 유지하기 위해 초과 표본들을 랜덤하게 제거하는 작업을 말한다. 이러한 과정에서 가구단위의 산입확률(inclusion probability)은 당연히 바뀌게 된다. 현 표본설계에서는 같은 EIER 지역에 속해 있고 같은 순환그룹에 포함되어 있는 모든 가구단위들을 안정화 지역(stabilization area)으로 정의하고, 각 안정화 지역 a에 대해서 표본크기

가 결정된다. 안정화 지역 a의 표본크기를 b_a라고 나타내자. 안정화 지역에서 안정화작업을 거치지 않은 표본크기를 n_a라 할 때, n_a가 b_a를 초과한다면 $n_a - b_a$만큼의 표본가구가 랜덤하게 제거된다. LFS에서는 임의의 집락이 부차추출 되었을 경우에는 이 집락을 안정화 작업에서 제외시키기 때문에 이러한 집락에 포함된 가구단위들은 안정화 가중치(stabilization weight)의 영향을 받지 않는다. 이러한 집락을 제외시킨 안정화 지역 a의 가구단위의 총계를 c_a라 할 때, 지역 a에 있는 조사가구의 안정화 가중치는 $s_a = (n_a - c_a)/(b_a - c_a)$가 이용된다.

통계조사에서 무응답은 항목 무응답(item nonresponse)과 단위 무응답(unit nonresponse)의 두 가지 유형으로 분류된다. LFS에서는 항목 무응답은 대체법(imputation)으로, 단위 무응답은 전체적인 가중치 조정을 통해 처리하고 있다. 항목 무응답은 지리적으로 또는 인구 통계적으로 유사한 특성을 갖는 응답자의 응답패턴을 이용하여 대체된다. LFS에서는 단위 무응답을 처리하기 위한 무응답 층을 " 같은 EIER 지역에 속하고, 같은 유형의 지역적 특성을 가지며, 같은 표본순환그룹 내에 있는 가구들" 로써 정의한다. 무응답 층을 올 바르게 구성하였을 경우 동일한 무응답 층에 속한 응답 가구와 무응답 가구는 서로 비슷한 속성을 갖기 때문에 응답 가구가 무응답 가구를 대표한다고 가정할 수 있게 된다. 단위 무응답에 대한 보정은 설계 가중치에 보정요인

$$f_b = \frac{\sum_{k=1}^{n} \pi_k^{-1}}{\sum_{k=1}^{r} \pi_k^{-1}}$$

을 곱하여 계산한다. 여기에서 π_k^{-1}은 각 표본가구에 부여된 설계 가중치, n은 무응답 층 b에 있는 표본가구 수, r은 응답가구 수를 나타낸다.

LFS에서의 최종 가중치는 특별한 경우를 제외하고는 설계 가중치와 보조정보로부터 얻게 되는 사후 가중치(g-factor)의 곱으로 계산되며, 여

기에서 사후 가중치 계산은 일반적인 회귀추정방법이 이용된다. 자세한 추정절차는 Lemaitre and Dufour (1987)에 소개되어 있다. 먼저 다음과 같은 기호를 정의하자.

$p=1,2,\cdots,10$: 주 지역을 나타내는 기호,

$u=1,2,\cdots,U$: p번째 주 내에 있는 EIER 지역,

$f=1,2,\cdots,F$: u번째 EIER 지역 내에 있는 추출틀의 형태,

$h=1,2,\cdots,H$: 추출틀 f 내에 있는 층,

$r=1,2,\cdots,6$: h번째 층 내에 있는 순환그룹(패널),

$j=1,2,\cdots,J$: r순환그룹의 집락,

$k=1,2,\cdots,K$: 집락 j에서의 가구,

$i=1,2,\cdots,c_k$: k번째 가구 내에 있는 구성원,

SC(subclustering) 방법은 우선 해당 집락을 여러 개의 작은 집락들로 재 구성한 후 표본 추출을 위한 집락들을 선정하고 전체 표본 수를 참고하여 선정된 집락 내에서 표본가구를 추출한다. 원래의 집락 추출률이 $R_{pufh\cdot j}$, 해당 집락 추출률이 $R^*_{pufh\cdot j}$라 하면 해당 집락의 부차 가중치는 $c_{pufh\cdot j}=R^*_{pufh\cdot j}/R_{pufh\cdot j}$이다.

SRC(self-representing cluster) 방법은 층 내에서 성장 집락을 분리하여 새로운 층(h)으로 구성하고 h층 내에서 여러 개의 집락들을 재 구성한 후 표본가구를 추출하는 방법이다. 원래의 층 추출률이 R_{pufh}, 새로 형성된 층의 추출률이 R^*_{pufh}라 하면 이때 새로 형성된 층에서 가구단위들에 배정되는 집락 부차가중치(cluster subweight)는 $c_{pufh}=R^*_{pufh}/R_{pufh}$이다. 성장집락을 제외한 나머지 집락의 가구단위에 대해서는 보정된 집락 부차가중치 $c_{pufh}=R^R_{pufh}/R_{pufh}$를 적용한다. 여기에서 R^R_{pufh}은 나머지 층의 추출률을 나타낸다.

CSS(cluster Subsampling) 방법은 추출된 가구단위들이 부차추출되고 이러한 부차추출된 가구단위들만 조사하는 방법이다. 원래의 집락 추출률

이 $R_{puflh\cdot j}$이고 부차추출하기 위한 해당 집락 추출률이 $R^{*}_{puflh\cdot j}$라면 이때 집락 부차가중치는 $c_{puflh\cdot j}=R^{*}_{puflh\cdot j}/R_{puflh\cdot j}$이다.

안정화 가중치(stabilization weight)는 앞서 언급되었던 안정화지역(stabilization area) 내에서만 적용된다. 각 안정화 지역 내에서의 표본크기를 $b_{pu\cdot\cdot r}$, 실제 추출된 표본크기를 $n_{pu\cdot\cdot r}$, 안정화 지역에서 CSS 방법으로 추출된 표본의 크기를 $c_{pu\cdot\cdot r}$이라 할 때, 안정화 가중치는

$$s_{pu\cdot\cdot r}=\frac{n_{pu\cdot\cdot r}-c_{pu\cdot\cdot r}}{b_{pu\cdot\cdot r}-c_{pu\cdot\cdot r}}$$

로 계산된다.

이상의 가중치들을 이용하여 조사가구에 대한 설계가중치를 산출하며 LFS에서는 다음과 같은 설계가중치를 고려한다.

$$\pi^{-1}_{puflsrjk}=w_{puflh}\times c_{puflh\cdot j}\times s_{pu\cdot\cdot r}$$

여기에서 w_{puflh}는 같은 층 내에 있는 모든 가구단위들에 대해서 동일한 가중치를 배정했던 표본설계 당시의 가중치를 나타낸다. 표기상의 편의를 위해 앞으로 $\pi^{-1}_{puflsrjk}$를 π^{-1}_{k}로 나타내기로 한다.

LFS에서는 무응답 층에 대해서 무응답 보정을 실시하며, 보정 가중치로써

$$f_{puf\cdot r}=\frac{\sum_{k\in s}\pi^{-1}_{k}}{\sum_{k\in r}\pi^{-1}_{k}}$$

을 이용한다. 여기에서 분자의 s에 대한 합은 무응답 층에 있는 모든 가구들에 대한 합을 나타내며, 분모의 r에 대한 합은 무응답 층에 있는 모든 응답가구들에 대한 합을 나타낸다. 같은 무응답 층에 속해 있는 모든 가구단위들은 동일한 무응답 가중치를 갖는다.

무응답 보정요소가 추가될 경우 부차 가중치는 $a_k=f_{puf\cdot\cdot r}\times\pi^{-1}_{k}$과 같이 설계가중치와 무응답 보정가중치의 곱으로 주어진다. 즉 같은 조사

가구 내의 모든 구성원들은 동일한 부차가중치를 갖는다.

위에서 언급한 부차가중치를 이용하여 고용 인구Y에 대한 총계 추정값을 산출해 보자. 모집단에서 고용인구의 총계를

$$t_y = \sum_U y_i$$

라고 하자. 여기에서 U에 대한 합은 모집단에서 관심영역 내에 있는 모든 구성원들의 합을 나타내며, y_i는 조사대상자가 고용일 경우 1, 아닐 경우 0의 값을 갖는다. 이때 표본조사에 의한 총계 추정치는 부차가중치에 의존하며

$$\hat{t}_{yn} = \sum_s y_i a_i$$

와 같이 표현될 수 있다. 여기에서 s에 대한 합은 표본으로 추출된 조사대상자들에 대한 합을 나타내고, a_i는 부차가중치를 의미한다. 위의 t_y와 $\hat{t}_{yn}$는 각각 다음과 같이 다시 표현할 수 있다.

$$t_y = \sum_{k=1}^{N} \sum_{i=1}^{c_k} y_i = \sum_{k=1}^{N} y_k \ ,$$

$$\hat{t}_{yn} = \sum_{k=1}^{n} a_k \sum_{i=1}^{c_k} y_i = \sum_{k=1}^{n} y_k a_k \ .$$

여기에서 c_k는 k번째 조사가구의 구성원의 수, N은 모집단의 가구 수, n은 표본 가구 수, y_k는 $y_k = \sum_{i \in k} y_i$를 의미하며, k는 가구 총 수, i는 구성원을 나타낸다.

LFS에서 적용하고 있는 마지막 단계의 가중치로써 사후층화 가중치(g-factor)를 들 수 있다. 사후층화 가중치는 사후층화를 통한 보조정보로부터 획득하며 회귀추정방법을 이용하여 산출한다. 각 주 단위의 성별-연령대별 그룹, ER 지역과 CMA 지역에 대한 인구 총계, 센서스 결과에 의한 인구 추계 정보 등이 보조정보로 활용되었다. 추가적인 논의를 위해

다음과 같은 기호를 정의하기로 하자.

y_i : i번째 조사자에 대한 특성치,

y_k : k번째 가구에 대한 특성치 총계,

Q : 추정에 이용된 보조변수의 수, $q=1,2,\cdots,Q$,

x_{qi} : 조사자 i에 대한 q번째 지표변수의 값, 지표변수는 조사자 i가 j번째 범주에 속할 경우 1, 기타 0의 값을 갖는다.

x_{qk} : k번째 가구단위에 속하는 조사자들에 대한 q번째 지표변수값의 총계,

x_k : q번째 원소가 x_{qk}인 $Q\times 1$벡터,

c_k : k번째 가구의 크기,

$\hat{t}_{y\pi}$: 위에서 언급한 부차가중치에 근거한 추정치,

$\hat{t}_{x_q\pi}$: q번째 보조변수에 대한 부차가중치에 근거한 추정치.

사후층화 가중치를 산출하기 위해

$$\hat{t}_{yr}=\hat{t}_{y\pi}+\sum_{q=1}^{Q}\hat{B}_q(t_{x_q}-\hat{t}_{x_q\pi})$$

와 같은 회귀추정량을 이용한다. 여기에서

$$\hat{t}_{x_q\pi}=\sum_s x_{qi}a_i\ ,$$

$$\hat{B}=(\hat{B}_1,\ \cdots,\hat{B}_Q)^T$$

$$=\left(\sum_{k=1}^{n}\frac{x_kx_k^Ta_k}{c_k}\right)^{-1}\sum_{k=1}^{n}\frac{x_ky_ka_k}{c_k}$$

이며, $\left(\sum_{k=1}^{n}\frac{x_kx_k^Ta_k}{c_k}\right)^{-1}$은 $Q\times Q$행렬, $\sum_{k=1}^{n}\frac{x_ky_ka_k}{c_k}$는 $Q\times 1$벡터를 나타낸다.

회귀추정량 $\hat{t}_{yr}$은

$$\hat{t}_{yr}=\sum_{k\in s}y_ka_kg_k$$

와 같이 사후층화 가중치를 포함하는 식으로 재 표현 할 수 있다. 여기서에서

$$g_k = 1 + (t_x - \hat{t}_{xa})^t \left(\sum_{k \in s} \frac{x_k x_k^t a_k}{c_k}\right)^{-1} \frac{x_k}{c_k}$$

로써 일명 g-factor로 불리우는 사후층화 가중치이다. 가구 구성원에 대한 사후층화 가중치는

$$g_i = 1 + (t_x - \hat{t}_{xa})^t \left(\sum_{i \in s} z_i z_i^t a_i\right)^{-1} z_i$$

이며, 여기에서 $z_i = \frac{1}{c_k} \sum_{i=1}^{c_k} x_i$이다. 즉 사후층화 가중치의 특징은 가구에 대한 가중치와 가구 내의 구성원에 대한 가중치가 일치하여 모든 구성원들이 동일한 가중치를 갖는다는 점이다.

LFS의 최종가중치는 설계 가중치, 무응답 조정 가중치와 사후층화 가중치의 곱으로 표현되며, 이러한 최종가중치를 이용하여 주 지역 및 전국 단위의 경제활동인구 총계, 취업자 총계, 실업자 총계, 취업 및 실업률 등이 추정된다.

LFS에서 추정량의 분산계산은 잭나이프 방법을 이용한다. 좀 더 일반적인 경우의 잭나이프 방법에 대한 기술은 Wolter(1985)를 참조할 수 있으며, 여기에서는 LFS에서 적용하는 잭나이프 알고리즘을 소개하기로 한다.

(i) 잭나이프 방법을 적용하기 위해 h번째 층은 J_h개의 반복표본을 갖는다고 가정한다($a = 1, 2, \cdots, J_h$). 우선 특정 반복표본에 해당하는 모든 가구들을 제거한다. 여기에서 해당 표본에서 반복 총계는 $J = \sum_{h=1}^{H} J_h$이고, H는 해당 표본에서 층의 총계를 나타낸다.

(ii) 주어진 층에서 나머지 $J_h - 1$개의 반복표본의 모든 가구들에 대해 부차가중치에 대한 보정이 이루어진다. 보정된 가중치의 값은

$a_k^{adj} = \frac{J_h}{J_h - 1} a_k$이다.

(iii) 보정된 부차가중치와 남아있는 해당표본을 이용하여 관심 추정치 $\hat{t}_{yr(ha)}$를 계산하기 위한 최종가중치를 계산한다. 여기에서 (ha)는 h번째 층으로부터 a번째 반복이 제거되었다는 것을 나타낸다.

해당표본의 모든 반복에 대해서 위의 (i)~(iii)의 절차가 반복되며, 결과로써 관심 추정치에 대한 J개의 서로 다른 추정값을 얻게 된다. 이러한 추정값을 이용하여 추정값의 분산을 계산하며 다음과 같은 분산 추정 공식을 이용한다.

$$\hat{V}(\hat{t}_{yr}) = \sum_{h=1}^{H} \frac{J_h - 1}{J_h} \sum_{a=1}^{J_h} (\hat{t}_{yr(ha)} - \hat{t}_{yr})^2.$$

실업률에 대한 분산 추정은 다음의 추정공식을 이용할 수 있다.

$$\hat{V}(100\frac{\hat{t}_{yr}}{\hat{t}_{zr}}) = 100^2 \sum_{h=1}^{H} \frac{J_h - 1}{J_h} \sum_{a=1}^{J_h} \left(\frac{\hat{t}_{yr(ha)}}{\hat{t}_{zr(ha)}} - \frac{\hat{t}_{yr}}{\hat{t}_{zr}} \right)^2.$$

여기에서 y는 실업자 총계, z는 경제활동인구 총계를 나타내며, 위의 결과는 실업률 $100(y/z)\%$에 대한 잭나이프 분산 추정공식이다.

월 변화량의 추정치에 대한 잭나이프 분산추정을 다음과 같이 산출할 수 있다. 연속되는 두 달의 월 추정치로부터 다음의 차분추정치 $\hat{D}_{yr} = \hat{t}_{yr}^{(2)} - \hat{t}_{yr}^{(1)}$을 고려하자. 여기에서 윗 첨자는 연속되는 월을 나타낸다. 대응되는 잭나이프 추정치는 $\hat{D}_{yr(ha)} = \hat{t}_{yr(ha)}^{(2)} - \hat{t}_{yr(ha)}^{(1)}$ 으로 표현할 수 있다. 이때 분산 추정치는 다음과 같이 주어질 수 있다.

$$\hat{V}(\hat{D}_{yr}) = \sum_{h=1}^{H} \frac{J_h - 1}{J_h} \sum_{a=1}^{J_h} (\hat{D}_{yr(ha)} - \hat{D}_{yr})^2 .$$

LFS의 연속된 두 달의 조사에서는 5/6의 표본이 일치한다. 공통표본을 이용하여 두 달 간의 월 변화량의 차를 추정하는 것이 위의 추정방법에 비해 훨씬 효율적일 수 있다. Singh et al.(1997)은 이러한 공통표본을 이용하여 다음과 같은 합성추정량을 제안하였다.

$$est^{C}_{(t+1)} = K \times est_{(t+1)} + (1-K) \times [est^{C}_{(t)} + change_{common}]$$

$$est^{C}_{(t+1)} = K \times est_{(t+1)} + (1-K) \times [est^{C}_{(t)} + change_{common}],$$

여기에서 윗첨자 C는 복합추정법을 의미하며, $change_{common}$은 공통표본을 이용한 변화량을 나타낸다. 이 추정량은 1998년 현재 캐나다 통계청에서 채택하고 있지는 않지만 새로운 표본설계에서는 사용될 전망이다.

위와 유사한 방법으로 n개의 월 추정치의 평균에 대한 분산추정을 고려할 수 있다. n개의 월 추정치의 평균은

$$\hat{A}_{yr} = \sum_{i=1}^{n} \frac{\hat{t}^{(i)}_{yr}}{n}$$

이고, 이에 대응하는 잭나이프 추정값은

$$\hat{A}_{yr(ha)} = \sum_{i=1}^{n} \frac{\hat{t}^{(i)}_{yr(ha)}}{n}$$

으로 계산할 수 있으며, 추정치의 분산은 다음의 추정공식을 이용할 수 있다.

$$\hat{V}(\hat{A}_{yr}) = \sum_{h=1}^{H} \frac{J_h - 1}{J_h} \sum_{a=1}^{J_h} (\hat{A}_{yr(ha)} - \hat{A}_{yr})^2.$$

3.2.6 데이터 관리

모든 표본조사에서와 마찬가지로 LFS 추정값들도 표본 오차와 비표본 오차를 수반한다. 따라서 조사 추정값들이 올바르게 해석되기 위해서는 추정값들의 정도를 나타내는 측도에 관한 점검이 요구된다.

표본오차에 직접적으로 영향을 미치는 것은 표본크기라 할 수 있다. 일반적으로 표본크기가 증가함에 따라 표본오차는 감소한다. 표본의 크기 뿐만 아니라 모집단의 변동, 추정 및 표본설계의 방법과 같은 요인들이 표본오차에 관련된 요인들이라 할 수 있다. 표본오차는 층화방법, 표본할당, 추출단위의 선택에서 뿐 만 아니라 다 단계 표본설계에서 매 단계에서 선택된 표본 추출방법 및 추정방법 등과 같은 요인들에 크게 의존한

다.

표본설계 및 추정방법에 대한 효율성을 점검하기 위한 측도로는 평균제곱오차 *MSE*가 이용된다. *MSE*는 추정값과 모집단의 실제값과의 편차제곱합의 평균으로 보통 정의된다. 표본오차와 관련된 또 다른 중요한 측도로써 변동계수 *CV*가 이용된다. *Y*가 어떤 특성치에 대한 추정치이고, *d*가 추정치의 표준오차라 할 때, *CV*값은 $(d/Y)\times 100$으로 정의된다. 또한 추정치의 신뢰구간은 표준오차 *d*로부터 추론될 수 있다. 한편 시간에 따른 표본설계의 낙후성을 평가하기 위한 지표로써 설계효과(design effect)가 이용된다. 설계효과는 기존 표본설계로부터 조사된 추정값의 분산과 단순임의표본으로부터 계산된 추정값의 분산의 비로써 정의되며, LFS에서는 표본설계효과(sample design effect)와 전체설계효과(overall design effect)와 같은 두 가지 형태의 설계효과를 계산한다. 표본설계효과는 모총계에 대한 가중치의 조정 없이 부차가중치만을 이용하여 계산되며, 전체설계효과는 앞서 언급되었던 최종가중치를 반영하여 계산된다. 따라서 표본설계효과는 표본설계의 효율성만이 반영되며, 전체설계효과는 층화, 다단계추출, 사후층화 및 추정 등의 표본설계의 전반적인 사항들이 반영된다고 보면 된다.

1997년 1월부터 7월까지의 LFS 조사자료에 대한 주 단위 및 전국 단위의 고용 및 실업관련 추정치들의 *CV* 값이 다음 <표3.16>에 주어졌다.

<표3.16> LFS 고용 및 실업 추정치들의 CV 값 (1997)

Province	Employed CV(%)	Unemployed CV(%)	Province	Employed CV(%)	Unemployed CV(%)
Newfoundland	2.2	6.1	Manitoba	0.91	6.5
Prince Edward Island	1.7	6.5	Saskatchewan	1.1	7.4
Nova Scotia	1.2	5.3	Alberta	0.76	5.9
New Brunswick	1.2	5.5	British Columbia	0.90	5.1
Quebec	0.79	3.5	Canada	0.32	1.72
Ontario	0.54	3.0			

고용과 실업에 대한 월 변화 추정값의 표준오차는 <표3.17>에, 고용과 실업에 관한 설계효과는 <표3.18>에 주어졌다.

<표3.17> LFS 고용 및 실업에 대한 월변화 추정값의 표준오차(1997)

Unit: thousands

Province	SE (employed)	SE (Unemployed)	Province	SE (employed)	SE (Unemployed)
Newfoundland	3	2	Manitoba	4	3
Prince Edward Island	1	1	Saskatchewan	3	2
Nova Scotia	4	3	Alberta	9	6
New Brunswick	3	2	British Columbia	12	9
Quebec	18	14	Canada	32	24
Ontario	20	15			

<표3.18> LFS 고용 및 실업에 대한 설계효과 (1997)

Province	Employed		Unemployed	
	Sample	Overall	Sample	Overall
Newfoundland	2.7	0.83	1.4	1.3
Prince Edward Island	2.0	0.53	1.1	1.1
Nova Scotia	2.2	0.51	1.2	1.1
New Brunswick	2.0	0.56	1.4	1.4
Quebec	2.1	0.55	1.1	1.0
Ontario	3.3	0.50	1.2	1.1
Manitoba	2.2	0.41	1.1	1.1
Saskatchewan	2.4	0.63	1.2	1.2
Alberta	4.1	0.40	1.1	1.1
British Columbia	2.1	0.50	1.2	1.1
Canada	2.8	0.51	1.2	1.1

비표본오차는 표본조사의 매 단계에서 발생할 수 있으며 주로 조사원의 무관심, 오해 및 잘못된 해석 등의 사유에 기인하며, 추정값의 편향 및 변동에 직접적인 영향을 미친다. 관측값의 수가 많거나 혹은 대영역의 조사에서는 비표본오차에 기인한 효과는 무시될 수도 있는 양이나, 소지역 추정의 문제에서는 민감한 문제로 인식되어진다. 비표본 편향 및 분산은 조사원에 대한 교육 및 조사원의 태도, 설문지 설계 상의 문제 또는 무응답을 처리하기 위해 이용되는 대체방법 등에서 발생될 수 있으며, 여기에서는 LFS에서의 적용범위 오차(coverage error), 무응답 오차, 빈집 오차(vacancy errror), 응답 오차(response error), 처리과정 오차(processing error)등에 관해서 설명하기로 한다.

적용범위 오차는 표본추출틀의 조사단위들이 목표모집단을 제대로 반영하지 못할 경우 발생할 수 있다. 조사단위들이 표본추출틀에서 누락되어 있는 경우, 목표모집단에 속하지 않은 단위들이 표본추출틀에 포함되어 있는 경우 또는 조사단위들이 표본추출틀에 중복되어 있는 경우 등이 적용범위 오차를 유발할 수 있는 일반적인 유형들이며, 이 중 조사단위들이 표본추출틀에서 누락되어 있는 경우가 LFS에서 가장 빈번히 발생하는 유형이라 할 수 있다. 나머지 유형의 문제는 LFS에서는 거의 무시된다. LFS에서는 적용범위 오차를 측정하기 위한 지표로써 손실률(slippage rate)을 이용한다. 손실률은 LFS 인구 추정치와 최근의 센서스 인구 추정치와의 차이에 대한 센서스 인구 추정치의 비율로써 정의된다. LFS에서는 CMA 지역, ER 지역, 주 및 전국 단위와 캐나다 지역의 성별(남, 녀)-연령대별(15-19, 20-24, 25-29, 30-39, 40-54, 55+) 범주에 대한 손실률을 매월 정기적으로 작성하고 있다. 손실률로부터 발생하는 비표본오차에 대한 보정은 추정과정에서 처리하고 있다. LFS 조사에서 평균적인 손실률의 양은 다음 <표3.19>와 같이 주어진다.

<표3.19> LFS 평균 손실률(Average Slippage Rate(%))

Provinces		Average	Provinces	Average
Canada	all	9.3	Nova Scotia	8.6
	15-19세	6.1	New Brunswick	10.4
	20-24세	15.6	Quebec	8.0
	25-29세	16.1	Ontario	9.7
	30-39세	9.8	Manitoba	6.1
	40-54세	8.0	Saskatchewan	10.7
	55이상	6.9	Alberta	7.4
Newfoundland		9.8	British Coumbia	12.4
Prince Edward Island		11.6		

조사가구에 대한 무응답 발생요인으로써 가구 구성원의 부재, 가구 구성원의 비 정상적인 거주 환경, 인터뷰 거절 등의 요인을 들 수 있다. 여기에서 인터뷰 거절의 비율은 매월 조사에서 1~2% 정도로 매우 낮게 나타나며, 주 지역에 대해서도 월 조사와 비슷한 비율을 보이나 높게는 약 3% 정도까지 나타난다. 단위 무응답에 대해서는 바로 전 달의 정보를 이용할 수 있다면 이를 이용하여 대체되며, 항목 무응답에 대해서는 표본 대체법이 이용된다. 가구 구성원이 거주하고 있지 않는 결측 가구 및 건물 철거 등으로 인한 비 존재 가구들에 대해서 발생하는 무응답은 편향에 영향을 미치지는 않지만 표본 분산에는 영향을 미치게 되므로 LFS에서는 이러한 유형의 오차 정보를 파악하기 위한 VC(vacancy check) 프로그램을 운영하고 있다. 다음 <표3.20>은 1997년 LFS 조사에서 발생한 평균 무응답률을 나타낸다.

<표3.20> LFS 평균 무응답률(1997)

Provinces	Average(%)	Maximum(%)	Minimum(%)
Newfoundland	4.2	5.4	3.0
Prince Edward Island	3.5	4.8	2.4
Nova Scotia	6.3	7.3	4.6
New Brunswick	4.6	5.4	3.1
Quebec	5.4	6.6	3.7
Ontario	4.8	5.7	3.7
Manitoba	3.6	5.4	2.1
Saskatchewan	3.6	4.6	2.4
Alberta	4.9	6.3	3.1
British Columbia	5.7	6.7	4.5
Canada	4.9	5.5	3.8

LFS에서 결측가구(dwelling vacant)는 사람이 거주하고 있지 않는 가구, 계절 가구 또는 공사 중인 가구로 정의되어 분류된다. 철거 또는 조사가구가 상점 등으로 용도가 변경된 경우는 비 존재 가구(dwelling non-existence)로 분류된다. 결측가구로 확인된 가구들은 LFS 추정치의 편향에 영향을 미치지는 않지만 표본 조사단위가 줄어들므로 추정 분산은 커지게 된다. 결측가구들은 새로운 입주자들이 상주할 가능성이 항상 존재하므로 매월 조사 대상에 포함된다. 그러나 LFS 조사에서 비 존재 가구로 확인된 조사가구들은 표본추출틀에서 일제히 삭제된다. 1997년 LFS 조사에서 발생한 평균적인 결측가구에 대한 비율이 다음 <표3.21>에 주어졌다.

<표3.21> LFS 평균 결측률(vacant rate)(1997)

Provinces	Average(%)	Maximum(%)	Minimum(%)
Newfoundland	15.4	14.9	16.4
Prince Edward Island	20.5	18.6	23.0
Nova Scotia	16.8	15.2	18.7
New Brunswick	14.1	13.5	15.2
Quebec	14.0	11.9	15.8
Ontario	10.8	10.0	11.3
Manitoba	17.1	16.4	17.7
Saskatchewan	14.7	12.5	15.5
Alberta	8.7	8.1	9.8
British Columbia	9.5	8.7	9.8
Canada	13.0	12.2	13.5

응답 오차(response error)는 설문지 설계, 문항 구성, 응답자의 인지력, 인터뷰 방식, 조사가 수행되는 상황 및 조사정보가 수집되고 집계되는 과정 등에 기인할 수 있다. 조사정보가 수집되고 집계되는 과정에서 발생하는 응답오차는 CAI 시스템에 의해 어느 정도 보완되었다고 볼 수 있다.

처리과정 오차(processing error)는 자료 획득, 편집, 코딩, 가중치를 산출하는 과정 및 목록화 작업 등의 매 단계에서 발생할 수 있다. LFS 조사에서는 이러한 매 단계의 처리과정을 전산화 작업으로 통합하여 자료처리과정에서 발생하는 오차를 최소화하고 있다. 전산화 통합 모드는 1993년부터 채택되어 시행되고 있으며 앞서 언급되었던 CAI 모드는 조사행정에서 조사원들의 조사과정을 보조하는 일종의 컴퓨터 보조관리 시스템을 말한다.

3.2.7 소지역 추정법

캐나다 노동력 조사에서는 표본설계 단계에서 EIER 지역과 CMA 지역과 같은 소지역 단위에 대한 추정을 고려하여 층화, 표본추출, 표본 배정 등이 이루어진다. 소지역 통계 추정을 위한 추정량은 크게 설계 기반 추정량(design-based estimator), 간접추정량(indirect estimator), 모형 기반 추정량(model-based estimator)이 이용된다. 소지역 통계 작성 시 설계 기반 추정량이 목표 요구정도를 만족한다면 우선적으로 설계 기반 추정량을 이용하며 그렇지 못할 경우에는 추정량의 신뢰도를 확보할 수 있는 다른 추정방법을 이용한다.

(1) 설계 기반 추정량(Design-Based Estimator)

일반적으로 설계 기반 추정량은 직접추정량(direct estimator)과 수정된 직접추정량(modified direct estimator)으로 구분된다. 관심변수와 밀접한 관련이 있는 보조정보가 있는 경우에 이를 이용하는 사후층화추정량(post stratified estimator), 비추정량(ratio estimator), 회귀추정량(regression estimator) 등은 직접추정량의 일종이다. 직접추정량은 편향이 없는 추정량이지만 해당 소지역에 배정된 표본의 크기가 작은 경우에는 추정량의 분산이 커져서 신뢰성이 떨어지게 된다. 한편 수정된 직접추정량(modified direct estimator)은 해당 소지역 이외의 다른 지역의 조사결과를 추정과정에 추가적으로 이용하며 추정량의 불편성은 근사적으로 유지된다.

직접추정량(direct estimator)은 보통 해당 소지역에서 조사된 자료만을 이용하여 추정되며, 간혹 센서스나 행정자료로부터 획득된 보조정보를 조사자료에 추가하여 추정되기도 한다. 가장 간단한 총계추정에 대한 직접추정량으로써 다음과 같은 단순추정량(expansion estimator)이 이용된다.

$$\hat{Y}_{c,a} = \sum_{i \in s_a} w_i y_i \tag{3.18}$$

여기에서 s_a는 소지역 a의 표본들의 집합, w_i는 조사단위 i에 대한 가중치를 나타낸다. 위의 직접추정량은 불편추정량이나 소지역 a의 표본크기가 작을 경우에는 분산이 커지기 때문에 신뢰성에 문제가 있을 수 있다.

소지역 a의 모집단의 크기 N_a를 알고 있을 경우에는 다음과 같은 사후층화추정량이 이용된다.

$$\begin{aligned}\hat{Y}_{pst,a} &= N_a \frac{\sum_{i \in s_a} w_i y_i}{\sum_{i \in s_a} w_i} \\ &= N_a \frac{\hat{Y}_{c,a}}{\hat{N}_{c,a}} \\ &= N_a \bar{y}_{c,a}\end{aligned} \tag{3.19}$$

위의 사후층화추정량은 먼저 언급된 단순추정량보다는 안정적이나 보다 복잡한 조사에서는 비추정편향(ratio estimation bias)이 발생할 가능성이 있다.

표본이 층화추출되고 층 h에서 소지역 a의 모집단의 크기 $N_{h,a}$가 알려져 있을 경우에는 다음과 같은 유형의 사후층화추정량이 소지역 추정에 이용된다.

$$\begin{aligned}\hat{Y}_{st,pst,a} &= \sum_h \left(N_{h,a} \frac{\sum_{i \in s_{h,a}} w_i y_i}{\sum_{i \in s_{h,a}} w_i} \right) \\ &= \sum_h N_{h,a} \frac{\hat{Y}_{h,c,a}}{\hat{N}_{h,c,a}} \\ &= \sum_h N_{h,a} \bar{y}_{h,a}\end{aligned} \tag{3.20}$$

여기에서 층 h는 표본설계 시 반영된 층이라기 보다는 사후층화에 의해 형성된 층을 말한다.

비추정법(ratio estimation)은 사후층화추정법과 유사하나 모집단 총계 N_a와 $N_{h,a}$대신에 보조정보에 의해 획득된 소지역 총계 X_a와 $X_{h,a}$를 이용하며, 이 값들을 알고 있을 경우 비추정량은 다음과 같이 정의된다.

$$\hat{Y}_{r,a} = X_a \hat{R}_a \ , \quad \hat{Y}_{st,r,a} = \sum_h X_{h,a} \hat{R}_{h,a} \tag{3.21}$$

여기에서 $\hat{R}_a = \hat{Y}_{c,a}/\hat{X}_{c,a}$는 Y_a/X_a의 추정값, $\hat{R}_{h,a} = \hat{Y}_{h,c,a}/\hat{X}_{h,c,a}$를 나타낸다.

회귀추정법(regression estimation)이 소지역 총계 추정에 이용되기도 한다. 이 방법은 관심변수 y와 공변량 x사이의 관계에서 회귀모수를 추정하여 소지역 총계 추정에 이용하는 방법으로써 추정량은 다음과 같은 형태로 주어진다.

$$\hat{Y}_{reg,a} = \hat{Y}_a + \hat{\beta}_a (X_a - \hat{X}_a) \tag{3.22}$$

여기에서 $\hat{Y}_a$는 직접추정 또는 사후층화추정법에 의해 추정된 소지역 a에 대한 총계 추정값이며 $\hat{X}_a$는 보조정보를 통해 $\hat{Y}_a$와 유사한 방법으로 추정된다. 추정모수 $\hat{\beta}_a$는 관심변수 y와 공변량 x의 관계로부터 추정되며

$$\hat{\beta}_a = \sum_{i \in s_a} v_i^{-1} w_i y_i x_i^t \left(\sum_{i \in s_a} v_i^{-1} w_i x_i x_i^t \right)^{-1}$$

과 같이 주어진다. 여기에서 v_i는 회귀가중치로써 주어지는 값이며, $v_i = x_i$이고 x가 상수일 경우에는 $\hat{\beta}_a = \hat{R}_a$인 관계가 성립한다. 회귀추정량의 불편성은 $\hat{Y}_a$와 $\hat{X}_a$의 불편성에 의존한다.

한편, 회귀추정량을 변형한 일종의 수정된 직접추정량(modified direct estimator)이 소지역 특성치 추정에 이용되기도 한다. 수정된 직접추정량은 해당 지역 외의 조사자료를 특성치 추정에 이용하며, 추정량의 불편성

은 회귀추정량과 마찬가지로 근사적으로 만족된다. 예를 들면 식(3.22)에서 추정 회귀모수 $\hat{\beta}_a$대신에 회귀모수에 대한 합성추정량의 일종인

$$\hat{\beta} = \sum_{i \in s} v_i^{-1} w_i y_i x_i^t \left(\sum_{i \in s} v_i^{-1} w_i x_i x_i^t \right)^{-1}$$

이 이용되었다면 이러한 추정량을 수정된 직접추정량(modified direct estimator)이라 부른다. 일반적으로 소지역 추정 시 $\hat{\beta}$이 $\hat{\beta}_a$보다 안정적인 것으로 알려져 있으며, $\hat{\beta}$과 $\hat{\beta}_a$의 가중평균 $\lambda_a \hat{\beta}_a + (1-\lambda_a)\hat{\beta}$이 추정 회귀모수로 이용되기도 한다. 여기에서 λ_a는 적절히 선택되는 값이다. x가 상수이고 $v_i = x_i$인 경우에는 $\hat{\beta}$대신 $\hat{R} = \hat{Y}_a / \hat{X}_c$이 이용될 수도 있다.

(2) 간접추정량(Indirect Estimator)

간접추정량은 합성추정량(synthetic estimator), 복합추정량(composite estimator), 표본수 의존 복합추정량(sample size dependent estimator) 등의 유형으로 구분되며, 해당 지역의 조사자료 뿐만 아니라 해당 지역을 포함하고 있는 더 큰 지역의 조사자료를 소지역 추정과정에 이용하여 소지역 추정의 신뢰성을 확보하는 방법이다.

합성추정법(synthetic estimation)은 소지역 추정 시 소지역을 포함하는 대영역의 정보를 함께 이용하는 방법으로써 소지역과 대영역의 특성구조가 유사하다는 가정 하에서 이용된다. 합성추정량의 분산은 직접추정량의 분산에 비해 작으나 전제한 가정이 성립하지 않을 경우에는 심각한 편향이 발생할 수 있다.

소지역의 특성치 평균이 전체 지역의 특성치 평균과 같다는 가정 하에서 이용되는 가장 간단한 형태의 합성추정량은 다음과 같다.

$$\hat{Y}_{syn,m,a} = N_a \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i} = N_a \bar{y} \qquad (3.23)$$

층화 또는 사후층화에 근거한 합성추정량은 보통 다음과 같은 형태로 주어진다.

$$\hat{Y}_{syn,st,m,a} = \sum_h N_{h,a} \frac{\sum_{i \in s_h} w_i y_i}{\sum_{i \in s_h} w_i} = \sum_h N_{h,a} \bar{y}_h \qquad (3.24)$$

직접추정법에서와 마찬가지로 합성추정법에서도 비합성추정법(ration synthetic estimation)이 고려될 수 있다. 비합성추정법은 모집단의 크기 N_a또는 $N_{h,a}$외에 소지역 추정을 위한 보조정보로써 공변량 x를 이용하며 추정량(ratio synthetic estimator)은 다음과 같은 형태로 정의된다.

$$\hat{Y}_{syn,r,a} = X_a \frac{\hat{Y}_e}{\hat{X}_e}, \quad \hat{Y}_{syn,st,r,a} = \sum_h X_{h,a} \frac{\hat{Y}_{h,e}}{\hat{X}_{h,e}} \qquad (3.25)$$

여기에서 $\hat{Y}_e = \sum_{i \in s} w_i y_i$는 y에 대한 모집단 총계 추정량이고, $\hat{Y}_{h,e} = \sum_{i \in s_h} w_i y_i$는 층의 총계 추정치를 나타낸다. 기타 비합성추정량들에 대한 내용은 Gonzalez(1973), Gonzalez and Waksberg(1973), Ghangurde and Singh(1977, 1978)에 자세히 소개되어있다. 한편, Singh and Tessier(1976)는 (3.25)식의 $\hat{Y}_{syn,r,a}$에서 $\hat{X}_e$대신에 X를 이용한 비합성추정량의 대체식 $\bar{Y}_{syn,r,a} = X_a \frac{\hat{Y}_e}{X}$를 제안하였다. 여기에서 $\hat{Y}_{syn,r,a}$와 $\bar{Y}_{syn,r,a}$는 모두 같은 양의 편향을 가지며, $\hat{Y}_{syn,r,a}$의 편향은 표본의 크기가 클 경우에는 무시될 수 있다. 두 추정량 중 하나의 추정량을 선택하는 문제에서는 $\hat{Y}_e$와 $\hat{X}_e$의 상관계수 ρ를 참조하도록 하였다. 일반적으로 표본의 크기가 클 경우, 두 추정량의 분산은 상관계수의 값이 $\rho \geq 0.5 c_x/c_y$ 일 때 $V(\hat{Y}_{syn,r,a}) \leq V(\bar{Y}_{syn,r,a})$인 관계가 성립한다. 여기에서 c_x와 c_y

는 각각 $\hat{X}_a$와 $\hat{Y}_a$의 변동계수(coefficient of variation)를 나타낸다. 상관계수 ρ의 값이 크거나 모집단의 분포가 한쪽으로 치우쳐져 있을 경우에는 $\hat{Y}_{syn,r,a}$가 선호되며, 변동계수 c_x의 값이 크거나 상관계수 ρ의 값이 적당할 경우에는 보통 $\overline{Y}_{syn,r,a}$를 선택한다.

소지역의 보조정보로써 이용된 공변량 x외에 추가적인 보조변수 z를 도입하여 소지역 특성치를 추정하는 다음과 같은 이변량 비합성추정량이 소지역 추정에 이용된다.

$$\hat{Y}^{(2)}_{syn,r,a} = \gamma_a X_a \frac{\hat{Y}_a}{\hat{X}_a} + (1-\gamma_a) Z_a \frac{\hat{Y}_a}{\hat{Z}_a} \tag{3.26}$$

여기에서 γ_a는 적절히 선택되는 값이다. 보다 일반적인 다변량 비합성추정량은 Olkin(1958)에서 참조할 수 있다.

회귀합성추정법은 비합성추정법과 유사하며 추정량은 다음과 같이 주어진다.

$$\hat{Y}_{syn,reg,a} = \hat{\beta} X_a \ ,$$

$$\hat{\beta} = \sum_{i \in s} v_i^{-1} w_i y_i x_i^t \left(\sum_{i \in s} v_i^{-1} w_i x_i x_i^t \right)^{-1} \tag{3.27}$$

회귀합성추정법은 표본설계의 층 내에서 또는 사후층화에 의해 형성된 층 내에서도 응용이 가능하며, Royall(1979)은 이러한 내용을 반영한 다음과 같은 회귀합성추정량을 제안하였다.

$$\hat{Y}_{syn,Ray,a} = \sum_{i \in s_a} y_i + \hat{\beta}(X_a - \sum_{i \in s_a} x_i) \tag{3.28}$$

복합추정량(composite estimator)은 직접추정량의 불안정성과 합성추정량의 잠재적 편향 가능성을 보완하기 위해 두 추정량의 가중평균을 취하며 일반적인 형태는 다음과 같다.

$$\hat{Y}_{com,a} = \lambda_a \hat{Y}_{dir,a} + (1-\lambda_a) \hat{Y}_{syn,a} \tag{3.29}$$

여기에서 가중치 λ_a는 적절히 선택되는 값이다.

가중치 λ_a를 결정하는 방법은 크게 세 가지 정도로 구분될 수 있다. 첫 번째 방법은 가장 간단한 방법으로써 가중치 λ_a를 고정계수로 두는 방법인데 추정량의 신뢰성에 문제가 있어 많이 사용되지는 않는다. 두 번째 방법은 추정하고자하는 소지역의 표본크기를 반영하는 방법이다. 이 경우 가중치 λ_a는 $\hat{N}_{c,a}/N_a$의 함수로 표현된다. Drew et al.(1982)은 표본크기에 의존하는 복합추정량으로써 다음과 같은 추정량을 제안하였다.

$$\hat{Y}_{ssd,r,a} = \lambda_a \hat{Y}_{r,a} + (1-\lambda_a)\hat{Y}_{syn,r,a} \qquad (3.30)$$

여기에서

$$\lambda_a = \begin{cases} 1 & \text{if } \hat{N}_{c,a} \geqq \delta N_a \\ \dfrac{\hat{N}_{c,a}}{\delta N_a} & otherwise \end{cases}$$

이며, δ는 합성추정량 부분의 편향을 보정하기 위해 주관적으로 결정되는 값이다. 캐나다 노동력조사에서는 $\delta = 2/3$를 이용한다. 식 (3.30)의 복합추정량은 직접추정량의 신뢰도가 완전히 확보된 지역에 대해서는 합성추정량의 가중치가 0이 되기 때문에 이러한 지역의 경우 직접추정값이 곧 복합추정값으로 선택된다고 볼 수 있다. 그렇지 않은 기타 지역에 대해서는 직접추정값과 합성추정값의 가중평균값으로 복합추정값이 계산된다. 캐나다 노동력조사에서 이러한 기타 지역들에 대한 합성추정량의 평균 가중치는 약 10% 정도이며 많아야 20%를 초과하지는 않는다. 이때 δ의 값은 [2/3 , 3/2]의 범위에 있는 것으로 알려져 있다. 이외의 표본크기 의존 복합추정량으로써 Sandal(1984)의 추정량

$$\hat{Y}_{ssd,reg,a} = \lambda_a \hat{Y}_{sreg,a} + (1-\lambda_a)\hat{Y}_{syn,reg,a}$$

를 들 수 있다. 사용된 가중치는 $\lambda_a = \hat{N}_{c,a}/N_a$이다. Rao(1986)는 위와 동일한 추정량에 대해 가중치를 약간 달리 적용할 것을 제안하였다. Rao의

가중치는 $\hat{N}_{e,a} \geq N_a$인 지역에 대해서는 $\lambda_a = 1$, 기타 지역에 대해서는 Sandal의 가중치와 동일하다. Sandal and Hidiroglou(1989)는 Rao의 가중치에서 $\hat{N}_{e,a} < N_a$일 때 $\lambda_a = (\hat{N}_{e,a}/N_a)^{h-1}$을 사용할 것을 제안하였다. 여기에서 h는 합성추정량의 편향을 감안하여 적절히 선택되는 값이다. 가중치를 결정하는 세 번째 방법은 직접추정량과 합성추정량의 평균제곱오차와 두 추정량의 공분산을 자료로부터 추정하여 최적가중치를 산정하는 방법이다. 복합추정량의 평균제곱오차는 다음 식과 같이 나타낼 수 있다.

$$MSE(\hat{Y}_{com,a}) = \lambda_a^2 MSE(\hat{Y}_{dir,a}) + (1-\lambda_a)^2 MSE(\hat{Y}_{syn,a}) + 2\lambda_a(1-\lambda_a)E(\hat{Y}_{dir,a} - Y_a)(\hat{Y}_{syn,a} - Y_a) \quad (3.31)$$

(3.31)식에서 $\hat{Y}_{com,a}$의 MSE를 최소화하는 가중치 λ_a는 다음 식과 같이 주어질 수 있다.

$$\hat{\lambda}_a = \frac{\widehat{MSE}(\hat{Y}_{syn,a}) - E(\hat{Y}_{syn,a} - Y_a)(\hat{Y}_{dir,a} - Y_a)}{\widehat{MSE}(\hat{Y}_{syn,a}) + \widehat{MSE}(\hat{Y}_{dir,a}) - 2E(\hat{Y}_{syn,a} - Y_a)(\hat{Y}_{dir,a} - Y_a)} \quad (3.32)$$

식 (3.32)에서 $\hat{Y}_{dir,a}$와 $\hat{Y}_{syn,a}$의 공분산의 항이 $\widehat{MSE}(\hat{Y}_{syn,a})$와 $\widehat{MSE}(\hat{Y}_{dir,a})$에 비해 매우 작다고 가정할 수 있다면 다음과 같은 근사적인 가중치를 이용할 수도 있다.

$$\hat{\lambda}_a^* = \frac{\widehat{MSE}(\hat{Y}_{syn,a})}{\widehat{MSE}(\hat{Y}_{syn,a}) + \widehat{MSE}(\hat{Y}_{dir,a})} \quad (3.33)$$

(3) 모형 기반 추정량(Model-Based Estimator)

소지역 추정에 자주 이용되는 모형 기반 추정법(model-based estimation)으로는 $EBLUP$(empirical best linear unbiased prediction), EB(empirical Bayes), HB(hierarchical Bayes) 접근법 등이 있다. 최근에는 소지역 통계 작성을 위해 횡단면 조사자료(cross-sectional data)와 시

계열자료(time series data)를 함께 추정과정에 이용하는 방법에 관한 연구가 활발히 진행되고 있다. 이 절에서는 횡단면 조사자료를 이용한 모형기반 추정량과 횡단면 조사자료와 시계열자료를 함께 이용하는 모형기반 추정량에 대한 내용들을 소개한다.

y_i를 i번째 소지역의 관심모수 θ_i에 대한 직접추정량, x_i를 모수 θ_i의 추정에 필요한 설명변수이고, 모형 $y_i = \theta_i + e_i$, $E(e_i) = 0$를 가정할 때, 소지역 i에 대한 다음과 같은 선형회귀모형(linear regression model)을 고려할 수 있다.

$$\theta_i = \beta_0 + \beta_1 x_i \ , \quad i = 1, 2, \cdots, I \tag{3.34}$$

여기에서 β_0와 β_1은 회귀모수를 나타낸다. 이때 θ_i에 대한 회귀합성추정량(regression synthetic estimator)은 다음과 같이 주어질 수 있다.

$$\bar{\theta}_{i(reg)} = \bar{\beta}_0 + \bar{\beta}_1 x_i \tag{3.35}$$

여기에서 $\bar{\beta}_0$와 $\bar{\beta}_1$은 결합모형 $y_i = \beta_0 + \beta_1 x_i + e_i$로부터 계산되는 최소제곱추정량을 나타낸다. 조사 추정량 y_i들의 공분산을 추정할 수 있을 경우에는 일반화 가중최소제곱추정량을 이용할 수도 있다. 위의 회귀합성추정량은 조사 추정량 y_i들에 대한 가중치가 반영되지 않기 때문에 큰 편향이 발생할 수 있다. 반면, EB추정량이나 $EBLUP$추정량은 적당한 가중치가 부여되어 편향 발생이 다소 억제되는 결과를 얻을 수 있다.

이러한 편향에 대한 문제점을 해결하기 위해 Fay and Herriot(1979)는 모형 (3.34)를 다음과 같이 해당 소지역에 대한 랜덤효과 v_i를 갖는 모형으로 보완하였다.

$$\theta_i = \beta_0 + \beta_1 x_i + v_i \tag{3.36}$$

여기에서 v_i는 평균이 0 이고 분산이 σ_v^2을 갖는 서로 독립인 정규분포를 따르는 확률변수, e_i는 평균이 0 이고 분산이 σ_i^2인 서로 독립인 정규 확률변수를 나타내며, σ_i^2은 기지인 값으로 가정된다. 이때 결합모형은 다음

과 주어진다.

$$y_i = \beta_0 + \beta_1 x_i + v_i + e_i \tag{3.37}$$

위의 모형으로부터 θ_i의 EB 추정량은 직접조사추정량 y_i와 회귀합성 추정량 $\hat{\theta}_{i(reg)} = \hat{\beta}_0 + \hat{\beta}_1 x_i$의 가중합으로 표현되며 다음 식과 같이 주어진다.

$$t_i(\hat{\sigma}_v^2, y) = w_i y_i + (1 - w_i)\hat{\theta}_{i(reg)} \tag{3.38}$$

여기에서 $w_i = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \sigma_i^2)$, $\hat{\beta}_0$와 $\hat{\beta}_1$은 결합모형 하에서 추정된 가중 최소제곱추정량을 나타내며, $\hat{\sigma}_v^2$은 σ_v^2의 적률추정량 또는 최대우도추정량 등이 이용될 수 있다. Fay and Herriot(1979)는 1970년 미국의 인구주택 총조사 자료로부터 인구 1000 미만의 소지역에 대한 소득관련 추정에 식 (3.38)의 EB추정량을 이용하였고, EB추정량이 직접 조사 추정량이나 합성추정량에 비해 표본오차가 작다는 사실을 수치적으로 제시하였다.

횡단면자료를 이용한 소지역 추정방법은 조사 시기가 상이한 조사자료들의 정보를 모형에 반영시키는 것은 사실상 어렵다. Scott et al.(1977), Jones(1980), Tiller(1989) 등은 이러한 단점을 보완하기 위해 반복적인 월별 조사자료들의 정보와 센서스 및 행정자료를 모형에 포함시킨 횡단면 시계열 모형들을 소지역 추정 문제에 도입하였다.

θ_{it} , y_{it}와 x_{it}를 각각 조사시기 t에서 소지역 i에 대한 모평균, 직접조사추정값, i번째 소지역과 관계가 있는 연관변수라 할 때, 우선 다음과 같은 모형을 고려한다.

$$y_{it} = \theta_{it} + e_{it}\ , \quad i = 1, 2, \cdots, I\ ;\ t = 1, 2, \cdots, T \tag{3.39}$$

여기에서 표본오차 e_{it}의 평균은 0, 분산공분산행렬은 기지인 블럭대각행렬 Σ_i ($T \times T$ 행렬)로 가정한다. 모평균 θ_{it}에 관한 모형은 다양한 유형으로 설정될 수 있으며 다음과 같은 모형들이 고려될 수 있다.

(I) $\theta_{it} = \beta_0 + \beta_1 x_{it} + v_i + \epsilon_{it}$, $\quad i = 1, 2, \cdots, I\ ;\ t = 1, 2, \cdots, T$,

여기에서 v_i는 소지역 고정효과, $\epsilon_{it} \overset{ind}{\sim} N(0, \sigma^2)$.

(Ⅱ) $\theta_{it} = \beta_0 + \beta_1 x_{it} + v_i + \epsilon_{it}$, $i = 1, 2, \cdots, I$; $t = 1, 2, \cdots, T$,

여기에서 $v_i \overset{ind}{\sim} N(0, \sigma_v^2)$, $\epsilon_{it} \overset{ind}{\sim} N(0, \sigma^2)$, $\{v_i\}$와 $\{\epsilon_{it}\}$는 서로 독립이며, 모형(Ⅰ)과는 달리 v_i들이 랜덤효과로 가정되었다.

