제 1 장

지역별고용조사의 무응답 가중치 작성 방법

김서영 · 안다영

제1절 서 론

1. 연구배경

대부분의 표본조사에서는 단위 무응답이 발생한다. 단위 무응답이란 개인 또는 가구 등 조사대상이 응답을 하지 않아서 발생한 응답 단위에서의 결측을 의미한다. 표본조사에서 단위 무응답(이하 무응답)이 발생할 때, 만약 응답자와 무응답자 그룹 간에 어떤 측면에서 특성이 다르게 존재하다면 이렇게 조산된 조사 추정치는 편향이 발생할 수 있다. 이러한 단위 무응답으로 인해 발생한 편향은 많은 경우에 있어서 무응답은 재조사(call-back 또는 follow up survey)나 무응답 가중치 조정을 통해 처리된다. 표본 중 조사에 협조한 응답자만을 이용하여 모집단의 총수나 평균을 추정할 경우 크게 두 가지 문제점이 발생할 수 있는데, 하나는 모수에 대해 편향된 추정을 한다는 것이다. 편향이 0이 되는 경우는 응답자 평균과 무응답자의 평균이 같게 되는 경우뿐이다. 또 다른 문제는 조사의 표본수가 줄어들기 때문에 추정치의 효율성이 떨어진다는 것이다. 따라서 무응답 처리의 기본적인 생각은 대부분의 표본조사에서 발생하는 무응답 편향을 보정하고 감소된 표본의 효율성을 높이고자 하는 것이다(김재광, 2008).

무응답 조정을 위해 가장 일반적으로 사용되는 방법은 무응답 층을 이용한 가중치 조정(weighting) 방법이다. 이 방법은 전체 표본에서 응답자와 무응답자는 보조정보에 기초하여 보정을 위한 조정 셀(adjustment cell)로 나뉘고, 이때 무응답 가중치는 모든 셀에 걸쳐 해당 셀에 포함된 개체들에 대한 응답률의 역수로 계산된다. 이때 보조변수는 응답과 무응답 그룹을 구분할 수 있는 특성이 많이 반영될수록 좋고, 응답자와 무응답자 모두를 포함한 표본 단위에 대해 알려진 값이어야 한다. 계산된 가중치는 설계 가중치에 곱해져서 최종 가중치를 생성하게 된다(Oh and Scheuren, 1983). 무응답 가중치 조정 방

법과 관련이 있는 방법으로 사후층화(post-stratification)방법이 있다. 이 방법은 전체 조정 셀에 대한 모집단의 분포가 센서스와 같은 기준 자료로부터 이용할 수 있을 때 적용 가능한 방법이다. 이 경우 해당 셀에서 응답자 수는 그 셀에서의 모집단 총수의 비(ratio)에 비례하게 된다.

무응답 가중치 조정 방법의 가장 큰 목적은 무응답 편향을 줄이고자 하는 것이다. 그렇지만 그 대신 분산이 증가하는 경우도 발생한다. 즉, 가중치 조정 방법은 무응답 편향을 줄이는 효과는 있지만 분산을 줄이는 효과는 기대하기가 어려울 수 있다. 좋은 보조정보를 사용할 수 있다면 분산이 증가하지 않으면서 편향을 줄일 수도 있다고 알려져 있다. 무응답 조정을 위한 보조정보는 조사결과를 예측할 수 있는 것이면 더 좋다.

무응답 처리를 잘하기 위해서는 무응답 발생 원리를 잘 이해하면 좋다. 무응답이 발생하는 메커니즘은 몇 가지가 있다. 무응답이 조사 결과는 물론이고 보조정보 또는 표본설계와 관련이 없는 경우 MCAR(missing completely at random)이라 한다. 이 경우 표본은 랜덤하고 대표성이 있다고 볼 수 있는데 이때 결과변수와 응답 여부는 서로 독립이다. 만약 응답 여부가 조사 결과와 연관되어 있고, 이런 관계를 관련이 있는 보조정보를 이용하여 모형을 세울 수 있다면 이 자료는 MAR(missing at random)이라 한다. 이때 조사결과는 응답 여부에 대해 조건부 독립이다.

2. 연구내용

표본조사에서 무응답은 조사환경의 변화와 밀접한 관련이 있다. 조사내용이 어렵고 복잡할수록 또는 자료수집방법이 응답자 친화적이지 못할수록, 무응답은 그렇지 않은 경우에 비해 더 발생하기 쉽다. 최근 조사환경이 빠르게 변함에 따라 조사기관에서 무응답 처리에 대한 관심에 더욱 커지고 있다. 통계청의 경우, 국가통계를 생산하는 전문기관으로 무응답 축소 차원에서 조사통계의 신뢰성을 높이기 위한 노력을 지속하고 있다. 하나의 방법으로 더 좋은 조사방법을 개발한다거나 무응답자들의 특성을 파악하여 응답 전환 전략을 세우는 등 다양한 방법을 강구하고 있다. 현재 통계청에서 실시하고 있는 많은 조사는 무응답을 완전히 허용하지 않고 있으나, 일부 가구 대상 조사에서 단위 무응답 처리를 활용하고 있다. 예를 들면 '가계동향조사'는 가구 단위 무응답을 허용하고 무응답 가중치 조정 기법을 적용하고 있다. '지역별고용조사'의 경우 무응답 가구 중 부재로 인한 무응답 가구는 대체(substitution)를, 거부로 인한 무응답 가구는 무응답을 허용하고 있으며, 사후층화 방법에 의해 무응답이 어느 정도 보정되고 있다(김서영과 안다영, 2010). 사후층화 방법은 앞에서 이미 언급한 바와 같이 모집단 분포가 알려져 있어야 하고, 각 층 내에는 관측치가 어느 정도는 포함되어야 하며 사후층화변수 외의 다른 변

수들에 대해서는 경우에 따라서 더 큰 편향을 초래할 수 있다는 점을 사전에 고려해야 한다.

본 연구는 무응답 가중치를 적용하여 '지역별고용조사'의 무응답 편향을 줄이는 최적의 방법을 찾고자 한다. 무응답 조정 방법은 일반적으로 널리 사용되고 있는 무응답층 조정방법과 응답성향을 반영한 성향가중치 방법을 중심으로 검토한다. 성향점수 방법은 Rubin과 Rosenbaum(1983)에 의해 소개되었다. 일반적으로 응답 대상자들의 응답성향을 모르기 때문에 이들의 응답성향을 추정해야 하는데, 이때 로지스틱 회귀모형을 자주 사용한다. Little(1986)은 MAR 자료에 대해 응답성향을 이용한 조정 셀 내에서의 조정 방법을 제시하였다. Little은 응답성향 조정 방법은 무응답 편향은 줄일 수 있지만 분산은 줄일 수 없다는 점을 밝히고 있다.

본 연구의 최종 목적은 무응답 가중치 조정을 통해 지역별고용조사의 무응답 편향을 줄이기 위한 방법을 찾는 것으로, 최적의 무응답 가중치 조정 방법을 찾고자 하는 것이다. 가구 대상 조사의 경우 무응답은 일반적으로 부재(non-contact) 또는 거부(refusal)에 의해 발생한다. 지역별고용조사에서도 이와 같은 현상이 발생하고 있지만 실제 이 조사에서는 부재 가구는 표본대체에 의해 교체하는 방식을 채택하고 있다. 따라서 지역별고용조사는 거부 가구만을 무응답 가구로 허용하고 있는 셈이다. 이러한 맥락에서 본 연구도 거부 가구만을 우선 무응답으로 간주하고자 한다. 그러나 부재 가구는 무시할 수 없는 결측값으로 응답하지 않은 가구와 성향이 완전히 동일한 가구로 대체되지 않는다면 오히려 더 큰 편향을 초래할 수 있을 것이다. 추후 최대한 부재 가구 정보를 파악하여 연구에 반영할 예정이다.

본 연구의 구성은 다음과 같다. 무응답 가중치 적용 방법을 중심으로 무응답 처리 방법을 소개한다. 가중치 조정 방법은 무응답 층을 이용하는 방법, 회귀모형을 이용하는 방법, 성향점수 모형을 이용하는 방법을 중심으로 소개한다. 3장에서는 지역별고용조사에 무응답 가중치를 조정하여 추정치를 구한다. 이때 가중치 조정 방법은 무응답 층과 성향가중치 방법을 고려한다. 이 방법들에 대해 무응답가중치 분포의 기초통계량값을 비교함으로써 각 방법의 특성이 분석한다. 4장에서는 본 연구에서 선택한 무응답 층 가중치 방법들에 대해 실제 지역별고용조사 자료에 적용한다. 시나리오별 가중치 조정 결과에 대해 가중치 조정한 추정치와 추정치의 MSE를 계산하고, 각각의 조정 결과를 비교함으로써 지역별 고용조사에 가장 적합한 시나리오를 찾는다. 마지막으로 각 방법의 특성을 요약하고 향후 연구과제를 통해 앞으로 해야 할 무응답 가중치의 연구 방향을 논의한다.

3. 기대 및 한계

본 연구는 지역별고용조사의 무응답 편향을 보정함으로써 조사 추정치의 신뢰성을 향상시킬 수 있다는 점에서 그 의의가 있다. 본 연구가 이론적으로 잘 알려진 무응답 가중치 조정 방법들을 통해, 지역별고용조사 자료에 가장 적합한 무응답 가중치 조정 방법을 찾고자 한다는 점에서 보면 어느 정도 안정적인 결과를 기대할 수 있을 것이다. 지역별고용조사는 동일한 모집단의 부모집단으로부터 추출된 각각의 표본을 하나의 표본으로 통합하는 방법을 사용한다는 점도 서로 다른 성격의 표본을 복합적으로 사용할 수 있는 계기가 될 것이다. 이는 복합추정량의 개념으로서 최근 서로 다른 성격의 조사를 통합하여 추정하는 방법에 대한 좋은 사례가 될 수 있을 것으로 기대한다.

그러나 모든 연구가 그렇듯이 본 연구도 연구 진행을 어렵게 하는 몇 가지 요인을 가지고 있다. 즉, 무응답 가구에 대한 정보 파악이 여전히 어렵다는 것이다. 이는 연구를 가장 어렵게 하는 요인이기도 하다. 그나마 다행스럽게도 지역별고용조사는 '가구관리종합표'를 통해 무응답 가구에 대한 매우 제한적인 정보를 파악할 수 있도록 하였지만, 표현 그대로 무응답 조정에 핵심적인 정보는 파악이 어렵고, 그렇게 수집된 정보자체가 정확하지 못할 수도 있는 점은 연구의 큰 제약이 되고 있다.

게다가 부재 가구가 표본교체에 의해 대체되고 있다는 점에서도 지역별고용조사의 무응답 편향의 잔재는 여전히 남아 있다. 무응답 가중치 조정은 무응답 가구에 대해 얼마나 좋은 정보를 습득하느냐가 가장 중요할 것이다. 이는 향후 파라테이터(paradata) 수집을 통해 어느 정도 극복할 수 있는 문제이며, 이러한 파라데이터 수집은 통계청의 조사통계의 신뢰성 향상을 위한 좋은 연구 자료로 사용될 수 있을 것이다.

제2절 단위 무응답 조정 방법

무응답이 모수 추정에 대해 편향을 발생시킨다는 것은 이미 잘 알려진 사실이다. 이는 조사 자료가 모집단에 대해 잘못된 추정을 하지 않도록 어떠한 형태로든 무응답을 보정해야 한다는 것을 의미한다. 이를 위한 몇 가지 조정 방법이 있는데, 일반적으로 자주 사용되는 방법으로는 항목무응답은 무응답 대체(imputation)방법, 단위 무응답은 자주 사용되는 조정 방법은 가중치 조정 방법(weighting adjustment)을 들 수 있다. 본 연구에서는 단위 무응답 조정방법에 대해서만 언급할 것이다. 단위 무응답에 대한 가중치 조정 방법은 응답한 개체들에 대해 가중치를 부여하는 방법이다. 이들 가중치는 과소하게 대표되는 그룹보다 과대하게 대표되는 그룹에서 더 작은 가중값을 갖도록 계산된다. 또 다

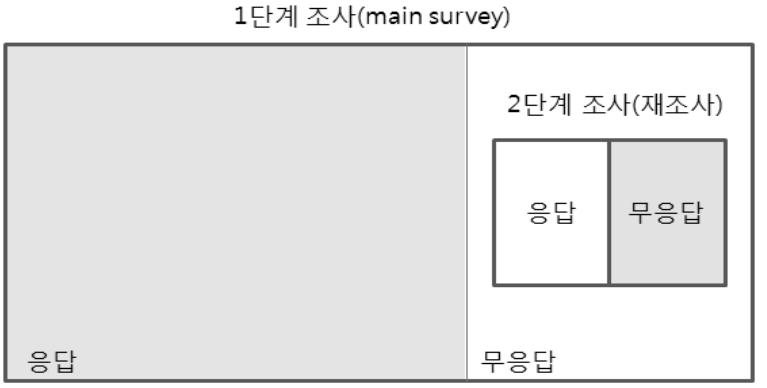
른 방법은 재조사 방법이다. 무응답이 편향된 추정량을 초래하는지를 평가하기 위해서는 무응답자들에 대한 어떠한 형태의 정보가 필요하다. 그러나 무응답자의 정의가 조사과정에서 정보를 제공하지 않은 개체를 의미한다는 점에서 무응답자들에 대해 어떤 정보를 파악한다는 것은 매우 어려운 일이다. 재조사는 무응답 문제를 최소한 부분적으로나마 해결하기 위한 시도이다.

1. 재조사

무응답이 모수 추정치에 대해 편향된 값을 초래한다는 사실은 Hansen과 Hurwitz(1946)에 의해서 처음 거론되었다. 이들이 제안한 방법은 본 조사인 우편조사에서 발생한 무응답자 중에서 표본을 뽑아 이 표본에 대해 다시 대면조사 방법으로 정보를 수집한다는 것이다. 이렇게 2단계에서 수집된 정보는 1단계에서 무응답한 모든 개체들에 대해 대표성을 갖게 되고, 이렇게 수집된 정보를 통해 응답자와 무응답자들 간의 차이를 얻을 수 있다. 이로부터 응답자와 무응답자의 차이라는 정보를 이용해서 무응답 편향을 보정할 수 있다. 만약 1단계에서의 자료수집방법이 우편조사였다면, 2단계의 자료수집방법은 우편조사가 아니라 더 숙련된 전문 조사원을 이용한 대면조사 방법을 이용할 수 있다. 물론 이 경우 조사비용이 증가하게 된다. 즉, 1단계 조사와 2단계 조사의 자료수집 방법이 다르고, 2단계에서의 자료수집 방법은 1단계에서의 방법보다는 더 응답률을 높일 수 있는 방법으로 선택하는 것이 바람직하다 하겠다.

재조사 방법에서 목표모집단은 1단계 응답결과에 대해 응답자 층과 무응답자 층으로 구성된다. 그런 다음 1단계 조사의 무응답자 층 그룹을 부모집단으로 간주하고 이로부터 표본을 랜덤하게 선택한다. 이렇게 선택된 랜덤표본을 대상으로 1단계와 다른 자료수집방법으로 조사를 실시한다. 즉, 재조사 방법은 조사에 반드시 응답하지 않는 개체들로 구성된 부모집단이 있다는 것을 전제로 한다. 재조사 방법은 응답층과 다른 무응답층으로 구성된 다른 부차 모집단(1단계 조사의 무응답 그룹)을 가정한다([그림 1-1]). 만약 2단계 재조사에서의 무응답을 무시할 수 있다면 무응답 층에서의 모수 추정치는 비편향 추정치가 될 것이다. 그러나 실제 상황에서 이러한 조건은 만족되기 매우 어렵고 실제로는 모집단은 응답자, 유연한 응답자(2단계 재조사에서 응답), 완전한 무응답자로 구성된 3개의 층을 구성한다.

예를 들어 전체 n_M명의 무응답자 중에서 랜덤하게 K명에게 다시 재조사를 실시하여 최종 표본을 얻었다고 할 때 자료구조는 <표 1-1>과 같을 것이다.



[그림 1-1] 재조사

<표 1-1>을 간단하게 설명해 보자. 우선 1단계 조사에서 총 표본크기는 N이고, 이중에서 무응답 크기는 n_M이다. 따라서 2단계에서는 재표본추출은 n_M을 모집단으로 간주하여 K개의 표본을 재추출하고 이를 대상으로 2단계 자료수집을 실시하게 된다. 이때 2단계 조사에 총 응답자는 최대 K이거나 그 보다 작게 될 것이다. 따라서 재조사를 포함하여 최종 응답자수는 r이 되고, 각 단계별 조사 추정치는 1단계는 $\hat{y_1}$, 2단계는 $\hat{y_2}$가 된다. 이 두 조사추정치를 이용하여 최종 추정치를 계산할 수 있다.

〈표 1-1〉 재조사 자료 구조

층	모집단크기	1차 표본크기	최종표본크기	최종 추정치
응답자	N_R	n_R	$r1 = n_R$	$\hat{y}_1$
무응답자	N_M	n_M	$r2 = K$	$\hat{y}_2$
전체	N	n	r	

이 두 조사추정치로부터의 불편 추정량은 다음과 같다.

$$\hat{Y} = \frac{n_R}{n}\hat{y_1} + \frac{n_M}{n}\hat{y_2}$$

추정치의 분산을 포함하여 재조사에 대한 구체적인 내용은 Bethlehem(2009)과 김재광(2008)을 참고할 수 있다.

2. 가중치 조정 방법

무응답 가중치 조정 방법은 재조사를 하지 않고 1단계 조사에서의 응답자만의 결과에 대해 가중치를 조정하여 추정량의 편향을 줄이고자 한다. 가중치 조정 방법은 보조변수를 사용하는 것이 기본이다. 보조변수는 조사를 통해 정보를 측정할 수 있거나 모집단 분포를 통해 이용할 수 있는 변수들이 해당될 수 있다. 즉, 응답결과에 영향을 줄 것으로 판단되는 보조변수에 대해 모집단 분포와 조사에서의 응답자 분포를 비교하고, 이를 통해 응답결과가 갖는 모집단에 대한 대표성 유무를 평가할 수 있다. 만약 이 두 자료의 분포가 매우 다르다면 무응답이 어떤 특정 그룹에 대해 선택적으로 이루어졌다고 볼 수 있다. 보조정보는 보정을 위한 가중치 계산에 사용할 수 있고 가중치는 응답한 개체에게 부여하게 된다. 우리가 최종적으로 기대하는 모집단 특성 추정치는 가중된 값을 이용하여 얻어질 수 있는데, 이때 보정변수에 대한 모집단 특성은 오차가 수반되지 않아야 한다.

표본설계 당시의 표본 가중치 $d_i = 1/\pi_i$, π_i는 개체 i에 대한 표본추출확률을 말한다. 만약 무응답이 없다면 HT(Horvitz-Thompson) 추정량은 다음과 같다.

$$\hat{Y}_{HT} = \frac{1}{N}\sum_{i=1}^{N} d_i y_i.$$

무응답이 있는 경우라면 무응답 가중치 보정에 의해 다음과 같이 표현할 수 있다.

$$\hat{y}_w = \frac{1}{N}\sum_{i=1}^{n} w_i y_i.$$

여기서 $w_i = d_i \times c_i$로서 c_i는 무응답 가중치 조정 기법에 의한 보정용 가중치를 나타낸다. 만약 응답이 다양한 보조변수에 대해 대표성이 있고 이들 변수들 모두가 조사된 항목들과 강한 상관관계가 있다면 가중된 표본 또한 이들 항목들에 대해 대표성을 갖게 되고, 따라서 모집단 특성 값에 대한 추정치가 더 정확해질 것이다.

