한국통계학회논문집
2009, 16권, 1호, 103–113

표본조사에서 무응답 가중치 조정층 구성방법에 따른 효과

김영원[1,a], 남시주[a]

[a]숙명여자대학교 수학통계학부

요 약

표본조사에서 무응답은 비표본추출오차를 발생시키는 중요한 원인 중 하나이다. 단위무응답이 발생하는 경우 무응답에 의한 편향을 줄이는 동시에 추정의 정도를 향상시키기 위해 단위무응답 조정층을 구성해 무응답 가중치 조정을 하는 것이 일반적이다. 본 연구에서는 무응답 조정층 구성과 관련된 기존의 이론들을 정리하고 어업총조사 자료를 이용한 실증적인 모의실험을 통해 효과적으로 무응답 조정층을 구성하는 방안에 대해 살펴본다. 모의실험결과 응답성향에 따른 조정층 구성보다는 예측평균을 기준으로 한 조정층 구성이 효율성 측면에서 효과적인 것으로 나타났으며, 아울러 다른 관심변수에도 적용될 수 있는 로버스트한 조정층 구성을 위해서는 예측평균만을 고려하는 것보다 응답성향과 예측평균을 모두 고려한 조정층 구성방법이 효과적인 것으로 나타났다. 한편 무응답 조정을 위한 응답률 산출에 있어서 설계가중치의 적용 필요성에 대해 살펴본 결과 설계가중치 적용 여부는 추정결과에 거의 영향을 주지 않는다는 사실을 확인할 수 있었다.

주요용어: 무응답 가중치, 무응답 조정층, 비편향 추정량, 어업총조사, 편향.

1. 서론

표본조사에서 무응답은 비표본추출오차(non-sampling error)를 발생시키는 주요한 원인 중 하나이다. 무응답은 면접거절 또는 접촉 불능 등으로 조사단위로부터 전혀 정보를 얻을 수 없는 단위무응답(unit nonresponse)과 조사는 이루어졌지만 일부 항목에 대해 결측이 발생한 항목무응답(item nonresponse)으로 구분될 수 있다. 무응답이 발생하는 경우 응답 자료만을 갖고 분석을 하면 심각한 편향(bias)이 발생할 가능성이 높다. 이런 문제를 해결하기 위해 항목무응답의 경우 조사된 항목들을 활용해 김영원과 조선경 (1996)에 정리된 것과 같이 다양한 형태의 대체(imputation) 방법이 이용된다. 단위무응답의 경우 표본설계 시점에 갖고 있던 제한적인 정보를 제외하고는 조사를 통해 추가적인 정보를 얻을 수 없기 때문에 응답률에 따른 가중치 조정(weighting adjustment)을 통해 편향을 줄이게 된다.

본 연구에서는 단위무응답에 대한 처리 방법으로 추정량의 정확성(accuracy)을 향상시키기 위해 단위무응답 조정층(adjustment cell)을 활용하여 편향을 줄이는 동시에 효율성도 제고할 수 있는 조정층 구성 방안에 대하여 논의하고자 한다. 무응답 조정층을 구성하는 데 있어서 일반적으로 사용되는 방법은 두 가지로 정리된다. 첫 번째 방법은 응답 가능성(확률)이 유사한 단위들로 조정층을 구성하여 가중치를 조정하는 방법으로 이 방법은 편향을 줄이는 것에 초점을 맞추고 있다. 두 번째 방법은 관심변수가 비슷한 값을 가질 것으로 예측되는 단위들로 조정층을 구성하여 가중치를 조정함으로써 추정량의 분산을 줄이는 것에 초점을 맞춘 것이다. 여기서 첫 번째 방법은 응답성향(response propensity)을 기준으로 무응답 조정층을 구성하는 방법이고, 두 번째 방법은 관심변수에 대한 예측평균(predictive

본 연구는 숙명여자대학교 2007학년도 교내연구비 지원에 의해 수행되었음.
[1] 교신저자: (140-742) 서울 용산구 효창원길 52, 숙명여자대학교 수학통계학부, 교수. E-mail: ywkim@sm.ac.kr

mean)을 기초로 무응답 조정층을 구성하는 방법이다. 조정층을 구성한 후에는 각 층내에서 응답한 단위들에 대한 설계가중치(design weight)의 합이 응답한 단위와 무응답한 단위들에 대한 설계가중치의 합이 되도록 응답 단위들의 가중치를 조정함으로써 각 층내에서의 응답 단위가 무응답 단위들을 보상해 주도록 하는 것이다

응답성향을 기초로 무응답 조정층을 구성하면 편향을 줄일 수 있지만 분산이 오히려 증가할 수도 있기 때문에 평균제곱오차 측면에서 효율성이 떨어진다는 문제를 갖고 있으며, 예측평균을 기준으로 조정층을 구성하는 경우 해당 관심변수에 대해서는 효율성을 높일 수 있지만 다른 관심변수에 대해서는 효율성이 보장되지 않는다는 문제가 있다. 대부분의 표본조사는 많은 수의 변수에 대해 동시에 관심을 갖는 다목적조사를 염두에 두고 있기 때문에 예측평균을 기준으로 조정층을 적용하는데 한계가 있다. Vartivarian과 Little (2002)은 응답성향 또는 예측평균을 기준으로 조정층을 구성하는 경우 어떤 장단점이 있는지 보여주고 있다. 한편 Little과 Vartivarian (2005)는 무응답 가중치 조정을 하는 경우 편향을 줄일 수 있지만 분산이 증가할 우려가 있다는 기존의 주장과는 달리 편향과 분산을 동시에 줄이는 것이 가능하다는 결과를 보여주고 있다. 하지만 이들 연구결과들은 실제 조사자료를 기반으로 한 것이 아니라 확률모형을 설정해 이론적인 시뮬레이션을 통해 이런 결과를 보여주고 있기 때문에 매우 가상적인 조건하에서 이런 현상을 설명하고 있다는 한계가 있다. 실제 조사자료를 분석하는 경우 어떤 변수가 응답성향을 설명해 주고 어떤 변수가 예측평균을 설명해 주는지 명확하지 않고, 특정 변수는 응답성향과 예측평균 모두에 영향을 주는 경우도 있기 때문에 실제 자료에서 무응답 조정층을 구성하는 문제는 그리 간단한 문제가 아닐 수 있다.