(Ⅲ) $\theta_{it} = \beta_0 + \beta_1 x_{it} + v_i + u_t + \epsilon_{it}$, $i = 1, 2, \cdots, I$; $t = 1, 2, \cdots, T$,

여기에서 $v_i \overset{ind}{\sim} N(0, \sigma_v^2)$, $u_t \overset{ind}{\sim} N(0, \sigma_u^2)$, $\epsilon_{it} \overset{ind}{\sim} N(0, \sigma^2)$이고, $\{v_i\}$, $\{u_t\}$와 $\{\epsilon_{it}\}$는 서로 독립이다. v_i는 소지역에 대한 랜덤효과, u_t는 조사시기에 대한 랜덤효과이다.

(Ⅳ) $\theta_{it} = \beta_0 + \beta_1 x_{it} + v_i + u_{it}$, $u_{it} = \rho u_{i,t-1} + \epsilon_{it}$, $|\rho| < 1$,

여기에서 $v_i \overset{ind}{\sim} N(0, \sigma_v^2)$, $\epsilon_{it} \overset{ind}{\sim} N(0, \sigma^2)$이고, $\{v_i\}$와 $\{\epsilon_{it}\}$는 서로 독립이며, $\{u_{it}\}$는 $AR(1)$과정을 따른다. 모형 (Ⅳ)는 다음과 같이 시차모형(lag model)으로 재표현 가능하다.

$$\theta_{it} = \rho\theta_{i,t-1} + (1-\rho)\beta_0 + \beta_1 x_{it} - \beta_1 \rho x_{i,t-1} + (1-\rho)v_i + \epsilon_{it} \quad (3.40)$$

모형 (Ⅳ)는 모평균 θ_{it}와 보조변수 x_{it}가 이전 조사시기에서의 값들을 모형에 반영시키기 때문에 위의 네가지 모형들 중에서는 가장 현실적인 추정 모형이라 할 수 있다. 따라서 모형 (Ⅳ)를 이용한 결합모형을 고려한다면 다음과 같이 주어진다.

$$y_{it} = \beta_0 + \beta_1 x_{it} + v_i + (e_{it} + u_{it}) \quad (3.41)$$
$$u_{it} = \rho u_{i,t-1} + \epsilon_{it} , \quad |\rho| < 1$$

여기에서 $v_i \overset{ind}{\sim} N(0, \sigma_v^2)$, $\epsilon_{it} \overset{ind}{\sim} N(0, \sigma^2)$, e_{it}는 평균이 0이고, 기지인 블럭대각공분산행렬 $\Sigma = diag(\Sigma_1, \Sigma_2, \cdots, \Sigma_I)$를 갖는다. 캐나다 노동력조사에서는 표본 공분산행렬 Σ를 이용할 수 없기 때문에 Tiller(1989)는 복합오차 $w_{it} = e_{it} + u_{it}$를 $AR(1)$과정으로 취급한 후 다음과 같은 결합모형을 이용하여 소지역 추정을 시도하였다.

$$y_{it} = \beta_0 + \beta_1 x_{it} + v_i + w_{it} \ , \tag{3.42}$$

$$w_{it} = \rho w_{i,t-1} + \epsilon_{it} \ , \quad |\rho| < 1$$

여기에서 $\theta_{it} = \beta_0 + \beta_1 x_{it} + v_i$, $v_i \overset{ind}{\sim} N(0, \sigma_v^2)$, $\epsilon_{it} \overset{ind}{\sim} N(0, \sigma^2)$.

$\{y_{it}\}$를

$$\begin{aligned} y &= (y_{11}, y_{12}, \cdots, y_{1T}; \cdots ; y_{I1}, y_{I2}, \cdots, y_{IT})^t \\ &= (y_1^t, y_2^t, \cdots, y_I^t)^t \end{aligned}$$

로 표현하면 위의 모형 (3.42)는 다음과 같은 일반적인 혼합모형(mixed model)의 일종으로 볼 수 있다.

$$y = X\beta + Zv + w \ , \ v \sim (0, \sigma_v^2 I), \ w \sim (0, \sigma^2 (I \otimes \Gamma)) \tag{3.43}$$

여기에서

$$X^t = (X_1^t, X_2^t, \cdots, X_I^t) \ ,$$

$$Z = I \otimes 1_T \ ,$$

$$\beta = (\beta_0, \beta_1)^t$$

이고 X_i는 t번째 행이 $(1, x_{it})$로 주어지는 $T \times 2$행렬, I는 크기 I인 항등행렬, 1_T는 1을 원소로 갖는 $t-$벡터, Γ는 (i,j)번째 원소가

$$\gamma_{ij} = (1-\rho^2)^{-1} \rho^{|i-j|}$$

인 $T \times T$행렬을 나타낸다.

β와 v의 일차결합 $\tau = k^t\beta + m^t v$에 대한 $\bar{\tau}(=\bar{\theta}_{it})$의 $BLUP$와 $BLUP$의 MSE는 다음과 같이 주어진다(Henderson, 1975).

$$\bar{\tau} = k^t\bar{\beta} + m^t Z^t \Sigma^{-1}(y - X\bar{\beta})(\sigma_v^2/\sigma^2) \tag{3.44}$$

$$MSE(\bar{\theta}_{it}) = \sigma^2\{k^t(X^t\Sigma^{-1}X)^{-1}k + \frac{\sigma_v^2}{\sigma^2}m^t m - \left(\frac{\sigma_v^2}{\sigma^2}\right)^2 m^t Z^t \Sigma^{-1} AZm$$

$$-2\frac{\sigma_v^2}{\sigma^2}k^t(X^t\Sigma^{-1}X)^{-1}X^t\Sigma^{-1}Zm\} \tag{3.45}$$

여기에서

$$\Sigma = I \otimes \{(\sigma_v^2/\sigma^2)J + \Gamma\},$$

$$\bar{\beta} = (X^t\Sigma^{-1}X)^{-1}(X^t\Sigma^{-1}y),$$

$$A = I - X(X^t\Sigma^{-1}X)^{-1}X^t\Sigma^{-1}$$

이고, J는 원소들이 1로 구성된 $T \times T$행렬을 나타낸다.

식 (3.44)를 이용하여 결합모형 (3.42)의 $\theta_{it}(=\tau)$에 대한 $BLUP$추정량을 계산하면 주요 항들은 다음과 같이 주어질 수 있다.

$$k^t = (1, x_{it}),$$

$$m^t = (0, \cdots, 0, 1, 0, \cdots, 0)_i,$$

$$m^t Z^t \Sigma^{-1}(y - X\bar{\beta}) = 1_T^t\{(\sigma_v^2/\sigma^2)J + \Gamma\}^{-1}(y_i - X_i\bar{\beta})$$

식 (3.44)의 $BLUP$는 미지인 분산비(variance ratio) σ_v^2/σ^2과 자기상관계수(autocorrelation) ρ에 의존하므로 이들 값들은 추정되어야 한다. 모수들에 대한 $BLUP$는 다음과 같이 주어진다(Pantula and Pollack, 1985).

$$\hat{\rho} = \left\{\sum_{i=1}^{I}\sum_{t=1}^{T-2}\hat{e}_{it}(\hat{e}_{i,t+1} - \hat{e}_{i,t+2})\right\}\left\{\sum_{i=1}^{I}\sum_{t=1}^{T-2}\hat{e}_{it}(\hat{e}_{it} - \hat{e}_{i,t+1})\right\}^{-1} \tag{3.46}$$

$\hat{\sigma}_v^2$과 $\hat{\sigma}^2$은 다음 (3.47), (3.48), (3.49), (3.50)식의 정의로부터 유도될 수

있다.

$$z_{it}^{(1)} = z_{it} - z_{it}^{(2)}. \tag{3.47}$$

여기에서

$$z_{it} = \begin{cases} y_{it} - \hat{\rho} y_{i,t-1} & , t \geq 2 \\ f_1 y_{it} & , t = 1 \end{cases},$$

$$z_{it}^{(2)} = c^{-1} d_i f_t ,$$

$$c = (1-\hat{\rho})\{T - (T-2)\hat{\rho}\},$$

$$f_t = \begin{cases} 1-\hat{\rho}^2 & , t = 1 \\ 1-\hat{\rho} & , t \geq 2 \end{cases},$$

$$d_i = \sum_{t=1}^{T} f_t z_{it}$$

로 주어진다.

$$h_{0it}^{(1)} = h_{0it} - h_{0it}^{(2)}. \tag{3.48}$$

여기에서

$$h_{0it} = \begin{cases} 1-\hat{\rho} & , t \geq 2 \\ f_1 & , t = 1 \end{cases},$$

$$h_{0it}^{(2)} = c^{-1} d_i f_t ,$$

$$f_t = \begin{cases} 1-\hat{\rho}^2 & , t = 1 \\ 1-\hat{\rho} & , t \geq 2 \end{cases},$$

$$d_i = \sum_{t=1}^{T} f_t h_{0it}$$

로 주어진다.

$$h_{1it}^{(1)} = h_{1it} - h_{1it}^{(2)}. \tag{3.49}$$

여기에서

$$h_{1it} = \begin{cases} x_{it} - \hat{\rho} x_{i,t-1} & , t \geq 2 \\ f_1 x_{it} & , t = 1 \end{cases},$$

$$h_{1it}^{(2)} = c^{-1} d_i f_t ,$$

$$f_t = \begin{cases} 1-\hat{\rho}^2 & , t=1 \\ 1-\hat{\rho} & , t \geqq 2 \end{cases},$$

$$d_i = \sum_{t=1}^{T} f_t h_{1it}$$

로 주어진다.

$$g_i = \sum_{t=1}^{T} f_t z_{it} , \quad f_{0i} = \sum_{t=1}^{T} f_t h_{0it} , \quad f_{1i} = \sum_{t=1}^{T} f_t h_{1it} \tag{3.50}$$

종속변수를 $z_{it}^{(1)}$, 설명변수를 $h_{0it}^{(1)}$와 $h_{1it}^{(1)}$를 갖는 절편항이 없는 회귀식을 적합시켰을 때 얻어지는 잔차제곱합을 $\hat{e}^t\hat{e}$라 하고, 종속변수 g_i, 설명변수 f_{0i}, f_{1i}를 갖는 절편항이 없는 회귀식에서 얻어지는 잔차제곱합을 $\hat{u}^t\hat{u}$라 할때, σ_v^2과 σ^2의 $BLUP$는 최종적으로 다음과 같이 주어진다.

$$\hat{\sigma}_v^2 = c^{-1}(I-2)^{-1}\{\hat{u}^t\hat{u} - \hat{\sigma}^2(I-2)\}, \tag{3.51}$$

$$\hat{\sigma}^2 = \{I(T-1)-2\}^{-1}\hat{e}^t\hat{e}$$

따라서 θ_{it}의 $EBLUP$ $\hat{\theta}_{it}$는 식 (3.44)에서 $\hat{\rho}$, $\hat{\sigma}_v^2$, $\hat{\sigma}^2$을 대체하여 얻을 수 있다.

만약 $p-1\,(\geqq 2)$개의 x변수들이 모형에 포함되어 있다면, θ_{it}의 $EBLUP$ $\hat{\theta}_{it}$는 다음과 같이 추정하면 된다. 우선 y_{it}와 $x_{1it}, \cdots, x_{p-1,it}$의 회귀식으로부터 $\{\hat{e}_{it}\}$를 추정한다.

다음으로 $\{h_{jit}, h_{jit}^{(1)}, h_{jit}^{(2)}; j=0,1,\cdots,p-1\}$을 앞서 언급된 바와 같이 1, x_{1it}, $\cdots$, $x_{p-1,it}$의 원소들로 정의한 후, $z_{it}^{(1)}$과 $h_{0it}^{(1)}$, $h_{1it}^{(1)}$, $\cdots$, $h_{p-1,it}^{(1)}$의 절편항이 없는 회귀식으로부터 $\hat{e}^t\hat{e}$을 추정한다. 같은 방법으로 $f_{ji} = \sum_{t=1}^{T} f_t h_{jit}$ $(j=0,1,\cdots,p-1)$를 계산한 후, g_i와 $f_{0i}, f_{1i}, \cdots, f_{p-1,i}$의

절편항이 없는 회귀식을 적합시켜 $\hat{u}'\hat{u}$를 구한다. 마지막으로 (3.51)식에서 $I(T-1)-2$를 $I(T-1)-p$로, $I-2$를 $I-p$로 대체하여 $BLUP$ $\hat{\sigma}_v^2$, $\hat{\sigma}^2$과 $\hat{\rho}$를 계산하면 $EBLUP$ $\hat{\theta}_{it}$를 얻을 수 있고, (3.45)식으로부터 $EBLUP$ $\hat{\theta}_{it}$의 MSE 값을 계산할 수 있다. 한편, 모형 (3.42) 하에서 조사 추정량 y_{it}의 MSE는 다음과 같이 주어진다.

$$MSE(y_{it}) = E(y_{it}-\theta_{it})^2 = V(w_{it}) = \frac{\sigma^2}{1-\rho^2} \tag{3.52}$$

$MSE(y_{it})$의 추정량은 위의 (3.52)식에서 σ^2, ρ를 각각 $\hat{\sigma}^2$, $\hat{\rho}$로 대체하여 얻을 수 있다.

(4) 소지역 추정량들의 효율

$\hat{Y}_{M,a}(r)$을 소지역 추정법 M을 이용하여 추정된 r번째 반복에서 특성치 Y_a의 몬테카를로 추정값이라 하자. 이 때 n개의 소지역에 대한 MSE 추정값의 평균은 다음 식을 이용하여 계산할 수 있다.

$$Avg\ \widehat{MSE}_M = \frac{1}{n}\sum_a\sum_{r=1}^{R}\frac{(\hat{Y}_{M,a}(r)-Y_a)^2}{R} \tag{3.53}$$

소지역 추정법 M을 이용하여 추정된 추정량들의 효율을 직접추정법 M_0를 이용한 추정량과 비교하여 나타낸다면 상대 효율은 다음 식을 이용하여 구할 수 있다.

$$Eff(M\ vs\ M_0) = \frac{Avg\ \widehat{MSE}_M}{Avg\ \widehat{MSE}_{M_0}} \tag{3.54}$$

여기에서 $Avg\ \widehat{MSE}_{M_0}$는 직접추정법 M_0에 의해 추정된 추정값들에 대한 평균제곱오차의 평균을 나타낸다.

(5) 표본 수 의존 추정(sample size dependent estimation)

현재 캐나다에서 소지역 통계 작성을 위해 사용되고 있는 추정량은 Drew et al.(1982)에 기초한 표본수 의존 추정량(sample size dependent estimator)이다. LFS자료에 근거하여 우선 소지역 a에 대해서 일반화 회귀추정량으로 관심변수의 총계를 추정한다.

$$\hat{Y}_{GREG,a} = \hat{Y}_{c,a} + \hat{\beta}_a (X_a - \hat{X}_{c,a}) \qquad (3.55)$$

여기에서

$$\hat{Y}_{c,a} = \sum_{i \in s_a} w_i y_i \ ,$$

$$\hat{X}_{c,a} = \sum_{i \in s_a} w_i x_i \ ,$$

$$\hat{\beta}_a = \sum_{i \in s_a} w_i y_i x_i^t \left(\sum_{i \in s_a} w_i x_i x_i^t \right)^{-1}$$

이다.

관심변수의 총계를 추정할 수 있으면 관심의 대상인 경제활동인구수, 실업자 수, 취업자 수, 실업률, 취업률 등을 모두 추정할 수 있게 된다. 추정과정에 이용된 가중치 w_i는 설계 가중치와 무응답 가중치 조정이 반영된 것이며, LFS의 최종 가중치를 의미하는 것은 아니다. 일반적으로 ER 지역의 소지역 통계 작성에 사용되는 보조정보는 주(Province) 단위의 노동력 통계 작성에 이용될 수 있는 보조정보에 비해 제한적이라 할 수 있다.

Drew et al.(1982)는 LFS에 알맞은 소지역 추정량을 찾기 위해 여러 종류의 추정량들을 비교하였다. 여기에서 사용된 소지역은 CD(Census Division) 지역이고 각 소지역에 대해 다음과 같은 세 개의 범주에 대한 보조정보를 추정과정에 이용하였다.

(ⅰ) 연령 15~16세, 65세 이상 전체인구

(ⅱ) 연령 17 ~64세 여성인구

(ⅲ) 나이 17~64세 남성인구 .

소지역 a를 포함하는 주(Province) 단위에서 $\hat{\beta}$을 구하여 합성추정량 $\hat{Y}_{SYN,GREG,a} = \hat{\beta} X_a$를 계산하였다. 여기에서 $\hat{\beta} = \sum_{i \in s} w_i y_i x_i^t \left(\sum_{i \in s} w_i x_i x_i^t \right)^{-1}$ 이며, 주 단위에서 사용된 보조정보는 30개의 연령 및 성별그룹의 총 수, 각 ER 지역과 CMA 지역에 대한 인구 총계 등이다.

LFS에서 사용되고 있는 표본수 의존 추정량은 앞서 언급한 두 추정량의 가중평균

$$\hat{Y}_{SSD,a} = \lambda_a \hat{Y}_{GREG,a} + (1 - \lambda_a) \hat{Y}_{SYN,GREG,a}$$

을 취한다. 여기에서 가중치 λ_a는

$$\lambda_a = \begin{cases} 1 & , \hat{N}_{c,a} \geq \delta N_a \\ \hat{N}_{c,a} / \delta N_a & , otherwise \end{cases}$$

이고, δ값은 2/3를 사용하고 있다. 만약 소지역에 대한 직접추정량이 목표 요구정도를 만족하고 있다면 합성추정량 부분에 대한 가중치는 0이 된다.

현재 LFS에서 소지역의 취업률과 실업률에 대한 추정값은 표본수 의존 추정량으로 구해진 추정치의 3개월 간의 평균값이 이용되고 있다. 캐나다 LFS에서는 6개월의 표본순환 체계를 따르고 있기 때문에 3개월의 결과를 평균하게 되면 결과적으로 1/3표본 수를 늘리는 효과를 갖게 된다. 만약 조사시점 간에 표본들이 정확히 일치한다면 추정의 효율 측면에서 이러한 이득은 기대할 수 없다.

연속조사에서 추정량의 정확도를 향상시키기 위해 서로 다른 시점에서 조사된 몇 개의 조사 자료를 풀링(pooling)하는 경우를 흔히 볼 수 있다. 특히 시점이 다른 몇 차례의 조사결과를 결합하거나 이들의 평균을 구하는 것이 보통의 방법인데, 이러한 방법은 소지역 통계 작성 시 해당 소지역에 배당된 표본 크기가 매우 작아서 추정의 정확도가 크게 떨어지는 경우에 유용하다. 그러나 다른 시점의 조사결과들을 결합하여 산출되는 추정값은 개념상의 문제점을 항상 내재하고 있다.

대영역 내에서 추정된 소지역 추정값들의 합계는 대영역의 추정값과 일치해야 하지만, 대개의 경우 소지역 추정값들의 합계는 대영역의 추정값과 일치하지는 않는다. 따라서 최종적으로 총계를 일치시키는 다음과 같은 보정이 이루어진다. 소지역 a에 대한 총계 Y_a의 추정량을 $\overline{Y}_a$, 해당 소지역을 포함하는 대영역 A에 대한 $\overline{Y}_a$의 합계를 $\overline{Y}(A)$라 하자. 이때 소지역 추정값들은 다음 (3.56)식의 $\overline{Y}_a^{ADJ}$를 통해 보정된다.

$$\overline{Y}_a^{ADJ} = \frac{\overline{Y}_a}{\overline{Y}(A)} \hat{Y}(A) \tag{3.56}$$

여기에서 $\overline{Y}(A) = \sum_{a \subset A} \overline{Y}_a$이다.

(6) AK 복합추정

LFS에서 m번째 달의 노동력 관련 특성치에 대한 총계를 Y_m이라 하고, i번째 패널에서 Y_m에 대한 단순 비추정량을 $y_{m,i}$라 하자. 여기에서 i는 m번째 달에 조사되는 6개 패널 각각을 의미한다. $m-1$번째 달과 m번째 달의 총계에 대한 변화량 $Y_m - Y_{m-1}$의 추정량 $d_{m,m-1}$은 6개의 패널 중 공통인 5개의 패널을 이용하여

$$d_{m,m-1} = \frac{\sum_{j=2}^{6} (y_{m,j} - y_{m-1,j-1})}{5} \tag{3.57}$$

와 같이 정의된다고 하자.

이때 LFS 자료를 이용하여 추정되는 총계 Y_m에 대한 AK복합추정량 y_m'은 다음 (3.58)식의 형태로 주어질 수 있다.

$$y_m' = \frac{(1-K+A)y_{m,1}}{6} + \frac{(1-K-A/5)\sum_{j=2}^{6} y_{m,j}}{6} + K(y_{m-1}' + d_{m,m-1}) \quad (3.58)$$

여기에서 K와 A는 상수들이고, K는 $0 \leq K < 1$을 만족하는 값이다.

위의 (3.58)식은 AK 복합추정량의 일종이다. 한편, 위의 AK 복합추정량은 $A=0$인 경우에는 K 복합추정량으로 불리우며, $A=0$, $K=0$인 경우에는 6개의 패널 추정량들에 대한 평균인 $\bar{y}_m = \sum_{i=1}^{6} y_{m,i}/6$의 값을 갖는다.

단순 비추정량 $y_{m,i}$에 대한 패널 간의 편향

$$\alpha_i = E(y_{m,i}) - Y_m \quad (3.59)$$

이 조사 달 m과 서로 독립이라고 가정하면 AK 복합추정량 y_m'의 편향은 다음과 같은 과정에 의해 유도될 수 있다.

(3.59)식으로부터 $E(y_{m,i}) = Y_m + \alpha_i$이고, $y_{m,i}$는 조사 달 m과는 서로 독립이므로

$$E(\bar{y}_m) = Y_m + \sum_{i=1}^{6} \frac{\alpha_i}{6}$$

$$= Y_m + \bar{\alpha}$$

이 성립하며, LFS에서는 보통 $\bar{\alpha}=0$이 가정된다. 한편, 식 (3.58)은 다음 (3.60)식과 같이 나타낼 수 있다.

$$y_m' = y_m + K(y_{m-1}' + d_{m,m-1}) \quad (3.60)$$

여기에서

$$y_m = \frac{(1-K+A)y_{m,1}}{6} + \frac{(1-K-A/5)\sum_{j=2}^{6} y_{m,j}}{6}$$

$$= (1-K)\bar{y}_m + \frac{A(y_{m,1} - \bar{y}_m)}{5}$$

따라서 y_m의 기댓값은

$$E(y_m) = (1-K)(Y_m + \bar{\alpha}) + \frac{A(\alpha_1 - \bar{\alpha})}{5} \tag{3.61}$$

으로 주어지며, $\bar{\alpha} = 0$일 경우에는 다음 식으로 축약될 수 있다.

$$E(y_m) = (1-K)Y_m + \frac{A\alpha_1}{5}$$

한편, (3.57)식으로부터 $d_{m,m-1}$의 기댓값을 계산하면 다음 (3.62)식과 같다.

$$E(d_{m,m-1}) = E\left(\frac{\sum_{j=2}^{6}(y_{m,j} - y_{m-1,j-1})}{5}\right)$$

$$= (Y_m - Y_{m-1}) + \frac{\alpha_6 - \alpha_1}{5} \tag{3.62}$$

위와 같은 과정을 n회의 이전 조사 달까지 반복 적용하면 AK 복합추정량 y_m'은

$$\begin{aligned} y_m' = {} & y_m + Ky_{m-1} + K^2 y_{m-2} + \cdots + K^{n-1} y_{m-n+1} + K^n y_{m-n}' \\ & + Kd_{m,m-1} + K^2 d_{m-1,m-2} + \cdots + K^n d_{m-n+1,m-n} \end{aligned} \tag{3.63}$$

와 같은 표현식으로 나타낼 수 있다. (3.63)식을 이용하여 y_m'의 기댓값을 계산하면 다음 (3.64)식과 같이 주어진다.

$$\begin{aligned} E(y_m') = {} & (1-K)(Y_m + KY_{m-1} + K^2 Y_{m-2} + \cdots + K^{n-1} Y_{m-n+1}) \\ & + K^n E(y_{m-n}') + \left\{(1-K)\bar{\alpha} + \frac{A(\alpha_1 - \bar{\alpha})}{5}\right\} \frac{(1-K^n)}{(1-K)} \\ & + K(Y_m - Y_{m-1}) + K^2(Y_{m-1} - Y_{m-2}) + \cdots \end{aligned}$$

$$+K^n(Y_{m-n+1}-Y_{m-n})+\left(\frac{\alpha_6-\alpha_1}{5}\right)\left(\frac{K(1-K^n)}{1-K}\right)$$

$$=Y_m+K^n\{E(y'_{m-n})-Y_{m-n}\}$$

$$+\left\{(1-K)\bar{\alpha}+\frac{A(\alpha_1-\bar{\alpha})}{5}+\frac{K(\alpha_6-\alpha_1)}{5}\right\}\left(\frac{1-K^n}{1-K}\right)$$

$$=Y_m+K^n\{E(y'_{m-n})-Y_{m-n}\}$$

$$+\left\{(1-K-A/5)\bar{\alpha}+\frac{A\alpha_1}{5}+\frac{K(\alpha_6-\alpha_1)}{5}\right\}\left(\frac{1-K^n}{1-K}\right) \quad (3.64)$$

여기에서 만약 n이 충분히 크고, $\bar{\alpha}=0$을 가정할 수 있다면, 위의 (3.64) 식은 다음과 같이 축약된다.

$$E(y'_m)=Y_m+\frac{A\alpha_1+K(\alpha_6-\alpha_1)}{5(1-K)} \quad (3.65)$$

또한 y_m' 의 편향은 월별 편향들과는 서로 독립이라고 가정하였기 때문에 이 전 조사달들과의 차이에 대한 변화량 $y'_m-y'_{m-r}$은 불편성을 만족한다. 즉,

$$E(y'_m-y'_{m-r})=Y_m-Y_{m-r} \quad (3.66)$$

의 관계가 성립한다.

LFS의 연동표본교체 체계를 도시하면 <표3.22>와 같다.

<표3.22> LFS의 표본교체 체계

m	m-1	m-2	m-3	m-4	m-5	m-6	m-7	m-8	m-9	m-10	m-11
1	(6)	(5)	(4)	(3)	(2)	(1)	((6))	((5))	((4))	((3))	((2))
2	1	(6)	(5)	(4)	(3)	(2)	(1)	((6))	((5))	((4))	((3))
3	2	1	(6)	(5)	(4)	(3)	(2)	(1)	((6))	((5))	((4))
4	3	2	1	(6)	(5)	(4)	(3)	(2)	(1)	((6))	((5))
5	4	3	2	1	(6)	(5)	(4)	(3)	(2)	(1)	((6))
6	5	4	3	2	1	(6)	(5)	(4)	(3)	(2)	(1)

<표3.22>의 m조사달과 $m-5$조사달 사이에서는 패널 간의 차 $i-j$의 값이 1과 5사이에 있을 경우, m조사달의 i패널($i=1,2,\cdots,6$)은 $m-j$조사달의 $i-j$패널과 같다. m조사달을 기준으로 바로 이전의 첫 번째 표본순환이 이루어지고 있는 6개월 간($m-1$에서 $m-6$)을 살펴보면, $6+i-j$의 값이 1과 6사이에 있을 경우, m조사달의 i패널은 $m-j$조사달의 $6+i-j$패널과 같다. 두 번째 표본순환이 이루어지고 있는 $m-7$조사달과 $m-12$조사달 사이에서는 $12+i-j$의 값이 1과 6사이에 있는 경우, m조사달의 i패널은 $m-j$조사달의 $12+i-j$패널과 같으며, 일반적으로 이전의 r번째 표본순환에서 m조사달의 i패널은 $m-j$조사달의 $6r+i-j$패널과 같다.

따라서 AK 복합추정량 y_m'의 분산 $V(y_m')$은 위의 연동교체표본 체계에서 추정된 패널 추정량들의 분산공분산 값들을 포함하게 되며 $V(y_m')$을 계산하기 위해 가정된 패널 추정량들의 분산공분산 구조는 다음과 같다.

(i) $V(y_{m,i})=\sigma^2,\quad i=1,2,\cdots,6$,

(ii) $Cov(y_{m,i},\ y_{m-j,6r+i-j})=\gamma_j^{(r)}\sigma^2,\quad i=1,2,\cdots,6,$

$$j>0,\ r\geq 0,\ 6\geq 6r+i-j\geq 1\ ,$$

$$\text{단, } \gamma_j^{(r)}=\begin{cases}\rho_j & ,r=0\ \ (i.e.,\ 6>i-j\geq 1)\\ \gamma_j & ,r=1\ \ (i.e.,\ 6\geq 6+i-j\geq 1)\\ 0 & ,r\geq 2\ (i.e.,\ 6\geq 6r+i-j\geq 1)\end{cases}$$

(iii) ρ_j와 γ_j는 정상성(stationary)을 만족한다. 즉 ρ_j와 γ_j는 j의 함수이고, m조사달과는 독립이다.

(iv) (ii)와 (iii)의 가정에 포함되지 않은 다음과 같은 공분산들은 0으로 가정한다.

$$Cov(y_{m,i},\ y_{m,j})=0\ ,\quad i\neq j,$$

$$Cov(y_{m,i},\ y_{m-1,j})=0\ ,\ i=1,j\neq 6,$$

$$Cov(y_{m,i},\ y_{m-1,j}) = 0\ ,\ i \neq 1, j \neq i-1,$$

$$Cov(y_{m,i},\ y_{m-g,j}) = 0\ ,\ g \geq 12.$$

추가적으로 AK 복합추정량 $y_m^{'}$의 월 추정값들이 매우 안정적이라면 $V(y_{m-1}^{'}) = V(y_m^{'})$을 가정할 수 있다. 위의 (ⅰ)~(ⅳ)의 가정과 $y_m^{'}$의 월 추정값들이 매우 안정적이라는 가정 하에서 복합추정량$y_m^{'}$의 분산을 유도하면 다음과 같다.

먼저 (3.60)식의 AK 복합추정량 $y_m^{'}$에 대한 분산은 다음 (3.67)식과 같이 주어진다.

$$V(y_m^{'}) = [V(y_m) + K^2 V(d_{m,m-1}) + 2KCov(y_m,\ d_{m,m-1}) + 2KCov(y_m,\ y_{m-1}^{'}) + 2K^2 Cov(d_{m,m-1},\ y_{m-1}^{'})]/(1-K^2) \tag{3.67}$$

(3.63)식에서 m을 $m-1$로, n을 12로 대체하면 $y_{m-1}^{'}$은 다음과 같이 나타낼 수 있다.

$$y_{m-1}^{'} = \sum_{g=1}^{12}(K^{g-1}y_{m-g} + K^g d_{m-g,m-g-1}) + K^{12}y_{m-13}^{'} \tag{3.68}$$

(3.63)식을 (3.68)식에 대입하고 가정에 따라 0의 값을 갖는 항들을 제거하면 $V(y_m{'})$은

$$\begin{aligned} V(y_m^{'}) = [&V(y_m) + K^2 V(d_{m,m-1}) + 2KCov(y_m,\ d_{m,m-1}) \\ &+ 2\sum_{g=1}^{12} K^g\{Cov(y_m,\ y_{m-g}) + KCov(d_{m,m-1},\ y_{m-g}) \\ &+ KCov(y_m,\ d_{m-g,m-g-1}) \\ &+ K^2 Cov(d_{m,m-1},\ d_{m-g,m-g-1})\}]/(1-K^2) \end{aligned} \tag{3.69}$$

와 같이 주어진다. 여기에서 $V(y_m)$, $V(d_{m,m-1})$, $Cov(y_m, d_{m,m-1})$은 (3.57)식과 (3.60)식으로부터 유도될 수 있고 계산결과는 다음과 같다.

$$V(y_m)=\left\{\frac{(1-K)^2}{6}+\frac{A^2}{30}\right\}\sigma^2, \tag{3.70}$$

$$V(d_{m,m-1})=\frac{2\sigma^2(1-\rho_1)}{5}, \tag{3.71}$$

$$Cov(y_m, d_{m,m-1})=\frac{(1-K)(1-\rho_1)\sigma^2}{6}-\frac{A(1-\rho_1)\sigma^2}{30}, \tag{3.72}$$

$$\begin{aligned}Cov(y_m, y_{m-g}) &= \sigma^2\rho_g I(g,5)\{(1-K)^2(6-g)/36\\ &\quad + A(1-K)(g-3)/90 - gA^2/900\}\\ &\quad + \sigma^2\gamma_g\{(1-K)^2(6-|g-6|)/36\\ &\quad + A(1-K)(|g-6|-3)/90 - |g-6|A^2/900\}\\ &\quad + \sigma^2\gamma_g I(g,6)I(6,g)A(1-K+A)/30,\end{aligned} \tag{3.73}$$

$$\begin{aligned}Cov(d_{m,m-1}, y_{m-g}) &= \sigma^2(\rho_g-\rho_{g-1})I(g,5)\{(1-K)(6-g)/30\\ &\quad + gA/150\} + \sigma^2(\gamma_g-\gamma_{g-1})\{(1-K)(6-|g-6|)/30\\ &\quad + |g-6|A/150 - (1-K+A)I(g,6)/30\},\end{aligned} \tag{3.74}$$

$$\begin{aligned}Cov(y_m, d_{m-g,m-g-1}) &= \sigma^2(\rho_g-\rho_{g+1})I(g,5)(1-K-\frac{A}{5})\\ &\quad \times\left(\frac{5-g}{30}\right)+\sigma^2(\gamma_g-\gamma_{g+1})\{(1-K-\frac{A}{5})(6-I(6,g)\\ &\quad - |g-6|)/30 + AI(g,5)/25\},\end{aligned} \tag{3.75}$$

$$\begin{aligned}Cov(d_{m,m-1}, d_{m-g,m-g-1}) &= \sigma^2\{(5-g)(2\rho_g-\rho_{g-1}-\rho_{g+1})\\ &\quad \times I(g,5)+(5-|g-6|)(2\gamma_g-\gamma_{g-1}-\gamma_{g+1})\}/25.\end{aligned} \tag{3.76}$$

단, $I(a,b)=\begin{cases}1 & , \text{if } a\leq b\\ 0 & , otherwise\end{cases}$.

한편, $y'_m-y'_{m-1}$의 분산도 (3.60)식과 위의 결과들로부터 유도될 수 있다. 먼저 (3.60)식을

$$y_m' - Ky_{m-1}' = y_m + Kd_{m,m-1}$$

와 같이 나타내고 양변에 대해서 분산을 취하면 다음과 같은 결과를 얻을 수 있다.

$$(1+K^2)V(y_m') - 2KCov(y_m', y_{m-1}') = V(y_m) + K^2V(d_{m,m-1}) + 2KCov(y_m, d_{m,m-1})$$

$K \neq 0$인 경우를 고려해 보자. $Cov(y_m', y_{m-1}')$는 위의 식으로부터 계산이 가능하다. 앞서 언급된 (3.70), (3.71) 및 (3.72)의 결과와

$$V(y_m' - y_{m-1}') = 2V(y_m') - 2Cov(y_m', y_{m-1}')$$

의 관계를 이용한다면 $y_m' - y_{m-1}'$의 분산은 다음 (3.77)식과 같이 주어진다.

$$V(y_m' - y_{m-1}') = \sigma^2\{A^2/30 - (1-\rho_1)KA/15 + (1-K)^2/6 + (1-\rho_1)K(K+5)/15\}/K - (1-K)^2V(y_m')/K \quad (3.77)$$

$K=0$인 경우에 대해서는

$$y_m' = \left(1 - \frac{A}{5}\right)\bar{y}_m + \frac{Ay_{m,1}}{5},$$

$$Cov(y_m', y_{m-1}') = Cov\left\{\left(1 - \frac{A}{5}\right)\bar{y}_m + \frac{Ay_{m,1}}{5}, \left(1 - \frac{A}{5}\right)\bar{y}_{m-1} + \frac{Ay_{m-1,1}}{5}\right\} \quad (3.78)$$

을 이용하여 $y_m' - y_{m-1}'$의 분산 공식을 유도할 수 있고 결과는 다음 (3.79)식과 같이 주어진다.

$$V(y_m' - y_{m-1}') = \sigma^2\{(1/15 + \rho_1/450 + \gamma_1/90)A^2 + 2(\rho_1 - \gamma_1)A/45 + 1/3 - (5\rho_1 + \gamma_1)/18\} \quad (3.79)$$

위의 결과들에 대한 적용사례로써 캐나다 Ontario 지역의 LFS자료 (1980~1981년)에 근거한 추정결과를 소개하기로 한다. $V(y_m')$의 표현식

에서 σ^2, ρ_j와 γ_j는 그들의 추정값으로 대체하여 계산되었다. 캐나다 LFS에서 $j \geq 6$의 범위에 대해서는 겹치는 패널들이 없기 때문에 이 범위에서는 ρ_j값들이 존재하지 않으며, γ_j도 $j \geq 12$의 범위에서는 값들이 존재하지 않는다. LFS자료에 근거하여 추정된 $\hat{\rho}_j$의 값이 다음 <표3.23>에 주어졌다. 여기에서 $\hat{\rho}_5$의 값은 표본으로부터 직접 계산할 수 없기 때문에 외삽법으로 계산된 값이다.

<표3.23> 특성치들에 대한 추정상관계수 $\hat{\rho}$ (1980-1981, Ontario)

구 분	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$	$\hat{\rho}_4$	$\hat{\rho}_5$
노동력인구	0.843	0.782	0.717	0.674	0.631
취업자	0.852	0.779	0.709	0.664	0.619
농업인구	0.955	0.926	0.901	0.861	0.821
비농취업인구	0.861	0.791	0.724	0.678	0.632
실업자	0.580	0.445	0.334	0.286	0.238

예상했던 바와 같이 노동력인구, 취업자, 농업인구, 비농취업인구 및 실업자의 5가지의 관심영역에 대한 추정값 $\hat{\rho}_j$은 j값이 증가함에 따라 감소하는 경향을 보인다. 특히 실업자를 제외한 모든 특성치들에 대해서 $\hat{\rho}_j$의 값들은 매우 큰 값을 나타낸다.

5가지의 노동력 관심영역에 대한 추정값 $\hat{\gamma}_j$는 다음 <표3.24>에 주어졌다. 여기에서 $\hat{\gamma}_5$는 내삽법에 의해 계산된 값이고, $\hat{\gamma}_{11}$은 외삽법에 의해 계산된 결과이다.

<표3.24> 특성치들에 대한 추정상관계수 $\hat{\gamma}$ (1980-1981, Ontario)

구분	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$	$\hat{\gamma}_5$	$\hat{\gamma}_6$	$\hat{\gamma}_7$	$\hat{\gamma}_8$	$\hat{\gamma}_9$	$\hat{\gamma}_{10}$	$\hat{\gamma}_{11}$
노동력인구	0.16	0.14	0.13	0.13	0.14	0.14	0.13	0.13	0.12	0.12	0.13
취업자	0.16	0.14	0.14	0.14	0.15	0.15	0.15	0.15	0.15	0.14	0.15
농업인구	0.48	0.48	0.47	0.49	0.48	0.47	0.46	0.43	0.39	0.32	0.25
비농취업인구	0.18	0.15	0.15	0.16	0.16	0.17	0.17	0.17	0.17	0.16	0.17
실업자	0.14	0.07	0.08	0.06	0.06	0.05	0.05	0.06	0.08	0.14	0.07

일반적으로 각 특성치들에 대해서 j가 증가할 때 γ_j는 감소할 것으로 예측되지만 <표3.24>에서 γ_j의 추정값들을 살펴보면 반드시 그러한 경향을 보이지는 않으며, 오히려 $\hat{\gamma}_{j+1}-\hat{\gamma}_j$의 값이 양의 값을 갖는 경우도 발생하는 것을 확인할 수 있다. 여기에서 $\hat{\gamma}_{j+1}-\hat{\gamma}_j$의 값이 양의 값을 갖는 경우에 대해서는 실제 상황으로 받아들이기 보다는 표본추출 오차에 기인하는 것으로 판단하는 것이 옳을 것 같다.

단순 비추정량의 추정분산에 대한 AK 복합추정량 및 K 복합추정량의 추정분산의 비를 계산한 상대효율 값들과 AK 복합추정량 또는 K 복합추정량의 분산을 최소화하는 최적상수 A와 K의 값들이 다음 <표3.25>와 <표3.26>에 주어졌다. 여기에서 분산에 대한 상대효율값들은 다음과 같이 정의되었다.

$$\text{상대효율} = \frac{Var(\text{단순비추정량})}{Var(AK\text{복합추정량 또는 }K\text{복합추정량})} \times 100$$

<표3.25>와 <표3.26>에서 최적 상수값들은 각 노동력 관심영역에 대해 분산을 최소화하는 값들로써 적절히 선택된 값들이다. <표3.25>의 결과를 살펴보면, 실업자를 제외한 나머지 노동력 관심영역들에 대해서 K 복합추정량은 단순 비추정량에 비해 약 $18 \sim 21\%$의 효율이득이 발생하였고, AK 복합추정량은 약 $26 \sim 30\%$의 효율이득이 발생하였다. 상대적으로 AK 복합추정량의 효율이득이 K 복합추정량의 효율이득보다 다소

높게 나타난다.

<표3.25> 추정량들의 상대효율 및 최적 상수($\gamma_i \neq 0$)

구 분	K 복합추정		AK 복합추정		
	최적K값	상대효율	최적K값	최적A값	상대효율
노동력인구	0.7	118.8	0.8	0.48	128.4
취업자	0.7	118.5	0.8	0.49	128.1
농업인구	0.8	120.6	0.8	0.38	126.9
비농취업인구	0.7	119.4	0.8	0.47	129.3
실업자	0.3	102.8	0.5	0.38	105.2

<표3.26> 추정량들의 상대효율 및 최적 상수($\gamma_i = 0$)

구 분	K 복합추정		AK 복합추정		
	최적K값	상대효율	최적K값	최적A값	상대효율
노동력인구	0.7	125.5	0.8	0.50	138.4
취업자	0.7	125.3	0.8	0.51	137.9
농업인구	0.8	167.3	0.9	0.46	187.9
비농취업인구	0.7	126.9	0.8	0.49	140.2
실업자	0.4	104.4	0.6	0.51	108.4

<표3.26>은 상대효율에 있어서 γ_j의 효과를 알아보기 위해 γ_j의 값들을 0으로 놓고 산출한 결과이다. <표3.25>와 <표3.26>을 비교해 보면 γ의 값이 0이 아닌 양의 값을 취할 때 5개 노동력 관심영역들의 분산 값들이 $\gamma=0$인 경우보다 훨씬 큰 값을 가지며 효율이득도 그만큼 감소한다. 따라서 $\gamma_j > 0$인 실제상황에서 $\gamma_j = 0$이 가정된다면 효율이득이 과대추정될 수 있다는 사실을 이들의 결과로부터 주목해야 한다.

앞서 언급한 바와 같이 복합추정량의 결점은 추정량의 편향 가능성에 있다. 따라서 편향된 추정량의 분산만을 비교하여 상대효율을 거론하는 것보다는 MSE 값들을 산출하여 효율이득을 비교하는 것이 보다 바람직

하다. AK 복합추정량 $y_m^{'}$의 편향은 (3.65)식에서와 같이 패널 간의 편향 α_i를 포함한다. 추정편향 $\hat{\alpha}_i = y_{m,i} - \hat{Y}_m$이 α_i의 불편추정량이고(즉, $\hat{Y}_m$이 Y_m의 불편추정량), 단순 비추정량 $\bar{y}_m$이 Y_m의 불편추정량이라고 가정하여 계산된 $\hat{\alpha}_i = y_{m,i} - \sum_{i=1}^{6} y_{m,i}/6$ 값들이 다음 <표3.27>에 주어졌다. 노동력인구, 취업자, 비농취업인구의 3가지 관심영역에 대해서 $\hat{\alpha}_1$의 값은 다른 값들에 비해 상당히 큰 음의 값을 가지며, 다른 $\hat{\alpha}_i$의 값들은 대부분 양의 값을 갖는다.

<표3.27> 편향 α_i의 추정결과(단위:천명)

구 분	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$
노동력인구	-135.3	39.8	41.1	31.1	15.4	7.9
취업자	-141.7	35.5	34.9	31.3	25.4	14.8
농업인구	-4.2	-2.6	2.2	-0.1	4.2	0.5
비농취업인구	-137.5	38.0	32.7	31.3	21.2	14.3
실업자	6.4	4.3	6.2	-0.1	-9.9	-6.9

5개의 노동력 관심영역들 각각에 대해서 분산 및 MSE 값을 최소화하는 최적상수 K, A값들 및 이들 최적상수값들에 의해 계산된 분산과 MSE값들이 다음 <표3.28>에 주어졌다. MSE에 대한 상대효율값들은 다음 식을 이용하여 계산되었다.

$$\text{상대효율} = \frac{MSE(\text{단순비추정량})}{MSE(AK\text{복합추정량 또는 } K\text{복합추정량})} \times 100$$

<표3.28> 단순 비추정량, K 복합추정량, AK 복합추정량의 분산과 MSE 값들의 비교

(ⅰ) 노동력인구

노동력인구	단순 비추정	K복합추정		AK복합추정		(공통상수) $K=0.4$ $A=0.4$
		(min MSE) $K=0$	(min Var) $K=0.7$	(min MSE) $K=0.7$ $A=0.7$	(min Var) $K=0.8$ $A=0.5$	
월추정값(10^3)	4480.7	4480.7	4547.5	4484.4	4527.6	4481.7
분산(10^6)	432.0	432.0	363.8	358.1	336.5	391.8
편향(10^3)	0	0	66.8	3.7	46.9	1.0
MSE(10^6)	432.0	432.0	4284.5	371.5	2532.4	392.9
상대효율		100.0	9.0	116.3	17.1	110.0

(ⅱ) 취업자

취업자	단순 비추정	K복합추정		AK복합추정		(공통상수) $K=0.4$ $A=0.4$
		(min MSE) $K=0$	(min Var) $K=0.7$	(min MSE) $K=0.8$ $A=0.9$	(min Var) $K=0.8$ $A=0.5$	
월추정값(10^3)	4186.0	4186.0	4259.0	4183.6	4240.3	4188.0
분산(10^6)	473.3	473.3	399.6	397.7	369.5	428.9
편향(10^3)	0	0	73.0	-2.4	54.3	2.0
MSE(10^6)	473.3	473.3	5732.2	403.2	3320.9	432.8
상대효율		100.0	8.3	117.4	14.3	109.4

(ⅲ) 농업인구

농업인구	단순 비추정	K복합추정		AK복합추정		(공통상수) $K=0.4$ $A=0.4$
		(min MSE) $K=0.6$	(min Var) $K=0.8$	(min MSE) $K=0.8$ $A=0.6$	(min Var) $K=0.8$ $A=0.4$	
월추정값(10^3)	142.0	143.4	145.7	143.2	144.1	142.1
분산(10^6)	85.7	75.6	71.1	68.7	67.6	80.8
편향(10^3)	0	1.4	3.7	1.2	2.1	0.1
MSE(10^6)	85.7	77.6	85.1	70.2	71.8	80.8
상대효율		110.5	110.7	122.2	119.4	106.1

(iv) 비농취업인구

비농취업인구	단순 비추정	K복합추정		AK복합추정		(공통상수) $K=0.4$ $A=0.4$
		(min MSE) $K=0$	(min Var) $K=0.7$	(min MSE) $K=0.8$ $A=0.9$	(min Var) $K=0.8$ $A=0.5$	
월추정값(10^3)	4043.9	4043.9	4114.7	4041.6	4096.6	4045.8
분산(10^6)	498.9	498.9	417.8	418.0	385.9	452.8
편향(10^3)	0	0	70.8	-2.3	52.7	1.9
MSE(10^6)	498.9	498.9	5436.1	423.3	3161.7	456.4
상대효율		100.0	9.2	117.9	15.8	109.3

(v) 실업자

실업자	단순 비추정	K복합추정		AK복합추정		(공통상수) $K=0.4$ $A=0.4$
		(min MSE) $K=0.2$	(min Var) $K=0.3$	(min MSE) $K=0.4$ $A=0.4$	(min Var) $K=0.5$ $A=0.4$	
월추정값(10^3)	294.8	294.1	293.7	293.9	293.2	293.9
분산(10^6)	117.5	114.9	414.3	112.5	111.7	112.5
편향(10^3)	0	-0.7	-1.1	-0.9	-1.6	-0.9
MSE(10^6)	117.5	115.4	115.7	113.3	114.4	113.3
상대효율		101.9	101.6	103.7	102.7	103.7

여기에서 최적상수 K는 0.0에서부터 0.9까지 0.1의 간격으로 선택된 10개의 값들 중에서 MSE 값을 최소화시키는 상수 값이다. AK 복합추정량에서 상수(K, A)는 $K(=0.0 \sim 0.9,$ 간격=0.1$)$와 $A(=0.0 \sim 1.0,$ 간격=0.1$)$값들의 조합에 의해 최적가중치가 결정되었다. K 복합추정량과 AK 복합추정량의 편향은 (3.65)식을 이용하여 계산하였다(α_1과 α_6 대신 <표 3.27>에서 추정된 $\hat{\alpha}_1$과 $\hat{\alpha}_6$을 대입하여 계산).

<표3.28>에서 MSE를 최소화하는 상수값들에 의해 계산된 추정결과들을 살펴보자. K 복합추정량은 "농업인구"와 "취업자"에서 단순 비추정량에 비해 약간의 효율이득이 발생하였다. 분산을 최소화하는 상수값들에

의해 계산된 추정결과에서 "노동력 인구", "취업자", "비농취업인구"에 대해서는 매우 낮은 10%미만의 상대효율값을 갖는 것으로 나타났다. 이는 편향의 영향에 기인하는 것으로 나타났다.

MSE를 최소화하는 상수값들에서 계산된 AK 복합추정값은 "실업자"를 제외한 나머지 4개 노동력 관심영역에서 단순 비추정값에 비해 16 ~ 22%의 상대효율이득이 발생하였다. 4개 노동력 관심영역에 대한 편향 값은 매우 작게 나타났다. 분산을 최소화하는 상수값들에 의해 계산된 추정결과를 살펴보자. "노동력인구", "취업자", "비농취업인구"에 대해서는 K 복합추정량과 마찬가지로 편향의 영향에 의해 매우 낮은 상대효율값을 나타낸다.

결론적으로 "실업자"를 제외한 나머지 4개의 노동력 관심영역에서 AK 복합추정량의 상대효율이 K 복합추정량의 상대효율보다 높게 나타났다(MSE를 최소화하는 상수값들에 의해 계산된 추정값들을 기준). 따라서 주어진 LFS 자료에서는 K 복합추정량보다는 AK 복합추정량의 효율이득이 높은 것으로 결론지을 수 있다.

한편, K 복합추정량 또는 AK 복합추정량의 월별변화량 $y_m^{'} - y_{m-1}^{'}$은 $Y_m - Y_{m-1}$의 불편추정량이다. <표3.29>에 이들 월별변화량에 대한 추정결과가 주어졌다.

<표3.29> 월별변화량에 대한 복합추정량의 상대효율

구 분	K 복합추정		AK 복합추정			공통상수 $K, A = 0.4$
	최적K값	상대효율	최적K값	최적A값	상대효율	상대효율
노동력인구	0.9	146.6	0.9	0.1	147.9	113.3
취업자	0.9	151.0	0.9	0.1	152.3	114.1
농업인구	0.9	234.7	0.9	0.0	234.7	112.3
비농취업인구	0.9	154.0	0.9	0.1	155.2	114.1
실업자	0.4	106.0	0.6	0.2	106.4	102.9

월변화량에 대해서는 K 복합추정량 및 AK 복합추정량의 상대효율(단순 비추정량에 대한 상대효율)이 "실업자"를 제외한 나머지 관심영역에서 매우 높게 나타나며, "노동력인구", "취업자", "비농취업인구"의 영역에서는 약 46 ~ 55%의 효율이득이 발생하였고, "농업인구"영역에서는 매우 높은 약 135%의 효율이득이 발생하였다.

(7) 노동력 추정값의 분산추정

(가) 배 경

캐나다 노동력 조사(LFS)는 캐나다 통계청에서 실시하는 가장 큰 규모의 월 단위 가구 조사로써 주로 전국 단위, 주 단위 및 주 내의 소지역 단위에 대한 다양한 노동력 특성에 대한 추정값들을 생산하고 있다. 캐나다의 LFS는 6개의 순환 패널을 갖는 층화 다단계 연동교체 표본설계를 따른다. 매번 인구 센서스 후 LFS는 표본설계에서 부분적인 보완이 이루어져 왔으며, 특히, 1981년에는 표본추출, 데이터 수집 및 추정 방법론 등에서 포괄적인 보완이 이루어졌다. 이 시기에 주 내의 소지역에 대한 추정치의 신뢰도를 높이기 위한 사후층화 비추정 절차가 새로이 마련되었다.

여기에서는 분산추정의 방법론에 대한 연구결과를 요약하였다. 과거 LFS의 분산추정은 Keyfitz 절차를 일반화한 Woodruff의 계산법이 이용되었다(Woodruff 1971). 이 방법은 Keyfitz 방법으로 불리우기도 한다. LFS에서는 다음 소개되는 세가지 유형의 지역들이 표본설계에 반영된다. 주요도시들로 구성되어 있는 SR지역(self-representing area), 소규모 도시들과 시골을 포함하는 NSR지역(non-self-representing area)과 군사지역 등과 같은 특수지역들이 이러한 유형의 지역들이다. NSR지역과 특수지역들에 대한 분산추정은 비추정방법에 Keyfitz 방법을 혼합하여 적용하였다. 이 단계 랜덤그룹 표본설계가 반영된 SR지역들에 대해서는 Rao, Hartley and Cochran(1962)과 Rao(1975)의 분산 추정량을 이용하였고, 이 방법을 Keyfitz 방법을 이용한 분산추정량과 비교하였다. 한편, 추정값들

에 대한 분산 추정량들을 비 보정(ratio adjustment) 분산추정량들과 편향 및 안정성의 측면에서 비교하였다. 또한 반복 수의 증가에 따른 Keyfitz 분산추정량의 영향도 살펴보았다.