본 절에서는 일반적으로 많이 사용되는 무응답 가중치 조정 방법들 중 몇 가지 방법에 대해 장단점을 위주로 간략하게 소개하기로 한다. 먼저 가장 간단하면서 공통적으로 사용되고 있는 방법 가운데 무응답 층을 이용하는 방법과 사후층화 방법을 설명한다. 그리고 선형 가중치 방법을 설명하고 마지막으로 응답성향 가중치 방법을 소개한다. 실제로 본 연구에서는 무응답 층 방법과 성향가중치 방법만을 적용하는 것을 목표로 한다.

따라서 선형 가중치 방법에 대한 자세한 설명은 Bethlehem(2009), Cobben(2009), 그리고 김재광(2008) 등을 참고하기 바란다.

가. 무응답 층을 이용한 가중치 조정 방법

우선 표본을 응답자와 무응답자 모두에서 이용할 수 있는 변수를 바탕으로 전체 표본을 G개의 서로 배타적인 무응답 층으로 나누었다고 하자. 이때 C를 무응답 층 변수라고 하자. 만약 $C=g$인 경우, n_g는 각 층에서의 표본수, r_g는 각 층에서의 응답자수에 해당된다. 이때 g번째 층에서의 i번째 개체가 갖는 응답확률은 단순하게 $\hat{\phi}_i=\frac{r_g}{n_g}$로 계산된다. 따라서 무응답 층 g내에서 응답자들은 다음과 같은 가중값을 갖는다(송주원과 안형진, 2009).

$$w_i = \frac{r(\pi_i\hat{\phi}_i)^{-1}}{\sum_{k=1}^{r}(\pi_k\hat{\phi}_k)^{-1}}$$

〈표 1-2〉 $g=2$인 경우에 대한 무응답 층 구조

무응답 층($G=2$)	표본수	응답자수	응답확률($\hat{\phi}_g$)
$g=1$	n_1	r_1	$\frac{r_1}{n_1}$
$g=2$	n_2	r_2	$\frac{r_2}{n_2}$

이러한 무응답 가중치 조정 방법은 HT 추정량보다 분산이 크지만 각 무응답 층 내에서 y_i들이 동질적이면 이 추정량의 분산이 작아지게 되어 추정량의 효율이 높아지게 된다.

나. 사후층화

사후층화(post stratification)는 그 자체로서 잘 알려진 방법이고 게다가 가중치 조정 방법으로 자주 사용되는 방법이다. 사후층화는 무응답 편향을 줄이는데도 효과적일 수 있다. 사후층화를 위해서는 한 개 이상의 좋은 보정변수가 필요하다. L개의 범주를 갖

는 보조변수 X가 있다고 하면 그 변수에 대해서 모집단 U는 L개의 층으로 구분된다. 즉, 전체 모집단은 $U_1, U_2, \cdots, U_L$와 같이 분할될 수 있다. 모집단 층 U_h에 포함된 모집단 개체수를 $N_h (h = 1,2,\cdots,L)$라 하자. 그러면 $N = N_1 + N_2 + \cdots + N_L$과 같다. 이 모집단으로부터 n개의 표본이 추출되었고, U_h층에서 n_h개의 표본이 추출되었다고 하자. 이때 $n = n_1 + n_2 + \cdots + n_h$와 같다.

사후층화는 동일한 층 내에서는 모든 개체들에 대해 동일한 가중치를 부여한다. 비복원 랜덤 추출인 경우, U_h층에서 관측된 개체 i에 대한 조정 계수 c_i는

$$c_i = \frac{N_h/N}{n_h/n},$$

와 같다. 설계가중치 $d = n/N$와 조정가중치 c_i를 이용하면 사후층화 추정량은

$$\hat{y}_{pw} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}^{(h)} \quad , \tag{1}$$

와 같다. 여기서 $\bar{y}^{(h)}$는 h층에 포함된 응답자들의 평균을 나타낸다.

사후층화는 무응답 층에 대해 모집단 비율, 즉 N_h/N이 알려져 있다면 무응답 층 조정 방법의 대안으로 사용될 수 있다. 층 내에서의 편향이 작다면 사후층화 추정량은 편향이 작다. 사후층화에서 층은 목표모집과 응답확률 모두에서 동질적이면 좋다. 층 내에서 개체들이 유사할수록 편향은 더 작아질 수 있다. 한편 사후층화는 몇 가지 단점을 포함한다. 즉 층 변수의 개수가 많아지면 셀(cell)의 개수가 더 많이 증가한다. 그렇게 되면 셀 내에 포함된 개체수가 매우 적거나 개체가 하나도 포함되지 않을 수 있다. 따라서 이런 경우 가중치 계산을 할 수 없거나 매우 극단적인 가중값이 생성될 수 있다. 이러한 현상은 층을 이용한 가중치 계산에서 빈번하게 나타날 수 있는 현상이다. 이를 해결하기 위해서는 모집단에 대해 더 많은 정보가 필요하게 되고, 모집단내의 결합분포가 필요할 수 있다.

다. 성향점수 방법

성향점수(propensity score)방법은 조사 자료 분석에서 응답자와 무응답자에 대한 보조변수의 분포 차이를 설명할 목적으로 사용되어 왔다. 이러한 성향점수 방법은 처음에 Rubin과 Rosenbaum(1983)에 의해 개발되었다. 조사 응답 측면에서 성향점수는 어떤 특

성 변수 X_i에 대해 개체 i가 응답할 확률로 적용된다. 즉, i $(i=1,\cdots,n)$번째 개체의 성향점수는 보조변수가 주어졌을 때 어떤 개체가 그 조사에 응답할 확률로 다음과 같이 표현될 수 있다.

$$\rho(\boldsymbol{x})=\Pr(R=1|\boldsymbol{x_i}, i=1,\cdots,n)$$

이러한 응답성향은 일반적으로 알려져 있지 않기 때문에 다음과 같은 로짓 모형과 같은 모형에 의해 추정될 수 있다. 모형은 다음과 같다.

$$\log\left(\frac{\rho(x_i)}{1-\rho(x_i)}\right)=\alpha+\boldsymbol{\beta}'\boldsymbol{x}_i+\epsilon_i,\ i=1,\cdots,n.$$

물론 로짓모형 외에 다른 모형도 사용될 수 있지만, 이미 다른 연구들(Dehija and Whaba, 1999)을 통해서 모형이 달라져도 유사한 결과를 얻는 경우가 많다고 알려져 있다. 자료연계(matching), 층화(stratification), 공분산 조정(covariance adjustment), 가중치(weighting) 방법이 이용될 수 있다(Steinmetz and Tijdens, 2010). 이 중에서 우리는 층화방법과 성향점수의 역수를 이용하여 무응답 가중치 조정에 사용하고자 한다.

① 응답성향 가중치 방법

이 방법은 성향점수의 역수를 가중치로 사용하는 것이다(Rosenbaum, 1984; Schonlau 등, 2007). 응답확률 ρ의 추정치를 조정가중치로 사용할 수 있다. 응답성향은 기본추정량(HT추정량)에 대해 다음과 추정량으로 표현될 수 있다. 이 추정량을 응답성향가중치 추정량(response propensity weighting estimator)이라 부른다.

$$\bar{y}_{ht}^{r}=\frac{1}{N}\sum_{i\in r}\frac{d_i y_i}{\hat{\rho}_i},\ r:\text{응답자}$$

응답성향가중치 추정량이 사후층화추정량 (1)과 유일하게 다른 점은 응답확률 대신에 응답성향을 가중치 조정 계수로 사용한다는 것이다.

Cobben과 Bethlehem(2005)은 이 응답성향 가중치가 항상 좋은 결과를 내는 것은 아니라는 것을 보여주었다. 이들은 모수 추정치가 불안정하다는 것을 예제를 통해 증명한바 있다. 이런 상황은 추정치가 성향점수 계산에 사용된 모형에 매우 의존적이기 때문일 수

있다. 또한 응답성향가중치 방법은 $\hat{\rho}$가 0에 근사해지면 이 값의 역수는 즉, $1/\hat{\rho}$이 무한대가 되어 수렴하지 못하는 문제가 발생할 수 있다.

② 응답성향 층화

성향점수가 직접적으로 사용될 수 있는 또 다른 방법은 응답성향에 기초하여 표본을 층화하는 것이다. 이 경우 층 내에서의 응답성향을 반영하는 확률 $\hat{\rho}$은 같은 값을 갖는다. 이때 응답확률은 보조변수 X에 영향을 받지 않는다. 이는 이들 층 내에서 응답자와 무응답자의 응답 행위가 거의 같다는 것을 의미한다.

층화 방법은 우선 추정된 성향점수에 기초하여 표본을 $\hat{\rho}$의 크기순으로 정렬한 후, 표본을 F개의 층으로 나눈다. 이때 각 층은 거의 동일한 개체수를 포함한다(동점인 경우가 있을 경우 층에 따라 약간 개체수가 다를 수 있다). Cochran(1968)은 분위수 분기점(quartile point)을 이용하여 층의 개수, $F=5$를 제안하였다. 실제로 5개 보다 많은 층을 사용할 경우, 층 구간이 더 좁게 형성되기 때문에 층 내에서 개체들은 훨씬 동질적일 수 있다는 의견도 있다. 본 연구는 5개의 층을 사용하였다. 층은 $s_1, s_2, ..., s_5$이라 하고, 층 f에서의 표본 크기는 n_f라 하자. 이때 표본크기 $n_f(= n_f^{nr} + n_f^r$, n_f^{nr}은 층 f내에 포함된 무응답자수, n_f^r은 응답자수를 각각 나타낸다)는 랜덤변수이다. 사후층화는 같은 층에서 모든 개체들에 대해 동일한 가중치를 부여한다. 층 f내 포함된 개체 i에 대한 조정 가중치 c_i는 다음과 같다.

$$c_{fi} = \frac{n_f}{n_f^r}.$$

따라서 응답성향 층화 방법에 의해 조정된 가중치는

$$w_i^{ps} = c_{fi} \times d_i,$$

이고, 이를 적용한 응답성향 층화(response propensity stratification) 추정량은

$$\bar{y}_{ps}^{\hat{\rho}} = \frac{1}{N} \sum_{f=1}^{F} n_f^r \bar{y}_r^{(f)},$$

과 같다. 여기서 $\bar{y}_r^{(f)}$는 층 f내에서 조사항목에 대한 가중되지 않은 평균을 나타낸다.

그러나 Cobben과 Bethlehem(2005)은 이러한 성향점수를 이용한 층화 조정 방법도 편향을 완전히 조정할 수 없다는 사실을 언급하기도 하였다. 또한 잘 알려진 층의 개수 5보다도 더 많은 층의 개수를 사용하면 약간 더 좋은 조정 효과를 얻을 수 있다고 하였다.

제3절 지역별고용조사와 가구관리종합표 자료분석

1. 지역별고용조사

지역별고용조사는 우리나라 시군 단위에서의 고용 현황을 파악하기 위한 조사로서 매년 조사를 목적으로 2008년 10월에 처음 시작되었고, 2010년부터는 3분기 조사(분기내 마지막 월)를 실시하고 있다. 이 조사는 시군 지역에 대해 세분화된 고용정보를 제공하여 지역의 고용정책 수립에 필요한 정보를 제공하고, 시도별 고용구조 분석 자료와 산업 및 직업에 대한 세분된 자료를 제공하는 것을 목적으로 하고 있다. 지역별고용조사(이하 고용조사)는 조사대상월 15일 현재 대한민국에 상주하는 만 15세 이상인 자를 대상으로 취업, 실업 및 비경제활동 등과 관련된 항목을 조사하고 있다.

가. 표본구조

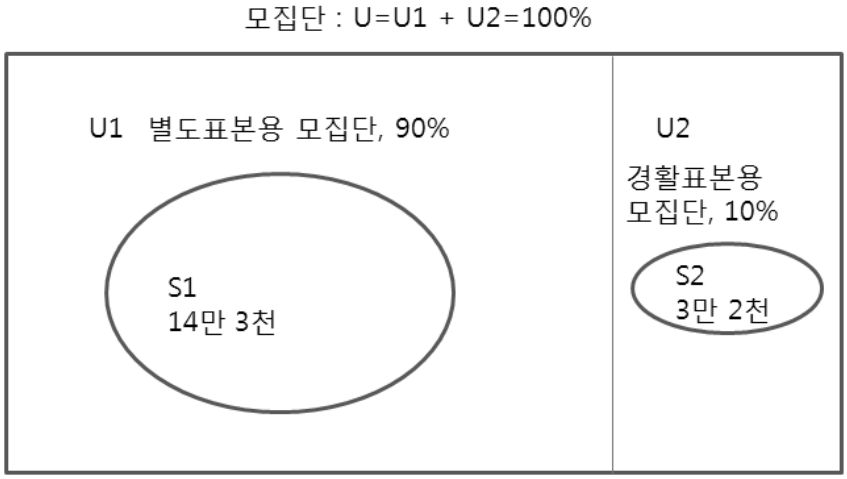
고용조사의 표본은 경제활동인구조사를 목적으로 추출된 표본(이하 경활표본)과 고용조사를 목적으로 별도로 추출된 표본(이하 별도표본)으로 구성되었다. 표본추출은 층화 2단 집락추출방법을 사용하였으며, 1차 추출단위인 조사구는 확률비례추출을 사용하였고, 2차 추출단위인 가구는 단순임의추출방법을 사용하였다. 추출된 가구내의 15세 이상 가구원은 모두 조사대상이 된다. 이렇게 추출된 고용조사의 총 표본규모는 경활표본과 별도표본을 합하여 약 17만 5천여 가구를 사용하고 있다. 2010년 고용조사의 표본규모는 <표 1-3>과 같다(통계청 표본과 2010년 12월).

〈표 1-3〉 2010년 9월 고용조사 표본규모

	별도표본	경활표본	전체
조사구수	7,171	1,629	8,800
가구수	143,000	33,000	176,000

<표 1-3>에 나타난 바와 같이 고용조사 전체 176,000여 가구의 표본은 경활표본과 별

도표본으로 각각 구성되었다. 기본적으로 경활표본은 동일한 조사 시점에서 동일한 항목을 조사하기 때문으로 표본을 그대로 사용하고, 별도표본은 추가로 추출된 개념을 적용하고 있다. 이때 경활표본은 2005년 센서스 자료의 10% 표본을 모집단으로 사용하였고, 별도표본은 2005년 센서스 자료의 표본에 해당되는 10% 표본을 제외한 나머지 전체 90% 자료 (즉, 전수자료(Short from data))를 모집단으로 사용하여 추출되었다. 두 표본은 2005년 센서스라는 동일한 모집단을 사용하고 있지만, 엄밀히 말하면 2005년 센서스를 두 개의 부모집단으로 나누어 각각의 부모집단으로부터 표본을 추출했다고 볼 수 있다. 이때 두 모집단은 서로 배반사건으로 두 부모집단이 겹치는 부분은 존재하지 않는다. 고용조사의 표본추출 원리를 그림으로 설명하면 [그림 1-3]과 같다. [그림 1-3]에서 2005년 센서스 자료의 전체집합을 U라 하면, U는 다시 센서스 자료의 90%에 해당하는 $U1$과 10%에 자료에 해당하는 $U2$로 나눌 수 있다. 이때 $U1 \cap U2 = \varnothing$이다. 부모집단 $U1$과 $U2$로부터 별도표본 $S1$과 경활표본 $S2$가 각각 추출되었다. 물론 $S1$과 $S2$의 추출 과정이 동시에 이루어진 것은 아니고, 조사의 순서상 경활표본이 먼저 추출되어 사용되었고, 추가로 별도표본이 추출되었다고 볼 수 있겠다. 즉, 고용조사는 100% 센서스 자료에 대해 두 개 서로 다른 독립표본이 통합되어 있는 상태라 할 수 있다.



[그림 1-2] 고용조사 표본추출 원리

나. 고용조사의 무응답 발생 현황

대부분의 표본조사가 그렇듯이 고용조사에서도 단위 무응답이 발생한다. 무응답이 MAR이 아니라면 이 무응답은 조사 추정치의 편향을 초래하는 것으로 알려져 있다. 따

라서 조사 추정치의 신뢰성을 높이기 위해서는 무응답 때문에 발생하는 편향을 줄일 필요가 있다. 여기서 우리가 더 신중하게 접근해야 할 것은 고용조사가 경활표본과 별도표본이 통합된 형태의 자료라는 점이다. 크게는 동일한 모집단으로부터 추출된 표본이라 하더라도 두 표본에 대한 조사가 별도로 이루어지기 때문에 두 표본을 이용한 조사는 완전히 독립적이라 할 수 있다. 이렇게 구성된 각각의 조사에서 발생하는 무응답 환경도 매우 다른 조건을 갖는다고 볼 수 있다. 경활표본을 이용한 조사(이하 경활조사)와 별도표본을 이용한 조사(이하 별도조사)는 여러 가지 측면에서 주어진 조건이 다르기 때문이다. 즉, 경활조사는 우선 훈련이 잘 되어 있는 숙련된 조사원이 조사에 투입된다는 점과 일회성 조사가 아닌 한 번 표본에 포함되면 36개월 동안 표본으로 조사에 참여하게 된다는 점이다. 즉, 조사원과 표본이 상당히 안정적인 조사라는 점을 기억할 필요가 있다. 반면에 별도조사는 임시조사원이 조사에 투입되며 일회성 조사의 성격을 지니게 된다. 게다가 별도조사는 소지역 통계 생산이 목적인만큼 경활조사에 비해 농어촌 지역들이 많이 포함되었다는 점에서도 무응답 발생 환경이 경활조사와는 어떤 면에서는 상당히 다를 것으로 판단된다. 따라서 경활조사와 별도조사는 각각 독립적인 무응답 발생환경으로 간주하여 무응답을 보정해야 할 것이다.

고용조사에서 무응답이 발생하는 이유는 크게 부재, 거부, 기타의 이유를 들 수 있다(김서영과 안다영, 2010). 현재 경활조사는 무응답을 허용하고 있지만 무응답 조정은 하지 않고 있다. 경활조사 무응답은 부재와 거부, 기타 등 어떤 이유에서든 발생한 무응답은 실질적 무응답으로 인정하고 있지만 그 비율은 그렇게 크지 않다('10년 9월 경활조사 실제 무응답률 약 5.9%). 별도조사는 전체 무응답 이유에 대해서 일부만 무응답으로 허용하고 있다. 즉, 전체 무응답 중 부재에 의한 무응답은 무응답한 가구와 특성이 유사한 가구로 교체하여 조사가 이루어지고 있는 반면, 거부 가구는 교체하지 않고 무응답 그대로 인정하고 있다. 어쨌건 실무통계에서 무응답은 랜덤하게 발생하지 않는다는 점에서 무응답 편향을 일으키는 원인이 되고 있다. 사실 부재한 가구의 성향이 고용특성 측면에서 응답 가구의 성향과 많이 닮아있다는 점을 볼 때(김서영과 안다영, 2010), 통계의 신뢰성을 높이기 위해 실제로 거부 가구와 마찬가지로 부재 가구도 무응답 가구로 허용되는 것이 더 바람직할 것이다.

