통계연구(2022), 제27권 제4호, 30-46/ DOI https://doi.org/10.22886/jkos.2022.27.4.30

자기기입 조사자료에 대한 히핑 현상 보정 방안 연구[1)]

박승환[2)]

요약

자기응답으로 조사된 자료에서 주제에 대한 민감성, 기억에 의한 오류 등으로 인하여 참값과 측정값간의 오차가 발생한다. 이러한 오차 중 반올림된 응답값 혹은 0 또는 5와 같은 선호하는 자리수에 응답값이 집중되는 경우를 히핑(heaping)현상이라고 한다. 이러한 히핑 현상은 우연히 발생할 수도 있으나 무작위로 발생하지 않는 경우 히핑 자료를 사용하여 통계분석을 한다면 편향된 분석결과를 도출할 수 있다. 이 논문에서는 히핑으로 인한 측정오차를 보정하기 위한 확률적 대체 방법을 제안하였다. 제안한 방법을 1차 년도 한국복지패널조사 자료의 경상소득 변수에 적용한 결과, 보정 후의 값들의 산포가 전체적으로 줄어들었으며 조사된 응답값과의 차이도 일정 구간에 속하였다. 본 연구결과를 바탕으로 조사자료의 측정오차를 보정할 수 있는 방법을 다양하게 모색할 수 있을 것이다.

주요용어 : 측정오차 모형, 히핑, 결측값 대체

1. 서론

히핑(heaping)이란 조사자료에 있어 응답자가 자기기입 등 스스로 응답한 흡연량, 음주횟수 등 범주형 변수의 빈도가 특정 응답값의 배수에 집중 되거나 임금, 소득, 소비 등 연속형 변수에 대하여 관측된 분포가 특정 응답값의 배수에 지나치게 크게 집중되는 현상을 뜻한다.

히핑 현상이 발생한 특정 값의 배수에 응답한 값 모두에 측정오차가 존재한다고 보기 보다는 정확하게 응답한 경우와 측정오차가 있는 응답 두 경우가 섞여있는 일종의 혼합분포의 형태로 히핑에 대한 측정오차를 해석할 수 있다. 이러한 혼합분포 형태의 측정오차는 추후 관심 변수에 대한 추정에 있어 편향을 발생시키고 분산을 증가시킨다. 특히 소득이 관심 변수인 경우 특정 응답값에 집중된 분포는 소득 불평등 지수 등 양극화 지수의 일종을 계산할 때 왜곡된 값을 제공하게 되는 원인이 될 수 있다.

히핑은 측정오차가 있는 경우 응답값을 활용한 모수 추정 문제와 유사하다. 즉 자료의 히핑 현상을 보정하기 위해 불완전하게 관측된 응답값을 사용하여 히핑이 발생한 부분의 응답값을 대체하는 방법을 생각해볼 수 있다. 히핑이 발생하는 응답을 모두 대체할지 일부만 대체할지는 히핑에 대한 혼합분포 가정을 통하여 조절할 수 있다.

---

1) 이 논문은 이혜정, 신지영, 박승환, 지희정, 오미애. (2021). 조사 자료의 품질 검증 연구보고서의 일부 내용을 중심으로 수정, 보완 및 추가 구성하여 작성하였음.

2) 교신저자. 춘천시 강원대학길 1, 강원대학교 정보통계학과, 조교수.
E-mail: stat.shpark@kangwon.ac.kr

히핑 현상이 존재하는 자료에 대한 통계 분석은 여러 측면에서 연구되어 왔다. T. H. Cummings et al. (2015)에서는 주로 가산 자료(count data)에서 발생하는 히핑 자료에 대하여 혼합분포를 가정하여 통계적 추정을 수행하였다. 히핑이 존재하는 지점의 응답값은 정확한 응답값이 아니며 히핑 지점을 기준으로 양측 절단(censored)되었다고 가정한 모형을 히핑 자료 분석에 적용하였다. Groß, M., and Rendtel, U. (2016)에서는 커널 함수 추정법을 활용하여 베이지안 방법론을 적용한 히핑 데이터 보정 방법을 다루고 있다. 커널 함수 추정법을 활용한 히핑 보정 방법은 R 프로그램 내부에 패키지 형태로 구현되어 있다. “Kernelheaping” 패키지 내부의 “dheaping” 함수를 사용하여 앞선 방법을 구현 할 수 있다. 대체 방법 중 Kim and Hong (2012)에서는 MCEM 방법을 이용하여 히핑된 자료값을 대체하였다. 히핑을 보정하게 되면 일반적으로 특정 응답값의 배수에 지나치게 집중된 분포가 주변으로 퍼지는 효과가 생겨 응답값의 분포가 부드러워진다.

본 연구의 목적은 복지패널의 경상소득 변수에 대하여 히핑 현상이 있는지 살펴본 후 히핑이 존재한다면 히핑 현상을 보정할 수 있는 기존의 방법을 검토하고 기존의 방법을 개선하는 새로운 방법론을 제안하여 실제 복지패널에 적용하는 것이다.

본 논문의 구성은 다음과 같다. 2장에서는 한국복지패널 개요 및 히핑 현황을 분석한다. 3장에서는 확률 대체 법을 통한 히핑 보정 방법론을 제안한다. 4장에서는 1차년도 한국복지패널 조사 자료에 대하여 히핑 보정 방법론을 적용하고 그 결과를 해석한다. 5장에서는 본 연구의 분석 내용을 정리하고 그 시사점 및 의의를 기술한다.