(나) SR 설계에 대한 분산추정

① SR 설계(SR Design)

LFS 표본설계에서 SR 지역들은 이단계 랜덤그룹 설계방식을 취하며, 일단계 추출단위(PSU)들은 확률비례추출로 추출되며 이단계 추출단위들은 계통추출이 이루어진다. 하나의 층에 N개의 일차추출단위(PSU)가 있고, j번째 PSU에 대한 크기 측도를 x_j $(j=1,2,\cdots,N)$, 거주단위의 수를 M_j, 층에서의 추출률을 $1/W$, 층으로부터 추출된 PSU의 수를 n이라 하자. N개의 PSU는 n개의 그룹으로 랜덤하게 분할되고 i번째 랜덤 그룹은 N_i개의 PSU들을 포함한다. 여기에서 $\sum_{i=1}^{n} N_i = N$이다. 우선 다음의 수식을 정의하자.

$$p_j = \frac{x_j}{\sum_{t=1}^{N} x_t}, \quad j=1,2,\cdots,N_t$$

$$\delta_{ij} = \begin{cases} 1 & , \text{if } the\ j^{th}\ PSU\ is\ in\ the\ i^{th}\ group \\ 0 & , otherwise \end{cases}$$

이때 $\pi_j = \sum_{j=1}^{N} \delta_{ij} p_j$는 i번째 랜덤 그룹의 상대적인 크기를 나타낸다.

$a_{ij} = \delta_{ij} W p_j / \pi_{ij}$, $r_{ij} = a_{ij} - [a_{ij}]$이라 하고, $\{r_{ij}, j=1,2,\cdots,N\}$가 내림차순으로 정렬되었다고 할 때, 계통추출의 추출간격 W_{ij}는 다음과 같이 정의할 수 있다.

$$W_{ij} = [a_{ij}] + 1, \quad j=1,2,\cdots,R$$
$$= [a_{ij}], \quad j=R+1,\cdots,N$$

여기에서 $R=\sum_{j=1}^{N} r_{ij}$, $\sum_{j=1}^{N} W_{ij}=W,\ i=1,2,\cdots,n$..

하나의 PSU는 n개의 랜덤 그룹 각각으로부터 추출률 W_{ij}에 비례하는 확률로 추출된다. i번째 랜덤 그룹으로부터 추출된 j번째 PSU는 $1/W_{ij}$의 비율로 부차추출된다. 이때 전체 추출률은 $1/W$이 된다. 각각의 랜덤 그룹은 1에서 6까지의 패널로 할당되고, 랜덤 그룹의 수 n은 보통 6의 배수이고 각각의 패널은 같은 수의 랜덤 그룹을 갖는다. 각 랜덤 그룹으로부터 하나의 PSU가 추출되므로 i번째 랜덤 그룹으로부터 추출된 PSU에서 부차추출률은 $1/W_i$이 된다. 랜덤 그룹 i에서 선택된 거주단위의 수는 m_i로 나타내기로 한다.

② 분산 추정량(Variance Estimator)

하나의 층에 대한 특성치 y의 총계에 관심이 있다고 하자. j번째 PSU에서 k번째 거주단위에 대한 y값을 y_{ik} $(k=1,2,\cdots,M_j)$라 할 때, 총계 $Y=\sum_{j=1}^{N}\sum_{k=1}^{M_j} y_{ik}$는 $\hat{Y}=W\sum_{i=1}^{n} y_i$로 추정될 수 있다. 여기에서 y_i는 i번째 그룹에서 선택된 PSU로부터 추출된 m_i개의 거주단위에 대한 y값들의 합을 나타낸다. $\hat{Y}$의 분산추정량은 다음과 같은 방법으로 추정될 수 있다.

(i) Keyfitz의 분산 추정량(1957)

과거 표본설계에서 이용하였던 총계 추정량에 대한 분산 추정공식은 다음 (3.80)식과 같다.

$$\hat{V}_1(\hat{Y})=W^2\left(\sum_{a} y_i-\sum_{e} y_i\right)^2 \qquad (3.80)$$

여기에서 $\sum_{a}$는 홀수 패널들에 대한 합, $\sum_{e}$는 짝수 패널들에 대한 합을 나타낸다. 위의 (3.80)식을 일반화시킨 일반화 Keyfitz 분산 추정공식은

다음 (3.81)식과 같다.

$$\hat{V}_2(\hat{Y}) = W^2 \frac{n}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (3.81)$$

여기에서 $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$이며, $\hat{V}_2(\hat{Y})$가 효율성이나 안정성 측면에서 $\hat{V}_1(\hat{Y})$보다 선호될 수 있다.

(ii) Rao, Hartley and Cochran의 분산 추정량(1962)

Rao, Hartley and Cochran의 분산 추정공식은 i번째 그룹으로부터 추출된 m_i개의 거주단위의 수가 고정되어있고, 거주단위들은 단순임의추출이 가정된 상태에서 유도된다. 분산 추정공식은 다음과 같이 주어진다.

$$\hat{V}_3(\hat{Y}) = A \sum_{i=1}^{n} \pi_i \left(\frac{M_i}{m_i} \frac{y_i}{p_i} - \hat{Y} \right)^2 + \sum_{i=1}^{n} \frac{\pi_i}{p_i} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i} \right) s_i^2 \qquad (3.82)$$

여기에서

$$A = \frac{\sum_{i=1}^{n} N_i^2 - N}{N^2 - \sum_{i=1}^{n} N_i^2},$$

$$s_i^2 = \frac{1}{m_i - 1} \sum_{k=1}^{m_i} (y_{ik} - \bar{y}_i)^2$$

이고, M_i는 i번째 그룹에서 선택된 PSU에 속해있는 거주단위들의 수이고 M_i개의 거주단위들 중 m_i개의 거주단위들이 계통추출로 추출되나 분산추정값은 단순임의추출 하에서 계산된다. i번째 그룹에서 선택된 PSU로부터 추출된 k번째 거주단위에 대한 y값이 y_{ik}이며 $\bar{y}_i$는 $\bar{y}_i = y_i/m_i$로 주어진다.

$\pi_i/p_i = W/W_i$이고, $M_i/m_i = W_i$이므로 (3.82)의 분산 추정공식은 다음

(3.83)식과 같이 주어질 수 있다.

$$\hat{V}_3(\hat{Y}) = A\sum_{i=1}^{n}\pi_i\left(W\frac{y_i}{\pi_i} - \hat{Y}\right)^2 + W\sum_{i=1}^{n}\left(1-\frac{m_i}{M_i}\right)M_i s_i^2 \qquad (3.83)$$

(iii) Rao의 분산 추정량(1975)

Rao의 분산 추정공식에서는 m_i개의 거주단위들이 단순임의추출로 추출되나, 표본크기 m_i를 확률변수로 취급하여 분산 추정공식을 유도하였다. Rao의 분산 추정공식은 다음 (3.84)식과 같이 주어진다.

$$\begin{aligned}\hat{V}_4(\hat{Y}) &= A\sum_{i=1}^{n}\pi_i\left(W\frac{y_i}{\pi_i} - \hat{Y}\right)^2 + \sum_{i=1}^{n}\left\{\frac{\pi_i^2}{p_i^2} - A\left(\frac{\pi_i}{p_i^2} - \frac{\pi_i^2}{p_i^2}\right)\right\}\frac{M_i^2 s_i^2}{m_i} \\ &\quad - \sum_{i=1}^{n}\frac{\pi_i}{p_i}M_i s_i^2 \\ &= \hat{V}_3(\hat{Y}) + W^2\sum_{i=1}^{n}m_i s_i^2\left\{\left(1-\frac{W_i}{W}\right) - A\left(\frac{1}{\pi_i}-1\right)\right\}\end{aligned} \qquad (3.84)$$

Rao의 분산 추정공식은 이 단계 표본추출에서 확률표본크기가 가정되기 때문에 음의 값이 나올 가능성이 있음에 주의해야한다.

③ Monte Carlo Study

네 가지의 분산 추정량들의 편향과 상대적인 안정성을 검토하기 위한 몬테카를로 연구가 수행되었다. 이용된 자료는 1981년 센서스 자료 중 조사지역 내의 약 20%의 계통추출 표본이며 Halifax의 CMA(Census Metropolitan Area) 지역으로부터 19개의 층에 대해 검토가 이루어졌다. 추출률 $1/W$은 0.04로 주어진다. 19개 층에 대한 PSU의 수, 추출된 PSU의 수, 거주단위들의 수 및 기대 표본크기가 <표3.30>에 주어졌다.

<표3.30> 몬테카를로 연구를 위해 이용된 층

층	거주단위 수	PSU 수	추출된 PSU 수	기대 표본크기
1	737	49	6	29.5
2	490	33	4	19.6
3	745	45	6	29.8
4	720	34	6	28.8
5	621	37	6	24.8
6	630	38	6	25.2
7	503	31	4	20.1
8	340	23	4	13.6
9	472	33	4	18.9
10	468	33	4	18.7
11	367	28	4	14.7
12	390	23	4	15.6
13	626	36	6	25.0
14	650	39	6	26.0
15	350	22	4	14.0
16	736	46	6	29.4
17	573	35	6	22.9
18	773	48	6	30.9
19	866	64	8	34.6
합계	11,056	697	100	442.3

몬테카를로 방법을 이용하여 각 층에서 1,000 개의 표본이 독립적으로 생성되었다. t번째 몬테카를로 표본생성으로부터 층 h에 대한 총계 Y_h의 추정값을 $\hat{Y}_{ht}$ $(h=1,2,\cdots,19,\ t=1,2,\cdots,1000)$, $\hat{V}_{jht}(j=1,2,3,4)$를 $\hat{Y}_{ht}$의 네 개의 분산추정량이라 하고 다음을 정의하였다.

$$Y=\sum_{h=1}^{19} Y_h\ ,$$

$$\hat{Y}_t=\sum_{h=1}^{19}\hat{Y}_{ht}\ ,\quad t=1,2,\cdots,1000$$

$$\hat{V}_{jt}=\sum_{h=1}^{19}\hat{V}_{jht}\ ,\quad j=1,2,3,4$$

여기에서 $\hat{Y}_t$는 t번째 몬테카를로 표본생성에서 얻어지는 총계 Y의 추정값, $\hat{V}_{jt}(j=1,2,3,4)$는 분산 추정값을 나타낸다.

몬테카를로 기대값과 분산을 각각 E^*와 V^*로 표기할 때 T개의 몬테카를로 표본생성에 대한 기대값과 분산은 각각 다음과 같이 주어진다.

$$E^*(\hat{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_t ,$$

$$V^*(\hat{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\{\hat{\theta}_t - E^*(\hat{\theta})\}^2$$

위의 정의를 이용하여 $\hat{Y}$의 몬테카를로 분산 $V^*(\hat{Y})$과 분산추정량 $\hat{V}_j$의 몬테카를로 기대값 $E^*(\hat{V}_j)$와 몬테카를로 분산 $V^*(\hat{V}_j)$를 얻을 수 있다.

분산추정량 $\hat{V}_j$의 편향과 백분위 편향은 각각 다음과 같이 정의될 수 있다.

$$B_j = E^*(\hat{V}_j) - V^*(\hat{Y}),$$

$$PB_j = 100\frac{B_j}{V^*(\hat{Y})}, \quad j=1,2,3,4$$

이때 $\hat{V}_j$의 평균제곱오차 MSE는 다음과 같이 주어진다.

$$MSE_j = V^*(\hat{V}_j) + B_j^2 , \quad j=1,2,3,4$$

Keyfitz 분산추정량 $\hat{V}_1$에 대한 $\hat{V}_j$의 상대효율은 다음과 같이 정의할 수 있다.

$$Rel.\,Eff(\hat{V}_j\ vs.\ \hat{V}_1) = \sqrt{\frac{MSE_1}{MSE_j}} , \quad j=1,2,3,4$$

분산추정량들에 대한 상대편향과 효율에 대한 결과는 <표3.31>과 <표3.32>에 주어졌다.

<표3.31> 총계 추정값에 대한 분산추정량들의 백분위 편향

구 분	백분위 편향(PB_j)			
	$\hat{V}_1$	$\hat{V}_2$	$\hat{V}_3$	$\hat{V}_4$
취업인구	23.4	24.5	-4.7	-6.3
실업인구	6.3	6.6	3.7	1.2
노동력인구	24.2	25.2	-5.1	-6.7

<표3.32> $\hat{V}_1$에 대한 $\hat{V}_2$, $\hat{V}_3$, $\hat{V}_4$의 상대효율

구 분	백분위 편향(PB_j)		
	$\hat{V}_2$	$\hat{V}_3$	$\hat{V}_4$
취업인구	1.51	3.22	3.11
실업인구	1.52	1.71	1.76
노동력인구	1.49	3.24	3.12

편향에 대해서는 분산추정량 $\hat{V}_1$과 $\hat{V}_2$가 유사하며, $\hat{V}_3$와 $\hat{V}_4$도 비슷한 결과를 보인다. 또한 $\hat{V}_1$과 $\hat{V}_2$의 편향은 취업인구와 노동력인구에서 비교적 큰 양의 편향값을 보이는 반면 $\hat{V}_3$와 $\hat{V}_4$의 편향은 상대적으로 작은 값을 나타낸다. 효율성 측면에서 $\hat{V}_3$와 $\hat{V}_4$가 서로 유사하며 $\hat{V}_1$과 $\hat{V}_2$에 비해서는 월등한 효율을 보인다. $\hat{V}_1$과 $\hat{V}_2$중에서는 $\hat{V}_2$의 효율이 더 좋게 나타난다.

전체인구에 대한 비 추정값들의 분산추정량에 대한 결과가 다음 <표3.33>과 <표3.34>에 주어졌다. 비 추정값들에 대한 분산추정량은 $\hat{V}_j^{(R)}$ $(j=1,2,3,4)$로 표기하였다.

<표3.33> 총계 추정값에 대한 분산추정량들의 백분위 편향

구 분	백분위 편향(PB_j)			
	$\hat{V}_1^{(R)}$	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
취업인구	3.7	4.3	-1.1	-3.1
실업인구	5.3	5.5	4.0	1.4
노동력인구	4.5	5.0	-0.5	-2.5

<표3.34> $\hat{V}_1^{(R)}$에 대한 $\hat{V}_2^{(R)}$, $\hat{V}_3^{(R)}$, $\hat{V}_4^{(R)}$의 상대효율

구 분	백분위 편향(PB_j)		
	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
취업인구	2.13	2.59	2.52
실업인구	1.57	1.71	1.76
노동력인구	2.08	2.56	2.51

$\hat{V}_1^{(R)}$과 $\hat{V}_2^{(R)}$의 편향이 취업인구와 노동력인구에서 $\hat{V}_1$과 $\hat{V}_2$에 비해 훨씬 작아진 사실을 확인할 수 있다. $\hat{V}_3^{(R)}$과 $\hat{V}_4^{(R)}$의 편향도 취업인구와 노동력인구에서 $\hat{V}_3$와 $\hat{V}_4$보다 작은 값을 가지며 실업인구에서는 거의 변화가 발생하지 않았다.

몬테카를로 표본을 이용하여 네 가지 분산추정량들의 비 보정(ratio-adjustment) 추정값들에 대한 95% 신뢰구간을 살펴보았다. 추정값들의 95% 신뢰구간에 대한 포함비율이 다음 <표3.35>에 주어졌다.

<표3.35> 비보정을 갖는 총계 추정값들의 95% 신뢰구간 포함비율

구 분	포함 비율			
	$\hat{V}_1^{(R)}$	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
취업인구	93.6	95.4	94.6	94.2
실업인구	94.3	95.1	95.3	95.0
노동력인구	93.2	95.3	94.6	94.2

취업인구, 실업인구 및 노동력인구의 모든 부분에서 네 가지 분산추정량들의 수행결과가 모두 적합한 것으로 나타났다.

편향의 관점에서는 비 보정 추정값들의 분산추정량들이 서로 상이한 결과를 나타내지는 않는다. 상대효율에서는 $\hat{V}_3^{(R)}$과 $\hat{V}_4^{(R)}$이 $\hat{V}_2^{(R)}$에 비해 효율이 높은 것으로 나타났다. 한편, $\hat{V}_1^{(R)}$의 자유도는 19 이고(각 층별로 1개의 자유도를 가짐) $\hat{V}_3^{(R)}$은 각 PSU가 하나의 반복으로 처리되어 81 개의 자유도를 갖게 된다. 따라서 비 보정 추정값들에 대한 Keyfitz 분산 추정량은 반복 수를 증가시키면 추정량의 안정성을 확보할 수 있다.

④ 반복 수를 갖는 Keyfitz 분산 추정량

Keyfitz 방법의 효율성을 높이기 위해 6 개의 순환 패널들이 반복표본들로 채택되었다. 6 개의 순환 패널을 반복표본으로 이용한 분산 추정값들과 과거 표본설계를 이용한 2 개의 반복표본으로부터 계산된 분산 추정값들이 비교되었다. 순환 패널이 반복표본으로 처리됨에 따라 기인된 중요한 관심 사항은 패널 편향으로부터 발생할 수 있는 분산 추정값들의 증가부분이다. 이러한 부분을 살펴보기 위해 '85년 3월부터 '87년 2월까지의 24개월의 LFS 자료가 이용되었다. 취업인구, 실업인구, 노동력인구의 24개월에 대한 분산 추정값들에 대한 평균과 표준편차를 계산하였다. 2 개의 반복표본과 6 개의 반복표본 하에서 얻어진 분산들에 대한 평균과 표준편차의 비는 24 개의 CMA(Census Metropolitan Area) 지역들의 평균으로 산출하였고 다음 <표3.36>에 주어졌다.

<표3.36> 단순임의 분산추정값의 비교(LFS의 CMA지역 자료)

구 분	분산의평균에 대한 비 평균(average ratio) : 2 vs. 6 반복	분산의 표준편차에 대한 비 평균(average ratio) : 2 vs. 6 반복
취업인구	0.997	1.813
실업인구	0.995	1.515
노동력인구	1.003	1.833

6개의 반복표본을 이용한 분산이 2개의 반복표본의 분산보다 작은 값을 나타낸다. 순환 패널을 반복으로 채택할 경우 분산 추정값들의 편향에 거의 영향을 미치지 않으며, 6개의 반복표본을 이용한 분산이 2개의 반복표본보다는 훨씬 안정적임을 확인할 수 있다. 즉 Keyfitz 방법에서 6개의 순환 패널을 반복으로 사용할 경우 심각한 편향은 발생하지 않으며 2개의 반복표본을 이용했을 경우보다는 효율이 증가됨을 확인할 수 있다.

(다) 비 추정값 탐색을 위한 분산추정방법

① LFS에서 비 추정방법

과거 LFS에서는 사후층화 비 추정방법이 이용되었다. 무응답을 보정하기 위한 일종의 설계 가중치인 부차 가중치가 LFS 목표 모집단의 추정치들에 대해 비 보정되었다. 이러한 비 추정방법은 주 지역의 특성치에 대한 추정의 신뢰도를 높이는 결과를 보였으나 주 내의 소지역들에 대해서는 문제점을 안고 있었다. 주 내의 ER(Economical Region) 지역과 CMA(Census Metropolitan Area) 지역들에 대한 추정의 정확도를 높이기 위해 탐색적인 비 추정 절차가 채택되었다.

탐색적 추정절차는 보정값들의 수열을 통해 수행된다. 먼저 부차가중치가 주 내의 소지역의 인구를 참조하여 보정되며, 이 후 성별-연령대별 범주를 반영한 주 수준의 보정값이 최종 가중치에 적용된다. 이러한 절차는 한번 더 수행되어 한 쌍의 가중치가 추가적으로 생성된다. W_0를 부차가중치라 하고 (W_1, W_2)와 (W_3, W_4)를 2회 반복으로부터 생성된 가중치들의 쌍이라 하자. 노동력 특성값들은 W_4를 이용하여 추정된다. 주 지역의 성별-연령대별 그룹들에서 주변 총계 W_4는 상응하는 그룹들의 외부 인구 추정치와 정확히 일치하나 주 내의 소지역에 대해서는 반드시 그렇지는 않다. 그러나 그 차이는 매우 작게 나타난다.

② 1회 반복 비 추정값에 대한 분산공식

1회 반복 비 추정값들에 대한 분산공식을 유도하면 다음과 같다. 여기에서 적용되는 기본적인 방법론은 선형적인 형태의 부차 가중치를 얻을 때까지 테일러 전개 근사식을 연속적으로 적용하는 방법이다. 세부적인 유도과정을 소개하면 다음과 같다.

$Y^{(0)}$, $Y^{(1)}$, $Y^{(2)}$를 주 지역에서 W_0, W_1, W_2에 근거하여 추정된 노동력 특성값 y의 추정값들이라 하자. 이때 $Y^{(2)}$는 다음 (3.85)식과 같이 주어질 수 있다.

$$Y^{(2)} = \sum_a \frac{Y_a^{(1)}}{P_a^{(1)}} P_a \tag{3.85}$$

여기에서 $Y_a^{(1)}$은 주 지역에서 성별-연령대별 그룹 a에 대한 특성치 y의 W_1가중추정값, $P_a^{(1)}$은 주 지역에서 성별-연령대별 그룹 a에 대한 인구의 W_1가중 추정값, P_a는 주 지역에서 성별-연령대별 그룹 a에 대한 인구의 외부 추정치를 나타낸다.

F_a를 $F_a = Y_a^{(1)}/P_a^{(1)}$이라 할 때, $(E(Y_a^{(1)}), E(P_a^{(1)}))$에서 F_a에 대한 일계 테일러 근사식을 구하면 다음과 같이 주어진다.

$$F_a \doteq \frac{E(Y_a^{(1)})}{E(P_a^{(1)})} + \frac{1}{E(P_a^{(1)})}\{Y_a^{(1)} - E(Y_a^{(1)})\} - \frac{E(Y_a^{(1)})}{\{E(P_a^{(1)})\}^2}\{P_a^{(1)} - E(P_a^{(1)})\}$$

이때 $Y^{(2)}$의 분산에 대한 테일러 근사식은 다음 (3.86)식과 같이 주어질 수 있다.

$$\begin{aligned} V(Y^{(2)}) &= V\left(\sum_a F_a P_a\right) \\ &\doteq V\left\{\sum_a \frac{P_a}{E(P_a^{(1)})}(Y_a^{(1)} - R_{Y_a}^{(1)} P_a^{(1)})\right\} \end{aligned} \tag{3.86}$$

여기에서 $R_{Y_a}^{(1)} = E(Y_a^{(1)})/E(P_a^{(1)})$을 나타낸다.

다음으로 W_1 가중 추정값 $Y_a^{(1)}$과 $P_a^{(1)}$은 W_0 가중 추정값의 항으로 다음 (3.87)식과 같이 나타낼 수 있다.

$$\begin{aligned} Y_a^{(1)} &= \sum_s \frac{Y_{sa}^{(0)}}{P_s^{(0)}} P_s \\ P_a^{(1)} &= \sum_s \frac{P_{sa}^{(0)}}{P_s^{(0)}} P_s \end{aligned} \tag{3.87}$$

여기에서 s는 CMA 또는 ER 지역을 나타내며, P_s는 지역 s의 인구를 나타낸다. (3.87)식의 $Y_a^{(1)}$과 $P_a^{(1)}$을 (3.86)식에 대입하여 W_0 가중 추정값의 비에 대한 일계 테일러 근사식을 구하면 다음 (3.88)식과 같이 주어질 수 있다.

$$\begin{aligned} V(Y^{(2)}) \doteq V[\sum_a \frac{P_a}{E(P_a^{(1)})} \sum_s \frac{P_s}{E(P_s^{(0)})} \{(Y_{sa}^{(0)} - R_{Y_m}^{(0)} P_s^{(0)}) \\ - R_{Y_a}^{(1)} (P_{sa}^{(0)} - R_{P_m}^{(0)} P_s^{(0)})\}] \end{aligned} \tag{3.88}$$

여기에서

$$R_{Y_m}^{(0)} = \frac{E(Y_{sa}^{(0)})}{E(P_s^{(0)})},$$

$$R_{P_m}^{(0)} = \frac{E(P_{sa}^{(0)})}{E(P_s^{(0)})}$$

이다.

위의 (3.88)식은 다음 (3.89)식과 같이 축약된 형태로 다시 표현할 수 있다.

$$\begin{aligned} V(Y^{(2)}) &\doteq V\left\{\sum_s \sum_{h \in s} \sum_{i=1}^{n_h} \sum_a (Z_{Y_{shia}}^{(1)} - R_{Y_a}^{(1)} Z_{P_{shia}}^{(1)})\right\} \\ &= V\left(\sum_s \sum_{h \in s} \sum_{i=1}^{n_h} D_{shi}^{(0)}\right) \end{aligned} \tag{3.89}$$

여기에서

$$D_{shi}^{(0)} = \sum_{a}(Z_{Y_{shia}}^{(1)} - R_{Y_a}^{(1)} Z_{P_{shia}}^{(0)}),$$

$$Z_{Y_{shia}}^{(0)} = \frac{P_a}{E(P_a^{(1)})}\frac{P_s}{E(P_s^{(0)})}(Y_{shia}^{(0)} - R_{Y_a}^{(0)} P_{shi}^{(0)}),$$

$$Z_{P_{shia}}^{(0)} = \frac{P_a}{E(P_a^{(1)})}\frac{P_s}{E(P_s^{(0)})}(P_{shia}^{(1)} - R_{P_a}^{(0)} P_{shi}^{(0)})$$

이고, h는 s에 속하는 층을 나타내며, i는 층 h에서 반복을 나타낸다.

식 (3.89)에서 $\left\{\sum_{i=1}^{n_h} D_{shi}^{(0)}\right\}$는 부차가중치들에 의해 결정되므로 독립성을 가정할 수 있다. 따라서 식 (3.89)은 다음 (3.90)식과 같이 표현될 수 있다.

$$\begin{aligned} V(Y^{(2)}) &\doteq V\left(\sum_{h\in s}\sum_{h}\sum_{i=1}^{n_h} D_{shi}^{(0)}\right) \\ &= V\left(\sum_{h}\sum_{i=1}^{n_h}\sum_{s\in h} D_{shi}^{(0)}\right) \end{aligned} \tag{3.90}$$

여기에서 $\sum_{s\in h}$는 층 h를 포함하고 있는 주 내의 모든 소지역들에 대한 합을 나타낸다. $D_{hi}^{(0)}$를 $D_{hi}^{(0)} = \sum_{s\in h} D_{shi}^{(0)}$와 같이 정의하면, 위의 (3.90)식은 다음과 같이 주어진다.

$$V(Y^{(2)}) \doteq V\left(\sum_{h}\sum_{i=1}^{n_h} D_{hi}^{(0)}\right) \tag{3.91}$$

여기에서 $\left\{\sum_{i} D_{hi}^{(0)}\right\}$는 부차가중치에 의해 결정되므로 이 변수들은 독립성을 가정할 수 있으며, 분산은 다음 식으로부터 추정될 수 있다.

$$\hat{V}(Y^{(2)}) \doteq \sum_{h}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\left(D_{hi}^{(0)} - \overline{D}_h^{(0)}\right)^2 \tag{3.92}$$

여기에서 $\overline{D}_h^{(0)} = \frac{1}{n_h}\sum_{i=1}^{n_h} D_{hi}^{(0)}$이다. 그러나 이 표현식에서는 기대값이 포함되어있고 이러한 값들은 미지의 값이므로 추정값으로 대체하여 분산의 추정값을 근사적으로 계산할 수 있으며 이를 이용한 분산 추정값은 최종적으로 다음 (3.93)식과 같이 주어진다.

$$\hat{V} \doteq \sum_h \frac{n_h}{n_h - 1}\sum_{i=1}^{n_h}\left(D_{hi}^{(2)} - \overline{D}_h^{(2)}\right)^2 \qquad (3.93)$$

여기에서

$$D_{hi}^{(2)} = \sum_{s \in h} D_{shi}^{(2)},$$

$$\overline{D}_h^{(2)} = \frac{1}{n_h}\sum_{i=1}^{n_h} D_{hi}^{(2)},$$

$$D_{shi}^{(2)} = \sum_a \left(Z_{Y_{shia}}^{(2)} - R_{Y_a}^{(2)} Z_{P_{shia}}^{(2)}\right),$$

$$Z_{Y_{shia}}^{(2)} = \frac{P_a}{P_a^{(1)}}\frac{P_s}{P_s^{(0)}}\left(Y_{shia}^{(0)} - \frac{Y_{sa}^{(0)}}{P_s^{(0)}} P_{shi}^{(0)}\right)$$

$$= Y_{shia}^{(2)} - \frac{P_{shi}^{(0)}}{P_s^{(0)}} Y_{sa}^{(2)}$$

$$Z_{P_{shia}}^{(2)} = \frac{P_a}{P_a^{(1)}}\frac{P_s}{P_s^{(0)}}\left(P_{shia}^{(0)} - \frac{P_{sa}^{(0)}}{P_s^{(0)}} P_{shi}^{(0)}\right)$$

$$= P_{shia}^{(2)} - \frac{P_{shi}^{(0)}}{P_s^{(0)}} P_{sa}^{(2)}$$

$$R_{Y_a}^{(2)} = \frac{Y_a^{(1)}}{P_a^{(1)}} = \frac{P_a}{P_a^{(1)}}\frac{Y_a^{(1)}}{P_a} = \frac{Y_a^{(2)}}{P_a}$$

위의 분산 추정공식 (3.93)식은 노동력 특성값들의 W_2 가중 추정값들에 대한 추정공식이며 W_0와 W_2의 두 가중치의 값을 요구한다.

③ 2회 반복 비 추정값의 분산 추정

2회 반복 비 추정값에 대한 분산공식은 앞에서 언급한 내용을 응용하여 테일러 급수전개의 연속적인 적용으로 얻을 수 있다. 그러나 2회 반복에 기인한 분산 추정공식은 매우 복잡한 형태를 취하기 때문에 1회 반복에 기인한 분산 추정공식이 오히려 합리적일 수 있다. 1회 반복 분산 공식은 한 쌍의 가중치 (W_0, W_2)를 이용한다. 여기에서는 (W_0, W_2) 대신에 (W_0, W_4)를 이용하였다. W_2보다는 W_4가 노동력 추정값들의 CV값들에 강한 영향력을 주지 않기 때문이다. 주 지역인 Nova Scotia 지역의 1981년 센서스 자료를 이용하여 몬테카를로 시뮬레이션이 수행되었다. 각각의 몬테카를로 표본에서 LFS 표본설계가 매 단계의 표본 추출을 통해 검증되었다. 1,000개의 몬테카를로 표본들이 독립적으로 추출되었다. 각각의 몬테카를로 표본들에 대해 2회 반복표본 비 추정값$(Y^{(4)})$, 1회 반복 분산 추정량과 이의 CV 추정값을 이용한 분산 추정값$(\hat{V}(Y^{(4)}))$과 95% 신뢰구간$\left(Y^{(4)} \pm 1.96\sqrt{\hat{V}(Y^{(4)})}\right)$을 주 지역과 주 지역 내의 소지역들에 대해 계산하였다. 또한 1,000개의 CV값들의 평균을 계산하여 실제값과 매우 유사한 몬테카를로 CV값들과 비교하였다. 결과는 <표3.37>에 주어졌다. <표3.37>의 모든 셀에 대해서 CV값의 차이는 8% 미만으로 나타났고, 21개의 셀 중 13개의 셀에서는 4% 미만의 값을 갖는다. 한편 신뢰구간의 포함범위는 <표3.38>에 주어졌다. 취업인구와 노동력인구에 대한 95% 신뢰구간의 포함범위는 만족한 값을 보이나 실업인구에 대해서는 다소 낮은 포함범위를 나타내나 여전히 받아들일만한 결과를 보인다.

<표3.37> 1회 반복 분산추정량의 평균 *CV*값과 몬테카를로 *CV*값

구 분	ER 210	ER 220	ER 230	ER 240	ER 250	CMA Halifax	Province (Nova Scotia)
	Average CV's						
취업인구	3.52	3.46	3.14	3.05	1.96	2.01	1.08
실업인구	10.36	12.28	13.13	13.43	10.35	10.55	5.27
노동력인구	2.98	3.17	2.85	2.73	1.77	1.83	0.91
	Monte Carlo CV's						
취업인구	3.48	3.35	2.95	2.86	1.97	1.99	1.11
실업인구	10.90	12.71	13.28	13.37	11.12	11.31	5.59
노동력인구	2.76	3.08	2.76	2.53	1.72	1.74	0.92

<표3.38> 1회 반복 분산추정량에 의한 95% 신뢰구간의 포함범위

구 분	ER 210	ER 220	ER 230	ER 240	ER 250	CMA Halifax	Province (Nova Scotia)
취업인구	94.5	92.8	94.0	94.7	94.7	94.9	92.5
실업인구	92.1	90.7	91.4	91.8	92.7	92.7	93.1
노동력인구	96.2	93.0	93.6	95.2	95.2	96.0	94.0

(라) 요 약

비 보정을 하지 않은 추정값들의 Keyfitz 분산 추정법은 매우 큰 양의 편향을 가지며 효율성도 크게 떨어진다. 반면에 탐색적 비 보정 추정방법은 상대적으로 작은 편향을 가지며 효율성도 크게 향상되는 것으로 확인되었다. 이 논문에서 소개된 비 보정 추정값들에 대한 분산 추정방법들은 무시할 수 있을 정도의 작은 편향을 갖는다. 한편 Keyfitz 방법은 반복수를 증가시킬 경우 다른 분산 추정방법에 비해 효율을 크게 향상시킬 수 있었다. LFS 자료에서 6개의 순환 패널을 반복으로 취급하여 Keyfitz 방법을 적용시켜본 결과 순환 패널 편향에 기인한 Keyfitz 추정분산의 편향은 발생하지 않았다. Keyfitz 방법에 의해 유도된 1회 반복 분산공식은 2회 반복의 탐색적 비 추정값들에 대해서 매우 합리적인 분산 추정값들을 제공하며 신뢰구간에 대한 포함범위도 좋은 특성을 나타낸다.

3.3 영 국

3.3.1 서 론

정부 공식통계는 경제 정책, 재원 분배 및 정책 결정 등에 참고 자료로 이용된다. 대 영역에 대해서는 공식 통계의 정보가 이용자들에게 제공되나 소지역에 대해서는 그렇지 못한 실정이다. 최근 영국 내에서는 소지역 통계 작성에 대한 요구가 꾸준히 제기되고 있고 특히 노동시장 동향에 대한 측도 개발이 시급히 요구되고 있다. 노동력 조사(LFS)는 노동시장 정보 파악에 중요한 역할을 담당하고 있으나 직접조사 추정값들은 소지역 추정에는 한계를 안고 있다. 이 논문은 LFS로부터 소지역 추정의 신뢰도를 향상시킬 수 있는 모형 기반 추정량들을 소개한다.

영국의 LFS는 세 달을 주기로 연속 조사가 실시된다. LFS는 약 60,000 조사가구 단위를 갖는 대규모 조사로써 16세 이상의 약 150,000명의 인구에 대해 조사가 이루어진다. 영국의 LFS는 국제노동기구(ILO)의 요구조건을 만족하도록 표본설계되어 있으며 이를 통해 실업통계가 작성된다. LFS 표본설계에서 표본은 단순임의추출로 추출되며 주로 국가 수준의 추정값을 생산하도록 설계되어 있다. 일년에 한번 소지역 단위인 UA지역(Unitary Authority)과 LAD지역(Local Authority)에대한 추정값들이 작성된다. 이 연구에서 소개되는 내용은 현재 영국 통계국(ONS)에서 진행하고 있는 소지역 추정법과 밀접한 관계가 있다.

소지역 추정법은 소지역에 대한 직접 조사 추정값들이 신뢰성에 문제가 있거나 계산될 수 없을 때 이용할 수 있는 통계적인 기법으로써 인근 지역의 보조정보를 빌려 소지역의 특성값을 추정하는 간접 추계 방법이다. 이 연구에서 이용한 주요 보조정보는 실업보험을 청구한 사람들의 수이다. 실업보험 자료는 행정 시스템에 의해 획득되기 때문에 표본오차가 없고 지역적 범주로 또는 성별-연령대별 범주들로 다양하게 분류될 수 있다. 실업보험 지급 청구자 수와 ILO 실업자 수와는 시기에 따라 약간의

차이는 있지만 강한 상관성을 나타낸다. 시기에 따라 발생하는 차이는 주로 행정 시스템의 변경 또는 경제 사이클의 변화 등에 기인한다. ONS는 Southampton 대학과 연계하여 LFS자료와 실업보험 청구자 수의 자료를 결합하여 소지역 추정값의 신뢰성을 확보할 수 있는 연구를 진행하고 있으며, 특히 UA 또는 LAD 지역에 대한 실업자 수를 추정하는 SPREE 방법(Purcell and Kish, 1980), 로지스틱 모형에 근거한 일종의 변형된 Fay-Herriot 방법(1979)과 Multi-level 모형화 방법(Goldstein, 1995)과 같은 세 가지 추정방법에 대해 연구를 진행하고 있다. SPREE 방법보다는 로지스틱 모형에 근거한 변형된 Fay-Herriot 방법과 Multi-level 모형화 방법이 더 좋은 효율을 나타내며, 여기에서는 변형된 Fay-Herriot 방법에 초점을 맞추어 소개한다.

3.3.2 소지역 추정 방법

앞으로 소개되는 수식에서 첨자 i와 j는 각각 UA 지역과 LAD 지역에 대한 성별-연령대별 그룹을 나타내며, 첨자 g와 h는 각각 UA 지역과 LAD 지역들을 나타낸다. 표본 자료는 LFS 추정값들과 각 지역 내에서 성별 연령대별 그룹으로 분류된 각 셀들에 대한 실업보험 청구자 수의 자료들로 이루어져 있다.

N_{ig}를 셀 (i,g)에서의 인구 총계, U_{ig}를 같은 셀에서의 실업자 총계라 할 때, 이 셀에서의 실업률은 $Z_{ig} = U_{ig}/N_{ig}$이다. 일반적으로 실업률 Z_{ig}는 g번째 지역의 특성값들에 의해 결정된다. Z_{ig}의 기대값과 분산을 각각

$$E(Z_{ig}) = \pi_{ig} \ ,$$

$$Var(Z_{ig}) = \frac{\pi_{ig}(1-\pi_{ig})}{N_{ig}}$$

라 하자. g번째 지역의 특성값들이 $E(Z_{ig})$의 값에 미치는 영향을 열거하기 위해 로지스틱 모형이 이용되었다. 이용된 로지스틱 모형은

$$logit(\pi_{ig}) = x_{ig}^{t}\beta$$

이다. 여기에서 벡터 x_{ig}는 소지역 g에서 i번째 성별-연령대별 그룹에 대한 속성들을 나타내며 알고있는 값이다.

N_{ig}^{*}를 셀 (i,g)에서의 인구 총계에 대한 LFS 추정값이라 하고, U_{ig}^{*}를 실업자 수에 대한 추정값이라 할 때, LFS 실업률 추정값은 $Z_{ig}^{*}=U_{ig}^{*}/N_{ig}^{*}$로 나타낼 수 있다. 성별-연령대별 그룹들이 합리적으로 정의되어 그룹 내에서 추출된 조사단위들에 대한 표본 가중치들에서 변동이 거의 발생하지 않는다고 가정할 수 있다면 셀 (i,g)에서의 실업률에 대한 LFS 추정값들은 표본 실업률로 근사될 수 있다. LFS 표본은 단순임의추출 표본이므로 다음 식들이 성립한다.

$$E(Z_{ig}^{*}|\,Z_{ig})=Z_{ig}\ , \tag{3.94}$$

$$\begin{aligned} V(Z_{ig}^{*}|\,Z_{ig}) &= \frac{N_{ig}-n_{ig}}{N_{ig}-1}\ \frac{Z_{ig}(1-Z_{ig})}{n_{ig}} \\ &= \frac{Z_{ig}(1-Z_{ig})}{n_{ig}^{*}} \end{aligned} \tag{3.95}$$

여기에서

$$n_{ig}^{*}=\frac{n_{ig}^{o}(N_{ig}^{*}-1)}{N_{ig}^{*}-n_{ig}^{*}}$$

이고, n_{ig}는 셀 (i,g)의 LFS 표본크기를 나타낸다. 주어진 Z_{ig}에 대해 Z_{ig}^{*}와 x_{ig}의 독립성을 가정한다면 위의 식들은 다음과 같이 주어질 수 있다.

$$\begin{aligned} E(Z_{ig}^{*}|\,Z_{ig}) &= E\{E(Z_{ig}^{*}|\,Z_{ig},x_{ig})\,|\,x_{ig}\} \\ &= E(Z_{ig}|\,x_{ig}) \\ &= \pi_{ig} \end{aligned} \tag{3.96}$$

$$\begin{aligned} V(Z_{ig}^{*}|\,x_{ig}) &= E\left\{\frac{Z_{ig}(1-Z_{ig})}{n_{ig}^{*}}\,|\,x_{ig}\right\}+\ V(Z_{ig}|\,x_{ig}) \\ &= \frac{\pi_{ig}(1-\pi_{ig})}{n_{ig}^{**}} \end{aligned} \tag{3.97}$$

여기에서

$$n_{ig}^{**} = \frac{n_{ig}^{oo}}{1 + \frac{n_{ig}^{oo} - 1}{N_{ig}^{*}}}$$

이다. 일반적으로 n_{ig}는 N_{ig}에 비해 상대적으로 작은 값을 갖기 때문에 식 (3.96)와 (3.97)은 Z_{ig}^{*}에 대한일종의 근사 이항 로지스틱 모형을 정의하기 위하여 π_{ig}에 대한 로지스틱 항과 결합될 수 있다. 이 모형은 표본크기 $n_{ig}^{o} = round(n_{ig}^{**})$와 표본 실업자 수 $m_{ig} = round(n_{ig}^{o} \times Z_{ig}^{*})$를 입력값으로 갖는 로지스틱 회귀 소프트웨어를 이용하여 실제 표본 자료에 적합될 수 있다. 여기에서 β의 추정값과 $Var(\beta)$의 추정값 $v(\beta)$를 이끌어 낸다. 이때 Z_{ig}의 추정량은 $\pi_{ig} = antilogit(x_{ig}^{t}\beta)$이 된다. 그러나 이 추정량은 불편성을 만족하지는 않는다. 따라서 편의를 보정한 형태의 추정량은 다음 (3.98)식과 같이 주어진다.

$$\pi_{ig} = \pi_{ig}\left\{1 - \frac{1}{2}(1 - \pi_{ig})(1 - 2\pi_{ig})(x_{ig}^{t} v(\beta) x_{ig})\right\} \qquad (3.98)$$

이때 소지역 g에서 실업자 총계에 대한 추정량은 다음 (3.99)식과 같이 주어진다.

$$\theta_{g} = \sum_{i \in g} \alpha_{ig} N_{ig}^{*} \pi_{ig} \qquad (3.99)$$

소지역 g와 h에 대한 모형기반 추정량들 간의 추정 공분산은 다음 (3.100)식을 통해 계산될 수 있다.

$$\widehat{Cov}(\theta_{g}, \theta_{h}) = \sum_{i \in g} \sum_{j \in h} \alpha_{ig} N_{ig}^{*} \pi_{ig}(1 - \pi_{ig})(x_{ig}^{t} v(\beta) x_{jh}) \pi_{jh}(1 - \pi_{jh}) N_{jh}^{*} \alpha_{jh} \qquad (3.100)$$

위에서 언급된 합성추정 형태의 모형기반 추정량은 소지역들 간의 변동을 설명할 수 없는 문제점을 안고 있으며, 일반적으로 이러한 방법으로

추정된 추정오차는 과소 추정되는 경향이 있는 것으로 밝혀지고 있다. 이러한 문제점은 소지역에 대한 랜덤효과를 모형에 반영한 Multi-level 모형을 통해 어느 정도 해소할 수 있으며 이러한 연구가 현재 영국 통계국에서 진행되고 있다. 대상 모형은 $logit(\pi_{ig}) = x_{ig}^t\beta + u_g$와 같은 모형이다. 여기에서 지역 명시 변수 $\{u_g\}$는 평균이 0이고 분산이 σ_u^2인 확률변수로 가정된다. *EBLUP* 형태의 성분 추정값 $\pi_{ig} = antilogit(x_{ig}^t\beta + u_g)$에 기반을 둔 소지역 추정값들은 표본 크기가 큰 UA 및 LAD 지역에서는 LFS 추정값들과 유사하며, 표본 크기가 작은 UA 및 LAD 지역에서는 고정효과를 같는 추정값들과 유사한 경향을 나타낸다. 이러한 추정량이 갖는 실제적인 문제는 추정량의 *MSE* 계산이 쉽지만은 않다는 데에 있다. 현재 영국 통계국에서는 하나의 절충안으로써 다음과 같은 분산 추정공식을 고려하고 있다.

$$\widehat{Var}(\theta_g) = \sum_{i\in g}\sum_{h\in g}\alpha_{ig}N_{ig}^*\pi_{ig}(1-\pi_{ig})\{\sigma_u^2 + x_{ig}^t v(\beta)x_{ig}\} \times \pi_{hg}(1-\pi_{hg})N_{hg}^*\alpha_{hg} \quad (3.101)$$

여기에서 σ_u^2은 랜덤효과 모형 적합에서 추정되는 소지역 간의 추정분산을 나타낸다. 이상에서 고려된 방법론은 추정값의 정확도를 개선시키기는 하나, 모형에 랜덤효과를 포함시키는 문제와 *EBLUP* 형태의 추정량들의 *MSE*를 추정하는 방법 및 비 추정 방법 등은 여전히 해결되어야 할 문제점으로 남게 된다. 이러한 문제를 해결하기 위한 연구가 현재 영국 통계국 및 Southampton 대학 연구진들에 의해 진행되고 있다.

3.4 프랑스

노동력조사로부터 국가 수준의 노동 통계는 일년에 한 번씩 발표되고 있으며, 부차관심영역에서 실업 통계를 생산할 때 일정한 오차 범위 내로 기준을 충족하도록 하는 소지역 추정법에 관한 연구가 진행되고 있다.

프랑스의 표본 설계는 지역별 층화를 기준으로 하였으나 실업자 수 또는 실업률 등의 집적된 자료의 이용 시에는 무응답을 보정하여 국가 수준으로 통계를 작성한 후 성별-연령대별로 분할했으므로 지역 통계의 신뢰성에 대해서 유의해야 한다.

프랑스 통계청에서 취업과 실업에 관한 지역 통계의 생산을 위해서 노동력 조사와 센서스 뿐 만 아니라 실업 보험 신청자료 등 행정 업무 자료를 이용하는 체계를 발전시켜왔으나 소지역이나 세분화된 범주의 통계 작성은 미흡하다. 특히 취업 통계는 연말에 각 회사로부터 취업자 수에 대한 자료를 수집하고 있으나 전년말 기준으로 변화율에 대한 자료로 활용하여 국가 수준에서 비율에 대한 통계를 작성하고 있으며 실업자의 수에 대한 직접 통계를 작성하지는 않는다. 그래서 취업자 통계는 실업 보험과 센서스 자료 등이 핵심적인 역할을 하고 있다. 취업 통계는 해당 익년에 이용가능하다.

실업 통계는 분기별로 지역 통계와 국가 통계를 동시에 발표하며, 주로 이용되는 자료는 노동력 조사와 구직 등록자의 수이다. ILO 기준의 실업자와 실제 구직 등록자 및 노동력 조사에서 실업자와 차이를 반영하기 위해서 노동력 조사에서 추정된 ILO 실업자와 구직 등록자 수의 국가적 비율을 추정하여 활용하고 있다.

경제활동인구는 실업자와 취업자를 합해서 추정하지만 취업자는 직장을 갖고 있는 사람과 가사를 돌보는 사람을 합산해야 하므로 매 해마다 연말에 외삽법으로 조정하여 분기별 통계를 수정하여 시계열 통계로 관리한다.

취업자 통계에서 문제점은 취업자 통계를 작성 시 연말에 센서스 자료를 이용하고 있으나 센서스 간격이 너무 길어서 인구 변동과 상황 변화를 제대로 반영할 수 없다는 점이다 (센서스:1968, 1975, 1982, 1990, 1999). 특히, 지역 취업 통계 작성 시에는 더 심각해질 수 있다.

실업 통계 작성에서 문제점은 노동력 조사에서 국가 단위의 ILO 실업 통계는 작성할 수 있으나 지역 단위 또는 소지역 단위의 실업 통계는 생

산할 수 없다. 또한, 인구 사회학적 구성과 같은 구조적인 특성에서 각 지역이 국가와 동일하다는 전제에서 지역별 실업통계가 작성되며 표준오차는 계산하지 못하고 있다.

취업 통계 작성의 취약점을 보완하기 위해서 "ESTEL"이라는 프로젝트가 진행 중에 있으며 1996년 통계를 시험 중에 있다. 이 프로젝트의 특징은 매년 말을 기준으로 조사자료 및 행정자료를 기반으로 전국 및 지역단위 취업 통계의 질과 신뢰도를 제고하는데 있다. 센서스 자료는 센서스 해에만 적용하고 다른 해에는 연말의 행정 업무와 조사 자료를 기준으로 한다는 것이 큰 특징이다.

핵심적인 과제는 경제 활동 관련 통계의 기준은 센서스 정보이므로 센서스 실시를 일정한 간격으로 실시하는 계획일 것이다. 매 5년마다 센서스에 준하는 대규모 통계 조사 계획을 추진하고 있다.

3.5 일 본

3.5.1 서 론

일본의 노동력조사는 매월 전국의 약 4만 세대, 10만 여명을 대상으로 실시하는 대규모 표본조사로써 노동력관련 특성치들인 취업자 수, 실업률 등은 이러한 월별 노동력조사를 통해 작성된다. 제한된 표본에서 추정결과의 정확도를 높이기 위해 표본설계에서는 국세조사구를 1차 추출단위로 하고, 가구를 2차 추출단위로 하는 층화 2단 추출법이 적용되며 이러한 층화 2단 추출법에 의해 전국의 11개 권역에 대해 조사객체를 선정하고 있다.

1차 추출단위의 추출은 가중치가 부여된 계통추출이고, 2차 추출단위의 추출은 등확률 계통추출이다. 또한 표본교체는 연동교체표본 설계를 따르며, 1차 추출단위인 국세조사구는 5년 마다 새로운 조사구들로 개편된다. 1차 추출단위의 추출틀은 조사구 명부이고, 조사구 명부는 그 특성에 의해서 층화되며 일정 순서에 의해 배열되어 있다. 2차 추출단위의 추출틀은 조사구 내의 가구명부가 이용된다.

월 노동력조사의 표본크기는 1차 추출단위가 약 2,900개 조사구이고, 2차 추출단위가 약 40,000개 가구로 구성된다. 노동력조사 결과에 대해서 시계열 분석을 시행할 경우 가장 많이 적용되는 방법은 전월과의 비교 및 전년 동월과의 비교이다. 일본 노동력조사의 표본설계는 시계열의 정도 향상을 위해 다음과 같은 특성을 갖는 연동교체표본 방식을 채택하고 있다. 첫째, 추출된 표본조사구는 4개월 간 계속해서 조사가 이루어지며 이후 매월 1/4씩 새로운 조사구로 교체된다. 즉, 추출된 표본조사구를 4개로 나누어 3/4이 전월과 같은 조사구, 1/4이 새로운 조사구로 대체되어 조사가 이루어진다. 둘째, 표본조사구는 4개월 간의 조사를 마친 후 8개월 간 조사에서 분리되어 익년 동기에 재 조사된다. 셋째, 표본조사구는 4개월 간 계속해서 조사를 시행하지만 조사구 내의 조사가구는 전반의 2개월과 후반의 2개월로 나누어 조사객체(조사가구)를 교체한다. 따라서 1차 추출단위로써 추출된 표본조사구는 4개월 간 계속해서 조사가 이루어지고, 이 기간이 끝나면 다른 표본조사구로 교체되지만 다음 해의 동기간에는 앞의 표본조사구에서 다시 조사가 이루어진다. 표본조사구의 교체는 일제히 이루어지지 않고, 전체를 A, B, C, D의 4개로 구분하고 매월 1구분씩 교체가 이루어진다. 2차 추출단위로써 추출된 가구는 2개월 간 계속해서 조사가 이루어지지만 표본조사구의 조사기간 4개월 중 전반의 2개월과 후반의 2개월로 가구를 교체한다. 따라서 추출된 가구에 거주하는 세대는 같은 가구에 거주하고 있으면 2개월 간 계속해서 조사가 이루어지고 또 다음 해의 같은 기간에 다시 2개월 간 계속해서 조사가 이루어진다.