가구 단위 무응답 문제를 해결하기 위해서는 무응답 가구에 대해 정보를 파악하는 것이 가장 중요하다. 그러나 현실적으로 무응답 가구는 조사에 협조하지 않는 가구라는 점에서 이들 가구들로부터 어떠한 정보를 획득한다는 것은 매우 어려운 일이 아닐 수 없다. 통계조사에 있어서 무응답 가구의 정보를 파악할 수만 있다면 이는 통계의 신뢰성 향상을 위한 연구 자료로 상당한 가치가 있을 것이다. 현재 고용조사에서 무응답 가구의 성향 파악은 매우 일부이긴 하지만 고용조사의 '가구관리종합표'에 기록된 자료를 통해

파악할 수 있다. 이 가구관리종합표는 조사원들의 가구방문실태를 기록하기 위한 것으로 조사원이 방문한 가구에 대해 그 가구의 주택유형(주택/아파트/기타)과 방문회차별 가구들의 응답상태(응답/거부/부재/기타 등)를 주로 기록하도록 되어 있다. 이로부터 무응답 가구에 대해서 주택유형, 방문횟수, 가구의 응답성향 등을 파악할 수 있다. 여기서 응답성향은 그 가구에 초기 응답 가구로서 조사에 협조적인 가구인지, 아니면 1, 2, 3회차 부재 또는 거부 등의 이유로 응답을 얻지 못하다가 마지막 4회 차에서 응답 또는 거부 및 부재 등에 따른 무응답 가구들인지를 파악할 수 있다. 따라서 이러한 자료들은 이후 무응답 조정을 위한 보조변수로서 활용될 수 있게 된다. 참고로 경활 또는 별도조사는 해당 가구에 대해 4번의 방문을 허용하고 마지막 4번째 방문에서 응답을 얻지 못할 경우 다른 유사가구로 대체하여 조사하는 방식을 취하고 있다.

2. 가구관리종합표 자료

가. 가구관리종합표 자료분석

별도조사에 대한 무응답 가구의 정보는 가구관리종합표에 기록된 정보를 이용하여 파악할 수 있다. 이 자료 분석의 목적은 본 연구에서 사용될 보조변수를 탐색하기 위한 것으로서 실제로 고용특성, 특히 실업과 다른 변수들과는 어떤 관계가 있으며, 특정 층별로 응답률은 어떻게 다른지 또는 같은지를 파악하려는데 그 목적이 있다. 특히 지리적 층 변수(시도 또는 시군구)는 무응답 층 변수의 기본으로서 일반적으로 사용되고 있다는 점에서 이들 지역 간의 실업률 또는 응답률 실태를 분석해 볼 필요가 있겠다.

1) 분석자료 설명

본 연구에 사용된 자료는 고용조사 원자료 및 가구관리종합표에 기록된 정보를 연계하여 분석한다. 고용조사 결과를 공표할 때 사용하는 원시자료와 가구관리종합표 자료는 별도로 관리되기 때문이다. 이를 연계하기 위하여 행정구역부호, 조사구번호, 거처 및 가구번호가 조합된 가구의 고유코드를 활용하게 된다. 본 연구에 사용된 가구관리종합표자료는 2010년 9월에 실시된 고용조사에 따른 것으로, 이 시점에서 지역별고용조사는 지방자치단체 일부와 공동조사로 이루어졌으며(경기, 군산, 창원) 이들 지역에서의 조사구 내 가구관리종합표는 작성되지 않았다. 그리고 경활표본 가구에 대한 기록도 포함되지 않았다. 가구관리종합표의 자료는 가구단위 정보이며 세부적으로는 행정구역부호 등 가구고유코드와 거처종류, 가구방문 회차별 응답상태, 조사방법선호도 등의 정보가 포함되어 있다.

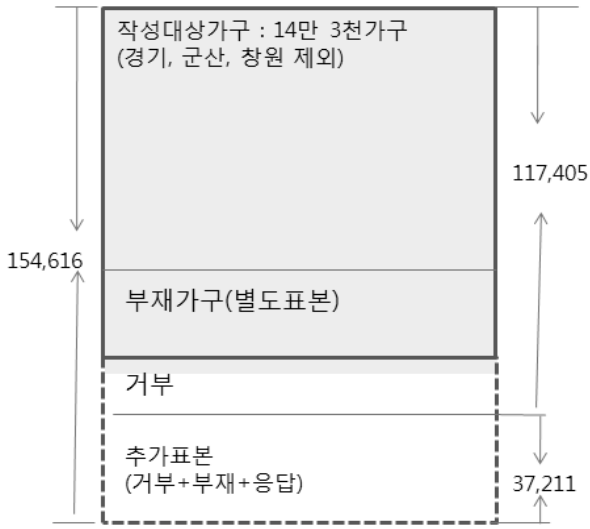
2) 기초자료 구성

고용조사 별도표본조사의 가구관리종합표 자료 분석을 위한 기초자료 자체는 간단하다. 즉, 별도표본 추출당시에 표본가구수에 대해 방문정보를 수집하고, 조사에 대해 응답하지 않은 가구에 대해서는 기본적으로 4번째 방문이 완료되고 난 이후에 교체표본으로부터 조사를 실시하게 된다. 따라서 가구관리종합표 자료는 표본추출 당시의 별도표본 가구수와 무응답 가구를 대체하기 위해 조사된 추가표본 가구로 구성되었다. 그리고 이렇게 추가된 표본 중 일부는 실제로 거부로 인한 무응답 가구의 대체용으로 사용되었다. 가구관리종합표의 레코드수와 실제 고용조사 집계에 사용한 자료의 레코드수가 완벽하게 일치하지는 않는다. 이는 가구관리종합표 자료가 여러 가지 이유에서 완전하지 않을 수 있다는 점을 그 이유가 될 수 있다. 따라서 실제 분석용 자료는 다소 복잡한 자료 연계과정을 거쳐 구성되고, 실제로 정확하게 모든 레코드 연계가 성립하지 않는다. 이에 대해서는 가구관리종합표 자료로부터 얻을 수 있는 파라데이터 정보를 활용하기 위해서는 좀 더 신중하게 자료수집에 접근할 수 있도록 장치를 마련할 필요가 있을 것이다.

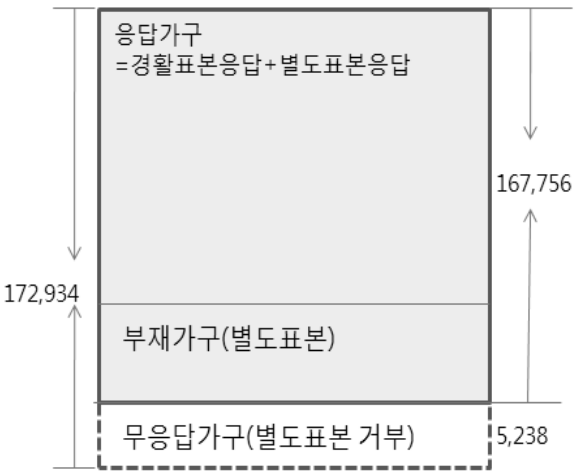
이제 가구관리종합표 자료를 정리해 보자. 우선 가구관리종합표 자료에 포함된 전체 가구수는 154,886가구(이 중 '주택유형'변수와 연계가 되지 않는 270가구 제외)로 이는 추출당시의 표본과 추가 조사된 표본을 합한 규모이다. 그리고 실제 고용조사의 조사표 자료 내에는 응답 가구와 거부로 인한 무응답 가구가 포함되어 있다. 이론적으로 거부가구 레코드는 가구관리종합자료의 거부 가구 레코드와 완벽하게 일치해야 하지만 실제로는 완벽하게 일치하지는 않는다. 조사표 자료 내에 포함된 거부 가구의 레코드 수는 5,237가구에 해당된다. 가구관리종합표자료의 구성 현황을 정리하면 [그림 1-3]과 같다. 가구관리종합표가 작성되어야 할 대상 가구수는 약 14만 3천여 가구이지만, 해당시점에서 경기, 창원, 군산 등의 지자체 공동 실시로 인해 이 지역은 가구관리종합표가 작성되지 않았고, 게다가 경활은 가구관리종합표가 작성되지 않았다. 실제로 가구관리종합표 자료와 조사표자료내의 거부 가구 레코드가 완전히 연계되는 가구수는 4,431여 가구수로 약 806여 가구가 연계되지 않는다. 이는 가구관리종합표 기록이 정확하지 않거나 하는 등의 이유에서 비롯된 것으로 짐작할 수 있다.

최종적으로 가구 단위 무응답을 조정을 위해서는 고용조사에 해당하는 조사표의 가구 레코드가 필요하다. 가구관리종합표 자료의 무응답 가구를 제외한, 즉 거부 가구 5,237가구를 제외한 조사표 레코드 수는 168,143으로 여기에는 경활표본 중 응답한 가구 레코드와 가구관리종합표에는 기록되지 않은 경기도, 군산, 창원을 포함한 수치이다. 이 중에서 부재 가구수가 정확히 몇 가구인지는 알 수 없지만, 가구관리의 부재 가구에 근사한 수치 정도라고 미루어 짐작할 수 있다. 따라서 최종 무응답조정을 위한 가구 레코

드는 고용조사 조사표 레코드에 포함된 168,143가구와 거부 무응답 레코드 5,237가구를 포함하여 173,380가구에 해당된다.



[그림 1-3] 가구관리종합표 자료 구성



[그림 1-4] 무응답 조정을 위한 최종자료 형태

다. 가구관리종합표 자료 분석

본 절에서는 무응답 가구 조정을 위해 필요한 무응답 층 변수 또는 보조변수를 찾기 위한 과정으로, 가구관리종합표를 통해 수집된 자료에 대해 분석한다. 실제로는 무응답 가구에 대한 정보가 매우 부족한 것이 연구의 어려운 현실이기도 하다. 무응답 조정을 위한 보조변수는 응답 가구와 무응답 가구 모두에 대해 정보를 파악할 수 있어야 한다. 현재 이용할 수 있는 보조변수 후보로는 조사대상가구들에 대한 주택유형(주택/아파트/기타)과 응답을 얻기까지 조사원이 가구를 방문한 횟수, 그리고 대상가구들의 응답성향변수가 있다. 응답성향변수는 4번째 방문까지 가구의 조사협조에 대한 속성으로 크게는 거부 또는 부재 속성으로 분류될 수 있다. 가구 속성을 구분하기는 쉽지 않다. 구분을 위한 기준이 명확하지 않기 때문이다. 본 연구는 4번 방문 중 3차 방문까지에서 응답을 얻은 가구는 조사에 대해 쉽게 협조하는 협조성향이 있는 가구, 그리고 마지막 방문에서 응답 가구와 거부 가구에 대해 각각 거부가 더 많았으면 거부 속성 가구, 부재가 더 많았으면 부재 속성 가구 등 3가지 속성으로 분류하기로 하였다. 이와 다른 분류도 가능할 것이다.

고용조사의 소지역별 고용동향을 파악한다는 점을 주목하여 실업률과 고용률이 주택유형 또는 지리적 단위별로 어떻게 다른지를 파악하고자 하였다. 그리고 이들 변수를 중심으로 응답률이 어떻게 다른지를 살펴보았다. 본 연구의 핵심이 무응답 조정 방법으로 무응답 가중치 방법을 사용한다는 측면에서, 주택유형, 응답속성, 지리적 특성 변수 등이 무응답 층 형성을 위한 변수와 응답성향변수로 활용할 수 있는지 여부를 파악하는데 초점을 맞추었다. 분석결과는 다음과 같다.

1) 응답유형별, 주택유형별 분포 및 고용 현황

〈표 1-4〉 가구관리종합표 모든 응답유형별 응답 분포

	빈도(가구)	%
응답	118,769	76.82
거부	5,866	3.79
부재	9,547	6.17
빈집	18,350	11.87
불능	219	0.14
철거	914	0.59
예비	951	0.62
총합	154,616	

<표 1-4>는 추가표본을 포함한 가구관리종합표에 기록된 모든 레코드에 대해서 각 응답유형별 분포를 나타낸 것이다. 전체 가구 중에서 응답이 76.82%로 가장 많고, 빈집이 11.87%로 상당히 높은 비중을 차지하고 있다. 이는 고용조사가 시군구 대상으로 이루어진다는 점에서 우리나라 주택구조가 농촌으로 갈수록 주택이 많고 사람이 살지 않은 빈집이 많다는 점에서 이해가 되는 부분이다. 실제 표본추출 당시에는 해당 가구에 사람이 거주하는지의 여부를 파악하기 어렵다는 점도 이러한 상황에 직면케 한다고 보인다. 그 다음으로 부재와 거부 가구 순으로 분포되는 것으로 나타났다. 빈집이나 철거로 인한 가구에 대한 처리여부에 대해서는 추후 논의될 필요가 있을 것이다. 이러한 가구들을 대체 대상으로 할지 아니면 불능가구로 처리할 지의 여부에 따라 조사의 실제 응답률이 달라질 수 있기 때문이다.

〈표 1-5〉 주택유형별 응답분포 현황

단위: 가구수, %

	주택		아파트		기타		총합
	빈도	%	빈도	%	빈도	%	
응답	73,877	62.20	36,245	30.52	8,647	7.28	118,769
거부	1,965	33.50	3,251	55.42	650	11.08	5,866
부재	4,108	43.03	4,167	43.65	1,272	13.32	9,547
빈집	15,279	83.26	1,630	8.88	1,441	7.85	18,350
불능	138	63.01	70	31.96	11	5.02	219
철거	709	77.57	151	16.52	54	5.91	914
예비	586	61.62	236	24.82	129	13.56	951
총합	96,662	62.52	45,750	29.59	12,204	7.89	154,616

<표 1-5>는 주택유형별 응답분포를 나타낸 것이다. 전체 대상가구 중 62.52%가 주택가구이고, 29.59%는 아파트, 나머지는 기타유형에 속한다. 주택유형별 응답분포는 전체 응답 현황인 <표 1-4>와 크게 다르지 않지만, 주택 가구에서 빈집이 다른 유형에 비해 훨씬 높다는 것을 알 수 있다.

그렇다면 주택유형별 고용특성이 어떻게 다른지 살펴보자. <표 1-6>은 가구관리종합표 자료를 바탕으로 조사된 값을 기준으로 작성된 수치이다. 주택유형별로 보면, 기타 주택유형이 전체 가구유형 중 차지하는 비중이 제일 작지만 실업률이 가장 낮은 것으로 나타났다. 아파트가 주택가구보다 실업률이 더 높고, 고용률도 낮은 것을 알 수 있다. 즉, 아파트 가구의 비경제활동 비율(이하 비경률)은 높다. 상대적으로 주택 가구는 농어촌 지역에 많이 분포되어 있고, 대부분 농림어업에 종사하는 것을 응답하는 경향이 있기 때문에 고용률이 아파트에서보다 더 높을 것으로 짐작할 수 있다.

〈표 1-6〉 주택유형별 고용률, 실업률, 비경률

	고용률	실업률	비경률
주택	61.26	1.90	37.56
아파트	55.04	2.82	43.36
기타	56.97	3.58	40.92
전체	58.89	2.03	39.71

◎ 고용률 = (취업자÷15세이상인구)×100
실업률 = (실업자÷경활인구)×100
비경률 = (비경인구수÷15세이상인구)×100

이제 가구를 응답속성별로 구분해 보자. 응답 속성 그룹에 따라 고용현황이 다를 수 있는지의 여부를 살펴보고자 한다. 우선 전체 가구 중 부재와 거부성향이 강한 그룹의 분포는 <표 1-7>과 같다. 부재성향과 거부성향이 강한 가구를 명확히 구분하기는 어렵지만, 여기서는 4회차까지 방문이 이루어진 가구 중 3번 이상 부재한 가구와 3번 이상 거부한 가구로 구분하였다. <표 1-7>에서 부재성향 가구 13,164가구 중 3번 연속 부재 후 응답 가구는 그 중 52.18%이고 그 중 나머지 47.82%는 부재인 것으로 나타났다. 거부 그룹은 총 1,280가구 중에서 연속 3번 거부 후 응답한 가구는 그 중 11.33%, 나머지 88.67%는 여전히 거부한 것을 알 수 있다. 이로부터 부재 가구는 지속적인 재접촉 시도를 통해 더 많은 응답을 얻을 수 있지만, 거부 가구는 응답전환 설득을 통해서는 쉽게 조사에 협조하지 않은 경향이 있다고 볼 수 있다. 전체 부재 또는 거부 그룹 1,444가구 중 3번 거부 후 마지막에 응답한 가구의 비율은 1%로 매우 낮은 반면, 3번 부재 후 응답한 가구는 47.56%로 상당히 높은 비율을 나타내고 있다.

이처럼 우리나라의 가구 특성상 부재 가구에 대해서 방문회차를 늘려서 더 많은 응답을 얻어내려는 노력이 필요하며, 거부 가구에 대해서는 효과적인 응답전환전략을 세우거나 차라리 무응답으로 인정하는 방안도 마련해 볼 필요가 있다.

〈표 1-7〉 무응답 성향 가구들의 마지막 회차에서의 응답 분포

	방문 1-3회차	4회차	빈도		그룹 내 (%)	전체 (%)
부재그룹	부재-부재-부재	응답	13,164	6,869	52.18	47.56
		부재		6,295	47.82	43.58
거부그룹	거부-거부-거부	응답	1,280	145	11.33	1.00
		거부		1,135	88.67	7.86
합계			14,444			100.00

〈표 1-8〉 응답성향별 고용현황

	주택	아파트	기타	총합
①-1 3회부재한 무응답 가구	3,984	4,791	1,361	10,136
(%)	39.31	47.27	13.43	
①-2 2회 이상 부재하고, 3 또는 4회차에 응답한 가구	12,797	10,014	2,251	25,062
(%)	51.06	39.96	8.98	
실업자수	356	314	79	749
실업률	2.28	2.40	2.90	2.39
고용률	61.41	58.65	59.37	60.05
비경률	37.15	39.91	38.86	38.48
②-1 3회 거부한 무응답 가구	910	1,499	278	2,677
(%)	33.99	55.62	10.38	
②-2 2회 이상 거부하고, 3 또는 4회차에 응답한 가구	285	353	84	722
(%)	39.47	48.89	11.63	
실업자수	2	6	9	17
실업률	5.99	1.38	8.41	1.94
고용률	59.71	56.02	50.78	56.71
비경률	39.93	43.19	44.56	42.17
③-1 원표본 거부 가구(B)	1,607	3,007	623	5,237
(%)	30.69	57.42	11.90	
③-2 거부대체가구(A)	1,713	3,016	510	5,239
(%)	32.70	57.57	9.73	
실업자수	69	122	30	221
실업률	3.66	3.22	4.82	3.51
고용률	52.77	53.11	55.21	53.21
비경률	45.22	45.11	41.99	44.85

◎ 고용률 = (취업자÷15세 이상 인구*)×100 (*15세 이상 인구: 고용조사대상)
실업률 = (실업자÷경활인구)×100, 비경률 = (비경인구수÷15세 이상 인구)×100

<표 1-8>은 응답성향별 분포와 고용현황을 나타낸 것이다. <표 1-8>에서 ①은 3회 부재한 후 마지막 회차에서 무응답한 가구와 2회 이상 부재 후 4회차에서 응답한 가구의 분포와 그 가구들의 고용상황을 나타낸 것이다. 부재 가구는 아파트가 가장 많고, 부재 후 응답 가구 중에서는 주택이 51.06%로 가장 높고, 아파트는 39.96%를 차지하였다. 이 중 실업률은 기타 가구가 가장 높고, 아파트, 주택 순으로 나타났다. 고용률은 주택이 가장 높고, 기타, 아파트 순으로 나타났다. 즉, 부재 후 응답 가구에서는 아파트에서 실업률은 가장 높고, 고용률은 가장 낮은 것으로 나타났다.

