연구보고서 2021-22

조사 자료의 품질 검증 연구

– 측정오차를 중심으로

이혜정
신지영·박승환·지희정·오미애

사람을
생각하는
사람들

KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

한국보건사회연구원
KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

▌연구진

연구책임자	**이혜정**	한국보건사회연구원 부연구위원
공동연구진	**신지영**	한국보건사회연구원 연구원
	박승환	강원대학교 정보통계학과 교수
	지희정	前 고려대학교 의학통계교실 박사
	오미애	한국보건사회연구원 연구위원

연구보고서 2021-22

조사 자료의 품질 검증 연구
- 측정오차를 중심으로

발 행 일 2021년 12월
발 행 인 이 태 수
발 행 처 한국보건사회연구원
주 소 [30147]세종특별자치시 시청대로 370
세종국책연구단지 사회정책동(1~5층)
전 화 대표전화: 044)287-8000
홈페이지 http://www.kihasa.re.kr
등 록 1999년 4월 27일(제2015-000007호)
인 쇄 처 고려씨엔피

ISBN 978-89-6827-819-8 93330
https://doi.org/10.23060/kihasa.a.2021.22

발|간|사

조사 자료는 관심 모집단에서 일부 대상만을 추출하여 표본조사를 통해 구축한 것이다. 구축된 조사 자료는 실증분석 연구 시 모집단 전체에 대해 추론한 추정값을 통해 의미 있는 연구 결과를 도출하는 데 활용된다. 이때 추정값과 참값 간 차이가 발생하기 마련이며, 추정값과 참값의 차이를 오차라고 정의한다.

보통 조사 자료는 오차를 포함하고 있으며 오차의 정도가 심각한 자료를 사용하여 통계분석을 한다면 심각하게 편향된 분석 결과를 도출하게 된다. 오차는 크게 표집오차와 비표집오차로 구분되며, 비표집오차가 차지하는 비중은 표집오차에 비해 훨씬 크기 때문에 축소하기 위한 노력이 필요하다. 이와 관련한 대체 방안 연구가 외국의 경우 오래전부터 진행되었으나, 국내에서는 활발하게 이루어지지 않았다.

이 연구에서는 표본조사 자료에서 발생할 수 있는 오차에 대해 살펴보았으며, 그중에서 측정오차와 관련하여 심도 있게 검증하였다. 표본조사 자료에 측정오차가 있는 경우에 대한 보정 방법으로 측정오차 보정을 통한 회귀계수 추정, 커널 함수 추정법을 활용한 히핑 보정, 비례핫덱대체 방법을 활용한 보정 등을 다양하게 모색하였고, 실제 자료에서의 활용 방안을 제시하였다. 한편, 기존 히핑 보정 방법을 확장하여 새로운 방법을 제안하였다. 마지막으로는 주요 해외 표본조사 자료에서의 측정오차 관리 방안 사례 조사와 전문가 자문회의 등을 통해 고품질 자료 생산을 위한 조사 자료 관리 방안을 제안하였다.

이 연구는 이혜정 부연구위원의 책임하에 오미애 연구위원, 신지영 연구원이 연구진으로 참여하였다. 외부 연구진으로 강원대학교 박승환 교수, 前 고려대학교 의학통계교실 지희정 박사가 참여하였다. 모든 연구진

의 노고에 감사드린다. 보고서 작성과 관련하여 유익한 의견을 주신 원내 이원진 부연구위원, 원외 고려대학교 송주원 교수, 한국리서치 박종선 이사, 그리고 익명의 검독위원들에게도 감사의 마음을 전한다.

마지막으로 이 보고서의 내용은 우리 연구원의 공식적인 견해가 아니라 연구진의 의견임을 밝힌다.

2021년 12월

한국보건사회연구원 원장

이 태 수

목 차

표 목차

그림 목차

Abstract

A Study on the Survey Data Quality – Focusing on Measurement Errors

Project Head: Lee, Hyejung

Survey data are usually constructed through sampling by extracting some subjects from the population of interest. Since we make inferences about the entire population with the survey data, there will be a difference between the sample estimate and the true population value. The difference between the sample estimate and the true population value is defined as an error, and the error can occur by various causes and situations.

The purpose of this study is to examine the quality of survey data and to suggest ways for quality improvement. This study analyzes the measurement errors occurred in the sample survey data, and proposes various methods for correcting them.

Main results of this study are as follows: Measurement error correction methods were performed by using the survey data. Measurement error correction methods include measurement error correction in the linear regression models with continuous outcomes, kernel density estimation for heaped data, and fractional hot deck imputation method. An R package that implements measurement error correction methods for re-

Co-Researchers: Shin, Ji-young·Park, Seunghwan·Jee, Hee-Jung·Oh, Miae

gression models with continuous outcomes is the function ‘mecor’ in the package ‘mecor’. An R package that implements measurement error correction methods for kernel density estimation for heaped data is the function ‘dheaping’ in the package ‘Kernelheaping’. An R package that implements measurement error correction methods for fractional hot deck imputation method is the function ‘FHDI’ in the package ‘FHDI’. As an R package has some packages for measurement error corrections, most researchers can use it without difficulty.

Meanwhile, a survey data management for producing high-quality data was proposed through a case study of measurement error management methods in major overseas sample survey data.

It is fundamentally difficult to eliminate measurement errors. However, this can be solved by active participation of the respondents, sincere conduct and appropriate management of the interviewers, and planning of the surveys reflecting the reality.

Keyword : measurement error, heaping, survey data

요약

1. 연구의 배경 및 목적

실증분석 연구는 조사 자료를 기반으로 하여 의미 있는 연구 결과를 도출하는 데 있다. 조사 자료는 보통 관심 모집단에서 일부 대상만을 추출하여 표본조사를 통해 구축한 것이다. 이러한 조사 자료를 가지고 모집단 전체에 대해 추론하므로 추정값과 참값 간 차이가 발생하기 마련이다. 추정값과 참값의 차이를 오차라고 정의하며 오차는 다양한 원인과 상황에서 발생할 수 있다.

사회과학 분야의 조사에서 참여자의 응답은 주관적인 판단 개입으로 정확하게 수집되지 않을 수 있다. 예를 들어, 가구소득이 얼마나 되는지에 관한 질문에 응답자는 응답에 대한 거부감 또는 세금 문제 때문에 솔직한 응답을 꺼리거나, 과거를 기억하지 못하거나, 설문을 잘못 이해하는 경우 등의 다양한 이유로 정확하게 응답하지 않는 경우가 이에 해당한다. 또한 조사원이 설문 진행을 거짓으로 하거나, 조사원이 잘못 이해하고 설문 응답을 받거나, 자료입력 과정에서의 단순 오류 등도 있다. 이렇듯 조사 자료는 보통 오차를 포함하고 있으며 오차의 정도가 심각한 자료를 사용하여 통계분석을 한다면 심각하게 편향된 분석 결과를 도출하게 된다.

조사에서 발생 가능한 모든 오차를 총오차라고 하며, 크게 표집오차와 비표집오차로 구분할 수 있다. 표집오차는 모집단 전체가 아니라 그중 일부인 표본을 조사하기 때문에 발생하는 오차이다. 비표집오차는 표집오차를 제외한 모든 오차, 측정오차, 포함오차, 무응답오차, 처리오차 등이 해당한다.

본 연구의 목적은 조사 자료의 품질을 검증하고, 이에 따른 관리 방안

을 마련하는 데 있다. 표본조사 자료에서 발생할 수 있는 오차를 살펴보고, 그중에서 측정오차에 대해 심도 있게 검증해 보고자 한다. 표본조사 자료에 측정오차가 있는 경우에는 보정하는 방안도 다양하게 모색하여 제안할 것이다. 또한, 주요 해외 표본조사 자료에서의 측정오차 관리 방안 사례를 조사하여 고품질 자료 생산을 위한 조사 자료 관리 방안을 마련하고자 한다.