따라서 여기서는 2000년 우리나라 어업총조사 자료를 기초로 실제 어가 관련 표본설계를 하는 경우 어떤 변수들을 사용해 무응답 조정층을 구성하는 것이 효과적인지 실증분석을 통해 논의해 보기로 한다. 2장에서는 무응답 가중치 조정과 관련된 이론을 정리하고, 3장에서는 어가조사를 위한 표본설계에서 무응답 조정층 구성방법에 따른 효율성을 비교분석해 보기로 한다.

2. 단위무응답 가중치 조정에 대한 이론적 배경

무응답 가중치 조정과정은 표본추출확률에 따른 설계 가중치를 확장한 것으로 볼 수 있다. 표본 자료 분석에서 비편향(unbised) 추정량을 얻기 위해서는 추출된 i번째 단위의 포함확률(π_i)의 역수를 설계가중치로 사용한다. y_i를 관심변수라고 하고 Y를 모집단에서 관심변수의 총계라 가정하면, 무응답이 발생하지 않는 경우 HT(Horvitz-Thompson) 추정량 $\widehat{Y_{HY}} = \sum \pi_i^{-1} y_i$가 모집단 총계 Y에 대한 비편향 추정량이 된다.

단위무응답이 발생하는 경우, 표본추출부터 응답을 얻는 과정은 먼저 전체 표본을 추출하고, 그 다음 표본으로 추출된 단위 중 응답자가 선택되는 과정으로 설명될 수 있다. 따라서 만약 표본으로 추출된 단위가 응답할 확률 ϕ_i를 알 수 있다고 하면, HT 추정량을 확장한 형태의 다음 추정량이 모집단 총계 Y에 대한 비편향 추정량이 된다.

$$\widehat{Y_R} = \sum \pi_i^{-1} \phi_i^{-1} y_i, \tag{2.1}$$

여기서 합은 모든 응답단위들에 대한 것이다. 흔히 응답률의 역수에 해당하는 ϕ_i^{-1}을 무응답 보정 상수라고 한다. 응답률을 기초로 한 추가적인 가중치는 추출확률에 따른 가중치와 유사한 개념으로 이해할 수 있다. 만약 각 단위들에 대한 응답률을 알 수 있으면 단위 무응답이 발생해도 비편향 추정이 가능하지만 실제로는 응답률 ϕ_i를 알 수 없기 때문에 이를 추정해야 한다.

조사자료 분석에 있어서 무응답 가중치 조정을 위해 ϕ_i를 추정하는 방법으로는 활용 가능한 보조정보들을 이용한 로지스틱 회귀모형을 사용하는 방법과 응답성향이 유사한 단위들로 무응답 조정층을 구

표 1: 조정층 구성 변수와 응답성향/관심변수와의 관련성이 추정결과에 미치는 영향

응답성향과의 관련성	관심변수와의 관련성	
	낮음	높음
낮음	[경우 1] 편향: --- 분산: ---	[경우 3] 편향: --- 분산: ↓
높음	[경우 2] 편향: ↓ 분산: ↑	[경우 4] 편향: ↓ 분산: ↓

성한 후 각 조정층내에서 응답률을 계산하는 방법이 사용되고 있다. 본 연구에서는 조정층을 구성해 ϕ_i를 추정한 값을 식 (2.1)에 대입해 설계가중치를 보정함으로써 편향을 줄이는 경우 어떤 방법으로 무응답 조정층을 구성하는 것이 효과적인지 살펴보고자 한다. 만약 단위 i가 조정층 c에 속하는 경우 가장 간단한 방법은 다음과 같이 ϕ_i를 해당 조정층 내의 전체 단위 중 응답한 단위의 비율로 추정하는 것이다.

$$\widehat{\phi_c} = \frac{r_{c+}}{n_{c+}}. \tag{2.2}$$

이 경우 설계가중치를 무시한 것이기 때문에 조정층 내의 단위들의 설계가중치가 같지 않으면 응답률에 대한 비편향 추정량이 아니다. 만약 최소한 근사적으로 응답률에 대한 비편향성을 만족하는 추정결과를 얻으려면 설계가중치를 반영한 다음 가중치를 사용할 수 있다.

$$\widehat{\phi_c} = \frac{\sum_{i \in R_c} \pi_i^{-1}}{\sum_{i \in S_c} \pi_i^{-1}}, \tag{2.3}$$

여기서 S_c는 조정층 c에 포함된 전체 단위들의 집합을 나타내고, R_c는 조정층 c에서 응답 단위들의 집합을 의미한다. 가중된 응답률 (2.3)을 무응답 가중치 조정에 사용하는 것이 일반적이지만, 단순한 형식의 (2.2) 대신 가중 응답률 (2.3)을 사용하는 것이 얼마나 효과적인지에 대해서도 검토가 필요하다 (Chapman 등, 1986; Little과 Vartivarian, 2003).