일본 노동력조사의 표본설계 변천과정을 살펴보면 다음과 같이 요약될 수 있다. 일본 노동력조사는 매월 만 15세 이상의 노동력인구를 대상으로 1946년 9월에 시험적으로 개시되었다. 이때의 노동력조사는 1차 추출단위가 시, 군 단위, 2차 추출단위는 구, 동, 읍면 단위, 3차 추출단위는 세대단위인 3단 추출법에 의해 추출된 약 15,000세대의 50,000여명의 고정표본에 대해 이루어졌다. 1948년 10월에는 2차 추출단위를 상주인구조사구로 조정하여 약 1,000조사구, 16,000여 세대의 56,000여명의 표본에 대해 조사

가 이루어졌으며, 표본조사구를 4개월 마다 일제히 교체하는 표본교체방식이 처음으로 실시되었다. 1950년 6월 노동력조사부터 2차 추출단위가 국세조사구로 개편되어 약 1,000개 조사구, 16,000여 세대의 51,000여명의 표본에 대해 조사가 이루어졌다. 1952년 11월에는 기존의 3단 추출법이 2단 추출법으로 변경되었다. 1차 추출단위는 1950년의 국세조사구, 2차 추출단위는 세대로 하여 추출된 약 1,000개 조사구, 11,000여 세대의 50,000여명의 표본에 대해 조사가 이루어졌고, 추출된 조사구는 3개월 간 계속 조사가 이루어진 후 매월 1/3의 조사구가 갱신되는 표본교체방식이 적용되었다. 1961년 10월부터 노동력조사 표본이 확대 개편되었다. 15세 이상의 노동력 인구를 대상으로 약 2,000개 조사구, 25,000여 세대의 70,000여명의 표본에 대해 조사가 이루어졌다. 추출된 표본조사구는 4개월 간 계속하여 조사가 이루어지고 매월 1/4의 조사구가 교체되는 표본교체방식이 채택되었고, 현재의 표본교체방식은 이 시기의 표본교체방식을 따르고 있다. 1982년 10월, 노동력조사에 대한 표본확대 개편이 시행되었다. 대영역을 기반으로 하는 기존의 표본설계 방식이 지역별 노동력 관련 통계를 생산하기위한 체제로 확대 개편되었으며, 약 2,900개 조사구, 40,000여 세대의 100,000여명의 표본에 대해 노동력조사가 실시되었다. 현재의 노동력조사의 표본크기 및 표본교체방식은 이 시기의 표본설계에 기반을 두고 있다.

3.5.2 층화 및 추출단위 구성

일본의 노동력조사에서는 국세조사의 조사구를 1차 추출단위로 한다. 1개 조사구는 약 50세대로 구성되며 전국에 걸쳐 중복되지 않게 설정되어 있다. 95년 국세조사의 조사구 수는 전국적으로 약 88만개이다. 국세조사구를 추출단위로 사용하면 국세조사구 명부를 추출틀로 이용할 수 있고 그 밖에 국세조사의 조사결과를 이용해서 조사구를 그 특성에 의해 분류하는 일이 용이하게 될 수 있는 이점도 있다. 또한 조사구가 특별시, 광역시, 도 및 시, 구, 동/읍면의 행정구획 중에서 설정되어 있는 것도 조사

시 이점으로 작용하고 있다. 2차 추출단위는 표본조사구 내에 있는 주택과 건물의 각 호에서 1세대가 거주할 수 있도록 되어있는 건물 또는 건물의 일구획으로 정의된다. 이와 같이 정의된 2차 추출단위를 가구라고 한다. 2차 추출단위에 대한 결정규칙은 다음 <표3.39>와 같다.

<표3.39> 2차 추출단위의 결정규칙

건물의 종류	추출단위의 결정 규칙
단독주택	건물전체를 하나의 추출단위로 함.
연립주택, 아파트	하나의 세대가 사용하도록 만들어진 각각의 구획을 추출단위로 함.
기숙사 (하숙집, 식당, 수용시설, 간이 숙박소 등을 포함)	기숙인 등의 각 거주실 및 가구주와 관리인 등의 세대 거주부분을 추출단위로 함. 식당, 수용시설 등에서 1실에 다수(약 10인 이상)의 거주자를 수용할수 있는 경우에는 1실을 5인 이하가 되도록 분할한 것을 하나의 추출단위로 함.
여관, 호텔	전 객실을 합쳐서 하나의 추출단위로 함
학교, 공장, 사무소 등	고용원 등의 세대가 거주할 수 있도록 되어있는 부분을 추출단위로 함.
병원, 요양소 등	입원환자의 각 병실 및 의사와 간호사 등의 거주부분을 각각 하나의 추출단위로 함. 병원 등의 1실에 다수의 입원환자가 들어가는 경우는 식당, 수용시설 등과 같이 분할한 것을 하나의 추출단위로 함.
주인집에서 숙식하는 사용인의 방이 3개 이상 있는 상점, 여관 등	사업주의 거주부분과는 별도로 주인집에서 사는 사용인의 각 방을 추출단위로 함. 사용인의 방이 3개 미만의 경우는 사업주의 세대에 포함함.
바깥채, 헛간 등의 부속 건물	주건물에 포함해서 하나의 추출단위로 함.
빈집, 공사 중의 건물 (사람이 전혀 살 수 없는 건물은 제외)	사람이 살고있는 건물에 준해서 추출단위로 함.

* 빈집, 공사 중의 건물도 추출단위로 하는 이유는 표본조사구를 4개월 간 계속해서 조사하는 것으로 하고 있기 때문에 2차 추출단위의 각 명부작성 시점에 거주자가 있지 않아도 조사시점에서는 거주자가 있을 가능성이 있기 때문임.

일본 노동력조사의 표본설계에서는 1차 추출단위인 조사구의 층화에

중점을 두고 2차 추출단위인 가구에 대해서는 추출단위 명부 상의 배열을 고안하여 층화에 준한 효과를 얻고 있다. 더욱이 조사구의 층화는 전국을 11개 권역으로 나누고 각각의 권역에서 층화가 이루어지며, 중요도가 높은 통계항목인 산업별, 종사지위별 취업자 수에서 추정결과의 정도가 높도록 조사구의 산업별, 종사지위별 취업자 구성을 층화의 기준으로 이용하고 있다. 이밖에 기숙사, 합숙소, 병원, 요양소, 사회시설 등에 거주하고 있는 사람의 취업상태가 조사구의 특징을 결정하는 경우가 많기 때문에 조사구에 대한 주거형태도 층화기준에 더해진다. 일본 노동력조사의 표본설계에서 층화에 대한 구체적인 방법은 다음과 같다.

o 층부호 01

조사의 범위에서 제외되고 있는 주둔군 지역 및 추출대상에서 제외되고 있는 형무소, 구치소 등이 있는 구역, 자위대 구역 및 수면 조사구는 종합해서 하나의 층으로 한다.

o 층부호 02, 03

사람이 살고 있지 않은 무인 조사구나 환산세대수가 15이하의 조사구에 대해서는 층화의 효과가 적기 때문에 산업별, 종사지위별 층화는 실시하지 않고 각각을 통합해서 하나의 층으로 한다. 환산세대수는 다음과 같이 산출한다.

$$\text{환산세대수} = (\text{세대인원이 2인이상인 보통세대수}) + \frac{1}{3}\{(\text{세대인원이 1인인 보통세대수}) + (\text{준세대 인원})\}$$

보통세대는 주거와 생계를 함께하고 있는 사람들의 모임 또는 가구를 이루고 살고 있는 단독자를 말하며, 준세대는 자취, 하숙, 독신숙소의 단독자를 말한다. 위 식에서 세대인원이 1인인 보통세대수와 준세대 인원의 합을 3으로 나누는 것은 세대인원이 2인 이상인 일반세대의 15세 이상의 세대인원이 평균 3명을 가정한것에 기인한다.

o 층부호 04

독신직원이 약 50인 이상 거주하는 기숙사, 합숙소가 있는 구역에 대

해서는 조사구 내에서 기숙사, 합숙소마다 단위구를 설정한다. 단위구마다 산업특성이 다르기 때문에 산업별 인구의 층 내 조사구 간의 분산을 줄이기 위해 층을 더욱 세분한다. 층의 세분화 과정에서 종사지위는 층화의 지표로 하지는 않는다. 학교의 독신학생이 약 50인 이상 거주하는 기숙사, 합숙소가 있는 조사구 및 사회시설이나 큰 병원이 있는 구역에 대해서는 시설의 종류에 따라 층을 설정한다.

o 층부호 05~24

환산세대수가 16이상의 조사구는 조사구 내의 15세 이상 인구에 대한 산업별, 종사지위별 취업자 수의 비율에 의해 층화한다.

이상의 방법에 의해 설정된 조사구의 층화기준은 <표3.40>과 같다.

<표3.40> 조사구의 층화기준

분류부호		분 류 기 준
대분류	소분류	
01		후치번호가 5(형무소, 구치소 등이 있는 지역), 6(자위대 구역), 7(주둔군 지역), 9(수면조사수)인 조사구
02		인구가 0인 조사구
03		환산세대수가 15이하인 조사구
04		후치번호가 4(사회시설, 큰 병원이 있는 구역)인 조사구
		후치번호가 8(50인 이상의 단독자가 거주할 수 있는 기숙사, 합숙소 등이 있는 구역)인 조사구
		후치번호가 4와 7이외의 환산세대 중 점유할 수 있는 주택의 일반세대 비가 0.75이상인 조사구
	01	학생의 합숙소, 기숙사(단, 50인 이상의 세대)가 있는 표본단위구
	02	병원, 요양소(단, 50인 이상의 세대)가 있는 표본단위구
	03	사회시설(단, 50인 이상의 세대)이 있는 표본단위구
	04	후치번호가 4인 주택, 상기의 어디에도 속하지 않는 표본단위구
	11	합숙소 등에 거주하는 광업의 취업자가 50인 이상의 표본단위구
	12	광업의 세대비가 0.6이상의 표본단위구
	21	합숙소 등에 거주하는 건설업의 취업자가 50인 이상의

		표본단위구
	22	건설업의 세대비가 0.6이상의 표본단위구
	31	합숙소 등에 거주하는 제조업의 취업자가 50인 이상의 표본단위구
	32	제조업의 세대비가 0.6이상의 표본단위구
	41	합숙소 등에 거주하는 도소매업, 음식점의 취업자가 50인 이상의 표본단위구
	42	도소매업, 음식점의 세대비가 0.6이상의 표본단위구
	51	기숙사 등에 거주하는 금융·보험업, 금융업의 취업자가 50인 이상의 표본단위구
	52	광업의 세대비가 0.6이상의 표본단위구
	61	합숙소 등에 거주하는 전기, 가스, 열공급, 수도업, 운수 및 통신업의 취업자가 50인 이상의 표본단위구
	62	전기, 가스, 열공급, 수도업, 운수 및 통신업의 취업자의 세대비가 0.6이상의 표본단위구
	71	합숙소 등에 거주하는 서비스업의 취업자가 50인 이상의 표본단위구
	72	서비스업의 세대비가 0.6이상의 표본단위구
	81	합숙소 등에 거주하는 공무의 취업자가 50인 이상의 표본단위구
	82	공무의 취업자의 세대비가 0.6이상의 표본단위구
	91	후치번호가 8인 조사구가 있는 주택, 상기의 어디에도 속하지 않는 표본조사구
	92	급여주택에 살고 있는 일반세대수의 비가 0.75이상의 조사구가 있는 주택, 상기의 어디에도 속하지 않는 표본조사구
05		광업의 취업자의 비가 0.1이상의 조사구
06		어업의 취업자의 비가 0.2이상의 조사구
07		어업의 취업자의 비가 0.1이상-0.2미만의 조사구
08		건설업, 제조업의 업주의 비가 0.1이상의 조사구
09		도소매업, 음식점의 업주의 비가 0.1이상의 조사구
10		전기, 가스, 열공급, 수도업, 운수, 통신업, 금융, 보험업, 부동산업, 서비스업의 업주의 비가 0.1이상의 조사구
11		농림업의 취업자의 비가 0.3이상의 조사구
12		농림업의 취업자의 비가 0.1이상-0.3미만의 조사구
13		공무의 취업자의 비가 0.1이상의 조사구
14		금융, 보험업, 부동산업의 고용자의 비가 0.1이상의 조사구
15		제조업의 고용자의 비가0.3이상의 조사구
16		건설업 고용자의 비가 0.1이상의 조사구

17		도소매업, 음식점의 고용자의 비가 0.2이상의 조사구
18		서비스업의 고용자의 비가 0.2이상의 조사구
19		전기, 가스, 열공급, 수도업, 운수, 통신업의 고용자의 비가 0.1이상의 조사구
20		제조업의 고용자의 비가 0.2이상-0.3미만의 조사구
21		제조업의 고용자의 비가 0.1이상-0.2미만의 조사구
22		도소매업, 음식점 고용자의 비가 01.이상-0.2미만의 조사구
23		가스업의 고용자의 비가 0.1이상-0.2미만의 조사구
24		상기의 어디에도 속하지 않은 조사구

*단독세대는 사회시설, 큰 병원 및 약 50인 이상의 단독자가 거주하고 있는 기숙사, 합숙소 등을 말한다.

*층부호 04에 속하는 조사구에 대한 단위구의 설정

-후치번호가 4 또는 8의 조사구로 단독세대에 속하는 인원이 50인 이상이 되는 단독세대가 있는 경우는 단독세대별로 단위구를 설정한다.

-후치번호가 4, 8이외의 조사구에 대해서는 그 조사구 전체를 하나의 단위구로 한다.

*층부호 0411, 0421, 0431, 0441, 0451, 0461, 0471, 0481의 단위구는 후치번호가 8이다.

*층부호 0412, 0422, 0432, 0442, 0452, 0462, 0472, 0482의 단위구는 후치번호가 8이외이고 환산세대수 중에 차지하는 각 해당 산업 종사 세대주의 비가 0.6 이상의 단위구이다..

3.5.3 표본배분, 추출, 교체방식

일본 노동력조사에서는 조사구 추출 시 확률비례추출을 적용한다. 각 조사구 추출에 대한 가중치 산정은 다음공식을 적용한다.

$$\text{가중치} = \frac{\text{산출세대수}}{15}$$

층부호 01, 02의 조사구에 대해서는 가중치 1이 적용되며, 2차 추출단위의 추출율은 가중치의 역수를 사용한다.

노동력조사에서는 전국 추정결과를 산출하는 것을 주 목적으로 하며 10개 지역별 추정결과는 이를 기반으로 사분기마다 작성된다. 지역별 추정결과에 대한 목표정도를 만족할 수 있도록 다음과 같은 기준에 의해 조사구 수가 배분되고 있다.

o 최소지역(시코쿠 지역)의 표본조사구 수

인구 규모가 가장 작은 시코쿠 지역의 표본조사구 수는 추정결과의 정도를 일정하게 유지할 수 있도록 표본조사구를 배분하며, 표본조사구의 추출은 8개 조의 부표본마다 행하기 때문에 부표본의 정수배로 조정하여 152개의 표본조사구를 배분한다.

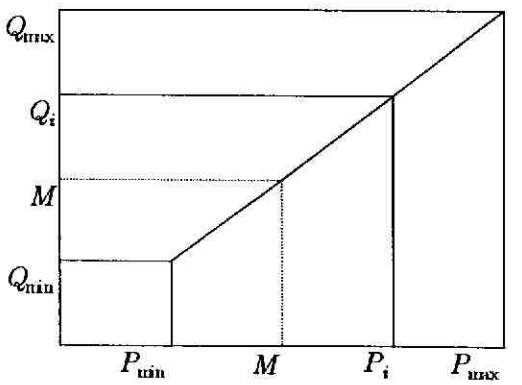
o 오끼나와 지역의 표본조사구 수

오끼나와 지역의 도별 추정결과의 정도를 일정하게 유지할 수 있도록 부표본의 정수배인 144개 표본조사구를 배분한다.

o 시코쿠 지역 이외지역에 대한 표본조사구 수

시코쿠 지역에 배분한 152개 조사구와 오키나와 지역에 배분한 144개 조사구를 제외한 나머지 2,584개 조사구는 선형변환법에 의해 인구비례적으로 각 지역별로 부표본의 정수배가 되도록 배분한다. 선형변환법은 인구비에 의한 배분값 P_i를 다음의 일차식에 의해 재배분하는 방법이다.

$$Q_i = \frac{M - Q_{\min}}{M - P_{\min}} (P_i - P_{\min}) + Q_{\min}$$



여기에서 N은 "전국의 표본조사구 수", M은 "1개 지역 당 평균 조사구 수($= N/10$)", P_i는 "i지역의 비례배분값", $P_{\min}$은 "최소지역의 비례배분값", Q_i는 "변환 후 i지역의 표본조사구 수", $Q_{\min}$은 "최소지역의 표본조사구 수"를 나타낸다.

지역별 표본조사구 수는 다음 <표3.41>과 같다. 각 지역별 표본조사구 수는 각 층의 조사구의 가중치 합계에 따라 비례배분하고 부표본의 정수배로 조정해서 층별 표본조사구 수를 결정한다. 조사구의 가중치 합계가 현저하게 작은 층은 특성이 유사한 층과 합병한다. 단, 층부호가 02, 03, 0401, 0402, 0403, 0404의 각 층은 서로 상이한 속성을 갖기 때문에 합병을 시행하지 않고 각각 독립으로 표본조사구의 추출을 시행한다.

<표3.41> 지역별 표본조사구 수

지 역	특광역시 및 도	조사구 수
홋카이도	홋카이도	176
토오호쿠	아오모리, 이와테, 미야기, 아키타, 야마가타, 후쿠시마	232
미나미칸토	사이타마, 치바, 토쿄, 카나가와	576
기타칸도・코오신	이바라키, 토치기, 군마, 야마나시, 나가노	232
호쿠리쿠	니이가타, 토야마, 이시카와 후쿠이	168
토오카이	기후, 시즈오카, 아이치, 미에	304
킨키	시가, 쿄토, 오사카, 효고, 나라, 와카야마	400
츄우고쿠	톳도리, 시마네, 오키야마, 히로시마, 야마구치	208
시코쿠	토쿠시마, 카가와, 에히메, 코오치	152
큐슈	후쿠오카, 사가, 나가사키, 구마모토, 오이타, 미야자키, 가고시마	288
오키나와	오키나와	144
계		2,880

일본 노동력조사에서 표본조사구의 추출은 추출작업의 간편성, 정도의 향상 및 표본의 지역별 분산 등을 고려하여 계통추출법을 채택하고 있다. 계통추출법은 추출단위를 일정의 순서로 배열하고 이것에 추출용 일련번호를 붙여 추출기부호(계통추출의 출발점이 되는 부호)와 추출간격에 의해 추출용 일련번호를 순차적으로 추출하는 방법이다. 표본조사구는 각 지역별, 지역 내에서는 층별, 층 내에서는 8개 조의 부표본별로 독립적으

로 추출된다.

각 층 내에서 조사구의 배열은 "합병 전의 층"→"특·광역시 및 도"→"시·구·동/읍, 면"→"조사구 번호의 주 번호"→"조사구 번호의 단위구 번호"의 순으로 한다. 각 층별 조사구들에 대한 가중값의 누적합은 j조사구의 가중값을 W_j라 할 때 위의 조사구 배열 순서에 따라 $A_1 = W_1$, $A_2 = A_1 + W_2$, …, $A_N = A_{N-1} + W_N$ (N은 조사구 총수)와 같이 산출한다. 위식에서 A_j를 "추출용 층내 누적부호"라 부른다. 층별 추출간격은 다음 식으로 계산한다.

$$i\text{층의 추출간격} = \frac{i\text{층 조사구의 가중값 합계}}{i\text{층의 표본조사구수}/8}$$

단, 층부호가 02, 0401, 0402, 0403, 0404인 6개의 각 층에 대해서는 "추출간격=조사구의 가중값 합계"로 추출간격이 결정된다.

<표3.41>에서 홋카이도 지역에 대한 층별 표본조사구 수 및 관련값들을 살펴보면 다음 <표3.42>와 같다.

<표3.42> 홋카이도 지역의 층별 표본조사구 등 일람표

합병전층부호		95년 국세조사		비례배분값	합병후층부호		표본조사구수	추출간격	추출기부호								교체조사구추출간격
		조사구수	가중값						AK	AL	BK	BL	CK	CL	DK	DL	
02		502	502	0.62	02		1	502						224			41
03		3098	3098	3.83	03		3	3098			699				777	2493	258
04	01	121	336	0.42	04	01	1	336	127								28
04	02	255	930	1.15	04	02	1	930					640				77
04	03	356	920	1.14	04	03	1	920				61					76
04	04	117	243	0.30	04	04	1	243		234							20
06		732	2149	2.66													
07		461	1387	1.71													
12		1208	3458	4.27	06		16	6264	1069	1945	1963	2239	2771	3150	4002	5583	522
11		2440	5534	6.84													

04	31	14	38	0.05	04	31	16	5319	126	490	1567	1784	2723	2838	3926	4972	443
04	32	1	2	0.00													
15		150	475	0.59													
20		234	764	0.94													
21		2669	9265	11.45													
04	11	0	0	0.00													
04	12	0	0	0.00													
05		27	95	0.12													
04	21	1	2	0.00	04	21	32	6322	707	762	888	1377	2750	3417	4146	5534	526
04	22	0	0	0.00													
16		7407	25282	31.25													
08		2	6	0.01													
04	41	1	5	0.01	04	41	8	10011	657	1430	1703	4802	5242	6950	7842	7862	834
04	42	0	0	0.00													
17		2832	8445	10.44													
09		594	1561	1.93													
22		*10982*	*39256*	*48.53*	*22*		*48*	*6542*	*242*	*1137*	*1706*	*1987*	*4297*	*4331*	*5568*	*6483*	*545*
04	71	17	45	0.06	04	71	16	7425	108	1371	1968	3915	4407	4657	5072	5932	618
04	72	2	5	0.01													
18		4560	14801	18.3													
23		3191	10978	13.57	10		16	5577	149	620	857	1361	1474	2378	4429	4760	464
10		71	177	0.22													
04	51	1	2	0.00	04	51	8	4539	288	435	1518	2387	2530	2771	3389	4152	378
04	52	0	0	0.00													
14		369	1078	1.33													
04	61	5	12	0.01													
04	62	0	0	0.00													
19		1016	3447	4.26													
04	81	2	5	0.01	04	81	8	8082	29	444	1477	1752	2923	5213	5703	6587	673
04	82	1	2	0.00													
13		1570	5375	6.64													
04	91	48	106	0.13													
04	92	3	6	0.01													
24		863	2588	3.20													
계		45923	142380	176			176										

여기에서 22층에 대한 추출부호 산출과정을 세부적으로 살펴보면 다음

<표3.43>과 같다.

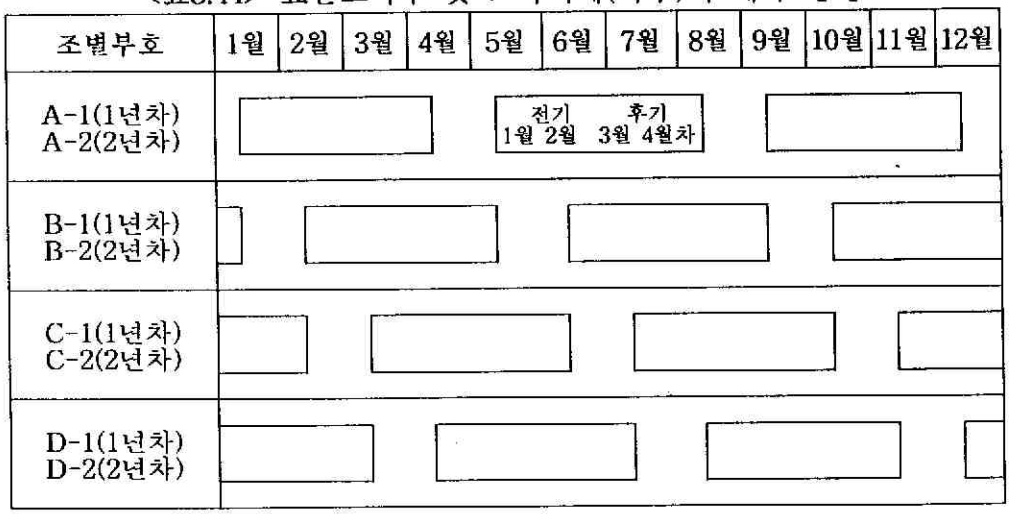
<표3.43> 홋카이도의 22층에서 추출부호의 산출

구 분		부 표 본							
		AK	AL	BK	BL	CK	CL	DK	DL
가중값 합계		39256							
표본조사구수		6 (48÷8)							
추출간격(F)		6542.66 (39256÷8)							
추출기번호(G)		242	1137	1706	1987	4297	4331	5568	6483
추출번호	$S_1 = G$	242	1137	1706	1987	4297	4331	5568	6483
	$S_2 = S_1 + F$	6785	7680	8249	8530	10840	10874	12111	13026
	$S_3 = S_2 + F$	13328	14223	14792	15073	17383	17417	18654	19569
	$S_4 = S_3 + F$	19871	20766	21335	21616	23926	23960	25197	26112
	$S_5 = S_4 + F$	26414	27309	27878	28159	30469	30503	31740	32655
	$S_6 = S_5 + F$	32957	33852	34421	34702	37012	37046	38282	39198

2차 추출단위인 가구조사 단위에 대한 추출은 추출단위 명부를 이용하여 계통추출한다. 이 경우 추출률은 조사구별로 정해져 있는 가중값의 역수를 이용한다.

노동력조사의 표본설계에서 표본교체방법을 살펴보기로 한다. 처음 추출한 표본조사구를 기준조사구라 하고, 소정의 조사기간을 마치고 교체되는 조사구를 교체조사구라 하자. 원칙적으로 기준조사구가 표본으로 선정되면 4개월 간 조사된 후 익년 동일 시기에 포함시키며 4개월 간 조사하고 표본에서 삭제한다. 교체조사구의 추출은 기준조사구의 추출용일련번호를 시작번호로 하고 교체조사구 추출간격을 이 번호에 더하는 계통추출법을 적용하며 구체적인 내용은 <표3.44>와 같다. 표본은 8개 조의 부표본으로 구성되어 있고, A, B, C, D의 구분은 조사를 새로 시작하는 달에 따라 붙여진 것이며, 후치번호 1은 금년에 새로 표본조사구에 포함된 조사구, 후치번호 2는 작년에 표본조사구로 포함되어 금년에 재조사한 후 내년에는 조사에서 삭제되는 조사구를 나타낸다.

<표3.44> 표본조사구 및 조사객체(가구)의 계속 상황



보다 구체적으로 각 조별 교체시기를 살펴보면 <표3.45>와 같다.

<표3.45> 조별 교체시기 현황

조별 부호		조사를 새로 시작하는 달
1년차 조사구	2년차 조사구	
A-1	A-2	1월 5월 9월
B-1	B-2	2월 6월 10월
C-1	C-2	3월 7월 11월
D-1	D-2	4월 8월 12월

A조에 속하는 조사구는 5월에 조사를 새로 시작해서 4개월 간 조사를 실시하고 새로운 A조의 조사구를 9월에 교체조사구로 추출하여 표본조사구로 포함시킨다. 그러나 5월에 조사를 시작한 조사구에 속하는 표본가구들은 5, 6월의 전반 2개월(전기) 조사분과 7,8월의 후반 2개월(후기) 조사분으로 분리 추출하여 조사한다. 즉 2차 추출단위인 표본가구들은 2개월 간 계속 조사한 후 1년 간 쉬었다가 익년의 같은 달에 조사된다. 하나의 기준조사구에 대응되는 교체조사구는 예비조사구를 포함하여 11개 조사구

를 추출하며 따라서 추출간격은 기준조사구 추출간격의 1/12로 한다. 2차 추출단위인 표본가구의 교체는 표본조사구 내에서 가구단위를 추출할 때 전기분과 후기분으로 나누어 별도의 독립적인 추출시작번호에 의해 계통추출한 표본가구들을 반씩 나누어 표본에서 삭제시키고 새로운 표본가구를 포함시킨다.

3.5.4 추정방법

표본조사 결과를 도출해 내는 가장 통상적인 방법은 각 조사결과에 추출률의 역수를 곱하여 합산한 후 추정값을 구하는 방법이다. 이와 같은 방법에 의해 구해진 추정값을 선형추정값이라 한다. 표본조사는 미지의 정보를 표본으로부터 획득하기 위해 시행된다. 이 경우 추정대상이 되는 목적항목과 연관성이 있는 보조정보가 있을 때 이것을 표본설계에 반영하면 조사결과의 정보량이 많아지고, 또한 결과의 정도도 향상될 수 있다. 일본 노동력조사에서는 목적항목에 대한 추정방법으로 보조정보를 이용한 다음과 같은 비추정식을 이용한다.

$$\text{목적항목의 비추정식} = \text{목적항목의 선형추정식} \times \frac{\text{보조항목의 참값(벤치마크)}}{\text{보조항목의 선형추정값}}$$

목적항목과 보조항목의 관련성은 목적항목의 종류에 따라 다르지만 양자의 상관성이 높을 때 추정결과의 정도는 좋아진다. 일본 노동력조사에서는 보조항목으로써 조사인구를 채택하고 있다. 노동력조사의 목적항목 및 보조항목의 선형추정값 X는 다음 일차식에 의해 구해진다.

$$\hat{X} = \sum_{h=1}^{10} \hat{X}_h = \sum_{h=1}^{10} \sum_{i=1}^{L_h} \hat{X}_{hi} \ ,$$

단,

$$\hat{X}_{hi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} \frac{1}{P_{hij}} r_{hij} \frac{1}{\left(\frac{1}{f_{hij}}\right)} X_{hij}$$

$$= \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} \frac{W_{hi}}{w_{hi}} r_{hij} f_{hij} X_{hij}$$

$$= F_{hi} \sum_{j=1}^{m_{hi}} r_{hij} X_{hij}$$

여기에서

h = 지역 번호$(1, 2, \cdots, 10)$,

i = 층 번호$(1, 2, \cdots, L_h)$,

j = 표본조사구 번호$(1, 2, \cdots, m_{hi})$,

$\hat{X}$ = 속성 X의 선형추정값,

$\hat{X}_h$ = h지역에 대한 속성 X의 선형추정값,

L_h = h지역의 층의 수,

$\hat{X}_{hi}$ = h지역의 i층에 대한 속성 X의 선형추정값,

m_{hi} = h지역에서 i층의 표본조사구 수,

W_{hi} = h지역의 i층에 대한 모든 조사구의 가중값 합계,

F_{hi} = h지역의 i층에 대한 선형추정 승률$(= W_{hi}/m_{hi})$,

X_{hij} = h지역의 i층에서 j표본조사구에 대한 속성 X의 조사인구,

P_{hij} = h지역의 i층에서 j표본조사구의 추출확률,

w_{hij} = h지역의 i층에서 j표본조사구의 가중값,

r_{hij} = h지역의 i층에서 j표본조사구의 보정률(조사구 세부 분할에 따른 변경 추출률),

f_{hij} = h지역의 i층에서 j표본조사구의 조사구 내 추출률의 역수 $(f_{hij} = w_{hij})$.

일본 노동력조사에서 목적항목에 대한 추정값은 다음 그림과 같은 과정을 통해 산출된다.

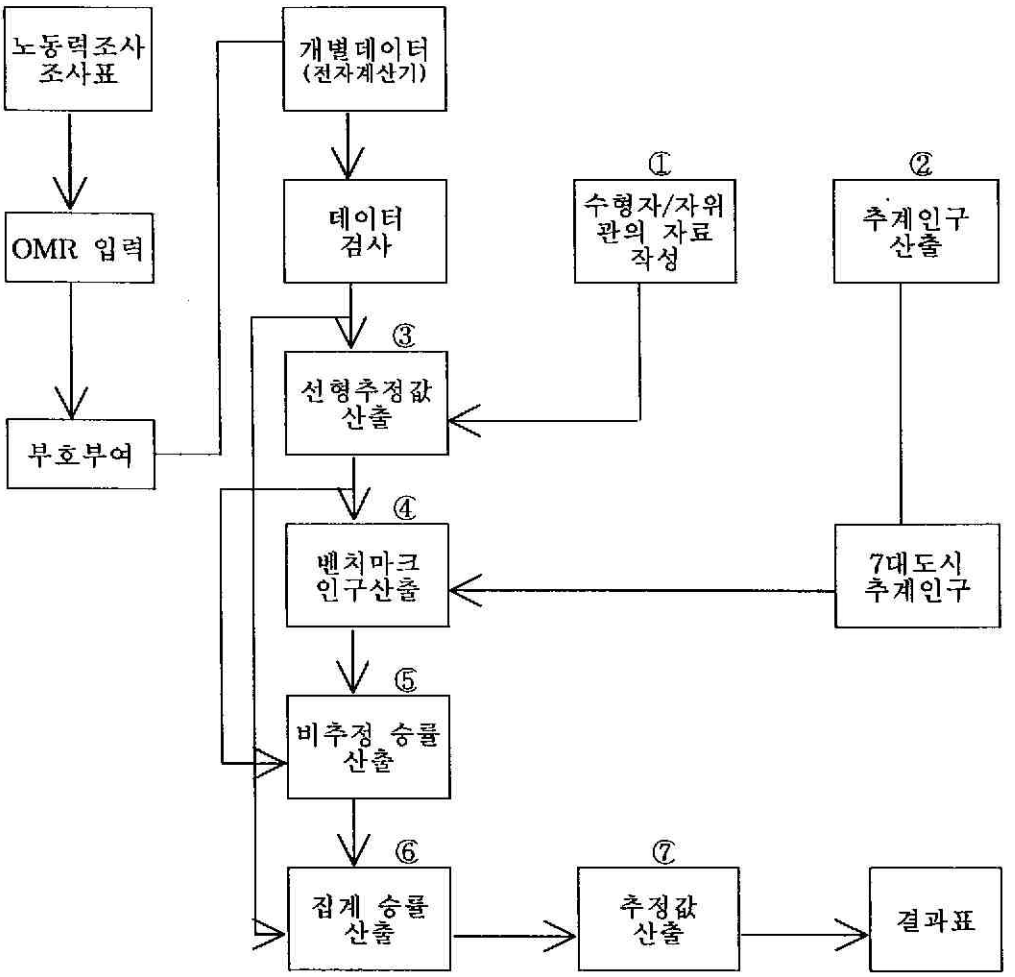


그림3.1 추정값 산출 순서

① 수형자 및 자위관의 자료 작성

노동력조사에서는 층화 2단추출법에 의해 조사가구가 추출된다. 그러나 형무소 등의 교정시설 수용자(수형자) 및 자위대의 당사 내 거주자(자위관)에 대해서는 조사원에 의한 조사가 곤란하므로 법무성 및 방위청으로부터 자료를 얻어 통계청에서 직접 직접하고 있다.

② 추계인구의 산출

국세조사에 의한 인구를 기준인구로 하고 매월 인구동태 통계의 출생아 수, 사망자 수, 출입국관리 통계의 출국자 수, 입국자 수의 통계 숫자을 가감하여 매월 말일 현재 전국의 남녀, 연령계급별 추계인구를 산출하고 있다. 추계인구의 산출식은 다음과 같다.

$$P = P_0 - P_d + P_m + P_s$$

여기에서 P_0는 "전월말 현재 추계인구", P_d는 "P_0인구 중 이 달에 사망한 인구", P_m은 "금월 중 연령계급의 이동에 의한 증가 수", P_s는 "금월 중 입국자수에 출국자 수를 감한 인구"를 나타낸다.

③ 인구의 선형추정값 산출

각 조사가구에는 조사가구가 속한 조사구의 선형추정 승률이 부여되어 있다. 각 조사가구의 선형추정 승률을 남녀별(2범주), 지역별(토쿄, 요코하마, 나고야, 쿄토, 오오사카, 고베, 기타큐슈의 7대도시와 7대도시 이외의 지역의 2범주), 연령계급별(14범주)의 각 범주에서 합산하고 각 범주의 인구 선형추정값을 산출한다.

④ 벤치마크 인구 산출

앞의 ②에서 산출된 추계인구는 전국의 남녀(2범주)- 연령계급(14범주)별 인구와 남녀(2범주)-지역(7대도시와 7대도시 이외 지역의 2범주)별 인구의 총 32개 범주에 대해서 작성된다. 벤치마크 인구는 32개 범주의 추계인구와 ③에 의해 산출된 선형추정값을 이용하여 지역별-남녀별-연령계급별의 총 56개 범주에 대해 산출된다.

다음 <표3.46>를 참고하여 구체적인 벤치마크 인구 산정에 대해 설명하기로 한다. 우선 <표3.46>에서 ○에는 ②의 추계인구, △에는 ③에서 산출된 선형추정값을 기입한다. △의 숫자의 세로합((Ⅲ))이 지역별 추계인구((Ⅰ))와 일치하지 않을 경우에는 (Ⅳ)의 보정 승률을 산출하여 △의 각 숫자에 곱해서 종계를 추계인구와 일치시킨다. 같은 방법으로 △의 보정숫자의 가로합이 추계인구와 같도록 보정한다. 이와 같은 방법을 각 범주에 대해 순차적으로 계속 적용하여 가로와 세로행렬에 대한 보정을 시

도하면 △의 숫자는 점점 안정되고, 가로와 세로의 관계에 모순이 없는 숫자로 수렴하게 된다. 최종적으로 얻어진 산출값을 56범주에 대한 벤치마크인구로 이용한다.

<표3.46> 56개 범주에 대한 벤치마크 인구의 산출표(남녀별)

구 분		(ⅰ)	(ⅱ)	(ⅲ)	(ⅳ)	(ⅴ)
		전국	7대도시	7대도시 이외지역	총계	짝수회째 보정승률 (ⅰ)÷(ⅳ)
(Ⅰ)	총수	○	○	○	-	-
(Ⅱ)	0-14세	○	△	△		
	15-19세	○	△	△		
	20-24세	○	△	△		
	25-29세	○	△	△		
	30-34세	○	△	△		
	35-39세	○	△	△		
	40-44세	○	△	△		
	45-49세	○	△	△		
	50-54세	○	△	△		
	55-59세	○	△	△		
	60-64세	○	△	△		
	65-69세	○	△	△		
	70-74세	○	△	△		
	75세이상	○	△	△		
(Ⅲ)	총계	-				-
(Ⅳ)	홀수회째 보정승률 (Ⅰ)÷(Ⅲ)	-			-	-

⑤ 비추정 승률의 산출

남녀별, 지역별, 연령계급별로 다음 식에 의해 비추정을 위한 승률을 산출한다.

$$\text{비추정 승률} = \frac{\text{벤치마크인구}}{\text{선형추정값}}$$

참고로 일본 노동력조사에서 1998년 1월부터 12월까지의 15세 이상 비추정 승률의 평균값은 1.108이다.

⑥ 집계 승률의 산출

비추정값은 목적항목의 56개 범주의 선형추정값에 비추정 승률을 곱하여 구할 수 있다. 그러나 계산의 편의 상 각 조사가구마다 집계 승률을 산출해서 계산하고 있다.

각 조사가구의 집계 승률= 조사가구가 속한 조사구의 선형추정 승률
×조사가구가 속한 연령계급, 남녀, 지역의 비추정 승률

⑦ 추정값의 산출

위 ⑥의 "각 조사가구의 집계 승률"을 각 결과표의 목적항목마다 합산해서 구한 결과를 산출한다. 지역별 목적항목에 대한 추정결과는 월별 추정값들의 3개월 분을 평균하여 전국결과와 일치하도록 보정하여 산출한다.

3.5.5 오차관리

일본 노동력조사에서 표본오차의 계산은 조사결과를 사용하여 계산되기 때문에 표본오차에 비표본오차도 일부 포함되어 계산된다. 모든 표본을 랜덤하게 8등분하여 얻어지는 각 부표본들은 전국에 대한 랜덤표본으로 간주되기 때문에 각 부표본에서도 동일 항목의 추정을 시행할 수 있다. 전국의 취업자 수에 대해서 표본 전체에서 추정한 결과를 $\hat{X}$, 각 부표본으로부터 추정한 결과를 $\hat{X}_1, \hat{X}_2, \cdots, \hat{X}_8$이라 하면 부표본에 의한 추정값 $\hat{X}_i$는 전체 표본에서 추정된 결과와 대체로 비슷한 수치를 가질 것으로 기대되지만 $\hat{X}_i$의 표본오차가 클 경우 흐트러짐의 정도가 크게 나타난다. 이러한 흐트러짐의 정도는 8개의 부표본에 의한 평균제곱오차

$$\frac{1}{7}\sum_{i=1}^{8}(\hat{X}_i-\hat{X})^2$$

로 계산된다.

표본이론에 의하면 추정값의 분산은 표본의 크기에 역비례한다. 따라서 각 부표본에 의한 추정값의 분산은 전체 표본의 분할 수에 역비례하게

된다. 이것을 8개의 부표본별 추정값과 표본 전체에 의한 추정값과의 관계에 적용시키면, 표본 전체는 하나의 부표본의 8배가 되므로 표본 전체에 의한 추정값의 평균제곱오차는 각 부표본에 의한 추정값의 분산의 1/8이 된다. 즉 $\hat{X}$의 평균제곱오차는

$$\left(\frac{1}{7}\sum_{i=1}^{8}(\hat{X}_i-\hat{X})^2\right)/8=\frac{1}{56}\sum_{i=1}^{8}(\hat{X}_i-\hat{X})^2$$

로 계산되며, 표준오차는 $\hat{X}$의 평균제곱오차의 제곱근으로 구해진다. 한편 변동계수 CV값은

$$\frac{\sqrt{\frac{1}{56}\sum_{i=1}^{8}(\hat{X}_i-\hat{X})^2}}{\hat{X}}$$

로 계산된다.

1998년 1월부터 12월까지의 일본 노동력조사의 주요항목별 추정값, 표준오차, CV값들에 대한 12개월 분의 단순 평균값들이 다음 <표3.47>에 주어졌다.

<표3.47> 주요항목별 추정결과

주 요 항 목		추정값 (만명)	표준오차 (만명)	CV (%)
노동력 인구		6793	27.2	0.4
산업별 취업자	전산업	6514	26.1	0.4
	농림업	317	10.5	3.3
	비농림업	6197	24.8	0.4
	어업	27	1.4	5.3
	광업	6	0.3	5.7
	건설업	662	13.2	2.0
	제조업	1382	16.6	1.2
	전기,가스,열공급,수도업, 운수 및 통신업	442	9.3	2.1
	도소매업,음식점,금융, 보험업,부동산업	1741	20.9	1.2
	서비스업	1685	16.9	1.0

	공무		217	5.9	2.7
농림업, 비농림업종상의 지위별 취업자	전산업	자영업주	761	12.2	1.6
		가족종업자	367	9.5	2.6
		고용자	5368	21.5	0.4
	농림업	자영업주	156	5.0	3.2
		가족종업자	128	5.9	4.6
		고용자	34	2.3	6.7
	비농림업	자영업주	605	9.7	1.6
		가족종업자	240	7.9	3.3
		고용자	5334	21.3	0.4
완전 실업자			279	7.0	2.5
비노동인구			3924	27.5	0.7

1998년 1월부터 12월까지의 항목별 추정자료를 이용하여 추정값과 *CV*값을 곡선 모형에 적합시켜 추정값의 크기에 따른 *CV*값을 산정한 결과가 다음 <표3.48>에 주어졌다. 산정결과는 추정값의 규모에 따른 *CV*값의 목표허용정도를 결정하는 기준으로 참고한다.

<표3.48> 추정값의 크기별 *CV*값

(i) 전국 연평균 추정값의 *CV*값

추정값의 크기 (만명)	5000	3000	2000	1000	700	500	300	200	100	70	50	30	20	10
표준오차 (만명)	10.8	8.3	6.7	4.7	3.9	3.3	2.5	2.0	1.4	1.2	1.0	0.8	0.6	0.4
CV (%)	0.2	0.3	0.3	0.5	0.6	0.7	0.8	1.0	1.4	1.7	2.0	2.5	3.1	4.3

(ii) 월별 추정값의 *CV*값

추정값의 크기 (만명)	5000	3000	2000	1000	700	500	300	200	100	70	50	30	20	10
표준오차 (만명)	19.7	15.4	12.6	9.0	7.6	6.5	5.1	4.2	3.0	2.5	2.1	1.7	1.4	1.0
CV (%)	0.4	0.5	0.6	0.9	1.1	1.3	1.7	2.1	3.0	3.6	4.3	5.5	6.8	9.8

(iii) 10개 지역별 연평균 추정값의 *CV*값(단위:%)

지역	추정값의 크기(만명)													
	2000	1000	500	300	200	100	50	30	20	10	5	3	2	1
홋카이도				0.80	1.00	1.40	2.00	2.50	3.10	4.30	6.00	7.70	9.30	1.30
토오호쿠			0.60	0.80	1.00	1.40	1.90	2.50	3.00	4.10	5.80	7.40	8.90	1.25
미나미칸토	0.40	0.60	0.80	1.10	1.30	1.80	2.50	3.20	3.90	5.40	7.50	9.50	11.5	1.60
기타칸토·코오신			0.70	0.90	1.10	1.50	2.10	2.70	3.20	4.40	6.10	7.70	9.30	1.29
호쿠라쿠				0.70	0.80	1.10	1.60	2.00	2.40	3.40	4.80	6.10	7.40	1.06
토오카이		0.40	0.60	0.80	1.00	1.30	1.90	2.50	3.00	4.30	6.00	7.80	9.60	1.35
킨키		0.60	0.80	1.00	1.20	1.60	2.30	2.90	3.50	4.90	6.70	8.60	10.4	1.44
츄우고쿠			0.70	0.90	1.00	1.50	2.00	2.60	3.10	4.30	5.90	7.50	9.10	1.26
시코쿠				0.60	0.80	1.10	1.50	1.90	2.40	3.30	4.60	6.00	7.30	1.02
큐슈		0.60	0.80	1.00	1.20	1.60	2.20	2.80	3.40	4.60	6.30	8.00	9.60	1.31

(iv) 10개 지역별 사분기 평균 추정값의 *CV*값

지역	추정값의 크기(만명)													
	2000	1000	500	300	200	100	50	30	20	10	5	3	2	1
홋카이도				1.40	1.70	2.30	3.30	4.20	5.20	7.30	10.3	13.3	16.2	22.8
토오호쿠			1.10	1.40	1.70	2.40	3.40	4.40	5.30	7.50	10.5	13.4	16.4	23.0
미나미칸토	0.70	1.00	1.40	1.80	2.20	3.00	4.30	5.40	6.60	9.30	13.0	16.7	20.3	28.4
기타칸토·코오신			1.10	1.40	1.80	2.50	3.40	4.40	5.40	7.50	10.5	13.5	16.4	23.0
호쿠라쿠				1.20	1.40	2.00	2.80	3.60	4.40	6.20	8.70	11.2	13.7	19.4
토오카이		0.70	1.10	1.40	1.70	2.40	3.40	4.40	5.40	7.60	10.8	14.0	17.1	24.3
킨키		1.00	1.30	1.70	2.10	2.90	4.00	5.20	6.30	8.80	12.3	15.7	19.0	26.6
츄우고쿠			1.10	1.40	1.70	2.40	3.30	4.30	5.20	7.40	10.4	13.3	16.3	22.9
시코쿠				1.10	1.40	1.90	2.70	3.50	4.30	6.10	8.60	11.0	13.5	19.0
큐슈		0.90	1.20	1.60	1.90	2.60	3.70	4.70	5.70	7.90	11.0	14.0	16.9	23.6

제4장 소지역 추정법

4.1 개 요

지리적으로 구분된 영역 또는 좀 더 일반적으로 임의의 부분모집단이 소지역으로 간주될 수 있다. 지리적으로 구분된 도 단위가 소지역이 될 수도 있고, 도 단위 내의 시군구 단위가 소지역이 될 수도 있다. 통상적으로 소지역에 대한 표본의 크기는 소지역의 크기에 따라 증가하는 경향이 있으나, 대영역을 기준으로 작성된 표본설계에서 소지역을 추정할 경우에는 반드시 이러한 경향을 보이는 것은 아니다. 소지역에 대한 추정값들은 정부의 지역별 예산 배분 등의 정책 입안 시에 절대적인 참고자료로 활용될 수 있기 때문에 최근에는 신뢰할 수 있는 소지역 통계 작성에 많은 관심들을 보이고 있다.

5년 또는 10년 마다 실시되는 센서스 자료, 실업자의 구직 등록 자료나 실직 보험 자료와 같은 행정보고 자료, 다양한 사회적 관심에 의해 실시되는 표본조사 자료 등이 소지역 추정을 위해 활용된다. 그러나 많은 경우 표본조사에 의해 확보된 직접 추정값들은 대영역에 대해서는 신뢰할 만한 수준이나, 소지역에서는 확보된 표본이 작을 경우가 대부분이므로 신뢰할만한 정확도를 기대하기는 어려운 실정이다. 이러한 이유 때문에 유사한 특성을 갖는 인근 지역으로부터 정보를 취득하여 소지역의 추정의 정확도를 높일 수 있는 간접 추정방법을 생각하게 되는데, 이러한 추정방법이 소지역 추정에 활용된다.

간접 추정량은 최근의 센서스 자료라든가 행정 보고 자료와 같은 보조자료를 소지역들과 연계할 수 있는 암시모형(Implicit Model) 또는 명시모형(Explicit Model)을 설정하여 추정된다. 암시모형 하에서 추정되는 간접 추정량으로 합성 추정량(Synthetic Estimator)과 복합 추정량(Composite Estimator)을 들 수 있고, 모형에 근거한 추정량으로는 경험적 베이즈(*EB*) 추정량, 계층적 베이즈(*HB*) 추정량 등을 들 수 있다.

이번 장에서는 소지역 추정기법에 관한 일반적인 이론들을 요약 정리한다. 대표적인 간접 추정법인 인구통계학적 방법, 합성 추정법과 복합 추

정법을 살펴보고, 모형 기반 추정법으로는 *EB* 추정법과 *HB* 추정법을 소개한다.

4.2 인구통계학적 방법(Demographic Method)

미국의 경우에서와 같이 10년 주기로 센서스를 할 경우, 지방 도시나 county의 중간 해당 년도의 인구를 추정하기 위해서 사용하는 추정법으로 센서스 자료와 인구수에 관련된 징후 변수(출생자수, 사망자수, 주택수, 등록한 학생 수 등)의 변동을 분석하여 얻은 예측값을 결합하는 추정법을 인구 통계학적 방법(Demographic Method)이라 말한다.

4.2.1 생명률법(Vital Rates Method: VR Method)

*VR*법은 출생과 사망에 관련된 자료를 이용하여 인구의 변동률보다는 징후 변수의 영향만을 분석하여 활용한다. 가장 최근에 센서스를 실시한 해를 기준 년도로 하고, 기준해로부터 t년 후에 소지역의 인구 수를 추정한다고 하자. 이 때 전제 조건은 추정 대상인 소지역을 포함하는 대지역의 특성과 소지역의 특성이 동일하다는 것이며, 전제 조건에서 많이 벗어나는 경우에는 추정량의 편향이 커져서 신뢰도가 낮아진다.

t 년 후의 소지역의 출생률과 사망률을 γ_{bt}와 γ_{dt}로 표현하고 대영역의 출생률과 사망률을 R_{bt}와 R_{dt}라 나타내면 다음과 같은 관계가 주어진다.

$$\gamma_{bt} = \gamma_{bo}\left(\frac{R_{bt}}{R_{bo}}\right), \quad \gamma_{dt} = \gamma_{do}\left(\frac{R_{dt}}{R_{do}}\right), \tag{4.1}$$

여기에서 γ_{bo}와 γ_{do}는 기준 해의 소지역의 출생률과 사망률이고 R_{bo}와 R_{do}는 기준 해의 대지역의 출생률과 사망률을 의미한다.

센서스를 실시한 기준해로부터 t 년 후의 인구 수는 다음 식에 의해서 추정할 수 있다.

$$p_t = \frac{1}{2}\left(\frac{b_t}{\gamma_{bt}} + \frac{d_t}{\gamma_{dt}}\right). \tag{4.2}$$

단, b_t와 d_t는 소지역의 t년 후의 출생자수와 사망자수를 뜻한다.

4.2.2 성분법(Components Method)

성분법은 출생과 사망 인구수 및 유입, 유출 인구에 관한 자료를 이용하여 소지역의 인구 수를 추정하기 위해 고안된 방법이다. 센서스를 실시한 기준해로부터 t년 동안의 출생 인구, 사망인구 및 총 이주인구를 각각 $b_{0,t}$, $d_{0,t}$와 $m_{0,t}$로 나타냈을 때 t년 후의 인구수는 다음식에 의해 추정한다.

$$p_t = p_0 + b_{0,t} - d_{0,t} + m_{0,t} \ , \tag{4.3}$$

여기에서 $m_{0,t} = i_{0,t} - e_{0,t} + n_{0,t}$로 주어지며, $i_{0,t}$는 유입인구, $e_{0,t}$는 유출인구, $n_{0,t}$는 주 간의 총 이주인구를 나타내며 행정보고자료에 의해 주어진다.

4.2.3 회귀 징후법(Regression Symptomatic Procedures)

회귀 징후법은 다중선형회귀모형을 이용하여 소지역의 인구를 추정하는 방법으로써 징후변수들을 독립변수로 선택하여 소지역 추정에 이용한다. 비 상관계수(ratio correlation), 차분 상관계수(difference correlation), 표본 회귀법(sample regression method) 등은 이러한 회귀징후법의 일종이다. 여기에서는 다른 두 방법보다는 비교적 자주 사용되고 있는 표본 회귀법을 설명하기로 한다. 먼저 종속변수와 독립변수를 다음과 같이 정의하자.

$Y_i = (p_{it}/P_t)/(p_{i0}/P_0) =$ 소지역 i의 인구비 변화량,

$x_{ij} = (s_{ijt}/S_{jt})/(s_{ij0}/S_{j0}) =$ 소지역 i에 대한 j번째 징후변수 s_j의 변화량,

여기에서 P_t , P_0 , S_{jt} 와 S_{j0} 는 소지역 i를 포함하는 대지역에서의 값들이고, x_{ij}는 행정자료로부터 얻는다($j = 1, 2, \cdots, p$).