2. 한국복지패널 개요 및 히핑 현황 분석

2.1 한국복지패널 개요

한국복지패널(KOWEPS)은 제주도, 농어가 및 읍·면을 포함한 전국을 대상으로 하는 종단면 조사이다. 최초의 한국복지패널 원표본 가구의 규모는 7,072가구이며 조사 대상은 표본 가구, 표본 가구에 속하는 15세 이상 가구원, 그리고 부가조사 대상으로 구분된다. 최초의 한국복지패널 원표본 가구를 선정하기 위한 기초 자료로는 ‘2006년 국민생활실태조사’를 활용하였으며, ‘국민생활실태조사’의 경우 ‘2005년 인구센서스 자료 90% 조사구’에서 추출하였다. 표본 배분은 복지 욕구를 보다 효과적으로 파악하기 위하여 저소득층을 과대 표집하였다. 즉 중위소득 60%(OECD 상대빈곤선) 이하의 저소득층을 3,500가구 추출하고, 중위소득 60% 이상에 해당하는 일반 가구를 3,500가구 추출하였다. 신규 패널 표본으로는 사전 조사 결과를 기초로 일반 및 저소득 가구를 구분하여 전체 조사구에서 완료 목표 1,800가구의 3배수인 5,400가구를 추출하였다. 일반 가구와 저소득 가구의 비율은 1차 연도(2006년)와 동일하게 저소득 가구를 과대 표집하였다. 지역별 표본 배분 또한 1차 연도 조사 당시의 지역별 가구 비율과 유사하게 표본 가구를 배분하여 패널의 동질성을 최대한 유지하였다.

한국보건사회연구원은 저소득층의 복지 수요 및 욕구를 적절히 조사하기 위해 2006년도에 조사 대상 가구로 일반 가구와 저소득층 가구를 각각 50%씩 추출하여 패널 표본 가구를 구축하였다. 2006년도의 패널 구축 과정을 먼저 살펴보면, 소득집단에 따라 패널을 구축하기 위해서는 표본 대상 가구의 소득 자료가 필요하였다. 이를 위해 '2005년 인구주택총사 90% 자료'에서 확률 비례 추출한 '2006년 국민생활실태조사'의 최종 조사 완료 가구인 2만 4,711가구의 소득 자료를 얻을 수 있었다. 이를 기준으로 일반 가구와 저소득층 가구를 구분하여 두 층에서 각각 3,500가구씩 총 7,000가구를 표본으로 선정하였으며, 이 중 최종 패널 가구로 구축된 표본 가구는 7,072가구였다. 표본 추출 과정에서 저소득층 가구는 향후 패널 탈락과 통계적 유의미성을 고려하여 과대 표집하였다.

한국보건사회연구원에서는 '2006년 국민생활실태조사'에 필요한 표본 가구 약 3만 가구를 조사하기 위해 2005년도 인구주택총조사 90% 조사구인 23만여 개의 조사구 중 517개를 지역별 조사구의 규모에 따라 층화 확률 비례 추출하였다. 전체 517개 표본 조사구 중 수해와 같은 천재지변으로 조사가 불가능한 지역을 제외한 487개 조사구에 대해 조사를 완료하였다. 조사 대상 지역은 섬을 제외한 전국의 각 시·도이며, 제주도를 포함한다. '2006년 국민생활실태조사'에서 최종적으로 조사가 완료된 2만 4,711가구의 소득을 기준으로 7,000가구를 2단계 표본 추출하여 최종적으로 7,072가구를 패널로 구축하였다. 1단계 표집 자료인 '2006년 국민생활실태조사' 자료를 바탕으로 패널 가구의 대표성을 확보하기 위해 중위소득 60% 이하인 저소득층 가구 3,500가구와 중위소득 60% 이상인 일반 가구 3,500가구를 각각 표본으로 추출하여 조사하였다. 소득 규모별로 구분된 2개의 층에서 지역별, 조사구별로 확률 비례 계통 추출에 따라 일반 가구와 저소득 가구를 표본으로 추출하였다. 상대적으로 작은 규모를 가진 저소득층 가구에 대해서는 추출률을 상향 조정하여 일반 가구와 동일한 수준인 3,500가구를 표본 가구로 선정하였다. 패널로 구축된 표본 가구를 당초 층별로 3,500가구씩 배분하였으나 조사 거절, 패널 참여 거부 등의 사유로 저소득층에서는 표본 설계 당시의 3,500가구를 약간 밑도는 규모인 3,283가구가 패널로 구축되었다.

한국복지패널의 6차 연도 조사 이후에 원표본 가구 유지율이 감소하는 상황에서 신규 표본 가구를 추가할 필요성이 제기되었다. 따라서 7차(2012년) 조사에서는 1차 조사 패널 표본 규모를 유지하고자 약 1,800가구를 추가하여 신규 패널을 구축하였다. 신규 표본 추가를 위한 표본 추출 방법도 1차 조사와 동일한 방식을 고려하였다.

2.2 히핑 현황 분석

한국복지패널 가구 자료 중 자기 기입으로 작성된 경상 소득에 대한 히핑 현상을 분석한다. 복지패널에서 경상 소득은 근로소득, 사업 및 부업소득, 재산소득, 사적이전 소득, 공적이전 소득의 합으로 계산되며 연간 소득을 의미한다. 경상소득과 이를 구성하는 하위 소득 변수들은 모두 연속형 변수이다. 복지패널 1차부터 15차까지 가구 경상 소득에 대하여 히핑 현상이 있는지 살펴본다. 7차에 추가된 신규패널을 따로 구분하여 원패널과 신규패널을 나누어 살펴보도록 한다.

