The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.3.375
32(3), 375–389

Composite estimation type weighting adjustment for bias reduction of non-continuous response group in panel survey

Hyunga Choi[a] · Youngwon Kim[b,1]

[a]Employment Statistics Survey Team, Korea Employment Information Service;
[b]Department of Statistics, Sookmyung Women's University

(Received April 9, 2019; Revised April 23, 2019; Accepted April 23, 2019)

Abstract

Sample attrition according to a long-term tracking reduces the representativeness of the sample data in a panel study. Most panel surveys in South Korea and other countries have prepared response adjustment weights in order to solve problems regarding representativeness due to sample attrition. In this paper, we divided the panel data into continuous response group and non-continuous response group according to response patterns and considered a weighting adjustment method to reduce the bias of the non-continuous response group. A simulation indicated that the proposed composite estimation type weighting method, which reflected the characteristics of non-continuous response groups, could be more efficient than other weighting methods in terms of reducing non-response bias. As a case study, the proposed methods are applied to the Korean Longitudinal Study of Ageing (KLoSA) data of the Korea Employment Information Service.

Keywords: panel survey, weight, composite estimation type weight, non-continuous response group

1. 서론

매년 혹은 반기 등 일정 주기로 동일응답자에 대해 추적조사가 진행되는 패널조사의 경우, 최초 모집단을 대표했던 원표본은 장기 추적에 따른 표본 이탈로 인해 특정 시점에서의 관심변수 추정 시 편향이 생기기 쉽다. 이러한 문제에 대한 해결책으로 국내·외 거의 모든 패널 조사에서 가중값을 사용하고 있다.

본 논문에서는 패널조사의 원표본에서 응답 패턴에 기반하여 연속해서 조사에 참여한 연속응답(continuous response) 그룹과 조사 비참여가 있었던 비연속 응답(non-continuous response) 그룹으로 구분하고, 비연속 응답 그룹을 중심으로 전체 편향을 줄일 수 있는 가중값 산출방법에 대해 다룬다. 이를 위해 2장에서는 기존 국내·외 패널조사에서의 횡단 가중값 산출방법 및 비연속 응답 그룹에 대한 가중값 부여 방법에 대해 살펴보고, 3장에서는 비연속 응답 그룹을 중심으로 편향을 줄일 수 있는 가중값 산출을 위한 새로운 방법을 제안한다. 4장에서는 모의실험을 통해 횡단 가중값 작성에 있어서 기존방법과 새로운 방법의 효율성을 편향을 중심으로 비교하고, 5장에서는 실제로 한국고용정보원의 고령화연구패널(Korean Longitudinal Study of Aging; KLoSA) 자료에 4장에서 제시한 방법들을 적용한 결과를 보여줄 것이다.

[1]Corresponding author: Department of Statistics, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-Gu, Seoul 04310, Korea. E-mail: ywkim@sookmyung.ac.kr

Wave 1	Wave 2	Wave 3	Wave 4 ~ t-1	Wave t	Type of respondent at wave t	Group	Notes
○	○	○	○	○	①	Continuous response Group	continuous responders up to t-1
○	○	○	○	×			
○	○	×	···	×		Non-continuous response Group	complementary set of continuous group
○	○	×	···	○	②		
○	×	○	···	○			
○	×	×	···	○			
○	×	×	···	×			
○	×	○	···	×			

Figure 2.1. Continuous and non-continuous response groups at wave t.

2. 비연속 응답 및 횡단 가중값 작성방법

패널조사에서는 고정된 동일표본에 대해 주기적으로 추적조사가 진행된다. 그에 따라 응답 패턴이 발생하는데, 계속해서 조사에 참여하는 그룹과 간헐적으로 응답에 참여하는 그룹으로 크게 나눌 수 있다. 본 논문에서는 이 두 그룹의 응답성향이 같지 않다고 보고 두 그룹을 구분하기로 한다. t 시점에서의 그룹 구분은 $t-1$ 시점까지의 응답 패턴에 따른다. 즉, $t-1$ 시점까지 연속하여 모두 응답한 경우는 연속 응답 그룹, 그 외의 응답 패턴은 비연속 응답 그룹으로 나눌 수 있고, t 시점 가중값 부여 대상 응답자는 Figure 2.1의 ①, ②와 같이 연속/비연속 응답자로 구분될 수 있다.

t 시점에서 종단과 횡단 가중값 작성 대상을 살펴보면, 종단 가중값은 위의 응답 패턴 그림에서 ① 응답자에게 무응답자 보정을 한 가중값을 부여한다. 즉, $t-1$ 시점까지 계속해서 응답한 표본을 대상으로 일반적으로 작성하게 된다. 반면 횡단 가중값은 응답자 ①, ② 모두에게 가중값을 부여한다. 일반적인 횡단 가중값 작성 과정은 t 시점에서 무응답 보정을 한 뒤, 사후보정 단계를 거친다. 본 연구에서는 사후보정 과정은 논외로 하고, 비연속 응답 그룹의 가중값 산출방법에 대해 초점을 두고자 한다. 기존 연구에서 횡단 가중값 작성 대상인 연속 응답자와 비연속 응답자에게 가중값을 어떻게 부여했는지 살펴보면 다음과 같다.

한국청소년패널(Korean Youth Panel Survey; KYPS)에서는 $t-1$ 시점 대비 t 시점에서의 응답률의 역수를 $t-1$ 시점의 횡단 가중값에 곱하여 산출한다. 이때 $t-1$시점의 횡단 가중값이 없는 비연속 응답자는 가장 최근 횡단 가중값을 사용하였다 (Park 등, 2011). 청년패널(Youth Panel; YP)에서는 연속/비연속 응답자 전체에 대해 1차 변수를 이용하여 t 시점에서 로지스틱 회귀 모형으로 응답률을 추정한 뒤, 그 역수를 1차 설계가중값에 곱하여 횡단 가중값을 산출하였다 (Park과 Kim, 2013; Shin 등, 2017). 고령화연구패널에서는 동일 거주지역 내 성과 연령대가 같은 경우 연속 응답자와 비연속 응답자의 가중값은 같다고 가정하고, 이런 가정에 따라 비연속 응답자에게 그들과 거주지역, 성, 연령대가 같은 연속 응답자들의 종단 가중값의 평균을 부여하고 있다 (Kim 등, 2015).

영국의 고령화패널(English Longitudinal Study of Ageing; ELAS)에서는 t 시점에서의 응답 여부에 응답 그룹 간 차이가 있다고 보고, 연속/비연속 응답 그룹 지시변수를 설명변수에 포함하여 응답모형에 영향을 주는 변수를 찾아 활용하였다. 거주권, 인종, 교육수준, 혼인상태, HSE 참여 여부 등 5개의 변수에 대해 t 시점의 연속/비연속 전체 응답자의 분포와 연속 응답자의 종단 가중값이 반영된 분포가 일치되도록 확대승수를 구하는 방식을 사용하였다 (Shaun 등, 2008). ELSA는 3, 4, 6, 7, 9차에 신규 코

호트가 추가되었지만, 고정표본에 대한 비연속 응답자들에 대한 횡단 가중값의 기저 값을 살펴보면, 6차에서는 가장 최근 종단 가중값을 사용하였지만, 8차는 가중값을 1로 설정하는 등 차수별로 가중값 작성 방식에 차이가 있다 (Shaun 등, 2008; Sally 등, 2015; Josiane 등, 2018).

국내에서 가장 오래된 가구 단위 패널조사는 1998년 시작된 한국노동패널조사(Korean Labor and Income Panel Study; KLIPS)로, t 시점의 횡단 가중값 산출 방식을 살펴보면, t 시점의 무응답 보정은 $t-1$ 시점의 응답자 기준 로지스틱 회귀모형을 통해 개인 단위로 응답률을 추정한다. 추정된 응답률의 역수를 $t-1$ 시점의 종단 가중값에 곱하여 최종 가구원별 종단 가중값을 만든다. 비연속 응답자의 경우 $t-1$ 시점의 종단 가중값이 없어서, 이들에 대해서는 가장 최근 개인 종단 가중값을 가져와 사용한다. 이를 토대로 가구 내 가구원들의 종단 가중값을 합산하여 전체 가구원 수로 나누어 구한 평균을 가구 가중값으로 사용한다. 이때 신규 진입한 원가구원은 전체 가구원수 산출 시 제외하고 비원가구원 가중값은 0으로 놓고 전체 가구원수에는 포함한다. 이렇게 산출된 가구 가중값을 t 시점에서 비원가구원을 포함한 개인 응답자 모두에게 적용한다 (Park 등, 2013; Baek과 Shim, 2012).