Vartivarian과 Little (2003)이 제시한 것과 같이 조정층 구성과정을 정리해 보면, 응답자과 무응답자 모두에 대해 알려진 변수들의 집합을 D라고 했을 때(D에는 층화 변수와 모든 다른 보조변수 등을 포함함), 만약 D값이 같다는 조건하에서, 응답자와 무응답자에 대한 관심변수 Y의 분포가 같다면, 무응답이 MAR(missing at random)인 것으로 볼 수 있다 (Little과 Rubin, 2002). 이런 경우 R이 응답 여부를 나타내는 지시변수이고, Y가 관심변수라면 다음의 관계가 성립한다.

$$R \amalg Y | D, \tag{2.4}$$

여기서 $\amalg$는 두 변수 사이의 관계가 독립이라는 것을 의미한다.

만약 D가 취할 수 있는 모든 경우에 대해 별도의 조정층을 구성하면 일부 조정층에 포함되는 표본의 수가 너무 작아서 무응답 조정 상수 산출과정이 불안정해 질 수 있다. 따라서 D를 기초로 조정층의 개수를 적정한 수준으로 조정하면서 무응답 편향을 제거하고 동시에 분산을 줄이는 조정층 구성방안을 모색하게 된다. 이런 목적을 달성하기 위해서는 응답경향을 설명해 주는 변수와 예측평균을 설명해 주는 변수를 조화롭게 선택해 조정층을 구성하는 것이 필요하다.

Little (1986)은 만약 $\widehat{Y}(D)$이 D가 주어졌을 때 Y의 예측평균이고, $\widehat{P}(D)$이 D가 주어졌을 때 예측된 응답확률이라면 (2.4)가 만족되며, 다음과 같은 조건이 성립한다는 것을 보여주었다.

$$Y \perp\!\!\!\perp R \mid \hat{Y}(D) \tag{2.5}$$

$$Y \perp\!\!\!\perp R \mid \hat{P}(D), \tag{2.6}$$

여기서 식 (2.5)는 예측평균을 기초로 하여 조정층을 구성하는 경우를, 식 (2.6)은 응답경향을 기초로 하여 조정층을 구성하는 경우를 나타낸다. 식 (2.5)와 (2.6)에서처럼 각 조정층 내에서 무응답이 무시할 수 있는 무응답(ignorable nonresponse)인 경우, 무응답 편향을 제거해 주게 된다. 특히 식 (2.6)과 같이 응답성향만을 고려해 조정층을 구성하는 경우 편향은 줄어들지만 관심변수 Y값의 변동이 작아진다는 보장이 없기 때문에 추정량의 분산이 커지는 결과가 발생할 수 있다. 반면에 식 (2.5)와 같이 예측평균을 기준으로 조정층을 구성하는 경우 관심변수마다 조정층이 달라진다는 한계가 있다. 관심변수가 여러 개인 경우, 조정층을 구성하는데 있어 인자분석이나 주성분분석을 사용하는 방법도 고려 할 수 있으나 현실적으로 매우 복잡하다.

한편, $\hat{P}(D)$과 $\hat{Y}(D)$을 결합하는 방식으로 조정층을 구성하면 다음과 같은 "이중의 로버스트성(double robustness)" 성질을 가지게 된다. 첫째, 만약 $\hat{P}(D)$이 정확하게 지정되고 $\hat{Y}(D)$이 부정확하게 지정되는 경우에도 결합분류를 통해 편향을 통제할 수 있다. 둘째, 만약 $\hat{P}(D)$이 부정확하고 $\hat{Y}(D)$이 정확하게 지정되는 경우에도 결합분류를 통해 편향을 통제하고 분산의 크기를 줄여 효율성의 증가를 가져올 수 있다. 결론적으로 $\hat{P}(D)$와 $\hat{Y}(D)$을 모두 적용한 결합분류는 편향을 줄이는 동시에 무응답 가중치 조정의 효율성도 향상시키는데 도움이 된다 (Vartivarian과 Little, 2003). Little과 Vartivarian (2005)는 무응답 조정층 구성에 사용된 변수와 응답성향 또는 관심변수와의 관련성에 따라 가중치 조정이 편향과 분산에 미치는 영향을 표 1과 같이 정리하고 있다. 이를 통해 무응답 조정에 있어서 단순히 편향을 줄이는 것에 초점을 맞추는 것보다 편향과 분산을 모두 설명해 주는 MSE (Mean Squared Error)를 줄일 수 있는 조정층을 구성하는 것이 중요하다는 점을 알 수 있다. 여기서 '↑'는 증가, '↓'는 감소, '---'는 변동이 없음을 의미한다.

하지만 표 1은 실제 상황을 매우 단순화한 가상적인 것이고, 실제 조사자료의 무응답 처리에 있어서 변수간의 관련성은 매우 복합적이기 때문에 표 1에 제시된 것보다 훨씬 복잡한 양상을 보인다. 따라서 실제 사례를 이용해 무응답 조정층 구성방식에 따라 무응답 가중치 조정이 추정결과에 미치는 영향을 살펴보는 것이 조정층 구성에 따른 효과를 이해하는데 도움이 될 것이다. 이런 점을 고려해 다음 절에서는 2000년 어업총조사 자료를 이용해 표본조사에서 무응답 가중치 조정층 구성방법이 추정의 정확성에 미치는 영향을 살펴보기로 한다.

3. 효율성 비교를 위한 모의실험

3.1. 모의실험 개요

단위무응답 가중치 조정을 위한 조정층 형성에 따른 효과를 분석하기 위해 2000년 어업총조사 자료를 이용해 모의실험을 수행했다. 모의실험을 위해 전체 어가 모집단에서 표본을 추출하고, 이 중 일부 표본에 대해서 주어진 무응답 패턴 및 응답률에 따라 무응답을 발생시키고, 다양한 변수를 활용해 무응답 조정층을 구성한 후 가중치를 조정하는 경우, 편향과 RMAE (Root Mean Squared Error)를 기준으로 조정층 구성방법에 따른 효과를 실증적으로 분석해 보고자 한다. 실제 조사 자료를 이용한 모의실험 과정을 정리하면 다음과 같다.