2. 주요 연구 결과

2장의 주요 연구 결과를 보면, 해외 주요 패널조사(미국 PSID, SIPP, 유럽 SHARE, 영국 BHPS)의 조사 품질 향상 방안 및 측정오차 처리 방법은 사후적인 보정에 앞서 측정오차로 인한 편향을 줄이기 위해 근본적으로 조사 설계 및 진행, 조사참여자의 응답, 데이터 처리 과정에서의 발생 가능성을 낮추려는 다양한 기술적, 시스템적 접근이 이루어지고 있었다. 그리고 기술적 접근으로 컴퓨터를 이용한 전화조사(CATI) 또는 컴퓨터를 이용한 대면조사(CAPI), 이전 조사에서 응답한 결과를 바탕으로 공통된 내용에 대한 설문을 대체하는 방식인 종속형 설문(Dependent Interviewing: DI), 응답자의 회고오차 발생을 억제하는 데 도움을 주는 Event History Calendar(EHC)의 방식 등을 도입하고 있었다. 시스템적 접근으로 응답자의 응답 부담 감소 및 적극적인 참여 유도를 위해 별도의 인센티브(사례비) 제공, 조사응답률 및 응답자의 설문 이해도를 높이기 위해 조사 시작 전에 별도의 우편물 발송, 조사원에 대한 체계적인 교육 프로그램 구축 및 감독관의 감독·평가 의무화를 시행하고 있었다. 또한 조사원이 체감하는 애로사항을 시스템 개선에 반영하기 위하여 정기적인 조사원 대상 설문조사를 실시하는 것으로 나타났다. 이러한 내용은 결론

에서 측정오차 관리 방안 제안 시 포함하여 다루었다.

3장의 주요 연구 결과는 다음과 같다. 자기응답으로 조사된 자료에서 발생하는 히핑 현상에 대해 1~15차 연도 한국복지패널조사 자료의 경상소득을 중심으로 살펴본 바, 주로 처음 응답한 조사차수, 1차 연도와 신규 표본의 첫해 조사에서 다른 조사 차수에 비해 높은 비율로 히핑 현상이 나타났다. 경상소득에 대한 히핑 존재 여부와 가구 특성 변수 간의 관계를 로지스틱 회귀분석을 통하여 살펴본 결과, 강원, 충북, 광주, 전남, 전북, 제주는 다른 지역에 비하여, 저소득 가구는 일반 가구에 비하여, 가구원 수가 증가할수록, 가구주 나이가 많을수록, 중졸 이하의 교육 수준일 때 히핑 발생 확률이 작게 나타났다. 홍민기 등(2014)은 히핑 보정 방법을 확장한 새로운 방법을 제안하였는데, 히핑 발생 확률을 반영한 대체를 통한 히핑 보정 방법이다. 이 방법을 한국복지패널조사 자료의 경상소득에 적용한 결과, 보정 후의 값들이 전체적으로 줄어들었으며 조사된 응답값과의 차이도 일정 구간에 속하였다. 선행연구에서 가정하듯이, 히핑은 응답값 기준 일정 범위에서 발생한다는 가정하에서는 조금 더 타당한 보정 방법이라고 생각한다. 일반 가구와 저소득 가구를 나누어 히핑 보정된 소득의 분포를 살펴보면 저소득 가구의 경상소득이 증가하는 방향으로, 또 표준편차는 줄어드는 방향으로 보정이 실행되었음을 알 수 있었다. 한편, 경상소득에 대하여 (중위값×0.6)의 값을 기준으로 하위소득 가구의 비율은 히핑 보정 전 1차 연도와 크게 차이가 없는 것으로 나타나, 히핑이 1차 연도의 저소득 가구의 비율이 높은 것의 큰 원인은 아니라고 생각한다.

4장에서는 2017년 가계금융복지조사 자료에서 가구주의 개인 근로소득, 가구의 근로자녀장려금, 가구원의 기초연금에 대한 응답값에 측정오차를 포함하고 있는지를 살펴보았고, 측정오차 보정을 통한 회귀계수 추정, 히핑 보정, 비례핫덱대체 방법을 사용하여 측정오차를 보정하였다.

근로소득, 근로자녀장려금과 기초연금에 대한 측정오차의 구조는 차이가 있는 측정오차(differential measurement error)였다. 차이가 있는 측정오차는 응답값이 관심 설명변수에 따라 행정보완값에 대해 체계적인 경우를 의미하며, 이에 대한 측정오차 보정 방법을 사용하여 회귀계수를 추정하였다. 관심 설명변수의 회귀계수는 모두 행정보완값의 회귀계수와 정확하게 일치하는 것을 확인하였다. 그러나 대부분의 보통 자료에서는 참값을 알고 있는 경우가 드물어서 일반 연구자는 참값을 알고 있지 않을 가능성이 크므로 이러한 측면에서 봤을 때 유용한 방법이라고 볼 수 있다. 한편 응답값과 행정보완값 간의 차이가 크고 연관성이 낮을수록 표준오차는 커지는 것으로 나타났다. 이로 인해 근로자녀장려금과 기초연금은 보정 후 회귀계수의 유의성이 보정 전과 다른 결과를 가졌다. 차이가 있는 측정오차 보정을 통한 회귀계수 추정 분석의 한계점은 설명변수 형태(type)를 이분형 범주만 사용 가능하다는 점, 측정오차에 영향을 미치는 설명변수에 대한 회귀모형 분석만 가능하여 더 많은 관심 설명변수의 영향을 파악할 수 없다는 점을 들 수 있다. 그러나 여러 설명변수가 함께 오차 없는 측정값과 연관되는 것을 현실적으로 모형화하기 어렵기 때문에, 본 연구와 같이 하나의 변수에 국한하는 경향이 있다고 볼 수 있다.

한편 측정오차 보정을 통한 회귀계수 추정 시 내부보정방법을 사용하였는데, 이때 가계금융복지조사 자료는 응답자가 응답값과 행정보완값을 모두 가지고 있어서 측정오차 보정 시 모두 활용하였다. 보통 한정적인 예산과 실측 자료 수집의 어려움 등으로 이와 같은 형태의 자료를 사용하기는 쉽지 않은 편이다. Nab, Groenwold, Welsing, and van Smeden(2019)의 논문에서는 보정 표본(calibration sample)의 크기, 측정오차가 있는 값과 없는 값 간의 상관관계 정도에 따른 측정오차의 보정 효과에 대한 모의실험을 통해 다음과 같은 결과를 도출하였다. 약한

상관관계(R^2=0.2)를 가지는 경우, 외부보정 표본의 규모를 50개까지 늘리기 전에는 측정오차의 보정 결과가 효과적이지 않았다. 그러나 보통 이상의 상관관계(R^2=0.5 또는 0.8)인 경우, 외부보정 표본의 규모가 15개로 적은 편이어도 회귀계수 보정 결과가 향상됨을 보였다. 이렇듯 측정오차를 판단할 수 있는 실측 자료(참값, 실측값 등)를 전체가 아닌 일부라도 구축하고 계속해서 측정오차 정도를 모니터링하는 방안을 고려해 볼 수 있다고 생각한다.

다음으로 근로소득에 대해 히핑 보정을 하였는데, 히핑 보정 후 결과는 전반적으로 부드러운 분포 함수 형태를 보였다. 히핑 보정 후 평균도 행정보완값과의 차이가 작았다. 히핑 보정 전 평균은 행정보완값과의 차이가 더 큰 편이었다. 그러나 히핑 보정 후 표준편차가 행정보완값 보다 작게 나타나, 내부보정 또는 외부보정 자료를 추가 정보로 사용하여 히핑 보정을 고려해 볼 수 있을 것이다.

마지막으로 근로자녀장려금의 응답값이 0원인 경우에 대해 0원을 무응답으로 간주한 다음, 비례핫덱대체 방법을 사용하여 적절한 값으로 대체하였다. 근로자녀장려금의 무응답 비율은 81.9%로 아주 높은 편이었다. 대체 시 활용한 대체 방법은 3가지로, 비례핫덱대체 방법에서 대체군 활용 변수가 3개, 4개인 경우와 회귀대체 방법이었다. 무응답 비율이 매우 높은 편이라는 점에서 대체 시 활용할 수 있는 자료(응답값 정보)가 충분하지 않아서 전반적으로 모든 대체 방법의 효과에 영향을 주었다. 그러나 대체 방법 중에서 비례핫덱대체 방법의 대체군 활용 변수가 4개인 경우의 대체 효과가 우수한 결과를 가지는 것으로 나타났다. 이는 무응답 변수와 연관성이 높은 행정보완값을 대체군 활용 변수로 사용하였고, 대체값을 여러 개 생성한 다음 최종 하나의 값으로 산출하여 대체값에 대한 변동 부분을 고려하였기 때문이라고 볼 수 있다. 회귀대체 방법도 행정보

완값을 설명변수로 사용하였으나, 최적의 회귀모형 적합이 되지 않은 점이 원인이라고 생각한다.

3. 결론 및 시사점

실제 조사 자료에서 연속형 관심 변수에 대한 측정오차를 보정하는 방안은 다음과 같다. 먼저 관심 변수에 대해 측정오차(히핑 포함)를 가지는지 현황을 파악하고, 관심 변수값이 측정오차를 가지는 경우 측정오차를 보정한다.