한국재정패널의 경우 가구유형별 적정 횡단 가중값 산출 방식이 다르다. 가구유형으로는 원가구, 대체가구, 분가가구로 나누어지며, 원가구와 대체가구에 대해서는 로지스틱 회귀를 활용하여 추정된 응답률의 역수를 1차 설계 가중값에 곱하여 종단 가중값을 산출하였고, 이때 2차부터 조사된 대체가구에는 대응되는 원패널 가구의 지역, 가구주 성, 가구주 연령, 가구원수, 가구 연간 소득 총액을 사용하였다. 분가가구에 대해서는 분가 사유에 따라 '결혼'으로 분가한 경우 앞서 작성된 가구 가중값의 1/2을, 그 외의 사유로 분가한 경우는 원가구의 가중값을 그대로 활용하여 횡단 가구 가중값을 산출하였다. 이렇게 작성된 횡단 가구 가중값을 소득이 있는 15세 이상 가구원에게 부여하였다 (Korea Institute of Public Finance, 2012).

한국아동패널의 경우 횡단 가중값 작성은 무응답 가구의 보정을 위해 가구와 어머니 특성을 고려한 응답모형을 적용하였다. 이때 1차 설명변수를 활용하여 1차 대비 t차 응답 여부에 대한 응답모형을 적합 시켰다. 아동패널의 경우 표본수가 크지 않아 응답률 추정 시 로지스틱 모형을 직접 활용하지 않고, 응답 여부에 영향을 주는 요인을 선정하는 과정에서만 활용하였다. 여기서 걸러진 주요 요인으로 무응답 조정 계급(non-response adjustment class)을 만들어 무응답 보정을 하였다. 주 요인으로는 거주지 권역, 출생순위, 어머니 취업여부가 사용되었다. 결과적으로 20–24개 층을 구성하고, 층별 가중응답률의 역수를 설계가중값에 곱하는 방식으로 무응답 보정 가중값을 산출하였다 (Lee 등, 2011).

3. 비연속 응답 그룹에 대한 적정가중값 산출 방법

3.1. 고령화 연구패널에서의 기존 횡단 가중값 산출방법

대부분 패널조사에서는 연속 응답 그룹만을 대상으로 종단 가중값을 작성하지만, 횡단 가중값은 비연속 응답 그룹을 포함한 해당 시점의 모든 응답자를 대상으로 작성한다. 앞장에서 살펴본 선행연구들은 연속과 비연속 응답자 그룹에 대한 구별 없이 가중값을 산출한다. 청년패널, 재정패널, 청소년패널의 경우 1차 대비 t 시점 혹은 $t-1$ 시점 대비 t 시점에서의 전체 응답자들(연속/비연속)에 대해 응답률을 추정해 가중값을 산출하고, 아동패널이나 영국고령화패널의 경우 응답 여부에 영향을 주는 요인으로 무응답 조정 층을 만들어서 무응답 보정을 하고 있으며, 고령화연구패널에서는 특정 변인으로 층을 구분하여 각 층에서의 연속/비연속 응답자 특성이 같다고 가정하여 가중값을 산출한다. 이들 모든 경우 결과적으로 연속 응답자와 비연속 응답자 그룹 간에 응답 특성에 있어서 차이를 없다는 것을 가정하는 것이다.

실제로 패널조사에서 연속 응답 그룹과 비연속 응답 그룹의 응답자 특성이 같다고 볼 수 있는가? 이런 질문에 대한 해답을 고령화연구패널 자료를 갖고 찾아보기로 한다. 응답 패턴에 따라 t 시점에서의 연속

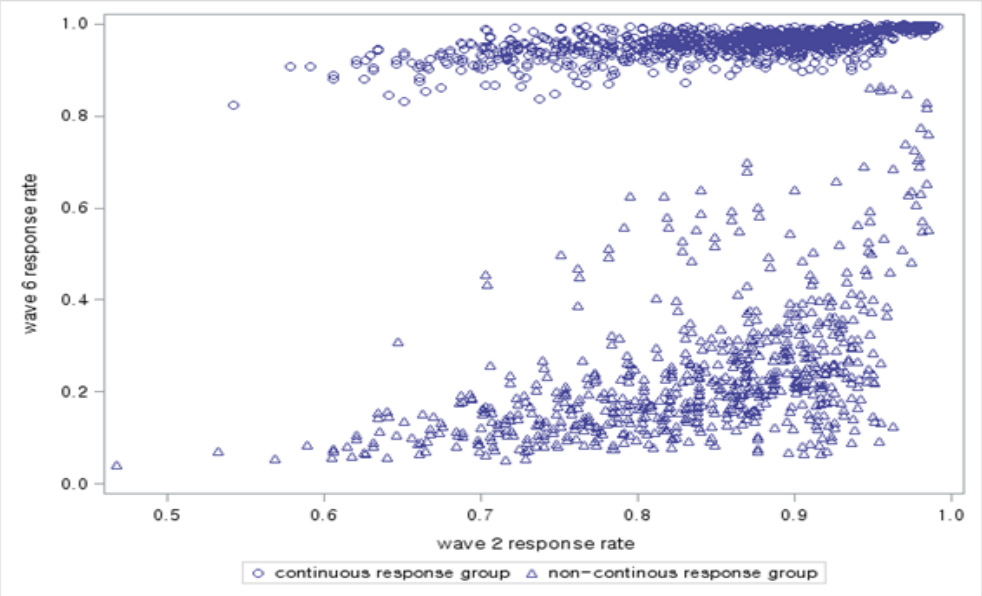


Figure 3.1. Wave t response rate versus wave 2 response rate by group.

응답 그룹과 비연속 응답 그룹은 다음과 같이 표기하자.

$$\text{연속 응답 그룹: } S_{\text{con},t} = \{i : R_2 = 1, \ldots, R_{t-1} = 1, i \in S\}; \qquad \text{비연속 응답 그룹: } S^c_{\text{con},t},$$

여기서 R_t는 t 시점에서 응답 여부를 나타내는 지시변수로 응답은 '1', 무응답은 '0'이다. 1차 조사의 전체표본은 모두 응답 그룹에 해당하므로, 연속/비연속 응답 그룹 구분은 2차(R_2)부터 응답 여부에 따라 구분된다. 따라서, t 시점에서 연속 응답 그룹은 응답자 중 $t-1$ 시점까지 모든 시점에 연속 응답한 경우를 말하며, 그 외 응답자는 비연속 응답 그룹으로 본다. Figure 3.1은 고령화연구패널에서 2차 조사 응답률 대비 6차 조사(t 시점)에서의 응답률을 나타낸 것으로, 산점도를 보면 연속/비연속 응답 그룹에 따라 응답 패턴이 명확히 구분된다. 연속 응답 그룹은 6차에서 응답률이 0.8–1 사이에 분포되고, 비연속 응답 그룹의 응답률은 0.4 아래에 집중되어 있다. 여기서 응답률은 고령화연구패널에서 무응답 보정을 위해 사용하고 있는 로지스틱 회귀모형에 연속/비연속 그룹을 나타내는 지시변수를 추가해 산출한 것이다.