히핑 현상은 조사된 소득 값의 분포가 특정 값의 배수에 특히 많이 치우진 것을 의미한다. 따라서 먼저 각 차수별 특정 응답값에 치우진 분포가 있는지 소득 응답값 별 비중을 분석하였다. <표 2.1>과 원표본 1차, 2차, 14차, 15차 경상소득 응답값 비중 상위 10개에 대한 경상소득 값을 제시하였다. <표 2.2>에는 신규표본 7차, 8차, 14차, 15차 경상소득 응답값 비중 상위 10개에 대한 경상소득 값을 제시하였다.

원표본 패널의 경우 1차 년도와 2차 년도 응답 비중 상위 10개의 응답값을 보면 5 또는 10의 배수 혹은 120의 배수임을 알 수 있다. 이러한 10의 배수의 형태로 응답 비중이 높은 현상은 14차, 15차 년도에서는 찾아 볼 수가 없다. 신규패널 표본의 경우에서는 7차 년도의 응답 비중 상위 10개의 응답값은 5의 배수의 형태로 나타나지만 8차 년도 이후로는 특정 배수의 형태로 나타나는 것을 찾을 수 없다.

<표 2.1> 원표본 1차, 2차, 14차, 15차 경상소득 응답 비중 상위 10개

1차년도		2차년도		14차년도		15차년도	
가구소득	비중	가구소득	비중	가구소득	비중	가구소득	비중
3,000	0.6%	3,000	0.6%	1,208	0.1%	3,603	0.2%
1,800	0.5%	1,800	0.5%	1,238	0.1%	969	0.1%
2,400	0.5%	2,400	0.5%	1,444	0.1%	5,580	0.1%
1,560	0.4%	1,560	0.4%	1,469	0.1%	894	0.1%
1,920	0.4%	1,920	0.4%	1,531	0.1%	948	0.1%
3,600	0.4%	3,600	0.4%	1,577	0.1%	1,023	0.1%
1,440	0.4%	1,440	0.4%	1,723	0.1%	1,139	0.1%
2,160	0.4%	2,160	0.4%	770	0.1%	1,327	0.1%
1,200	0.3%	1,200	0.3%	828	0.1%	1,984	0.1%
3,120	0.3%	3,120	0.3%	874	0.1%	2,304	0.1%

<표 2.2> 신규표본 7차, 8차, 14차, 15차 경상소득 응답 비중 상위 10개

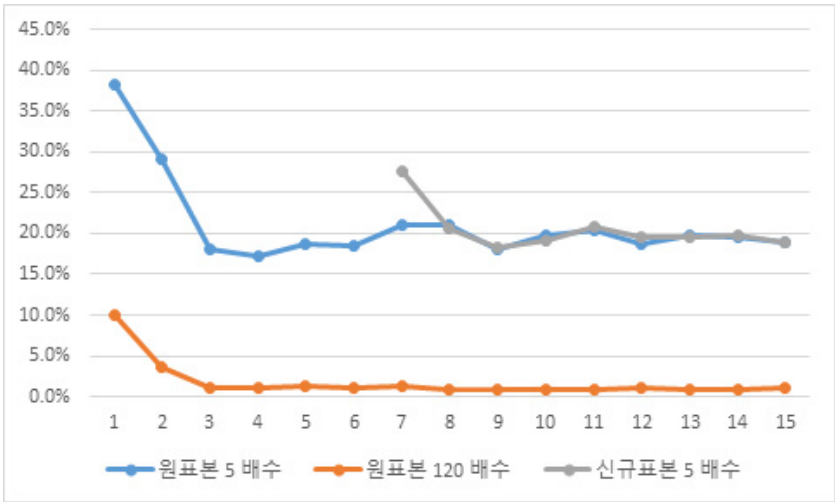
7차년도		8차년도		14차년도		15차년도	
가구소득	비중	가구소득	비중	가구소득	비중	가구소득	비중
649	0.4%	1,337	0.4%	1,531	0.3%	1,139	0.4%
775	0.3%	613	0.2%	770	0.3%	1,270	0.4%
529	0.2%	624	0.2%	874	0.3%	882	0.3%
569	0.2%	647	0.2%	1,723	0.3%	1,109	0.2%
589	0.2%	664	0.2%	1,226	0.2%	1,158	0.2%
609	0.2%	765	0.2%	1,303	0.2%	1,289	0.2%
703	0.2%	823	0.2%	1,444	0.2%	1,508	0.2%
995	0.2%	1,243	0.2%	1,522	0.2%	1,908	0.2%
1,340	0.2%	1,628	0.2%	1,596	0.2%	2,304	0.2%
1,545	0.2%	2,182	0.2%	1,663	0.2%	3,603	0.2%

원표본의 경우 1차부터 15차까지 경상소득이 5또는 120의 배수의 형태로 나타나는 비율과 신규표본의 경우 7차부터 15차까지 5의 배수의 형태로 경상소득이 나타나는

비율을 분석하였다. <표 2.3>과 <그림 2.1>을 보면 원표본에서 1차 년도에서 5의 배수인 경상소득의 비율은 약 38% 2차 년도에서는 29%로 그 후 차수에서 비율인 약 20% 보다 상당히 높게 나타남을 알 수 있다. 120의 배수 비율도 1차 년도에서는 약 10%로 그 이후 차수에 비해서 상당히 높게 나타난다. 신규표본에서 5의 배수인 경상소득의 비율은 7차 년도 약 27%로 그 이후 차수에서 약 20%로 나타나는 점에서 다소 높게 나타남을 알 수 있다. 이로부터 1차 년도와 7차 년도에는 경상 소득이 5또는 120의 배수 형태로 응답한 경우가 다른 차수에 비해서 높음을 알 수 있다.