관심 변수에 대한 측정오차의 포함 여부는 그림(plot)을 통해 응답값과 실제값의 분포 파악, 대응 표본 t-검정 실시, 다항 로지스틱 회귀모형 분석 등을 통해 파악할 수 있다. 또한 히핑 현상을 살펴볼 때는 관심 변수값에 대한 상위 응답 비중, 특정 배수의 형태를 가지는지에 대한 분석, 그림을 통해 확인할 수 있다.

측정오차 보정 방법으로는 측정오차 보정을 통한 회귀계수 추정, 커널함수 추정법을 활용한 히핑 보정, 비례핫덱대체(FHDI) 방법을 활용한 보정 등이 있다. 측정오차 보정을 통한 회귀계수 추정을 위해서는 실측 자료가 갖춰져 있어야 하지만, 해당하는 모든 응답값에 대해 실측 자료를 가지고 있을 필요는 없다. 측정오차 보정을 통한 회귀계수 추정은 R 통계패키지에 'mecor' 패키지 내부의 'mecor' 함수를 사용하면 된다. 커널함수 추정법을 활용한 히핑 보정은 R 통계패키지에 'Kernelheaping' 패키지 내부의 'dheaping' 함수이다. 비례핫덱대체 방법을 활용한 보정은 R 통계패키지에 'FHDI' 패키지 내부의 'FHDI' 함수이다. 이외에 적합한 다른 대체 방법을 활용하여 관심 변수에 대한 측정오차를 보정할 수 있다. 이렇듯 R 통계패키지에는 측정오차 보정을 위한 패키지가 있어 일반

연구자도 어렵지 않게 활용할 수 있다는 점에서 유용하다고 생각한다.

또한 측정오차로 인한 편향을 줄이기 위해 사후적인 보정에 앞서 근본적으로 조사 설계 및 진행, 조사참여자의 응답, 데이터 처리 과정에서의 발생 가능성을 낮추려는 노력이 필요하다. 마지막으로, Kasprzyk(2005)가 언급한 대로 측정오차의 원인을 4가지로 구분하여 관리 방안을 제안하고자 한다. 측정오차의 원인 4가지는 설문 내용, 자료 수집 방법, 조사원, 응답자로 구분되며, 이에 따른 관리 방안은 〈요약표 1-1〉과 같다.

〈요약표 1-1〉 측정오차 관리 방안

구분	관리 방안
설문 내용	- 설문 구성에 있어서 사전·사후 정성 조사: 설문 문구 수정 및 사후 결과 해석에 활용함 - 과감한 설문 축소 - 설문 문항을 응답자가 이해하기 쉽도록 최대한 쉽게 표현함 - 종속형 설문 및 Event History Calendar 도입
자료 수집 방법	- 리뷰 또는 검증 등을 통한 사후 자료 확인 및 재조사 - 일부의 실측 자료(참값, 실측값 등) 구축 및 활용 - Paradata를 활용하여 조사 과정 분석 - 관련 영수증 제출, 실제값 측정 등을 통한 자료 수집 - 행정 자료 등을 다양하게 활용하여 조사 자료와 연계
조사원	- 조사원의 교육 강화 - 조사원 교육 후 테스트 - 조사 초기 리뷰 강화로 조사원에 대한 피드백 제공 - 조사에 대한 조사원 태도를 사후 파악하여 피드백 제공 - 조사원의 처우 개선
응답자	- 금전적 인센티브의 현실화 필요성 - 정서적 인센티브 고취: 조사의 중요성을 강조하여 책임감이나, 국가 정책 수립에 관여한다는 효능감 등 고취 - 제도적 인센티브 : 세금 감면, 전기료 인하, 봉사활동 점수 제공 등

자료: 2장 내용 및 필자 작성

근본적으로 측정오차를 모두 해결하기는 어려우나, 응답자의 적극적인 참여, 조사원의 성실한 진행 및 조사원에 대한 적절한 관리, 연구자의 현

실을 고려한 기획 등을 통해 상당 부분 해결이 가능할 것이므로 모든 조사 단계가 철저하게 관리되어야 할 것이다.

키워드 : 측정오차, 히핑, 조사 자료

제1장

서론

제 1 장 서론

제1절 연구 배경 및 목적

실증분석 연구는 조사(survey)를 통하여 수집한 자료(조사 자료)를 기반으로 하여 의미 있는 연구 결과를 도출하는 데 있다. 보통 조사 자료는 관심 모집단에서 일부 대상만을 추출하여 표본조사를 통해 구축한 것이다. 이러한 조사 자료를 가지고 모집단 전체에 대해 추론하므로 추정값(sample estimate)과 참값(true population value) 간 차이가 발생하기 마련이다. 추정값과 참값의 차이를 오차라고 정의하며 오차는 다양한 원인과 상황에서 발생할 수 있다.

현실에 편재하는 통계분석에서 정확한 측정값을 얻는 것은 어렵거나 현실적이지 않을 것이다(Guo, 2010). 사회과학 분야의 조사에서 참여자의 응답은 주관적인 판단 개입으로 정확하게 수집되지 않을 수 있기 때문이다. 예를 들면 가구소득이 얼마나 되는지에 관한 질문에 응답자는 응답의 거부감, 세금 문제 때문에 솔직한 응답을 꺼리거나, 과거를 기억하지 못하거나, 설문을 잘못 이해하는 경우 등의 다양한 이유로 정확하게 응답하지 않는 경우가 이에 해당한다. 또한 조사원이 설문 진행을 거짓으로 하거나, 조사원이 잘못 이해하고 설문 응답을 받거나, 자료입력 과정에서의 단순 오류 등도 있다. 임상 연구나 역학연구에서도 모호한 조사, 부정확한 도구 사용, 잘못 설계된 설문 문항, 생물학적인 변동, 고비용의 표준 방법 등과 같이 다양한 원인으로 측정오차를 포함하고 있을 가능성이 큰 편이다(Guo, 2010).

이렇듯 조사 자료는 보통 오차를 포함하고 있으므로 오차의 정도가 심각한 자료를 사용하여 통계분석을 한다면 심각하게 편향된 분석 결과를 도출하게 된다. 보통 측정오차 보정(measurement error correction), 측정오차가 있는 자료를 결측치로 간주하고 대체(imputation)하는 방법, 가중치 조정(weighting adjustment) 등의 방법이 편향을 보정하는 데 사용된다(Zalsha, 2020).

조사에서 발생 가능한 모든 오차를 총오차(total survey error)라고 하며, 크게 표집오차(sampling error)와 비표집오차(non-sampling error)로 구분할 수 있다. 표집오차는 모집단 전체가 아닌 그중 일부인 표본을 조사하기 때문에 발생하는 오차이다. 비표집오차에는 측정오차(measurement error), 포함오차(coverage error), 무응답오차(non-response error), 처리오차(processing error) 등이 있다. 즉 표집오차를 제외한 모든 오차를 비표집오차라고 볼 수 있다.

비표집오차는 표집오차에 비해 총오차에서 차지하는 비중이 훨씬 크기 때문에(김정섭, 임경은, 2010) 비표집오차를 축소하기 위한 노력이 필요하다. 표집오차는 표본 수가 증가하면 작아지는 특성이 있으므로, 발생하는 오차의 정도를 이론적으로 측정할 수 있다. 그러나 비표집오차는 측정이 쉽지 않고 이론적인 설명이나 추정 또한 어려운 편이다. 또한 비표집오차는 조사 기획 및 실사의 전 과정에서 발생한다. 비표집오차 중에서 측정오차에 대한 전통적인 보정 방법은 오차 분포의 모수를 타당성 자료 또는 반복 측정 자료에서 추정하는 것이다. 이러한 방법으로 조정된 추정량은 비편향이거나 일치성을 가지게 된다.

외국의 경우 비표집오차를 줄이기 위한 노력의 일환으로 측정오차에 대한 문제를 인지하고 그 대처 방안에 관한 연구가 오래전부터 진행해 왔다. 그러나 국내의 경우 측정오차 관련 연구는 초기 단계라고 할 수 있다

고(김준원, 신동균, 2016) 하였으나, 이러한 상황에도 불구하고 측정오차 관련 국내연구 동향은 활발하지 않은 편으로 나타났다. 홍민기, 김재광, 한치록, 김기민(2014)은 한국노동패널조사 자료를 사용하여 히핑(heaping) 현상에 대해 연구하였다. 히핑 현상은 응답자가 대략적인 값으로 응답하여 발생하는 측정오차의 한 가지 형태로 볼 수 있다. Kim and Hong(2012)이 제시한 MCEM방법을 이용하여 히핑된 자료값을 대체하여 히핑을 보정한 결과, 소득분포가 상당히 부드럽게 되었다. 측정오차가 있는 자료에서 측정오차를 줄여서 분석하고자 할 때 자료변환의 방법으로 대체방법론을 활용하는 것을 제안하여 모수적 비례대체(parametric fractional imputation) 방법을 사용하였고, 분석 결과 응답자료 보다는 대체자료를 사용하여 분석한 추정량이 편향을 더 줄여주는 것으로 나타났다. 박인호, 전경배, 주성제(2017)는 비표집오차와 관련한 기존 연구를 검토하고 한국은행 서베이 통계를 연구하여 포함오차, 무응답오차, 기타 표집오차들에 대해 살펴보았으며, 비표집오차 축소를 위한 체크리스트 예시를 제시하였다. 그리고 기관 차원의 서베이 통계에 대한 지속적인 품질향상을 위한 고려사항에 대해 논의하였다.