또한 주요 관심변수 관점에서 연속/비연속 응답 그룹별로 차이가 있는지 살펴보기 위해 몇가지 관심변수의 평균을 비교해 보았다. 여기서 각 차수별 관심변수 값은 무응답자의 경우 관측되지 않기 때문에 모든 응답자에 대한 관측값이 존재하는 1차 조사에서 해당 변수 관측값의 평균으로 비교하기로 한다. 연속/비연속 그룹 구분은 $t-1$차까지의 응답 패턴을 기준으로 정해지기 때문에, Figure 3.2는 그룹 비교가 가능한 3차부터 6차까지 고혈압 유병률, 대졸자 비율, 아파트 거주비율, 자가 비율, 연간 개인 총소득, 나이를 비교한 것이다.

Figure 3.2를 보면 모든 변수에서 연속/비연속 응답 그룹 간에 차이가 있다는 것을 볼 수 있다. 고혈압 유병률과 자가 비율, 평균 나이는 비연속 응답 그룹이 연속 응답 그룹의 평균보다 낮았다. 그 외 대졸자 비율, 아파트 거주비율, 연간 개인 총소득은 비연속 응답 그룹에서 높았다. 그룹별 평균은 차수가 진행되면서 일정 수준의 차이를 그대로 유지하거나 초기 격차보다 더 벌어지는 현상을 보여주고 있다. 제시

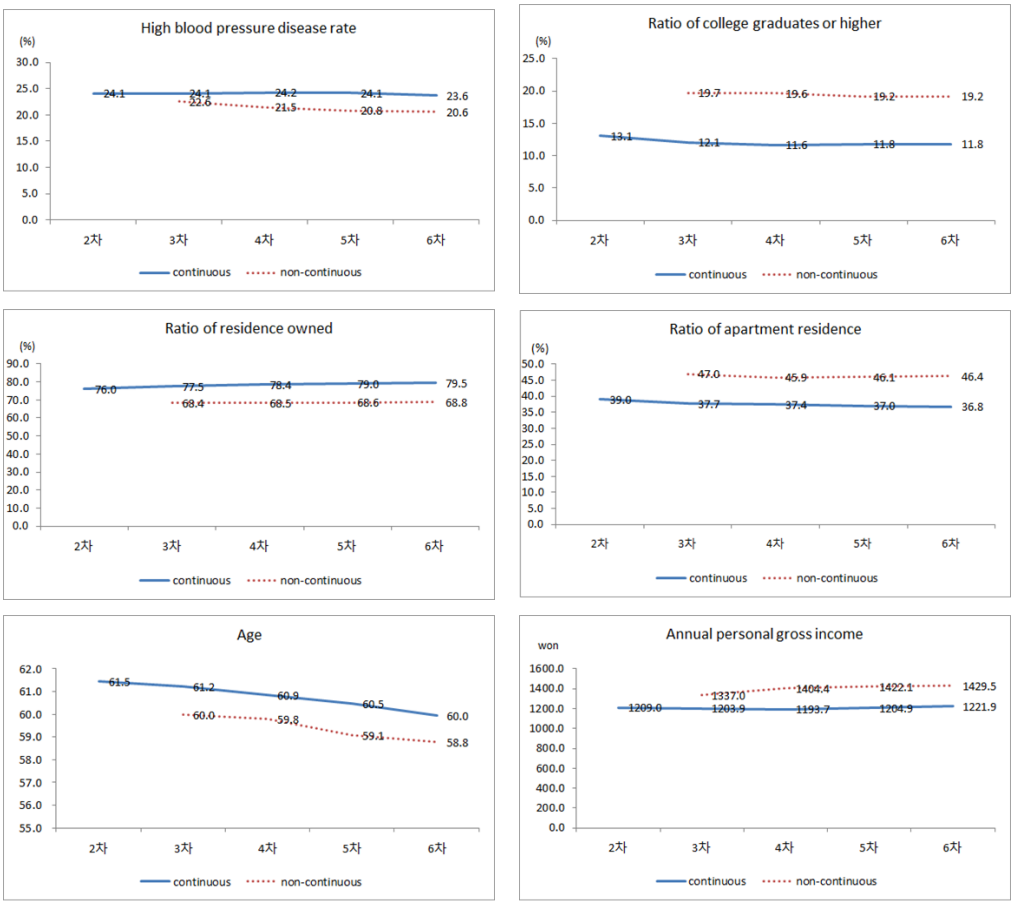


Figure 3.2. Difference of mean by continuous and non-continuous response group.

된 Figure 3.1과 Figure 3.2를 통해 연속/비연속 응답자 그룹 간에는 특성상 뚜렷한 차이가 있음을 확인할 수 있다.

3.2. 새로운 복합 가중값 산출방법

패널조사에서 보통 비연속 응답에 해당하는 표본 수는 적은데, 이렇게 표본 수가 적을 때 가중값 산출을 위해 별도의 무응답 모형을 적합하면 신뢰도는 상당히 떨어질 수 있다. 따라서 대부분의 패널조사에서 비연속 응답자들에 대한 횡단 가중값 추정을 별도로 하지 않고 있다. 그러나 추적 기간이 길어짐에 따라 비연속 응답 그룹에 해당하는 표본 수는 어느 정도 증가하게 되고, 더욱이 연속/비연속 응답 그룹 간 특성이 다르다면, 편향 문제는 점점 더 심각해질 수 있다. 따라서 이런 편향을 줄이기 위한 비연속 응답자에 대한 적정 가중값 작성방법에 관한 연구가 필요하다. 본 연구에서 새롭게 제안하는 방법은 소지역 추정에 자주 활용되는 복합추정량(composite estimator) 방식의 가중값이다. ‘borrow strength’라고도 불리며 소지역 관심변수 추정에 있어 적은 표본 수로 인한 직접 추정량(direct estimator)이 갖는 한계를 극복하고자, 외부 행정자료나 관련 조사자료로부터 동일 관심변수에 대한 합성추정량(synthetic estimator)을 가져와 직접추정량과 결합한 형태로 추정하는 것을 말한다 (Takis, 2010). 즉, 비연속 응

답 그룹의 응답률이 대체로 낮아 표본 수가 적기 때문에 직접추정 방식으로 추정된 응답률은 가중값 작성 시 극단 가중값을 산출하게 된다. 따라서, 본 연구에서는 비연속 응답 그룹에 대해 소지역(small area) 추정과 같은 접근 방식을 통해 그룹의 특성을 반영하면서 극단 가중값을 줄이는 무응답 보정 방식을 제시한다.

합성추정량으로는 연속응답 그룹(S_{con})에서 추정된 응답모형의 회귀계수 값을 활용하여 비연속 응답 그룹(S^c_{con})에 속하는 단위의 응답률을 추정($\hat{p}^s_{i,S^c_{\text{con}},t}$)한 것을 사용하고, 직접추정량으로는 비연속응답 그룹만을 대상으로 로지스틱 회귀모형을 적용하여 추정한 응답률($\hat{p}^d_{i,S^c_{\text{con}},t}$)을 사용한다. 비연속그룹 내 각 단위에 대해 산출된 2개의 응답률을 결합하여 복합추정 응답률($\hat{p}^c_{i,S^c_{\text{con}},t}$)을 구하고, 응답률의 역수를 설계가중값에 곱해 비연속그룹 응답자의 횡단 가중값을 산출하는 방식이다. 여기서, $\hat{\beta}_{S^c_{\text{con}},t}$는 비연속응답 그룹의 응답모형으로 추정된 회귀계수, $\hat{\beta}_{S_{\text{con}},t}$은 연속응답 그룹에서 추정된 회귀계수를 말한다.