<표 1-8>에서 ②는 거부 가구들에 대한 응답성향에 대한 정보를 나타낸다. 3회 거부한 후 응답한 가구는 아파트에서 55.62%로 가장 높고, 2회 이상 거부 후 응답 가구는 아

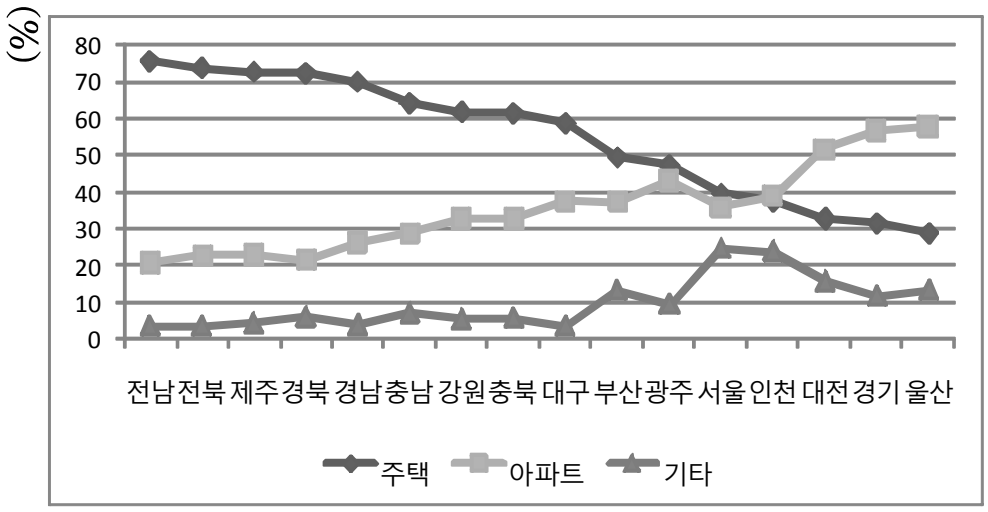
파트에서 48.89%로 가장 높다. 이러한 현상은 부재와는 다른 양상으로서, 부재 가구에서는 2회 이상 부재 후 응답한 가구의 비율이 주택에서 월등히 높은 반면, 2회 이상 거부 후 응답 가구는 아파트에서 높은 것을 알 수 있다. 2회 이상 거부 후 응답 가구 중 실업률이 기타에서 가장 높고, 다음이 주택, 아파트 순이고, 고용률은 주택에서 가장 높고, 다음으로 아파트, 기타 순으로 나타났다. 마지막으로 <표 1-8>의 ③은 추출당시 표본으로부터 발생한 거부 가구와 이를 대체하기 위해 사용한 거부대체 가구에 대한 결과이다. ③-1은 원표본에서의 거부 가구 중 57.42%는 아파트에서 발생하였다. 이들 대체 가구 중에서 실업률은 기타에서 가장 높고 다음으로 주택과 아파트인 것으로 나타났지만, 주택과 아파트는 비율적 측면에서 실업률이 큰 차이를 보이지는 않는다.

<표 1-8>로 부터 알 수 있는 것은 부재로 인한 무응답 가구는 연속적으로 부재한 후에 응답한 가구와 그 성향과 비슷하고, 반대로 거부로 인한 무응답 가구는 연속적으로 거부한 후에 응답한 가구와 그 성향이 비슷하다고 할 때, 대체로 부재성 가구는 주택에서 실업률이 낮고, 고용률이 높은 반면, 거부성 가구는 주택에서는 실업률과 고용률이 높다는 것이다. 한편, 거부 가구에 대한 대체 가구는 대체로 거부성 가구와 유사하게 주택에서 실업률은 높으나 고용율이 상대적이 낮은 경향이 있으며 그 정도에 있어서도 차이가 있다. 즉, 거부 대체 가구의 성향이 원 표본의 거부성 가구와는 고용 특성이 대체로 다를 것으로 예상할 수 있다.

물론 부재 가구에 대해 이를 대체하기 위해 사용된 가구의 정보는 파악하지 않았지만, 우선 무응답 가구의 성향이 대체가구의 성향과 다를 것이라는 것은 틀림없는 사실이다. 따라서 무응답 가구에 대한 보완책으로 무응답 가구를 대체할 경우, 신중한 판단이 뒤따라야 할 것이며 많은 경제적・시간적 노력이 필요하게 된다. 오히려 무응답 가중치 조정 방법을 통해 이러한 무응답 편향을 줄임으로써 통계의 신뢰성을 높이는 것이 조사의 효율성 측면에서 더 타당할 것이다.

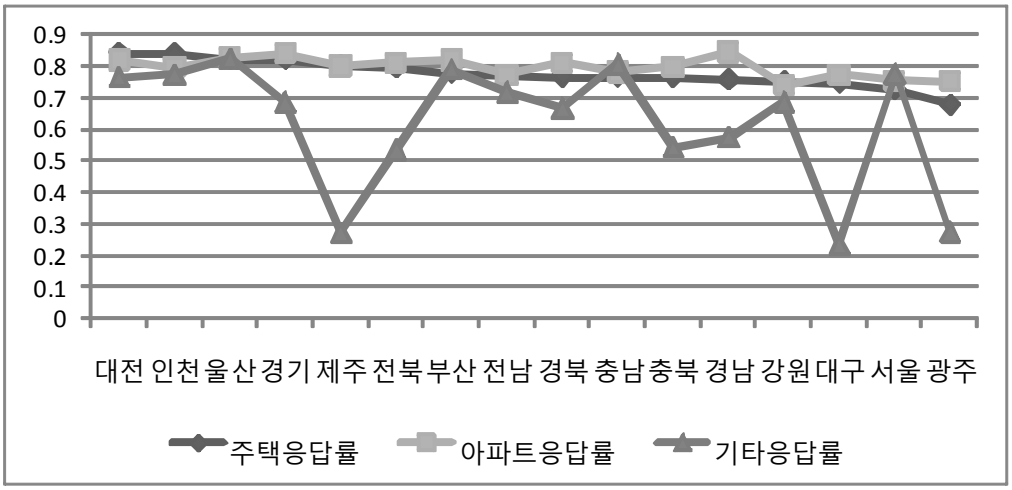
2) 지역별, 주택유형별, 응답 현황

이제 행정구역단위별 또는 주택유형별 응답 현황을 살펴보자. <표 1-5>는 16개 시도별 주택유형별 분포를 나타낸다. 그림에서 x축은 %, y축은 시도를 나타내고 주택비율이 큰 순으로 정리한 것이다. <표 1-5>에서 보면 대전, 경기, 울산을 제외한 16개 시도는 대체로 주택이 차지하는 비율이 높다. 특히 경기도를 제외한 모든 도 단위 지역에서 주택비율이 현저하게 높다.



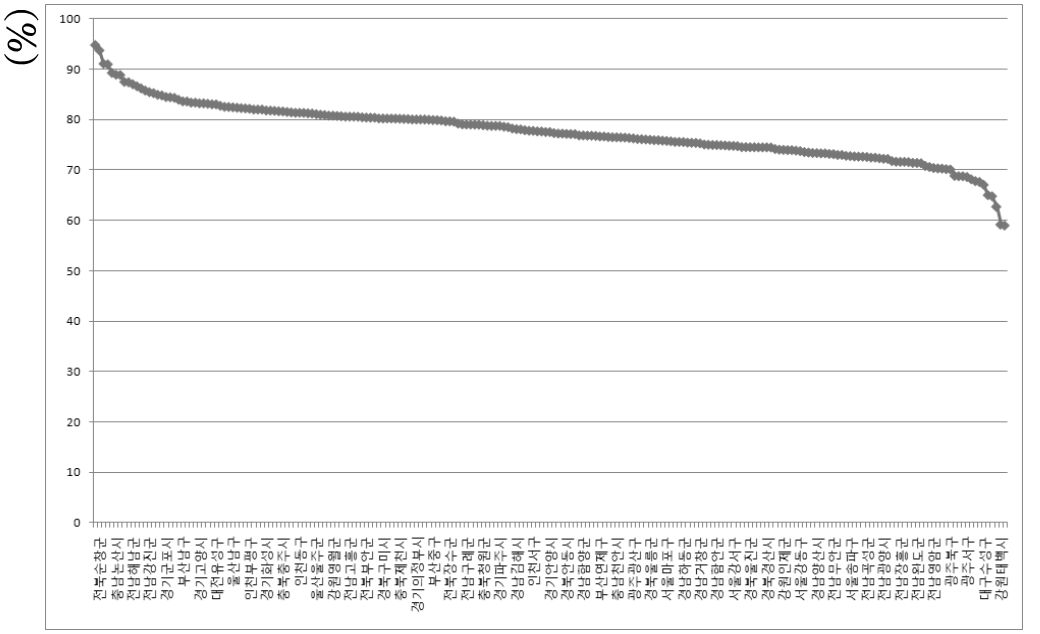
[그림 1-5] 16개 시도별 주택유형별 분포

[그림 1-6]은 16개 시도별 주택유형별 응답률을 나타낸 것이다. 전체 시도에 걸쳐 아파트 응답률이 높고, 그 다음으로 주택응답률이 높은 것으로 나타났다. 좀 더 자세하게는 특정 시도에서는 주택유형별 응답률 경향이 다른 지역들과 약간씩 다른 것을 알 수 있다. 따라서 시도에 따라 주택유형별 응답률 상황을 고려하여 이를 층 구분 변수로 사용할 필요가 있을 것으로 판단된다.



[그림 1-6] 16개 시도별 주택유형별 응답률

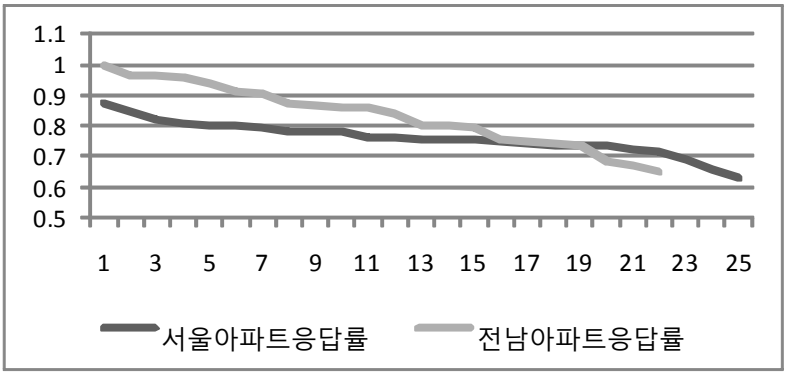
[그림 1-7]은 시군구별 응답률을 나타낸다. 시군구별로 응답률 차이가 분명한 것을 알 수 있다. 따라서 시군구별 응답률을 무응답 층 변수로 고려해 볼 수 있다.

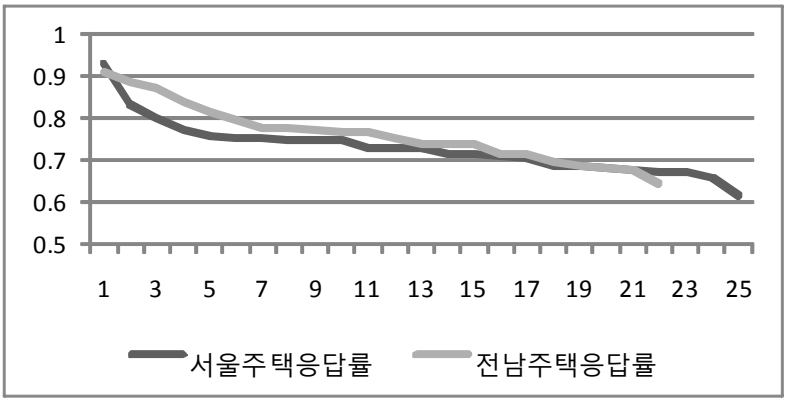


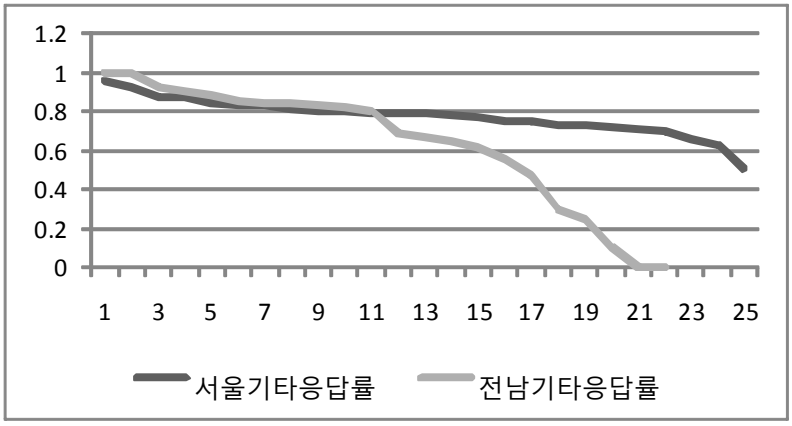
[그림 1-7] 시군구별 응답률

다음으로 시군구별, 주택유형을 고려한 응답률을 비교한 결과는 [그림 1-8]과 같다. 지면한계 상 모든 지역에 대해 나타내기는 어렵지만, 가장 대도시라 할 수 있는 서울과 농어촌 지역을 많이 포함하는 지역 중 전남 지역에 대해서만 응답률을 비교해 보았다. [그림 1-8]의 그림은 주택, 아파트, 기타 유형에 따른 서울과 전남 지역의 응답률을 각각 나타낸 것이다. 대체로 서울 지역이 전남 지역보다 주택유형에 상관없이 응답률이 낮고 그 격차는 아파트와 기타에서 크게 나타났다. 각 시군구별 주택유형별로는 본문에서 나타내지는 않았지만, 응답률에 차이가 크게 나타나는 것을 알 수 있었다. 따라서 시군구별 주택유형별 응답률을 고려하여 무응답 조정을 함으로써 더 큰 무응답 조정 효과를 얻을 수 있을 것으로 기대한다.

이러한 결과를 바탕으로 본 연구는 시도, 시군구, 주택유형을 무응답 층 변수를 사용하여 무응답 가중치 조정을 시도하기로 한다. 그리고 응답성향변수도 성향점수모형에서 포함하여 무응답 가중치 작성에 활용하고자 한다.







[그림 1-8] 주택유형에 따른 서울과 전남의 시군구별 응답률

제4절 무응답 가중치 조정 결과

1. 무응답 가중치 적용 시나리오

2010년 9월 시군구 고용조사 통계는 사후층화에 따른 사후층화 추정량이 사용된 것이다. 본 연구는 무응답 가중치를 작성하는 것이 주된 목적이며 무응답 가중치 적용 이전에 설계가중치를 기본적으로 사용하였다. 설계가중치는 이미 작성된 결과를 그대로 반영하고 본 연구는 이후 무응답 가중치 적용에 대한 방법을 제시하는데 목표를 두었다.

고용조사는 앞에서 설명된 바와 같이 경활표본과 별도표본으로 구성되어 있기 때문에 이 두 표본을 어떻게 통합할 것인지가 중요한 사항이다. 이와 같은 문제는 최근 표본조사 추정치 작성과정에서 중요하게 다루어지는 문제이기도 하다. 특히 표본을 유지하기가 어려운 패널조사에서는 매우 중요한 문제로 인식되고 있고 이에 관한 많은 연구가 진행되고 있다. 본 연구는 하나의 통합된 고용조사 표본을 구성하기 위해 우선 경활조사와 별도조사가 독립적으로 실시된 조사로 간주하였다. 이는 2005년 센서스 전체 자료를 모집단으로 하여 서로 겹치는 부분이 없이 부모집단을 형성하여 이로부터 경활표본과 별도표본을 추출했다는 점을 고려할 때, 이러한 가정은 큰 무리가 없을 것이다.

고용조사의 무응답 가중치는 경활표본과 별도표본 각각에 대해 무응답 가중치를 따로 계산한 후 이것을 하나로 통합하는 과정을 거치고자 한다. 이 두 표본을 하나로 통합할 때 사용할 통합계수는 고용조사 전체표본 중 각 표본이 차지하는 비중을 사용할 수 있다. 이처럼 표본비중을 사용하는 경우는 두 조사의 분산이 동일하다고 가정하는 경우이다. 그런데 만약 두 조사의 분산이 다르다고 한다면 두 조사의 정도가 서로 다르다고 할 수 있기 때문에 정도가 좋은 조사에 더 많은 가중치를 반영하는 방법도 고려할 수 있다. 본 연구에서는 두 가지 방법을 모두 사용하여 통합계수를 작성한다.

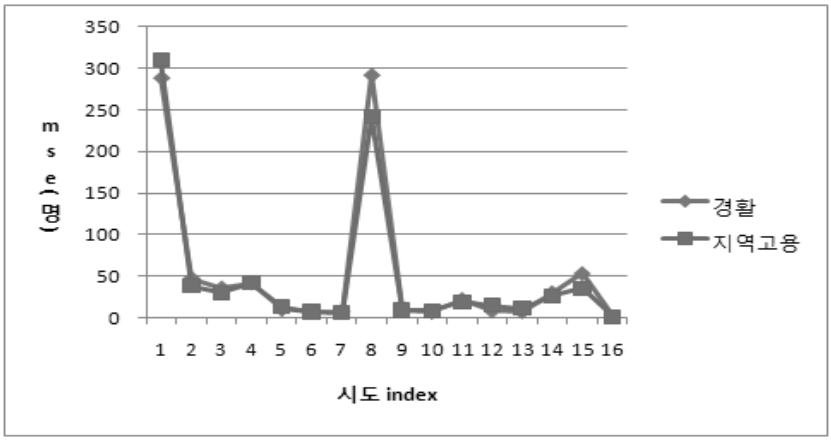
1) 표본비중을 고려한 경활표본과 별도표본 통합계수

j_1 = 경활표본 가구수/지역별고용조사 총 표본가구수, 경활자료에 사용
j_2 = 별도표본 가구수/지역별고용조사 총 표본가구수, 지역별고용조사에 사용

2) 두 조사의 분산비중을 고려한 경우

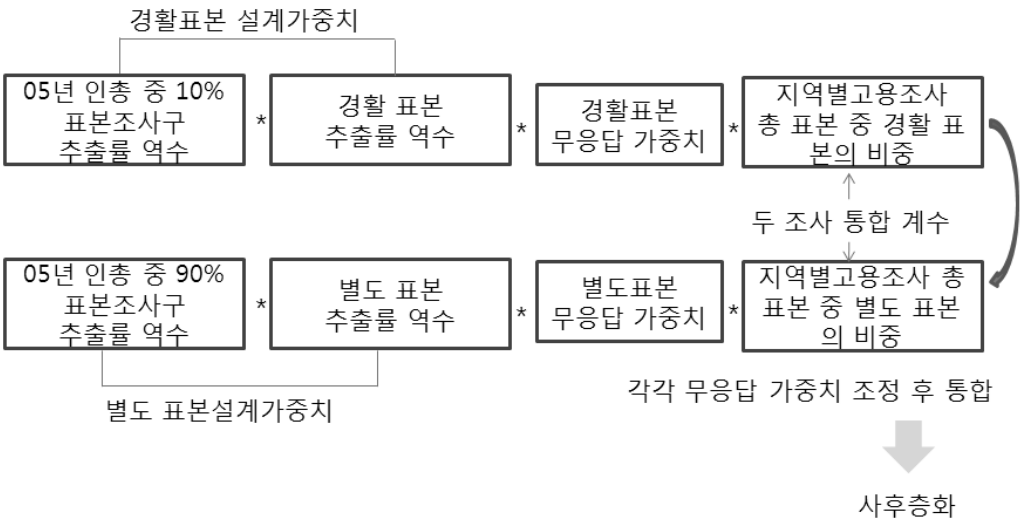
두 조사의 분산을 고려하여 '경활조사분산/지역별고용조사분산' 의 비율만큼을 경활조사에 반영하는 방법을 사용하였다.

경활조사와 고용조사의 평균제곱오차(MSE)를 비교한 결과, 16개 시도별 두 조사의 분산은 고용조사가 경활조사에 비해 대체로 약간 작거나 유사한 것으로 나타났다([그림 1-9]). 분산은 표본크기에 영향을 받기 때문에 표본이 매우 큰 지역별고용조사의 분산이 원칙적으로 작은 것이 당연하다. 따라서 엄밀한 의미에서 두 조사의 분산을 고려하고자 한다면, 지역별고용조사의 표본이 경활표본 크기와 동일하도록 설정하고, 이를 여러번 반복한 다음 반복적으로 계산한 분산의 평균값을 지역별고용조사의 분산으로 간주하는 것이 두 조사의 분산을 비교하는 의미가 있다고 판단되었다. 따라서 본 연구에서는 표본크기가 작은 경활조사의 표본크기에 대해 지역별고용조사에서 100번 반복 추출하여 각각으로부터 계산된 분산의 평균값을 사용하였다.