표 2: 어업형태에 따른 무응답 패턴

어업형태			양식어업	자기어선	남의 어선 승선	어선 비사용
전체 표본에서 무응답률 20%	2%차이	패턴①	0.20	0.22	0.18	0.16
	5%차이	패턴②	0.20	0.25	0.15	0.10
	10%차이	패턴③	0.20	0.30	0.10	0.00
전체 표본에서 무응답률 30%	2%차이	패턴④	0.30	0.32	0.28	0.26
	5%차이	패턴⑤	0.30	0.35	0.25	0.20
	10%차이	패턴⑥	0.29	0.39	0.19	0.09

(1) 모집단과 표본추출

실제 상황에서 무응답 가중치 조정을 하기 위해서 어떤 방식으로 조정층을 구성하는 것이 효과적인지 분석하기 위해 2000년 어업총조사 자료를 기초로 모집단을 설정하고, 모의실험을 수행하기로 한다. 2000년 어업총조사 자료에 의하면 전국의 어가수는 81,571개이다. 이 중에서 전남이 26,936개로 전체 어가의 33.0%이고, 경남이 14,009개로 17.2%, 충남이 9,444개로 11.6%이며, 이들 3곳이 전체 어가의 61.8%를 차지하고 있다. 어업형태별로 보면 어로어업이 56,761가구로 전체 어가의 69.6%를 차지하고 있다 (류제복 등, 2002). 어업총조사의 어가 중 2종 겸업(어가는 전업과 1종 겸업, 2종 겸업으로 구분됨)이라든지 내륙에 위치한 어가의 경우 실제로 거의 어업활동을 하지 않거나 단순히 선박을 보유한 특수한 경우로 일반적인 어업생산 활동을 하는 어가로 볼 수 없기 때문에 모의실험에서는 전국 어가 중 전업어가와 1종 겸업어가를 대상으로 하고, 실제 어업활동이 활발하지 않는 지역에 해당하는 5가구 이하의 어가가 있는 조사구를 배제한 56,633어가를 모집단으로 보고 모의실험을 수행했다.

모의실험을 위한 표본은 편의상 전국에서 1,200어가를 층화추출하는 것으로 가정했다. 시도별로 최소 40가구 이상의 표본 어가를 확보하도록 하기 위해 각 시도별로 25가구씩 배정한 후 나머지 925가구에 대해서는 각 시도별 어가수에 비례배분하는 것으로 시도별로 표본크기를 정하고 어가를 층화추출했다. 어가가 적은 울산과 경기의 경우 40어가, 41어가가 표본으로 배정되었고, 어가가 많은 전남, 경남의 경우 329어가, 201어가가 표본으로 배정되었다. 이 경우 시도별로 표본추출률이 다르기 때문에 비편향 추정을 위해서는 설계가중치를 반영한 HT 추정량을 사용하는 것이 필요하다. 이런 방식으로 추출된 1,200어가로 구성된 표본을 1,000번 반복 추출하고, 이들 1,000개 표본에서 얻어지는 결과를 기초로 조정층 구성에 따른 효과를 분석했다.

(2) 표본자료에서 무응답 생성

어업총조사에서는 어업형태, 연간 판매액, 선박보유톤수, 비동거 종사자수, 양식면적, 전·겸업 여부 등이 조사되고 있다. 이 중 어업형태가 어가의 특성을 구분하는 가장 대표적인 변수로 볼 수 있다. 따라서 모의실험에서는 어업형태에 따라 무응답률을 달리하는 방식을 통해 전체 표본에서는 무응답이 무시할 수 없는 무응답(non-ignorable nonresponse)이 되도록 무응답을 생성했다. 참고로 어업형태는 '양식어업', '자기어선(빌린 어선 포함)', '남의 어선승선', '어선 비사용'으로 분류된다. 아울러 어가경제조사 등에서 관심이 높은 소득 관련 변수인 연간 판매액을 관심변수로 설정하고, 무응답 패턴에 따라 표본 어가 중 일부 어가에 대해서는 연간 판매액이 결측된 것으로 처리했다.

본 연구에서는 1,200 표본 어가에 대한 전체 무응답률이 20%와 30%인 경우를 고려했으며, 각각의 경우 어업형태별로 2%, 5% 또는 10%씩 무응답률이 차이가 나도록 결측이 있는 표본 자료를 생성했다. 구체적인 무응답 발생 패턴을 정리하면 표 2와 같다. 표 2를 보면 6가지의 무응답 패턴을 고려하고 있다. 예를 들어, 첫 번째 경우는 어업형태별로 무응답률을 20%, 22%, 18%, 16%로 설정하여 전체적으로 무응답률이 20% 정도가 되도록 표본에서 무응답을 발생시킨 것이다.