· 비연속응답 그룹 직접추정 응답률($\hat{p}^d_{i,S^c_{\text{con}},t}$)

$$\hat{p}^d_{i,S^c_{\text{con}},t} = \frac{\exp\left(X'_{S^c_{\text{con}}}\hat{\beta}_{S^c_{\text{con}},t}\right)}{1+\exp\left(X'_{S^c_{\text{con}}}\hat{\beta}_{S^c_{\text{con}},t}\right)}. \tag{3.1}$$

· 비연속응답 그룹 합성추정 응답률($\hat{p}^s_{i,S^c_{\text{con}},t}$)

$$\hat{p}^s_{i,S^c_{\text{con}},t} = \frac{\exp\left(X'_{S^c_{\text{con}}}\hat{\beta}_{S_{\text{con}},t}\right)}{1+\exp\left(X'_{S^c_{\text{con}}}\hat{\beta}_{S_{\text{con}},t}\right)}. \tag{3.2}$$

· 비연속응답 그룹 복합추정 응답률($\hat{p}^c_{i,S^c_{\text{con}},t}$)

$$\hat{p}^c_{i,S^c_{\text{con}},t} = \alpha\hat{p}^s_{i,S^c_{\text{con}},t} + (1-\alpha)\hat{p}^d_{i,S^c_{\text{con}},t}, \tag{3.3}$$

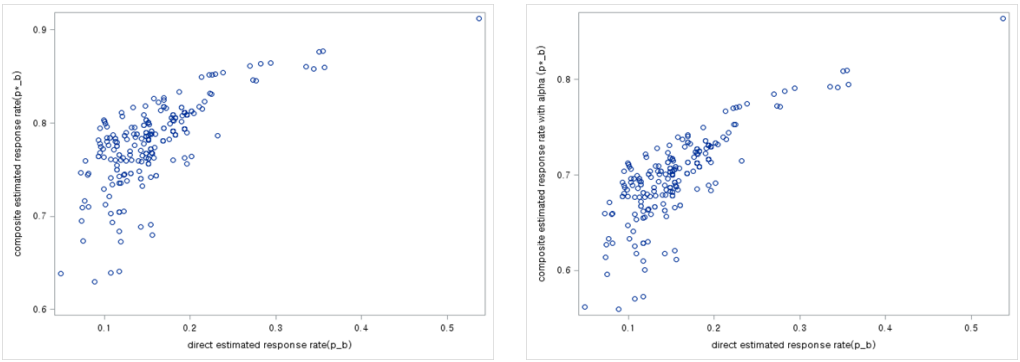
여기서 α는 두 추정량을 결합하기 위한 가중값이다. 본 연구에서는 그룹별 표본크기와 평균제곱오차(mean square error; mse)를 사용한 다음 두 가지 방법을 이용한다.

$$\alpha_1 = \frac{n_{S_{\text{con}}}}{n_{S_{\text{con}}} + n_{S^c_{\text{con}}}}, \qquad \alpha_2 = \frac{\text{mse}_{S^c_{\text{con}}}}{\text{mse}_{S_{\text{con}}} + \text{mse}_{S^c_{\text{con}}}}.$$

먼저 복합추정 방식을 사용함에 따라 직접추정 방식과 비교해 가중값에 어떤 변화가 있을지 살펴보았다. Figure 3.3은 비연속 응답자의 직접추정과 복합추정으로 산출한 응답률의 산점도이다. 여기서 복합추정은 결합가중값 α에 따라 두 가지 방법으로 추정했다. 산점도를 보면 x축이 직접추정, y축이 복합추정 응답률을 나타내는데, 직접추정 응답률은 모두 0.5 이하로 추정되었지만, 복합추정 응답률은 0.5–0.9 수준으로 직접추정에 비해 증가하는 것을 볼 수 있다. 일반적으로 비연속 응답자의 경우 직접추정 가중값을 사용하면 극단 가중값이 과다하게 발생하기 때문에 기존 연구에서는 비연속 응답 그룹을 연속응답 그룹과 구분하지 않고 가중값을 작성해왔다. 결과적으로 복합추정 방식을 통해 비연속 응답자에 대한 극단 가중값 발생 가능성을 줄이고 동시에 편향도 줄일 수 있다면 제시한 가중값 산출방식은 패널조사에서 보다 안정적으로 가중값을 산출할 수 있는 효과적인 방법이라 할 수 있다.

3.3. 비연속응답 그룹에 대한 다양한 형태의 가중값 작성방법

횡단 가중값 산출방법과 관련하여 본 연구에서 제안한 복합추정 방식과 기존 연구에서 적용된 가중값 산출방법들을 비교하고자 한다. 또한, 기존에 사용하던 가중값 작성방법은 아니지만 검토 가능한 가중값 작성방법들도 함께 정리하였다.



(a) combined weight α: sample size (b) combined weight α: mse

Figure 3.3. Scatter plot of direct estimate (p_b) and composite estimate (p^*_b) for response rate of non-continuous response group (Data: KLoSA, KEIS).

· [방법 1] 성별, 연령대, 거주지역 등 동일한 범주의 연속 응답자와 비연속 응답자의 가중값은 같다고 가정하는 방법이다. 비연속 응답자에 대한 적정 가중값은 범주가 동일한 연속 응답자의 평균 종단 가중값이 된다. 이 방법은 Kim 등 (2015)이 고령화연구패널에서 적용한 것이다.

· [방법 2] 기존 패널자료에서 사용한 방법은 아니지만 [방법 1]보다는 좀 약한 가정으로, 동일한 범주의 연속 응답자와 비연속 응답자의 응답률이 같다고 가정하는 방법이다. 비연속 응답자의 응답률은 동일 범주에 속하는 연속응답 그룹의 평균 응답률이 된다.

· [방법 3] 청년패널에서 적용한 방법으로 1차 대비 t 시점 응답률은 연속/비연속 응답 그룹 구분 없이 추정하는 방식이다. 추정된 응답률의 역수를 설계가중값에 곱하여 가중값을 산출한다 (Park과 Kim, 2013).

· [방법 4] 패널조사에서 실제 적용된 방법은 아니지만 극단 가중값을 줄이기 위해 [방법 3]에서 추정된 응답률을 5개의 분위수 층으로 나누고, 층별 무응답 보정을 한다. 층별 확대승수를 설계가중값에 곱하여 가중값을 산출한다 (Qixuan 등, 2012).

· [방법 5] 본 연구에서 새롭게 제안한 방법으로 비연속 응답 그룹의 응답률은 식 (3.1)–(3.3)의 복합추정 방식으로 산출한다. 추정된 응답률의 역수를 설계가중값에 곱하여 비연속 응답자의 가중값을 산출한다. 복합추정에서 결합가중값으로 α_1 사용.

· [방법 6] 새롭게 제안한 [방법 5]의 복합추정에서 결합가중값으로 α_2 사용.

참고로 연속 응답자의 종단 가중값은 무응답 보정으로 인해 재진입한 비연속 응답자까지 이미 반영하고 있기 때문에 연속/비연속 응답 그룹을 구분해 가중값을 작성하는 [방법 5]와 [방법 6]에서는 t 시점의 연속 응답자들의 종단 가중값의 스케일을 조정하는 과정이 필요하다. 이를 위해 t 시점에서 비연속 응답자를 포함한 전체 응답자들을 응답률에 따라 5개의 그룹으로 나누고, 그룹 내에서 재진입한 비연속 응답자의 비중만큼 연속 응답자의 가중값에서 감해주는 작업을 했다. 아울러 [방법 1]–[방법 6]의 모든 경우 가중값을 작성한 뒤, 작성된 가중값 총합은 원표본 가중값 총합과 일치되도록 스케일 조정을 하였다.

4. 모의실험

4.1. 모의실험 설정

모의실험을 통해 앞에 정리한 가중값 작성방법들의 효율성을 비교해 보고자 한다. 모의실험을 위해 우

Table 4.1. Scenario of three-year panel data for simulation

Group	Wave 1 response rate	Wave 2 response rate vs wave 1	Wave 3 response rate vs wave 2 (vs wave 1)
Continuous	100%	average response rate 86.0%, around 4,300 samples	average response rate 88.0%, around 3,784 samples
Non-continuous	100%	-	average response rate 30.0%, around 210 samples
Total	100%, 5,000 samples	average response rate vs wave1 86.0%, around 4,300 samples	average response rate vs wave1 80%, around 3,994 samples

선 1차 원표본 5,000명에 대해 차수별로 무응답을 발생시켜 총 3개 시점에 대한 패널자료를 구현하였다. 1차 대비 2차 응답률은 86%, 2차 대비 3차 응답률은 연속 응답 그룹은 88%, 비연속 응답 그룹은 30%가 되도록 설정하였다. 1차 원표본 대비 3차 응답률은 80%이다. 이런 시나리오는 고령화연구패널 표본 마모 및 무응답 패턴에 기초한 것이다. 3차에서 무응답 발생 시 그룹별 응답 특성이 상이하다는 가정에 따라 응답모형을 다르게 설정하였다. 제시된 응답률은 반복 실험에서 평균 응답률로 볼 수 있다 (Brady, 2013).