[그림 1-9] 지역별고용조사의 모의실험 분산과 경활조사 분산의 비교

위의 두 방법으로 계산된 계수를 두 조사를 통합하는 계수로 사용하여 두 조사를 하나의 조사로 통합할 수 있다. 정리하면 가중치 적용 시나리오를 그림으로 나타내면 다음 [그림 1-10]과 같다. 조사 당시의 인구분포가 이사로 인한 전입 또는 전출, 연동표본의 새로운 추가로 인해 표본설계당시와는 약간 달라질 수 있다. 이러한 현상은 사후층화 과정을 통해 보정함으로써 통계의 신뢰성을 높일 수 있다. 본 연구는 무응답 가중치 적용 전·후의 결과를 비교할 때, 두 표본을 통합한 후 성별, 연령별 인구분포를 고려하여 사후층화한 결과를 사용하였다.



[그림 1-10] 무응답 가중치 적용 시나리오

2. 무응답 가중치 작성 및 적용 결과

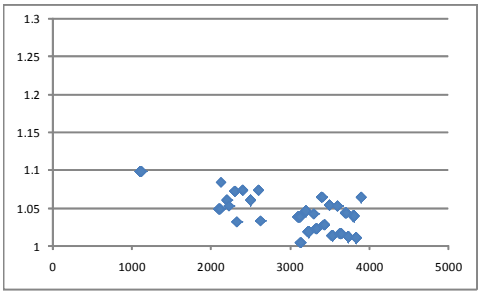
무응답 가중치는 경활표본과 별도표본에 대해서 각각 같은 방법과 다른 방법을 통해 작성하기로 한다. 두 표본에 사용될 수 있는 무응답 조정 변수가 다를 수 있기 때문이다. 별도표본에서 무응답은 거부 가구로 정의되고, 경활표본에서 무응답은 거부와 부재 가구 모두로 정의되었다. 무응답 가중치는 무응답 층 조정 방법과 응답성향 조정 방법을 이용한다.

가. 별도표본 무응답 가중치

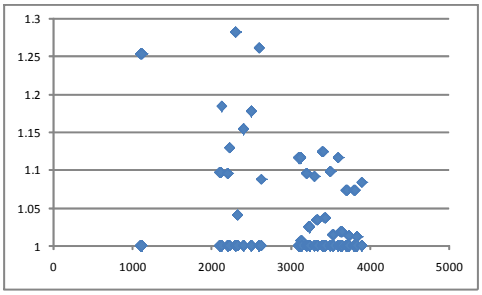
1) 무응답 층 방법

3절에서 분석한 기초분석결과를 바탕으로 무응답 층 변수는 지역변수와 주택유형 변수를 사용한다. 지역구분 변수는 우리나라 행정구역단위로서 16개 시도와 230여개 시군구 구분을 사용한다. 16개 시도는 다시 시구·군(2개 범주)을 사용할 수 있다. 주택유형은 단독주택, 아파트, 기타 등 3개 범주를 사용한다. 무응답 층으로는 다음의 4가지 방법을 시도하였다. 이는 지역층의 세분화 그리고 주택유형 고려 여부에 따라 무응답 가중치의 변화를 살피기 위한 것이다. 이들 각각에 대해서 각 층별로 응답률의 역수인 무응답 가중치의 산점도와 기초통계량을 통해 특성을 살펴본다. 그리고 이 무응답 가중치를 설계가중치에 적용한 결과에 대해서도 살펴본다.

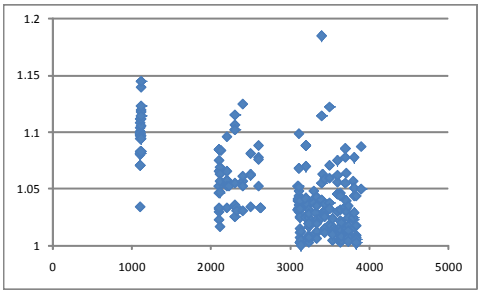
① 16개 시도 내 시구·군 28개 층
② 16개 시도 내 시구·군(28개 층)×주택유형(3개 층)=총 83개 층
③ 시군구 232개 층
④ 232개 시군구 층×주택유형(3개 층)=656개 층



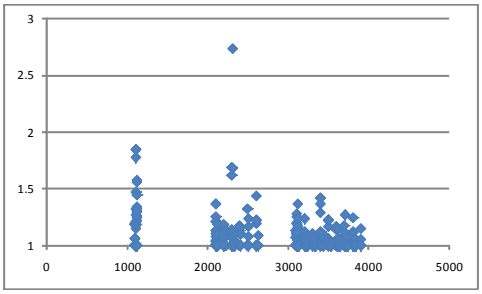
① 16개시도×시구/군=28개 층



② 16개시도×시구/군×주택유형=83개 층



③ 232개시도×시구/군=232개 층



④ 232개시도×시구/군×주택유형=656개 층

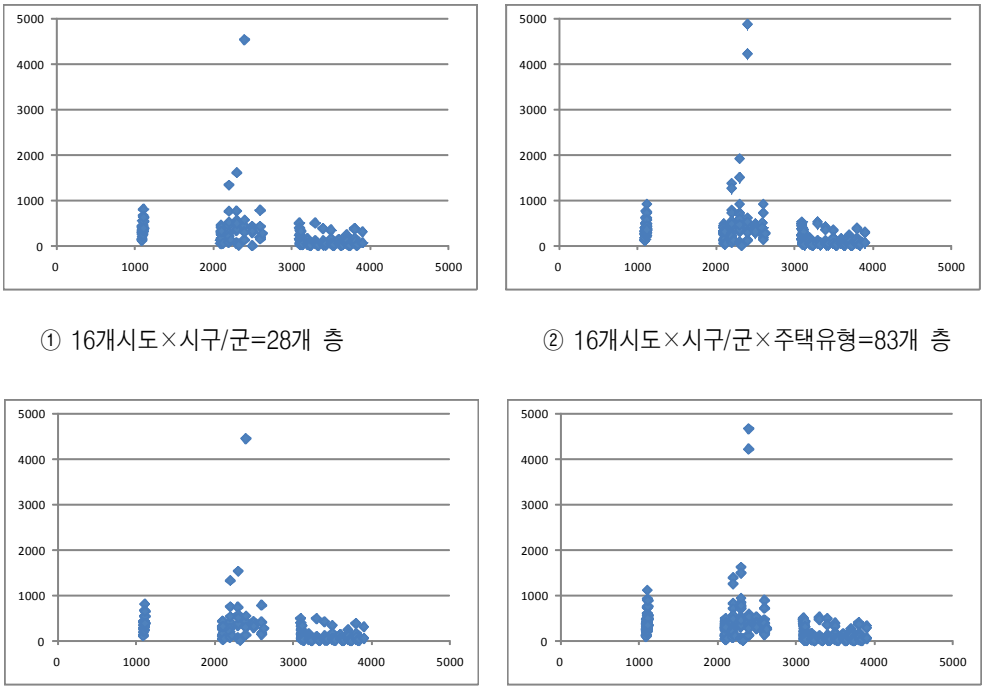
[그림 1-11] 방법별 무응답 가중치의 산점도

〈표 1-9〉 무응답 층화 방법별 무응답 가중치의 기초통계량

	응답률역수		응답률역수	
	①	②	③	④
최대값	1.098	1.282	1.184	2.736
3사분위수	1.060	1.034	1.062	1.027
중위수	1.039	1.000	1.034	1.000
평균	1.043	1.035	1.044	1.043
1사분위수	1.018	1.000	1.016	1.000
최소값	1.004	1.000	1.000	1.000

위의 [그림 1-11]로부터 지역을 시도만 구분했을 때(①)에 비해 232개 층으로 구분했을 때(②)의 무응답 가중치가 상대적으로 크고 변동성이 있다는 것을 알 수 있다. 게다가 각 지역층에 주택유형을 고려하게 되면 무응답 가중치는 더 커지게 되고, 특히 시도층에서 변동성 또한 상당히 커지는 것을 알 수 있다. 이와 같은 현상은 지역층에 따라 응답률이 다르고, 주택유형에 따라 응답률이 상당히 다르다는 사실에서 비롯된 것으로 볼 수 있다. 한편 무응답 가중치가 상대적으로 그렇게 큰 경우는 존재하지 않는 것 같다. 이와 같은 사실로 비추어 볼 때 무응답 층을 가급적 세분화할수록 편향을 더 섬세하게 조정할 수 있을 것으로 기대할 수 있다.

여기에 설계가중치를 적용한 결과는 다음 [그림 1-12]와 같다. 전체적으로 볼 때 4가지 방법 모두에서 설계가중치를 적용한 결과는 유사해 보인다. 이들 방법에 대한 특성은 최종 추정치를 구한 이후에 상대편향이나 MSE를 통해 비교해 볼 수 있을 것이다.



① 16개시도×시구/군=28개 층 ② 16개시도×시구/군×주택유형=83개 층

③ 232개시도×시구/군=232개 층 ④ 16개시도×시구/군×주택유형=656개 층

[그림 1-12] 무응답 층화 방법별 무응답 가중치×설계가중치의 산점도

〈표 1-10〉 무응답 층화 방법별 무응답 가중치×설계가중치의 기초통계량

	응답률역수×설계가중치		응답률역수×설계가중치	
	①	②	③	④
최대값	4528.682	4866.816	4439.300	4661.280
3사분위수	288.811	270.130	279.280	271.924
중위수	64.056	63.421	60.850	62.443
평균	188.299	178.324	185.148	181.317
1사분위수	29.512	29.681	29.195	29.854
최소값	0.000	5.890	0.000	5.890

2) 응답성향점수 방법

Cobben(2008)은 최종 응답 여부와 조사결과 값에 대한 모형을 응답선택모형(response selection model)으로 설명하였다. 응답선택모형은 두 개의 방정식으로 구성된다. 응답성향모형을 구축하기 위해 이용 가능한 보조변수를 이용하였다. 응답 여부를 나타내는 변수는 응답을 했으면 $R=1$, 그렇지 않으면 $R=0$의 값을 갖는다. 반응변수와 보조변수와의 연관성 정도는 χ^2 통계량을 이용한 유의성 검정을 통해 확인하고, 이 결과 최종 모형은 다음과 같다. 각 모형에서 첨자는 각 변수에 대한 범주의 개수를 나타낸다.

$$주택유형_3 + 방문횟수_4 + 지역별응답률_{230}$$

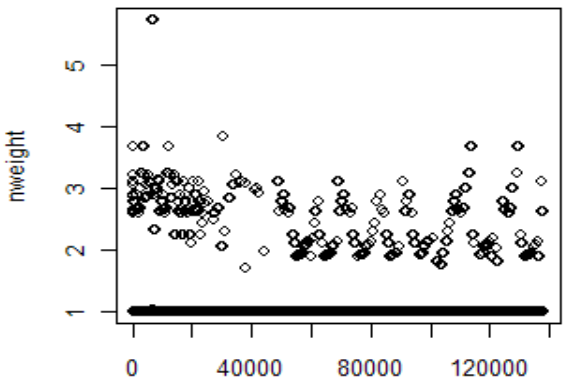
각각 모형에 대해 응답성향점수를 계산한 후 응답성향점수의 역수와 층화방법을 이용하여 성향점수를 이용한 무응답 가중치를 구하고, 이를 설계가중치에 곱해서 최종 별도표본에 대한 무응답 가중치 조정을 시도한다. 성향가중치는 가구의 특성을 반영하는 정보를 활용할 수 있는 경우 유용한 방법이 될 수 있다.

〈표 1-11〉 성향점수 역수가중치 계산 결과 (가구별)

최소값	1사분위수	중위수	평균	3사분위수	최대값
1.000	1.000	1.000	1.038	1.000	5.719

<표 1-11>을 보면 성향점수 역수에 의해 계산된 가중치는 최소값 1.000, 최대값 5.719, 평균 1.038의 분포를 갖는다. 무응답 층에서와 유사하게 유사 층별로 거의 동일한

성향가중치 값을 갖는다는 것을 알 수 있고, 대체로 안정된 가중치이긴 하나 무응답 층화방법에 의한 가중치에 비해 약간 큰 값이 존재하기도 한다([그림 1-13]). <표 1-11>에서 가로축은 지역에 해당되는 레코드 번호, 세로축은 가중치값을 의미한다.

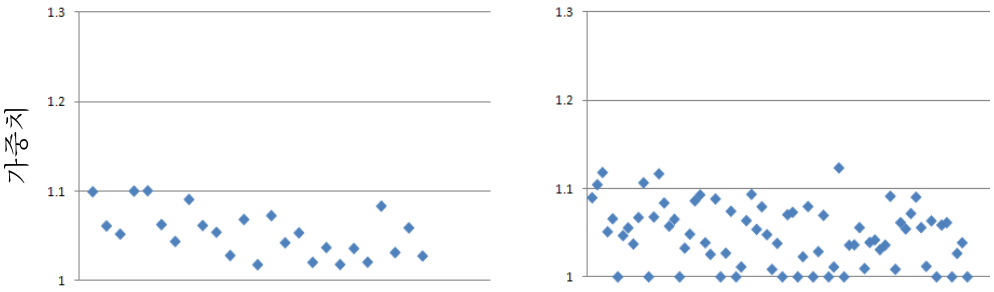


[그림 1-13] 성향점수 역수 방법에 의한 가중치 산점도

따라서 전체적으로 5가지 방법에 의해 작성된 무응답 가중치는 성향점수 역수방법에서 약간 큰 값이 존재하지만 전체적으로 크게 두드러진 값은 없는 것으로 판단되며, 이 결과를 전체자료에 적용 후 최종 추정치를 구함으로써 각 방법의 특성을 논할 수 있을 것으로 보인다.

나. 경활표본 무응답 가중치

경활표본의 경우 별도표본조사와 유사하게 무응답 층을 구성하였다. 이때 경활조사는 시도 기반 조사라는 점을 고려하여 무응답 층은 표본설계 층과 동일하게 25개 지역층을 사용하고, 여기에 3개 주택유형을 고려한 층화 방법을 사용하였다. [그림 1-14]에서 보면 그림 주택유형을 고려하지 않은 그림(a)에 주택유형을 층 변수로 고려한 (b)의 경우가 가중치값이 약간 큰 경우가 있는 것으로 보인다. 이는 층을 세분화 할수록 층의 무응답 속성을 더 잘 이해할 수 있다는 장점도 있지만, 그 층에서의 조사대상자수가 매우 적다면 그에 따라 무응답률이 매우 크게 작용할 수 있게 된다. 따라서 층을 세분화할 때는 이러한 점을 고려하여 층의 개수를 정할 필요가 있다. 다행스럽게도 [그림 1-14]는 각 층화방법별 가중치의 분포가 1 근처에서 매우 안정적인 것을 알 수 있다. 이들 가중치에 대한 기초통계량값은 아래 산점도를 통해 충분히 설명가능하기 때문에 여기서는 별도로 제시하지 않기로 한다.



(a) 25개 층　　(b) 25개 층×주택유형

[그림 1-14] 각 층화방법별 가중치 산점도

3. 각 방법별 가중치 조정결과 비교

가. 비교 시나리오

본 연구에서 고려한 다양한 방법에 대해서 각 방법별 가중치 조정 효과는 여러 개의 시나리오를 작성하여 비교한다. 우선 무응답 가중치 작성을 위해 경활조사와 별도조사에 대한 각 층화방법을 정리하면 다음 <표 1-12>, <표 1-13>과 같다. 이때 두 조사의 각 무응답 층을 조합할 때 주택유형 여부를 고려하였다. 즉, 주택유형 층을 고려한 경우와 그렇지 않은 경우로 나누어 주택유형 변수의 영향을 살펴보고자 하였다.

<표 1-12>에서 표본은 별도표본과 경활표본으로 구분되었고, 각 표본 내에서 층 개수는 별도표본의 경우 4가지 방법, 경활표본은 2가지 방법으로 사용하였다. 참고로 별도표본의 층 개수를 보면 ①의 경우 층 개수가 28개인데 엄밀하게 말하면 16개 시도*2개 시구/군으로 구분했을 때 32개가 되어야 할 것이다. 그렇지만 이렇게 층을 나누었을 때, 무응답이 존재하지 않은 층도 존재한다는 것이다. 따라서 전체 층의 개수가 32개가 되지 않고 28개가 된 것이다. 나머지 경우에 대해서도 무응답 층 구성 방법에 의한 개수만큼 층의 개수가 되지 않은 경우는 위와 동일한 개념으로 이해할 수 있고, 층을 세분화할 수록 이러한 가능성은 더 커질 수 있을 것이다.

〈표 1-12〉 각 표본별 무응답 층화 방법

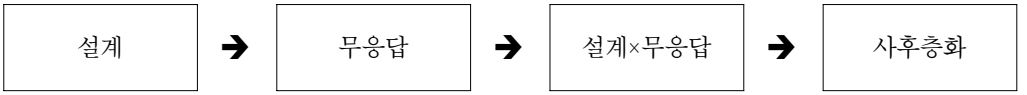
표본구분	층 개수	무응답 층 구성
별도표본	① 28	16개 시도×시구/군
	② 83	16개 시도×시구/군×주택유형(주택/아파트/기타)
	③ 232	232개 시군구
	④ 656	232개 시군구×주택유형(주택/아파트/기타)
경활표본	ⓐ 25	16개 시도×동/읍면부
	ⓑ 83	16개 시도×동/읍면부×주택유형(주택/아파트/기타)

※ 시도 내 시구/군이 모두 존재하지 않거나, 주택유형에서 누락되는 유형이 있는 경우가 있어 층 개수와 무응답 층 산식결과와 층 개수가 일치하지 않음

〈표 1-13〉 각 층화방법에 따른 통합 시나리오

시나리오	별도표본	경활표본
가	① 16개 시도×시구/군	ⓐ 16개 시도×동/읍면부
나	② 16개 시도×시구/군×주택유형	ⓑ 16개 시도×동/읍면부×주택유형
다	③ 232개 시군구	ⓐ 16개 시도×동/읍면부
라	④ 232개 시군구×주택유형	ⓑ 16개 시도×동/읍면부×주택유형
마	⑤ 성향	ⓑ 16개 시도×동/읍면부×주택유형

〈표 1-14〉 가중치 적용 단계



<표 1-14>는 본 연구의 단계별 적용가중치를 요약한 것이다. 우선 두 조사에 대해 각각의 설계가중치를 적용하고, 여기에 각각 무응답 가중치를 적용한 후, 그 다음과정에서 두 조사의 통합계수를 적용한 다음, 마지막으로 사후층화에 의해 보정을 거치기로 하였다.

나. 각 방법별 비교

지금부터 각 시나리오에 대한 가중치 조정결과를 살펴보자. 비교를 통해 파악하고자 하는 내용은 다음과 같다. 각 방법별 비교는 각 조정에 따른 평균제곱오차(MSE, 또는 CV)를 사용하였다. 기존방법과의 비교는 모든 조정결과에 대해 사후층화를 하고 이 결과에 대해 상대적 편향과 MSE를 비교하였다. 편향과 MSE는 작을 수록 좋다. 본 절을 통해서 무응답 층화방법에 따른 효과, 가중치 방법에 따른 효과, 통합계수 사용에 따른 효과 등을 각각 살펴볼 수 있다. 우선 각 방법에 따른 가중치 특성을 살펴보기로 하자.