연구 목적은 조사 자료의 품질을 검증하고, 이에 따른 관리 방안을 마련하는 데 있다. 통계청에서는 통계품질진단 시 5가지 차원[1]의 품질 수준이 어느 정도인지를 측정하고, 각 차원의 품질 수준을 높이기 위해 통계를 어떻게 개선해야 하는지 방향을 제시한다(통계청, 2021, pp.4-5). 이 연구에서는 5가지 차원 중에서 정확성 측면을 살펴보는데, 정확성은 오차가 작을수록 정확성이 높은 통계라고 할 수 있다. 표본조사 자료에서 발생할 수 있는 오차를 살펴보고 그중에서 측정오차에 대해 심도 있게 검

1) 5가지 차원은 관련성, 정확성, 시의성/정시성, 비교성/일관성, 접근성/명확성으로 정의하고 있다.

증해 보고자 한다. 표본조사 자료에 측정오차가 있는 경우에는 보정하는 방안도 다양하게 모색하여 제안할 것이다. 또한, 주요 해외 표본조사 자료에서의 측정오차 관리 방안 사례를 조사하여 고품질 자료 생산을 위한 조사 자료 관리 방안을 마련하고자 한다.

이러한 연구 결과는 정책 수립 시 정책입안자나 연구자가 편의가 없고 신뢰성 있는 통계적 분석 결과를 활용할 수 있도록 조사 자료의 품질을 향상하는 데 기여할 것이다.

제2절 연구 내용 및 방법

1. 연구 내용

이 연구는 조사 자료의 품질 검증 연구로 표본조사 자료에서 발생하는 측정오차를 중심으로 살펴본다. 각 장에서의 연구 내용은 다음과 같다.

제2장은 표본조사에서의 총오차에 대한 구성요소 소개 및 개념 정리, 각 오차에 따른 축소 방안을 살펴본다. 주요 해외 조사 자료인 미국의 Panel Study of Income Dynamics, Survey of Income and Program Participation, 유럽의 Survey of Health, Ageing and Retirement in Europe, 영국의 British Household Panel Survey의 표본 및 조사 방식, 측정오차 관리, 조사 품질 향상 방안에 대해 알아보고 이를 통해 고품질 자료 생산을 위한 조사 자료 관리 방안에 대해 모색한다.

제3장은 한국복지패널조사 자료의 경상소득 변수에 대하여 히핑 현상이 있는지 살펴본 후, 히핑이 존재한다면 히핑 현상을 보정할 수 있는 기존의 방법을 검토하고 개선하는 새로운 방법론을 제안하여 실제 한국복

지패널조사 자료에 적용하는 것이다.

제4장은 가계금융복지조사 자료에서 가구주의 개인 근로소득, 가구의 근로·자녀장려금, 가구원의 기초연금 변수에 대하여 응답값과 행정보완값 간의 관계를 파악하여 측정오차가 존재하는지 살펴본다. 측정오차가 존재하는 경우 측정오차를 보정할 수 있는 여러 가지 방법을 시도해 본다. 그 방법으로는 측정오차 보정을 통한 회귀계수 추정, 히핑 보정, 비례 핫덱대체(Fractional Hot Deck Imputation)방법을 사용한 측정오차 보정이 이에 해당한다. 이 방법에 대한 이론적 개념과 적용 방법을 검토하고, 실제 자료에 활용하여 결과를 살펴본다.

결론으로 연구 결과를 요약한 다음, 측정오차 축소를 위한 관리 방안에 대해 제안하고자 한다.

2. 연구 방법

국내외 문헌 연구, 실제 조사 자료 분석, 전문가 자문회의 등을 다양하게 활용하여 연구한다. 각 장에서 사용한 조사 자료를 보면 제3장 자기기입 소득에 대한 히핑 보정은 2006~2020년(1~15차 연도) 한국복지패널조사 자료이고, 제4장 금액 변수에 대한 측정오차 보정은 2017년 가계금융복지조사 자료이다. 통계패키지는 SAS 및 R을 사용한다.

제2장

오차의 개념 및 해외 조사 자료 사례 연구

제 2 장 오차의 개념 및 해외 조사 자료 사례 연구

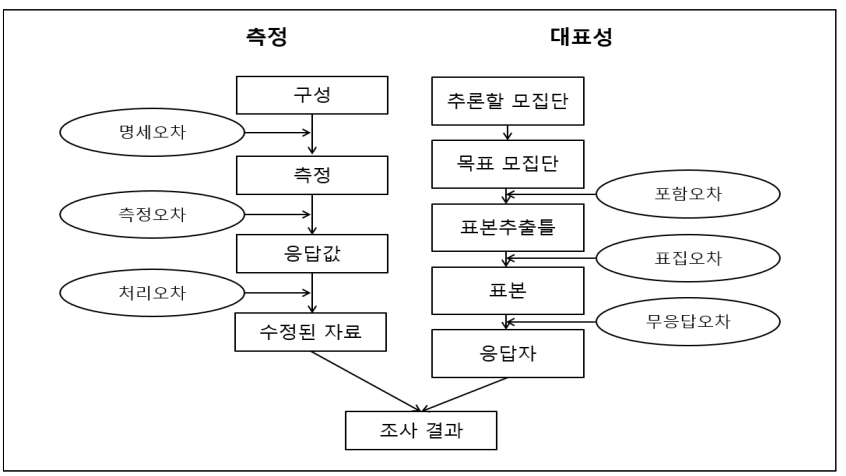
조사는 일부 대상만을 조사하여 모집단 전체에 대하여 추론하기 때문에, 추정값(sample estimate)과 알려지지 않은 참값(true population value) 간의 차이가 불가피하다. 이러한 차이는 총조사오차(total survey error, TSE)로, 표본추정량과 모수 간의 제곱의 기댓값인 평균제곱오차(mean squared error, MSE)로 정의되며, 크게 표집오차(sampling error)와 비표집오차(non-sampling error)로 구분할 수 있다(박인호 등, 2017).

표집오차는 조사 대상 일부만을 선택함에 따라 발생하는 확률적 오차(random error)로, 발생 및 규모 제어가 가능하다. 그러나 비표집오차는 조사의 기획 및 수행의 전 과정에서 발생할 수 있다. 조사 주기별로 선택되는 절차 또는 방법은 비표집오차의 잠재적 발생 원인이 될 수 있는데, 비확률적·비의도적 오류 및 시스템 과실에 해당하는 이러한 비표집오차는 측정 및 제어가 용이하지 않은 것으로 알려져 있다. 패널조사 과정에서 발생하는 총조사오차는 해당 통계의 정확성과 신뢰성에 중대한 영향을 미칠 수 있다. 특히, 비표집오차는 표집오차보다 총조사오차에서 차지하는 비중이 큰 것으로 알려져 있으며(김정섭, 임경은, 2010), 이에 따라 체계적인 오차 관리를 통해 비표집오차를 축소하기 위한 노력이 필요하다.

Groves and Lyberg(2010)는 총조사오차를 조사의 설계, 자료 수집, 처리 및 분석 등 전반의 과정에서 발생할 수 있는 모든 오차의 누적으로 정의한다. 〔그림 2-1〕에 제시된 바와 같이 총조사오차는 크게 측정 관련

오류와 대표성 관련 오류로 구분되는데, 조사 주기별로 발생 가능한 각각의 오차 요소들을 구분하여 식별하면 전반적인 조사 품질 향상이 가능할 것으로 기대된다. 즉, 전체적인 비표집오차 축소의 효율성 개선을 위해서는 비표집오차를 오차의 발생 시점 및 요인에 따라 측정오차, 무응답오차, 처리오차 등으로 세부 구분하여 각 오차의 특성을 고려할 필요가 있다.

[그림 2-1] 조사 주기별 총조사오차의 구조



자료: Groves, R. M., Lyberg, L. (2010). p.856.