응답 여부에 영향을 주는 변수로 범주형 변수 x(범주 1, 2, 3), 연속형 변수 z을 고려하였다. 차수별로 x변수는 같다고 가정하였다. z변수는 개인별 응답속성과 관련한 기저(base)이며, 값이 클수록 응답 가능성이 높아지도록 설정하였다. 원표본 5,000명에 대하여 x변수의 범주별 구성비는 20%, 30%, 50%가 되도록 생성하였고, z변수는 지수분포에 로그를 취해 0보다 크고 왼쪽으로 긴 꼬리를 갖는 비대칭 분포를 가정하였다. 응답모형으로는 다음과 같은 로짓모형을 사용하였다.

$$\text{logit}\left(\frac{\hat{p}_{i,t}}{1-\hat{p}_{i,t}}\right) = \beta_0 + \beta_1 x_i + \beta_2 z_i.$$

2차 시점 응답모형에서는 값이 '3' 범주로 갈수록 응답 확률이 높아지도록 하였고, 3차 시점에서 연속 응답 그룹의 경우는 2차와 동일한 모형을 사용하였는데, 다만 응답률 조정을 위해 계수만 다르게 하였다. 구체적으로 보면, 2차 응답모형은 모형에서 x(범주 1, 2, 3)에 따라 β_1은 (0.2, 0.8, 1), β_2는 0.13 가정하였다. 3차 응답모형의 경우 연속 응답 그룹에서는 β_1은 (0.8, 1.2, 1.4), β_2는 0.11, 비연속 응답 그룹에서는 β_1은 $(-1.2, -1.6, -1.8)$, β_2는 0.08로 설정하고, β_0는 연속/비연속 그룹 모두 동일한 정규분포에 따른다고 가정하였다. 여기서 비연속 응답 그룹은 연속 응답 그룹 응답자들 범주 속성과 β_1을 반대로 설정하기 위해 x가 '1'일 때 응답 가능성이 높아지도록 한 것이다.

관심변수로는 이항 변수와 연속형 변수를 고려하였다. 변수 생성 시 연속/비연속 응답 그룹 간 차이를 두었고, 비연속 응답 그룹에서의 관심변수가 연속 응답 그룹보다 큰 값을 갖도록 설정하였다. 관심변수 생성에 따라 다섯 가지 유형을 고려하였는데, 유형5로 갈수록 그룹 간 관심변수의 평균 차이가 증가한다. 유형1(T1)은 그룹별 관심변수 차이를 절편에만 의존하도록 하였고, 설계가중값은 균등하다고 놓은 것이다. 유형2(T2)는 T1에서 관심변수 생성 시 x범주에 의존하도록 하였고, 범주가 '1'일수록 큰 값을 갖도록 하였다. 이때, 그룹별 x범주에 따른 β_1은 같게 놓았다. 유형3(T3)은 T2에서 x범주에 따른 β_1를 그룹별로 달리 적용하였고, 비연속 응답 그룹에서 β_1의 영향이 더 커지게 하였다. 유형4(T4)에서는 T3의 설정에 설계가중값을 x범주별로 불균등하게 하였다. x가 '1'일수록 높은 설계가중값을 갖고, 같은 범주에서 그룹 간 차이는 없다. 유형5(T5)에서는 T4의 불균등 설계가중값을 비연속 응답 그룹에서 더 큰 값을 갖도록 그룹별로도 차이가 있게 설정하였다. 구체적인 내용은 Table 4.2에 정리하였다.

Table 4.2. Generating variable of interest (y_i) for simulation

	Binomial variable $\text{logit}\left(\frac{\hat{\phi}_{i,t}}{1-\hat{\phi}_{i,t}}\right) = \alpha + \beta_1 x_i$	Continuous variable $y_i = \alpha + \beta_1 x_i + e_i$
T1	$\beta_1 = 0$, $\alpha \sim N(0.3, 0.01)$, $i \in S_{\text{con}}$ $\alpha \sim N(0.5, 0.01)$, $i \in S_{\text{con}}^c$	$\beta_1 = 0$, $\alpha \sim N(100, 10)$, $e_i \sim N(0, 10)$, $i \in S_{\text{con}}$ $\alpha \sim N(120, 10)$, $e_i \sim N(0, 10)$, $i \in S_{\text{con}}^c$
T2	$\alpha \sim N(0.1, 0.01)$, $\beta_1 = (0.4, 0.2, 0)$ for $x_i = (1, 2, 3)$, $i \in S_{\text{con}}$ $\alpha \sim N(0.2, 0.01)$, $\beta_1 = (0.4, 0.2, 0)$ for $x_i = (1, 2, 3)$, $i \in S_{\text{con}}^c$	$\alpha \sim N(100, 10)$, $e_i \sim N(0, 10)$, $\beta_1 = (30, 20, 10)$ for $x_i = (1, 2, 3)$, $i \in S_{\text{con}}$ $\alpha \sim N(110, 10)$, $e_i \sim N(0, 10)$, $\beta_1 = (30, 20, 10)$ for $x_i = (1, 2, 3)$, $i \in S_{\text{con}}^c$
T3	$\alpha \sim N(0.1, 0.01)$, $\beta_1 = (0.4, 0.2, 0)$ for $x_i = (1, 2, 3)$, $i \in S_{\text{con}}$ $\alpha \sim N(0.2, 0.01)$, $\beta_1 = (0.6, 0.4, 0)$ for $x_i = (1, 2, 3)$, $i \in S_{\text{con}}^c$	$\alpha \sim N(100, 10)$, $e_i \sim N(0, 10)$, $\beta_1 = (30, 20, 10)$ for $x_i = (1, 2, 3)$, $i \in S_{\text{con}}$ $\alpha \sim N(110, 10)$, $e_i \sim N(0, 10)$, $\beta_1 = (50, 30, 20)$ for $x_i = (1, 2, 3)$, $i \in S_{\text{con}}^c$
T4	Under the same condition as T3, but unequal weight(w_{i1}: design weight), regardless of response group $w_{i1} = (30, 20, 10)$ for $x_i = (1, 2, 3)$	
T5	Under the same condition as T3, but unequal weight by response group(w_{i1}: design weight) $i \in S_{\text{con}}$, $w_{i1} = (30, 20, 10)$ for $x_i = (1, 2, 3)$ $i \in S_{\text{con}}^c$, $w_{i1} = (40, 25, 15)$ for $x_i = (1, 2, 3)$	

한편 가중값 산출방법에 따른 차이를 비교하기 위해 다음의 기준을 사용하였다. 여기서 $\bar{y}_1$은 1차 전체 원표본 자료에서 설계가중값을 적용한 평균(이항변수의 경우 비율)을 나타내고, $\bar{y}^m$은 일부 무응답이 발생한 3차 시점에 응답 자료만을 갖고 [방법 1]–[방법 6]에 따라 무응답 보정 가중값을 구한 후 산출한 평균(이항변수의 경우 비율) 추정값이다. 모의실험 결과는 원표본 크기가 5,000인 자료를 1,000번 반복적으로 생성해 비교한 것이다.

$$\text{편향(bias)} : \frac{1}{n}\sum_{i=1}^{n}\left(\bar{y}_i^m - \bar{y}_1\right), \qquad \text{오차의 절대값(absolute error)} : \frac{1}{n}\sum_{i=1}^{n}\left|\bar{y}_i^m - \bar{y}_1\right|,$$

$$\text{오차의 제곱(squared error)} : \frac{1}{n}\sum_{i=1}^{n}\left(\bar{y}_i^m - \bar{y}_1\right)^2, \qquad \text{상대 오차(relative error)} : \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\bar{y}_i^m - \bar{y}_1\right|}{\bar{y}_1}.$$

4.2. 모의실험 결과

Table 4.3에서 편향을 보면, 모든 경우 [방법 1]–[방법 4]의 가중값을 적용한 경우 유사한 수준의 편향이 발생하지만, 복합추정 방법을 이용한 [방법 5]와 [방법 6]의 경우 다른 방법에 비해 편향이 대폭 감소한다는 것을 볼 수 있다. 오차의 절대값과 편향을 함께 비교해 보면, [방법 1]–[방법 4]에서는 반복 시행에서 항상 일정 수준의 편향이 존재함을 알 수 있다. 하지만 [방법 5]와 [방법 6]의 편향은 [방법 1]–[방법 4]에 비해 반복 시행에서 평균 추정값이 완전자료에서 구한 평균과 가깝게 추정됨을 볼 수 있었다. 연속형 변수에서도 가중값 방법별로 이산형 변수와 동일한 결과를 보였는데, 상대 오차에서 [방법 5], [방법 6]의 결과는 0.1% 수준으로 다른 방법들 대비 매우 작았다.