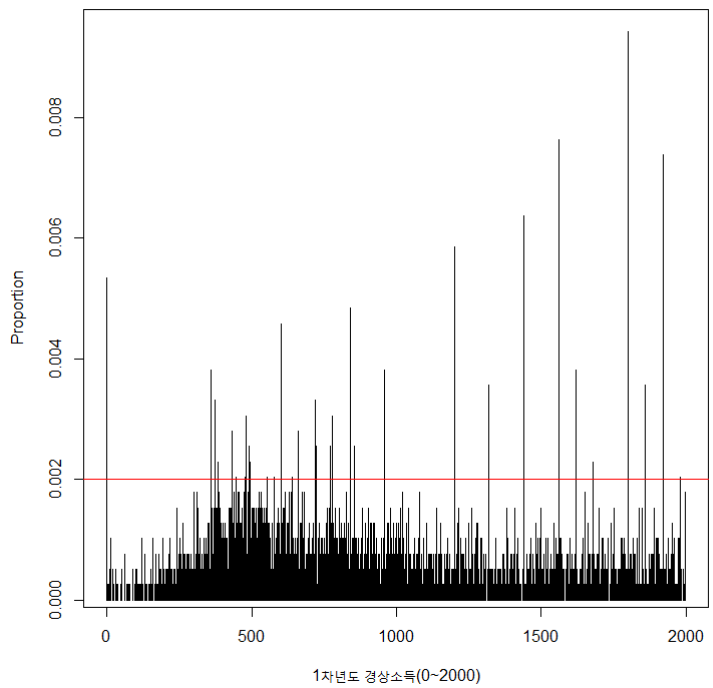
<표 2.3> 1차부터 15차까지 5또는 120 배수인 경상소득 응답값의 비율

차수	원표본		신규표본
	5의 배수	120의 배수	5의 배수
1차	38.3%	10.1%	
2차	29.2%	3.7%	
3차	18.0%	1.0%	
4차	17.3%	1.0%	
5차	18.7%	1.3%	
6차	18.4%	1.1%	
7차	21.1%	1.2%	27.7%
8차	21.0%	0.9%	20.6%
9차	18.1%	0.9%	18.3%
10차	19.8%	0.9%	19.0%
11차	20.3%	0.9%	20.7%
12차	18.8%	1.1%	19.5%
13차	19.7%	0.9%	19.6%
14차	19.5%	0.9%	19.8%
15차	18.9%	1.1%	19.0%

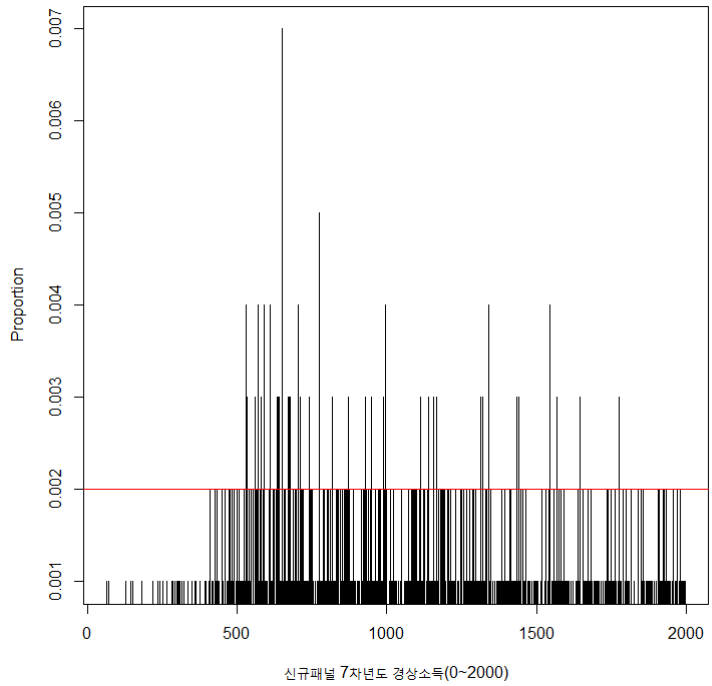


<그림 2.1> 1차부터 15차까지 5또는 120 배수인 경상소득 응답값의 비율

1차 년도의 경상소득 값이 5또는 120의 배수 형태로 응답한 경우가 많다는 앞선 결과에 대하여 자세히 살펴보기 위하여 spike plot을 그려보았다. 해석을 용이하게 하기 위하여 1차 년도 경상소득 중 0보다 크고 2,000보다 작은 응답값들에 대하여 spike plot을 작성하였고 이는 <그림 2.2>에 나타나 있다. 그림을 보면 응답 비율 약 0.2%가 넘어가는 경우 특정 응답값들에 분포가 집중되어 비율 값이 튀는 것을 확인할 수 있다. 신규패널에 대해서도 7차 년도 경상소득에 대해서 spike plot을 작성하였다. <그림 2.3>을 보면 응답 비율 약 0.2%가 넘어가는 경우 특정 응답값들에 분포가 집중되어 비율 값이 튀는 것을 확인할 수 있다.



<그림 2.2 > 1차 년도 경상소득(0~2,000)에 대한 spike plot



<그림 2.3> 신규패널 7차 년도 경상소득(0~2,000)에 대한 spike plot

원표본과 신규표본 1차 년도에서 응답 비율 약 0.2%가 넘어가는 경우의 응답값들에 대하여 각각 <표 2.4>와 <표 2.5>에 경상소득 크기순으로 표시하였다. 원표본 1차 년도 경상소득 중 응답 비율이 특히 높은 응답값은 260, 600, 840 등으로 모두 120의 배수인 것을 확인 할 수 있다. 신규 표본의 경우 경상소득의 끝자리가 9, 0, 5로 끝나는 응답값의 비율이 높은 점을 확인 할 수 있다.