이 장에서는 총조사오차에서 표집오차와 비표집오차의 정의 및 구분, 축소 방안을 알아보고, 비표집오차 중 하나인 측정오차의 원인과 축소 방안에 대해 구체적으로 살펴보았다. 그리고 주요 해외 조사 자료에서 측정오차를 줄이기 위한 관리 방안에 관한 사례를 미국의 Panel Study of Income Dynamics, Survey of Income and Program Participation, 유럽의 Survey of Health, Ageing and Retirement in Europe, 영국의 British Household Panel Survey를 중심으로 조사하였다.

제1절 표집오차

1. 표집오차의 정의 및 구분

조사의 목적은 모집단(population)에서 추출한 표본(sample)을 바탕으로 모집단의 특성을 추론하는 것이다. 표집오차(sampling error)란 모집단 전체가 아닌 일부 표본에 한해서만 자료를 수집하는 조사의 특성상 발생 가능한 오류이며, 표본으로부터 추정된 값과 조사의 관심 대상인 모집단의 실제 값과의 차이를 나타낸다(Statistical Policy Office, 2001). 이승희(2010)에 따르면, 각 표본에서 추출한 통계치, 반복된 표본에서 추출한 통계치의 평균값, 그리고 전체 모집단에서 얻어지는 결과물이 일치하지 않는 것은 표집오차가 발생하기 때문이다. 따라서 표집오차는 〔그림 2-1〕에 제시된 바와 같이 표집이 모집단을 어느 정도까지 대표하는가(representation)와 크게 연관된다.

2. 표집오차의 축소 방안

표집오차의 크기를 결정하는 주된 요소는 표본의 수이다. 일반적으로 표본의 크기가 클수록 표집오차가 감소한다. 또한 이 장에서 논의될 주요 해외 조사 사례와 같이 표본에 대하여 군집 표집(cluster sampling)을 활용하는 경우에는, 군집 간 동질성이 클수록 표집오차의 크기가 줄어드는 것으로 알려져 있다. 그러나 표본의 크기를 조절하는 것 이외에도, 표본추출방법에 따라 표집오차의 크기가 상이할 수 있다. 이승희(2010)는 표본추출방법으로 크게 단순임의추출(simple random sampling), 층화추출(stratified sampling), 계통추출(systematic sampling) 등을 제

시하고 있는데, 실제 조사에서는 이들을 서로 결합하거나 더 세부화된 방법을 사용하고 있다. 주요 해외 패널조사의 경우, 표본 크기 조정과 더불어 적절한 표본 설계 방안을 선택함으로써 모수 추정의 정확도를 향상하고 표집오차의 축소를 모색하고 있다(제4절 참조).

제2절 비표집오차

1. 비표집오차의 정의 및 구분

총조사오차에서 표집오차가 아닌 다른 오차들은 모두 비표집오차로 정의된다. 〈표 2-1〉은 비표집오차의 종류를 정리한 것이다. 비표집오차는 조사의 설계, 진행 및 자료 취합 등 조사 전반의 과정에서 발생하며, 표본의 대표성과 관련된 포함오차(coverage error) 및 무응답오차(non-response error)와 자료의 측정과 연관된 명세오차(specification error), 측정오차(measurement error), 처리오차(processing error) 등으로 세분할 수 있다.

〈표 2-1〉 비표집오차 종류

구분	비표집오차
표본의 대표성	포함오차 무응답오차
자료의 측정	명세오차 측정오차 처리오차

자료: 필자 작성

포함오차는 표본이 목표모집단(target population)을 너무 적게 또는 불필요하게 많이 포함(과소포함, 과대포함)하거나, 복수의 목표모집단 단위가 표본에 존재하는 경우 발생한다. 무응답오차는 조사 대상이 응답하지 않아 자료의 부재로 발생하는 오차를 말한다. 명세오차는 조사의 목적이 조사항목에 반영된 개념과 상이할 때 발생하며, 주로 연구진과 설문지 개발자 간의 의사소통 미흡으로 발생하는 것으로 알려져 있다(박인호 등, 2017). 측정오차는 자료를 수집하는 과정에서 발생하는 오차로, 응답값과 알려지지 않은 실제값과의 차이로 발생한다(Kasprzyk, 2005). 처리오차는 자료의 최종 처리 과정에서 발생하는 오차로, 수집된 데이터의 입력 및 편집 과정에서 발생할 수 있다.

2. 비표집오차의 축소 방안

표집오차와 달리 비표집오차는 조사의 기획 및 수행의 전 과정에서 발생 가능하며, 측정과 제어가 용이하지 않다. 가중치 부여(weighting), 사후층화(post stratification), 자료 대체(imputation) 등 비표집오차에 사후적으로 대응하는 방안이 제시되고 있으나, 이는 비표집오차의 발생 방지 및 최소화를 위한 근본적인 해결책이 될 수 없는 것으로 사료된다. 비표집오차 축소와 관련된 정확하고 보편적인 해결 방안은 지금까지 제시된 바 없으나, Organisation for Economic Cooperation and Development(2003)는 비표집오차를 조사 주기별로 세분화하고, 각 비표집오차에 대한 대응 방안을 간략히 제시하고 있다. 예를 들어, 데이터 입력 및 편집 과정에서의 처리오차를 축소하기 위하여 컴퓨터 지원 대면조사(computer assisted personal interview, CAPI) 또는 컴퓨터 지원 전화조사(computer assisted telephone interview, CATI)의 도입

을 권고한다. 아울러 조사 도구의 타당성 검토, 설문지 설계 확인, 조사원에 대한 적절한 교육·훈련 프로그램 제공 등은 조사의 전반적인 품질 향상에 필수적인 요소로 판단되며, 이와 관련한 측정오차의 축소 방안은 다음 절에서 자세히 다루고자 한다.

제3절 측정오차

1. 측정오차의 원인

보통 가구를 대상으로 한 조사 자료는 다양한 방법을 통해 수집된다. 이러한 자료 수집의 전제는, 측정 대상의 특성과 개념에 대한 정의가 명확하며 일련의 체계적인 자료 수집 절차가 동반된다는 것이다. 그러나 전술된 가정에도 불구하고 실제 조사를 통해 수집되는 변수들의 경우, 설문의 설계, 측정 방법 및 입력의 오류, 또는 반복되는 조사에 따른 응답자의 응답 부담(response burden) 등에 기인한 오차가 발생할 수 있다. 이러한 오차는 비표집오차의 하나인 측정오차(measurement error)로, 자료를 수집하는 과정에서 주로 발생하며 특정 변수에 대해 응답자가 제공하는(또는 실제 기록되는) 값과 알려지지 않은 실제 값 간의 차이를 의미한다(Kasprzyk, 2005). 측정오차는 통계의 편향(bias)과 변동(variance)을 초래할 수 있으며, 때로는 분석의 결과를 왜곡시킬 수 있다.

측정오차는 응답자, 조사원, 그리고 설문 내용 때문에 주로 발생한다. 예컨대 응답자는 의도와 무관하게 잘못된 정보를 제공할 수 있으며, 조사원 또한 설문 과정에서 잘못된 응답을 유도하거나 올바른 응답을 잘못 입력할 수 있다. 모호한 설문 내용 또한 측정오차의 원인이 될 수 있다. 해

외의 경우 측정오차의 다양한 원인과 대처 방안에 관한 연구가 오래전부터 진행되어 왔다. Biemer and Lyberg(2003)는 측정오차의 원인을 정보 시스템(information system), 조사의 설정(setting), 자료 수집 방법(mode of data collection), 응답자(respondent), 면접(interview), 그리고 조사 도구(instrument) 등 5가지로 구분하여 제시하고 있다. 이와 유사하게 Kasprzyk(2005)는 측정오차의 원인을 아래와 같이 4개 요인으로 세분화하여 구분하고 있다.