설명변수가 x에 의존하도록 관심변수를 설정한 T3를 T1과 비교해 보면 다른 방법의 경우 편향과 오차가 대폭 증가하는 데 비해 [방법 5]와 [방법 6]의 경우 편향과 오차가 큰 차이가 없음을 볼 수 있다. 연속형 변수에서의 상대오차도 [방법 5]와 [방법 6]의 경우 T1–T3 모든 경우 0.1% 수준을 그대로 유지하는

Table 4.3. Simulation result

		Binomial variable			Continuous variable			
		Bias	Absolute value of error	Square of error	Bias	Absolute value of error	Square of error	Relative error
T1	[Method 1]	−0.446	0.460	0.286	−1.634	1.634	2.681	0.016
	[Method 2]	−0.490	0.498	0.322	−1.786	1.786	3.200	0.017
	[Method 3]	−0.483	0.493	0.319	−1.641	1.641	2.702	0.016
	[Method 4]	−0.498	0.506	0.332	−1.645	1.645	2.717	0.016
	[Method 5]	−0.059	0.361	0.198	0.049	0.115	0.021	0.001
	[Method 6]	−0.059	0.361	0.199	0.049	0.115	0.021	0.001
T2	[Method 1]	−0.376	0.405	0.233	−0.707	0.707	0.510	0.006
	[Method 2]	−0.451	0.465	0.289	−0.839	0.839	0.713	0.007
	[Method 3]	−0.483	0.494	0.322	−0.920	0.920	0.854	0.008
	[Method 4]	−0.503	0.513	0.342	−0.939	0.939	0.891	0.008
	[Method 5]	−0.007	0.364	0.204	0.002	0.136	0.029	0.001
	[Method 6]	−0.021	0.365	0.205	−0.021	0.136	0.029	0.001
T3	[Method 1]	−0.624	0.629	0.481	−2.647	2.647	7.029	0.022
	[Method 2]	−0.726	0.728	0.615	−2.983	2.983	8.913	0.025
	[Method 3]	−0.734	0.736	0.629	−2.880	2.880	8.313	0.024
	[Method 4]	−0.755	0.756	0.660	−2.908	2.908	8.474	0.024
	[Method 5]	0.012	0.352	0.205	0.121	0.174	0.047	0.001
	[Method 6]	−0.005	0.353	0.206	0.086	0.154	0.038	0.001
T4	[Method 1]	−0.769	0.774	0.731	−2.879	2.879	8.332	0.023
	[Method 2]	−0.916	0.917	0.970	−3.338	3.338	11.177	0.027
	[Method 3]	−0.905	0.906	0.952	−3.156	3.156	9.994	0.025
	[Method 4]	−0.935	0.936	1.006	−3.195	3.195	10.241	0.025
	[Method 5]	−0.045	0.413	0.267	0.072	0.183	0.051	0.001
	[Method 6]	−0.060	0.414	0.268	0.037	0.172	0.046	0.001
T5	[Method 1]	−1.172	1.172	1.513	−4.430	4.430	19.670	0.035
	[Method 2]	−1.133	1.134	1.428	−4.246	4.246	18.073	0.033
	[Method 3]	−1.081	1.081	1.319	−3.957	3.957	15.699	0.031
	[Method 4]	−1.133	1.133	1.434	−4.030	4.030	16.288	0.032
	[Method 5]	0.054	0.470	0.360	0.191	0.250	0.094	0.002
	[Method 6]	0.032	0.466	0.355	0.145	0.222	0.076	0.002

안정적인 결과를 보여주고 있다.

한편 불균등 설계가중값 설정에 따른 영향을 보기 위해 고려한 T4와 T5의 모의실험 결과를 보면, 우선 균등 설계가중값을 가정한 T1–T3에 비해 [방법 1]–[방법 4]의 경우는 편향과 오차가 상당폭 증가하지만 [방법 5]와 [방법 6]의 경우에는 불균등 설계가중값의 경우에도 편향과 오차가 대폭 감소하여 균등 설계가중값의 경우와 큰 차이를 보이지 않는다. 결과적으로 Table 4.3의 모의실험 결과를 통해 복합추정 방식의 가중값 산출방식을 적용한 [방법 5]와 [방법 6]이 어떤 유형의 무응답 모형에서도 [방법 1]–[방법 4]에 비해 패널조사에서 보다 안정적이고 효율적인 무응답 가중값 보정 방법으로 결론을 내릴 수 있다.

5. 실증분석

고령화연구패널 1차-3차-4차년도 자료에 앞서 제시한 가중값 산출방법들을 적용하였다. 여기서는 편의

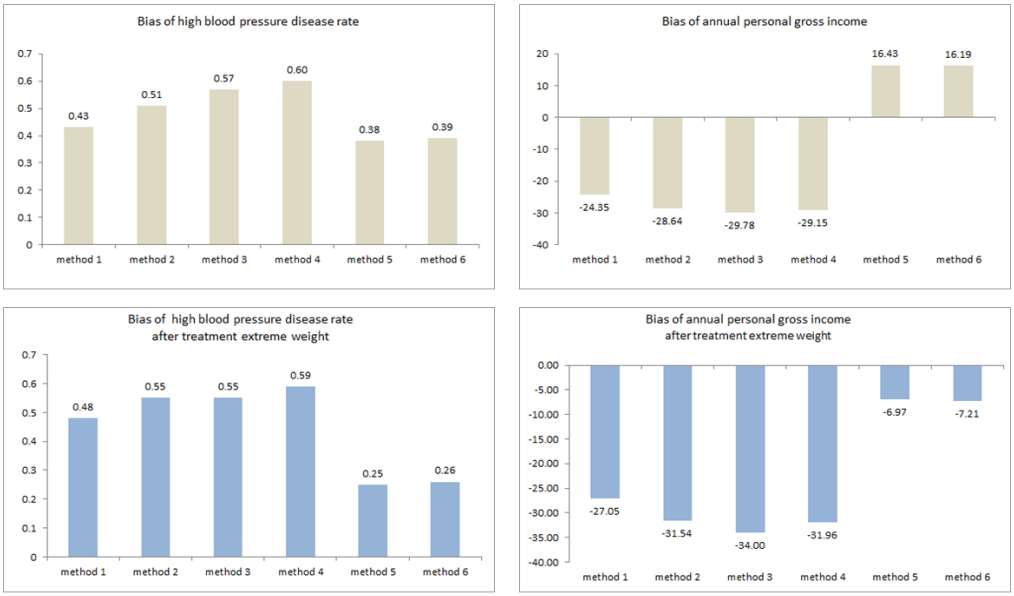


Figure 5.1. Comparison of bias with extreme weight (above) and bias after treatment of extreme weights (below).