회귀 표본법은 종속변수 Y_i가 징후변수 $x_{i1}, x_{i2}, \cdots, x_{ip}$의 일차결합으

로 표현될 수 있다는 것을 가정하며, 이때 Y_i의 값은 조사된 직접추정값 $\hat{Y}_i$을 이용하여 m개의 소지역 중 k개의 소지역에 대하여 선형회귀식을 적합시켜 회귀계수들을 추정한 후, Y_i의 추정값으로 다음의 표본회귀 추정량을 이용한다.

$$\tilde{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \ , \ i=1,2, \cdots , m \qquad (4.4)$$

소지역 i에 대한 인구수는 (4.4)식의 표본 회귀추정량을 이용하여 다음 식으로 추정한다.

$$\bar{p}_{it} = \tilde{Y}_i \left(\frac{p_{i0}}{P_0} \right) \hat{P}_t \ , \quad i=1,2, \cdots , m \qquad (4.5)$$

표본 회귀추정량은 표본으로부터 직접 추정된 값이 아니라 다중 선형회귀를 거쳐 얻어진 보정된 추정량이며, 표본 회귀법은 이를 이용하여 소지역의 인구를 추정하는 방법이다. 그러나 이러한 방법은 추후 논의될 모형에 근거한 소지역 추정보다는 효율성이 상당히 떨어지는 것으로 밝혀지고 있다.

4.3 합성 추정법(Synthetic Estimation)

추정하고자 하는 소지역과 특성이 유사한 소지역들의 정보를 이용하여 추정값의 정도를 높이고자하는 추정방식을 합성 추정법이라 하며, 주변이나 유사지역의 정보를 이용하므로 "Borrow Strength"라고 말하기도 한다. 표본조사의 설계 시에는 대영역에 대해서만 직접 추정값을 구하고자 하였으나 대영역을 분할한 소지역의 추정값이 필요한 때에는 대영역과 소지역의 구조적 특성이 같다는 조건하에서 소지역의 연구변수에 대한 추정값을 구할 수 있는데, 이때 대영역의 분할은 지리적인 분할보다는 연령대별 또는 교육정도별과 같은 특성에 따른 분할을 말한다.

대영역을 I개 소지역으로 분할하며 또한 대영역을 특성 기준에 따라 J개의 범주로 분류한다면 i 소지역의 추정값은 다음 식으로 구할 수 있다.

$$\hat{Y}_{i\cdot} = \sum_{j}^{J} p_{ij} \hat{Y}_{\cdot j} \tag{4.6}$$

단, p_{ij}는 소지역 i의 j범주에 대한 가중값이며 센서스나 행정자료에서 구해진다. $\hat{Y}_{\cdot j}$는 대영역에서 j범주에 대한 표본 추정값이다. 단, 대영역의 표본의 수는 충분하게 많아서 신뢰성 있는 추정값을 구할 수 있다고 가정한다. i소지역의 실업자 추정에 관한 경우를 생각해 보자.

Y_{ij} : i소지역의 j범주(연령대별 또는 교육정도별)의 실업자수,

X_{ij} : i소지역의 j범주(연령대별 또는 교육정도별)의 경제활동인구,

$Y_{\cdot j} = \sum_{i} Y_{ij}$: 대영역에서 j범주의 실업자 수,

$Y_{i\cdot} = \sum_{j} Y_{ij}$: i소지역의 실업자 수.

$Y_{\cdot j}$의 직접 추정값 $\hat{Y}_{\cdot j}$는 표본조사 자료만으로 추정가능하고, X_{ij}는 센서스 또는 행정자료 등 보조변수의 정보에서 계산 가능한 것으로 가정한다면 합성추정량은 다음과 같이 나타낼 수 있다.

$$\hat{Y}_{i\cdot}^{S} = \sum_{j} \left(\frac{X_{ij}}{X_{\cdot j}} \right) \hat{Y}_{\cdot j} \tag{4.7}$$

만약 $\hat{Y}_{\cdot j}$가 비추정량의 형식을 갖는다면,

$$\hat{Y}_{\cdot j} = \left(\frac{\hat{Y}_{\cdot j}}{\hat{X}_{\cdot j}} \right) X_{\cdot j}$$

로 나타낼 수 있으므로 (4.7)식은 다음과 같이 표현될 수 있다.

$$\hat{Y}_{i\cdot}^{S} = \sum_{j} X_{ij} \left(\frac{\hat{Y}_{\cdot j}}{\hat{X}_{\cdot j}} \right) = \sum_{j} \left(\frac{X_{ij}}{\hat{X}_{\cdot j}} \right) \hat{Y}_{\cdot j} \tag{4.8}$$

여기에서 $\hat{Y}_{i\cdot}^{S}$가 불편추정량이 되기 위해서는 $\frac{Y_{\cdot j}}{X_{\cdot j}} = \frac{Y_{ij}}{X_{ij}}$를 만족해야 하고, 이를 만족하지 못할 경우에는 편향추정량이 되며, 이때 $\hat{Y}_{i\cdot}^{S}$의 편향의

크기는 $B(\hat{Y}_{i.}^{S}) = E(\hat{Y}_{i.}^{S} - Y_{i.})$이다. 즉,

$$B(\hat{Y}_{i.}^{S}) = \sum_{j} X_{ij}\left(\frac{Y_{.j}}{X_{.j}} - \frac{Y_{ij}}{X_{ij}}\right).$$

$\hat{Y}_{i.}^{S}$의 평균제곱오차 $MSE(\hat{Y}_{i.}^{S})$의 근사적 불편추정량은 다음과 같이 주어질 수 있다.

$$\widehat{MSE}(\hat{Y}_{i.}^{S}) = (\hat{Y}_{i.}^{S} - \hat{Y}_{i.})^2 - \widehat{Var}(\hat{Y}_{i.}) \tag{4.9}$$

4.4 복합 추정법(Composite Estimation)

소지역에 배정된 표본수가 적기 때문에 표본 조사만을 이용한 직접 추정량의 불안정에서 오는 낮은 신뢰성과 합성추정량의 편향을 보완하기 위해서 직접 추정값과 합성 추정값의 가중평균은 사용하는데 이를 복합 추정량(Composite Estimator)이라 한다.

$$\hat{Y}_{i.}^{C} = w_i \hat{Y}_{i.} + (1 - w_i)\hat{Y}_{i.}^{S} \tag{4.10}$$

여기에서 $\hat{Y}_{i.}$는 표본조사에서 직접 계산한 추정값이며, $\hat{Y}_{i.}^{S}$는 합성 추정값을 나타낸다. w_i는 가중값으로 0과 1사이의 값이다.

먼저 평균제곱오차 $MSE(\hat{Y}_{i.}^{C})$를 최소화하는 w_i는 아래와 같다.

$$w_{i(opt)} = \frac{MSE(\hat{Y}_{i.}^{S})}{MSE(\hat{Y}_{i.}^{S}) + Var(\hat{Y}_{i.})} \tag{4.11}$$

최적 가중값 $w_{i(opt)}$의 추정값은 다음 식으로 계산될 수 있다.

$$\hat{w}_{i(opt)} = \frac{mse(\hat{Y}_{i.}^{S})}{(\hat{Y}_{i.}^{S} - \hat{Y}_{i.})^2} \tag{4.12}$$

모든 소지역에 공통 가중값을 부여하는 방법으로써 초기 공통 가중값 w를 이용하여 $MSE(\hat{Y}_{i.}^{S})$들의 평균을 최소화하는 가중값은 아래와 같다.

$$\hat{w}_{(opt)} = 1 - \frac{\sum_i \widehat{Var}(\hat{Y}_{i\cdot})}{\sum_i (\hat{Y}^S_{i\cdot} - \hat{Y}_{i\cdot})^2} \tag{4.13}$$

각 소지역에 배정된 표본 크기에 의존하는 가중값은 다음과 같이 계산된다.

$$w_i(\delta) = \begin{cases} 1, & \hat{N}_i \geq \delta N_i \\ \dfrac{\hat{N}_i}{\delta N_i}, & otherwise \end{cases} \tag{4.14}$$

단, N_i는 i소지역의 크기이며 $\hat{N}_i = N(n_i/n)$ 이다. $\hat{N}_i$는 직접추정량이며 δ는 합성추정량의 기여도를 조정하는 값이므로 주관적으로 결정한 값이다. 예를 들어 캐나다 노동력 통계조사에서는 2/3으로 한다.

어떤 추정법에 의해서 소지역의 추정값을 구하더라도 대영역을 소지역으로 분할하여 각 소지역의 추정값을 추정하므로 소지역의 추정값의 합계는 대영역의 추정값과 같아야 할 것이다. 왜냐하면 매월 정부기관에서 발표하는 광역시와 도의 실업자 수와 해당 소지역의 추정값의 합계가 같도록 조정하지 않으면 서로 상이한 통계수치로 인하여 혼란을 줄 수 있기 때문에 한 가지 통계수치가 되도록 조정된 추정량을 계산해야 할 것이다. 각 소지역의 추정량을 생명률법, 합성추정법 또는 복합 추정법 등의 어느 한 방법으로 계산한 것으로 간주할 때 조정된 소지역 추정량은 다음과 같다.

$$\hat{Y}^A_{i\cdot} = \left(\frac{\hat{Y}^*_{i\cdot}}{\sum_i \hat{Y}^*_{i\cdot}} \right) \hat{Y} \tag{4.15}$$

단, $\hat{Y}$는 대영역에 대한 직접 추정값이며, $\hat{Y}^*_{i\cdot}$는 i소지역을 * 추정법으로 추정한 것이다.

4.5 모형기반 간접추정법

4.5.1 기본적인 지역 수준 모형

소지역 추정시 모형에 근거한 추정방법이 많은 사람들의 관심을 끌고 있는 것은 다음과 같은 몇가지 장점에 기인한다. 먼저 모형 기반 추정법은 소지역들을 연결하고 있는 모형 구조가 소지역 간의 복잡한 오차구조를 내포하고 있기 때문에 소지역 간의 변동을 반영하여 소지역 추정의 정확도를 높일 수 있다는 점이며, 또한 표본자료로부터 모형의 유용성이 확인될 수 있고, 연속형의 자료뿐만 아니라 범주형 자료 및 시계열 자료와 같은 다양한 경우들에 대해서도 모형화하여 추론할 수 있으며, 모형 기반 추정법으로 소지역 추정량들과 연관있는 많은 측도들이 얻어질 수 있다는 장점들을 들 수 있다.

지역 간의 공변량을 포함하고 있는 지역 수준 모형을 이용하여 경험적 최량선형불편예측($EBLUP$) 추정량, 경험적 베이즈(EB) 추정량, 계층적 베이즈(HB) 추정량에 대해 소개하기로 한다. 지역 수준 모형은 기본적으로 두가지의 성분들로 이루어진다. 즉, 소지역에 대한 직접추정량 $\hat{\theta}_i$과 소지역의 보조변수들로 표현되는 θ_i의 두 가지 성분들을 모형으로 연결하여 모형 기반 추정량을 찾아내게 된다.

지정된 함수 $g(\cdot)$에 대하여 직접 추정량 $\hat{\theta}_i = g(\hat{\overline{Y}}_i)$은 모집단의 값 $\theta_i = g(\overline{Y}_i)$와 표본추출오차 e_i에 의해 다음과 같이 표현될 수 있다.

$$\hat{\theta}_i = \theta_i + e_i \ , \qquad i = 1, 2, \cdots, m \tag{4.16}$$

여기에서 표본추출오차 e_i는 서로 독립이며, 평균이 0, 분산이 ψ_i임이 가정되며, 보통 ψ_i는 기지인 값으로 가정된다.

θ_i는 소지역의 정보를 나타내는 보조변수 $z_i = (z_{i1}, z_{i2}, \cdots, z_{ip})^T$를 이용하여 선형회귀모형을 통해 표현한다.

$$\begin{aligned}\theta_i &= z_{i1}\beta_i + z_{i2}\beta_2 + \cdots + z_{ip}\beta_p + v_i \\ &= z_i^T\beta + v_i\end{aligned} \tag{4.17}$$

여기에서 모형오차 v_i는 서로 독립이며, 평균이 0, 분산 σ_v^2을 갖고, 표본추출오차 e_i와는 서로 독립임을 가정한다.

마지막으로 (4.16)과 (4.17)의 두 성분들을 결합하면 다음과 같은 결합모형을 얻을 수 있다.

$$\hat{\theta}_i = z_i^T\beta + v_i + e_i \tag{4.18}$$

위의 결합모형은 고정효과 β와 소지역 랜덤효과 v_i를 갖는 선형혼합효과모형의 일종이며, 특히 설계 기반 확률변수(design-based random variable) e_i와 모형 기반 확률변수(model-based random variable) v_i를 동시에 포함하고 있는 모형이다. 여기에서 모수 σ_v^2은 소지역들의 동질성을 나타내는 측도이다.

4.5.2 경험적 베이즈(*EB*)방법

경험적 최량선형불편예측(*EBLUP*) 방법, 경험적 베이즈(*EB*) 방법 및 계층적 베이즈(*HB*) 방법은 모형에 근거한 소지역 추정문제에 많이 활용되고 있는 방법이다. 특히 경험적 최량선형불편예측 방법은 선형혼합모형을 이용한 추론에 응용되어 왔고, 경험적 베이즈 방법 및 계층적 베이즈 방법은 좀 더 일반적인 모형을 이용한 소지역 추정에 활용되고 있다.

EBLUP 추정량은 랜덤오차 e_i와 v_i의 분포에 대한 가정을 필요로 하지 않으나, *MSE* 추정을 위해 정규분포를 가정하기도 한다. 또한, *EBLUP* 추정량과 *EB* 추정량은 e_i와 v_i를 정규분포로 가정했을 경우에는 동일하며, *HB* 추정량과는 근사적으로 같게 나타난다. 그러나 추정량들의 변동을 나타내는 측도들은 동일하지는 않다.

고정계수 l_t를 갖는 θ_i의 선형추정량 $\sum_t l_t\hat{\theta}_t$가 모형 (4.18)에 대해서

$\sum_t l_t\hat{\theta}_t - \theta_i$의 기대값이 0을 만족할 때, $\sum_t l_t\hat{\theta}_t$를 θ_i의 선형불편예측(LUP)추정량이라 한다. θ_i의 최량선형불편예측($BLUP$)추정량은 선형불편예측(LUP) 추정량들 중 최소평균제곱오차를 갖는 추정량을 말한다.

모형 (4.18) 하에서 θ_i의 $BLUP$추정량은 다음과 같이 주어진다(Prasad and Rao,1990).

$$\tilde{\theta}_i(\sigma_v^2) = \gamma_i\hat{\theta}_i + (1-\gamma_i)\, z_i^T\tilde{\beta}(\sigma_v^2) \tag{4.19}$$

여기에서 $\gamma_i = \sigma_v^2/(\sigma_v^2+\psi_i)$ 이고, $\tilde{\beta}(\sigma_v^2)$ 은 가중치 $(\sigma_v^2+\psi_i)^{-1}$을 갖는 가중최소제곱추정량으로 아래와 같이 주어진다.

$$\tilde{\beta}(\sigma_v^2) = (\sum_i \gamma_i z_i z_i^T)^{-1}(\sum_i \gamma_i z_i y_i) \tag{4.20}$$

(4.19)식의 $BLUP$추정량은 가중치 γ_i를 갖는 직접추정량 $\hat{\theta}_i$과 가중치 $1-\gamma_i$를 갖는 회귀합성추정량 $z_i^T\tilde{\beta}(\sigma_v^2)$의 가중결합으로 볼 수 있다. 또한, 표본분산 ψ_i가 작을때(σ_v^2이 클 경우) $BLUP$추정량은 직접추정량 $\hat{\theta}_i$에 큰 가중치가 부여되고, 반대의 경우에는 회귀합성추정량 $z_i^T\tilde{\beta}(\sigma_v^2)$에 큰 가중치가 부여된다. 표본이 추출되지 않은 지역들에 대해서는 $BLUP$ 추정량은 회귀합성추정량만으로 주어질 수 있다.

$BLUP$추정량의 변동의 측도는 추정량의 $MSE = E(est.-\theta_i)^2$에 의해 주어지며 다음과 같다.

$$MSE\left(\tilde{\theta}_i(\sigma_v^2)\right) = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2)\ , \tag{4.21}$$

여기에서

$$g_{1i}(\sigma_v^2) = \gamma_i\psi_i\ ,$$

$$g_{2i}(\sigma_v^2) = \sigma_v^2(1-\gamma_i)^2 z_i^T(\sum_i \gamma_i z_i z_i^T)^{-1} z_i$$

로 주어진다.

식(4.19)와 (4.21)은 랜덤오차 v_i와 e_i에 관한 분포의 가정을 필요로 하지는 않는다.

주요 항 $g_{1i}(\sigma_v^2) = \gamma_i \psi_i$ 는 $O(1)$, $g_{2i}(\sigma_v^2)$ 은 $O(m^{-1})$ 에 유계인 항이며, 이로부터 $BLUP$추정량의 MSE 값은 γ_i나 모형분산 σ_v^2이 표본분산 ψ_i에 비해 작을 경우 직접추정량의 MSE 값보다 훨씬 작아질 수 있다는 사실을 알 수 있다. 따라서 소지역 추정의 정확도는 표본 분산에 비해 모형 분산을 작게 할 수 있는 보조변수에 크게 의존한다고 볼 수 있다.

대부분의 문제에서는 모형분산 σ_v^2은 미지이므로 적절한 $\hat{\sigma}_v^2$을 추정하여 $EBLUP$추정량 $\tilde{\theta}_i = \tilde{\theta}_i(\hat{\sigma}_v^2)$ 을 산출한다. 이때 소지역의 평균 $\overline{Y}_i$의 추정량은 $g^{-1}(\tilde{\theta}_i)$ 로, σ_v^2 의 추정량은 $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$ 로 주어진다. 여기에서 $\tilde{\sigma}_v^2$은 다음 식을 만족한다.

$$(m-p)\tilde{\sigma}_v^2 = \sum_i (\hat{\theta}_i - z_i^T \beta^*)^2 - \sum_i \psi_i h_{ii} \qquad (4.22)$$

(4.22)식에서 $h_{ii} = z_i^T (\sum_i z_i z_i^T)^{-1} z_i$이고, β^*는 β의 OLS(ordinary least squares) 추정량이다. 한편, $\tilde{\sigma}_v^2$은 다음과 같은 비선형 방정식의 반복적인 해로써 구할 수도 있다.

$$a(\sigma_v^2) = \sum_i \{\hat{\theta}_i - z_i^T \tilde{\beta}(\sigma_v^2)\}^2 / (\sigma_v^2 + \psi_i) = m - p \qquad (4.23)$$

여기에서 $\tilde{\beta}(\sigma_v^2)$ 은 (4.20)식에 주어졌고, (4.23)식의 가운데 항은 가중잔차제곱합, $m-p$는 가중잔차제곱합과 관계가 되는 자유도이다. 만약 $\hat{\sigma}_v^2 = 0$ 이면 $EBLUP$추정량 $\tilde{\theta}_i$는 회귀합성추정량 $z_i^T \hat{\beta}$로 축약된다. 단, $\hat{\beta} = \tilde{\beta}(\hat{\sigma}_v^2)$ 이며, (4.20)식에서 σ_v^2 대신에 $\hat{\sigma}_v^2$으로 대체하여 산출한다. 물론 위의 (4.22), (4.23)식으로부터 얻게 되는 추정량들도 v_i와 e_i의 분포에

대한 가정을 필요로 하지는 않는다.

만약 랜덤오차 v_i와 e_i가 정규분포를 따른다고 가정한다면, $\hat{\theta}_i$는 평균이 $z_i^T\beta$ 이고 분산이 $\sigma_v^2+\psi_i$인 서로 독립인 정규분포를 따르게 된다. 이러한 분포에 대한 가정 하에서 계산된 β와 σ_v^2의 최대우도추정량을 제한최대우도추정량($REML$)이라 하며, 선형혼합모형 하에서도 근사적으로 유효하다. 따라서 $\tilde{\theta}_i$의 $BLUP$추정량을 산출 시 σ_v^2의 $REML$추정량을 이용하여도 근사적으로 타당하다.

경험적 베이즈(EB) 추정법은 랜덤오차 v_i와 e_i가 정규분포를 따른다는 가정 하에서 출발한다. $(\hat{\theta}_i, \theta_i)$의 결합분포가 평균이 $(z_i^T\beta, z_i\beta)$, 분산이 $(\sigma_v^2+\psi_i, \sigma_v^2)$, 상관계수가 γ_i인 이변량 정규분포를 따른다고 가정하자. 이때, θ_i 의 평균제곱오차를 최소로 하는 베이즈 추정량은 다음과 같다.

$$\begin{aligned}\tilde{\theta}_i^B(\beta, \sigma_v^2) &= E(\theta_i \mid \hat{\theta}_i, \beta, \sigma_v^2) \\ &= \gamma_i\hat{\theta}_i + (1-\gamma_i)z_i^T\beta \end{aligned} \tag{4.24}$$

(4.24)식의 베이즈 추정량은 선형성 또는 불편성을 만족하지는 않는다. 여기에서 모수 β와 σ_v^2을 제한최대우도($REML$)추정량으로 대체하여 다음과 같은 θ_i의 경험적 베이즈(EB) 추정량을 얻는다.

$$\tilde{\theta}_i^{EB} = \tilde{\theta}_i^B(\hat{\beta}, \hat{\sigma}_v^2) \tag{4.25}$$

경험적 베이즈(EB) 추정량 $\tilde{\theta}_i^{EB}$는 정규분포의 가정 하에서는 $EBLUP$ 추정량 $\tilde{\theta}_i$와 같다. 그러나 경험적 베이즈 방법은 $\hat{\theta}_i$과 θ_i의 임의의 결합분포에 대해서도 일반적으로 응용할 수 있다는 점을 장점으로 들 수 있다.

$EBLUP$추정량 $\tilde{\theta}_i = \tilde{\theta}_i(\hat{\sigma}_v^2)$의 MSE추정량은 (4.21)식에서 σ_v^2 대신 $\hat{\sigma}_v^2$을 대체하여 얻어질 수 있으나, 이 경우에는 σ_v^2에 대한 추정효과가 무시

되기 때문에 MSE의 추정값은 과소추정되는 경향을 보인다. 이러한 문제 때문에 Prasad and Rao(1990)는 $\{v_i\}$와 $\{e_i\}$에 대해 정규성을 가정하여 근사적으로 불편인 $EBLUP$추정량 $\tilde{\theta}_i$의 MSE 추정량을 제안하였다. Prasad and Rao(1990)가 제안한 MSE추정량은 (4.22)식의 σ_v^2의 적률추정량을 사용하였을 경우 다음과 같이 주어진다.

$$mse(\tilde{\theta}_i) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2) \tag{4.26}$$

여기에서

$$g_{1i}(\sigma_v^2) = \gamma_i \psi_i \ ,$$

$$g_{2i}(\sigma_v^2) = \sigma_v^2 (1-\gamma_i)^2 z_i^T (\sum_i \gamma_i z_i z_i^T)^{-1} z_i \ ,$$

$$g_{3i}(\sigma_v^2) = [\psi_i^2 / (\sigma_v^2 + \psi_i)^3] h(\sigma_v^2) \ ,$$

$$h(\sigma_v^2) = 2m^{-2} \sum_i (\sigma_v^2 + \psi_i)^2$$

로 주어진다.

최근들어 Jing, Lahiri and Wan(1999)은 근사적으로 불편인 잭나이프 MSE 추정량을 제안하였다. 잭나이프 방법은 랜덤인 지역효과들을 갖는 로지스틱 회귀와 같은 좀 더 복잡한 모형들에 대해서도 쉽게 적용할 수 있다는 장점을 갖고 있다.

θ_i의 EB추정량 (4.25)를 $\tilde{\theta}_i^{EB} = k(\hat{\theta}_i, \hat{\phi})$ 로 표현할 때, 잭나이프 절차는 다음과 같다. 여기에서 $\phi = (\beta, \sigma_v^2)$ 은 모형에서의 모수 β와 σ_v^2을 나타낸다.

(i) l 번째 지역의 자료 $(\hat{\theta}_l, z_l)$을 제외한 ϕ의 추정량 $\hat{\phi}(l)$을 계산한다. 이때의 EB추정량을 $\tilde{\theta}_i^{EB}(l) = k(\hat{\theta}_i, \hat{\phi}(l))$로 나타내자.

(ii) $\hat{M}_{2i} = \frac{m-1}{m} \sum_{l=1}^{m} [\tilde{\theta}_i^{EB}(l) - \tilde{\theta}_i^{EB}]^2$를 계산한다.

(iii) $\hat{M}_{1i} = g_{1i}(\hat{\sigma}_v^2) - \frac{m-1}{m}\sum_{l=1}^{m}[g_{1i}(\hat{\sigma}_v^2(l)) - g_{1i}(\hat{\sigma}_v^2)]^2$을 계산한다.

(iv) 잭나이프 MSE추정량 $mse_J(\tilde{\theta}_i^{EB}) = \hat{M}_{1i} + \hat{M}_{2i}$를 계산한다.

$\hat{M}_{1i}$은 ϕ가 기지일 때 MSE에 대한 추정량이며, $\hat{M}_{2i}$는 모형 모수 ϕ를 추정할 때 추가적으로 발생하는 MSE에 대한 변화량을 추정한다.

4.5.3 계층적 베이즈(HB) 방법

계층적 베이즈(HB) 방법을 이용한 추론은 비교적 추론의 정확도가 높고, 복잡한 유형의 문제들에서도 최근에 개발된 $MCMC$(Monte Carlo Markov Chain)방법을 이용하여 해결할 수 있다. 깁스 샘플러가 이러한 문제 해결을 위해 제공된다. HB방법에서는 모형 모수 $\phi = (\beta, \sigma_v^2)$ 뿐만 아니라 모집단의 값 θ_i가 랜덤으로 간주되며, 모형 모수들에 대한 사전분포가 명시된다. θ_i의 추론은 주변 사후분포에 의해 결정된다. 즉, 주어진 자료 $\{(\hat{\theta}_i, z_i), i=1,2,\cdots,m\}$에 대한 조건부 분포 $f(\theta_i|\hat{\theta})$에 의해 추론이 행해진다. 여기에서 $\hat{\theta}$은 직접추정값 $\hat{\theta}_i$의 벡터이다. 특히 θ_i는 사후분포의 평균 $E(\theta_i|\hat{\theta})$에 의해 추정되며, 추정량의 변동은 사후분포의 분산 $V(\theta_i|\hat{\theta})$에 의해 추정된다.

먼저 σ_v^2이 기지인 상태를 가정하고 β에 관한 사전분포를 배정하기로 한다. β의 사전분포가 상수에 비례하고(i.e improper prior), v_i와e_i 가 정규분포를 따른다고 가정한다면, 이때 사후평균 $E(\theta_i|\hat{\theta}, \sigma_v^2)$ 은 (4.19)식의 $BLUP$추정량 $\tilde{\theta}_i(\sigma_v^2)$과 동일하다. 더욱이 사후분산 $V(\theta_i|\hat{\theta}, \sigma_v^2)$은 (4.21)식의 $BLUP$추정량의 MSE와 같다. 따라서 σ_v^2이 기지인 상태에서는 HB 방법과 $EBLUP$방법은 동일한 추론을 이끌어 낸다고 볼 수 있다.

실제의 문제에서는 σ_v^2은 대부분 미지의 값으로 나타난다. 이러한 경우

에는 β 뿐만 아니라 σ_v^2에 관한 사전분포를 고려해야 하며, 또한 서로 독립임을 가정하여 주변사후분포 $f(\sigma_v^2 | \hat{\theta})$을 이끌어 낸다. 만약 σ_v^2에 관한 사전분포를 불완전(improper) 사전분포로 배정한다면, θ_i의 사후분포가 불완전 사후분포가 될 수 있기 때문에 이러한 문제를 피하기 위해서 $\tau_v = \sigma_v^{-2}$의 사전분포를 $G(a, b)$와 같이 배정 한다 (여기에서 $G(a, b)$: $f(\tau_v) \propto exp(-a\tau_v)\, \tau_v^{b-1}$). 주변사후분포 $f(\sigma_v^2 | \hat{\theta})$를 이용한 HB 추정량 $E(\theta_i | \hat{\theta})$은 다음 식과 같이 주어진다.

$$\tilde{\theta}_i^{HB} = E(\theta_i | \hat{\theta}) = \int \tilde{\theta}_i(\sigma_v^2) f(\sigma_v^2 | \hat{\theta})\, d\sigma_v^2 \qquad (4.27)$$

위의 (4.27)식을 $E_{\sigma_v^2 | \hat{\theta}}[\tilde{\theta}_i(\sigma_v^2)]$으로 표현하면, 사후분산 $V(\theta_i | \hat{\theta})$은 다음과 같다.

$$V(\theta_i | \hat{\theta}) = E_{\sigma_v^2 | \hat{\theta}}[g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2)] + V_{\sigma_v^2 | \hat{\theta}}[\tilde{\theta}_i(\sigma_v^2)] \qquad (4.28)$$

여기에서 $V_{\sigma_v^2 | \hat{\theta}}$는 $f(\sigma_v^2 | \hat{\theta})$에 관한 분산을 의미한다.

위에서 소개한 (4.27)식과 (4.28)식은 일차원 수치적분으로 계산된다. 좀 더 복잡한 모형에 대한 고차원 수치적분은 $MCMC$방법을 이용하여 계산할 수 있다. (4.27)식으로부터 $\tilde{\theta}_i^{HB}$는 $EBLUP(EB)$추정량 $\tilde{\theta}_i(\hat{\sigma}_v^2)$과 근사적으로 같다는 것을 알 수 있다.

깁스 샘플링은 위의 (4.27)식와 (4.28)식을 결정하기 위해 사용될 수 있는 일종의 $MCMC$방법이다. 깁스 샘플링을 수행하기 위해서는 다음과 같은 깁스 조건부 분포들이 필요하다.

(i) $[\beta | \theta, \sigma_v^2, \hat{\theta}] \sim N_p[(\sum_i z_i z_i^T)^{-1} (\sum_i z_i \theta_i), \sigma_v^2 (\sum_i z_i z_i^T)^{-1}]$

(ii) $[\theta_i | \beta, \sigma_v^2, \hat{\theta}] \sim N[\tilde{\theta}_i^B(\beta, \sigma_v^2), g_{1i}(\sigma_v^2) = \gamma_i \psi_i]$

(iii) $[\tau_v = \sigma_v^{-2} | \beta, \theta, \hat{\theta}] \sim G(\tilde{a}, \tilde{b})$,

단, $\tilde{a}=\frac{1}{2}\sum_i(\theta_i-z_i^T)^2+a$, $\tilde{b}=\frac{m}{2}+b$.

깁스 앨고리즘은 다음 절차에 의해 이루어진다.

(a) $\theta_i=\theta_i^{(0)}$, $\sigma_v^2=\sigma_v^{2(0)}$을 초기값으로 하여 (i)로부터 $\beta^{(1)}$을 계산

(b) $\beta=\beta^{(1)}$, $\sigma_v^2=\sigma_v^{2(0)}$을 이용하여 (ii)로부터 $\theta_i^{(1)}$을 계산

$(i=1,2,\cdots,m)$

(c) $\theta_i=\theta_i^{(1)}$ 과 $\beta=\beta^{(1)}$을 이용하여 (iii)으로부터 $\sigma_v^{2(1)}$을 계산

(d) 절차 (a), (b), (c)를 한 사이클로 하여 반복 수행

수렴이 이루어지는 시점 t까지 충분히 반복한 후, 이 후부터 얻어지는 J개의 표본 $\{\beta^{(t+j)},\sigma_v^{2(t+j)},\theta_1^{(t+j)},\cdots,\theta_m^{(t+j)};j=1,2,\cdots,J\}$을 β, σ_v^2 , θ_1 , $\cdots$, θ_m의 결합 사후분포로 얻은 표본으로 간주한다. 초기값은 보통 $\theta_i^{(0)}=\tilde{\theta}_i^{EB}$, $\sigma_v^{2(0)}=\sigma_v^2$의 $REML$추정량을 사용한다.

위에서 계산된 J개의 표본을 이용하여 θ_i의 사후평균, 사후분산을 다음과 같이 추정한다.

$$\begin{aligned}\tilde{\theta}_i^{HB}&\approx\frac{1}{J}\sum_j\tilde{\theta}_j(\sigma_v^{2(t+j)})\\&=\frac{1}{J}\sum_j\tilde{\theta}_i(j)\\&=\bar{\theta}_i(\cdot)\,,\end{aligned}\tag{4.29}$$

$$V(\theta_i\mid\hat{\theta})\approx\frac{1}{J}\sum_j[g_{1i}(\sigma_v^{2(t+j)})+g_{2i}(\sigma_v^{2(t+j)})]+\frac{1}{J}\sum_j[\tilde{\theta}_i(j)-\bar{\theta}(\cdot)]^2\tag{4.30}$$

제5장 시군구 경활통계 작성

5.1 개 요

통계청에서는 취업, 실업 등과 같은 경제적 특성을 조사하여 국가의 고용정책 입안과 평가에 필요한 기초자료를 수집할 목적으로 매월 3만 표본 가구 내에 거주하는 만 15세 이상인 사람들을 대상으로 경제활동인구 조사를 실시하고 있다. 매월 15일을 포함하는 주 중에 표본 가구 내에 거주하는 사람들의 취업, 실업 및 비 경제활동인구 관련 사항을 방문면접이나 컴퓨터면접 방식으로 조사한 후 조사된 자료를 직접 추정방법에 의해 추정 계산하여 매 익월 말경에 대영역인 7개 광역시와 9개 도에 대해 조사 결과를 발표하고 있다.

1995년부터 시작된 지방자치제도와 1997년 IMF 사태에 기인하여 최근에는 실업 관련 통계뿐만 아니라 다양한 분야의 통계 작성도 시군구의 소지역 단위까지 작성해야 한다는 인식이 높아지고는 있지만 현재와 같은 대영역 표본설계를 기반으로 하는 통계 작성 방법으로는 신뢰성 있는 소지역 통계 생산은 불가능한 실정이다. 왜냐하면 소지역 통계의 작성은 단순히 추정단계에서 추정량의 선택을 통해서 해결될 수 있는 것이 아니라 통계조사의 계획, 표본설계, 추정 등 통계조사 전 과정을 종합적으로 고려할 때 가능한 일이기 때문이다.

통계청의 경제활동인구조사는 대영역인 광역시와 도별 통계 작성을 목적으로 표본설계 되었기 때문에 소지역인 시군구 단위는 표본설계에 반영된 관심영역이 아니다. 따라서 현재 활용되고 있는 대영역 기반의 표본설계를 이용하여 소지역 통계를 직접 생산할 경우 시군구 지역에 배정된 표본조사구 수가 불균형적이고 특정 시군구 단위의 소지역에 대해서는 표본조사구의 수가 하나 내지 두 개 정도로 너무 작게 배정되어 있기 때문에 신뢰할만한 소지역 단위의 통계생산은 어렵게 된다.

우리의 목적은 대영역 기반 표본설계의 구조 하에서 직접 생산된 시군구 단위의 직접추정값들을 소지역 추정에 이용되는 명시적인 모형을 통해

보정하여 어느 정도 신뢰성을 확보하는 일이다. 이를 위하여 사전에 소지역에 대한 층화, 표본 배정, 집락화의 수준에 대한 검토 후 직접추정값을 보정할 수 있는 다양한 보조정보들을 고려하였다.

이 보고서에서는 현재 통계청에서 실시하고 있는 대영역 기반 경제활동인구조사의 표본설계 하에서 직접적으로 생산된 시군구 단위의 소지역에 대한 직접추정값들을 합성추정법 또는 복합추정법과 같은 설계 기반 소지역 추정법을 통해 보정하여 시군구의 실업자 총계를 추론한다. 표본조사구 수가 불균형으로 배정된 상태에서 추정된 소지역의 월별 직접추정값들에 대해 센서스 및 행정보고자료를 통해 선택된 보조정보를 이용하여 직접추정값들에 대한 보정을 시도하였다.

"Borrow Strength"를 적용하기 위하여 대영역별 경제활동인구조사 자료를 크게 시군구 그룹으로 크게 3개 그룹으로 분할하고, 각 그룹 내에서는 유사성질 범주를 성별(남, 여)-연령대별(15-29세, 30세 이상)로 총 4개의 범주로 구분하였다. 각 그룹별 4개 범주에 대해서는 범주별 실업자 총계를 산출하여 대영역 내의 시군구 단위에 대한 실업자 총계 추정을 위한 보조정보로 활용하였다.

5.2 소지역 추정을 위한 집락화 방안

경제활동인구조사는 대영역인 특·광역시 및 도지역의 총 16개 대영역에 대한 경제활동인구 특성을 파악할 목적으로 표본설계 되었다. 현재의 경제활동인구조사에서 특·광역시 및 도지역에 배정된 표본조사구 수는 총 1,271개이며 이들 대영역에 대한 표본조사구 수의 분포는 다음 <표5.1>과 같다.

<표5.1> 특·광역시 및 도지역의 표본조사구 수 분포

시도	주민등록 인구	시군구 수	표본조사구		시도	주민등록 인구	시군구 수	표본조사구	
			조사구 수	구성비 (%)				조사구 수	구성비 (%)
서울	10,331,244	25	161	12.67	충북	1,504,518	12	66	5.19
부산	3,786,033	16	93	7.32	충남	1,928,088	16	71	5.59
대구	2,539,587	8	73	5.74	전북	2,013,923	14	63	4.96
인천	2,581,557	10	95	7.47	전남	2,104,052	22	72	5.66
광주	1,387,360	5	88	6.92	경북	2,802,597	23	75	5.90
대전	1,408,809	5	65	5.11	경남	3,124,123	20	65	5.11
울산	1,060,378	5	43	3.38	제주	547,964	4	37	2.91
경기	9,612,036	31	125	9.83	합계	48,289,173	234	1,271	100.00
강원	1,556,904	18	79	6.22					

* 주민등록인구는 2001년 12월 기준

<표5.1>에서 대영역 내의 시군구 단위 소지역들의 수는 총 234개이며 이 들 시군구 단위 소지역들 각각에 대해 실업자 총계를 추정할 경우 시군구 단위 소지역들에 배정된 표본조사구의 수가 충분하다면 이러한 소지역들의 실업자 총계 추정값들은 신뢰성을 확보할 수 있다. 그러나 경제활동인구조사의 표본설계는 대영역의 경제활동 특성을 파악할 목적으로 설계되었기 때문에 시군구 단위의 소지역들에 배정된 표본조사구의 수는 매우 불균형적인 분포를 나타낸다. 따라서 표본조사구 수가 충분히 배정되지 못한 특정 소지역들의 경우 실업자 총계 추정값들의 추정오차는 신뢰할 수 없을 정도로 매우 큰 값을 나타낼 가능성이 있다. 표본조사구 수가 충분하지 못한 시군구단위 소지역 추정값들의 정확성을 제고하기 위한 방안으로 대영역 내의 시군구단위들을 연구변수들에 의해 집락화하여

특성이 유사한 인근지역의 정보를 합성하여 실업자 총계를 추정하는 방안이 고려될 수 있다.

대영역인 강원도의 시군단위 소지역들을 예를 들어 보자. 다음 <표5.2>는 강원도 내의 18개 시군단위 소지역들에 대한 표본조사구 수의 분포를 나타낸다.

<표5.2> 강원도 시군단위 소지역들의 표본조사구 수 분포

순번	시군구	주민등록 인구	표본조사구		순번	시군구	주민등록 인구	표본조사구	
			조사구 수	구성비 (%)				조사구 수	구성비 (%)
1	춘천시	252,019	12	15.19	11	고성군	34,728	2	2.53
2	홍천군	74,224	3	3.80	12	양양군	30,276	3	3.80
3	철원군	52,224	2	2.53	13	동해시	103,933	3	3.80
4	화천군	25,026	2	2.53	14	태백시	56,254	2	2.53
5	양구군	23,274	0	0.00	15	삼척시	79,862	4	5.06
6	인제군	33,133	2	2.53	16	영월군	52,151	4	5.06
7	원주시	275,217	13	16.46	17	평창군	46,564	3	3.80
8	횡성군	45,701	2	2.53	18	정선군	49,111	4	5.06
9	강릉시	232,664	16	20.25	합계		1,556,904	79	100.00
10	속초시	90,543	2	2.53					

표본조사구 수가 3개 이상 배정된 시군단위 소지역들에 대한 실업자 총계 추정값들은 앞으로 소개될 소지역추정법인 5.3절의 합성추정법과 5.4절의 복합추정법을 적용하면 추정값들의 신뢰성은 상당 부분 확보될 수 있다. 그러나 표본조사구 수가 2개 이하인 철원군, 화천군, 인제군, 횡성군, 속초시, 고성군, 태백시의 7개 시군단위 소지역들(표본조사구 수가 배정되지 않은 양구군은 추정에서 제외함)은 5.3절과 5.4절에서 소개되는 소지역추정법(합성추정법, 복합추

정법)을 적용하더라도 실업자 총계 추정값의 추정오차는 매우 큰 값을 나타낸다.

이와 같이 대영역 내에서 표본조사구 수가 충분하지 못한 시군 단위 소지역들의 실업자 총계를 추정할 때 다음과 같은 대안이 고려될 수 있다. 예를 들어 표본조사구 수가 2개 이하인 횡성군(2개)과 고성군(2개)의 경우, 이들 두지역의 실업률이 매우 유사하다고 가정할 수 있다면, 이 때 횡성군과 고성군을 하나의 그룹으로 묶어 실업자 총계를 좀 더 안정적으로 추정한 후 인구비례로 횡성군과 고성군에 각각 추정값을 배분하여 주는 방법을 생각할 수 있다. 이러한 방법은 표본조사구 수가 적은 소지역들에 대한 실업자 총계 추정 시 매우 합리적인 대안으로 이용될 수 있다. 대영역 내에서 실업률이 유사한 시군단위 소지역들을 그룹으로 묶을 경우 집락분석 결과를 이용할 수 있다.

집락분석의 정확성을 기하기 위해 2000년 「인구주택 총 조사」 자료를 이용하였다. 「인구주택 총 조사」 자료를 이용하여 강원도 내에서 실업률이 유사한 시군단위 소지역들을 집락화하면 다음 <표5.3>과 같이 주어진다. 보다 정확한 분석을 위해 "남자 15~29세 실업률", "남자 30세이상 실업률", "여자 15~29세 실업률", "여자 30세이상 실업률"의 4개 연구변수들을 집락분석에 이용하였다.

<표5.3> 강원도 시군단위 소지역들의 집락분석 결과

( ):표본조사구 수

구 분	집락1	집락2	집락3	집락4
시군구	철원군(2), 화천군(2), 양구군(0), 인제군(2)	강릉시(16), 삼척시(4), 횡성군(2), 영월군(4), 고성군(2)	원주시(13), 동해시(3), 태백시(2), 속초시(2)	춘천시(12), 홍천군(3), 평창군(3), 정선군(4), 양양군(3)

집락분석 결과에 의하면 강원도의 시군단위 소지역들은 실업률이 유사한 4개의 집락들로 구분되었다. 따라서 위의 집락분석 결과를 참조하여 표본조사구 수가 2개 이하인 7개 시군단위 소지역들("철원군(2개)", "화천군(2개)", "인제군(2개)", "횡성군(2개)", "속초시(2개)", "고성군(2개)", "태백시(2개)")을 실업률이 유사한 소지역들끼리 묶어 그룹별 실업자 총계를 추정한다면 좀 더 안정적인 추정결과를 얻을 수 있다. 같은 집락으로 분류된 "철원군(2)+화천군(2)+인제군(2)", "횡성군(2)+고성군(2)", "태백시(2)+속초시(2)"를 각각 하나의 그룹으로 묶어 충분한 표본조사구 수를 확보한 후에 소지역추정법을 적용한다면 보다 안정적으로 실업자 총계를 추정할 수 있고 추정오차를 줄일 수 있다.

경제활동인구조사에서 시군구단위에 배정된 표본조사구 수가 2개 이하인 전국의 시군구단위들에 대해서 이상의 방법을 적용한 결과를 요약하면 다음과 같다.

① 서울특별시

<표5.4> 서울특별시 구단위 소지역들의 표본조사구 수 분포

순번	시군구	주민등록 인구	표본조사구		순번	시군구	주민등록 인구	표본조사구	
			조사구 수	구성비 (%)				조사구 수	구성비 (%)
1	종로구	187,038	4	2.48	14	마포구	382,195	5	3.11
2	*중구*	*146,335*	*1*	*0.62*	15	양천구	486,095	7	4.35
3	용산구	250,550	3	1.86	16	강서구	523,542	7	4.35
4	성동구	343,471	3	1.86	17	구로구	417,453	9	5.59
5	광진구	390,090	4	2.48	18	금천구	263,061	4	2.48
6	동대문구	383,822	9	5.59	19	영등포구	409,920	6	3.73
7	중랑구	449,965	10	6.21	20	동작구	407,793	7	4.35
8	성북구	453,517	9	5.59	21	관악구	529,741	8	4.97
9	강북구	352,317	4	2.48	22	서초구	397,983	7	4.35
10	도봉구	372,318	8	4.97	23	강남구	546,038	8	4.97
11	노원구	648,615	11	6.83	24	송파구	658,242	9	5.59
12	은평구	469,242	5	3.11	25	강동구	490,585	8	4.97
13	서대문구	371,316	5	3.11	합계		10,331,244	161	100.00

* 표본조사구 수가 2개 이하인 지역: "중구"

<표5.5> 서울특별시 구단위 소지역들의 집락분석 결과

( ): 표본조사구 수

구분	집락1	집락2	집락3	집락4	집락5
시군구	종로구(4), 서대문(5)	중구(1), 용산구(3), 성동구(3), 금천구(4)	동대문구(9), 성북구(9), 강북구(4), 노원구(11), 양천구(7)	중랑구(10), 도봉구(8), 은평구(5), 강서구(7), 구로구(9), 동작구(7), 관악구(8), 서초구(7), 강남구(8), 송파구(9), 강동구(8)	광진구(4), 마포구(5), 영등포구(6)

<표5.4>와 <표5.5>의 결과로부터 표본조사구 수가 2개만이 배정된 "중구"지역은 실업률이 유사한 같은 집락 내의 "용산구(3)"지역과 합성하여 소지역추정법을 적용하여 추정한다. 즉, "중구(1)+용산구(3)"의 그룹에서 실업자 총계를 추정한 후 각 지역별 실업자 수는 상주추정인구비례로 배분한다.

② 부산광역시

<표5.6> 부산광역시 구/군단위 표본조사구 수 분포

순번	시군구	주민등록 인구	표본조사구		순번	시군구	주민등록 인구	표본조사구	
			조사구 수	구성비 (%)				조사구 수	구성비 (%)
1	*중구*	*57,658*	*1*	*1.08*	10	사하구	385,967	9	9.68
2	서구	152,752	4	4.30	11	금정구	287,730	5	5.38
3	동구	125,085	4	4.30	12	강서구	60,773	3	3.23
4	영도구	180,626	5	5.38	13	연제구	226,768	6	6.45
5	부산진구	427,511	10	10.75	14	수영구	181,777	5	5.38
6	동래구	295,915	8	8.60	15	사상구	302,349	8	8.60
7	남구	303,301	8	8.60	16	*기장군*	*76,721*	*1*	1.08
8	북구	311,893	9	9.68	합계		3,786,033	93	100.00
9	해운대구	409,207	7	7.53					

* 표본조사구 수가 2개 이하인 지역: "중구(1개)", "기장군(1개)"

<표5.7> 부산광역시 구/군단위 소지역들의 집락분석 결과

( ): 표본조사구 수

구 분	집락1	집락2	집락3	집락4
시군구	중구(1), 동래구(8), 기장군(1)	서구(4), 영도구(5), 남구(8), 사하구(9), 금정구(5),연제구(6), 수영구(5)	동구(4), 부산진구(10), 북구(9), 해운대구(7), 사상구(8)	강서구(3)

<표5.6>과 <표5.7>의 결과를 참조하면 "중구(1)"와 "기장군(1)"지역은

각각 표본조사구 수가 1개 이므로 같은 집락에 속해있는 "동래구(8개)"지역을 이들 지역들과 합성하여 "중구(1)+기장군(1)+동래구(8)"로 묶어 하나의 그룹으로 추정한 후 상주추정인구 비례로 실업자 수를 배분한다.

③ 경기도

<표5.8> 경기도 시/군단위 표본조사구 수 분포

순번	시군구	주민등록 인구	표본조사구		순번	시군구	주민등록 인구	표본조사구	
			조사구 수	구성비 (%)				조사구 수	구성비 (%)
1	수원시	978,698	11	8.80	17	포천군	148,452	2	1.60
2	안양시	593,967	8	6.40	18	부천시	785,754	8	6.40
3	안산시	598,560	4	3.20	19	광명시	337,175	8	6.40
4	과천시	71,525	1	0.80	20	시흥시	342,351	2	1.60
5	오산시	115,161	0	0.00	21	김포시	183,156	9	7.20
6	군포시	270,326	2	1.60	22	평택시	361,992	5	4.00
7	의왕시	124,772	3	2.40	23	안성시	142,799	2	1.60
8	용인시	455,118	6	4.80	24	고양시	814,493	4	3.20
9	화성시	214,729	5	4.00	25	파주시	226,858	5	4.00
10	성남시	937,780	7	5.60	26	구리시	185,494	1	0.80
11	하남시	124,018	3	2.40	27	남양주시	376,231	6	4.80
12	광주시	154,808	1	0.80	28	가평군	56,211	2	1.60
13	의정부시	368,887	4	3.20	29	이천시	188,367	4	3.20
14	동두천시	75,699	1	0.80	30	여주군	105,084	2	1.60
15	양주군	138,748	3	2.40	31	양평군	82,921	4	3.20
16	연천군	51,902	2	1.60	합계		9,612,036	125	100.00

* 표본조사구 수가 2개 이하인 지역: "과천시(1)", "군포시(2)", "포천군(2)", "시흥시(2)", "안성시(2)", "구리시(1)", "광주시(1)", "동두천시(1)", "연천군(2)", "가평군(2)", "여주군(2)"

* 표본조사구가 배정되지 않은 오산시는 분석에서 제외함

<표5.9> 경기도 시/군단위 소지역들의 집락분석 결과

( ): 표본조사구 수

구분	집락1	집락2	집락3	집락4	집락5	집락6
시군구	평택시(5), 용인시(6), 파주시(5), 이천시(4), 안성시(2), 광주군(1), 포천군(2), 가평군(2), 양평군(4)	수원시(11), 성남시(7), 안양시(8), 부천시(8), 동두천시(1), 고양시(4), 과천시(1), 구리시(1), 군포시(2), 하남시(3)	연천군(2)	김포시(9), 여주군(2), 화성시(5)	의정부(4), 광명시(8), 의왕시(3), 양주군(3)	안산시(4), 남양주(6), 오산시(0), 시흥시(2)

<표5.8>과 <표5.9>를 참조하여 표본조사구 수가 2개 이하인 시/군단위 소지역들을 실업률이 유사한 지역들로 합성하면 "안성시(2)+광주군(1)", "포천군(2)+가평군(2)", "동두천시(1)+과천시(1)+구리시(1)+군포시(2)", "시흥시(2)+안산시(4)"의 그룹들로 묶을 수 있다. 단, "연천군(2)"은 동일집락 내에 실업률이 유사한 시/군단위 소지역들이 없으므로 유사 인근지역인 "여주군(2)"과 합성하여 하나의 그룹으로 묶은 후 실업자 총계를 추정한다.

④ 강원도

(강원도는 앞서 설명한 결과를 참조)

⑤ 충청북도

<표5.10> 충북 시/군단위 소지역들의 표본조사구 수 분포

순번	시군구	주민등록 인구	표본조사구		순번	시군구	주민등록 인구	표본조사구	
			조사구 수	구성비 (%)				조사구 수	구성비 (%)
1	청주시	594,716	24	36.36	8	음성군	88,783	6	9.09
2	청원군	125,221	5	7.58	9	*단양군*	*38,797*	*2*	*3.03*
3	*진천군*	*61,483*	*1*	*1.52*	10	*보은군*	*42,215*	*2*	*3.03*
4	괴산군	43,220	5	7.58	11	옥천군	59,836	3	4.55
5	*증평출장소*	*31,574*	*0*	*0.00*	12	영동군	57,319	3	4.55
6	충주시	216,036	11	16.67	합계		1,504,518	66	100.00
7	제천시	145,317	4	6.06					

* 표본조사구 수가 2개 이하인 지역: "진천군(1)", "증평출장소(0)", "단양군(2)", "보은군(2)"

* 표본조사구 수가 배정되지 않은 "증평출장소(0)" 지역은 분석에서 제외함

<표5.11> 충북 시/군단위 소지역들의 집락분석 결과

( ): 표본조사구 수

구분	집락1	집락2	집락3
시군구	청주시(24), 충주시(11), 제천시(4), 청원군(5), 옥천군(3), 진천군(1), 괴산군(5), 음성군(6), 단양군(2)	영동군(3), 증평출장소(0)	보은군(2)

<표5.10>과 <표5.11>로부터 표본조사구 수가 2개 이하인 시/군단위 소지역들 중 "진천군(1)"과 "단양군(2)"은 실업률이 서로 유사한 동일집락에 포함되므로 "진천군(1)+단양군(2)"의 그룹으로 합성하여 실업자 총계를

추정할 수 있다. 그러나 "보은군(2)"의 경우는 같은 집락 내에 유사 시/군 지역들이 없고 또한 실업률이 유사한 인근 시/군단위 소지역들을 선택하기가 곤란하므로 다소 추정오차가 커지더라도 단독으로 추정한다.