〈표 1-15〉 각 사후조정 후 가중치별 기초통계량

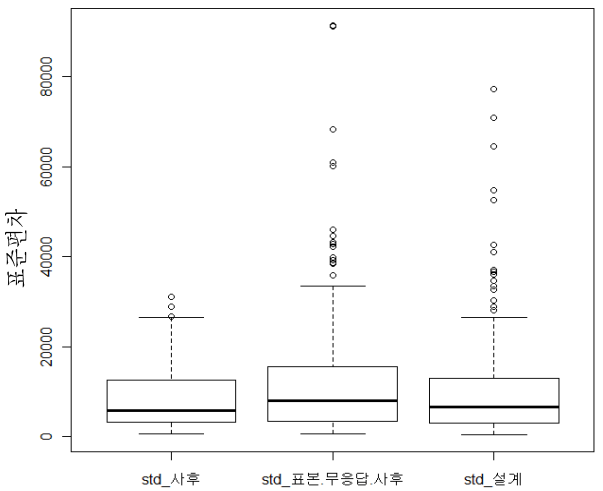
통계량	M2_A	M2_B가	M2_B나	M2_B다	M2_B라	M2_C가	M2_C나	M2_C다	M2_C라
최소값	6.02	0.83	0.83	0.84	0.84	6.02	5.94	6.02	6.00
최대값	3238.00	1011.08	1208.30	988.71	1278.70	3179.76	3316.08	3132.22	3200.33
평균	114.76	114.76	114.76	114.76	114.76	114.76	114.76	114.76	114.76
표준편차	120.86	120.32	120.68	120.48	121.69	120.56	121.04	120.52	121.64

통계량	M3_A	M3_B가	M3_B나	M3_B다	M3_B라	M3_C가	M3_C나	M3_C다	M3_C라
최소값	6.02	0.78	0.77	0.78	0.77	6.02	5.94	6.02	6.00
최대값	2679.02	1015.65	1213.59	993.27	1248.81	2631.80	2782.71	2593.22	2682.28
평균	114.76	114.76	114.76	114.76	114.76	114.76	114.76	114.76	114.76
표준편차	121.11	120.76	121.12	120.92	122.12	120.83	121.33	120.80	121.93

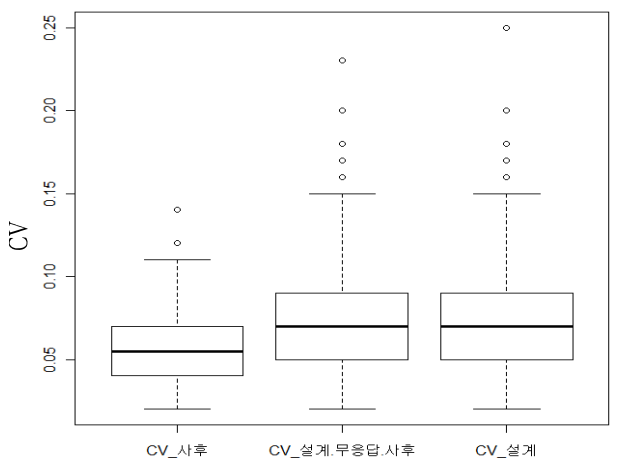
위의 <표 1-15>에서 보면 최소값과 최대값의 차이가 상당히 크다. 그런데 최소값을 갖는 사후조정 층을 보면, 울릉도의 60세 이상 연령그룹으로 현재 인구대비 표본으로 추출된 인구비가 크고 무응답 조정계수가 작기 때문에 사후조정 후 가중치가 그다지 크지 않다고 보인다. 가중치가 최대값인 지역은 몇 개의 광역시 지역(대구 달서, 인천 부평, 광주 북구)으로 실제로 기존 방법에 비해 추정치의 차이가 크게 발생하는 지역에 해당한다.

한편, 사후층화 조정만 한 경우(기존)와 설계가중치를 적용했을 때의 가중치들의 CV는 사후층화 조정만 하더라도 크게 변동이 발생하지는 않는다. 다만, 몇 개의 지역에서는 CV가 크게 증가하는 지역들도 있고, 이런 지역들에서는 사후층화 조정만 함으로써 표본조사구내의 가중치의 변동이 커지게 되어 결국은 추정치의 분산을 증가시키는 역할을 할 수 있다고 알려져 있다. 실제로 우리 자료에서의 가중치 변동성을 살펴보았다. 다음 그림은 사후가중치와 설계가중치 그리고 설계가중치×무응답가중치를 적용했을 때의

표준편차([그림 1-15])와 CV의 분포([그림 1-16])를 나타낸 것이다. [그림 1-15]에서 보면 연구에 사용된 자료의 경우 사후가중치와 설계가중치의 분산(std_사후, std_설계)은 거의 유사한 것으로 나타났다. 그런데 [그림 1-16]의 CV에 있어서는 사후가중치의 CV가 설계가중치의 CV에 비해 상대적으로 작은 것을 알 수 있다. 설계가중치×무응답가중치×사후조정한 경우 설계가중치와 크게 다르지 않은 것으로 나타났다.



[그림 1-15] 가중치들의 표준편차 분포



[그림 1-16] 가중치들의 변동성 분포

〈표와 그림에서 사용한 기호 정리〉

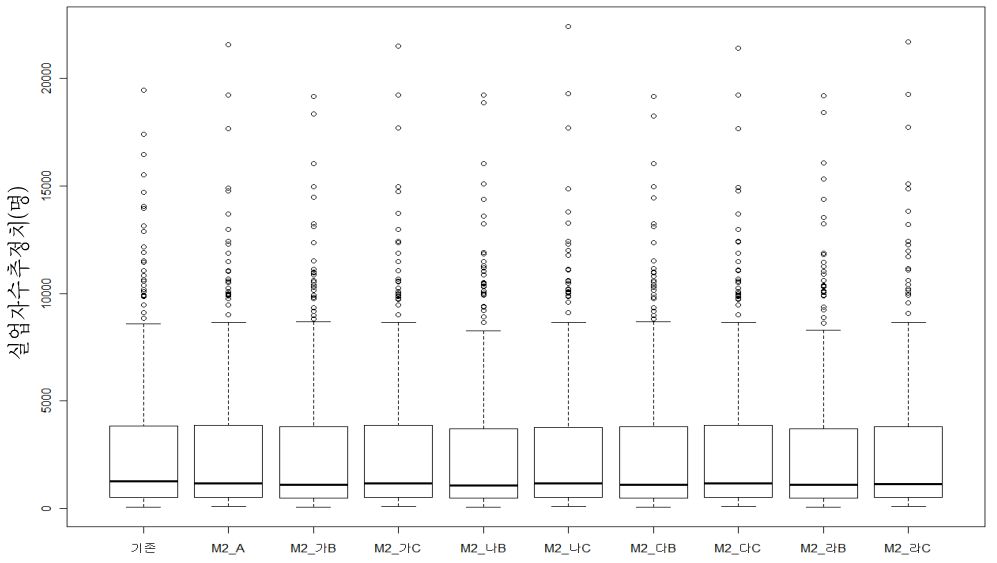
M2 : 통합계수를 표본크기만 고려
M3 : 통합계수를 표본크기와 분산비중을 고려
가 : 무응답 층 - 시도/시군 vs. 시도/동읍면
나 : 무응답 층 - 시군×주택유형 vs. 시도/동읍면×주택유형
다 : 무응답 층 - 시군구 vs. 시도/동읍면
라 : 무응답 층 - 시군구×주택유형 vs. 시도/동읍면×주택유형
A : 설계가중치
B : 무응답가중치
C : 설계+무응답가중치

본 연구가 다양한 무응답 층화 방법에 따른 여러 가지 시나리오를 사용하고 있기 때문에 비교결과를 정리하기다 다소 복잡하게 보일 수 있다. 따라서 위의 표는 독자들의 이해를 돕기 위해 결과 설명을 위해 사용된 기호를 정의한 것이다.

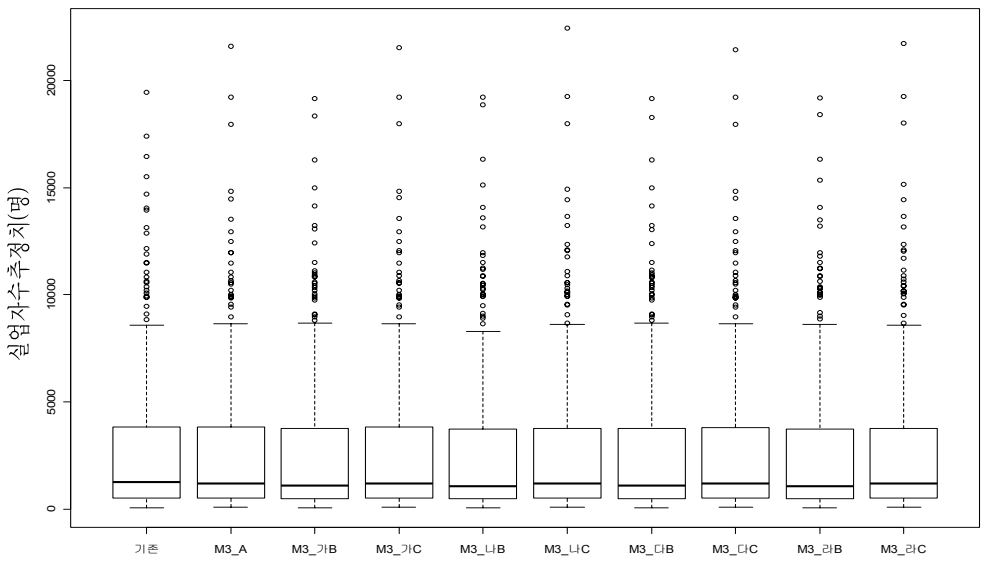
[그림 1-17]과 [그림 1-18]은 위 그림은 두 표본조사를 연결할 때, 통합계수로 표본크기만을 고려한 경우와 표본크기×분산비중을 고려한 경우에 대한 230여 시군구의 실업자 추정치의 분포를 각각 나타낸 것이다. 각 그림에 따르면 전체적으로 통합계수 사용방법에 관계없이 시군구의 추정치 분포는 거의 동일한 것을 알 수 있다 ([그림 1-17] VS. [그림 1-18] 비교). 각 그림의 방법 간에는 미세한 차이가 있기는 하지만, 추정치에 영향을 주는 정도는 아니라고 판단되었다. 따라서 각 가중치 적용방법에 상관없이 추정치들의 분포는 현재 공표되고 있는 방법(기존)과 크게 다르지 않을 것으로 보인다.

추정결과에 큰 차이가 없다면, 본 연구에서는 실무적용에 간단한 방법을 우선적으로 선택하기로 하였다. 따라서 표본크기 비중만으로 고려한 통합계수를 사용하기로 하였다. 더욱이, 분산비중까지 고려할 경우, 현재 각 조사의 분산계산방법이 경활표본의 크기와 동일하게 고용조사를 축소하여 반복적으로 계산한 결과이기 때문에 재현이 쉽지 않다는 점도 있다. 뿐만 아니라, 이러한 반복적인 분산 계산 방법이 반복횟수나 방법에 따라서도 약간씩 차이가 있을 수 있고 이 결과들이 최종 추정치에 영향을 줄 수도 있다고 판단되었기 때문에 간단하면서도 안정적인 방법으로서 표본크기를 두 조사의 통합계수로 선택한 이유라 할 수 있다.

전국 단위의 실업자 총계 추정치를 각 방법별로 보면 <표 1-16>과 같다.



[그림 1-17] 표본크기로 통합- 가중치 적용 후 사후조정 : 230개시군구의 실업자 추정치 분포



※ 가중치 적용 후 사후조정
- 230개시군구의 실업자 추정치 분포, x축은 실업자 추정치(명), y축은 방법을 나타냄

[그림 1-18] 표본크기×분산비중으로 통합; 230개시군구의 실업자 추정치 분포

〈표 1-16〉 각 방법에 따른 전국 실업자수 추정치 비교

기존	M2_A	M2_가B	M2_가C	M2_나B	M2_나C	M2_다B	M2_다C	M2_라B	M2_라C
785,993	780,327	769,249	780,782	773,817	785,194	769,033	780,479	774,011	784,934
기존	M3_A	M3_가B	M3_가C	M3_나B	M3_나C	M3_다B	M3_다C	M3_라B	M3_라C
785,993	777,892	767,909	778,276	772,450	782,658	767,690	777,975	772,668	782,434

<표 1-16>에서 전국 실업자 추정치를 이용하여 기존값과 비교하면 다음과 같다.

① 통합계수 사용여부 및 가중치 방법에 상관없이 모든 시나리오를 적용할 경우, 기존 방법에 비해 전국 실업자 총수 추정치가 줄어드는 효과가 있다.

② 무응답가중치(B)만 적용한 경우는 설계가중치(A 또는 C)를 적용한 경우에 비해 전국 실업자수 추정치의 크기가 작아진다.

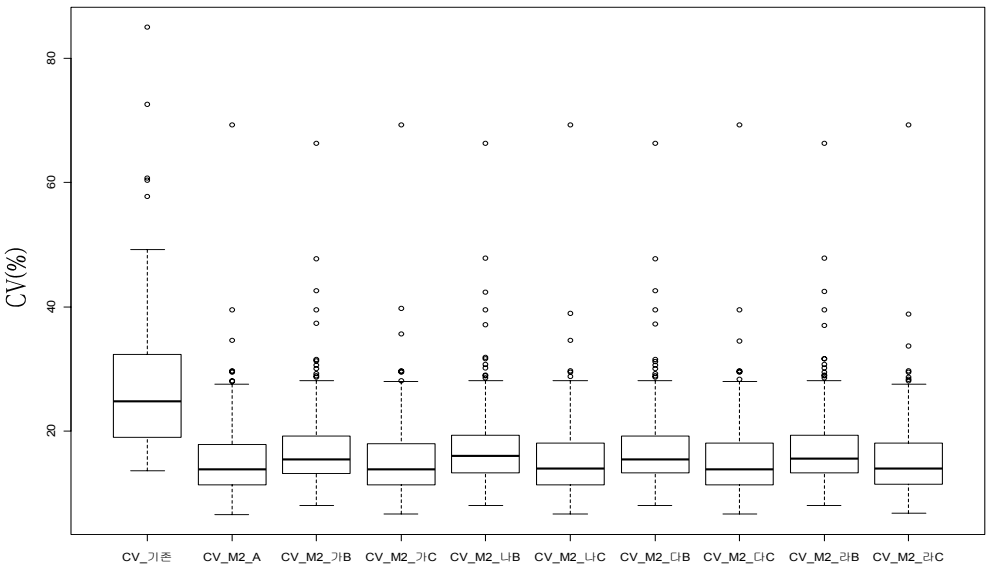
③ 설계가중치와 무응답가중치를 적용한 경우(C)는 설계가중치만 적용한 경우(A)에 비해 실업자수 추정치가 커진다.

④ 지역층×주택유형 층(나C 라C)을 고려한 경우는 지역층(가C 또는 다C)만 고려한 경우에 비해 전국 실업자수 추정치가 약간 커진다.

④ 지역층을 세분화 할수록 즉, 260개 시군구 지역층×주택유형(라C)이 25개 시도 층×주택유형(나C)을 고려한 경우에 비해 전국 실업자수가 약간 작아진다.

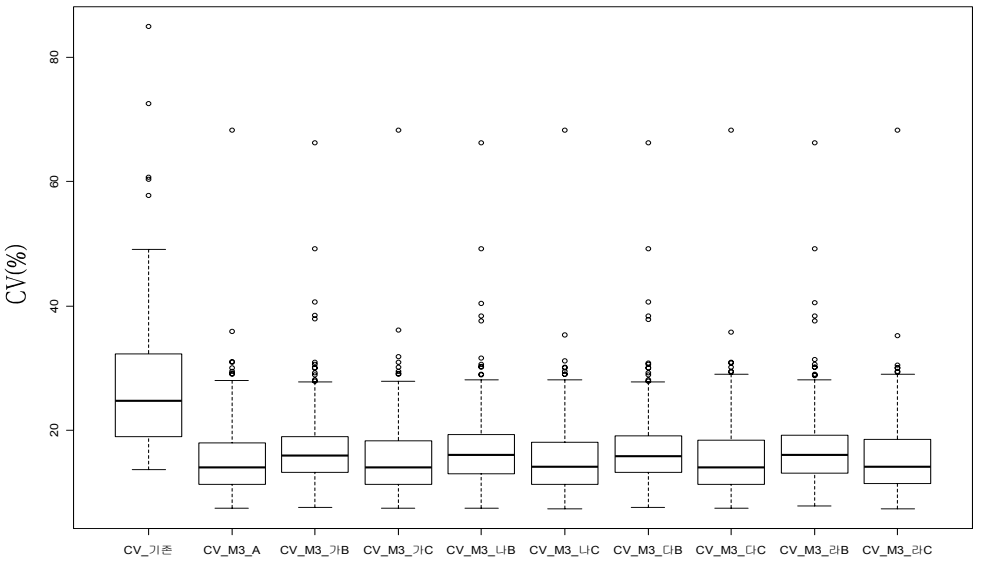
⑤ 통합계수를 표본크기만 고려한 경우는 표본크기(M2)와 분산비중(M3)을 고려한 경우에 비해 전국 실업자수 추정치가 크게 나타난다.

위의 결과에 따르면 통합계수는 표본크기×분산비중, 가중치방법은 설계가중치×무응답가중치, 무응답 층은 주택유형×230개 시군구 층을 고려할 때 좋은 추정치를 얻을 수 있을 것으로 보인다. 그렇지만 통합계수 여부에 따라서는 전국 실업자수의 차이가 매우 작고, 가중치의 분산도 크게 차이가 없다는 점에서 볼 때, 계산상의 편리한 점을 고려한 표본크기만으로 통합한 라C의 경우가 더 실용적일 수 있다.