상 1차-3차-4차 자료 총 3개년도 패널자료만 존재하는 것처럼 보고 제시된 방법을 적용해 보기로 한다. 고령화연구패널에는 매 차수별 자연적으로 발생하는 사망자가 있고, 이들은 가중값 작성 대상에서 제외된다. 따라서 1차의 10,254명 중 4차 조사에서 사망자에 해당하는 823명을 제외하고 1차 조사 가중값 작성 대상 원표본은 9,431명으로 구성된 것으로 보기로 한다. 1차 대비 2차의 응답률은 81.2%(7,920명), 2차 대비 3차의 연속/비연속 응답 그룹별 응답률은 각각 94.4%, 14.6%였다. 1차 원표본(9,431명) 대비 3차 조사에서 그룹별 응답률은 연속 응답 그룹은 76.6%(7,226명), 비연속 응답 그룹은 2.8%(260명)으로 3차 조사 전체 응답자수는 7,486명이다. 무응답 발생 후 무응답 보정 가중값 산출방법에 따라 1차 조사 원표본의 특성을 얼마나 잘 설명하는지를 살펴보기 위한 관심변수로는 고혈압 유병률과 연간 개인 총소득을 대상으로 분석하였다. 연속/비연속 응답 그룹별 평균은 고혈압 유병률이 각각 24.2%, 21.5%이었고, 연간 개인 총소득은 1143.01만원, 1438.68만원이었다. 추정의 정확성 비교 시 t 시점의 모집단 참값을 알 수 없기 때문에, 여기서는 1차 원표본을 기준으로 전체 표본에서의 평균과 4차 조사에서 응답 자료만을 이용해 앞 절에서 제시한 여섯 가지 무응답 보정 가중값 작성방법에 따라 산출한 가중값을 이용해 구한 평균 추정값을 비교하였다.

응답률 추정을 위한 로짓모형에서 독립변수로는 기존 고령화연구패널에서 활용하고 있는 5개 지역(제주 제외), 성, 연령대, 최종학력, 단독가구 여부, 자가 여부, 연금소득 여부, 농어업 소득 유무, 주거형태를 사용하였다 (Kim 등, 2015).

추정된 응답률을 활용하여 제시된 여섯 가지 방법으로 가중값을 산출해 관심변수 평균을 구한 결과는 Figure 5.1의 위쪽 그림을 보면, 그룹 구분 없이 추정한 [방법 1]–[방법 4]에 비해 연속/비연속 응답 그룹을 구분하고 복합추정 방식을 적용한 [방법 5]와 [방법 6]에서의 편향이 고혈압 유병률과 연간 개인 총소득에서 모두 감소하는 것을 확인할 수 있다. 이 결과는 모의실험에서 T4와 유사한 결과를 보여준다.

한편, 연속/비연속 응답 그룹을 구분하여 추정한 [방법 5]와 [방법 6]은 편향을 줄여 추정의 정확성은 높일 수 있었으나, 극단 가중값이 상대적으로 많이 산출되어 가중값의 변동계수가 커지는 한계를 갖고 있

Table 5.1. Weight statistic by method (bias of annual personal gross income)

	Before treatment extreme weights					After treatment extreme weights				
	Bias	CV	Max	Med	Min	Bias	CV	Max	Med	Min
Weight	-	47.90	6546.84	1405.70	277.71	-	47.90	6546.84	1405.70	277.71
Method 1	−24.35	57.36	11033.16	1649.75	306.03	−27.05	55.39	6409.67	1657.51	307.47
Method 2	−28.64	57.24	11149.73	1659.61	309.26	−31.54	55.20	6477.99	1667.57	310.74
Method 3	−29.78	55.07	11500.86	1676.38	314.35	−34.00	53.08	6139.97	1684.22	315.82
Method 4	−29.15	52.87	10444.02	1693.31	320.96	−31.96	51.00	5707.12	1701.13	322.45
Method 5	16.43	119.81	45320.40	1410.59	261.88	−6.97	103.05	15058.68	1444.83	268.23
Method 6	16.19	119.86	45293.66	1410.68	261.89	−7.21	103.17	15096.49	1444.70	268.21

었다. 따라서, 극단 가중값 처리 후에도 추정의 결과가 다른 방법에 비해 우수하다고 볼 수 있는지 확인할 필요가 있다. 극단 가중값 처리 방법에는 여러 가지 방법들이 있으나 본 연구에서는 간단하게 99% 절삭법을 적용하였다 (Kim 등, 2015). 극단 가중값 처리 후 결과는 Figure 5.1의 아래쪽 그림에서 볼 수 있는데, 연속/비연속 응답 그룹 구분 없이 추정한 [방법 1]–[방법 4]은 극단 가중값 처리 후에 편향이 증가하는 현상을 보여주고 있다. 반면에 연속/비연속 응답 그룹을 구분하여 추정한 [방법 5]와 [방법 6]은 다른 방법들과는 달리 극단 가중값 처리 후에 편향이 감소하고 추정 결과도 좋았다.

Table 5.1은 제시된 여섯 가지 가중값 산출방법에서 극단 가중값 처리 전과 후의 가중값 현황과 개인 총소득 평균에 대한 추정 결과를 비교해 보기 위해 기본적인 기술통계를 산출해 정리한 것이다. Table 5.1를 보면, 가중값 변동계수(coefficient of variation; CV)는 [방법 5]와 [방법 6]에서는 극단 가중값 처리 후에 13% 정도 수준 감소하였고, [방법 1]–[방법 4]의 경우는 대략 0.3% 수준 감소하였다. 아울러 연간 개인 총소득에 대한 추정 결과에 있어서 [방법 5]와 [방법 6]의 경우 다른 방법에 비해 편향이 대폭 감소함으로 확인할 수 있다. 고혈압 유병률의 경우도 매우 유사한 결과를 보여주기 때문에 분석결과는 논문에 제시하지 않았다.

한편 기존 응답모형에서 사용한 변수 외에도 각 단위의 응답성향을 설명해 줄 수 있는 파라데이터가 존재하면 이를 응답모형에 반영함으로써 보다 효과적인 무응답 보정 가중값 산출이 가능할 수 있다. 여기서는 실험적으로 고령화연구패널 사례에서 6차 조사까지의 응답 이력을 이용하여 응답성향의 잠재변수로 볼 수 있는 가상적 평균 응답률 변수 z_i를 응답률 추정 시 추가적으로 사용하는 경우를 고려해 보았다. z_i는 응답자의 응답횟수를 기준으로 만든 가상적인 변수로, 값이 클수록 응답할 가능성이 높다고 할 수 있다. 가상적 평균 응답률 z_i를 응답모형에 반영하여 [방법 5]와 [방법 6]의 방법으로 가중값을 작성한 결과 정리한 결과는 Figure 5.2와 같다. 응답성향을 나타내는 변수를 추가적으로 반영하게 되면 고혈압 유병률과 연간 개인 총소득에서 모두 편향이 대폭 개선될 수 있다는 것을 보여주고 있다.

6. 결론

패널 자료는 오랜 추적 기간에 따라 자료가 쌓이는 만큼 그 가치가 증대된다. 이와 맞물려 장기추적에 따른 표본이탈 등 자료의 대표성 보완을 위한 노력도 더욱 중요하다. 표본 마모에 따른 패널자료의 모집단 대표성 문제를 대부분의 패널조사에서는 가중값을 작성해 해결하고자 노력해 왔다. 본 논문에서는 패널조사에서 사용되는 기존 가중값 작성방법들을 살펴보고, 연속/비연속 응답 그룹의 응답자를 구분하고 극단 가중값의 발생을 줄이기 위해 결합추정 방식의 새로운 가중값 산출방법을 제시하였다.

기존의 가중값 작성방법들에 있어서는 연속/비연속 응답 그룹의 구분 없이 작성되기 때문에 비연속 응답 그룹이 연속 응답 그룹과 특성상 큰 차이가 없다는 것을 전제로 가중값을 산출한 것으로 볼 수 있다.