(3) 모의실험을 위한 관심변수 및 조정층 변수 선정

모의실험에서는 어가경제조사 등에서 주요 관심대상이 되는 어가소득과 비슷한 속성을 갖는 어가의 연간 판매액을 관심대상 변수로 설정했다. 결과적으로 표본조사를 통해 전국 어가의 연간 판매액에 대한 모집단 평균을 추정하는 것을 고려한 것이다. 무응답은 어업형태에 따라 응답률이 다른 것으로 설정했기 때문에 조정층 구성에 있어서 어업형태 변수가 응답경향을 나타내는 변수에 해당한다. Little과 Vartivarian (2005)이나 Vartivarian과 Little (2003) 등의 기존 연구와는 달리 가상적인 모형을 통해 모의실험 자료를 생성한 것이 아니기 때문에 여기서는 인의적으로 설정한 무응답 발생 패턴을 설명해 주는 어업형태 변수이외의 다른 변수들에 대해서는 이들 변수들이 예측평균과 관련성이 높은지 또는 응답성향과 관련성이 높은지를 표 1과 같이 단순화에서 설명할 수 없는 상황이다.

모의실험에서는 다양한 현상을 관측해 보기 위해 어업형태, 시도, 선박보유톤수(이하 선박톤수), 전·겸업 등의 변수를 이용한 조정층 구성방법을 비교대상으로 했다. 여기서 어업형태는 응답경향을 설명하는 변수로 표 1에서 [경우 2]에 해당하는 변수이고, 선박톤수로 조정층을 구성하는 경우는 표 1에서 [경우 3]에 가까운 것으로 볼 수 있다. 만약 어업형태와 선박톤수를 동시에 이용한 교차분류 형식의 조정층을 구성하게 되면 표 1의 [경우 4]와 유사한 상황으로 볼 수 있다. 한편 시도를 조정층 변수로 사용하는 경우는 표 1의 [경우 1]에 해당하는 것으로 볼 수 있다. 참고로 연속형 변수인 선박톤수의 경우 7개 범주(0톤 이하, 0~1톤, 1~2톤, 2~4톤, 4~5톤, 5~10톤, 10톤 이상)로 구분해 조정층 구성에 사용했다. 이와 같이 다양한 특성을 갖는 변수들을 기준으로 조정층을 구성하는 경우 무응답을 보정한다는 측면에서 어떤 차이가 발생하게 되는지 분석했다.

(4) 편향과 RMSE 계산

모의실험에서는 표본크기가 1,200인 표본을 반복적으로 1,000번 발생시켜 각 표본에서 무응답을 발생시킨 후 다양한 조정층 구성에 따른 가중치 조정을 통해 얻어지는 추정결과를 비교하고 있다. 각 조정층 구성방법에 따른 효과를 비교하기 위해 다음과 같은 방식으로 편향과 RMSE를 구했다.

$$\mathrm{Bias}\left(\hat{\bar{y}}\right) = \left(\frac{1}{1000}\sum \hat{\bar{y}}_i\right) - \bar{Y},$$

$$\mathrm{RMSE}\left(\hat{\bar{y}}\right) = \sqrt{\mathrm{MSE}\left(\hat{\bar{y}}\right)} = \sqrt{\frac{1}{1000}\sum\left(\hat{\bar{y}}_i - \bar{Y}\right)^2},$$

여기서 $\bar{Y}$는 모집단 평균이고, $\hat{\bar{y}}_i$는 1200개 어가로 구성된 i번째 표본에서 무응답 가중치 조정과정을 통해 산출된 모집단 평균에 대한 추정치이다. 참고로 어가의 연간 판매액 모집단 평균, 즉 $\bar{Y}$는 2530.611이고, 효율성 비교를 위해 편향과 RMSE을 살펴보기로 한다.

3.2. 조정층 구성방법에 따른 효율성 비교

앞 절에서 설명된 무응답 패턴(①~⑥)과 다양한 변수를 사용한 조정층 구성 방법에 따라 무응답 가중치 조정을 하는 경우 얻어지는 추정량의 편향과 RMSE를 비교한 결과는 표 3과 같다.

표 3을 보면 우선 [1]과 같이 무응답률을 20%, 30% 발생시킨 후 무응답을 무시하고 응답자료만을 사용해 추정하는 경우, 예상했던 것처럼 편향이 발생하고 있으며, RMSE가 커져 추정의 효율성이 떨어짐을 확인할 수 있다. [2]의 경우는 응답성향이나 예측평균과 무관한 층화변수인 시도를 기준으로 한 무응답 조정층을 이용해 가중치를 보정한 것으로 효율성 측면에서 전혀 도움이 되지 않음을 볼 수 있다.

[3]과 [4]를 보면 응답성향을 설명하는 어업형태(4개 범주)를 기준으로 조정층을 구성하는 경우 추

표 3: 무응답 조정층 구성에 따른 편향과 RMSE

무응답 조정층 구성 변수	전체 무응답률 20%					
	패턴 ①		패턴 ②		패턴 ③	
	bias	RMSE	bias	RMSE	bias	RMSE
[1] 무응답 무시	−19.4	148.8	−56.2	153.0	−106.0	175.8
[2] 시도	−18.9	148.4	−54.9	152.3	−102.0	173.4
[3] 어업형태(가중)	1.3	149.3	−2.9	146.8	2.4	149.3
[4] 어업형태(비가중)	1.2	149.3	−3.0	146.6	2.4	149.2
[5] 어업형태*시도	0.0	148.5	−4.5	145.8	−0.2	148.6
[6] 선박톤수(가중)	1.4	144.5	−4.4	144.2	0.5	144.3
[7] 선박톤수(비가중)	2.1	144.7	−3.0	144.7	3.5	144.4
[8] 선박톤수*시도	−1.8	145.1	−6.8	146.3	−7.3	145.6
[9] 어업형태*선박톤수(가중)	1.8	144.4	−3.4	144.4	0.5	144.4
[10] 어업형태*선박톤수(비가중)	1.9	144.7	−3.6	144.9	0.5	144.3
[11] 어업형태*전겸업*선박톤수(가중)	1.7	144.7	−3.3	144.7	0.1	143.8
[12] 어업형태*전겸업*선박톤수(비가중)	1.8	145.0	−3.6	145.2	0.1	143.8