1) 설문 내용(questionnaire)
2) 자료 수집 방법(data collection mode)
3) 조사원(interviewer)
4) 응답자(respondent)

설문 내용에 의한 측정오차는 설문의 설계, 시각적 레이아웃, 설문의 주제 및 문구의 오류로 인해 발생할 수 있다. Biemer, Groves, Lyberg, Mathiowetz, and Sudman(1991)에 따르면 설문의 표현이 모호할 경우 응답자가 설문의 의도를 상이하게 해석할 수 있으며, 각 설문 또는 전체 설문지의 길이 또한 측정오차 발생 원인이 될 수 있다. 예컨대 Kasprzyk(2005)는 세부적이고 긴 설문은 간단하고 짧은 설문보다 정확한 응답을 유도할 수 있으나, 이와 반대로 전체 설문지의 길이가 길수록 응답자의 응답 부담이 증가할 수 있다고 주장한다. 자료의 수집 방법 또한 그 차이에 따라 측정오차를 초래할 수 있다. 대개의 조사가 채택하고 있는 대면조사(face-to-face interview), 전화조사(telephone interview), 서면조사(mail interview) 방식은 조사원의 유무, 컴퓨터를 이용한 설문조사(computer-assisted paper interview, CAPI) 또는 컴퓨

터를 이용한 전화조사(Computer-Assisted Telephone Interview: CATI) 방식의 소프트웨어 사용 유무 등에 따라 크기가 다른 측정오차를 발생시킬 수 있다. 조사원은 조사나 자료의 입력과정에서 응답에 영향을 미칠 수 있으며, 응답자 또한 자기 의도와 관계없이 잘못된 응답을 할 수 있다. 특히 과거에 대한 설문의 경우, 시간의 경과에 따라 응답자는 실제보다 과대 또는 과소 추정하여 응답하는 것으로 나타났다(Kasprzyk, 2005).

2. 측정오차 축소 방안

조사에서 측정오차의 존재는 기존 문헌에서 광범위하게 연구되어 밝혀진 바 있다. 전술된 바와 같이 측정오차는 응답자, 조사원, 설문을 포함한 조사의 다양한 요소로 인해 발생할 수 있다. 응답자는 의도적이든 의도적이지 않든 잘못된 정보를 제공할 수 있으며, 조사원의 말투와 외모까지 응답에 영향을 미치는 요소로 작용할 수 있다. 응답 오입력, 데이터 위조, 조사 절차 미준수 또한 대표적인 측정오차 발생 요인으로 지목된다. 모호한 질문과 오해하기 쉬운 용어, 지침 등은 설문 내용의 잘못된 설계로 인해 발생할 수 있는 측정오차의 예시이다.

측정오차는 사후적으로 보정이 가능한데, 보정 방법은 외부데이터의 사용 여부에 따라 사용된 자료 외에 별도의 표본에서 측정오차의 분포를 얻어 내는 방법인 외부 보정(external calibration)과 현 자료 내에서 일부를 무작위로 추출하여 검증 표본(validation sample)의 측정오차를 얻어 내는 방법인 내부 보정(internal calibration)으로 구분할 수 있다(Guo, Little, & McConnell, 2012). 그러나 이는 사후적인 조치에 불과하며, 조사 과정에서 발생하는 측정오차의 근본적인 축소 방안으로 볼 수

없다. 측정오차를 축소하는 방안은 앞서 언급된 측정오차 발생의 세부 요인별로 고려하는 것이 용이하다(1. 측정오차의 원인 참조). 예컨대 조사원과 응답자로부터 발생하는 측정오차 축소를 위해 Kasprzyk(2005)는 응답자에 대한 응답 부담(response burden) 경감, 필요시 재조사(reinterview) 진행, 체크리스트를 통한 이중 체크(record check), 조사원에 대한 감독, 교육, 훈련 프로그램 마련 등을 제안하고 있다. 또한, 설문의 설계 과정에서 발생할 수 있는 측정오차는 단일 조사표(harmonized questionnaire)를 마련, 활용함으로써 일부 방지할 수 있다(Organisation for Economic Cooperation and Development, 2003).

제4절 해외 조사 자료의 사례 연구

패널조사(panel survey)는 특정한 관심집단의 역동성에 대한 지속적인 파악을 목적으로 다양한 분야에서 도입·운용되고 있다. 이와 같은 종단면(longitudinal) 패널조사는 여러 측면에서 유용하다고 할 수 있는데, 시간의 경과에 따라 표본이 축적되어 추정의 효율성(efficiency)과 정확도(precision)가 향상되고, 동일한 응답자에 대한 자료를 수집하여 시간의 흐름에 따른 표본의 변화 파악, 특정 집단의 역동성 연구가 가능하다는 것이다. 그러나 이러한 장점에도 불구하고 패널조사 통계는 몇 가지 주의사항을 동반한다. 예컨대 응답자의 측면에서는 주기적으로 반복되는 조사에 대한 응답 부담(response burden)이 발생할 수 있으며, 응답에 대한 피로감 누적으로 응답자가 무성의한 답변을 할 수도 있다. 또한, 조사의 설계, 진행 및 입력 단계에서 발생 가능한 다양한 원인에 따른 탈락

으로 오차가 발생할 수 있다는 점이 패널조사의 단점이라고 볼 수 있다.

이 절에서는 주요 해외 패널조사의 자료 처리 방안에 관한 사례를 연구하였다.

1. Panel Study of Income Dynamics

가. 개요

미국의 Panel Study of Income Dynamics(이하 PSID)는 현재까지 지속되고 있는 가장 오래된 종단면 가구 조사(longitudinal household survey)로, 린든 존슨 행정부가 시행한 빈곤과의 전쟁(War on Poverty) 정책의 정량적 평가와 소득과 빈곤의 역동성 파악이라는 목표 달성을 위해 도입되었다. 1966년 미국 통계청(U.S. Census Bureau)은 인구 전반을 대표하는 22,000가구의 표본과 비백인(non-white) 거주 조사구역의 15,000가구를 표본 가구로 설정하고, 이 중 약 30,000가구에 대한 경제기회조사(Survey of Economic Opportunity: SEO)를 계획하고 시행하였다. 이후 미시간대학교 조사연구센터(Survey Research Center at the University of Michigan: SRC)가 비빈곤가구(non-poor household)와 원표본 가구에서 분리된 가구원에 대한 추적을 포함하여 조사하고 있다 (McGonagle, Schoeni, Sastry, & Freedman, 2012). 1968년 PSID 최초 조사는 경제기회조사(SEO)의 저소득층 표본 1,872가구 및 미시간대학교 조사연구센터(SRC)가 생산가능 인구에 대하여 별도로 추출한 비빈곤층 2,930가구를 원표본으로 설정하고, 해당 가구에 속한 18,223명의 개인을 대상으로 진행하고 있다(Institute for Social Research, 2021).

나. 표본 및 조사 방식

PSID의 조사 대상 가구 또는 조사대상자 수는 〈표 2-2〉와 같으며, 차수에 따라 대체로 증가 추세를 보이고 있다. 그러나 실제 PSID의 표본 수는 〈표 2-2〉와 상이하며 역동적이다. 이는 원표본 가구에서 출생하거나 원표본 가구로 입양된 신규 가구원이 자동적으로 표본가구원에 포함되기 때문이며, 원표본 가구의 가구원이 분가하여 형성한 경제적으로 독립적인 가구(split-off) 또한 개별적인 조사 대상에 추가되기 때문이다. 조사의 대표성 확장을 위한 노력의 일환으로 미시간대학교 조사연구센터는 1990년부터 1995년까지 한시적으로 멕시코, 푸에르토리코, 쿠바 출신 라틴계 가구를 조사에 포함한 바 있으며, 1997년 조사 이후부터는 기존의 경제기회조사(SEO) 표본, 미시간대학교 조사연구센터(SRC) 표본 외에도 이민자 표본을 포함한 조사를 진행하고 있다.

〈표 2-2〉 표본 유형별 조사 대상 가구 및 조사대상자 수

(단위: 가구, 명)

조사 년도	가구					개인				
	SRC	SEO	라틴계	이민자	총계	SRC	SEO	라틴계	이민자	총계
1968	2,930	1,872	-	-	4,802	9,461	8,772	-	-	18,223
1969	2,643	1,817	-	-	4,460	8,643	8,569	-	-	17,212
1970	2,754	1,891	-	-	4,645	8,751	8,597	-	-	17,348
1980	3,589	2,944	-	-	6,533	10,034	9,713	-	-	19,747
1990	3,935	3,393	2,043	-	9,371	10,677	10,068	7,452	-	28,197
1995	4,565	4,002	1,834	-	10,401	12,314	11,615	5,955	-	29,844
1997	4,592	1,714	-	441	6,747	12,363	5,703	-	1,695	19,761
2001	4,970	1,945	-	491	7,406	13,340	6,232	-	1,828	21,400
2011	5,495	2,767	-	645	8,907	14,607	7,844	-	2,210	24,661
2019	5,255	3,172	-	1,142	9,569	13,925	8,401	-	3,758	26,084

주: 1990년 조사 – 1995년 조사까지 라틴계 표본 포함, 1997년 조사부터 격년 조사 시행 및 이민자 표본 포함

자료: Institute for Social Research. (2021). PSID Main Interview User Manual. University of Michigan.