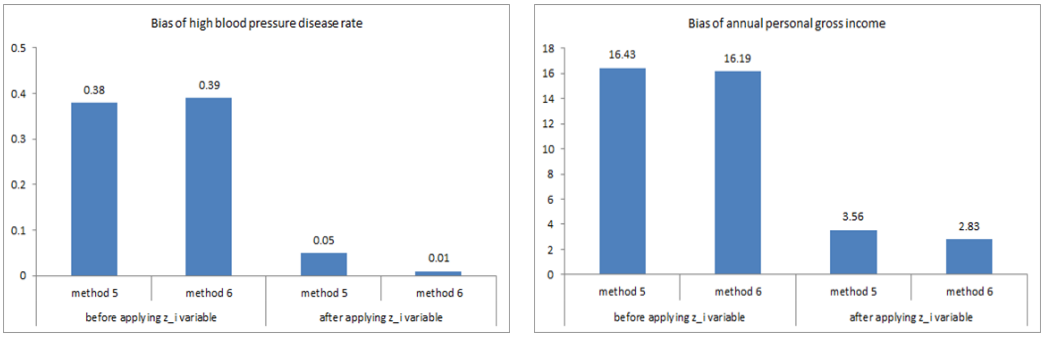


Figure 5.2. Bias of method 5 and method 6 with response propensity variable z_i.

하지만 지속적으로 패널조사에 적극적으로 참여하는 연속 응답 그룹과 조사 참여의향이 상대적으로 많이 떨어지는 것으로 볼 수 있는 비연속 응답 그룹이 동일한 특성을 갖는다고 가정하는 것은 현실적으로 매우 수용하기 어려운 가정일 수 있다. 실제 고령화연구패널 자료를 이용하여 연속 응답 그룹과 비연속 응답 그룹 간 주요 변수들의 평균에 있어서 차이가 있음을 보았고, 2차의 응답률 대비 t 시점 응답률의 산점도에서도 그룹 간 차이가 있음을 확인하였다. 따라서, 본 논문에서는 비연속 응답 그룹의 특성이 연속 응답 그룹과 동일하지 않다는 것을 전제로 연속/비연속 응답 그룹별 특성을 반영한 복합추정 방식을 적용한 새로운 패널조사 가중값 작성방법을 제안하였다. 복합추정 방식을 도입함으로써 비연속 응답 그룹의 응답률 산출과정에서 비연속 응답 그룹에서 산출한 직접추정량과, 연속 응답 그룹의 응답모형을 기반으로 한 합성 추정량을 결합하는 가중값 산출방식을 취함으로써 극단 가중값 발생 가능성을 상당폭 줄일 수 있었다.

다양한 시나리오의 모의실험을 통해 제안한 방법이 기존의 다른 가중값 방법들과 비교하여 얼마나 효율적인지를 살펴보았다. 결과적으로 그룹 구분 없이 작성해왔던 기존의 [방법 1]–[방법 4]에 비해 복합추정 방식의 [방법 5]와 [방법 6]은 편향을 대폭 감소시키는 효과가 있다는 것을 보일 수 있었다. 또한 실제 고령화연구패널 자료에 적용해 본 결과 기존의 [방법 1]–[방법 4]의 방법에 비해 연속/비연속 응답 그룹을 구분한 [방법 5]와 [방법 6]을 통해 고혈압 유병률과 연간 개인 총소득 추정에 있어서 무응답에 따른 편향을 대폭 줄일 수 있다는 것을 확인할 수 있었다.

패널 자료는 웨이브가 진행됨에 따라 1차 원표본에서 연속 응답 그룹의 비중은 점점 작아지고, 비연속 응답 그룹의 비중은 점점 늘어나게 된다. 또한, 연속/비연속 응답 그룹 간에 특성상 차이가 있는 경우 연속 응답 그룹만으로 최초 원표본을 대표하기는 힘들다. 이런 관점에서 본 논문에서 제시한 패널조사 가중값 산출방식은 비연속 응답 그룹을 응답 그룹과 구분해 관련 특성을 적극 반영함으로써 무응답에 따른 편향을 줄일 수 있다는 측면에서 의미 있는 결과를 제시하고 있다고 판단된다.

References

Baek, J. S. and Shim, K. S. (2012). How to create cross weights in household panel survey, *2012 Second-Half Research Report 3*, National Bureau of Korea.

Brady, T. W. (2013). The effects of errors in paradata on weighting class adjustments: a simulation study, *Improving Survey with Paradata: Analytic Uses of Process Information*, John Wiley & Sons Inc.

Josiane, B., David, H., Anni, O., and Melanie, N. (2018). *The Dynamics of Ageing : Evidence from the English Longitudinal Study of Ageing 2002-16(Wave 8) Ch5. Methodology*, National Center for Social Research.

Kim, Y. W., Lee, K. J., and Park, I. H. (2015). *Report of Recreating Weights for KLoSA Wave 1–4 and Weights for Wave 5*, Korea Employment Information Service.

Korea Institute of Public Finance (2012). *Weights for National Survey of Tax and Benefit* (Technical Report), Korea.

Lee, J. R., Choi, E. Y., Do, N. H., Song, S. Y., Wang, Y. H., and Jung, Y. H. (2011). *Panel Study on Korean Children 2011 Report*, Korea Institute of Child Care and Education.

Park, M. G. and Kim, Y. W. (2013). *Research Report of YP Wave 6 and New Response Model Development*, Korea Employment Information Service.

Park, M. G., Kim, Y. W., and Byun, J. S. (2013). *Weights Research of Korean Labor & Income Panel Survey*, Korea Labor Institute.

Park, M. G., Lee, K. S., Park, H. S., and Kang, H. C. (2011). A Study on the Construction of Weights for KYPS. *Survey Research 12(3) 173-186*. The Korean Association for Survey Research.

Qixuan, C., Andrew, G., Melissa, T., Fran, H. N. and Sandro, G. (2012). Weighting Adjustments for Panel Nonresponse. *Unpublished manuscript*. Columbia University.

Sally, B., David, H., and Margaret, B. (2015). *The Dynamics of Aging : The 2012 English Longitudinal Study of Ageing(Wave 6)*. National Center for Social Research.

Shaun, S., Kate, C., and Carli, L. (2008). *Living in the* 21^{st} *century : older people in England the 2006 English Longitudinal Study of Ageing (Wave 3) Ch9. Methodology*. National Center for Social Research.

Shin, J. G., Hwang, K. H., and Cho, M. S. (2017). *YP2007 Wave10 Analysis Report*, Korea Employment Information Service.

Takis, M. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **72**, 27–48.

패널조사에서 비연속 응답 그룹 편향 보정을 위한 복합가중값

최형아[a] · 김영원[b,1]

[a]한국고용정보원 고용통계조사팀, [b]숙명여자대학교 통계학과

(2019년 4월 9일 접수, 2019년 4월 23일 수정, 2019년 4월 23일 채택)

요 약

패널 자료는 자료가 축적되는 만큼 그 가치가 증대된다. 이와 동시에 장기추적에 따른 표본이탈은 자료의 신뢰성을 떨어뜨린다. 국내·외 대부분의 패널조사에서 가중값 보정을 통해 표본 이탈 문제를 해결하고 있다. 본 논문에서는 패널자료에서 차수별 응답여부에 따라 연속 응답 그룹과 비연속 응답 그룹으로 나누고, 비연속 응답 그룹에 대한 적정 가중값 산출방법을 검토하였다. 연속/비연속 응답그룹을 구분하여 비연속 응답 그룹의 응답자 특성을 반영한 복합추정 방식의 가중값 작성방법을 제안하고, 그룹의 구분 없이 작성하였던 기존의 가중값 작성방법과 새로 제안한 복합추정 방식의 가중값 산출방법의 효율성을 모의실험과 실증분석을 통해 살펴보았다. 결과적으로 새로 제안한 복합추정 방식의 가중값 산출방법은 기존 방법 보다 편향을 대폭 감소시킴을 모의실험을 통해 볼 수 있었다. 한편, 제시한 가중값 작성방법을 한국고용정보원 고령화연구패널에 적용한 결과도 제시하였다.

주요용어: 패널조사, 가중값, 복합추정 가중값, 비연속 응답 그룹

[1]교신저자: (04310) 서울 용산구 청파로47길 100, 숙명여자대학교 통계학과. E-mail: ywkim@sookmyung.ac.kr