무응답 조정층 구성 변수	전체 무응답률 30%					
	패턴 ④		패턴 ⑤		패턴 ⑥	
	bias	RMSE	bias	RMSE	bias	RMSE
[1] 무응답 무시	−22.2	157.7	−60.2	165.5	−60.2	190.5
[2] 시도	−21.7	157.3	−58.1	164.5	−113.3	187.5
[3] 어업형태(가중)	1.6	157.9	1.6	159.5	4.4	159.4
[4] 어업형태(비가중)	1.4	158.0	1.5	159.6	4.4	159.4
[5] 어업형태*시도	0.0	157.1	−0.4	159.6	3.2	159.2
[6] 선박톤수(가중)	−0.6	148.5	−1.5	149.8	0.9	148.7
[7] 선박톤수(비가중)	0.4	149.3	−0.2	150.6	4.4	149.0
[8] 선박톤수*시도	−8.8	152.4	−8.8	152.4	−8.5	153.8
[9] 어업형태*선박톤수(가중)	−0.2	149.2	−0.4	150.5	1.1	148.0
[10] 어업형태*선박톤수(비가중)	0.1	150.2	−1.0	151.5	1.0	148.0
[11] 어업형태*전겸업*선박톤수(가중)	−0.4	149.7	−1.0	151.6	0.4	149.4
[12] 어업형태*전겸업*선박톤수(비가중)	−0.2	150.8	−1.7	152.6	0.3	149.5

정에 미치는 영향을 알 수 있다. 이 경우 [1]에 비해 편향은 대폭 줄어들었지만 RMSE 측면에서 보면 별 도움이 되지 않는다. 이 경우 편향을 줄이기 위해 가중치를 조정함으로써 결과적으로 분산이 증가하게 되어 RMSE 측면에서 보면 효율성 향상에 도움이 되지 않음을 볼 수 있다. 이런 결과는 무응답 조정을 위해 가중치를 조정하게 되면 편향을 줄일 수 있지만 반면에 분산이 증가할 우려가 있다는 Kalton과 Kasprzyk (1986), Kish (1992) 등의 주장과 일치하는 결과를 보여준다.

반면 예측평균을 기준으로 조정층을 구성하는 것에 가까운 [6]과 [7]의 선박톤수(7개 범주)를 이용하는 경우 편향을 줄여줄 뿐만 아니라 RMSE가 감소하고 있음을 볼 수 있다. 아울러 응답성향을 나타내는 어업형태와 예측평균을 설명하는 선박톤수를 모두 사용한 교차분류 형태의 조정층을 사용한 [9]와 [10]의 경우 편향도 줄어들면서 RMSE도 감소하지만 선박톤수만을 사용하는 [6]이나 [7]과 큰 차이는 없는 것으로 나타났다. 반면 [11]이나 [12]와 같이 어업형태, 선박톤수 및 전겸업을 모두 사용해 조정층을 구성하는 경우 조정층의 개수가 56개로 너무 많기 때문에 오히려 효율성이 떨어질 수 있다는 것을 보여준다. 이런 현상들은 전체 무응답률이 20%인 경우와 30%인 경우에 있어서 큰 차이를 보이지 않고 있다.

아울러 표 3에서 [3]과 [4], [6]과 [7], [9]와 [10] 및 [11]과 [12]의 결과를 서로 비교해 보면, 조정층을 구성한 다음 조정층내의 응답률을 산출하는 과정에서 식 (2.2)와 같이 설계가중치를 적용하지 않는

방법(비가중)과 식 (2.3)과 같이 설계가중치를 적용하는 방법(가중)의 차이를 볼 수 있다. 결론적으로 모든 경우에 있어서 응답률을 산출할 때 설계가중치를 적용하는 것이 큰 의미가 없음을 알 수 있다. 한편 [5]와 [8]처럼 층화변수에 해당하는 시도가 조정층 구성 변수로 사용되면 조정층 내 단위들의 설계가중치가 모두 같아지기 때문에 식 (2.3)이나 식 (2.2) 어떤 응답률 산출방법을 사용해도 결과는 같다.

이와 같은 비교결과를 종합해 보면 우선 응답성향만을 고려해 조정층을 구성하면 편향을 줄이는 데는 도움이 되지만 예상과는 달리 RMSE를 기준으로 한 효율성 측면에서는 무응답 가중치 조정이 별 도움이 되지 않는다는 것을 알 수 있다. 반면에 예측평균을 기준으로 조정층을 구성하면 편향을 줄이는 동시에 분산을 줄일 수 있기 때문에 효율성 측면에서 효과적이라는 사실을 확인할 수 있다. 물론 이 경우 다른 변수를 관심대상으로 하는 경우 어떤 결과가 나오는 지에 대해서는 추가적인 검토가 필요하다. 아울러 [9]나 [10]처럼 응답성향을 나타내는 변수와 예측평균을 나타내는 변수를 모두 사용하는 방식이 상당히 효율적인 방법이 될 수도 있을 것이다. 또한 조정층 구성 후 무응답 보정 상수를 산출하는 과정에서 설계가중치를 적용한 식 (2.3)을 사용하는 대신 설계가중치를 무시한 식 (2.3)을 사용하더라도 큰 차이가 없음을 볼 수 있다.