⑥ 충청남도

<표5.12> 충남 시/군단위 소지역들의 표본조사구 수 분포

순번	시군구	주민등록 인구	표본조사구		순번	시군구	주민등록 인구	표본조사구	
			조사구 수	구성비 (%)				조사구 수	구성비 (%)
1	공주시	134,383	6	8.45	10	*서천군*	*72,903*	*2*	*2.82*
2	논산시	140,793	9	12.68	11	*청양군*	*40,086*	*2*	*2.82*
3	금산군	63,345	3	4.23	12	홍성군	93,790	3	4.23
4	*연기군*	*81,821*	*0*	*0.00*	13	서산시	150,504	7	9.86
5	천안시	436,708	14	19.72	14	태안군	67,947	3	4.23
6	아산시	188,372	7	9.86	15	*당진군*	*120,818*	*1*	*1.41*
7	예산군	100,602	4	5.63	16	*계룡출장소*	*28,883*	*0*	*0.00*
8	보령시	116,546	6	8.45	합계		1,928,088	71	100.00
9	부여군	90,587	4	5.63					

* 표본조사구 수가 2개 이하인 지역: "연기군(0)", "서천군(2)", "청양군(2)", "당진군(1)", "계룡출장소(0)"

* 표본조사구 수가 배정되지 않은 "연기군(0)"과 "계룡출장소(0)"는 분석에서 제외함

<표5.13> 충남 시/군단위 소지역들의 집락분석 결과

( ): 표본조사구 수

구분	집락1	집락2	집락3	집락4
시군구	공주시(6), 금산군(3), 청양군(2)	서천군(2)	보령시(6), 논산시(9), 예산군(4), 태안군(3), 당진군(1), 계룡출장소(0)	천안시(14), 아산시(7), 서산시(7), 연기군(0), 부여군(4), 홍성군(3)

<표5.12>와 <표5.13>의 시/군지역 표본조사구 수 분포와 집락분석결과를 토대로 표본조사구 수가 2개 이하인 "서천군(2)", "청양군(2)", "당진군(1)"지역들은 "서천군(2)+청양군(2)", "당진군(1)+태안군(3)"으로 합성하여 그룹별 실업자 총계를 추정한다.

⑦ 전라북도

<표5.14> 전북 시/군단위 소지역들의 표본조사구 수 분포

순번	시군구	주민등록 인구	표본조사구		순번	시군구	주민등록 인구	표본조사구	
			조사구 수	구성비 (%)				조사구 수	구성비 (%)
1	전주시	623,897	18	28.57	9	남원시	104,198	5	7.94
2	군산시	277,680	9	14.29	10	*임실군*	*37,548*	*1*	*1.59*
3	익산시	337,240	9	14.29	11	*순창군*	*34,003*	*2*	*3.17*
4	김제시	115,683	4	6.35	12	*진안군*	*32,750*	*1*	*1.59*
5	*완주군*	*87,081*	*2*	*3.17*	13	*무주군*	*30,385*	*0*	*0.00*
6	정읍시	152,452	4	6.35	14	*장수군*	*30,521*	*2*	*3.17*
7	고창군	74,413	3	4.76	합계		2,013,923	63	100.00
8	부안군	76,072	3	4.76					

* 표본조사구 수가 2개 이하인 지역: "완주군(2)", "임실군(1)", "순창군(2)", "진안군(1)", "무주군(0)", "장수군(2)"

* 표본조사구 수가 배정되지 않은 "무주군(0)"지역은 분석에서 제외함

<표5.15> 전북 시/군단위 소지역들의 집락분석 결과

( ): 표본조사구 수

구분	집락1	집락2	집락3
시군구	정읍시(4), 완주군(2), 진안군(1), 장수군(2), 임실군(1)	무주군(0), 순창군(2), 고창군(3)	전주시(18), 군산시(9), 익산시(9), 남원시(5), 김제시(4), 부안군(3)

위의 시/군단위 표본조사구 수 분포와 집락분석결과를 토대로 표본조사구 수가 2개 이하인 지역들은 실업률이 유사한 동일 집락 내의 시/군단위 소지역들로 합성하여 추정한다. 대상 시/군단위 소지역들은 "완주군(2)+진안군(1)", "장수군(2)+임실군(1)", "순창군(2)+고창군(3)"으로 합성하여 실업자 총계를 추정한다.

⑧ 전라남도

<표5.16> 전남 시/군단위 소지역들의 표본조사구 수 분포

순번	시군구	주민등록 인구	표본조사구		순번	시군구	주민등록 인구	표본조사구	
			조사구 수	구성비 (%)				조사구 수	구성비 (%)
1	담양군	53,672	2	2.78	13	나주시	106,431	3	4.17
2	곡성군	38,930	0	0.00	14	영암군	64,738	3	4.17
3	화순군	79,496	2	2.78	15	함평군	43,667	2	2.78
4	영광군	70,269	2	2.78	16	고흥군	95,960	4	5.56
5	장성군	54,958	1	1.39	17	보성군	59,526	1	1.39
6	목포시	245,666	14	19.44	18	장흥군	51,850	1	1.39
7	무안군	69,178	3	4.17	19	강진군	50,925	3	4.17
8	여수시	320,570	10	13.89	20	해남군	96,590	4	5.56
9	순천시	272,124	9	12.50	21	완도군	65,458	0	0.00
10	광양시	138,468	6	8.33	22	진도군	41,148	2	2.78
11	구례군	33,031	0	0.00	합계		2,104,052	72	100.00
12	신안군	51,361	0	0.00					

* 표본조사구 수가 2개 이하인 지역: "담양군(2)", "곡성군(0)", "화순군(2)", "영광군(2)", "장성군(1)", "구례군(0)", "신안군(0)", "함평군(2)", "보성군(1)", "장흥군(1)", "완도군(0)", "진도군(2)"

* 표본조사구 수가 배정되지 않은 "곡성군(0)", "구례군(0)", "신안군(0)", "완도군(0)"지역들은 분석에서 제외함

<표5.17> 전남 시/군단위 소지역들의 집락분석 결과

( ): 표본조사구 수

구분	집락1	집락2	집락3	집락4	집락5
시군구	목포시(14), 여수시(10), 순천시(9)	광양시(6), 곡성군(0), 구례군(0), 고흥군(4), 보성군(1), 화순군(2), 장흥군(1), 영광군(2), 완도군(0), 진도군(2)	나주시(3), 담양군(2), 해남군(4), 함평군(2), 장성군(1)	강진군(3), 무안군(3)	영암군(3), 신안군(0)

<표5.16>과 <표5.17>을 참조하여 표본조사구 수가 2개 이하인 시/군단위 소지역들을 다음과 같은 그룹으로 합성하여 실업자 총계를 추정한다. "보성군(1)+화순군(2)+장흥군(1)"을 하나의 그룹으로, "영광군(2)+진도군(2)"을 하나의 그룹으로, "담양군(2)+함평군(2)+장성군(1)"을 하나의 그룹으로 합성하여 그룹별 실업자 총계를 추정한다.

⑨ 경상북도

<표5.18> 경북 시/군단위 소지역들의 표본조사구 수 분포

순번	시군구	주민등록 인구	표본조사구		순번	시군구	주민등록 인구	표본조사구	
			조사구 수	구성비 (%)				조사구 수	구성비 (%)
1	*영천시*	*119,077*	*2*	*2.67*	13	*청송군*	*33,568*	*1*	*1.33*
2	경산시	218,638	3	4.00	14	*영양군*	*22,286*	*1*	*1.33*
3	*군위군*	*34,293*	*2*	*2.67*	15	*봉화군*	*41,452*	*2*	*2.67*
4	*청도군*	*51,471*	*0*	*0.00*	16	김천시	151,764	3	4.00
5	*고령군*	*37,498*	*1*	*1.33*	17	구미시	348,489	9	12.00
6	*성주군*	*50,933*	*1*	*1.33*	18	상주시	122,277	5	6.67
7	칠곡군	107,158	3	4.00	19	문경시	89,234	3	4.00
8	포항시	516,576	14	18.67	20	*예천군*	*58,217*	*2*	*2.67*
9	경주시	288,915	5	6.67	21	*영덕군*	*49,674*	*2*	*2.67*
10	안동시	182,082	4	5.33	22	*울진군*	*65,878*	*1*	*1.33*
11	영주시	128,924	8	10.67	23	*울릉군*	*9,950*	*0*	*0.00*
12	의성군	74,243	3	4.23	합계		2,802,597	75	100.00

* 표본조사구 수가 2개 이하인 지역: "영천시(2)", "군위군(2)", "청도군(0)", "고령군(1)", "성주군(1)", "청송군(1)", "영양군(1)", "봉화군(2)", "예천군(2)", "영덕군(2)", "울진군(1)", "울릉군(0)"

* 표본조사구 수가 배정되지 않은 "청도군(0)", "울릉군(0)" 지역은 분석에서 제외함

<표5.19> 경북 시/군단위 소지역들의 집락분석 결과

( ): 표본조사구 수

구분	집락1	집락2	집락3	집락4	집락5
시군구	영주시(8), 상주시(5), 의성군(3), 청송군(1), 영덕군(2), 청도군(0), 고령군(1), 봉화군(2)	포항시(14), 안동시(4), 영천시(2), 문경시(3), 군위군(2), 영양군(1)	경주시(5), 김천시(3), 경산시(3), 성주군(1)	울릉군(0)	구미시(9), 칠곡군(3), 예천군(2), 울진군(1)

<표5.18>과 <표5.19>로부터 표본조사구 수가 충분하지 못한 시/군단위 소지역들을 "청송군(1)+영덕군(2)", "고령군(1)+봉화군(2)", "영천시(2)+문경시(3)", "군위군(2)+영양군(1)", "경산시(3)+성주군(1)", "예천군(2)+울진군(1)"과같이 실업률이 유사한 그룹들로 묶어 그룹별 추정을 실시한다.

⑩ 경상남도

<표5.20> 경남 시/군단위 소지역들의 표본조사구 수 분포

순번	시군구	주민등록 인구	표본조사구		순번	시군구	주민등록 인구	표본조사구	
			조사구 수	구성비 (%)				조사구 수	구성비 (%)
1	창원시	528,152	8	12.30	12	*산청군*	*39,915*	*1*	*1.54*
2	마산시	434,912	11	16.92	13	통영시	135,845	4	6.15
3	*진해시*	*135,539*	*2*	*3.08*	14	거제시	180,496	3	4.62
4	김해시	357,149	3	4.62	15	*고성군*	*62,383*	*1*	*1.54*
5	밀양시	123,393	4	6.15	16	*남해군*	*58,039*	*2*	*3.08*
6	양산시	202,784	4	6.15	17	*하동군*	*60,370*	*2*	*3.08*
7	*함안군*	*66,054*	*2*	*3.08*	18	*함양군*	*45,350*	*1*	*1.54*
8	*창녕군*	*71,215*	*1*	*1.54*	19	거창군	68,713	5	7.69
9	진주시	340,669	8	12.31	20	*합천군*	*59,890*	*1*	*1.54*
10	*사천시*	*119,555*	*1*	*1.54*	합계		3,124,123	65	100.00
11	*의령군*	*33,700*	*1*	*1.54*					

* 표본조사구 수가 2개 이하인 지역: "진해시(2)", "산청군(1)", "함안군(2)", "창녕군(1)", "사천시(1)", "의령군(1)", "고성군(1)", "남해군(2)", "하동군(2)", "함양군(1)", "합천군(1)

<표5.21> 경남 시/군단위 소지역들의 집락분석 결과

( ): 표본조사구 수

구분	집락1	집락2	집락3	집락4
시군구	통영시(4), 함안군(2), 하동군(2), 거창군(5)	진주시(8), 사천시(1)	창원시(8), 마산시(11), 진해시(2), 김해시(3), 밀양시(4), 거제시(3), 양산시(4), 의령군(1), 합천군(1)	창녕군(1), 고성군(1), 산청군(1), 함양군(1), 남해군(2)

경상남도의 시/군단위 소지역들의 합성결과는 다음과 같다. 경상남도에서는 "함안군(2)+하동군(2)", "진주시(8)+사천시(1)", "진해시(11)+마산시(2)", "양산시(4)+의령군(1)+합천군(1)", "창녕군(1)+고성군(1)+산청군(1)", "함양군(1)+남해군(2)"과 같이 소지역들을 합성하여 그룹별 실업자 총계를 추정한다.

5.3 직접추정법

대영역 내에서 소지역 i에 대한 실업자 총계를 추정하기 위한 직접추정량은 다음과 같은 총계 추정공식이 이용된다.

$$\hat{Y}_{i\cdot} = \sum_{s=1}^{2} {}_s\hat{Y}_{i\cdot} \quad , \quad i=1,2,\cdots,I; s=1,2; h=1,2,\cdots,n_i$$

$$= \sum_{s=1}^{2}\sum_{h=1}^{n_i} {}_s\hat{Y}_{ih}$$

$$= \sum_{s=1}^{2}\sum_{h=1}^{n_i} {}_sM_i \, {}_sY_{ih} \quad , \qquad (5.1)$$

여기에서 s는 성별(남, 여)을 나타내는 첨자, n_i는 경제활동인구조사에서

소지역 i의 표본조사구 수, ${}_sY_{ih}$는 각 성별에 대해서 소지역 i의 표본 조사구에서 조사한 실업자 수를 나타낸다. 승수 ${}_sM_i = {}_s\hat{X}_{i\cdot} / {}_sX_{i\cdot}$은 $\hat{Y}_{i\cdot}$이 불편추정량이 되도록 산정 한다. 여기에서 ${}_s\hat{X}_{i\cdot}$은 소지역 i에 대한 15세 이상의 상주 추계인구를 나타내며, ${}_sX_{i\cdot}$는 경제활동인구조사에서 조사된 15세 이상의 조사인구를 나타낸다.

직접추정량 $\hat{Y}_{i\cdot}$의 분산은 다음과 같이 주어진다.

$$Var(\hat{Y}_{i\cdot}) = \sum_{s=1}^{2} Var({}_s\hat{Y}_{i\cdot}) + 2Cov({}_1\hat{Y}_{i\cdot}, {}_2\hat{Y}_{i\cdot})$$

$$= \sum_{s=1}^{2} {}_sM_i^2 \, Var(\sum_{h=1}^{n_i} {}_sY_{ih}) + 2 \, {}_1M_i \, {}_2M_i \, Cov(\sum_{h=1}^{n_i} {}_1Y_{ih}, \sum_{h=1}^{n_i} {}_2Y_{ih})$$

소지역 i에 대한 직접추정량 $\hat{Y}_{i\cdot}$의 분산에 대한 추정은 통계청의 연속차 분산추정공식을 적용하면 다음과 같이 계산된다.

$$\widehat{Var}(\hat{Y}_{i\cdot}) = \sum_{s=1}^{2} {}_sM_i^2 (\zeta_i \sum_{h=1}^{n_i} {}_sU_{ih}^2) + 2 \, {}_1M_i \, {}_2M_i (\zeta_i \sum_{h=1}^{n_i} {}_1U_{ih} \, {}_2U_{ih}) \qquad (5.2)$$

여기에서

$${}_sU_{ih} = d_{\,s}Y_{ih} - {}_s\rho_i \cdot d_{\,s}X_{ih} \ ,$$

$$d_{\,s}Y_{ih} = {}_sY_{ih} - {}_s\,Y_{i,h+1} \ ,$$

$$d_{\,s}X_{ih} = {}_s\,X_{ih} - {}_s\,X_{i,h+1} \ ,$$

$${}_s\rho_i = {}_sY_{i\cdot} / {}_sX_{i\cdot} \ ,$$

$$\zeta_i = [1 - n_i/(10N_i)]\, n_i/[2(n_i - 1)] \ ,$$

$N_i =$ 소지역 i에 대한 표본추출틀의 조사구 수.

경제활동인구조사에서 대영역에 포함된 소지역들은 표본설계에 반영된 관심영역이 아니다. 따라서 대영역 표본설계에 기반을 둔 표본 조사로부터 소지역들에 대한 관심 통계량들을 추정한다면 소지역에 할당된 표본 조사구 수가 충분하지 않기 때문에 신뢰할 만한 결과를 얻을 수 없게 된다. 이러한 관점에서 현행 경제활동인구조사 자료로부터 소지역 추정값들

의 신뢰성을 확보하기 위한 다음과 같은 합성추정법 및 복합추정법이 제안될 수 있다.

5.4 합성 추정법

대영역 표본설계에 기반을 둔 통계청의 직접추정량은 각 소지역에 할당된 표본 조사구의 수가 충분하지 않기 때문에 소지역 실업통계의 정확도를 제공하지는 못한다. 소지역 i에 대한 합성추정량 $\hat{Y}_{i\cdot}^{S}$는 소지역 i와 특성이 유사한 인근 지역의 정보를 추정에 이용하는 간접적인 설계 기반 추정량이다.

우선 대영역 내에 I개의 시군구 단위의 소지역들이 있다고 가정하자. "Borrow Strength"를 적용하기 위해 대영역을 특성이 유사한 시단위, 군단위 및 구단위들의 3개의 그룹으로 분할하고, 각 그룹들을 4개의 성별(남, 여)-연령대별(15-29세, 30세이상) 범주로 구분한다. 여기에서 $I=\sum_{k=1}^{3} I_k$이며, 시, 군 및 구 그룹들은 각각 I_1 , I_2 , I_3개의 동질적인 소지역 단위들로 구성된다.

"Borrow Strength"를 적용하여 정의되는 소지역 i에 대한 합성추정량 $\hat{Y}_{i\cdot}^{S}$는 해당 소지역과 유사한 정보를 갖는 인근 소지역들의 정보를 이용하여 추정되므로 추정오차는 직접추정량에 비해 현저하게 줄어들 수 있으나 해당 소지역과 인근 유사지역의 정보가 동질적이지 못할 경우 편향이 발생할 가능성이 있다.

합성추정량을 정의하기 위해 다음과 같은 기호들이 이용되었다.

N_i = 표본추출틀에서 소지역 i의 조사구 수,

n_i = 경제활동인구조사에서 소지역 i에 할당된 표본조사구 수,

${}_jP_{t,i}^{C}$ = t년 센서스로부터 추계된 j범주에 대한 소지역 i의 상주인구,

${}_jP_{t,i}^{R}$ = j범주에 대한 소지역 i의 t년 주민등록인구,

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

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

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

I_k개의 소지역들을 포함하고 있는 부차관심영역들인 각 시, 군, 구 그룹들을 특성 기준에 따라 유사성을 갖는 4개의 성별(남, 여)-연령대별(15-29세, 30세이상) 범주들로 구분할 때 각 그룹 내에서 소지역 i의 실업자 총계에 대한 합성추정량 $\hat{Y}^S_{i\cdot}$는 다음과 같이 주어질 수 있다.

$$\hat{Y}^S_{i\cdot} = \sum_{j=1}^{4} \frac{{}_j\hat{P}_i}{{}_j\hat{X}_\cdot} {}_j\hat{Y}_{dir}, \qquad i=1,2,\cdots,I_k, \qquad (5.4)$$

여기에서

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

$${}_j\hat{X}_\cdot = \sum_{i=1}^{I_k} {}_j\hat{X}_i,$$

$${}_j\hat{Y}_{dir} = \sum_{i=1}^{I_k} \sum_{h=1}^{n_k} {}_jM_i\, {}_jY_{ih}$$

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

합성추정량 $\hat{Y}^S_{i\cdot}$에 대한 정확도의 측도로써 다음과 같은 합성추정량의

평균제곱오차를 고려할 수 있다.

$$MSE(\hat{Y}^S_{i\cdot}) = Var(\hat{Y}^S_{i\cdot}) + [Bias(\hat{Y}^S_{i\cdot})]^2 .$$

위 식에서 $Var(\hat{Y}^S_{i\cdot})$는 ${}_j\hat{P}_i/{}_j\hat{X} = const$ 를 가정한다면 다음 식과 같이 주어질 수 있다.

$$\begin{aligned} Var(\hat{Y}^S_{i\cdot}) = &\sum_{j=1}^{J} \left(\frac{{}_j\hat{P}_i}{{}_j\hat{X}_{\cdot}}\right)^2 Var({}_j\hat{Y}_{dir}) \\ &+ 2\sum_{j<l} \left(\frac{{}_j\hat{P}_i}{{}_j\hat{X}_{\cdot}}\right)\left(\frac{{}_l\hat{P}_i}{{}_l\hat{X}_{\cdot}}\right) Cov({}_j\hat{Y}_{dir}, {}_l\hat{Y}_{dir}) \end{aligned} \tag{5.5}$$

j범주의 실업자 총계에 대한 직접추정량은

$${}_j\hat{Y}_{dir} = \sum_{i=1}^{I_k}\sum_{h=1}^{n_{ij}} M_j Y_{ijh}$$

로 주어지므로 직접추정량 ${}_j\hat{Y}_{dir}$의 분산과 공분산은 연속차 분산추정방법에 의해 다음 식으로부터 추정될 수 있다.

$$\widehat{Var}({}_j\hat{Y}_{dir}) = M_j^2 \zeta_j \sum_{i=1}^{I_k}\sum_{h=1}^{n_{ij}} U_{ijh}^2 ,$$

$$\widehat{Cov}({}_j\hat{Y}_{dir}, {}_l\hat{Y}_{dir}) = M_j M_l \zeta_j \sum_{i=1}^{I_k}\sum_{h=1}^{n_{ij}} U_{ijh} U_{ilh} ,$$

여기에서

$$U_{ijh} = d_j Y_{ijh} - \frac{Y_{\cdot j\cdot}}{X_{\cdot j\cdot}} \cdot d_j X_{ijh} ,$$

$$d_j Y_{ijh} = Y_{ijh} - Y_{ij,h+1} ,$$

$$d_j X_{ijh} = X_{ijh} - X_{ij,h+1} ,$$

$$\zeta_j = (1-f_j)\, n_j / [2(n_j - 1)] ,$$

$$f_j = n_j/(10N_j)$$

로 주어진다.

따라서 합성추정량 $\hat{Y}_{i.}^{S}$의 추정분산은 다음과 같이 주어진다.

$$\widehat{Var}(\hat{Y}_{i.}^{S}) = \left(\frac{\hat{P}_{ij}}{\hat{X}_{.j}}\right)^2 \left(M_j^2 \zeta_j \sum_{i=1}^{I_k} \sum_{h=1}^{n_{ij}} U_{ijh}^2\right) + 2\sum_{j<l} \left(\frac{\hat{P}_{ij}}{\hat{X}_{.j}}\right)\left(\frac{\hat{P}_{il}}{\hat{X}_{.l}}\right)\left(M_j M_l \zeta_j \sum_{i=1}^{I_k} \sum_{h=1}^{n_{ij}} U_{ijh} U_{ilh}\right) \tag{5.6}$$

소지역 i에 대한 합성추정량 $\hat{Y}_{i.}^{S}$의 추정분산은 위의 (5.6)식과 같이 명시된 절차에 의해 계산될 수 있으나, 현행 경제활동인구조사 체계에서 편향에 대한 추정은 결코 쉬운 문제가 아니다. 소지역 i의 실업자 총계에 대한 참값을 센서스 자료를 이용하여 결정할 수는 있으나 시점 상으로 서로 상이한 양상을 보일 가능성이 있기 때문에 편향 추정에 직접적으로 이용될 수는 없다.

이러한 문제점에 기인하여 Ghosh and Rao (1994)는 $Cov(\hat{Y}_{i.}, \hat{Y}_{i.}^{S}) = 0$의 가정 하에서 합성추정량 $\hat{Y}_{i.}^{S}$의 평균제곱오차를 추정할 경우 다음과 같은 근사적인 불편추정량을 이용할 것을 제안한 바 있다.

$$mse(\hat{Y}_{i.}^{S}) \approx (\hat{Y}_{i.}^{S} - \hat{Y}_{i.})^2 - \widehat{Var}(\hat{Y}_{i.}) .$$

그러나 위의 추정량은 소지역에 배정된 표본조사구 수가 충분하지 못할 경우 직접추정값의 불안정에 기인하여 합성추정값의 평균제곱오차의 추정값이 음의 값이 나올 가능성도 있다. 따라서 소지역에 배정된 표본조사구 수가 충분하지 못한 한국의 경활조사 체계에 적용하기에는 무리가 있는 추정공식이다. 또 다른 추정량의 정확도에 대한 측도로써 대영역 내의 모든 소지역들에 대해 평균제곱오차 추정값들의 평균을 취하는 방법이 이용될 수 있으나 이 측도는 전자보다는 좀 더 안정적일 수는 있지만 각 소지역에 대한 평균제곱오차 추정값들을 제공하지는 못한다.

위에서 언급된 사실들을 보완할 수 있는 하나의 대안으로써 잭나이프

추정 방법이 고려될 수 있다. 잭나이프 추정방법의 첫 번째 단계는 경활조사 자료로부터 반복표본을 생성하는 것이다. 우선 소지역 i 내에서 하나의 표본조사구가 교대로 선택되어 표본으로부터 제거된 후 나머지 표본조사구들에 대해서 승수가 보정된다. 반복표본들은 조사구의 갯수 만큼 생성되며, 이 들 반복표본들을 이용하여 새로운 합성추정값들이 다시 계산된다. 구체적인 절차를 설명하면 다음과 같다.

(Ⅰ) 대영역 내에서 분할된 유사성질을 갖는 부차그룹(시, 군 및 구 그룹)에대해서 소지역 i로부터 h번째 표본조사구를 제거한 후, 실업자 조사 총계 Y에 대한 다음과 같은 반복표본을 생성한다.

$$S_{i(h)} = \{Y_{11}, \cdots, Y_{1n_1}; \cdots; Y_{i1}, \cdots, Y_{i,h-1}, Y_{i,h+1}, \cdots, Y_{in_i}; Y_{I1}, \cdots, Y_{In_I}\}.$$

기호 $i(h)$는 새로운 합성추정값을 계산하기 위해 소지역 i 로부터 h번째 표본조사구가 제거되었다는 것을 나타낸다. 소지역 i에 대해서 총 n_i개의 반복표본이 생성되며, k번째 부차그룹에 대한 반복표본의 개수는 총 $n = \sum_{i=1}^{I} n_i$개이다.

(Ⅱ) 주어진 소지역 i에서 해당 소지역 내에 남아있는 $n_i - 1$개 표본조사구의 전체 조사가구들에 대해서 승수에 대한 보정이 이루어진다. 보정된 승수값은 다음과 같다.

$$_jM_i^{adj} = \frac{n_i}{n_i - 1}\, {}_jM_i\,, \qquad j = 1,2,3,4$$

(Ⅲ) k 번째 부차그룹 내에 남아 있는 $n-1$개 표본조사구들을 이용하여 소지역 i에 대한 새로운 합성추정값 $\hat{Y}_{i.}^{S}(h)$를 계산한다.

위의 절차는 해당 부차관심영역 내에 있는 모든 표본조사구들에 대해서 반복되며, 이로부터 n개의 서로 다른 실업자 총계에 대한 합성추정값들이 생성된다. 소지역 i에 대해서는 n_i개의 서로 다른 합성추정값들이 얻어진다.

소지역 i의 실업자 총계에 대한 잭나이프 평균제곱오차 추정식은 다음과 같이 주어진다.

$$mse_J(\hat{Y}_{i\cdot}^S) = \widehat{Var}_J(\hat{Y}_{i\cdot}^S) + [\widehat{Bias}_J(\hat{Y}_{i\cdot}^S)]^2 \tag{5.7}$$

여기에서

$$\widehat{Var}_J(\hat{Y}_{i\cdot}^S) = \frac{n_i - 1}{n_i} \sum_{h=1}^{n_i} [\hat{Y}_{i\cdot}^S(h) - \frac{1}{n_i} \sum_{l=1}^{n_i} \hat{Y}_{i\cdot}^S]^2 , \tag{5.8}$$

$$\widehat{Bias}_J(\hat{Y}_{i\cdot}^S) = (n_i - 1)[\frac{1}{n_i} \sum_{h=1}^{n_i} \hat{Y}_{i\cdot}^S(h) - \hat{Y}_{i\cdot}^S] \tag{5.9}$$

로 주어진다.

5.5 복합 추정법

경제활동인구조사 자료로부터 직접적으로 추정된 소지역 i에 대한 직접추정량 $\hat{Y}_{i\cdot}$은 해당 소지역에 할당된 표본 수가 충분하지 않기 때문에 추정값들의 신뢰성을 확보할 수 없다. 또한 인근 지역의 유사정보를 이용하여 추정되는 합성추정량 $\hat{Y}_{i\cdot}^S$는 잠재적인 편향 가능성이 항상 내재되어 있다. 따라서 소지역에 배정된 표본 수가 적을 경우, 표본조사 자료만을 이용하여 추정되는 직접추정량의 불안정성과 합성추정량의 편향 가능성을 보완하기 위해 직접추정량 $\hat{Y}_{i\cdot}$와 합성추정량 $\hat{Y}_{i\cdot}^S$의 가중평균을 이용한 다음과 같은 복합추정량 $\hat{Y}_{i\cdot}^C$가 고려될 수 있다.

$$\hat{Y}_{i\cdot}^C = w_i \hat{Y}_{i\cdot} + (1 - w_i)\hat{Y}_{i\cdot}^S , \quad i = 1, 2, \cdots, I_k \tag{5.10}$$

여기에서 가중치 w_i는 0과 1사이의 값을 취한다.

이 때 복합추정량 $\hat{Y}_{i\cdot}^C$의 평균제곱오차는 다음 식과 같이 주어진다.

$$\begin{aligned} MSE(\hat{Y}_{i\cdot}^C) = {} & w_i^2 MSE(\hat{Y}_{i\cdot}) + (1 - w_i)^2 MSE(\hat{Y}_{i\cdot}^S) \\ & + 2w_i(1 - w_i) E(\hat{Y}_{i\cdot} - Y_{i\cdot}^*)(\hat{Y}_{i\cdot}^S - Y_{i\cdot}^*) \end{aligned} \tag{5.11}$$

여기에서 $Y_{i\cdot}^*$는 소지역 i의 실업자 총계에 대한 참값을 나타낸다. 위의

(5.11)식을 w_i의 함수로 가정하여 가중치 w_i에 대해서 미분하면, 평균제곱오차를 최소화하는 다음과 같은 가중치를 산정할 수 있다.

$$w^*_{i(opt)} = \frac{MSE(\hat{Y}^S_{i.}) - E(\hat{Y}_{i.} - Y^*_{i.})(\hat{Y}^S_{i.} - Y^*_{i.})}{MSE(\hat{Y}^S_{i.}) + MSE(\hat{Y}_{i.}) - 2E(\hat{Y}_{i.} - Y^*_{i.})(\hat{Y}^S_{i.} - Y^*_{i.})} \quad (5.12)$$

여기에서 직접추정량 $\hat{Y}_{i.}$은 불편추정량이 되도록 산정하므로 (5.12)식에서 평균제곱오차 $MSE(\hat{Y}_{i.})$은 $Var(\hat{Y}_{i.})$과 동일한 값을 갖는다. 또한 (5.12)식에서 $Cov(\hat{Y}_{i.}, \hat{Y}^S_{i.}) = 0$ 를 가정한다면, 가중치 $w^*_{i(opt)}$는 다음 식으로 근사될 수 있다.

$$w_{i(opt)} = \frac{MSE(\hat{Y}^S_{i.})}{MSE(\hat{Y}^S_{i.}) + Var(\hat{Y}_{i.})} \quad (5.13)$$

위의 (5.13)식에서 최적가중치 $w_{i(opt)}$는 경제활동인구조사 자료로부터 추정되어야할 값이다. $MSE(\hat{Y}^S_{i.})$에 대해서는 (5.7)식에서 주어진 $mse_J(\hat{Y}^S_{i.})$으로, $Var(\hat{Y}_{i.})$에 대해서는 (5.2)식에서 주어진 $\widehat{Var}(\hat{Y}_{i.})$으로 추정될 수 있고, 이때 최적가중치에 대한 추정식은 다음과 같이 주어진다.

$$\hat{w}_{i(opt)} = \frac{mse_J(\hat{Y}^S_{i.})}{mse_J(\hat{Y}^S_{i.}) + \widehat{Var}(\hat{Y}_{i.})} \quad (5.14)$$

따라서 (5.10)식의 복합추정량은 경제활동인구조사 자료로부터 추정된 최적가중치 $\hat{w}_{i(opt)}$를 이용하여 다음 식으로부터 추정될 수 있다.

$$\hat{Y}^C_{i.} = \hat{w}_{i(opt)} \hat{Y}_{i.} + (1 - \hat{w}_{i(opt)}) \hat{Y}^S_{i.} \quad (5.15)$$

소지역 i에 대한 복합추정량의 평균제곱오차 $MSE(\hat{Y}^C_{i.})$는 경제활동인구조사 자료로부터 추정되어야 하며 좀 더 안정적인 평균제곱오차에 대한 추정값들을 산출하기 위해 잭나이프 추정방법이 적용될 수 있다. 소지역 i에서 실업자 총계에 대한 복합추정값들의 잭나이프 평균제곱오차는 다음 추정식으로부터 산정할 수 있다.

$$mse_J(\hat{Y}_{i\cdot}^{C}) = \widehat{Var}(\hat{Y}_{i\cdot}^{C}) + [\widehat{Bias}_J(\hat{Y}_{i\cdot}^{C})]^2 \qquad (5.16)$$

여기에서

$$\widehat{Var}_J(\hat{Y}_{i\cdot}^{C}) = \frac{n_i - 1}{n_i} \sum_{h=1}^{n_i} [\hat{Y}_{i\cdot}^{C}(h) - \frac{1}{n_i} \sum_{l=1}^{n_i} \hat{Y}_{i\cdot}^{C}(l)]^2, \qquad (5.17)$$

$$\widehat{Bias}_J(\hat{Y}_{i\cdot}^{C}) = (n_i - 1)[\frac{1}{n_i} \sum_{h=1}^{n_i} \hat{Y}_{i\cdot}^{C}(h) - \hat{Y}_{i\cdot}^{C}], \qquad (5.18)$$

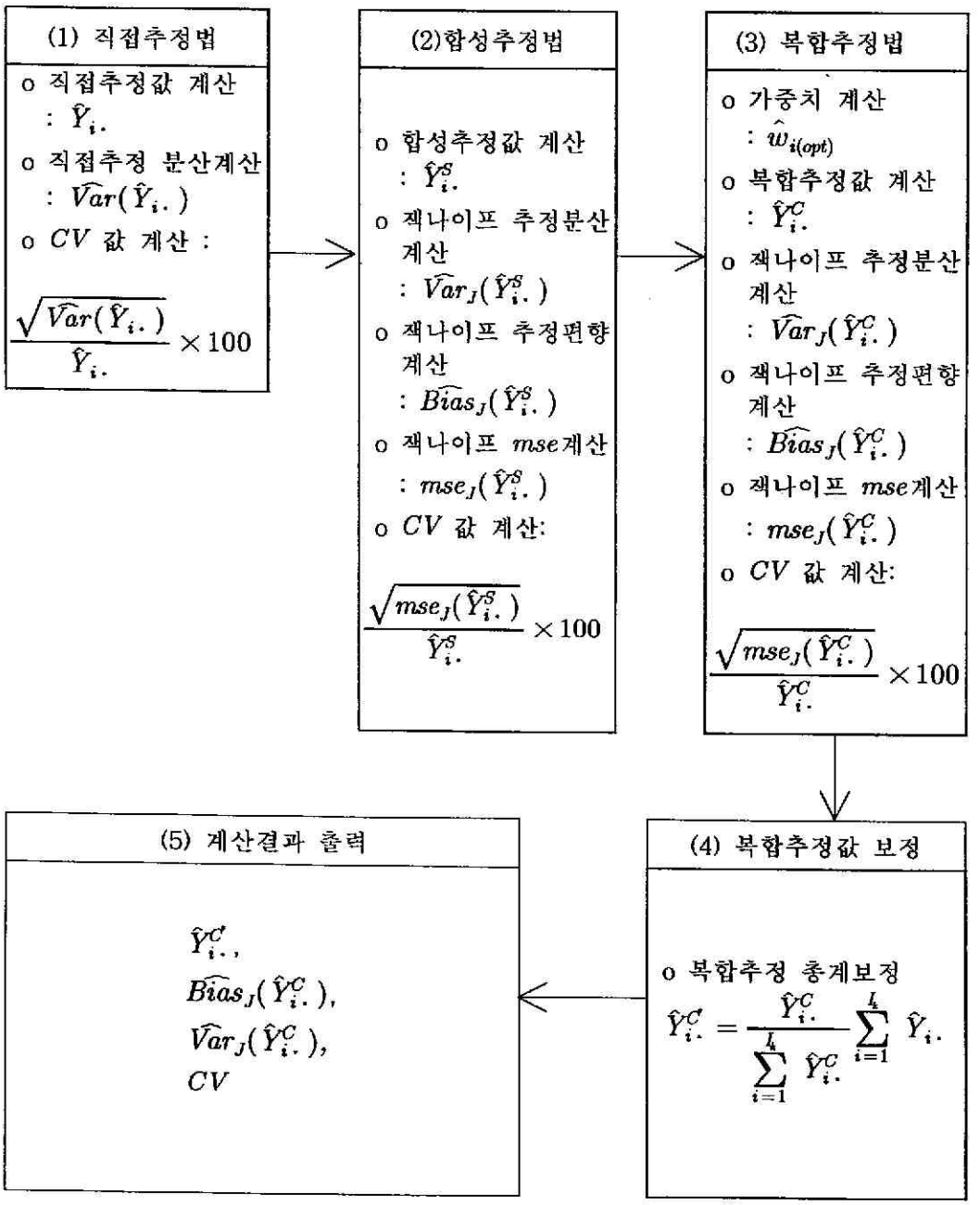
$$\hat{Y}_{i\cdot}^{C}(h) = \hat{w}_{i(opt)} \hat{Y}_{i\cdot} + (1 - \hat{w}_{i(opt)}) \hat{Y}_{i\cdot}^{S}(h) . \qquad (5.19)$$

로 주어진다.

5.6 SAS 프로그램 알고리즘

5.6.1 프로그램 순서도

SAS 프로그램을 순서도로 나타내면 다음과 같다.



5.6.2 세부 알고리즘

대영역인 충청북도는 3개의 시 단위 소지역(청주, 충주, 제천)들과, 8개의 군단위 소지역(보은, 옥천, 영동, 괴산, 진천, 음성, 청원, 단양)들로 구성되어있다. 충청북도의 경제활동인구조사 자료에 대한 시군구 단위 소지역 추정을 예로 설명한다.

시군 단위 소지역들의 지역코드는 "청주=1", "충주=2", "제천=3", "보은=4", "옥천=5", "영동=6", "괴산=7", "진천=8", "음성=9", "청원=10", "단양=11"로 나타내고, 각 시군 단위 소지역들에 대한 표본조사구의 수는 "청주=n_1", "충주=n_2", "제천=n_3", "보은=n_4", "옥천=n_5", "영동=n_6", "괴산=n_7", "진천=n_8", "음성=n_9", "청원=n_{10}", "단양=n_{11}"로 나타내기로 한다.

(1) 직접추정법

(Step1) 직접추정값 계산($\hat{Y}_{i.}$)

대영역 내의 I개의 소지역에 대한 실업자 총계 추정값은 경제활동인구조사 자료에 근거하여 다음 직접 추정 공식을 이용하여 계산한다. 경제활동인구조사자료는 조사구별로 자료가 정리되어 있고, 소지역 i의 자료는 동부와 읍면부 단위의 조사구들로 구성되어 있다.

충청북도 내의 시군 단위 소지역들을 예로 설명하면, 소지역 i가 청주시와 같이 시 단위 지역인 경우는 동단위의 조사구들로 구성되어 있고, 청원군과 같이 군 단위 지역인 경우 읍면 단위의 조사구들로 구성되어 있으나, 충주시의 경우에는 동단위와 읍면단위 조사구가 혼합되어 있다.

따라서 위의 3가지 경우를 고려하여 대영역인 하나의 집락 내에서 해당 소지역들에 대한 실업자 총계 추정 프로그램을 작성하여야 한다

$$\hat{Y}_{i.} = \sum_{s=1}^{2} {}_s\hat{Y}_{i.} \quad , \quad i=1,2,\cdots,I; s=1,2; h=1,2,\cdots,n_i$$

$$= \sum_{s=1}^{2}\sum_{h=1}^{n_i} {}_s\hat{Y}_{ih}$$

$$= \sum_{s=1}^{2} \sum_{h=1}^{n_i} {}_sM_i \; {}_sY_{ih} \;,$$

여기에서

s = 성별(남1, 여2)을 나타내는 첨자,

n_i = 경제활동인구조사에서 소지역 i의 표본조사구 수,

${}_sY_{ih}$ = 각 성별에 대해서 소지역 i의 표본조사구에서 조사한 실업자수,

${}_sM_i = {}_s\hat{X}_{i.} / {}_sX_{i.}$ (주어진 값:데이터 필드에 포함됨),

${}_s\hat{X}_{i.}$ = 소지역 i에 대한 15세 이상의 상주 추계인구,

${}_sX_{i.}$ = 경제활동인구조사에서 조사된 15세 이상의 조사인구.

※ 소지역 i의 표본조사구들이 동단위, 읍면단위, 또는 동과 읍면단위의 혼합으로 구성되어 있어도 일반적으로 적용가능

(Step2) 직접추정값의 분산 계산

소지역 i에 대한 직접 추정값의 분산은 다음 식을 적용하여 추정한다.

$$\widehat{Var}(\hat{Y}_{i.}) = \sum_{s=1}^{2} {}_sM_i^2 (\zeta_i \sum_{h=1}^{n_i} {}_sU_{ih}^2) + 2 \, {}_1M_i \, {}_2M_i (\zeta_i \sum_{h=1}^{n_i} {}_1U_{ih} \, {}_2U_{ih})$$

여기에서

$${}_sU_{ih} = d_{\,s}Y_{ih} - {}_s\rho_i \cdot d_{\,s}X_{ih} \;,$$

$$d_{\,s}Y_{ih} = {}_sY_{ih} - {}_s\,Y_{i,h+1} \;,$$

$$d_{\,s}X_{ih} = {}_s\,X_{ih} - {}_s\,X_{i,h+1} \;,$$

$${}_s\rho_i = {}_sY_{i.} / {}_sX_{i.} \;,$$

$$\zeta_i \approx n_i / [2(n_i - 1)] \;,$$

N_i = 소지역 i에 대한 표본추출틀의 조사구 수.

※ 소지역 i가 동단위 표본조사구들만으로 구성되어 있을 경우와, 소지역 i가 읍면단위 표본조사들만으로 구성되어 있을 경우는 위의 분산 추정공식을 이용하여 계산하고,

※ 소지역 i가 동단위와 읍면단위의 표본조사구들로 혼합되어 있을 경우에는(예를 들면 충북지역에서 충주시와 같은 경우임), 동단위 표본조사구들만으로 추정분산을 계산하고, 읍면단위 표본조사구들만으로 추정분산을 계산한 후, 이들을 합하여 소지역 i의 추정분산을 산출함

(Step3) 직접추정값들의 CV 값 계산

$$CV(\hat{Y}_{i.}) = \frac{\sqrt{\widehat{Var}(\hat{Y}_{i.})}}{\hat{Y}_{i.}} \times 100 \ , \quad i = 1, 2, \cdots, 11$$

(Step4) 이상의 직접추정법에 의한 계산 결과를 저장

$\hat{Y}_{1.}, \hat{Y}_{2.}, \cdots, \hat{Y}_{11.}$,

$\widehat{Var}(\hat{Y}_{1.}), \widehat{Var}(\hat{Y}_{2.}), \cdots, \widehat{Var}(\hat{Y}_{11.})$

$CV(\hat{Y}_{1.}), CV(\hat{Y}_{2.}), \cdots, CV(\hat{Y}_{11.})$

(2) 합성추정법

(Step1) 합성추정값을 계산하기 위한 사전작업으로써 대영역인 충북지역을 시지역(청주, 충주, 제천), 군지역(보은, 옥천, 괴산, 진천, 영동, 음성, 청원, 단양), 구지역(충북에는 구지역이 없음)으로 데이터 셀을 분할하고 표본조사구별로 경제활동인구조사 자료를 작성한다. 이 후 각각의 그룹에 대해서 다음을 계산한다.

① 범주별 상주추계인구 자료 작성(통계청 추계자료 이용)

o 시지역(청주, 충주, 제천)

청주$A1$ = 청주 1범주 상주추계인구, 청주$A2$ = 청주 2범주 상주추계인구,

청주$A3$ = 청주 3범주 상주추계인구, 청주$A4$ = 청주 4범주 상주추계인구,

충주$A1$ = 충주 1범주 상주추계인구, 충주$A2$ = 충주 2범주 상주추계인구,
충주$A3$ = 충주 3범주 상주추계인구, 충주$A4$ = 충주 4범주 상주추계인구,

제천$A1$ = 제천 1범주 상주추계인구, 제천$A2$ = 제천 2범주 상주추계인구,
제천$A3$ = 제천 3범주 상주추계인구, 제천$A4$ = 제천 4범주 상주추계인구,

o 군지역(보은, 옥천, 영동, 괴산, 진천, 음성, 청원, 단양)

보은$A1$ = 보은 1범주 상주추계인구, 보은$A2$ = 보은 2범주 상주추계인구,
보은$A3$ = 보은 3범주 상주추계인구, 보은$A4$ = 보은 4범주 상주추계인구,

옥천$A1$ = 옥천 1범주 상주추계인구, 옥천$A2$ = 옥천 2범주 상주추계인구,
옥천$A3$ = 옥천 3범주 상주추계인구, 옥천$A4$ = 옥천 4범주 상주추계인구,

.....

단양$A1$ = 단양 1범주 상주추계인구, 단양$A2$ = 단양 2범주 상주추계인구,
단양$A3$ = 단양 3범주 상주추계인구, 단양$A4$ = 단양 4범주 상주추계인구.

② 범주별 상주조사인구 카운트(경활조사에서 단순히 카운트된 수)

o 시지역(청주, 충주, 제천)

청주$B1$ = 청주 1범주 상주조사인구, 청주$B2$ = 청주 2범주 상주조사인구,
청주$B3$ = 청주 3범주 상주조사인구, 청주$B4$ = 청주 4범주 상주조사인구,

충주$B1$ = 충주 1범주 상주조사인구, 충주$B2$ = 충주 2범주 상주조사인구,
충주$B3$ = 충주 3범주 상주조사인구, 충주$B4$ = 충주 4범주 상주조사인구,

제천$B1$ = 제천 1범주 상주조사인구, 제천$B2$ = 제천 2범주 상주조사인구,
제천$B3$ = 제천 3범주 상주조사인구, 제천$B4$ = 제천 4범주 상주조사인구,

o 군지역(보은, 옥천, 영동, 괴산, 진천, 음성, 청원, 단양)
보은$B1$ = 보은 1범주 상주조사인구, 보은$B2$ = 보은 2범주 상주조사인구,
보은$B3$ = 보은 3범주 상주조사인구, 보은$B4$ = 보은 4범주 상주조사인구,

.....

단양$B1$ = 단양 1범주 상주조사인구, 단양$B2$ = 단양 2범주 상주조사인구,
단양$B3$ = 단양 3범주 상주조사인구, 단양$B4$ = 단양 4범주 상주조사인구.

③ 범주별 실업자 총계 카운트(경활조사에서 단순히 카운트된 수)
o 시지역(청주, 충주, 제천)
청주$C1$ = 청주 1범주 실업자 총계, 청주$C2$ = 청주 2범주 실업자 총계,
청주$C3$ = 청주 3범주 실업자 총계, 청주$C4$ = 청주 4범주 실업자 총계,

충주$C1$ = 충주 1범주 실업자 총계, 충주$C2$ = 충주 2범주 실업자 총계,
충주$C3$ = 충주 3범주 실업자 총계, 충주$C4$ = 충주 4범주 실업자 총계,

제천$C1$ = 제천 1범주 실업자 총계, 제천$C2$ = 제천 2범주 실업자 총계,
제천$C3$ = 제천 3범주 실업자 총계, 제천$C4$ = 제천 4범주 실업자 총계,

o 군지역(보은, 옥천, 영동, 괴산, 진천, 음성, 청원, 단양)
보은$C1$ = 보은 1범주 실업자 총계, 보은$C2$ = 보은 2범주 실업자 총계,
보은$C3$ = 보은 3범주 실업자 총계, 보은$C4$ = 보은 4범주 실업자 총계,

.....

단양$C1$ = 단양 1범주 실업자 총계, 단양$C2$ = 단양 2범주 실업자 총계,
단양$C3$ = 단양 3범주 실업자 총계, 단양$C4$ = 단양 4범주 실업자 총계,

④ 범주별 경제활동인구 총계 카운트(경활조사에서 단순히 카운된 수)

o 시지역(청주, 충주, 제천)

청주$D1$ = 청주 1범주 경활인구총계, 청주$D2$ = 청주 2범주 경활인구총계,
청주$D3$ = 청주 3범주 경활인구총계, 청주$D4$ = 청주 4범주 경활인구총계,

충주$D1$ = 충주 1범주 경활인구총계, 충주$D2$ = 충주 2범주 경활인구총계,
충주$D3$ = 충주 3범주 경활인구총계, 충주$D4$ = 충주 4범주 경활인구총계,

제천$D1$ = 제천 1범주 경활인구총계, 제천$D2$ = 제천 2범주 경활인구총계,
제천$D3$ = 제천 3범주 경활인구총계, 제천$D4$ = 제천 4범주 경활인구총계,

o 군지역(보은, 옥천, 영동, 괴산, 진천, 음성, 청원, 단양)

보은$D1$ = 보은 1범주 경활인구총계, 보은$D2$ = 보은 2범주 경활인구총계,
보은$D3$ = 보은 3범주 경활인구총계, 보은$D4$ = 보은 4범주 경활인구총계,

.....

단양$D1$ = 단양 1범주 경활인구총계, 단양$D2$ = 단양 2범주 경활인구총계,
단양$D3$ = 단양 3범주 경활인구총계, 단양$D4$ = 단양 4범주 경활인구총계,

(Step2) 합성추정값 계산

① 시지역 합성추정값 계산

o 범주별 실업자 총계 계산

$$ {}_1\hat{Y}_{dir} = (\text{청주}A1/\text{청주}B1)*\text{청주}C1 $$
$$ +(\text{충주}A1/\text{충주}B1)*\text{충주}C1 $$
$$ +(\text{제천}A1/\text{제천}B1)*\text{제천}C1 $$

$${}_2\hat{Y}_{dir} = (\text{청주}A2/\text{청주}B2)*\text{청주}C2$$
$$+(\text{충주}A2/\text{충주}B2)*\text{충주}C2$$
$$+(\text{제천}A2/\text{제천}B2)*\text{제천}C2$$

$${}_3\hat{Y}_{dir} = (\text{청주}A3/\text{청주}B3)*\text{청주}C3$$
$$+(\text{충주}A3/\text{충주}B3)*\text{충주}C3$$
$$+(\text{제천}A3/\text{제천}B3)*\text{제천}C3$$

$${}_4\hat{Y}_{dir} = (\text{청주}A4/\text{청주}B4)*\text{청주}C4$$
$$+(\text{충주}A4/\text{충주}B4)*\text{충주}C4$$
$$+(\text{제천}A4/\text{제천}B4)*\text{제천}C4$$

o 실업자 총계에 대한 합성추정값 계산

-청주시

$$\hat{Y}^S_{1.} = (\text{청주}A1*_1 \hat{Y}_{dir})/(\text{청주}A1+\text{충주}A1+\text{제천}A1)$$
$$+(\text{청주}A2*_2 \hat{Y}_{dir})/(\text{청주}A2+\text{충주}A2+\text{제천}A2)$$
$$+(\text{청주}A3*_3 \hat{Y}_{dir})/(\text{청주}A3+\text{충주}A3+\text{제천}A3)$$
$$+(\text{청주}A4*_4 \hat{Y}_{dir})/(\text{청주}A4+\text{충주}A4+\text{제천}A4)$$

-충주시

$$\hat{Y}^S_{2.} = (\text{충주}A1*_1 \hat{Y}_{dir})/(\text{청주}A1+\text{충주}A1+\text{제천}A1)$$
$$+(\text{충주}A2*_2 \hat{Y}_{dir})/(\text{청주}A2+\text{충주}A2+\text{제천}A2)$$
$$+(\text{충주}A3*_3 \hat{Y}_{dir})/(\text{청주}A3+\text{충주}A3+\text{제천}A3)$$
$$+(\text{충주}A4*_4 \hat{Y}_{dir})/(\text{청주}A4+\text{충주}A4+\text{제천}A4)$$

-제천시

$$\hat{Y}^S_{3.} = (\text{제천}A1*_1 \hat{Y}_{dir})/(\text{청주}A1+\text{충주}A1+\text{제천}A1)$$
$$+(\text{제천}A2*_2 \hat{Y}_{dir})/(\text{청주}A2+\text{충주}A2+\text{제천}A2)$$

$+$(제천$A3$*$_3$ $\hat{Y}_{dir}$)/(청주$A3$+충주$A3$+제천$A3$)

$+$(제천$A4$*$_4$ $\hat{Y}_{dir}$)/(청주$A4$+충주$A4$+제천$A4$)

(output) 합성추정값 계산결과 저장

$\hat{Y}^S_{1.}$, $\hat{Y}^S_{2.}$, $\hat{Y}^S_{3.}$

② 군지역 합성추정값 계산

위의 ①과 같은 방법으로 계산함

(Step3) 시지역(청주, 충주, 제천)에서 반복 합성추정량 생성

① 청주시에서 반복 합성추정량들을 생성

다음의 과정을 반복하면 청주시의 표본조사구 수($=n_1$) 만큼 반복 합성추정량들이 생성됨. 청주시에서 l번째 표본조사구를 제외하고 나머지 표본조사구들을 이용하여 새로운 합성추정값을 다음 절차에 의해 반복적으로 생성함($=\hat{Y}^S_{1.}(l)$, $l=1,2,\cdots,n_1$)

(절차1) 청주시 범주별 상주조사인구 다시 카운트(l번째 조사구를 제외하고 카운트 함)

청주$B1(l)$ $=$청주시 1범주 상주조사인구,

청주$B2(l)$ $=$청주시 2범주 상주조사인구,

청주$B3(l)$ $=$청주시 3범주 상주조사인구,

청주$B4(l)$ $=$청주시 4범주 상주조사인구.

(절차2) 청주시 범주별 실업자 총계 다시 카운트(l번째 조사구를 제외하고 카운트 함)

청주$C1(l)$ $=$청주시 1범주 실업자 총계,

청주$C2(l)$ $=$청주시 2범주 실업자 총계,

청주$C3(l)$ $=$청주시 3범주 실업자 총계,

청주$C4(l)$ $=$청주시 4범주 실업자 총계.