조사는 1997년까지 매년 진행되었으나, 그 이후는 2년마다 진행되고 있다. 현재까지 활용된 조사 방식은 크게 대면조사(face-to-face interview), 전화조사(telephone interview), 그리고 컴퓨터를 이용한 전화조사(Computer-Assisted Telephone Interview: CATI) 등 3가지로 구분된다. 1968년 최초 조사부터 1972년까지 95% 이상이 대면조사로 진행되었으며, 이후 전화조사 방식이 점진적으로 활용되었다. 1993년부터는 일부 조사에 CATI 방식이 활용되었으며, 조사의 품질 향상을 위해 2003년부터는 Blaise, SurveyTrak 등 소프트웨어 도입을 통해 설문 내용 프로그램화 및 응답 대상자에 대한 정보관리를 진행하고 있다. PSID의 경우, 1973년 전화조사의 활용을 본격화한 이후 최근까지도 CATI 방식을 활용하고 있는 것으로 판단된다. 1980년부터 1990년까지 약 90%의 조사가 전화조사로 진행되었으며, 1991년 이후는 95%를 상회하였다. 2019년 조사에서는 전체 조사의 95.6%가 CATI 방식의 전화조사로, 이외의 조사는 컴퓨터를 이용한 개별 면접(Computer-Assisted Personal Interview: CAPI) 형식으로 진행되었다.

다. 조사 품질 향상 방안

조사 내용은 고용, 소득 등에서 조사대상자의 건강, 지출, 시간 사용 등까지 점진적으로 확대되었으며, 이에 따라 평균 설문시간이 대폭 증가하였다. 평균 설문시간은 1973년 조사의 경우 20.1분으로 최저를 기록하였는데, 2019년 조사에서는 4배에 해당하는 80.3분을 기록하였다. 이는 주 조사 영역에 대한 평균 설문시간만을 산출한 수치이며, 추가 설문에 소요되는 12.7분을 고려하면 조사대상자에 대한 총 설문 부담(response burden)은 평균 93분으로 나타난다. 응답자는 고용 영역에 대한 응답에

가장 많은 시간(18.5분)을 할애하는 것으로 조사되었으며, 지출(16.1분), 건강 상태 등(14.7분)이 뒤를 이었다. 설문 내용의 지속적인 증가로 인한 조사대상자의 부담 감소를 위해 PSID는 이전 조사에서 응답한 결과를 바탕으로 공통된 내용에 대해 설문을 대체하는 방식인 종속형 설문(Dependent Interviewing: DI)을 활용하고 있다. 사용자 매뉴얼에 따르면 2019년 조사에서는 2017년보다 설문 내용을 일부 축소하였으며, 연금 영역의 설문에 대해서는 DI를 활용하여 평균 설문시간을 5.5분 단축하였다(Institute for Social Research, 2021).

라. 측정오차 처리 사례

PSID 내 측정오차의 존재는 오래전부터 연구되어 왔다. Duncan and Hill(1985)은 418명의 응답자를 대상으로 해당 노동자의 소득에 대한 응답과 기업의 실제 급여 자료를 비교하여 응답 내용과 실제 값의 차이 유무를 분석하였다. 해당 연구는 노동조합에 소속된 노동자의 경우 소득의 측정오차의 편차가 30%(1982년의 경우 15%) 더 크게 나타나는 것으로 밝혔다.

이처럼 PSID는 자료 검증 연구의 일환으로 PSID Validation Survey(이하 PSID VS)를 개별적으로 수행하였다. PSID를 진행하면서 주요 노동시장 변수들에 대해 응답자들로부터 정보를 획득하고, 향후 별도의 PSID VS를 실시하여 해당 응답자들이 소속되어 있는 직장으로부터 같은 변수에 대해 정보를 수집, 획득한 정보의 일치 여부를 검증하고 있다(Pischke, 1995).

PSID의 경우 사용자 지침(User Guide)에서 측정오차를 별도로 언급하고 있지 않다. 그러나 CATI로의 전환, Event History Calendar(EHC)

도입, 사례비 제공, 조사원 교육 등을 통해 조사의 품질향상을 도모하고, 측정오차를 축소하기 위한 노력을 지속하고 있는 것으로 나타났다. 특히, 응답자가 과거의 사건이나 상황을 정확히 인지하지 못하여 잘못된 정보를 제공하는 회고오차(recall error) 발생 방지를 위해 EHC를 도입하여 과거의 응답 정보를 전자적으로 관리하고 있다. 또한, 응답자의 응답 부담 감소 및 원활한 참여 유도를 위해 응답자에게 별도의 인센티브를 제공하고 있다. 1968년 최초 조사에서는 각 응답자에게 5달러가 지급되었으나 계속 증가하여 2015년도 조사에서는 70달러가 일괄적으로 지급되었고, 2017년도 조사부터는 최대 150달러(이민자 표본의 경우 최대 300달러)가 지급되었다. 아울러 표본이탈(sample attrition)을 방지하기 위하여 뉴스레터, 주소 업데이트 우편 등을 발송하고 있다. 그리고 조사원에 대한 사전교육 내용을 디지털로 전환하여 2007년부터는 정식 교육 이전에 온라인으로 수료하도록 의무화하고 있다. 조사원은 별도로 마련된 체크리스트에 따라 조사를 수행하여야 하며, 추가적인 전문성을 보유한 감독관(supervisor)이 각 조사원의 선발, 교육, 모의 실사 수행, 실제 조사 수행에 관여하고 있다. 또한, PSID는 데이터 수집을 용이하게 하고 비용을 절감하기 위해 표본 지역(primary sampling area)을 200개 이상의 군집(cluster)으로 구분하고, 가능한 응답자들에게 동일한 조사원을 고정적으로 배치하는 방안을 마련하였다(Institute for Social Research, 2021).

2. Survey of Income and Program Participation

가. 개요

미국 Survey of Income and Program Participation(이하 SIPP)은 미국의 개인 및 가구별 연간·반기별 소득 현황 파악과 정부의 소득 분배 정책에 관한 포괄적인 정보 제공 및 정책 효과 분석을 목표로 1983년부터 미국 통계청(US Census Bureau)에 의해 시행되고 있다. 〈표 2-3〉은 SIPP에서 조사하고 있는 항목을 보여주고 있다.

〈표 2-3〉 SIPP 조사항목 구분

구분	세부 조사항목(예시)
인구학적 특성 (Demographic Characteristics)	- 나이, 성별, 인종 등 - 교육 현황 - 사용 언어 - 결혼 및 자녀 유무 - 거주지역
고용 (Employment)	- 노동 시간 - 임금 - 직업 특성 - 노동조합 가입 여부
자산 및 부채 (Assets and Liabilities)	- 기타 자산 보유 현황 - 은퇴 계획 - 부채 - 공과금
건강 및 복지 (Health and Well-being)	- 자녀에 대한 지원 - 장애 유무 - 부양가족 유무 - 의료비 지출 - 건강보험
복지 프로그램 대상 여부 및 소득 분배 (Program Participation and Income Transfers)	- 퇴직 후 소득 - 소셜 시큐리티(social security) - 기타 지원 프로그램 등록 및 수급 여부

자료: US Department of Commerce, Economic and Statistics Administration, US Census Bureau. (2021). 2018 Survey of Income and Program Participation Users' Guide.

매년 또는 격년으로 시행되는 PSID, BHPS 등 타 패널조사와는 다르게 SIPP는 32개월에 걸쳐 4개월마다 동일한 조사대상에 대한 소득 현황 등을 조사하였다. 1983년 10월에 수행된 1차 조사에서 2만 가구 이상의 표본을 선정하여 15세 이상의 구성원에 대한 조사를 진행한 바 있다. 그러나 예산 부족을 포함한 기타 문제로 인해 1988년과 1989년에는 조사가 진행되지 않는 등, 조사의 설계 및 운영에 대한 문제점이 계속 제기되었다.

1996년 개편은 종단면 조사 품질 향상의 일환으로 진행되었다. 개편을 통해 표본의 크기가 대폭 증가하였고, 조사 기간 또한 개편 전 32개월에서 48개월로 연장되었으며, CAPI 방식의 조사를 도입하였다. 개편 이후 수행된 1996년도 조사의 표본 규모는 40,188가구로, 원표본 가구의 구성원과 신규 편입 대상자는 48개월의 패널조사 주기 동안 4개월에 한 번씩 조사에 응하게 되었다. 2000년 예산 부족으로 인한 조사 중단 이후, SIPP는 2014년 재개편(re-engineered SIPP)을 진행하여 비용 절감과 응답 부담 감소를 도모하였다. 이러한 노력의 하나로 Event History Calendar(EHC)를 도입하여 응답자로부터의 오류 발생을 방지하고 있다. 또한 비용 절감을 위해 조사주기를 기존 4개월에서 1년으로 연장하는 대신, 1996년 개편 이후 중단되었던 중복(overlapping) 패널 구조를 2018년 조사부터 재도입하여 표본 크기 확대와 표준오차 축소를 목표로 하고 있다(US Department of Commerce, Economic and Statistics Administration, & US Census Bureau, 2021).