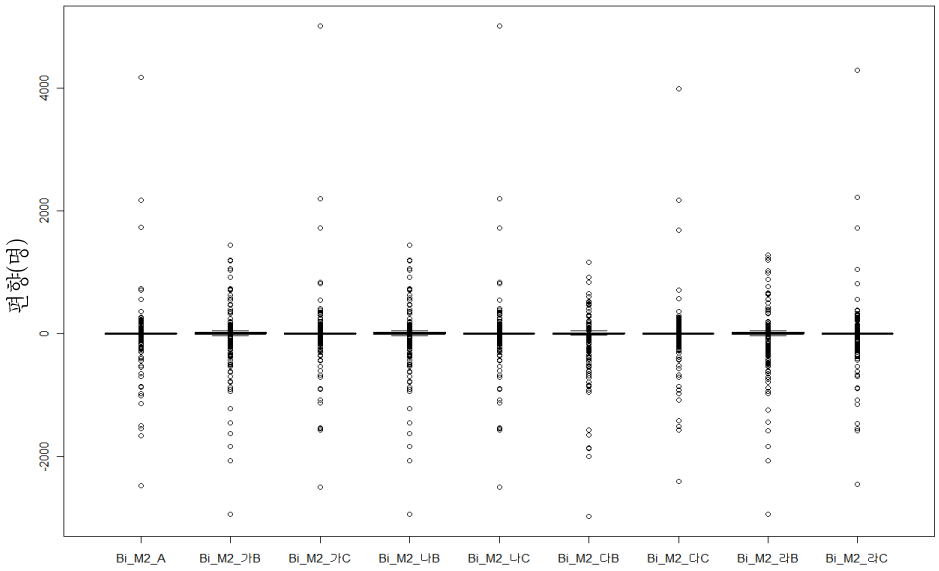
※ 가중치 적용 후 사후조정 - 230시군구 추정치의 CV 분포

[그림 1-19] 표본크기로 통합

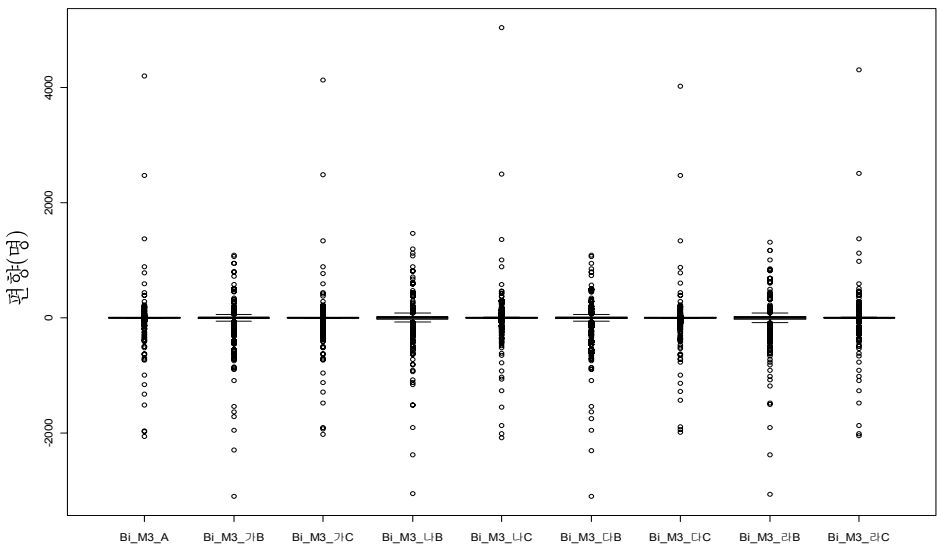


※ 가중치 적용 후 사후조정 - 230 시군구에 대한 추정치의 CV 분포

[그림 1-20] 표본크기+분산비중으로 통합



[그림 1-21] 표본크기로 통합 : 가중치 적용 후 사후조정- 230 시군구 추정치의 Bias 분포



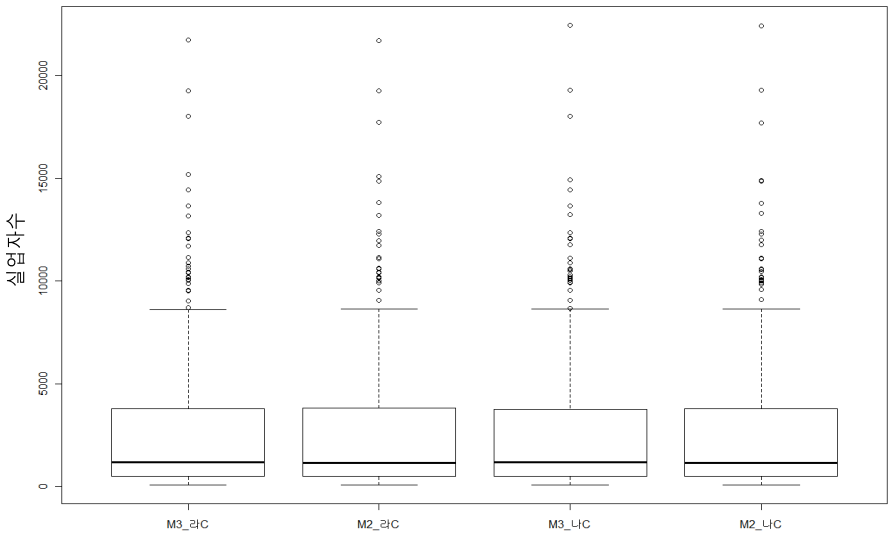
[그림 1-22] 표본크기×분산비중으로 통합, 가중치 적용 후 사후조정- 230 시군구 추정치의 Bias 분포

위 그림은 통합계수에 따른 각 추정치들의 CV와 편향을 각각 나타낸 것이다. 우선 [그림 1-19]와 [그림 1-20]을 보면 설계가중치와 무응답가중치를 적용하면 추정치들의 CV는 통합계수 방법 여부에 상관없이 기존의 방법에 비해 상당히 작아지는 경향을 보이는 것을 알 수 있다. 가중치 방법별로는 설계가중치×무응답가중치를 적용한 경우(M2_C, M3_C)의 CV가 다른 방법들에 비해 상대적으로 낮고 주택유형을 무응답 층화변수로 고려한 경우(나, 라)의 CV가 그렇지 않은 경우에 비해 낮다. 지역층 개수에 큰 차이는 없지만 구체적으로 볼 때 지역층이 세분화 될수록 CV가 약간 작아지는 것을 알 수 있다. 그렇지만 지역층이 세분화될수록 표본크기가 적게 할당된 층의 경우 가중치에 민감하게 반응할 수 있다는 점을 주지할 필요가 있을 것이다.

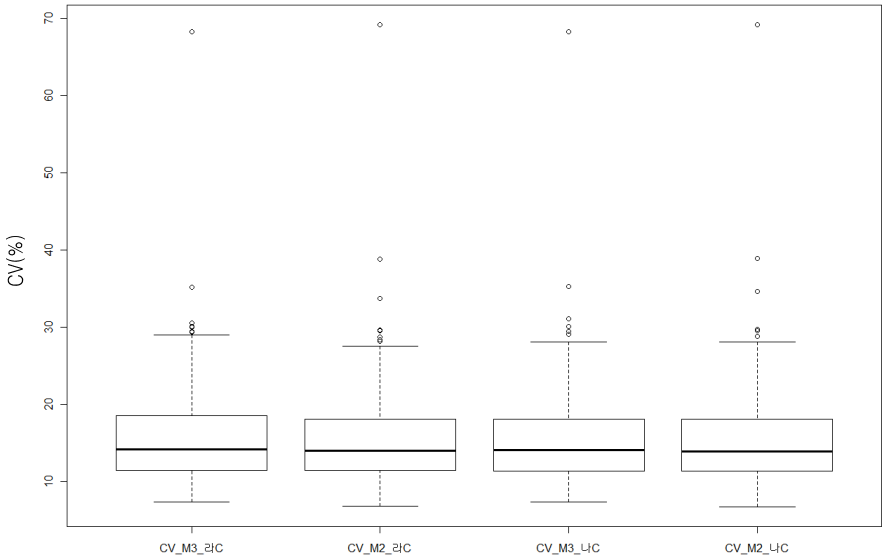
[그림 1-21]과 [그림 1-22]는 추정치의 기존값에 대한 편향을 나타낸 것이다. 기존값이 참값이 아니지만 이러한 가중치를 적용했을 때, 추정치가 기존값에 비해 어느 정도 차이가 있는 지를 살펴볼 수 있다. 주택유형 층을 고려한 경우 편향이 약간 줄어들지만, 설계가중치를 적용한 경우에는 전국 수준에서 큰 차이는 없는 것으로 판단된다. 하지만 230개 시군구별로 보면, 지역에 따라서는 기존값과 가중치 적용 후 추정치가 크게 차이가 있는 지역들도 있다. 따라서 통계의 신뢰성과 정도 향상을 위해서는 표본추출과 응답현황이 반영될 수 있도록 가중치를 적용하는 것이 적합하다고 판단된다.

지금까지의 결과를 간단하게 요약해보면, 가중치의 CV나 추정치의 CV 또는 편향 측면에서 볼 때, 설계가중치와 무응답가중치를 고려한 경우가 좋고, 무응답 층은 230개 시군×주택유형을 고려한 경우가 좋다고 판단된다. 이제 통합계수 방법의 효과를 살펴보자. 다른 방법들에 대한 설명은 생략하고, 위에서 언급한 대로 230×주택유형 층에 대해 설계가중치×무응답가중치를 적용한(라C) 방법을 중심으로 비교해 보자. 아래 그림은 주택유형 층을 고려한(나C와 라C) 방법에 대해 통합계수 방법(M2=표본크기만 고려한 경우, M3=표본크기와 분산비중을 고려한 경우)의 추정치와 CV를 나타낸 것이다. [그림 1-23]과 [그림 1-24]에서 보면, 방법 간 추정치 또는 CV값의 차이는 거의 없다고 판단된다. M3_라C의 경우가 M2_라C의 CV범위가 약간 넓지만 이 정도는 방법 간에 차이를 둘 정도는 아니다. 하지만 세부 지역별로 볼 때 분산비중을 고려하여 두 조사를 통합한 경우가 그렇지 않은 경우에 비해 더 미세한 효과조정은 있다고 볼 수 있다.

위에서 언급한 바와 같이 표본크기와 분산비중을 모두 고려한 가중치 적용이 지역별로는 더 미세한 조정이 가능하지만 그 차이가 거의 없기 때문에 실무적으로 계산이 더 간편한 방법을 선택하는 것도 하나의 방법이 될 것이다. 즉, 본 연구에 따르면 통합계수로서 표본크기 비중만을 고려하여 두 조사를 통합하는 것도 좋은 방법이 될 것으로 판단된다.



[그림 1-23] 통합계수 여부(표본크기 vs. 표본크기×분산비중)와 지역층 세분화(25개 층 vs. 230층)에 따른 시군구별 실업자 추정치 분포



[그림 1-24] 통합계수 여부(표본크기 vs. 표본크기×분산비중)와 지역층 세분화(25개 층 vs 230층)에 따른 시군구별 실업자 추정치의 CV

제5절 결론 및 요약

1. 요약

지금까지 지역별고용조사의 무응답가중치 적용 방법에 대해서 살펴보았다. 2010년 3분기 시점의 지역별고용조사 무응답률은 약 4%(거부가구만 의미)를 넘지 않으며 시군구별로 약간 차이는 있지만, 특정 지역 또는 그룹에서 무응답 가구가 집중되는 현상은 발생하지 않았다. 무응답은 랜덤하게 발생하지 않는 이상 조사추정치에 영향을 준다는 것이 이미 알려진 사실이고, 실제로 국가통계와 같은 대규모 조사에서의 무응답은 일반적으로 랜덤하지 않은 경향이 있다. 본문에서 살펴본바와 같이 우리나라 가구단위 조사의 경우 무응답 가구는 주택유형과 그리고 대도시 여부에 따라서는 무응답률 크기는 조금씩 다르다. 물론 무응답 가구 특성 분석을 위한 좋은 변수들이 있을 수 있겠지만, 본 연구는 현재 주어진 자료에 한하여 제한적으로 사용하였다. 게다가 지역별고용조사는 무응답 가구의 성향이 응답 가구 성향과 다르기 때문에 무응답조정은 조사추정치 질을 높이기 위해 필요한 과정이라 할 수 있다.

지역별조사의 무응답가중치 방법 연구의 시작점은 이 조사 자료의 구조를 이해하는 것으로부터 시작한다. 이 조사는 두 개의 서로 다른 조사가 혼합된 경우라 할 수 있다. 즉, 지역별고용조사 중 일부 표본은 경제활동인구조사 표본에 의한 것이고, 일부 표본은 이 조사를 위해 별도로 추출된 표본에 의한 것이다. 즉, 이 두 조사는 거의 동일한 조사내용에 대해 서로 다른 표본추출 프레임을 사용하고, 조사과정도 조사원 활용 등의 측면서 조금씩 다른 시스템을 사용한다고 할 수 있다. 따라서 본 연구는 지역별고용조사는 서로 독립적인 두 조사의 통합자료로 구성된 것으로 간주하였다. 이와 같은 문제는 가중치 작성과정에 있어서, 각 가중치는 두 조사에서 각각 독립적으로 작성·적용된 후 두 결과를 통합하기 위한 적절한 통합계수를 이용하여 하나의 자료로 통합하는 것과 같다.

본 연구에서는 무응답가중치 적용효과를 살펴보는 것이 일차적인 목적이다. 이를 위해 기본적으로 설계가중치를 적용한다는 것을 전제한다. 무응답가중치 작성에 있어서 무응답 층을 어떻게 구성할것인가는 적절한 가중치를 작성하기 위해 매우 중요하다. 그러나 현실적으로 무응답 가구에 대한 특성을 파악하는 것이 쉽지 않기 때문에 많은 경우에 있어서 지역정보나 기본적인 인구학적 정보를 사용하는 것으로 보인다. 본 자료의 경우도 이미 언급한 바와 같이 무응답 가구에 대해서 알 수 있는 정보는 지역과 표본가구의 주택유형 정보로 매우 제한적이다. 주택유형을 무응답 층 변수로 고려한 것은 이 변수가 고용정보를 파악하는 것과 밀접한 관련이 있다고 볼 수 있기 때문이다.

2. 결과

본 연구는 무응답 층에 대해 크게 지역과 주택유형을 주요정보로 사용하였다. 그리고 두 자료의 통합계수는 전체표본 중 각 조사표본이 차지하는 표본크기비와 두 조사의 분산의 비율을 사용하였다. 각 방법들에 대한 비교결과는 다음과 같다.

① 무응답 층은 지역과 주택유형 층을 이용하는 것이 좋다.

구체적으로 별도표본은 230개 시군구층, 경활표본은 25개 시도/동읍면 층을 이용한다. 이때 경활표본의 표본추출 시 25개 지역층을 고려한다는 점에서 무응답 층 또한 동일한 방법을 사용할 수 있다. 지역별고용조사의 목적이 시군통계작성이라는 점을 고려하여 무응답조정 시 경활표본의 경우도 별도표본과 같이 230개 층을 이용할 수 있을 것이다. 그렇지만 적은 표본 수에 대해 너무 세분화된 층으로 구분할 경우, 특정 층에서 큰 가중치가 발생할 수 있고 이는 오히려 추정치에 과다한 가중치가 적용되는 경우가 우려된다. 따라서 가능하면 지역별고용조사는 무응답 층은 표본설계 층과 동일하게 사용함으로써 가중치 작성을 가능하게 하도록 하는 것이 실무적으로 유용할 것이다.

② 가중치 적용방법은 설계가중치×무응답가중치 방법이 좋다.

현재까지 지역별고용조사의 무응답률은 낮은 편이기 때문에 무응답가중치를 적용하지 않는다 하더라도 전체적인 추정치에 큰 영향은 미치지 않을 것이다. 현재 지역별고용조사의 경우 부재한 가구는 표본대체를 하고 있기 때문에 실제 무응답률은 이 보다 훨씬 높은 것이 사실이다. 게다가 지역에 따라서는 특히 대도시의 경우 무응답 발생 비율이 비도시 지역과 다르기 때문에 세부 지역별 특성을 반영한다면 지역별 무응답 조정을 하는 것이 바람직할 것이다.

③ 두 독립조사를 통합할 때 사용하는 통합계수는 두 조사의 표본크기의 비를 이용하는 것이 중요하다.

이것을 기본으로 하고 추가적으로 좀 더 세부적인 조정을 위해 두 조사의 분산비도 고려할 수 있겠다. 본 연구에서는 표본크기비와 분산비 모두를 사용했을 때 그 결과가 표본크기비만 사용했을 때와 거의 차이가 없었다. 따라서 활용의 간단함을 추구한다면 표본크기만을 고려해도 무방할 것이다.

결론적으로, 지역별고용조사는 다음과 같은 가중치 적용절차를 거침으로써 조사통계

의 품질이 한층 더 향상될 것으로 보인다.

1. 지역별고용조사에 사용되는 경활표본과 별도표본은 독립표본으로 간주한다.
2. 모든 가중치 작업은 각 조사에 대해서 별도로 진행한다.
3. 별도표본의 무응답 층은 230개 시군구×주택유형 층을 사용하고, 경활표본은 25개 시・동/읍면×주택유형 층을 사용한다.
4. 설계가중치는 각 조사의 표본설계 당시의 추출 층내 추출률의 역수를 사용하고, 무응답가중치는 각 무응답 내에서의 응답률의 역수를 사용한다. 이때 경활의 경우는 연동표본을 고려하여 추출률을 반영한다.
5. 가중치 방법은 설계가중치×무응답가중치 방법을 사용한다.
6. 이렇게 작성된 가중치 결과에 대해 경활표본에는 '경활표본크기/지역별고용조사 전체표본크기'를, 별도표본에는 '별도표본크기/지역별고용조사 크기'를 각각 곱해서 두 자료를 통합하여 하나의 자료를 만든다.
7. 이렇게 통합된 결과에 대해 성과 연령변수를 이용하여 사후조정한 후 이를 최종 가중치로 사용하여 추정치를 만든다.

3. 논의 및 향후 연구

조사통계생산에 있어서 가중치 적용은 통계품질과 매우 밀접한 관계가 있다. 이에 대해 여러 가지 이유가 있겠지만, 일부 표본조사를 통해 전국 수준의 통계를 작성해야 한다는 점과 표본의 응답이 완벽하지 않다는 점이 가장 큰 이유가 될 수 있을 것이다. 가급적 표본추출시의 효과를 그대로 반영하고, 불완전한 응답 실태를 과학적인 방법으로 보정해줌으로써 통계의 질을 높일 수 있다는 것은 이미 이론적으로 또는 경험적으로 알려진 사실이다. 그런 점에서 국가통계의 품질향상을 위한 다양한 시도는 높이 평가되어야 하고, 이러한 노력들이 실제 통계생산에 반영될 수 있어야 할 것이다.

지역별고용조사의 무응답가중치 작업은 통계품질 향상에 큰 기여를 할 수 있을 것이다. 그렇지만 본 연구에서 보면 몇 가지 아쉬운 점이 있다. 현실적으로 무응답표본가구의 특성을 파악한다는 것이 매우 어렵다는 점이고, 이를 파악할 만한 변수가 매우 제한적이라는 것이다. 무응답가중치 조정을 위해 무응답조정변수는 매우 중요한 역할을 한다. 무응답 조정을 위한 좋은 변수가 가장 절실하다 하겠다. 지역별고용조사의 경우에서 무응답 층변수로 주택유형을 사용하였지만, 이 외에도 해당가구의 가구원수, 가구주 연령, 주택크기(평수 등) 등의 정보를 수집할 수 있다면 더 좋은 무응답조정이 가능할 것이다. 그렇지만 이와 정보 수집이 현실적으로 어렵고 정확하게 파악할 수 있느냐 하는

것도 관심사항이다. 이러한 정보는 파라데이터와 같은 자료를 통해 수집될 수 있으며, 이러한 파라데이터는 통계품질 향상을 위한 연구에 매우 중요한 자료이다. 현재 통계청의 파라데이터 수집 노력은 통계정책과 통계품질향상에 큰 기여가 있을 것으로 기대한다.

또한 무응답 가구의 허용 범위에 대해 논의가 있어야 한다. 일반적으로 가구 단위에서 조사에서 무응답 가구 유형은 크게 부재 가구, 거부 가구, 기타 가구로 구분하고 있다 (APPOR, 2009). 현재 지역별고용조사는 거부 가구를 무응답 가구로 정의하고 있고, 부재 가구 등은 현장에서의 표본대체를 허용하고 있다. 그렇지만 현실적으로 표본대체가 합리적으로 이루어지기는 매우 어려울 것이며, 이는 결국 원표본의 구조에 영향을 줄 수 있게 된다. 뿐만 아니라 현장조사환경이 점점 어려워지고 있고, 가구로부터 응답을 얻어내기가 점점 어려워지는 환경에서 표본대체가 앞으로 얼마나 가능할 지에 대한 고민을 할 필요가 있고, 이에 대한 대안을 준비해 나가야 할 것이다. 특히 지역별고용조사의 경우, 소지역단위 조사라는 점에서 높은 응답률에 대한 기대는 더 어려워질 것이다. 이에 대해 표본대체 비율을 점점 줄이되, 통계적 방법에 의한 대처방법을 통해 좋은 통계품질을 유지할 수 있을 것이다.

본 연구는 다양한 무응답조정방법을 통해 그 효과를 비교분석한 것으로, 통계적인 평가방법에 최적의 방법을 선택하는데 초점을 맞추었다. 본 연구에서 선택한 방법에 대한 몇 가지 미세한 검토를 통해 실무적 활용이 가능할 것으로 보인다. 실무적 활용을 위해서는 유사한 추정치를 유도할 수 있다면 가능한 간단한 자료처리방법과 가중치 작성방법을 선택하는 것이 좋을 것이다. 시의성 있는 통계작성이 그만큼 중요하기 때문이기도 하다.