(절차3) 청주시 범주별 경제활동인구 총계 다시 카운트(l번째 조사

구를 제외하고 카운트 함)

청주$D1(l)$=청주시 1범주 경제활동인구 총계,

청주$D2(l)$=청주시 2범주 경제활동인구 총계,

청주$D3(l)$=청주시 3범주 경제활동인구 총계,

청주$D4(l)$=청주시 4범주 경제활동인구 총계.

(절차4) 범주별 실업자 총계에 대한 직접추정값을 계산

o 1범주 실업자 총계에 대한 직접추정값

$$_1\hat{Y}_{dir}(l) = (\text{청주}A1/\text{청주}B1(l))*\text{청주}C1(l)$$
$$+(\text{충주}A1/\text{충주}B1)*\text{충주}C1$$
$$+(\text{제천}A1/\text{제천}B1)*\text{제천}C1$$

o 2범주 직접추정값

$$_2\hat{Y}_{dir}(l) = (\text{청주}A2/\text{청주}B2(l))*\text{청주}C2(l)$$
$$+(\text{충주}A2/\text{충주}B2)*\text{충주}C2$$
$$+(\text{제천}A2/\text{제천}B2)*\text{제천}C2$$

o 3범주 직접추정값

$$_3\hat{Y}_{dir}(l) = (\text{청주}A3/\text{청주}B3(l))*\text{청주}C3(l)$$
$$+(\text{충주}A3/\text{충주}B3)*\text{충주}C3$$
$$+(\text{제천}A3/\text{제천}B3)*\text{제천}C3$$

o 4범주 직접추정값

$$_4\hat{Y}_{dir}(l) = (\text{청주}A4/\text{청주}B4(l))*\text{청주}C4(l)$$
$$+(\text{충주}A4/\text{충주}B4)*\text{충주}C4$$
$$+(\text{제천}A4/\text{제천}B4)*\text{제천}C4$$

(절차5) l 번째 합성추정값 계산(= $\hat{Y}_1^S.(l)$)

$$\hat{Y}_1^S.(l) = (\text{청주}A1 \times_1 \hat{Y}_{dir}(l))/(\text{청주}A1+\text{충주}A1+\text{제천}A1)$$
$$+(\text{청주}A2 \times_2 \hat{Y}_{dir}(l))/(\text{청주}A2+\text{충주}A2+\text{제천}A2)$$

$$+(\text{청주}A3\times_3 \hat{Y}_{dir}(l))/(\text{청주}A3+\text{충주}A3+\text{제천}A3)$$

$$+(\text{청주}A4\times_4 \hat{Y}_{dir}(l))/(\text{청주}A4+\text{충주}A4+\text{제천}A4),$$

여기에서 $l=1,2,\cdots,n_1$

(절차6) 청주시에서 생성된 n_1개의 반복합성추정값들과 이들의 평균값을 저장

$$\hat{Y}_1^S.(1),\ \hat{Y}_1^S.(2),\ \cdots,\ \hat{Y}_1^S.(n_1),$$

$$\hat{Y}_1^S.(\cdot)=(\hat{Y}_1^S.(1)+\hat{Y}_1^S.(2)+\cdots+\hat{Y}_1^S.(n_1))/n_1$$

② 충주시에서 반복 합성추정량들을 생성

충주시에서 l번째 표본조사구를 제외하고 나머지 표본조사구들을 이용하여 새로운 합성추정값을 다음 절차에 의해 반복적으로 생성함 ($=\hat{Y}_2^S.(l)$, $l=1,2,\cdots,n_2$)

(절차1) 충주시 범주별 상주조사인구 다시 카운트(l번째 조사구를 제외하고 카운트 함)

$$\text{충주}B1(l)=\text{충주시 1범주 상주조사인구},$$

$$\text{충주}B2(l)=\text{충주시 2범주 상주조사인구},$$

$$\text{충주}B3(l)=\text{충주시 3범주 상주조사인구},$$

$$\text{충주}B4(l)=\text{충주시 4범주 상주조사인구}.$$

(절차2) 충주시 범주별 실업자 총계 다시 카운트(l번째 조사구를 제외하고 카운트 함)

$$\text{충주}C1(l)=\text{충주시 1범주 실업자 총계},$$

$$\text{충주}C2(l)=\text{충주시 2범주 실업자 총계},$$

$$\text{충주}C3(l)=\text{충주시 3범주 실업자 총계},$$

$$\text{충주}C4(l)=\text{충주시 4범주 실업자 총계}.$$

(절차3) 충주시 범주별 경제활동인구 총계 다시 카운트(l번째 조사구를 제외하고 카운트 함)

충주$D1(l)$=충주시 1범주 경제활동인구 총계,

충주$D2(l)$=충주시 2범주 경제활동인구 총계,

충주$D3(l)$=충주시 3범주 경제활동인구 총계,

충주$D4(l)$=충주시 4범주 경제활동인구 총계.

(절차4) 범주별 실업자 총계에 대한 직접추정값을 계산

o 1범주 실업자 총계에 대한 직접추정값

$$_1\hat{Y}_{dir}(l) = (\text{충주}A1/\text{충주}B1(l))*\text{충주}C1(l)$$
$$+(\text{청주}A1/\text{청주}B1)*\text{청주}C1$$
$$+(\text{제천}A1/\text{제천}B1)*\text{제천}C1$$

o 2범주 직접추정값

$$_2\hat{Y}_{dir}(l) = (\text{충주}A2/\text{충주}B2(l))*\text{충주}C2(l)$$
$$+(\text{청주}A2/\text{청주}B2)*\text{청주}C2$$
$$+(\text{제천}A2/\text{제천}B2)*\text{제천}C2$$

o 3범주 직접추정값

$$_3\hat{Y}_{dir}(l) = (\text{충주}A3/\text{충주}B3(l))*\text{충주}C3(l)$$
$$+(\text{청주}A3/\text{청주}B3)*\text{청주}C3$$
$$+(\text{제천}A3/\text{제천}B3)*\text{제천}C3$$

o 4범주 직접추정값

$$_4\hat{Y}_{dir}(l) = (\text{충주}A4/\text{충주}B4(l))*\text{충주}C4(l)$$
$$+(\text{청주}A4/\text{청주}B4)*\text{청주}C4$$
$$+(\text{제천}A4/\text{제천}B4)*\text{제천}C4$$

(절차5) l 번째 합성추정값 계산(= $\hat{Y}_2^S.(l)$)

$$\hat{Y}_2^S.(l) = (\text{충주}A1 \times_1 \hat{Y}_{dir}(l))/(\text{청주}A1+\text{충주}A1+\text{제천}A1)$$
$$+(\text{충주}A2 \times_2 \hat{Y}_{dir}(l))/(\text{청주}A2+\text{충주}A2+\text{제천}A2)$$

$+$(충주$A3 \times_3 \hat{Y}_{dr}(l)$)/(청주$A3$+충주$A3$+제천$A3$)

$+$(충주$A4 \times_4 \hat{Y}_{dr}(l)$)/(청주$A4$+충주$A4$+제천$A4$),

여기에서 $l=1,2,\cdots,n_2$

(절차6) 충주시에서 생성된 n_2개의 반복합성추정값들과 이들의 평균값을 저장

$$\hat{Y}_2^S.(1),\ \hat{Y}_2^S.(2),\ \cdots,\ \hat{Y}_2^S.(n_2),$$

$$\hat{Y}_2^S.(\cdot)=(\hat{Y}_2^S.(1)+\hat{Y}_2^S.(2)+\cdots+\hat{Y}_2^S.(n_2))/n_2$$

③ 제천시에서 반복 합성추정량들을 생성

제천시에서 l번째 표본조사구를 제외하고 나머지 표본조사구들을 이용하여 새로운 합성추정값을 다음 절차에 의해 반복적으로 생성함 $(=\hat{Y}_3^S.(l)\ ,\ l=1,2,\cdots,n_3)$

(절차1) 제천시 범주별 상주조사인구 다시 카운트(l번째 조사구를 제외하고 카운트 함)

제천$B1(l)$=제천시 1범주 상주조사인구,

제천$B2(l)$=제천시 2범주 상주조사인구,

제천$B3(l)$=제천시 3범주 상주조사인구,

제천$B4(l)$=제천시 4범주 상주조사인구.

(절차2) 제주시 범주별 실업자 총계 다시 카운트(l번째 조사구를 제외하고 카운트 함)

제천$C1(l)$=제천시 1범주 실업자 총계,

제천$C2(l)$=제천시 2범주 실업자 총계,

제천$C3(l)$=제천시 3범주 실업자 총계,

제천$C4(l)$=제천시 4범주 실업자 총계.

(절차3) 제천시 범주별 경제활동인구 총계 다시 카운트(l번째 조사구를 제외하고 카운트 함)

제천$D1(l)$=제천시 1범주 경제활동인구 총계,

제천$D2(l)$=제천시 2범주 경제활동인구 총계,

제천$D3(l)$=제천시 3범주 경제활동인구 총계,

제천$D4(l)$=제천시 4범주 경제활동인구 총계.

(절차4) 범주별 실업자 총계에 대한 직접추정값을 계산

o 1범주 실업자 총계에 대한 직접추정값

$$_1\hat{Y}_{dir}(l) = (\text{제천}A1/\text{제천}B1(l))*\text{제천}C1(l)$$
$$+(\text{청주}A1/\text{청주}B1)*\text{청주}C1$$
$$+(\text{충주}A1/\text{충주}B1)*\text{충주}C1$$

o 2범주 직접추정값

$$_2\hat{Y}_{dir}(l) = (\text{제천}A2/\text{제천}B2(l))*\text{제천}C2(l)$$
$$+(\text{청주}A2/\text{청주}B2)*\text{청주}C2$$
$$+(\text{충주}A2/\text{충주}B2)*\text{충주}C2$$

o 3범주 직접추정값

$$_3\hat{Y}_{dir}(l) = (\text{제천}A3/\text{제천}B3(l))*\text{제천}C3(l)$$
$$+(\text{청주}A3/\text{청주}B3)*\text{청주}C3$$
$$+(\text{충주}A3/\text{충주}B3)*\text{충주}C3$$

o 4범주 직접추정값

$$_4\hat{Y}_{dir}(l) = (\text{제천}A4/\text{제천}B4(l))*\text{제천}C4(l)$$
$$+(\text{청주}A4/\text{청주}B4)*\text{청주}C4$$
$$+(\text{충주}A4/\text{충주}B4)*\text{제천}C4$$

(절차5) l 번째 합성추정값 계산(= $\hat{Y}^{S}_{3.}(l)$)

$$\hat{Y}^{S}_{3.}(l) = (\text{제천}A1 \times_1 \hat{Y}_{dir}(l))/(\text{청주}A1+\text{충주}A1+\text{제천}A1)$$
$$+(\text{제천}A2 \times_2 \hat{Y}_{dir}(l))/(\text{청주}A2+\text{충주}A2+\text{제천}A2)$$

$+$(제천$A3\times_3 \hat{Y}_{dir}(l)$)/(청주$A3$+충주$A3$+제천$A3$)

$+$(제천$A4\times_4 \hat{Y}_{dir}(l)$)/(청주$A4$+충주$A4$+제천$A4$),

여기에서 $l=1,2,\cdots,n_3$

(절차6) 제천시에서 생성된 n_3개의 반복합성추정값들과 이들의 평균값을 저장

$$\hat{Y}^S_3.(1),\ \hat{Y}^S_3.(2),\ \cdots,\ \hat{Y}^S_3.(n_3),$$

$$\hat{Y}^S_3.(\cdot)=(\hat{Y}^S_3.(1)+\hat{Y}^S_3.(2)+\cdots+\hat{Y}^S_3.(n_3))/n_3$$

(Step4) 군지역(보은, 옥천, 영동, 괴산, 진천, 음성, 청원, 단양)에서 반복 합성추정값들을 생성(Step3의 과정과 같음)

① 보은군에서 반복 합성추정값들을 생성

.....

⑧ 단양군에서 반복 합성추정값들을 생성

(Step5) 이상에서 반복 계산된 합성추정값들을 이용하여 충북내의 각 시군단위 소지역들에 대한 합성추정값들의 잭나이프 *Bias*, 잭나이프 분산, 잭나이프 *mse*, *CV* 값들을 계산하고 이 결과를 저장함.

충북 내의 시군 단위 소지역 i에 대해서 다음을 계산하여 저장함.

$$\widehat{Bias}_J(\hat{Y}^S_{i\cdot})=(n_i-1)[\hat{Y}^S_{i\cdot}(\cdot)-\hat{Y}^S_{i\cdot}],\qquad i=1,2,\cdots,11$$

$$\widehat{Var}_J(\hat{Y}^S_{i\cdot})=\frac{n_i-1}{n_i}\sum_{h=1}^{n_i}[\hat{Y}^S_{i\cdot}(h)-\hat{Y}^S_{i\cdot}(\cdot)]^2,$$

$$mse_J(\hat{Y}^S_{i\cdot})=\widehat{Var}_J(\hat{Y}^S_{i\cdot})+[\widehat{Bias}_J(\hat{Y}^S_{i\cdot})]^2,$$

$$CV(\hat{Y}^S_{i\cdot})=\frac{\sqrt{mse_J(\hat{Y}^S_{i\cdot})}}{\hat{Y}^S_{i\cdot}}\times 100,$$

여기에서 $\hat{Y}_{i.}^{S}$ 는 (Step2)에 저장되어 있는 값들임.

(3) 복합추정법

(Step1) 각 시군 단위 소지역들에 대해서 최적 가중치 $\hat{w}_{i(opt)}$ 을 계산

$$\hat{w}_{i(opt)} = \frac{mse_J(\hat{Y}_{i.}^{S})}{mse_J(\hat{Y}_{i.}^{S}) + \widehat{Var}(\hat{Y}_{i.})} \quad , \quad i=1,2,\cdots,11$$

여기에서 $mse_J(\hat{Y}_{i.}^{S})$는 (2)의 (Step5)에 저장되어 있고, $\widehat{Var}(\hat{Y}_{i.})$은 (1)의 (Step4)에 저장되어 있는 값들임.

(Step2) 각 시군 단위 소지역들에 대한 복합추정값 $\hat{Y}_{i.}^{C}$ 을 계산

$$\hat{Y}_{i.}^{C} = \hat{w}_{i(opt)}\hat{Y}_{i.} + (1-\hat{w}_{i(opt)})\hat{Y}_{i.}^{S} \quad , \quad i=1,2,\cdots,11$$

여기에서 $\hat{Y}_{i.}$ 은 (1)의 (Step4)에 저장되어 있는 값들이고, $\hat{Y}_{i.}^{S}$ 는 (2)의 (Step2)에 저장되어 있는 값들임.

(Step3) 각 시군 단위 소지역 i에 대해서 소지역 i의 표본조사구 개수 $(=n_i)$ 만큼 새로운 복합추정값들과 이들의 평균을 계산

소지역 i에 대해서

$$\hat{Y}_{i.}^{C}(1) = \hat{w}_{i(opt)}\hat{Y}_{i.} + (1-\hat{w}_{i(opt)})\hat{Y}_{i.}^{S}(1), \quad i=1,2,\cdots,11$$

$$\hat{Y}_{i.}^{C}(2) = \hat{w}_{i(opt)}\hat{Y}_{i.} + (1-\hat{w}_{i(opt)})\hat{Y}_{i.}^{S}(2),$$

$$\cdots$$

$$\hat{Y}_{i.}^{C}(n_i) = \hat{w}_{i(opt)}\hat{Y}_{i.} + (1-\hat{w}_{i(opt)})\hat{Y}_{i.}^{S}(n_i),$$

$$\hat{Y}_{i.}^{C}(\cdot) = (\hat{Y}_{i.}^{C}(1) + \hat{Y}_{i.}^{C}(2) + \cdots + \hat{Y}_{i.}^{C}(n_i))/n_i$$

(Step4) 복합추정값들의 잭나이프 *Bias*, 잭나이프 분산, 잭나이프 *mse*, *CV* 값들을 계산하여 저장

소지역 i에 대하여

$$\widehat{Bias}_J(\hat{Y}_{i\cdot}^C) = (n_i - 1)[\hat{Y}_{i\cdot}^C(\cdot) - \hat{Y}_{i\cdot}^C], \quad i = 1,2,\cdots,11$$

$$\widehat{Var}_J(\hat{Y}_{i\cdot}^C) = \frac{n_i - 1}{n_i}\sum_{h=1}^{n_i}[\hat{Y}_{i\cdot}^C(h) - \hat{Y}_{i\cdot}^C]^2,$$

$$mse_J(\hat{Y}_{i\cdot}^C) = \widehat{Var}_J(\hat{Y}_{i\cdot}^C) + [\widehat{Bias}_J(\hat{Y}_{i\cdot}^C)]^2,$$

$$CV(\hat{Y}_{i\cdot}^C) = \frac{\sqrt{mse_J(\hat{Y}_{i\cdot}^C)}}{\hat{Y}_{i\cdot}^C} \times 100$$

여기에서 $\hat{Y}_{i\cdot}^C(h)$와 $\hat{Y}_{i\cdot}^C(\cdot)$은 위의 (Step3)에서 계산된 값들이고, $\hat{Y}_{i\cdot}^C$은 위의 (Step2)에서 계산된 값들임.

(4) 복합추정값들의 총계 보정 및 추정결과 출력

(Step1) 시군구 단위 소지역 i의 복합추정값 보정

대영역 내에서 분할된 3개의 시군구 그룹들 각각에 대해서 복합추정값들에 대한 총계 보정을 다음과 같이 실시한다.

① 시 그룹의 소지역들(청주, 충주, 제천)에 대한 총계 보정

$$\hat{Y}_{i\cdot(adj)}^C = \frac{\hat{Y}_{i\cdot}^C}{\sum_{i=1}^{3}\hat{Y}_{i\cdot}^C} \times \sum_{i=1}^{3}\hat{Y}_{i\cdot}, \quad i = 1,2,3$$

여기에서 $\sum_{i=1}^{3}\hat{Y}_{i\cdot}$은 시지역(청주, 충주, 제천시)의 직접추정값들에 대한 합을 나타낸다.

② 군 그룹의 소지역들(보은, 옥천, 영동, 괴산, 진천, 음성, 청원, 단양)에 대한 총계 보정

$$\hat{Y}^{C}_{i\cdot\,(adj)} = \frac{\hat{Y}^{C}_{i\cdot}}{\sum_{i=1}^{8}\hat{Y}^{C}_{i\cdot}} \times \sum_{i=1}^{8}\hat{Y}_{i\cdot} \quad , \quad i=1,2,\cdots,8$$

여기에서 $\sum_{i=1}^{8}\hat{Y}_{i\cdot}$ 은 군지역(보은, 옥천, 영동, 괴산, 진천, 음성, 청원, 단양군)의 직접추정값들에 대한 합을 나타낸다.

(Step2) 시군구 단위 추정결과 출력

구 분		직접추정법			합성추정법				복합추정법			
		$\hat{Y}_{i\cdot}$	$\sqrt{\widehat{Var}}$	CV	$\hat{Y}^{S}_{i\cdot}$	$Bias$	$\sqrt{mse}$	CV	$\hat{Y}^{C}_{i\cdot}$	$Bias$	$\sqrt{mse}$	CV
시지역	청주시(1) 충주시(2) 제천시(3)											
군지역	보은군(4) 옥천군(5) 영동군(6) 괴산군(7) 음성군(8) 청원군(9) 진천군(10) 단양군(11)											
합												

5.7 추정결과

5.7.1 시군구 추정결과

2001년 5월의 경제활동인구조사 자료로부터 대영역 내의 시군군단위 소지역들의 실업자 총계 추정결과를 요약하면 다음과 같다.

(1) 서울특별시

<표5.22> 서울특별시의 구단위 실업자 총계 추정결과

시군구	직접추정법			합성추정법			복합추정법		
	추정값	추정 오차	*CV* (%)	추정값	추정 오차	*CV* (%)	추정값	추정 오차	*CV* (%)
종로구	5,037	2,987	59.3	3,586	33	0.9	3,589	33	0.9
중구/용산구	8,569	1,229	14.3	7,602	65	0.8	7,609	65	0.8
중구							*2,788*		
용산구							*4,821*		
성동구	2,579	1,611	62.5	6,809	109	1.5	6,795	108	1.5
광진구	5,077	2,123	41.8	8,054	135	1.6	8,048	135	1.6
동대문구	13,686	4,587	33.5	7,940	116	1.4	7,950	116	1.4
중랑구	12,974	3,389	26.1	8,908	132	1.4	8,921	132	1.4
성북구	12,061	2,642	21.9	9,674	141	1.4	9,687	141	1.4
강북구	7,656	4,564	59.6	7,214	249	3.2	7,221	248	3.2
도봉구	8,690	3,428	39.4	7,024	90	1.2	7,031	90	1.2
노원구	11,228	2,671	23.8	12,071	135	1.0	12,078	135	1.0
은평구	5,117	1,720	33.6	9,115	113	1.2	9,104	112	1.2
서대문구	5,198	1,194	23.0	7,369	97	1.2	7,358	97	1.2
마포구	15,352	4,725	30.8	7,749	273	3.3	7,777	272	3.3
양천구	7,817	3,082	39.4	9,285	206	2.1	9,284	205	2.1
강서구	3,412	1,446	42.4	10,343	77	0.7	10,332	77	0.7
구로구	18,844	3,650	19.4	8,324	166	1.9	8,350	166	1.8
금천구	9,402	2,869	30.5	5,471	86	1.5	5,479	86	1.5
영등포구	2,498	900	36.0	8,176	50	0.6	8,165	50	0.6
동작구	6,944	3,039	43.8	8,571	159	1.7	8,573	159	1.7
관악구	13,606	4,735	34.8	11,303	266	2.2	11,317	265	2.2
서초구	6,984	1,097	15.7	7,993	64	0.7	7,996	64	0.7
강남구	8,529	2,328	27.3	11,338	191	1.6	11,326	190	1.6
송파구	2,579	1,583	61.4	13,074	149	1.1	12,992	148	1.1
강동구	12,854	2,101	16.3	9,700	111	1.1	9,715	111	1.1
계	206,693			206,693			206,693		

(2) 부산광역시

<표5.23> 부산광역시의 구/군단위 실업자 총계 추정결과

시군구	직접추정법			합성추정법			복합추정법		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
중구/동래구 /기장군	5,751	2,465	42.9	9,639	311	3.0	9,579	306	3.0
중구							2,226		
동래구							4,680		
기장군							2,673		
서구	2,305	885	38.4	3,507	89	2.4	3,495	89	2.4
동구	2,910	1,485	51.0	2,912	46	1.5	2,914	46	1.5
영도구	6,356	2,417	38.0	4,228	111	2.4	4,235	110	2.4
부산진구	14,297	3,393	23.7	9,764	388	3.7	9,817	383	3.6
남구	8,569	3,236	37.8	6,956	226	3.0	6,966	225	3.0
북구	6,333	3,275	51.7	6,566	220	3.1	6,568	219	3.1
해운대구	8,056	3,681	45.7	8,656	482	5.2	8,643	474	5.1
사하구	9,779	1,542	15.8	8,514	204	2.2	8,530	200	2.2
금정구	6,264	2,510	40.1	6,625	248	3.5	6,623	245	3.4
강서구	582	473	81.3	1,249	6	0.4	1,250	6	0.4
연제구	6,356	1,693	26.6	5,110	83	1.5	5,116	82	1.5
수영구	2,305	892	38.7	4,108	51	1.2	4,104	51	1.2
사상구	4,633	1,385	29.9	6,662	94	1.3	6,656	94	1.3
계	84,496			84,496			84,496		

(3) 인천광역시

<표5.24> 인천광역시의 구/군단위 실업자 총계 추정결과

시군구	직접추정법			합성추정법			복합추정법		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
중구	1,450	912	62.9	1,328	38	2.7	1,328	38	2.7
동구	2,540	1,283	50.5	1,478	23	1.5	1,478	23	1.5
남구	10,534	1,976	18.8	8,166	401	4.6	8,235	385	4.4
연수구	2,175	1,181	54.3	4,580	154	3.2	4,537	151	3.1
남동구	6,890	1,394	20.2	7,321	199	2.6	7,304	195	2.5
부평구	10,547	2,060	19.5	9,800	485	4.7	9,807	460	4.4
계양구	5,460	2,164	39.6	5,608	332	5.6	5,597	325	5.5
서구	5,832	1,478	25.3	5,993	190	3.0	5,985	187	2.9
강화군	0	0	0.0	1,155	120	9.8	1,156	120	9.8
계	45,428			45,428			45,428		

(4) 대구광역시

<표5.25> 대구광역시의 구/군단위 실업자 총계 추정결과

시군구	직접추정법			합성추정법			복합추정법		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
중구	1,860	524	28.2	1,806	20	1.1	1,808	20	1.0
동구	5,577	1,159	20.8	6,735	189	2.7	6,703	185	2.6
서구	11,617	1,889	16.3	5,742	160	2.7	5,785	159	2.6
남구	1,860	1,528	82.2	3,966	127	3.1	3,953	126	3.0
북구	6,042	1,823	30.2	7,710	326	4.0	7,656	316	3.9
수성구	5,575	2,119	38.0	8,848	370	4.0	8,750	359	3.9
달서구	13,477	2,204	16.4	11,243	684	5.8	11,393	624	5.2
달성군	2,787	1,738	62.4	2,744	154	5.3	2,746	153	5.3
계	48,795			48,795			48,795		

(5) 광주광역시

<표5.26> 광주광역시의 구단위 실업자 총계 추정결과

시군구	직접추정법			합성추정법			복합추정법		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
동구	2,807	976	34.8	2,237	85	3.5	2,262	84	3.4
서구	4,837	1,164	24.1	5,050	361	6.6	5,047	330	6.1
남구	5,289	1,137	21.5	4,265	212	4.6	4,328	205	4.4
북구	8,629	1,256	14.6	8,634	444	4.7	8,640	395	4.2
광산구	2,761	830	30.1	4,137	248	5.5	4,046	228	5.2
계	24,323			24,323			24,323		

(6) 대전광역시

<표5.27> 대전광역시의 구단위 실업자 총계 추정결과

시군구	직접추정법			합성추정법			복합추정법		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
동구	6,233	1,283	20.6	4,167	261	5.9	4,294	251	5.6
중구	4,774	806	16.9	4,169	189	4.3	4,244	179	4.0
서구	5,002	1,255	25.1	7,391	494	6.3	7,131	427	5.7
유성구	1,212	285	23.5	2,648	51	1.8	2,637	49	1.8
대덕구	4,774	1,405	29.4	3,620	200	5.2	3,689	196	5.1
계	21,995			21,995			21,995		

(7) 울산광역시

<표5.28> 울산광역시의 구/군단위 실업자 총계 추정결과

시군구	직접추정법			합성추정법			복합추정법		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
중구	5,628	1,394	24.8	3,105	467	14.6	3,270	420	12.2
남구	5,264	2,093	39.8	4,347	727	16.2	4,332	649	14.2
동구	1,403	1,562	111.3	2,466	323	12.7	2,368	310	12.4
울주군	0	0	0	2,377	620	25.3	2,325	620	25.3
계	12,295			12,295			12,295		

(8) 경기도

<표5.29> 경기도의 시/군단위 실업자 총계 추정결과

시군구	직접추정법			합성추정법			복합추정법		
	추정값	추정오차	CV (%)	추정값	추정오차	CV (%)	추정값	추정오차	CV (%)
수원시	11,375	3,520	30.9	15,583	453	2.9	15,547	446	2.9
성남시	17,166	4,045	23.6	15,541	778	5.0	15,632	750	4.8
의정부시	10,031	4,262	42.5	5,935	154	2.6	5,952	154	2.6
안양시	15,667	5,134	32.8	9,779	322	3.3	9,822	321	3.3
부천시	20,166	6,228	30.9	12,638	584	4.6	12,731	579	4.6
광명시	14,219	6,106	42.9	5,552	145	2.6	5,568	145	2.6
평택시	3,317	1,771	53.4	5,697	69	1.2	5,705	69	1.2
동두천/과천/구리/군포	14,323	5,104	35.6	9,388	370	4.0	9,433	368	3.9
동두천시							*1,194*		
과천시							*1,105*		
구리시							*2,776*		
군포시							*4,357*		
안산/시흥시	9,979	1,463	14.7	13,914	185	1.3	13,881	182	1.3
안산시							*8,941*		
시흥시							*4,940*		
고양시	4,292	2,880	67.1	12,490	501	4.0	12,276	486	4.0
남양주시	4,063	362	8.9	5,712	57	1.0	5,684	56	1.0
의왕시	0	0	0.0	2,083	249	12.0	2,087	249	12.0
하남시	2,896	1,692	58.4	2,067	9	0.5	2,071	9	0.5
용인시	2,484	1,578	63.5	6,756	181	2.7	6,714	179	2.7
파주시	4,536	424	9.3	3,260	29	0.9	3,272	29	0.9
이천시	1,196	733	61.3	3,006	16	0.5	3,011	16	0.5
안성/광주	0	0	0.0	4,149	326	7.9	4,157	326	7.9
안성시							*2,301*		
광주군							*1,856*		
김포시	422	418	99.1	2,585	13	0.5	2,588	13	0.5
화성군	843	538	63.8	849	56	3.8	916	56	3.7
양주군	2,415	877	36.3	1,548	409	15.0	1,643	336	12.6
여주/연천군	843	679	80.5	1,317	467	20.1	1,136	317	17.2
여주군							*812*		
연천군							*324*		
포천/가평군	1,242	504	40.6	1,454	236	9.2	1,429	193	8.3
포천군							*1,033*		
가평군							*396*		
양평군	422	496	117.5	597	60	5.7	641	60	5.7
계	141,897			141,897			141,897		

(9) 강원도

<표5.30> 강원도의 시/군단위 실업자 총계 추정결과

시군구	직접추정법			합성추정법			복합추정법		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
춘천시	5,618	2,123	37.8	3,558	487	15.2	3,698	463	13.9
원주시	1,825	900	49.3	3,658	196	5.9	3,589	187	5.8
강릉시	5,066	1,681	33.2	3,186	270	9.4	3,252	263	9.0
동해시	718	218	30.4	1,304	37	3.1	1,292	36	3.1
태백/속초시	1,067	367	34.4	1,905	74	4.3	1,880	71	4.2
태백시							*714*		
속초시							*1,166*		
삼척시	349	0	0.0	1,032	18	1.9	933	112	13.3
홍천군	256	242	94.5	367	57	12.9	359	54	12.5
횡성/고성군	482	277	57.5	403	63	13.1	403	60	12.5
횡성군							*226*		
고성군							*177*		
영월군	511	212	41.5	247	28	9.5	251	28	9.2
평창군	227	259	114.1	224	34	12.6	223	33	12.4
정선군	0	0	0.0	287	130	37.9	287	130	37.9
철원/화천/인제군	256	282	110.2	542	60	9.3	527	58	9.1
철원군							*248*		
화천군							*119*		
인제군							*160*		
양양군	511	102	20.0	174	30	14.3	194	27	11.8
계	16,886			16,886			16,886		

(10) 충청북도

<표5.31> 충청북도의 시/군단위 실업자 총계 추정결과

시군구	직접추정법			합성추정법			복합추정법		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
청주시	9,929	1,948	19.6	8,666	1,203	13.2	8,803	871	9.3
제천시	751	475	63.2	2,112	123	5.5	2,006	115	5.4
충주시	3,296	820	24.9	3,199	215	6.4	3,166	201	6.0
보은군	263	227	86.3	347	24	5.9	352	24	5.9
옥천군	790	333	42.2	584	59	8.7	598	57	8.4
영동군	252	189	75.0	650	48	6.5	635	45	6.3
괴산군	779	649	83.3	350	27	6.7	357	27	6.7
음성군	1,294	560	43.3	822	91	9.6	844	88	9.2
청원군	767	307	40.0	1,252	84	5.8	1,233	78	5.6
진천/단양군	779	658	84.5	919	241	22.8	906	213	20.7
진천군							*558*		
단양군							*348*		
계	18,900			18,900			18,900		

(11) 충청남도

<표5.32> 충청남도의 시/군단위 실업자 총계 추정결과

시군구	직접추정법			합성추정법			복합추정법		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
천안시	4,572	993	21.7	4,782	737	12.6	4,568	475	8.8
공주시	1,039	534	51.4	1,497	115	6.3	1,522	110	6.1
보령시	1,486	555	37.3	1,199	53	3.6	1,243	53	3.6
아산시	783	517	66.0	2,112	107	4.1	2,127	103	4.1
서산시	3,302	710	21.5	1,552	94	5.0	1,629	93	4.8
논산시	1,507	1,151	76.4	1,546	89	4.7	1,601	89	4.7
금산군	0	0	0.0	904	183	20.5	901	183	20.5
부여군	528	376	71.2	923	40	4.4	916	40	4.4
서천/청양군	528	587	111.2	1,176	76	6.6	1,161	75	6.5
서천군							*736*		
청양군							*425*		
홍성군	2,493	1,269	50.9	1,007	157	15.8	1,027	155	15.2
예산군	1,535	435	28.3	1,105	54	5.0	1,109	54	4.9
태안/당진군	2,063	1,876	90.9	2,033	776	38.7	2,034	663	32.9
태안군							*732*		
당진군							*1,302*		
계	19,836			19,836			19,836		

(12) 전라북도

<표5.33> 전라북도의 시/군단위 실업자 총계 추정결과

시군구	직접추정법			합성추정법			복합추정법		
	추정값	추정 오차	*CV* (%)	추정값	추정 오차	*CV* (%)	추정값	추정 오차	*CV* (%)
전주시	11,397	2,864	25.1	9,647	999	10.2	9,779	890	9.0
군산시	5,596	2,769	49.5	4,281	489	11.3	4,300	474	10.9
익산시	3,434	2,064	60.1	4,936	556	11.1	4,810	519	10.6
정읍시	886	542	61.2	1,811	36	2.0	1,799	36	2.0
남원시	500	474	94.8	1,355	18	1.3	1,348	18	1.3
김제시	1,701	794	46.7	1,484	57	3.8	1,478	57	3.8
완주/진안군	1,031	625	60.6	758	414	35.0	740	288	25.3
완주군							*534*		
진안군							*206*		
장수/임실군	0	0	0.0	523	202	24.7	531	202	24.7
장수군							*229*		
임실군							*302*		
순창/고창군	723	399	55.2	486	118	15.5	492	108	14.3
순창군							*156*		
고창군							*336*		
부안군	362	348	96.1	350	52	9.4	353	50	9.3
계	25,630			25,630			25,630		

(13) 전라남도

<표5.34> 전라남도의 시/군단위 실업자 총계 추정결과

시군구	직접추정법			합성추정법			복합추정법		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
목포시	12,071	2,251	18.6	5,306	575	11.7	5,739	540	10.1
여수시	4,983	978	19.6	6,175	364	6.4	6,034	320	5.7
순천시	2,842	859	30.2	5,391	243	4.9	5,182	225	4.6
나주시	1,250	1,564	125.1	2,059	120	6.3	2,041	119	6.3
광양시	334	270	80.8	2,549	40	1.7	2,485	39	1.7
담양/함평/장성군	2,434	791	32.5	1,435	315	21.5	1,537	272	17.1
담양군							*575*		
함평군							*435*		
장성군							*527*		
고흥군	1,842	1,219	66.2	725	108	14.7	721	108	14.4
보성/화순/장흥군	0	0	0.0	2,253	673	29.4	2,213	673	29.4
보성군							*698*		
화순군							*913*		
장흥군							*602*		
강진군	0	0	0.0	520	154	29.1	510	154	29.1
해남군	1,250	1,572	125.8	788	128	16.0	777	128	15.8
영암군	0	0	0.0	810	243	29.5	795	243	29.5
무안군	1,250	817	65.4	612	40	6.5	603	40	6.5
영광/진도군									
영광군	1,250	977	78.2	882	112	12.5	871	111	12.3
진도군									
계	29,506			29,506			29,506		

(14) 경상북도

<표5.35> 경상북도의 시/군단위 실업자 총계 추정결과

시군구	직접추정법			합성추정법			복합추정법		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
포항시	8,770	1,953	22.3	6,367	510	6.9	6,426	478	6.4
경주시	1,822	1,341	73.6	3,337	302	7.9	3,242	288	7.7
김천시	1,947	963	49.5	1,855	100	4.7	1,848	99	4.6
안동시	0	0	0.0	3,163	476	13.0	3,153	476	13.0
구미시	8,660	1,437	16.6	4,082	217	4.6	4,146	212	4.4
영주시	636	819	128.8	1,591	17	0.9	1,585	17	0.9
영천/문경시	649	587	90.4	2,517	105	3.6	2,449	101	3.6
영천시							*1,345*		
문경시							*1,104*		
상주시	0	0	0.0	1,338	198	12.8	1,334	198	12.8
경산/성주군	5,192	1,239	23.9	3,425	348	8.8	3,493	322	8.0
경산시							*2,900*		
성주군							*593*		
군위/영양군	0	0	0.0	548	126	24.7	540	126	24.7
군위군							*317*		
영양군							*223*		
의성군	0	0	0.0	680	149	23.5	670	149	23.5
청송/영덕군	0	0	0.0	778	171	23.6	767	171	23.6
청송군							*319*		
영덕군							*448*		
고령/봉화군	0	0	0.0	834	195	25.1	822	195	25.1
고령군							*395*		
봉화군							*427*		
칠곡군	2,097	2,452	116.9	893	284	34.1	898	280	33.0
예천/울진군	2,594	1,055	40.7	958	174	19.5	993	170	18.1
예천군							*469*		
울진군							*524*		
계	32,367			32,367			32,367		

(15) 경상남도

<표5.36> 경상남도의 시/군단위 실업자 총계 추정결과

시군구	직접추정법			합성추정법			복합추정법		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정. 오차	CV (%)
창원시	8,996	2,803	31.2	7,957	739	9.7	8,071	691	9.0
마산/진해시	11,849	3,561	30.1	9,480	606	6.7	9,586	589	6.5
마산시							*7,409*		
진해시							*2,177*		
진주/사천시	4,536	1,044	23.0	7,416	255	3.6	7,283	241	3.5
진주시							*5,486*		
사천시							*1,797*		
통영시	3,771	218	5.8	1,994	26	1.3	2,027	25	1.3
김해시	1,689	1,550	91.8	5,044	342	7.1	4,904	326	7.0
밀양시	3,775	1,501	39.8	1,851	48	2.7	1,857	48	2.7
거제시	1,346	566	42.1	2,645	42	1.7	2,645	42	1.7
양산/의령/합천	4,699	1,251	26.6	4,274	115	2.8	4,289	114	2.8
양산시							*2,988*		
의령군							*476*		
합천군							*825*		
함안/하동군	502	574	114.3	416	177	33.5	410	162	30.7
함안군							*218*		
하동군							*192*		
창녕/고성/산청군	0	0	0.0	1,482	556	29.5	1,466	556	29.5
창녕군							*600*		
고성군							*514*		
산청군							*352*		
남해/함양군	502	382	76.1	352	164	36.5	354	138	30.3
남해군							*198*		
함양군							*156*		
거창군	1,495	357	23.9	250	56	17.5	269	54	15.7
계	43,160			43,160			43,160		

(16) 제주도

<표5.37> 제주도의 시/군단위 실업자 총계 추정결과

시군구	직접추정법			합성추정법			복합추정법		
	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)	추정값	추정 오차	CV (%)
제주시	6,301	1,518	24.1	6,569	1,029	16.8	6,573	705	11.4
서귀포시	2,184	1,050	48.1	1,916	183	10.2	1,912	177	9.9
북제주군	666	225	33.8	349	143	35.2	377	102	21.2
남제주군	0	0	0.0	317	154	42.0	289	154	42.0
계	9,151			9,151			9,151		

5.7.2 효율비교

2001년 5월 시군구단위 소지역들의 추정결과를 기반으로 직접추정값 대비 합성 또는 복합추정값들의 상대효율이득을 요약하였다. 기존 직접추정값 대비 합성 또는 복합추정값들의 상대효율이득은 다음 식으로 주어진다.

$$\text{상대효율이득} = \frac{\text{직접추정값의 } CV\text{값} - \text{합성/복합추정값의 } CV\text{값}}{\text{직접추정값의 } CV\text{값}} * 100(\%)$$

(1) 서울특별시

<표5.38> 직접추정 대비 합성/복합추정의 상대효율이득

시군구	직접추정법		합성추정법			복합추정법		
	추정값	*CV* (%)	추정값	*CV* (%)	효율이득 (%)	추정값	*CV* (%)	효율이득 (%)
종로구	5,037	59.3	3,586	0.9	98.5	3,589	0.9	98.5
중구/용산구	8,569	14.3	7,602	0.8	94.4	7,609	0.8	94.4
중구						*2,788*		
용산구						*4,821*		
성동구	2,579	62.5	6,809	1.5	97.6	6,795	1.5	97.6
광진구	5,077	41.8	8,054	1.6	96.2	8,048	1.6	96.2
동대문구	13,686	33.5	7,940	1.4	95.8	7,950	1.4	95.8
중랑구	12,974	26.1	8,908	1.4	94.6	8,921	1.4	94.6
성북구	12,061	21.9	9,674	1.4	93.6	9,687	1.4	93.6
강북구	7,656	59.6	7,214	3.2	94.6	7,221	3.2	94.6
도봉구	8,690	39.4	7,024	1.2	97.0	7,031	1.2	97.0
노원구	11,228	23.8	12,071	1.0	95.8	12,078	1.0	95.8
은평구	5,117	33.6	9,115	1.2	96.4	9,104	1.2	96.4
서대문구	5,198	23.0	7,369	1.2	94.8	7,358	1.2	94.8
마포구	15,352	30.8	7,749	3.3	89.3	7,777	3.3	89.3
양천구	7,817	39.4	9,285	2.1	94.7	9,284	2.1	94.7
강서구	3,412	42.4	10,343	0.7	98.3	10,332	0.7	98.3
구로구	18,844	19.4	8,324	1.9	90.2	8,350	1.8	90.7
금천구	9,402	30.5	5,471	1.5	95.1	5,479	1.5	95.1
영등포구	2,498	36.0	8,176	0.6	98.3	8,165	0.6	98.3
동작구	6,944	43.8	8,571	1.7	96.1	8,573	1.7	96.1
관악구	13,606	34.8	11,303	2.2	93.7	11,317	2.2	93.7
서초구	6,984	15.7	7,993	0.7	95.5	7,996	0.7	95.5
강남구	8,529	27.3	11,338	1.6	94.1	11,326	1.6	94.1
송파구	2,579	61.4	13,074	1.1	98.2	12,992	1.1	98.2
강동구	12,854	16.3	9,700	1.1	93.3	9,715	1.1	93.3
계	206,693		206,693		평균 95.3	206,693		평균 95.3

(2) 부산광역시

<표5.39> 직접추정 대비 합성/복합추정의 상대효율이득

시군구	직접추정법		합성추정법			복합추정법		
	추정값	CV (%)	추정값	CV (%)	효율이득 (%)	추정값	CV (%)	효율이득 (%)
중구/동래구/기장군	5,751	42.9	9,639	3.0	93.0	9,579	3.0	93.0
중구						2,226		
동래구						4,680		
기장군						2,673		
서구	2,305	38.4	3,507	2.4	93.8	3,495	2.4	93.8
동구	2,910	51.0	2,912	1.5	97.1	2,914	1.5	97.1
영도구	6,356	38.0	4,228	2.4	93.7	4,235	2.4	93.7
부산진구	14,297	23.7	9,764	3.7	84.4	9,817	3.6	84.8
남구	8,569	37.8	6,956	3.0	92.1	6,966	3.0	92.1
북구	6,333	51.7	6,566	3.1	94.0	6,568	3.1	94.0
해운대구	8,056	45.7	8,656	5.2	88.6	8,643	5.1	88.8
사하구	9,779	15.8	8,514	2.2	86.1	8,530	2.2	86.1
금정구	6,264	40.1	6,625	3.5	91.3	6,623	3.4	91.5
강서구	582	81.3	1,249	0.4	99.5	1,250	0.4	99.5
연제구	6,356	26.6	5,110	1.5	94.4	5,116	1.5	94.4
수영구	2,305	38.7	4,108	1.2	96.9	4,104	1.2	96.9
사상구	4,633	29.9	6,662	1.3	95.7	6,656	1.3	95.7
계	84,496		84,496		평균 92.9	84,496		평균 93.0

(3) 인천광역시

<표5.40> 직접추정 대비 합성/복합추정의 상대효율이득

시군구	직접추정법		합성추정법			복합추정법		
	추정값	CV (%)	추정값	CV (%)	효율이득 (%)	추정값	CV (%)	효율이득 (%)
중구	1,450	62.9	1,328	2.7	95.7	1,328	2.7	95.7
동구	2,540	50.5	1,478	1.5	97.0	1,478	1.5	97.0
남구	10,534	18.8	8,166	4.6	75.5	8,235	4.4	76.6
연수구	2,175	54.3	4,580	3.2	94.1	4,537	3.1	94.3
남동구	6,890	20.2	7,321	2.6	87.1	7,304	2.5	87.6
부평구	10,547	19.5	9,800	4.7	75.9	9,807	4.4	77.4
계양구	5,460	39.6	5,608	5.6	85.9	5,597	5.5	86.1
서구	5,832	25.3	5,993	3.0	88.1	5,985	2.9	88.5
강화군	0	0.0	1,155	9.8	-	1,156	9.8	-
계	45,428		45,428		평균 87.4	45,428		평균 87.9

(4) 대구광역시

<표5.41> 직접추정 대비 합성/복합추정의 상대효율이득

시군구	직접추정법		합성추정법			복합추정법		
	추정값	CV (%)	추정값	CV (%)	효율이득 (%)	추정값	CV (%)	효율이득 (%)
중구	1,860	28.2	1,806	1.1	96.1	1,808	1.0	96.5
동구	5,577	20.8	6,735	2.7	87.1	6,703	2.6	87.5
서구	11,617	16.3	5,742	2.7	83.4	5,785	2.6	84.0
남구	1,860	82.2	3,966	3.1	96.2	3,953	3.0	96.4
북구	6,042	30.2	7,710	4.0	86.8	7,656	3.9	87.1
수성구	5,575	38.0	8,848	4.0	89.5	8,750	3.9	89.7
달서구	13,477	16.4	11,243	5.8	64.6	11,393	5.2	68.3
달성군	2,787	62.4	2,744	5.3	91.5	2,746	5.3	91.5
계	48,795		48,795		평균 86.9	48,795		평균 87.6

(5) 광주광역시

<표5.42> 직접추정 대비 합성/복합추정의 상대효율이득

시군구	직접추정법		합성추정법			복합추정법		
	추정값	CV (%)	추정값	CV (%)	효율이득 (%)	추정값	CV (%)	효율이득 (%)
동구	2,807	34.8	2,237	3.5	89.9	2,262	3.4	90.2
서구	4,837	24.1	5,050	6.6	72.6	5,047	6.1	74.7
남구	5,289	21.5	4,265	4.6	78.6	4,328	4.4	79.5
북구	8,629	14.6	8,634	4.7	67.8	8,640	4.2	71.2
광산구	2,761	30.1	4,137	5.5	81.7	4,046	5.2	82.7
계	24,323		24,323		평균 78.1	24,323		평균 79.7

(6) 대전광역시

<표5.43> 직접추정 대비 합성/복합추정의 상대효율이득

시군구	직접추정법		합성추정법			복합추정법		
	추정값	CV (%)	추정값	CV (%)	효율이득 (%)	추정값	CV (%)	효율이득 (%)
동구	6,233	20.6	4,167	5.9	71.4	4,294	5.6	72.8
중구	4,774	16.9	4,169	4.3	74.6	4,244	4.0	76.3
서구	5,002	25.1	7,391	6.3	74.9	7,131	5.7	77.3
유성구	1,212	23.5	2,648	1.8	92.3	2,637	1.8	92.3
대덕구	4,774	29.4	3,620	5.2	82.3	3,689	5.1	82.7
계	21,995		21,995		평균 79.1	21,995		평균 80.3

(7) 울산광역시

<표5.44> 직접추정 대비 합성/복합추정의 상대효율이득

시군구	직접추정법		합성추정법			복합추정법		
	추정값	CV (%)	추정값	CV (%)	효율이득 (%)	추정값	CV (%)	효율이득 (%)
중구	5,628	24.8	3,105	14.6	41.1	3,270	12.2	50.8
남구	5,264	39.8	4,347	16.2	59.3	4,332	14.2	64.3
동구	1,403	111.3	2,466	12.7	88.6	2,368	12.4	88.9
울주군	0	0	2,377	25.3	-	2,325	25.3	-
계	12,295		12,295		평균 63.0	12,295		평균 68.0

(8) 경기도

<표5.45> 직접추정 대비 합성/복합추정의 상대효율이득

시군구	직접추정법		합성추정법			복합추정법		
	추정값	CV (%)	추정값	CV (%)	효율이득 (%)	추정값	CV (%)	효율이득 (%)
수원시	11,375	30.9	15,583	2.9	90.6	15,547	2.9	90.6
성남시	17,166	23.6	15,541	5.0	78.8	15,632	4.8	79.7
의정부시	10,031	42.5	5,935	2.6	93.9	5,952	2.6	93.9
안양시	15,667	32.8	9,779	3.3	89.9	9,822	3.3	89.9
부천시	20,166	30.9	12,638	4.6	85.1	12,731	4.6	85.1
광명시	14,219	42.9	5,552	2.6	93.9	5,568	2.6	93.9
평택시	3,317	53.4	5,697	1.2	97.8	5,705	1.2	97.8
동두천/과천/구리/군포	14,323	35.6	9,388	4.0	88.8	9,433	3.9	89.0
동두천시						*1,194*		
과천시						*1,105*		
구리시						*2,776*		
군포시						*4,357*		
안산/시흥시	9,979	14.7	13,914	1.3	91.2	13,881	1.3	91.2
안산시						*8,941*		
시흥시						*4,940*		
고양시	4,292	67.1	12,490	4.0	94.0	12,276	4.0	94.0
남양주시	4,063	8.9	5,712	1.0	88.8	5,684	1.0	88.8
의왕시	0	0.0	2,083	12.0	-	2,087	12.0	-
하남시	2,896	58.4	2,067	0.5	99.1	2,071	0.5	99.1
용인시	2,484	63.5	6,756	2.7	95.7	6,714	2.7	95.7
파주시	4,536	9.3	3,260	0.9	90.3	3,272	0.9	90.3
이천시	1,196	61.3	3,006	0.5	99.2	3,011	0.5	99.2
안성/광주	0	0.0	4,149	7.9	-	4,157	7.9	-
안성시						*2,301*		
광주군						*1,856*		
김포시	422	99.1	2,585	0.5	99.5	2,588	0.5	99.5
화성군	843	63.8	849	3.8	94.0	916	3.7	94.2
양주군	2,415	36.3	1,548	15.0	58.7	1,643	12.6	65.3
여주/연천군	843	80.5	1,317	20.1	75.0	1,136	17.2	78.6
여주군						*812*		
연천군						*324*		
포천/가평군	1,242	40.6	1,454	9.2	77.3	1,429	8.3	79.6
포천군						*1,033*		
가평군						*396*		
양평군	422	117.5	597	5.7	95.1	641	5.7	95.1
계	141,897		141,897		평균 89.4	141,897		평균 90.0

(9) 강원도

<표5.46> 직접추정 대비 합성/복합추정의 상대효율이득

시군구	직접추정법		합성추정법			복합추정법		
	추정값	*CV* (%)	추정값	*CV* (%)	효율이득 (%)	추정값	*CV* (%)	효율이득 (%)
춘천시	5,618	37.8	3,558	15.2	59.8	3,698	13.9	63.2
원주시	1,825	49.3	3,658	5.9	88.0	3,589	5.8	88.2
강릉시	5,066	33.2	3,186	9.4	71.7	3,252	9.0	72.9
동해시	718	30.4	1,304	3.1	89.8	1,292	3.1	89.8
태백/속초시	1,067	34.4	1,905	4.3	87.5	1,880	4.2	87.8
태백시						*714*		
속초시						*1,166*		
삼척시	349	0.0	1,032	1.9	-	933	13.3	-
홍천군	256	94.5	367	12.9	86.3	359	12.5	86.8
횡성/고성군	482	57.5	403	13.1	77.2	403	12.5	78.3
횡성군						*226*		
고성군						*177*		
영월군	511	41.5	247	9.5	77.1	251	9.2	77.8
평창군	227	114.1	224	12.6	89.0	223	12.4	89.1
정선군	0	0.0	287	37.9	-	287	37.9	-
철원/화천/인제군	256	110.2	542	9.3	91.6	527	9.1	91.7
철원군						*248*		
화천군						*119*		
인제군						*160*		
양양군	511	20.0	174	14.3	28.5	194	11.8	41.0
계	16,886		16,886		평균 77.0	16,886		평균 77.0

(10) 충청북도

<표5.47> 직접추정 대비 합성/복합추정의 상대효율이득

시군구	직접추정법		합성추정법			복합추정법		
	추정값	*CV* (%)	추정값	*CV* (%)	효율이득 (%)	추정값	*CV* (%)	효율이득 (%)
청주시	9,929	19.6	8,666	13.2	32.7	8,803	9.3	52.6
제천시	751	63.2	2,112	5.5	91.3	2,006	5.4	91.5
충주시	3,296	24.9	3,199	6.4	74.3	3,166	6.0	75.9
보은군	263	86.3	347	5.9	93.2	352	5.9	93.2
옥천군	790	42.2	584	8.7	79.4	598	8.4	80.1
영동군	252	75.0	650	6.5	91.3	635	6.3	91.6
괴산군	779	83.3	350	6.7	92.0	357	6.7	92.0
음성군	1,294	43.3	822	9.6	77.8	844	9.2	78.8
청원군	767	40.0	1,252	5.8	85.5	1,233	5.6	86.0
진천/단양군	779	84.5	919	22.8	73.0	906	20.7	75.5
진천군						*558*		
단양군[2]						*348*		
계	18,900		18,900		평균 79.1	18,900		평균 81.7