<표 2.4> 1차 년도 원표본 경상소득 중 응답 비율 높은 경상소득

경상소득	비율	경상소득	비율
360	0.2%	2,160	0.4%
600	0.3%	2,400	0.5%
840	0.3%	2,500	0.2%
960	0.2%	2,640	0.2%
1,200	0.3%	3,000	0.6%
1,440	0.4%	3,120	0.3%
1,560	0.4%	3,600	0.4%
1,620	0.2%	4,000	0.2%
1,800	0.5%	4,200	0.2%
1,920	0.4%	4,800	0.3%
2,040	0.3%	6,000	0.3%

<표 2.5> 1차 년도 신규표본 경상소득 중 응답 비율 높은 경상소득

경상소득	비율
529	0.2%
569	0.2%
589	0.2%
609	0.2%
649	0.4%
703	0.2%
775	0.3%
995	0.2%
1,340	0.2%
1,545	0.2%

3. 확률 대체법을 통한 히핑 보정 방법

김재광 (2014)에서는 노동 패널 자료의 소득 변수에 대한 히핑 현상을 보정하기 위해 결측치 대체법을 활용하였다. Fractional imutation 방법에 기초하여 주어진 자료 $\boldsymbol{x}_i$, z_i를 통하여 참값 y_i를 생성하는 방법으로 히핑 자료를 보정하였다. 참값 y_i를 다음의 모형으로부터 생성한다.

$$f(y_i|\boldsymbol{x}_i, z_i) \propto f_1(y_i|\boldsymbol{x}_i) g(z_i|\boldsymbol{x}_i, y_i) \tag{3.1}$$

여기서 $f_1(y_i|\boldsymbol{x}_i)$에 대한 모형은 참값과 보조변수간의 모형을 뜻하며 구조오차 모형이라고 한다. $g(z_i|\boldsymbol{x}_i, y_i)$에 대한 모형은 y_i가 주어졌을 때 $\boldsymbol{x}_i$와 z_i는 조건부 독립이라는 가정 하에서 관측값과 참값간의 모형을 뜻하며 측정오차 모형이라고 한다..

김재광 (2014)에서는 히핑 보정을 모든 경우에 대하여 실행하였는데 본 연구에서는 히핑이 발생한 경우와 아닌 경우를 나누어 히핑 보정을 수행하는 방법을 고려하였다. 먼저 히핑을 나타내는 지시 변수 I_h를 생각하자.

$$I_h = \begin{cases} 1 \; y_i \neq z_i (\text{히핑}) \\ 0 \; y_i = z_i \end{cases}$$

이 I_h에 관련한 확률 값은 설명 변수 $\boldsymbol{x}_i$에 의존하는 어떤 함수로 표현될 수 있다고 가정하자. 즉, 로지스틱 혹은 프로빗 같이 알려진 함수 $\pi(\cdot)$에 대하여 다음과 같이 표현되고 (ϕ_0, ϕ_1)은 추정해야할 모수 값으로 볼 수 있다. 본 연구에서는 $\boldsymbol{x}_i$에만 의존하게 모형을 세웠으나 y_i에도 의존하게 모형을 세울 수 있을 것이다.

$$\Pr(I_h = 1 | \boldsymbol{x}_{i,} y_i) = \pi(\phi_0 + \boldsymbol{x}'_i \boldsymbol{\phi}_1)$$

또한 $\tilde{y}_i$를 y_i에서 측정 오차가 추가된 경우 얻어지는 값, 즉 히핑 현상에 의해 조정된 값이라고 한다면, 관측치 z_i는 $z_i = (1 - I_{hi}) y_i + I_{hi} \tilde{y}_i$ 로 표현될 수 있을 것이다. 이와 같은 경우 히핑 현상을 보정한 보정 응답값은 대체(impuation)방법을 통하여 생성할 수 있고 그 형태는 다음과 같다.

$$y^{**} = \Pr(I_h = 0 | \boldsymbol{x}_i) y_i + \Pr(I_h = 1 | \boldsymbol{x}_i) y_i^*,$$

여기서 y_i^*는 식 (3.1)의 분포로부터 생성된다. $f_1(y_i | x_i)$에 대한 모형과 $g(z_i | x_i, y_i)$ 모형은 다음과 같다:

$$y_i^* \sim f(y_i | \boldsymbol{x}_i, z_i) \propto f_1(y_i | \boldsymbol{x}_i) g(z_i | \boldsymbol{x}_i, y_i),$$

$$\log(y_i) = \beta_0 + \boldsymbol{x}_i' \boldsymbol{\beta}_1 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_1^2) \tag{3.2}$$

$$\log(z_i) \sim N(\gamma_0 + \gamma_1 \log(y_i), \sigma_2^2). \tag{3.3}$$

가구소득의 분포는 오른쪽 극단으로 치우쳐 있는 경우가 많아 식(3.2)와 같은 로그 변환을 사용하여 구조모형을 설정한다. 모수 $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1), \sigma_1^2$과 $\boldsymbol{\gamma} = (\gamma_0, \gamma_1), \sigma_2^2$은 EM 알고리즘을 통하여 추정할 수 있고 그 과정은 다음과 같다.