지금까지 본 연구에서는 우선적으로 무응답 층을 이용한 가중치 작성방법을 제시하였다. 이론적으로 성향가중치 방법은 무응답 가중치 방법으로 널리 알려진 방법이다. 연구를 시작한 시점에서는 성향가중치를 고려하고자 하였으나, 성향변수의 활용상 제약 또는 연구기간 제약상 이론적 검토만 하고 실제적으로 적용해보지 못한 아쉬움이 있다. 향후 무응답 가구의 성향조정을 위한 좋은 변수를 얻을 수 있다면, 이러한 성향가중치 방법은 보다 좋은 무응답조정효과를 기대할 수 있겠다. 뿐만 아니라 무응답 층에 대한 향후 이런 방법들에 대한 추가 연구가 있을 예정이다. 마지막으로 지역별고용조사의 경우 가구관리종합표를 통해 무응답 가구에 대한 정확한 정보를 추가적으로 수집함으로써 통계 품질 향상에 기여할 수 있기를 기대한다.

참고문헌

김서영, 안다영(2010). 「2010년 통계개발원 연구보고서」 "3장, 가구면접조사에서 무응답률과 무응답 편향", 111-164.

김재광 (2009). 「표본조사론」, 자유아카데미.

신민웅, 이상은 (2001). 「표본설계」, 교우사.

송주원, 안형진 (2009). 「무응답 자료 처리 및 분석」, 통계교육원.

통계청 (2007). 「가구부문 표본개편 결과」, 통계청.

AADOR(2009), "Standard Definitions : Final Dispositions of Case Codes and Outcome Rates for Surveys", The American Association for Public Opinion Research

Hansen, M. W.Hurwitz (1946). "The problem of ninresponse in sample surveys", Journal of the American Statistical Association, 41, 517-529.

Harrod, L.A. and Lesser, V. (2004). "The use of propensity scores to adjust for nonignorable nonresponse bias", ASA section on survey research methods.

Cobben, F. (2009). "New developments in nonresponse adjustment methods", Working paper.

Cobben, F. (2009). "Nonresponse in sample surveys : Methods for analysis and adjustment". Statistics Netherlands.

Cobben, F. and Bethlehem (2005). "Adjusting undercoverage and nonresponse bias in telephone surveys". Discussion paper 05006, Statistics Netherlands, Available at www.cbs.nl.

Cochran, W.G. (1968). "The effective of adjustment by subclassfication in removing bias in observational studies", Biometrics, 24, 205-213.

Merkoris, T. (2010). "Combining information from multiple surveys by using regression for efficient small domain estimation", J.R.Statist.Soc.B, 72, 27-48.Rosenbaum, P.R. and Rubin, D.B. (1983). "The central role of the propensity score in observational studies for causal effects". Biometrika, 70, 41-55.

Rosenbaum, P.R. and Rubin, D.B. (1984). "Reducing bias in observational studies using subclassfication on the propensity score". Journal of the American Statistical Association, 79, 516-524.

Schonlau, M., van Soest, A., Kapteyn, A. (2007). "Are Webographic' or Attitudinal questions useful for adjusting estimates from web surveys using propensity scoring ?". Journal of the American Statistical Association, 79, 516-524.

<부 록>

〈부표1〉 230시군구/25시도, 동읍면 × 주택유형 층을 이용한 시군구 실업자수 추정치, CV

cityname	기존	M2_A	M2_라B	M2_라C	CV 기존	CV M2_A	CV M2_라B.00	CV M2_라C
1	3645	3635	3664	3644	27.61	23.54	24.30	24.01
2	1789	2343	1551	2345	26.02	28.09	13.15	28.38
3	4458	3905	4825	3764	27.42	19.26	18.64	18.04
4	5223	5314	5063	5432	29.77	18.84	23.03	18.64
5	9471	9450	9901	9563	23.79	16.52	22.69	17.07
6	6832	6769	6517	6888	20.34	10.61	15.51	10.89
7	12882	12976	13534	13212	20.43	13.55	17.77	14.44
8	11467	11480	11801	11724	23.24	20.02	25.89	21.50
9	4949	5308	4338	5333	24.73	19.03	16.60	19.34
10	7389	7478	6857	7278	25.67	15.51	19.79	14.80
11	10623	9760	9916	10620	25.63	18.17	29.44	25.47
12	10195	10235	10330	10204	22.89	13.96	16.96	13.96
13	6003	5956	6644	6060	34.21	26.83	31.65	27.17
14	8006	8004	9242	8058	31.85	24.28	27.61	25.13
15	11048	11066	11246	11158	16.68	12.51	15.01	12.39
16	10081	9903	10116	10156	20.00	12.24	14.04	12.82
17	8582	8494	8289	8652	21.97	14.95	22.76	16.32
18	10362	9950	11241	10165	16.13	12.93	14.95	13.04
19	5998	6129	5522	6315	23.63	13.05	16.86	11.95
20	11904	11879	11461	12276	21.52	15.71	19.42	16.71
21	13134	12439	11884	11978	22.15	17.16	16.46	15.81
22	6784	6815	7440	6923	21.87	9.80	12.06	10.18
23	6013	5138	4429	5337	32.12	13.99	23.69	17.72
24	10821	10663	10317	10418	24.98	22.55	25.83	21.87
25	10567	10564	10375	10408	19.04	12.78	12.30	11.80
26	605	558	621	559	32.67	19.45	21.34	19.17
27	1967	1892	2057	1876	21.07	11.76	13.29	11.45
28	1607	1821	1367	1845	35.80	39.53	16.86	38.82
29	2894	2861	2932	2822	24.52	14.71	17.66	14.71
30	4222	4952	5250	5033	29.82	20.00	20.62	19.71
31	4853	4932	5181	4883	19.14	8.86	11.24	9.60
32	3971	3527	3026	3542	32.04	11.63	19.06	11.49
33	3402	3231	2908	3264	21.76	8.28	10.93	8.95
34	7145	6735	6725	6889	18.80	12.49	17.76	14.75
35	3907	4040	4283	4121	32.30	18.47	23.57	19.49
36	5974	6231	6740	6342	19.10	15.40	19.00	16.06
37	1042	1107	1030	1111	21.90	15.79	14.28	16.15
38	4184	4169	3962	4055	28.25	20.26	24.37	19.30
39	3037	3134	2659	3107	23.68	15.84	12.80	16.06
40	4192	4178	4106	4134	15.79	11.64	15.01	10.89
41	1181	1014	1219	1002	29.94	21.49	19.38	20.01
42	2111	1730	2214	1765	21.11	15.63	16.92	16.23

cityname	기존	M2_A	M2_라B	M2_라C	CV 기존	CV M2_A	CV M2_라B.00	CV M2_라C
43	7064	6760	6647	6787	26.45	13.69	13.48	13.81
44	4549	4491	4648	4490	21.60	13.41	12.92	13.36
45	4736	4682	4799	4704	15.86	11.28	11.27	11.20
46	6206	6382	6222	6310	19.81	17.35	12.46	17.61
47	6594	6345	6319	6294	23.32	14.28	15.14	15.31
48	15506	17676	16062	17721	18.39	16.21	10.82	16.26
49	3261	2740	3387	2728	22.04	12.59	11.78	11.92
50	1527	1441	1584	1459	22.52	19.90	14.95	20.59
51	3196	3158	3213	3162	16.18	18.12	15.05	18.60
52	8515	7011	7803	7046	23.52	15.65	13.10	15.82
53	5145	5036	4810	4947	14.69	8.48	9.50	8.34
54	11502	10522	10886	10605	18.23	13.45	12.62	13.34
55	17418	21590	18406	21703	16.52	25.28	13.29	25.37
56	7514	7692	7661	7592	16.83	11.05	11.62	11.86
57	8839	8653	8627	8597	19.46	12.01	14.14	14.47
58	609	555	621	556	38.84	24.16	18.20	26.03
59	.	329	328	328	.	17.54	17.06	17.06
60	3095	2757	3047	2783	31.16	21.82	22.15	21.90
61	2447	2435	2256	2400	26.39	8.67	9.19	7.77
62	3942	3993	4059	4065	26.16	17.34	17.50	17.78
63	6732	4255	6803	4274	29.88	15.47	14.37	15.00
64	6385	6649	6071	6671	21.56	17.59	14.05	17.78
65	4566	4551	4568	4591	20.93	14.10	16.14	14.90
66	4351	4297	4249	4290	23.00	12.42	12.97	11.82
67	6286	6286	6258	6258	22.73	9.84	9.94	9.94
68	3357	2968	3085	2989	22.84	8.35	9.64	9.28
69	2518	2259	2276	2256	24.69	8.57	8.53	8.51
70	3825	3917	3709	3810	23.25	13.04	13.57	12.89
71	6097	5839	6103	5828	20.23	16.55	13.32	17.34
72	2573	2499	2556	2504	22.68	12.28	10.76	11.97
73	2614	2624	2674	2678	17.66	12.17	9.69	12.31
74	2834	2838	2876	2854	32.42	19.98	21.16	19.67
75	19462	19239	19214	19264	14.68	12.23	14.43	12.25
76	14694	13682	13253	13815	18.70	12.69	17.12	13.20
77	6312	6357	5848	6444	20.89	12.60	17.51	13.55
78	7643	7364	6660	7266	18.29	10.91	15.90	10.70
79	13959	12291	11016	12416	16.18	11.15	15.53	11.02
80	9937	9763	10602	9922	13.86	9.82	11.71	10.22
81	4360	4360	4407	4446	20.01	10.09	12.55	10.02
82	1911	2394	1850	2441	24.56	34.63	22.60	33.74
83	9113	9002	8892	9078	17.42	9.35	12.08	9.42
84	12164	11020	10330	11083	17.62	10.04	13.77	9.91
85	598	699	581	706	29.13	26.78	18.07	27.06
86	2844	2764	2900	2756	22.31	14.53	17.02	14.15
87	7250	7338	6998	7335	18.71	12.89	14.33	12.59
88	2695	2707	2651	2676	17.59	10.38	10.61	10.71
89	3900	4007	3373	4046	21.99	10.54	14.34	10.52
90	3243	3262	3205	3258	18.79	9.27	12.10	9.13

cityname	기존	M2_A	M2_라B	M2_라C	CV 기존	CV M2_A	CV M2_라B.00	CV M2_라C
91	1991	2018	2010	2045	19.76	10.69	14.74	11.11
92	2532	2288	2714	2293	17.74	11.64	13.33	11.99
93	14054	14768	15328	15103	15.78	12.53	14.10	13.07
94	4525	4605	3782	4600	20.69	13.19	15.51	13.24
95	2961	2978	3151	2977	21.12	13.29	15.89	13.36
96	2714	2736	2650	2765	19.45	11.07	14.51	11.23
97	1846	2112	1642	2146	25.48	18.56	19.40	18.63
98	5605	5359	4726	5389	22.96	15.08	19.17	15.15
99	4129	4359	3798	4363	14.20	12.80	11.27	13.30
100	1178	1140	1204	1133	31.72	14.87	18.08	14.49
101	2602	2563	2788	2550	24.21	17.11	19.33	17.04
102	608	586	639	584	34.57	13.90	13.07	13.94
103	682	626	706	625	23.16	12.72	13.23	12.69
104	581	548	598	548	25.75	9.07	9.14	9.06
105	679	736	611	737	31.08	24.58	14.72	24.85
106	3685	3855	3842	3840	18.31	12.49	14.75	12.07
107	4794	4828	4891	4932	20.03	15.51	18.92	16.04
108	2635	2818	3133	2857	19.54	12.17	13.40	13.02
109	1043	1029	1089	1047	24.77	11.27	15.41	11.13
110	525	596	481	597	28.33	22.64	20.42	22.58
111	1570	1639	1454	1628	19.04	12.67	12.84	11.97
112	615	604	618	597	31.89	17.61	19.68	17.20
113	321	320	301	316	37.05	12.76	18.91	12.63
114	404	404	405	404	30.46	17.84	21.06	17.67
115	312	319	279	317	30.56	18.26	12.86	18.13
116	355	355	356	354	34.36	22.56	26.12	22.54
117	611	608	662	610	22.44	13.95	15.83	14.14
118	250	248	264	248	37.61	17.19	17.61	17.27
119	131	120	139	119	32.74	16.66	18.01	16.35
120	107	95	115	95	36.68	15.15	15.02	14.85
121	419	400	452	400	27.73	14.96	15.30	14.97
122	492	477	548	480	36.90	27.97	30.21	28.16
123	311	321	303	319	34.10	22.85	28.98	22.72
124	8181	9910	9379	9907	20.40	19.71	17.37	19.51
125	2754	3002	2857	2994	13.66	6.58	8.55	6.86
126	1566	1398	1582	1398	23.47	14.40	15.68	14.98
127	2062	2022	1985	2017	17.62	10.78	13.39	10.99
128	233	243	214	243	28.53	13.40	18.04	13.37
129	295	295	301	292	39.66	14.96	24.01	14.95
130	313	318	283	318	34.38	16.80	19.73	16.87
131	898	896	915	898	18.37	9.57	9.06	9.40
132	393	367	441	368	27.74	15.11	16.21	15.15
133	1004	980	854	986	24.92	13.91	13.84	13.65
134	204	198	216	198	27.71	8.40	12.95	8.33
135	507	486	534	488	22.40	15.21	14.93	14.95
136	9880	9979	10069	10137	15.71	16.33	18.17	19.04
137	1351	1456	1087	1449	23.67	17.33	12.92	17.23
138	782	814	770	822	28.88	10.29	17.73	10.72

cityname	기존	M2_A	M2_라B	M2_라C	CV 기존	CV M2_A	CV M2_라B.00	CV M2_라C
139	4206	4045	3852	4079	17.45	12.34	17.50	13.76
140	2336	2313	2326	2365	20.37	11.15	13.67	11.58
141	897	880	767	884	26.17	14.10	13.93	14.69
142	553	578	540	561	21.39	17.40	14.60	17.82
143	541	593	467	595	29.87	15.98	19.36	15.97
144	1277	1277	1268	1277	18.81	9.19	12.02	8.69
145	1206	1192	1294	1195	25.92	17.60	19.42	17.64
146	483	515	422	518	36.49	23.51	25.77	23.61
147	186	185	185	184	41.71	22.61	22.11	22.50
148	1262	1265	1337	1271	22.71	10.61	10.58	10.73
149	1146	1142	1246	1145	22.06	9.10	9.14	9.11
150	857	840	893	840	30.68	21.48	20.80	21.52
151	1586	1593	1653	1609	27.52	11.97	14.81	12.21
152	9870	10097	10065	10011	13.99	10.57	11.81	10.48
153	1994	2021	1970	1979	24.86	9.48	10.97	9.22
154	4061	3952	3929	3971	17.39	9.61	14.24	10.09
155	1800	1753	1885	1744	20.29	13.26	15.21	12.95
156	782	706	915	708	29.71	19.60	19.17	19.71
157	934	965	947	965	28.27	13.62	16.97	13.47
158	991	1005	1084	1004	19.78	7.76	8.06	7.66
159	132	141	123	141	44.45	13.30	28.59	13.12
160	103	112	93	113	39.41	18.41	25.91	18.48
161	121	117	123	117	36.29	17.38	16.18	17.43
162	246	259	229	259	49.14	27.33	39.52	27.27
163	299	289	309	288	38.62	12.84	14.14	12.56
164	204	203	192	204	42.73	11.48	25.45	11.54
165	509	510	530	510	36.09	18.98	18.22	19.05
166	3286	3270	3234	3265	20.75	14.41	18.87	14.47
167	3750	3812	3907	3839	16.40	11.11	14.51	11.87
168	2171	2268	2359	2255	25.24	15.06	19.05	15.21
169	554	580	602	584	34.46	15.83	20.65	15.66
170	1730	1698	1738	1700	20.63	13.99	15.82	13.57
171	647	645	662	645	27.20	14.28	14.70	14.29
172	150	146	158	146	39.40	17.04	18.27	17.09
173	276	268	287	267	34.53	24.11	23.28	24.15
174	342	363	407	362	47.61	23.54	21.09	23.59
175	164	165	174	164	44.30	16.10	17.00	16.38
176	987	997	1023	998	22.78	12.61	12.42	12.54
177	180	179	187	178	48.60	12.29	13.75	10.90
178	102	101	107	101	72.57	29.57	26.55	28.71
179	422	399	347	398	37.34	18.37	31.63	18.48
180	676	666	624	674	24.78	8.13	15.38	8.49
181	928	906	843	907	27.44	18.03	22.89	18.10
182	95	94	96	94	60.71	29.68	30.68	29.68
183	251	244	206	244	44.15	17.38	42.46	16.96
184	695	694	723	693	25.74	17.41	17.51	17.37
185	477	483	515	483	38.27	26.52	28.01	26.52
186	149	153	134	153	39.11	12.41	23.66	12.47

cityname	기존	M2_A	M2_라B	M2_라C	CV 기존	CV M2_A	CV M2_라B. 00	CV M2_라C
187	.	317	316	316	.	28.68	28.81	28.81
188	6582	5926	5793	5967	18.96	12.85	15.30	12.97
189	1710	1572	1511	1549	27.31	14.20	18.79	14.01
190	1105	1110	1164	1124	31.81	16.36	19.55	16.75
191	1235	1244	1299	1254	25.89	8.08	11.43	7.94
192	7171	7166	7571	7555	15.31	12.57	15.88	13.81
193	1552	1513	1574	1514	25.11	14.49	14.32	14.45
194	1061	1115	1067	1115	26.79	13.09	17.90	12.90
195	677	675	521	672	39.93	21.01	37.05	20.77
196	731	723	733	719	25.41	11.80	11.89	12.22
197	3401	3097	2751	3062	21.05	8.84	10.92	9.42
198	100	98	101	98	60.37	29.52	28.93	29.52
199	462	457	482	457	34.36	9.68	9.32	9.69
200	117	113	121	113	44.85	20.16	18.64	20.26
201	9	8	10	8	100.00	.	.	.
202	323	324	319	324	31.82	13.22	19.04	13.11
203	498	453	592	449	28.38	11.26	14.89	11.22
204	785	759	808	760	17.93	8.58	8.36	8.60
205	419	432	406	433	28.52	11.06	16.66	11.05
206	1835	1831	1830	1843	19.68	11.39	14.65	11.60
207	291	289	295	289	45.16	27.54	28.09	27.55
208	371	345	400	345	32.63	16.39	18.85	16.41
209	680	671	701	672	34.23	23.85	27.72	23.90
210	.	.	.	.	.	.	.	.
211	4020	4081	3985	4074	15.08	7.91	12.06	7.70
212	2246	2468	2162	2493	20.55	21.62	13.53	22.27
213	761	710	746	706	30.78	12.97	16.56	12.86
214	6036	6181	6462	6305	14.84	7.86	10.28	8.02
215	1262	1157	1275	1146	27.77	18.74	19.90	18.77
216	1351	1551	992	1593	32.23	17.90	25.71	18.47
217	2551	2567	2662	2651	22.79	12.13	14.49	12.99
218	16459	14915	14383	14876	15.29	13.24	14.78	13.05
219	65	83	53	83	57.78	20.97	47.83	20.99
220	655	667	632	675	25.70	8.93	16.03	9.10
221	528	503	583	503	31.60	13.26	12.79	13.19
222	525	526	518	529	31.45	13.90	19.06	14.06
223	175	172	181	172	84.99	69.21	66.29	69.24
224	437	418	466	418	34.12	16.88	16.26	16.88
225	289	272	303	271	30.01	9.33	10.24	9.33
226	415	398	431	398	33.84	19.58	18.37	19.47
227	702	703	702	704	28.25	17.00	18.59	17.05
228	262	273	234	274	34.38	11.72	18.31	11.82
229	4711	4963	4831	4982	18.44	13.83	12.07	14.87
230	716	736	726	730	37.15	12.62	15.92	11.75