3.3. 다른 관심변수의 무응답 조정에 미치는 효과

조정층 구성방법에 따른 효과를 비교해 본 결과 응답성향을 중심으로 한 조정층 구성보다는 예측평균을 기준으로 조정층을 구성하는 것이 효율적임을 볼 수 있었다. 하지만 예측평균을 기준으로 조정층을 구성하게 되면 관심변수가 달라지는 경우 조정층 구성도 변경되어야 하기 때문에 실제 적용에 있어서 현실적으로 상당한 어려움이 따를 수 있다. 따라서 여기서는 특정 변수에 초점을 맞추어 구성한 조정층을 변경하지 않고 다른 관심변수에 대해서도 동일한 조정층을 사용해 무응답 가중치 조정을 하는 경우 어떤 결과가 발생하는지 살펴보기로 한다.

앞 절에서는 관심변수가 연간 판매액인 경우 선박톤수를 조정층 구성 변수로 사용해 예측평균에 가까운 조정층을 구성했다. 이런 과정을 통해 만들어진 조정층을 선박톤수와 별 관계가 없는 양식면적이나 비동거 종사자수(이하 종사자수) 등에 대한 모집단 추정문제에서 무응답 조정 가중치 산출을 위해 사용하는 경우를 고려해 보기로 한다. 여기서는 표 3의 내용 중 전체 무응답률이 30%인 경우, 조정층 구성방법 중 주요 관심대상인 [3], [6] 및 [9]에 대한 분석결과만을 비교해 보기로 한다(모의실험결과 무응답률이 20%인 경우에도 큰차이가 없었음). 관심변수가 종사자수인 경우와 양식면적인 경우 제시된 조정층 구성 방법을 적용해 무응답 가중치 조정을 하는 경우 발생하는 편향과 RMSE를 정리해 보면 각각 표 4 및 5와 같다.

표 4의 결과를 보면 응답성향을 나타내는 어업형태에 따른 조정층은 편향을 줄이는 데 매우 효과적이지만 RMSE를 줄이지는 못한다는 점을 다시 한 번 확인할 수 있다. 관심변수가 종사자수인 경우에도 앞에서 예측평균에 의한 조정층 구성에 사용되었던 선박톤수에 의한 조정층 구성이 상당히 효과적임을 볼 수 있다. 어업형태와 선박톤수를 교차분류해 사용하는 경우 선박톤수를 사용하는 조정층과 큰 차이가 없는 것으로 나타났다(효율성 비교를 쉽게 하기 위해 표 4의 결과는 원래 변수에 1,000을 곱한 것임).

한편 양식면적을 관심변수로 설정하는 경우, 표 5와 같이 이번에도 어업형태에 따른 조정층은 편향을 줄이는 데 매우 효과적이라는 것을 확인할 수 있으며, 이 경우 상대적으로 RMSE도 일부 줄여주는 것으로 보인다. 여기서 특히 유의할 사항은 관심변수가 연간 판매액 또는 종사자수인 경우 효과적인 것으로 나타났던 선박톤수에 의한 조정층을 사용하는 경우, 오히려 무응답을 무시하는 경우에 비해 편향이 대폭 증가하는 특이한 현상을 볼 수 있다. 이런 현상은 무응답 처리에 있어서 조정층 구성을 상당히 신중하게 결정할 필요가 있다는 것을 시사해 준다. 하지만 어업형태와 선박톤수를 동시에 사용하는 경우 이런 문제를 상당 부분 해결할 수 있다는 것을 볼 수 있다. 특히 표 5와 같이 어떤 변수에 대해서

표 4: 관심변수가 종사자수인 경우 조정층 구성에 따른 편향과 RMSE

조정층 구성 변수	전체 무응답률 30%					
	패턴 ④		패턴 ⑤		패턴 ⑥	
	bias	RMSE	bias	RMSE	bias	RMSE
[1] 무응답 무시	−10.04	82.61	−28.00	84.38	−52.36	94.22
[3] 어업형태	−0.07	83.37	−2.30	82.96	−0.86	85.60
[6] 선박톤수	−1.07	78.49	−4.01	79.35	−3.64	82.52
[9] 어업형태*선박톤수	−1.24	78.80	−3.38	79.48	−2.13	82.21

표 5: 관심변수가 양식면적인 경우 조정층 구성에 따른 편향과 RMSE

조정층 구성 변수	전체 무응답률 30%					
	패턴 ④		패턴 ⑤		패턴 ⑥	
	bias	RMSE	bias	RMSE	bias	RMSE
[1] 무응답 무시	18.92	912.97	28.29	920.40	58.57	910.74
[3] 어업형태	2.81	909.93	−6.62	915.85	−11.28	902.06
[6] 선박톤수	83.91	924.01	206.69	964.41	458.74	1052.41
[9] 어업형태*선박톤수	−6.56	911.05	−16.43	918.97	−16.93	907.69

예측평균을 설명하는데 매우 효과적인 변수로 조정층을 구성하는 경우, 다른 관심변수에 대해서는 무응답 조정 때문에 오히려 오차가 늘어날 수도 있다는 점에 유의할 필요가 있다. 이런 문제점을 해결하기 위해서는 응답성향과 예측평균을 설명하는 변수를 함께 사용하는 것이 효과적임을 알 수 있다. 다시 말해 응답성향과 예측평균을 모두 고려한 조정층을 사용하는 것이 이중 로버스트성 관점에서 상당히 바람직한 것으로 판단된다.

4. 결론

단위무응답 가중치 조정을 효과적으로 하기 위해 조정층 구성과정에서 어떤 점들이 고려될 필요가 있는지 살펴보았다. 물론 제시된 제한적인 연구결과를 통해 이론적으로 완성된 해답을 얻을 수 있는 것은 아니다. 하지만 최소한 조사자료를 실제 분석할 때, 조정층을 구성하고 무응답 조정 상수를 산출해 가중치를 조정하는 과정에서 항상 고려해야 할 사항들이 무엇인지 실증자료를 통한 모의실험을 통해 살펴보았다. 본 연구를 통해 얻은 결론을 정리해 보면 다음과 같다.