[Step1] y_i 대신 z_i를 사용하여 $f(z_i | \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma_1^2)$의 모수를 추정하고 이를 $(\boldsymbol{\beta}, \sigma_1^2)$의 초

기치를 구한다. 이로부터 $y_i^{*(j)} \sim f(y|\boldsymbol{x};\boldsymbol{\beta},\sigma_1^2)$, $j=1,...,M$을 생성한다. 생성된 $y_i^{*(j)}$를 사용하여 (γ,σ_2^2)의 초기치를 구한다.

[Step2] 각 i에서 생성된 M개의 $y_i^{*(j)}$를 이용하여 fractional weights을 계산한다.

$$w_{ij}^{*} \propto f(y_i^{*(j)} \mid \boldsymbol{x}_i;\boldsymbol{\beta},\sigma_1^2) f(z_i \mid y_i^{*(j)};\gamma,\sigma_2^2),$$
$$\sum_{j=1}^{M} w_{ij}^{*} = 1.$$

[Step3] 현 step에서의 각 모수의 추정치와 그에 따른 fractional weights을 이용하여 다음의 스코어 함수를 0으로 만드는 새로운 모수 추정치를 구한다.

$$\sum_{i}^{n}\sum_{j=1}^{M} w_{ij}^{*} S_1(\boldsymbol{\beta};\boldsymbol{x}_i,y_i^{*(j)}) = 0$$
$$\sum_{i}^{n}\sum_{j=1}^{M} w_{ij}^{*} S_2(\sigma_1^2;\boldsymbol{x}_i,y_i^{*(j)}) = 0$$
$$\sum_{i}^{n}\sum_{j=1}^{M} w_{ij}^{*} S_3(\gamma;z_i,y_i^{*(j)}) = 0$$
$$\sum_{i}^{n}\sum_{j=1}^{M} w_{ij}^{*} S_4(\sigma_2^2;z_i,y_i^{*(j)}) = 0$$

[Step4] Step2과 Step3을 반복하여 모수 추정치가 수렴할 때까지 반복한다.

다음으로 모수 추정량에 대한 분산추정에 대하여 살펴본다. EM알고리즘을 통하여 추정하고자 하는 모수를 $\eta=(\beta,\sigma_1^2,\gamma,\sigma_2^2)$라 할 때, 모수 추정량 $\hat{\eta}=(\hat{\beta},\hat{\sigma_1^2},\hat{\gamma},\hat{\sigma_2^2})$은 다음의 식 $\overline{S}^{*}(\eta)=0$ 해이다.

$$\overline{S}^{*}(\eta) = \sum_{i}^{n}\sum_{j=1}^{M} w_{ij}^{*} s_{ij}^{*}(\eta) = \begin{pmatrix} \sum_{i}^{n}\sum_{j=1}^{M} w_{ij}^{*} S_1(\boldsymbol{\beta};\boldsymbol{x}_i,y_i^{*(j)}) \\ \sum_{i}^{n}\sum_{j=1}^{M} w_{ij}^{*} S_2(\sigma_1^2;\boldsymbol{x}_i,y_i^{*(j)}) \\ \sum_{i}^{n}\sum_{j=1}^{M} w_{ij}^{*} S_3(\gamma;z_i,y_i^{*(j)}) \\ \sum_{i}^{n}\sum_{j=1}^{M} w_{ij}^{*} S_4(\sigma_2^2;z_i,y_i^{*(j)}) \end{pmatrix}.$$

여기서 $\overline{S}^{*}(\eta)=\sum_{i=1}^{n} w_i \overline{S}_i^{*}(\eta)$, $\overline{S}_i^{*}(\eta)=(\overline{S}_{1i}^{*}(\beta), \overline{S}_{2i}^{*}(\sigma_1^2), \overline{S}_{3i}^{*}(\gamma), \overline{S}_i^{*}(\sigma_2^2))'$로 나타낼 수 있으며 $\overline{S}_{1i}^{*}(\beta)=\sum_{j=1}^{M} w_{ij}^{*} S_1(\beta, \boldsymbol{x}_i, y_i^{*(j)})$로 정의되며 $\overline{S}_{2i}^{*}(\sigma_1^2), \overline{S}_{3i}^{*}(\gamma), \overline{S}_i^{*}(\sigma_2^2)$ 각각에 대해서도 유사하게 정의 된다. $\hat{\eta}$에 대한 분산추정은 $\hat{I}_{obs}^{*}(\eta)$를 통해 이루어지며 다음과 같이 정의된다. 자세한 계산 방법은 Kim (2011)에 나타나 있다.

$$\hat{I}_{obs}^{*}(\eta)=-\sum_{i=1}^{n}\sum_{j=1}^{M} w_i w_{ij}^{*}(\eta) S_{ij}^{*}(\eta)\left\{S_{ij}^{*}(\eta)-\overline{S}_i^{*}(\eta)\right\}'-\sum_{i=1}^{n} w_i \dot{S}_i(\eta).$$

4. 한국복지패널 히핑 현상 보정 결과

4.1 히핑 발생 여부에 대한 로지스틱 회귀분석 결과

앞선 히핑 현상 현황 분석결과를 바탕으로 1차 년도 복지패널 중 경상소득에 대하여 경상소득값이 120의 배수인 경우 히핑 현상이 발생한 것으로 지정한다. 히핑 존재 여부와 여러 설명 변수들간의 관계를 살펴본다. 설명 변수로는 지역구분, 균등화 소득에 따른 가구 구분, 가구원수, 가구 중 성별, 가구주 교육 수준과의 관계를 사용하였다.