(11) 충청남도

<표5.48> 직접추정 대비 합성/복합추정의 상대효율이득

시군구	직접추정법		합성추정법			복합추정법		
	추정값	*CV* (%)	추정값	*CV* (%)	효율이득 (%)	추정값	*CV* (%)	효율이득 (%)
천안시	4,572	21.7	4,782	12.6	41.9	4,568	8.8	59.4
공주시	1,039	51.4	1,497	6.3	87.7	1,522	6.1	88.1
보령시	1,486	37.3	1,199	3.6	90.3	1,243	3.6	90.3
아산시	783	66.0	2,112	4.1	93.8	2,127	4.1	93.8
서산시	3,302	21.5	1,552	5.0	76.7	1,629	4.8	77.7
논산시	1,507	76.4	1,546	4.7	93.8	1,601	4.7	93.8
금산군	0	0.0	904	20.5	-	901	20.5	-
부여군	528	71.2	923	4.4	93.8	916	4.4	93.8
서천/청양군	528	111.2	1,176	6.6	94.1	1,161	6.5	94.2
서천군						*736*		
청양군						*425*		
홍성군	2,493	50.9	1,007	15.8	69.0	1,027	15.2	70.1
예산군	1,535	28.3	1,105	5.0	82.3	1,109	4.9	82.7
태안/당진군	2,063	90.9	2,033	38.7	57.4	2,034	32.9	63.8
태안군						*732*		
당진군						*1,302*		
계	19,836		19,836		평균 80.1	19,836		평균 82.5

(12) 전라북도

<표5.49> 직접추정 대비 합성/복합추정의 상대효율이득

시군구	직접추정법		합성추정법			복합추정법		
	추정값	*CV* (%)	추정값	*CV* (%)	효율이득 (%)	추정값	*CV* (%)	효율이득 (%)
전주시	11,397	25.1	9,647	10.2	59.4	9,779	9.0	64.1
군산시	5,596	49.5	4,281	11.3	77.2	4,300	10.9	78.0
익산시	3,434	60.1	4,936	11.1	81.5	4,810	10.6	82.4
정읍시	886	61.2	1,811	2.0	96.7	1,799	2.0	96.7
남원시	500	94.8	1,355	1.3	98.6	1,348	1.3	98.6
김제시	1,701	46.7	1,484	3.8	91.9	1,478	3.8	91.9
완주/진안군	1,031	60.6	758	35.0	42.2	740	25.3	58.3
완주군						*534*		
진안군						*206*		
장수/임실군	0	0.0	523	24.7	-	531	24.7	-
장수군						*229*		
임실군						*302*		
순창/고창군	723	55.2	486	15.5	71.9	492	14.3	74.1
순창군						*156*		
고창군						*336*		
부안군	362	96.1	350	9.4	90.2	353	9.3	90.3
계	25,630		25,630		평균 78.8	25,630		평균 81.6

(13) 전라남도

<표5.50> 직접추정 대비 합성/복합추정의 상대효율이득

시군구	직접추정법		합성추정법			복합추정법		
	추정값	*CV* (%)	추정값	*CV* (%)	효율이득 (%)	추정값	*CV* (%)	효율이득 (%)
목포시	12,071	18.6	5,306	11.7	37.1	5,739	10.1	45.7
여수시	4,983	19.6	6,175	6.4	67.3	6,034	5.7	70.9
순천시	2,842	30.2	5,391	4.9	83.8	5,182	4.6	84.8
나주시	1,250	125.1	2,059	6.3	95.0	2,041	6.3	95.0
광양시	334	80.8	2,549	1.7	97.9	2,485	1.7	97.9
담양/함평/장성군	2,434	32.5	1,435	21.5	33.8	1,537	17.1	47.4
담양군						*575*		
함평군						*435*		
장성군						*527*		
고흥군	1,842	66.2	725	14.7	77.8	721	14.4	78.2
보성/화순/장흥군	0	0.0	2,253	29.4	-	2,213	29.4	-
보성군						*698*		
화순군						*913*		
장흥군						*602*		
강진군	0	0.0	520	29.1	-	510	29.1	-
해남군	1,250	125.8	788	16.0	87.3	777	15.8	87.4
영암군	0	0.0	810	29.5	-	795	29.5	-
무안군	1,250	65.4	612	6.5	90.1	603	6.5	90.1
영광/진도군	1,250	78.2	882	12.5	84.0	871	12.3	84.3
영광군						*548*		
진도군						*323*		
계	29,506		29,506		평균 75.4	29,506		평균 78.2

(14) 경상북도

<표5.51> 직접추정 대비 합성/복합추정의 상대효율이득

시군구	직접추정법		합성추정법			복합추정법		
	추정값	CV (%)	추정값	CV (%)	효율이득 (%)	추정값	CV (%)	효율이득 (%)
포항시	8,770	22.3	6,367	6.9	69.1	6,426	6.4	71.3
경주시	1,822	73.6	3,337	7.9	89.3	3,242	7.7	89.5
김천시	1,947	49.5	1,855	4.7	90.5	1,848	4.6	90.7
안동시	0	0.0	3,163	13.0	-	3,153	13.0	-
구미시	8,660	16.6	4,082	4.6	72.3	4,146	4.4	73.5
영주시	636	128.8	1,591	0.9	99.3	1,585	0.9	99.3
영천/문경시	649	90.4	2,517	3.6	96.0	2,449	3.6	96.0
영천시						*1,345*		
문경시						*1,104*		
상주시	0	0.0	1,338	12.8	-	1,334	12.8	-
경산/성주군	5,192	23.9	3,425	8.8	63.2	3,493	8.0	66.5
경산시						*2,900*		
성주군						*593*		
군위/영양군	0	0.0	548	24.7	-	540	24.7	-
군위군						*317*		
영양군						*223*		
의성군	0	0.0	680	23.5	-	670	23.5	-
청송/영덕군	0	0.0	778	23.6	-	767	23.6	-
청송군						*319*		
영덕군						*448*		
고령/봉화군	0	0.0	834	25.1	-	822	25.1	-
고령군						*395*		
봉화군						*427*		
칠곡군	2,097	116.9	893	34.1	70.8	898	33.0	71.8
예천/울진군	2,594	40.7	958	19.5		993	18.1	
예천군						*469*		
울진군						*524*		
계	32,367		32,367		평균 81.3	32,367		평균 82.3

(15) 경상남도

<표5.52> 직접추정 대비 합성/복합추정의 상대효율이득

시군구	직접추정법		합성추정법			복합추정법		
	추정값	CV (%)	추정값	CV (%)	효율이득 (%)	추정값	CV (%)	효율이득 (%)
창원시	8,996	31.2	7,957	9.7	68.9	8,071	9.0	71.2
마산/진해시	11,849	30.1	9,480	6.7	77.7	9,586	6.5	78.4
마산시						*7,409*		
진해시						*2,177*		
진주/사천시	4,536	23.0	7,416	3.6	84.3	7,283	3.5	84.8
진주시						*5,486*		
사천시						*1,797*		
통영시	3,771	5.8	1,994	1.3	77.6	2,027	1.3	77.6
김해시	1,689	91.8	5,044	7.1	92.3	4,904	7.0	92.4
밀양시	3,775	39.8	1,851	2.7	93.2	1,857	2.7	93.2
거제시	1,346	42.1	2,645	1.7	96.0	2,645	1.7	96.0
양산/의령/합천	4,699	26.6	4,274	2.8	89.5	4,289	2.8	89.5
양산시						*2,988*		
의령군						*476*		
합천군						*825*		
함안/하동군	502	114.3	416	33.5	70.7	410	30.7	70.7
함안군						*218*		
하동군						*192*		
창녕/고성/산청군	0	0.0	1,482	29.5	-	1,466	29.5	-
창녕군						*600*		
고성군						*514*		
산청군						*352*		
남해/함양군	502	76.1	352	36.5	52.0	354	30.3	60.2
남해군						*198*		
함양군						*156*		
거창군	1,495	23.9	250	17.5	26.8	269	15.7	34.3
계	43,160		43,160		평균 75.4	43,160		평균 77.1

(16) 제주도

<표5.53> 직접추정 대비 합성/복합추정의 상대효율이득

시군구	직접추정법		합성추정법			복합추정법		
	추정값	CV (%)	추정값	CV (%)	효율이득 (%)	추정값	CV (%)	효율이득 (%)
제주시	6,301	24.1	6,569	16.8	30.3	6,573	11.4	52.7
서귀포시	2,184	48.1	1,916	10.2	78.8	1,912	9.9	79.4
북제주군	666	33.8	349	35.2	-4.1	377	21.2	37.3
남제주군	0	0.0	317	42.0	-	289	42.0	-
계	9,151		9,151		평균 35.0	9,151		평균 56.5

제6장 시·도 단위의 세부영역 통계 작성

6.1 개 요

6.2 소지역추정법 적용

6.3 SAS 프로그램 알고리즘

6.4 추정결과

제7장 결 언

부록1 시군구 추정결과(전국)

※ 실업자 총계, 취업자 총계, 경제활동인구 총계, 실업률

부록2 특·광역시 및 도지역에 대한 세부항목별 추정결과 (전국)

부록3 시군구 추정 SAS프로그램

(1) 데이터 셀 작성 프로그램

```
/*****************************************************
    각 시도별로 취업자/실업자를 카운트하는 프로그램
*****************************************************/
OPTIONS ls=132 ps=200 NOCENTER NODATE NONUMBER
FONT= 'Terminal' 8
/*********************************************************/
%LET file_1 = 'd:\mySASpgm\rawdata\m0105.txt' ;
              /* path지정:원자료가있는곳 */
%LET file_2 = 'd:\mySASpgm\rawdata\regioncode.txt' ;
              /* path지정:조사구코드가있는곳 */
%LET lib_1 = project ;
       /* method1으로 생성된 data set을 저장할 라이브러리명 */
%LET lib_2 = result ;
       /* method2으로 생성된 data set을 저장할 라이브러리명 */
%LET libpath_1 = 'd:\mySASpgm\sasdata' ;
       /* method1로생성된dataset의물리적경로 */
%LET libpath_2 = 'd:\mySASpgm\result' ;
       /* method2로생성된dataset의물리적경로 */
/*********************************************************/

LIBNAME &lib_1 &libpath_1 ;
LIBNAME &lib_2 &libpath_2 ;

DATA &lib_1..low ;
       INFILE &file_1;
       INPUT josagu 2-6 sex 21 age 22-24 educat 31-33 job 41
workable 50 rat 75-81 .3 ;
RUN

DATA josa1 ;
```

```
        INFILE &file_2;
        INPUT code 9-13 name $ 47-72 ;
RUN ;

DATA josa ;
        FORMAT cityname $12. ;/*cityname을문자형12byte로지정*/
        SET josa1 ;
        k = INDEX(name,' '); /*k:' '의 인덱스*/
        sigun = SUBSTR(name,k-2);
        sigungu = SUBSTR(sigun,1,2);
        ub = SUBSTR(sigun,5);
        ub = ub||' '/*읍면지역을나타내기 위해서*/
        j = INDEX(ub,' ');
        ub = SUBSTR(ub,j-2);
        ubmeun = SUBSTR(ub,1,2);
        IF (sigungu ='시' OR sigungu ='구') THEN citytype=1 ;
              /*citytype=1 : 시구지역*/
        IF (sigungu ='군') THEN citytype = 2 ;
              /*citytype=2 : 군지역*/
        IF (ubmeun ='읍' OR ubmeun='면') THEN towntype= 2 ;
             /*towntype=2 : 읍면지역*/
        IF towntype=. THEN towntype= 1 ;
            /*towntype=1 : 동지역*/
        IF 10<INT(code/1000)<30
            THEN cityname = SUBSTR(name,1,k-1);
            ELSE cityname = SUBSTR(name,1,6);
        keep code cityname citytype towntype;
RUN

DATA one ;
              SET &lib_1..low ;
              numjosa=INT(josagu/1000);
RUN ;

PROC SQL ; /* 조사구번호에 조사구 이름 부여 */
```

```
            CREATE TABLE two AS
            SELECT *
            FROM one  full JOIN josa
            ON one.josagu = josa.code;
            QUIT ;
RUN

DATA three ;
    /* 'm0105.txt'와 'regioncode.txt'에 서로 없는 행 삭제 */
            SET two ;
            IF josagu = . THEN DELETE ;
            IF cityname=' ' THEN DELETE
            DROP code;
RUN ;

DATA &lib_1..low_name ;
      SET three ;
      IF job = 0  THEN  jobtype = 1 ;
                  /* jobtype=1 : 취업자 */
      ELSE IF job = 1 AND workable = 0 THEN  jobtype = 1 ;
      ELSE IF job = 1 AND workable= 1 THEN jobtype= 2 ;
                  /* jobtype=2 : 실업자 */
      ELSE IF job = 2 THEN  jobtype=3 ;
                  /* jobtype=3 : 비경활인구 */
      ELSE DELETE ;
                  /* 잘못입력된 경우 삭제 */
      IF age < 30 THEN agetype=1 ;
      ELSE agetype = 2 ;
RUN ;

%MACRO population(region,i);
  •         DATA acount;
                  SET &lib_1..low_name;
                  IF numjosa = &i ;
            RUN;
```

```
PROC SORT
       DATA = acount; BY cityname citytype
        DESCENDING towntype ;
RUN;

DATA acount;
       SET acount;
       BY cityname;
       IF first.cityname=1
       KEEP cityname citytype towntype numjosa;
RUN;

PROC IMPORT
  DATAFILE= "d:\mySASpgm\rawdata\0105추계인구.xls"
       OUT= a&i
    DBMS=EXCEL2000 REPLACE;
RANGE=&region ;
RUN;

PROC TRANSPOSE
       DATA=a&i OUT=a&i ;
RUN;

DATA a&i;
       SET a&i;
       numjosa = &i ;
       name = COMPRESS(_LABEL_,' ');
       m15_30 = col13 ; m30_ = col14 + col15 ;
       f15_30 = col21 ; f30_ = col22 + col23 ;
       IF numjosa > 30
           THEN cityname = SUBSTR(name,1,6);
       IF numjosa = 26
           THEN cityname = '울산시'||name;
       IF numjosa < 26
```

```
                THEN cityname = name ;
        KEEP cityname m15_30 m30_ f15_30 f30_ ;
RUN ;

PROC SQL;
        CREATE TABLE a&i AS
        SELECT DISTINCT  *
        FROM acount JOIN a&i ON
             acount.cityname=a&i..cityname ;
        QUIT;
RUN;

DATA &lib_1..a&i ;
        SET a&i;
RUN;

%MEND

%population("서울$B5:AG28", 11);
%population("부산$B5:AG28", 21);
%population("대구$B5:AG28", 22);
%population("인천$B5:AG28", 23);
%population("광주$B5:AG28", 24);
%population("대전$B5:AG28", 25);
%population("울산$B5:AG28", 26);
%population("경기$B5:AG28", 31);
%population("강원$B5:AG28", 32);
%population("충북$B5:AG28", 33);
%population("충남$B5:AG28", 34);
%population("전북$B5:AG28", 35);
%population("전남$B5:AG28", 36);
%population("경북$B5:AG28", 37);
%population("경남$B5:AG28", 38);
%population("제주$B5:AG28", 39);
```

(2) 시군구 추정프로그램

```
/*****************************************************
      경제활동인구 계산 프로그램(서울특별시)
*****************************************************/

OPTIONS ls=132 ps=200 NOCENTER NODATE NONUMBER
FONT= 'Terminal' 8

/*****************************************************/
%LET file_1 = 'd:\mySASpgm\rawdata\m0105.txt' ;
      /* path지정:원자료가있는곳 */
%LET file_2 = 'd:\mySASpgm\rawdata\regioncode.txt' ;
      /* path지정:조사구코드가있는곳 */
%LET lib_1 = project ;
      /* method1으로 생성된 data set을 저장할 라이브러리명 */
%LET lib_2 = result ;
      /* method2으로 생성된 data set을 저장할 라이브러리명 */
%LET libpath_1 = 'd:\mySASpgm\sasdata' ;
      /* method1로생성된dataset의물리적경로 */
%LET libpath_2 = 'd:\mySASpgm\result' ;
      /* method2로생성된dataset의물리적경로 */
/*****************************************************/

LIBNAME &lib_1 &libpath_1 ;
LIBNAME &lib_2 &libpath_2 ;

      DATA &lib_1..seoul ;
            SET &lib_1..low_name ;
            IF numjosa=11 ;
      RUN ;

      PROC SORT
            DATA=&lib_1..seoul OUT=one; BY cityname;
      RUN
```

```
DATA one ;
        SET one ;
        BY cityname ;
        IF FIRST.cityname=1
RUN

DATA one;
        SET one;
          /* 소지역의 index부여******one iml을사용하기위해*/
        DO id = 1 TO _N_; END
                id = id-1
        KEEP cityname id;
RUN

DATA two ;
        SET &lib_1..seoul ;
        BY josagu ;
        IF FIRST.josagu = 1 ;
RUN ;

PROC FREQ
        DATA = two;    /* 소지역의 조사구수 카운트 */
        TABLES cityname /OUT=three NOPRINT ;
RUN ;

DATA three;      /* ******************** three */
        SET three;
        RENAME count=josagusu ;
RUN

PROC SQL
        CREATE TABLE all AS
        SELECT DISTINCT *
        FROM all JOIN three ON
```

```
            all.cityname=three.cityname;
        QUIT
RUN ;

PROC SORT
        DATA=&lib_1..seoul OUT=four; BY josagu sex ;
RUN ;

DATA four ;
          /* 소지역의 남녀에따른 숭수  ***four*/
        SET four ;
        BY josagu sex ;
        IF FIRST.josagu=1 OR FIRST.sex=1 ;
        KEEP josagu cityname sex rat citytype towntype;
RUN ;

DATA five ;
        SET two ;
        KEEP cityname josagu josagu  ;
RUN ;

PROC FREQ
        DATA=&lib_1..seoul ;
        TABLES cityname josagu*sex*jobtype*agetype/ OUT =
         six NOPRINT SPARSE
RUN ;

PROC SQL
        CREATE TABLE six AS
        SELECT DISTINCT *
        FROM six JOIN five ON six.josagu=five.josagu; RUN
        CREATE TABLE x1 AS
        SELECT DISTINCT  *
        FROM six JOIN one ON six.cityname = one.cityname
          JOIN three ON six.cityname=three.cityname
```

```
          FULL JOIN four ON six.josagu=four.josagu AND
          six.sex=four.sex;
        QUIT
RUN ;

DATA &lib_1..seoul ;
        SET x1 ;
        DROP percent ;
        LABEL josagusu ='i소지역의 조사구수'
                josagu='조사구번호'
                sex='성별'
                jobtype='1:취업자 2:실업자 3:비경활 '
                agetype='1:34세이하 2:35세이상'
                cityname='i소지역명'
                id= 'i소지역번호'
                count='각범주도수'
                towntype='1:동 2:면읍'
                citytype='1:시구지역 2:군지역'
                rat='승수'
RUN ;

DATA seoul;
        SET &lib_1..seoul;
RUN

PROC IML ;

/***********직접추정값계산*****************************/
        CREATE seoul.est1 VAR{id direct dir_var};
        USE seoul VAR{id josagu sex agetype citytype
         towntype jobtype josagusu count rat  };
        READ ALL INTO x ;
 /*직접 추정값의 계산*/
        si_com = shape(0,2,1) ;/*통합지역의조사구수를동면단위
                                로입력하여합성분산추정시사용*/
```

```
bb    = MAX(x[,1]);        /* 소지역수*/

START var;
        IF mb=1 THEN xi_i = 1 ;
        ELSE xi_i = mb/(2*(mb-1)) ;
        sdir_var = SHAPE(0,2,1) ;
        suih = SHAPE(0,mb,2);
        DO k=1 TO 2 /*성별에 따라서*/
            READ ALL WHERE( (jobtype=1 |
              jobtype=2)& sex=k & id=i ) INTO act;
            READ ALL WHERE(  jobtype=2 & id=i &
              sex=k ) INTO unemp;
              sxi=SUM(act[,9]);  /*성별에따른i소지역의
                                  15세이상 경활인구 총계*/
              syi=SUM(unemp[,9]);  /*성별에 따른i소지
                          역의 15세이상 실업자인구 총계*/
              spi=syi/sxi ;

              DO j=1 TO mb - 1 ;/*조사구에따라서*/
                idjosa1 = act[(j-1)*4+1,2];/*j번째조
                                  사구번호를idjosa에저장*/
                idjosa2 = act[j*4+1,2] ;   /*(j+1)번
                              째조사구번호를idjosa에저장*/
                READ ALL WHERE( sex=k & (jobtype=1
          | jobtype=2) & josagu=idjosa1) INTO two;
                READ ALL WHERE( sex=k & (jobtype=1
          | jobtype=2) & josagu=idjosa2) INTO three;
                sxih = SUM(two[,9]);
                sxih1 = SUM(three[,9]);
                dsxih = sxih - sxih1 ; /*경활인구연속
                                              차분산*/
                READ ALL WHERE( sex=k & jobtype=2 &
                      josagu=idjosa1) INTO four;
                READ ALL WHERE( sex=k & jobtype=2 &
                      josagu=idjosa2) INTO five;
```

```
            syih = SUM(four[,9]);
            syih1 = SUM(five[,9]);
            dsyih = syih - syih1 ; /*실업자연속차
                                        분산*/
            suih[j,k] = dsyih - spi*dsxih ;
            END ;
            sdir_var[k] =
   xi_i*(act[1,10]**2)*(SUM(suih[,k]#suih[,k])) ;
      END
      dir_var = SUM(sdir_var) +
   2*y[1,10]*y[3,10]*xi_i*SUM(suih[,1]#suih[,2]);
FINISH ;

DO i=1 TO bb;
      READ ALL WHERE(jobtype=2 & id=i) INTO y ;

      direct = SUM(y[,9]#y[,10]); /*직접추정값*/
      id = i ;
      mb = y[1,8];          /*i 소지역의조사구수*/

      /*직접 추정값의 분산계산*/
      IF MIN(y[,5])= MAX(y[,6]) THEN DO;/*통합지
                                  역이아닌경우*/
            RUN var; APPEND
      IF MIN(y[,5]) ^= MAX(y[,6]) THEN DO; /*통합
                                    지역의경우*/
            READ ALL WHERE(id=i) INTO select;
            d = ALL(select[,6]=2); /*시지역중읍면
                                  단위만있는경우*/
      IF d = 0 THEN DO;
            dir_var1 = SHAPE(0,2,1) ;/*동지역과
                  읍면단위지역의분산을합하기위해*/
            su = SHAPE(0,2,1) ;
      DO l=1 TO 2 ; /*towntype=1:동 towntype=2
                                        면.읍*/
```

```
READ ALL WHERE(id=i & towntype=1 )
        INTO A;
su[1] = NROW(A)/12 /*동또는읍면의조사
                                구수*/
mb = su[1] ;
IF mb = 1 THEN xi_i= 1 ;
ELSE xi_i = mb/(2*(mb-1)) ;

sdir_var = SHAPE(0,2,1);
suih = SHAPE(0,mb,2);

DO k=1 TO 2/*성별에따라서*/
      READ ALL WHERE(id=i & sex=k &
towntype=1 & (jobtype=1 | jobtype=2)) INTO act;
      READ ALL WHERE(id=i & sex=k &
towntype=1 & jobtype=2 ) INTO unemp;
      sxi=SUM(act[,9]); /*성별에따른
    i소지역의 15세이상 경활인구 총계*/
      syi=SUM(unemp[,9]);/*성별에
따른i소지역의 15세이상 실업자인구 총계*/
      spi=syi/sxi ;

      DO j=1 TO su[1] - 1 ;/*조사구
                        에따라서*/
        idjosa1=act[(j-1)*4+1, 2];
         /*one의1행2열이조사구번호*/
        idjosa2 = act[j*4+1, 2];
           READ ALL WHERE(
towntype=1 & sex=k & (jobtype=1 | jobtype=2) &
josagu=idjosa1 ) INTO two;
           READ ALL WHERE(
towntype=1 & sex=k & (jobtype=1 | jobtype=2) &
josagu=idjosa2 ) INTO three;
           sxih  = SUM(two[,9]);
           sxih1=SUM(three[,9]);
```

```
                                dsxih = sxih - sxih1 ;
                                READ ALL WHERE(
        towntype=1 & sex=k & jobtype=2 & josagu=idjosa1)
        INTO four;
                                READ ALL WHERE(
        towntype=1 & sex=k & jobtype=2 & josagu=idjosa2)
        INTO five;
                                syih  = SUM(four[,9]);
                                syih1 = SUM(five[,9]);
                                dsyih = syih - syih1 ;
                          suih[j,k]=dsyih-spi*dsxih ;
                                END ;
                                sdir_var[k] = xi_i *
        (act[1,10]**2)*(sum(suih[,k]#suih[,k])) ;
                    END
                    uih = SUM(suih[,1]#suih[,2]);
                    dir_var1[1] = SUM(sdir_var) +
         2*A[1,10]*A[3,10]*xi_i*uih ;
              END
              dir_var=SUM(dir_var1) ;
              si_com = si_com || su ;
              APPEND
        END
        IF d=1 THEN DO;
              RUN var; APPEND     END
        END
END


/*******합성추정값계산*****************************/

      CREATE seoul.est2 VAR{id compose com_var};
      USE seoul VAR {id sex agetype jobtype citytype towntype
       count rat };
      read all into sido;
```

```
START sistar;
DO j=1 TO 2/*agetype*/
        m = m+1 n = 0
        DO k=1 TO 3/*jobtype*/
                DO l=1 TO 2/*sex*/
                        n = n+1
                        READ ALL WHERE(citytype=i &
                  agetype=j & jobtype=k & sex=l ) INTO A ;
                        count[m,n]=sum(A[,7]);
                        table[m,n]=sum(A[,7]#A[,8]);
                END
        END
END
FINISH;

IF ALL(sido[,5]=1) =1 then DO;
        m = 0 i=1 table=SHAPE(0,2,6);
        count = SHAPE(0,2,6);
        RUN sistar ;
mact1=(table[1,1]+table[1,3]+table[2,1]+table[2,3]);
        msum1 = mact1 + (table[1,5]+table[2,5]) ;
fact1=(table[1,2]+table[1,4]+table[2,2]+table[2,4]);
        fsum1 = fact1 + (table[1,6]+table[2,6]) ;
        k1=mact1/msum1; k2=fact1/fsum1;
        su1 = table[1,3]/(table[1,1]+table[1,3]);
        su2 = table[2,3]/(table[2,1]+table[2,3]);
        su3 = table[1,4]/(table[1,2]+table[1,4]);
        su4 = table[2,4]/(table[2,2]+table[2,4]);
        s1 = table[1,1]+table[1,3];
s2=table[2,1]+table[2,3];
        s3 = table[1,2]+table[1,4];
s4=table[2,2]+table[2,4];

  USE all VAR{id citytype towntype m15_30 m30_ f15_30
```

```
            f30_ josagusu};
                    READ ALL INTO popul;
                    READ ALL WHERE(citytype=1 ) INTO si;
                    READ ALL WHERE(citytype=2 )INTO gun;

m1=SUM(si[,4])/(count[1,1]+count[1,3]+count[1,5]);
m2=SUM(si[,5])/(count[2,1]+count[2,3]+count[2,5]);
m3=SUM(si[,6])/(count[1,2]+count[1,4]+count[1,6]);
m4=SUM(si[,7])/(count[2,2]+count[2,4]+count[2,6]);
       END

       IF ALL(sido[,5]=1)^=1 then DO ;
              m = 0 table = SHAPE(0,4,6); count = SHAPE(0,4,6);
              DO i = 1 TO 2 /*citytype*/
                     RUN sistar;
              END
              mact1 =
(table[1,1]+table[1,3]+table[2,1]+table[2,3]);
              msum1 = mact1 + (table[1,5]+table[2,5]) ;
              fact1 =
(table[1,2]+table[1,4]+table[2,2]+table[2,4]);
              fsum1 = fact1 + (table[1,6]+table[2,6]) ;
              mact2 =
(table[3,1]+table[3,3]+table[2,1]+table[2,3]);
              msum2 = mact2 + (table[3,5]+table[4,5]) ;
              fact2 =
(table[3,2]+table[3,4]+table[2,2]+table[2,4]);
              fsum2 = fact2 + (table[4,6]+table[4,6]) ;

              k1=mact1/msum1; k2=fact1/fsum1; k3=mact2/msum2;
              k4=fact2/fsum2;

              su1 = table[1,3]/(table[1,1]+table[1,3]);
              su2 = table[2,3]/(table[2,1]+table[2,3]);
              su3 = table[1,4]/(table[1,2]+table[1,4]);
```

```
su4 = table[2,4]/(table[2,2]+table[2,4]);
gu1 = table[3,3]/(table[3,1]+table[3,3]);
gu2 = table[4,3]/(table[4,1]+table[4,3]);
gu3 = table[3,4]/(table[3,2]+table[3,4]);
gu4 = table[4,4]/(table[4,2]+table[4,4]);

s1 = table[1,1]+table[1,3];
s2=table[2,1]+table[2,3];
s3 = table[1,2]+table[1,4];
s4=table[2,2]+table[2,4];
g1 = table[3,1]+table[3,3];
g2=table[4,1]+table[4,3];
g3 = table[3,2]+table[3,4];
g4=table[4,2]+table[4,4];

USE a11 VAR{id citytype towntype m15_30
            m30_ f15_30 f30_ josagusu};
      READ ALL INTO popul;
      READ ALL WHERE(citytype=1 ) INTO si;
      READ ALL WHERE(citytype=2 )INTO gun;
      m1 =
SUM(si[,4])/(count[1,1]+count[1,3]+count[1,5]);
      m2=
SUM(si[,5])/(count[2,1]+count[2,3]+count[2,5]);
   m3=
SUM(si[,6])/(count[1,2]+count[1,4]+count[1,6]);
   m4=
SUM(si[,7])/(count[2,2]+count[2,4]+count[2,6]);
   mstar1=
SUM(gun[,4])/(count[3,1]+count[3,3]+count[3,5]);
   mstar2=
SUM(gun[,5])/(count[4,1]+count[4,3]+count[4,5]);
   mstar3=
SUM(gun[,6])/(count[3,2]+count[3,4]+count[3,6]);
   mstar4=
```

```
    SUM(gun[,7])/(count[4,2]+count[4,4]+count[4,6]);
END

USE seoul VAR {id josagu sex agetype citytype towntype
                              jobtype josagusu count rat };
       READ ALL INTO y ;
       sd=SHAPE(0,MAX(y[,8]),1);
              DO k=1 TO 2 /*성별에 따라서*/
                     DO n=1 TO 2 ; /*나이에따라서*/
                        auih=SHAPE(0,MAX(y[,8]),bb);
                        DO i=1 TO bb;/*소지역에따라서*/
                            READ ALL WHERE(id=i) INTO x;
                            READ ALL WHERE(id=i &
  (jobtype=1 | jobtype=2)& sex=k & agetype=n) INTO one;
                            READ ALL WHERE(id=i &
  jobtype=2 & sex=k & agetype=n) INTO one1;
                            axi=SUM(one[,9]);  /*성별과
                나이에따른i소지역의 15세이상 경활인구 총계*/
                            ayi=SUM(one1[,9]);  /*성별과
                나이에따른i소지역의 15세이상 실업자인구 총계*/
                            api=ayi/axi ;
                            mb=one[1,8];
                            DO j=1 TO mb - 1 ;/*조사구에
                                                따라서*/
                              idjosa1 = one[j*2-1, 2];
                                        /*j번째조사구번호를
                                          idjosa에저장*/
                              idjosa2 = one[j*2+1, 2] ;
                                        /*(j+1)번째조사구
                                       번호를idjosa에저장*/
                              READ ALL WHERE((jobtype=1 |
jobtype=2) & sex=k & agetype=n & josagu=idjosa1) INTO two;
                              READ ALL WHERE((jobtype=1 |
jobtype=2) & sex=k & agetype=n & josagu=idjosa2) INTO
three;
```

```
                        axih = SUM(two[,9]);
                        axih1 = SUM(three[,9]);
                        daxih = axih - axih1 ; /*경활
                                              인구연속차분산*/
                        READ ALL WHERE( jobtype=2 &
      sex=k & agetype=n & josagu=idjosa1) INTO four;
                        READ ALL WHERE( jobtype=2 &
      sex=k & agetype=n & josagu=idjosa2) INTO five;
                        ayih = SUM(four[,8]);
                        ayih1 = SUM(five[,8]);
                        dayih = ayih - ayih1 ;/*실업자
                                              연속차분산*/
                        auih[j,i]=dayih - api*daxih;
                        END ;
                  END
                        sd = sd||auih ;
            END
      END
      sd=sd[,2:4*bb+1] ;
      ui1h = sd[,1:bb];
       ui2h = sd[,bb+1:2*bb];
       ui3h = sd[,2*bb+1:3*bb];
       ui4h = sd[,3*bb+1:4*bb];
       z = 0 ;
START repeat ;
      e1=popul[i,3]*k1; e2=popul[i,4]*k1;
      e3=popul[i,5]*k2; e4=popul[i,6]*k2;
      com_var1 = ((e1/s1)**2)*(m1**2*SUM(ui1h#ui1h))+
       ((e2/s2)**2)*(m2**2*SUM(ui2h#ui2h))+
       ((e3/s3)**2)*(m3**2*SUM(ui3h#ui3h))+
       ((e4/s4)**2)*(m4**2*SUM(ui4h#ui4h))+
        2*e1*e2/s1/s2*m1*m2*SUM(ui1h#ui2h)+
        2*e1*e3/s1/s3*m1*m3*SUM(ui1h#ui3h)+
        2*e1*e4/s1/s4*m1*m4*SUM(ui1h#ui4h)+
        2*e2*e3/s2/s3*m2*m3*SUM(ui2h#ui3h)+
```

```
        2*e2*e4/s2/s4*m2*m4*SUM(ui2h#ui4h)+
        2*e3*e4/s3/s4*m3*m4*SUM(ui3h#ui4h) ;
      com_var = com_var1*xi;
      compose = e1*su1 + e2*su2 + e2*su3 + e4*su4 ;
      id=i;
FINISH ;

DO i=1 TO bb;  /*각소지역에따른합성추정과분산추정*/
      IF  popul[i,2]=1 & popul[i,3]=1 THEN DO;/*시지역인
                                              경우*/
          nj=SUM(si[,7]);
          xi=nj/(2*(nj-1));        /*시지역의조사구수*/
          RUN repeat; APPEND
      END
      IF popul[i,2] = 2 THEN DO; /*군지역인경우*/
        njstar=SUM(gun[,7]);
        xistar=njstar/(2*(njstar-1));/*군지역의조사구수*/
        e1=popul[i,3]*k3; e2=popul[i,4]*k3;
        e3=popul[i,5]*k4; e4=popul[i,6]*k4;
       com_var=((e1/g1)**2)*(mstar1**2)*SUM(ui1h#ui1h)+
        ((e2/g2)**2)*(mstar2**2)*SUM(ui2h#ui2h)+
        ((e3/g3)**2)*(mstar3**2)*SUM(ui3h#ui3h)+
        ((e4/g4)**2)*(mstar4**2)*SUM(ui4h#ui4h)+
         2*e1*e2/g1/g2*mstar1*mstar2*SUM(ui1h#ui2h)+
         2*e1*e3/g1/g3*mstar1*mstar3*SUM(ui1h#ui3h)+
         2*e1*e4/g1/g4*mstar1*mstar4*SUM(ui1h#ui4h)+
         2*e2*e3/g2/g3*mstar2*mstar3*SUM(ui2h#ui3h)+
         2*e2*e4/g2/g4*mstar2*mstar4*SUM(ui2h#ui4h)+
         2*e3*e4/g3/g4*mstar3*mstar4*SUM(ui3h#ui4h) ;
        com_var = com_var*xistar;
        compose = e1*gu1 + e2*gu2 + e3*gu3 + e4*gu4 ;
        id=i;
        APPEND
      END
      IF popul[i,2]=1 & popul[i,3]=2 THEN DO;/*통합지역
```

```
                                                    인경우*/
                    USE seoul VAR{id towntype};
                    READ ALL WHERE(id=i) INTO meun;
                    IF ALL(meun[,2]=2)=1 THEN DO;
                           RUN repeat;
                           APPEND
                    END
                    IF ALL(meun[,2]=2)=0 THEN DO;
                    /*동단위의합성분산*/
                           nj=SUM(si[,7]);
                          xi=nj/(2*(nj-1));  /*시지역의조사구수*/
                           njstar=SUM(gun[,7]);
                          xistar=njstar/(2*(njstar-1));/*군지역
                                                    의조사구수*/
                           z=z+1 ; h = si_com[1,z+1];
                           p = si_com[2,z+1];/*h:동단위조사구수
                                                p:읍면단위조사구수*/
                           e1=popul[i,3]*k1;
                          e2=popul[i,4]*k1;
                           e3=popul[i,5]*k2;
                           e4=popul[i,6]*k2;
                           com_var1=
((e1/s1*h*m1/(h*m1+p*mstar1))**2)*(m1**2*SUM(ui1h#ui1h))+
((e2/s2*h*m2/(h*m2+p*mstar2))**2)*(m2**2*SUM(ui2h#ui2h))+
((e3/s3*h*m3/(h*m3+p*mstar3))**2)*(m3**2*SUM(ui3h#ui3h))+
((e4/s4*h*m4/(h*m4+p*mstar4))**2)*(m4**2*SUM(ui4h#ui4h))+
2*e1*e2/s1/s2*h*m1*h*m2/((h*m1+p*mstar1)*(h*m2+p*mstar2))*
m1*m2*SUM(ui1h#ui2h)+
2*e1*e3/s1/s3*h*m1*h*m3/((h*m1+p*mstar1)*(h*m3+p*mstar3))*
m1*m3*SUM(ui1h#ui3h)+
2*e1*e4/s1/s4*h*m1*h*m4/((h*m1+p*mstar1)*(h*m4+p*mstar4))*
m1*m4*SUM(ui1h#ui4h)+
2*e2*e3/s2/s3*h*m2*h*m3/((h*m2+p*mstar2)*(h*m3+p*mstar3))*
m2*m3*SUM(ui2h#ui3h)+
2*e2*e4/s2/s4*h*m2*h*m4/((h*m2+p*mstar2)*(h*m4+p*mstar4))*
```

```
m2*m4*SUM(ui2h#ui4h)+
2*e3*e4/s3/s4*h*m3*h*m4/((h*m3+p*mstar3)*(h*m4+p*mstar4))*
m3*m4*SUM(ui3h#ui4h) ;
                        com_var2 =
     ((e1/g1*p*mstar1/(h*m1+p*mstar1))**2)*
     (mstar1**2*SUM(ui1h#ui1h))+
     ((e2/g2*p*mstar2/(h*m2+p*mstar2))**2)*
     (mstar2**2*SUM(ui2h#ui2h))+
     ((e3/g3*p*mstar3/(h*m3+p*mstar3))**2)*
     (mstar3**2*SUM(ui3h#ui3h))+
     ((e4/g4*p*mstar4/(h*m4+p*mstar4))**2)*
     (mstar4**2*SUM(ui4h#ui4h))+
      2*e1*e2/g1/g2*p*mstar1*p*mstar2/((h*m1+p*mstar1)*
    (h*m2+p*mstar2))*mstar1*mstar2*SUM(ui1h#ui2h)+
     2*e1*e3/g1/g3*p*mstar1*p*mstar3/((h*m1+p*mstar1)*
     (h*m3+p*mstar3))*mstar1*mstar3*SUM(ui1h#ui3h)+
     2*e1*e4/g1/g4*p*mstar1*p*mstar4/((h*m1+p*mstar1)*
     (h*m4+p*mstar4))*mstar1*mstar4*SUM(ui1h#ui4h)+
     2*e2*e3/g2/g3*p*mstar2*p*mstar3/((h*m2+p*mstar2)*
     (h*m3+p*mstar3))*mstar2*mstar3*SUM(ui2h#ui3h)+
     2*e2*e4/g2/g4*p*mstar2*p*mstar4/((h*m2+p*mstar2)*
     (h*m4+p*mstar4))*mstar2*mstar4*SUM(ui2h#ui4h)+
     2*e3*e4/g3/g4*p*mstar3*p*mstar4/((h*m3+p*mstar3)*
     (h*m4+p*mstar4))*mstar3*mstar4*SUM(ui3h#ui4h) ;
                com_var = com_var1*xi +com_var2*xistar;
                compose1 = e1*h*m1/(h*m1 + p*mstar1)*su1 +
                   e2*h*m2/(h*m2 + p*mstar2)*su2+
                   e3*h*m3/(h*m3 + p*mstar3)*su3+
                   e4*h*m4/(h*m4 + p*mstar4)*su4 ;
                compose2 = e1*h*m3/(h*m1 + p*mstar1)*gu1 +
                   e2*h*m2/(h*m2 + p*mstar2)*gu2+
                   e3*h*m3/(h*m3 + p*mstar3)*gu3 +
                   e4*h*m4/(h*m4 + p*mstar4)*gu4 ;
                compose = compose1 + compose2 ;
                id=i; APPEND
```

```
            END
        END
END
QUIT ;

DATA seoul._est;
        MERGE all seoul.est1 seoul.est2 ;
        BY id;
        KEEP id citytype direct dir_var compose com_var
                josagusu;
RUN

DATA seoul._weight ;
        SET seoul._est ;
                weight = com_var/(com_var + dir_var) ;
        IF direct = 0 THEN DO
                add = compose ;
                add_se = SQRT(com_var) ;
                END
        ELSE DO ;
            add = weight*direct + (1-weight)**2*compose ;
           add_se = SQRT(weight**2*dir_var+
                    (1-weight)**2*com_var) ;
                END
        dir_se = SQRT(dir_var);
        dir_cv = dir_se/direct ;
        com_se = SQRT(com_var);
        com_cv = com_se/compose ;
        add_cv = add_se/add ; RUN

PROC IML ;
        CREATE seoul._bojung VAR{id com_b add_b};
        USE seoul._weight VAR{id citytype direct compose
            add};
```

```
READ ALL WHERE( citytype=1 ) INTO si;
READ ALL WHERE( citytype=2 ) INTO gun;
bb = NROW(si) + NROW(gun) ;
READ ALL INTO A;
DO i=1 TO bb;
       IF (A[i,2]=1 | A[i,2]=3) THEN DO;
         com_b=A[i,4]*(SUM(si[,3])/SUM(si[,4]));
         add_b=A[i,5]*(SUM(si[,3])/SUM(si[,5]));
         id=i; APPEND
       END ;
       IF  A[i,2]=2 THEN DO;
        com_b=A[i,4]*(SUM(gun[,3])/SUM(gun[,4]));
        add_b=A[i,5]*(SUM(gun[,3])/SUM(gun[,5]));
        id=i; APPEND
       END
END
CLOSE seoul._weight ;
USE seoul._weight VAR{ citytype id direct dir_se
   dir_cv com_se com_cv add_se add_cv josagusu};
       READ ALL INTO est ;
CLOSE seoul._weight ;
USE seoul._bojung  VAR{com_b add_b};
       READ ALL INTO bojung ;
CLOSE seoul._bojung ;
estmate = est[,1:5] || bojung[,1] || est[,6:7] ||
       bojung[,2] ||est[,8:10] ;
cname = { citytype id direct dir_se dir_cv com_b
   com_se com_cv add_b add_se add_cv josagusu};

CREATE seoul.est FROM estmate [COLNAME = cname];
       APPEND FROM estmate [COLNAME = cname] ;
CLOSE seoul.iest ;
free;
QUIT
```

```
DATA &lib_1..seoul.est;
        MERGE one seoul.est ;
        by id ;
        drop id ;
        if citytype=1 then sigun='시지역'
        else sigun='군지역'
        LABEL cityname='시군구명'
          sigun   = '시.군지역'
          direct  = '직접추정값'
          dir_se  = '직접추정오차'
          dir_cv  = '직접변동계수'
          com_b   = '합성추정값(direct)'
          com_se  = '합성추정오차(direct)'
          com_cv  = '합성변동계수(direct)'
          add_b   = '복합추정값(direct)'
          add_se  = '복합추정오차(direct)'
          add_cv  = '복합변동계수(direct)'
          josagusu='표본조사구수'
       drop citytype ; RUN

PROC SORT data=&lib_1..seoul.est; BY descending sigun
     cityname; RUN ;

TITLE " seoul 의 추정값 결과"

PROC PRINT    ID sigun ;    RUNTITLE

/**********jacknife 추정**********************/
PROC IML
        CREATE seoul.dir var{num, ysidot};
        USE &lib_1..seoul VAR {id josagu sex agetype
jobtype citytype count josagusu };
        READ ALL INTO josa;
        bb= MAX(josa[,1]);
```

```
            citycount=SHAPE(0,bb,12);
            DO i=1 TO bb ;/*소지역에따라*/
                  n=0 ;
                  DO j=1 TO 2 ;/*sex*/
                        DO k=1 TO 2/*agetype*/
                              DO l=1 TO 3/*jobtype*/
                                    n= n+1 ;
                                    READ ALL WHERE(id=i &
sex=j & agetype=k & jobtype=l) INTO A ;
                                    citycount[i,n] =
SUM(A[,7]) ;
                              END
                        END
                  END
            END

      /*범주별직접추정값계산(범주별실업자총계)*/
USE all VAR{id citytype towntype m15_30 m30_ f15_30 f30_ };
            DO k = 1 TO MAX(josa[,6]) ;/*citytype*/
                  READ ALL WHERE(citytype = k) INTO sigungu;
                  yidir = SHAPE(0,1,4) ;
                  DO i=1 TO 4 ;/*각범주별*/
                        idir= SHAPE(0,NROW(sigungu),1);
                        DO j=1 TO NROW(sigungu) ;/*시군지역의
                                                소지역에따라*/
                          number = sigungu[j,1] ;
                          aj = sigungu[j,4+(i-1)] ;
                  bj = SUM(citycount[number,1+3*(i-1):3*i]);
                          cj = citycount[number,2+3*(i-1)];
                          idir[j,1] = (aj/bj)*cj ;
                        END
                        yidir[1,i] = SUM(idir[,1]) ;
                  END ;
                  ysidot=SHAPE(0,NROW(sigungu),4);
                  DO i=1 TO NROW(sigungu);
```

```
                    DO j=1 TO 4
                      ysidot[i,j] =
            (sigungu[i,j+3]*yidir[1,j])/SUM(sigungu[,j+3]);
                    END
                END
                num = sigungu[,1];
                ysidot = ysidot[,+];
                APPEND ;
            END
            CLOSE seoul.dir ;

      /*새로운합성추정값들을생성*/
START jack(area,josagusu,number,area_n,i,citycount,act,
noemp,popul);
                        habyidir= shape(0,josagusu,1);
                        DO j=1 TO josagusu;/*조사구수에따라*/
                            idir1= SHAPE(0,area_n,4);
                            DO k=1 TO area_n; /*시지역에
                                    해당하는소지역수*/
                                number=area[k,1];/*소지
                                        역번호*/
                                DO l=1 TO 4 /*네가지
                                        범주별*/
                                  aj=area[k,4+(l-1)];
                                  IF number=i THEN DO;
                                      bj1 =
 SUM(citycount[number,1+3*(l-1):3+3*(l-1)])- act[j,l] ;
                                      cj1 =
 citycount[number,2+3*(l-1)]-noemp[j,l];
                                  END
                                  IF number^=i THEN DO;
                                    bj1 =
 SUM(citycount[number,1+3*(l-1):3+3*(l-1)]);
                                    cj1 =
 citycount[number,2+3*(l-1)];
```

```
                                        END
                                    idir1[k,1]=(aj/bj1)*cj1;
                                    END
                                END ;
                                yidir1 = idir1[+,];
                                sumyidir=SHAPE(0,4,1);
                                DO k=1 TO 4 ;
                                    aj= popul[i,4+(k-1)] ;
                                    sumyidir[k,1] =
aj*yidir1[1,k]/SUM(area[,4+(k-1)]) ;
                                END
                                habyidir[j,1] = sumyidir[+,];
                        END
                        ssqyidir=SSQ(habyidir);/*제곱합*/
                        meanyidir=habyidir[:,];/*평균*/
                        APPEND ;
            FINISH;

            READ ALL INTO popul;
            READ ALL WHERE(citytype=1) INTO si;
            READ ALL WHERE(citytype=2) INTO gun;
            nsi= NROW(si);
            ngun=NROW(gun);

            CREATE seoul.com VAR{i meanyidir ssqyidir
josagusu};
            USE &lib_1..seoul VAR {id josagu sex agetype
jobtype citytype count josagusu };
            meanyidir=SHAPE(0,bb,1);
            DO i=1 TO bb;
                READ ALL WHERE(id=i) INTO one;/*각소지역선택*/
                        josagusu=one[1,8];/*i지역의조사구수*/
                        act    = SHAPE(0,josagusu,4) ;
                        noemp = SHAPE(0,josagusu,4) ;
                        number= one[1,1] ; /*i소지역번호*/
```

```
                        IF one[1,8] > 1 THEN DO ;/*i소지역
                            조사구수가두개이상일경우에한하여*/
                              DO j=1 TO one[1,8] ;
                                    idjosagu =
 one[12*(j-1)+1,2] ; m=0
                                    DO k=1 TO 2 ;
                                          DO l=1 TO 2 ;
m=m+1
                                                READ ALL
 WHERE(josagu=idjosagu & sex=k & agetype=l) INTO josagu ;
                                                act[j,m]=
 SUM(josagu[,7]);/*소지역범주별상주조사인구*/

noemp[j,m] = josagu[2,7];     /*소지역의실업자총계*/
                                          END
                                    END
                              END
                        END

                  IF one[1,6]=1 THEN DO;      /*시지역인경우*/
                        RUN jack( si, josagusu, number,
nsi, i, citycount, act, noemp, popul);
                  END

                  IF one[1,6]=2 THEN DO;      /*군지역인경우*/
                        RUN jack( gun, josagusu, number,
ngun, i, citycount, act, noemp, popul);
                  END
            END
      QUIT

      DATA est ;
            SET &lib_1..seoul.est ;
            keep cityname sigun direct dir_se com_se ;
```

```
RUN

PROC SORT DATA=seoul.dir; BY num  ; RUN

PROC SORT DATA=est; BY cityname  ; RUN

DATA seoul.jack;
      MERGE est seoul.dir seoul.com ;
      IF josagusu=1 THEN bias_jn = meanyidir - ysidot ;
            ELSE bias_jn = (josagusu-1)*(meanyidir -
ysidot);
      IF josagusu=1 THEN var_jn=
(ssqyidir/josagusu-meanyidir**2) ;
            ELSE var_jn=
(josagusu-1)*(ssqyidir/josagusu-meanyidir**2) ;
      mse_jn = var_jn + bias_jn**2 ;
      cv_jn = (SQRT(mse_jn)/ysidot)*100 ;

      mse_s = dir_se + bias_jn**2 ;
      wopt_1 = mse_jn/(mse_jn + dir_se) ;
      wopt_2 = mse_s /(mse_s + dir_se) ;
      IF wopt_1=. THEN wopt_1=0 ;
      IF wopt_2=. THEN wopt_2=0 ;

      y_c1 = wopt_1*direct + (1-wopt_1)*ysidot ;
      y_c2 = wopt_2*direct + (1-wopt_2)*ysidot ;
      y_c1dot =wopt_1*direct + (1-wopt_1)*meanyidir ;
      y_c2dot =wopt_2*direct + (1-wopt_2)*meanyidir ;


      bias_c1 =(josagusu-1)*(y_c1dot - y_c1) ;
      IF josagusu=1 THEN var_c1=(wopt_1*direct)**2 +
2*wopt_1*direct*(1-wopt_1)*meanyidir +
((1-wopt_1)**2)*ssqyidir/josagusu - y_c1dot**2 ;
            ELSE var_c1 =
```

```
(josagusu-1)*((wopt_1*direct)**2 +
2*wopt_1*direct*(1-wopt_1)*meanyidir +
((1-wopt_1)**2)*ssqyidir/josagusu - y_c1dot**2) ;
              mse_c1 = var_c1 + bias_c1**2 ;
              cv_c1 = 100*SQRT(mse_c1)/y_c1dot ;

              bias_c2 =(josagusu-1)*(y_c2dot - y_c2) ;
              var_c2=dir_se*wopt_2**2 + com_se*(1-wopt_2)**2 ;
              mse_c2 = var_c2 + bias_c2**2 ;
              cv_c2 = 100*SQRT(mse_c2)/y_c2dot ;
              DROP i num direct dir_se com_se meanyidir
ssqyidir ;
       RUN ;

       TITLE " seoul 지역의 합성추정 최종결과 출력"

       DATA &lib_2..seoul.com ;
              SET seoul.jack ;
              KEEP  cityname sigun ysidot bias_jn var_jn mse_jn
cv_jn ;
       RUN

       PROC SORT DATA=&lib_2..seoul.com ; BY DESENDING sigun ;
RUN

       PROC PRINT ID sigun; RUN ; TITLE

       TITLE " seoul 지역의 복합추정결과 출력(방법1)" ;

       DATA &lib_2..seoul.add_1;
              SET seoul.jack ;
              KEEP  cityname sigun y_c1dot bias_c1 var_c1
mse_c1 cv_c1 ;
       RUN
```

```
PROC SORT DATA=&lib_2..seoul.add_1;BY DESENDING sigun;RUN

PROC PRINT ID sigun ; RUN TITLE ;

TITLE "seoul 지역의 복합추정결과 출력(방법2)" ;

DATA &lib_2..seoul.add_2 ;
      SET seoul.jack ;
      KEEP  cityname sigun y_c2dot bias_c2 var_c2
mse_c2 cv_c2 ;
RUN

PROC SORT DATA=&lib_2..seoul.add_2;BY DESENDING sigun;RUN

PROC PRINT ID sigun ; RUN TITLE
```