첫째, 응답성향만을 고려해 조정층을 구성하는 경우 편향을 줄이는 데는 효과적이지만 분산까지 감소시켜 전반적인 효율성을 향상시키기는 어려운 것으로 판단된다. 반면에 예측평균을 고려해 조정층을 사용하는 경우 편향을 줄이는 동시에 분산도 줄일 수 있어서 상당히 효율적인 것으로 결론 내릴 수 있다.

둘째, 예측평균 조정층을 기초로 한 무응답 가중치 조정방법은 관심변수가 바뀌면 조정층을 다시 구성해야 한다는 한계를 갖고 있다. 이런 문제점은 실제 조사자료에 대한 무응답 처리에 있어서 현실적으로 매우 큰 제약조건이 될 수 있다. 하지만 예측평균 및 응답성향을 동시에 고려한 조정층을 구성하는 경우 관심변수가 달라지더라도 상당히 로버스트한 결과를 보여주기 때문에 이런 측면을 실제 무응답 가중치 처리 과정에서 적극적으로 활용할 필요가 있다. 하지만 이런 경우 너무 많은 수의 변수를 사용해 조정층을 구성하는 경우 조정층의 수가 과다해서 일부 조정층에 포함되는 표본의 수가 적어지기 때문에 오히려 무응답 가중치 조정의 효율이 떨어질 수 있다는 점에 유의해야 한다.

셋째, 일반적으로 자체가중(self-weighted) 표본이 아닌 경우 조정층내에서 무응답 조정 상수를 산출하기 위해 설계가중치를 반드시 적용하는 것이 필요하다는 기존의 주장은 타당성이 없는 것으로 보인다. 실제 모의실험을 통해 살펴본 결과 무응답 조정 상수 산출과정에서 설계가중치 부여 여부가 결

과에 미치는 영향은 극히 미미하기 때문에 실무적인 입장에서는 경우에 따라 단순한 형식으로 응답률을 산출하여도 문제가 없는 것으로 판단된다.

주요 관심변수가 여러 개인 경우 각 관심변수별로 예측평균에 의한 무응답 조정층이 달라져야 한다. 하지만 관심변수별로 별도의 예측평균에 따른 조정층을 구성하는 것은 실용적이지 못하고 많은 노력을 필요로 한다. 따라서 본 연구에서는 다루지 못했지만 이런 경우 인자분석이나 주성분분석 등의 다변량 기법을 이용해 주요 관심변수들을 동시에 설명할 수 있는 변수를 만들어 예측평균 조정층을 구성하는 방안을 연구해 볼 필요가 있다. 향후 이런 방법이 구현될 수 있으면 실제 적용이 간편하고 매우 효과적인 무응답 가중치 조정이 가능할 것으로 기대된다.

참고 문헌

김영원, 조선경 (1996). 표본조사에서 항목 무응답 대체 방법, <한국통계학회논문집>, **3**, 145–159.

류제복, 김영원, 박진우 (2002). < 2002년 어가경제조사 표본설계>, 한국통계학회.

Chapman, D. W., Bailey, L. and Kasprzyk, D. (1986). Nonreponse adjustment procedures at the U. S. Bureau of the Census, *Survey Methodology*, **12**, 161–180.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data, *Survey Methodology*, **12**, 1–16.

Kish, L. (1992). Weighting for unequal, *Journal of Official Statistics,* **8**, 183–200.

Little, R. J. (1986). Survey nonresponse adjustments for estimates of means, *International Statistical Review,* **54**, 139–157.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd Edition, John wiley & Sons, New York.

Little, R. J. and Vartivarian, S. (2003). On weighting the rates in non-response weights, *Statistics in Medicine,* **22**, 1589–1599.

Little, R. J. and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means?, *Survey Methodology,* **31**, 161–168.

Vartivarian, S. and Little, R. J. (2002). On the formation of weighting adjustment cells for unit nonresponse, In *Proceedings of the Survey Research Methods Section,* ASA.

Vartivarian, S. and Little, R. J. (2003). Weighting adjustments for unit nonresponse with multiple outcome variables, The University of Michigan, *Department of Biostatistics Working Paper Series*.

2008년 11월 접수; 2008년 12월 채택

Forming Weighting Adjustment Cells for Unit-Nonresponse in Sample Surveys

Young-Won Kim[1,a], Si-Ju Nam[a]

[a]Dept. of Statistics, Sookmyung Women's Univ.

Abstract

Weighting is a common form of unit nonresponse adjustment in sample surveys where entire questionnaires are missing due to noncontact or refusal to participate. A common approach computes the response weight as the inverse of the response rate within adjustment cells based on covariate information. In this paper, we consider the efficiency and robustness of nonresponse weight adjustment based on the response propensity and predictive mean. In the simulation study based on 2000 Fishry Census in Korea, the root mean squared errors for assessing the various ways of forming nonresponse adjustment cells are investigated. The simulation result suggest that the most important feature of variables for inclusion in weighting adjustment is that they are predictive of survey outcomes. Though useful, prediction of the propensity to response is a secondary. Also the result suggest that adjustment cells based on joint classification by the response propensity and predictor of the outcomes is productive.

Keywords: Fishery census, bias, nonresponce adjustment cell, nonresponce weight, unbiased estimator.

This Research was supported by the Sookmyung Women's University Research Grant 2007.

[1] Corresponding author: Professor, Department of Statistics, Sookmyung Women's University, Hyochangwon-Gil 52, Younsan-Gu, Seoul 140-742, Korea. E-mail: ywkim@sm.ac.kr