히핑 발생 여부에 대한 확률 $\Pr(I_h=1|x_i)$에 대한로지스틱 회귀분석 추정에 관한 최종 모형으로 지역구분, 일반/저소득 가구 구분, 가구원수, 가구주 나이, 가구주 고유 수준을 설명변수로 사용하였다. 로지스틱 회귀모형에 대한 변수별 회귀 계수값, 표준오차 및 유의확률이 <표 4.1>에 나타나 있다.

<표 4.1 > 히핑 존재 여부에 대한 로지스틱 회귀분석 결과

	Coefficient	S.E	유의확률
(Intercept)	0.205	0.256	0.423
인천/경기	-0.111	0.118	0.349
부산/울산/경남	-0.181	0.133	0.173
대구/경북	0.078	0.143	0.588
대전/충남	-0.177	0.174	0.307
강원/충북	-0.548	0.209	0.009
광주/전남북/제주	-0.300	0.149	0.044
저소득가구	-0.340	0.101	0.001
2인가구	-0.574	0.129	0.000
3인가구	-0.298	0.127	0.019
4인가구	-0.361	0.124	0.003
5인이상가구	-0.504	0.170	0.003
가구주 나이	-0.039	0.004	0.000
가구주 학력 고졸이하	0.357	0.122	0.003
가구주 학력 대학이상	0.000	0.137	0.999

로지스틱 분석 결과를 보면 다른 변수의 효과를 제거하였을 때, 강원, 충북, 광주, 전남, 전북, 제주는 다른 지역에 비하여 히핑이 발생할 확률이 작게 나타난다. 저소득 가구는 일반가구에 비하여 히핑 발생 확률이 작다. 가구원수가 증가할수록 히핑이 발생할 확률은 작아진다. 가구주 나이가 많을수록 중졸이하의 교유수준일 때 히핑 발생 확률은 작게 나타난다.

4.2 확률 대체법을 통한 히핑 보정 결과

1차 년도 복지패널 경상소득에 대한 히핑 보정을 수행하였다. 설명 변수로는 지역구분, 균등화 소득에 따른 가구 구분, 가구원수, 가구주 나이, 가구주 교육 정도를 사용하였다. 모형 (3.2)와 (3.3)에 대한 모수 추정 결과는 <표 4.2>와 <표 4.3>에 각각 나타나 있다.

<표 4.2> 구조오차 모형 모수 추정 결과

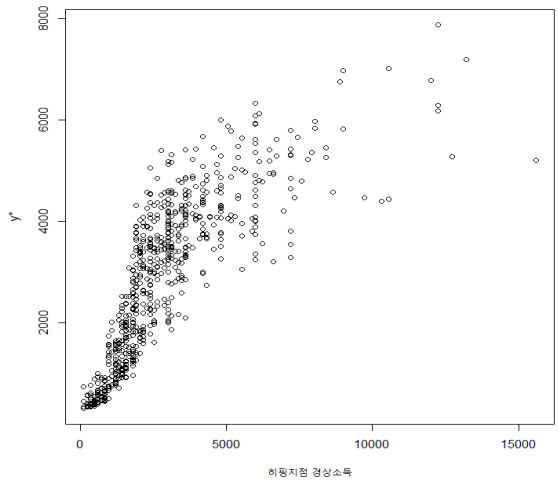
모수	Coefficient	S.E
상수항	7.606	0.012
인천/경기	-0.060	0.006
부산/울산/경남	-0.053	0.006
대구/경북	-0.156	0.007
대전/충남	-0.091	0.008
강원/충북	-0.133	0.008
광주/전남북/제주	-0.126	0.006
저소득가구	-1.333	0.004
2인가구	0.490	0.005
3인가구	0.802	0.006
4인가구	0.991	0.006
5인이상가구	1.120	0.008
가구주 나이	-0.004	0.000
가구주 학력 고졸이하	0.040	0.005
가구주 학력 대학이상	0.268	0.006
σ_1^2	0.023	

<표 4.3> 측정오차 모형 모수 추정 결과

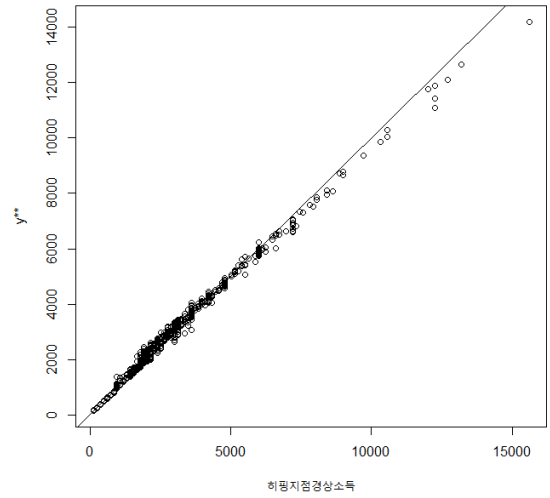
모수	Coefficient	S.E
γ_0	0.634	0.036
γ_1	0.907	0.005
σ_2^2	0.158	

추정된 모수와 최종 fractional weights를 사용하여 얻어진 히핑 보정 대체값 y^*와 조사된 응답값과의 관계가 <그림 4.1>에 나타나 있다. 히핑 보정을 통해서 값들이 전

체적으로 줄어들었지만 분산이 경상소득이 증가함에 따라 커지는 모습을 볼 수 있다. 히핑 확률까지 고려한 히핑 보정 대체값 y^{**}와 조사된 응답값과의 관계는 <그림 4.2>에 나타나 있다. 히핑 보정을 통해서 값들이 전체적으로 줄어들었으며 그 조사된 응답값과의 차이도 크지 않다. 히핑이 선행연구에서 가정하듯이 응답값 기준 일정 범위에서 발생한다는 가정하에서는 조금 더 타당한 보정 방법이라고 생각된다.



<그림 4.1> 응답 경상소득 값과 대체값 y^{*}의 산점도



<그림 4.2> 응답 경상소득 값과 확률 반영 대체값 y^{**}의 산점도

균등화 소득에 의한 저소득 가구의 비율은 1차 년도에 46.3%로 계산된다. 이후 차수년도에서는 41%에서 43% 사이로 저소득 가구의 비율이 나타나는 점으로 보아 1차

년도의 높은 저소득 가구 비율이 히핑으로 인하여 나타나는 현상인지 살펴보고자 한다. <표 4.4>에는 보정 전, 후의 저소득 가구에 대한 평균과 표준편차 값이 나타나 있다. 저소득 가구의 경상소득이 증가하는 방향으로 또 표준편차는 줄어드는 방향으로 보정이 실행되었음을 알 수 있다. 경상소득에 대하여 중앙값*0.6을 기준으로 하위 소득 가구의 비율을 살펴보았다. <표 4.5>에 각 차수별 기준 소득과 하위 소득 가구의 비율이 나타나 있다. 마지막 행에 1차 년도에 히핑 보정된 경상소득을 통해서 계산된 결과가 나타나 있다. 보정 전 1차 년도와 크게 차이가 없어 1차 년도의 저소득 가구의 비율이 높은 부분에 있어서 히핑은 큰 원인은 아니라고 판단된다.

<표 4.4> 히핑 보전 전, 후의 저소득 가구의 경상소득 평균, 표준편차

	보정 전		보정 후	
	평균	표준편차	평균	표준편차
저소득 가구	780.1	546.3	810.2	450.0

<표 4.5> 경상소득 중위값*0.6 기준 하위소득 가구 비율

	기준값	비율
1차	1,032.6	34.7%
2차	1,195.8	33.8%
3차	1,293.3	33.2%
4차	1,413.6	33.5%
5차	1,470.0	32.4%
6차	1,549.2	33.4%
7차	1,440.0	32.0%
8차	1,532.4	32.6%
9차	1,509.3	33.9%
10차	1,549.8	33.5%
11차	1,636.2	33.6%
12차	1,726.8	33.1%
13차	1,767.6	32.9%
14차	1,792.2	32.7%
15차	1,884.0	33.0%
1차_보정	1,049.4	34.5%

5. 결론 및 논의

본 논문에서는 자기응답으로 조사된 자료에서 발생하는 측정오차 보정에 대하여 살펴보았다. 이러한 측정오차 중 반올림된 응답값 혹은 0 또는 5와 같은 선호하는 자리수에 응답값이 집중되는 히핑 현상 보정을 중심으로 측정오차 보정 방법론을 제안하였다. 히핑 현상 보정을 위하여 관측값과 보조 변수들을 사용하여 참값을 생성하는 대체법을 이용하였다. 모든 응답값들을 대체 하는 것이 아니라 히핑 현상이 발생할 확률에 따른 확률 대체법을 제안하였다.

제안한 방법을 1차 년도 한국복지패널조사 자료의 경상소득 변수에 적용한 결과, 보정 후의 값들의 산포가 전체적으로 줄어들었으며 조사된 응답값과의 차이도 일정 구간에 속하였다. 본 연구결과를 바탕으로 조사자료의 측정오차를 보정할 수 있는 방법을 다양하게 모색할 수 있을 것이다. 향후 연구에서는 히핑 발생 확률이 참값에 의존하는 모형으로 확장하여 고도화된 확률 대체 방법론을 다루도록 한다.

(2022년 10월 25일 접수, 2022년 11월 8일 수정, 2022년 11월 8일 채택)

참고문헌

측정오차에 대한 통계적 보정 방법론 연구(패널자료 품질개선 연구(Ⅲ)), 김재광 (2014)

Cummings, T. H., Hardin, J. W., McLain, A. C., Hussey, J. R., Bennett, K. J. and Wingood, G. M. (2015). Modeling heaped count data. *The Stata Journal*, 15, 457-479.

Groß, M. and Rendtel, U. (2016). Kernel density estimation for heaped data. *Journal of Survey Statistics and Methodology*, 4, 339-361.

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98, 119-132.

Kim, J. K. and Hong, M. (2012). Imputation for statistical inference with coarse data. *Canadian Journal of Statistics*, 40, 604-618.

A Probablistic Imputation Approach for Handling Heaping in the Self-reported Survey Data

Seunghwan Park[1)]

Abstract

The true value and the measured value in the self-reported survey data differ because of subject sensitivity and memory issues. When the focus of the answer value is on the rounded response value or the number of desired digits, such as 0 or 5, heaping, a type of these errors, happens. However, statistical analysis using heaped data can result in biased analysis results if this heaping phenomena does not occur randomly. In this study, we suggest an imputation approach for reducing measurement errors due to heaping. Variance estimation is also investigated. Using the proposed method on the current income variable of the first Korean Welfare Panel Survey resulted in a narrower distribution of the corrected values, and the difference from survey respondents' responses likewise belonged to a particular range.

Key words : Imutation approach, Heaped data, Measurement error model

1) (Corresponding author) Professor, Dept. of Information Statistics, Kangwon National University, 1 GANGWONDAEHAKGIL, CHUNCHEON-SI, GANGWON-DO, 24341 REPUBLIC OF KOREA. E-mail: stat.shpark@kangwon.ac.kr