나. 표본 및 조사 방식

조사는 예산 부족, 개편 등의 사유로 주기적으로 진행되지 않았으나, 2018년 조사는 44,870가구의 64,191명의 구성원을 대상으로 현재 진행

중이다(〈표 2-4〉 참조). 기존에는 동일한 조사대상에 대하여 8~12회 조사를 진행하였으나, 2014년 개편 이후 연 1회 조사를 수행하고 있다. 조사주기마다 새로운 표본 가구를 설정하여 조사를 진행하는 것이 타 패널 조사와의 차이점이라고 볼 수 있다.

〈표 2-4〉 SIPP 차수별 응답 가구 수

(단위: 가구)

조사 시작 연도	1차 조사	마지막 조사	1차 조사 대상 가구 수	총 조사 차수
1984	83년 10월	86년 7월	20,897	9
1985	85년 2월	87년 8월	14,306	8
1986	86년 2월	88년 4월	12,425	7
1987	87년 2월	89년 5월	12,527	7
1988	88년 2월	90년 1월	12,725	6
1989	89년 2월	90년 1월	12,867	3
1990	90년 2월	92년 9월	19,800	8
1991	91년 2월	93년 9월	15,626	8
1992	92년 2월	95년 5월	21,577	10
1993	93년 2월	96년 1월	21,823	9
1996	96년 4월	00년 3월	40,188	12
2001	01년 2월	04년 1월	50,500	9
2004	04년 2월	08년 1월	51,379	12
2008	08년 9월	12년 12월	52,031	13
2014	14년 2월	17년 6월	42,348	4
2018	18년 2월	진행 중	44,870	진행 중

주: 2018년도 조사는 현재 진행 중

자료: US Department of Commerce, Economic and Statistics Administration, US Census Bureau. (2021). 2018 Survey of Income and Program Participation Users' Guide.

또한, 기존의 PAPI 방식에서 탈피하여 1996년 개편 이후 CAPI 방식의 조사가 점진적으로 도입되었다. 2014년 개편에 따라 2차 조사까지는 CAPI 방식을 활용하고 있으며, 이후 조사부터 CATI 방식의 조사를 수행하고 있다(National Research Council, 2009).

다. 조사 품질 향상 방안

SIPP는 조사의 전반적인 품질 높이기 위해 다양한 노력을 기울여 왔으며, 이 중 하나는 소득에 따른 스크린 설문(income screener questions)의 도입이라고 할 수 있다. Income screener questions의 도입으로 응답자가 특정 소득 기준 이상의 소득이 있다고 판단되는 경우, 미리 지정된 설문이 자동적으로 빠져서 응답자의 응답 부담을 줄일 수 있는 것으로 판단된다(US Department of Commerce et al., 2021). 더불어 응답 부담 축소를 위한 노력의 일환으로 SIPP는 특정 응답자를 'clump'별로 분류하고 있다. 해당 분류를 통해 'clump' 대표자의 응답이 'clump' 구성원 전체에 자동으로 적용되어 타 구성원들이 마주하는 설문 수와 응답 부담 증가를 방지하고 있다.

PSID와 마찬가지로 SIPP 또한 이전 조사에서 응답한 결과를 바탕으로 공통된 내용에 대해 설문을 대체하는 방식인 종속형 설문(Dependent Interviewing: DI)을 2004년도 조사부터 활용하고 있다. 2018년도 조사의 경우, SIPP는 2차 조사 이후부터 종속형 설문을 활용하였으며, 약 500개 항목이 성공적으로 대체되어 응답 부담 감소를 유도하고, 돌출 편향(seam bias)의 부정적 효과를 완화한 것으로 평가하고 있다. 특히, 종속형 설문을 최초 도입한 2004년도 조사의 경우, 조사의 모든 영역에서 돌출 편향 감소 효과가 나타난 것으로 평가된 바 있다(Moore, 2008).

라. 측정오차 처리 사례

SIPP의 측정오차 존재 여부는 연구 별로 상이하게 나타나고 있다. Kalton, Kasprzyk, and McMillen(1989)은 SIPP의 나이, 인종, 성별,

산업 및 직업 분야의 설문에서 측정오차가 존재함을 밝힌 바 있다. 그러나 Marquis and Moore(1990)는 SIPP 내 측정오차가 매우 드물게 존재하며 표집오차의 편차는 2% 미만이라고 주장한다. 해당 연구는 SIPP 내 측정오차의 원인 및 해결 방안은 알려진 바 없음을 나타내고 있으며, 2018년도 조사에 대한 사용자 지침(User Guide) 또한 응답자가 설문을 오인하는 경우 측정오차가 발생할 수 있음을 간략히 언급하고 있다(US Department of Commerce et al., 2021).

일반적으로 정부의 복지 프로그램에 참여하여 수혜를 입는 응답자는 더 적극적으로 조사에 참여하는 것으로 알려져 있으며, 정부의 복지 정책에 대한 정보를 조사 및 수집하는 SIPP의 특성상 표집오차는 타 패널조사보다 비교적 작게 나타나는 것으로 판단된다. PSID와 마찬가지로 임금 및 소득에 대한 SIPP의 자료 검증 연구(validation study) 또한 비정기적으로 실시되고 있다. 특히, 사회보장제도(social security)의 행정 자료와의 비교연구가 지속 수행되고 있는데, Abowd and Stinson(2011)의 조사는 SIPP 임금 통계 자료와 Details Earnings Records(DER)가 86% 일치함을 나타내어 SIPP 내 측정오차가 타 패널조사 대비 상대적으로 우수함을 보여주었다.

3. Survey of Health, Ageing and Retirement in Europe

가. 개요

유럽의 Survey of Health, Ageing and Retirement in Europe(이하 SHARE)은 노년층 인구의 건강 상태 및 사회적 위치 파악을 위해 2004년부터 50세 이상 인구를 대상으로 진행되고 있는 가장 포괄적인

국제 패널조사로서, 28개 유럽 국가들과 이스라엘이 참여하고 있다. European Research Infrastructure Consortium(ERIC) 주관으로 2021년 현재까지 8차 조사가 진행되었으며, 약 48만 회의 조사에 약 14만 명의 응답자가 참여하였다.

나. 표본 및 조사 방식

1차 조사는 오스트리아, 벨기에, 스위스, 독일, 덴마크, 스페인, 프랑스, 그리스, 이탈리아, 네덜란드, 스웨덴 등 11개 국가에서 시행되었으며, 이후 계속 증가하여 현재는 이스라엘을 포함한 29개국에서 진행되고 있다. 조사 주기는 국가별 상황에 따라 일부 상이할 수 있으나, 대개 격년 주기로 진행되고 있다. 1차 및 2차 조사의 주제는 건강, 사회·경제적 요소에 관한 내용이 주를 이루었으며, 3차 조사부터 SHARE LIFE로의 주제 전환을 통해 약 3만 명에 달하는 응답자의 생활 전반에 대한 조사를 진행하였다. 이후 최근 8차 조사에서는 코로나19 팬데믹 발생 및 지속에 따른 신체적 및 정신적 건강 상태, 사회적 및 경제적 변화를 중심으로 조사가 진행된 바 있다.

조사 방식은 PAPI, CATI, CAPI가 혼재되어 진행되고 있으나, 최근 8차 조사에서는 코로나19의 영향으로 CATI 방식이 가장 선호되었다 (European Research Infrastructure Consortium, 2021).

다. 조사 품질 향상 방안

SHARE는 조사 회차마다 지정된 국가를 대상으로 조사원에 대한 조사(interviewer survey)를 병행하여 진행하고 있다. 이를 통해 조사에 대

한 조사원의 태도를 사후적으로 파악하여, 조사의 품질 향상을 위해 대응 방안을 마련하고 있으며, 향후 개선 방향에 대한 조사원들의 의견 수렴을 진행하고 있다.

라. 측정오차 처리 사례

SHARE에 대한 측정오차 관련 연구는 비교적 찾아보기 어려우며, 8차 조사의 Release Guide에서도 측정오차 관련 내용은 언급되어 있지 않다. Bingley and Martinello(2014)는 SHARE 덴마크의 교육, 고용 시장 현황, 소득 자료와 덴마크의 행정 통계(Public Administrative Register)와의 자료 검증 연구(validation study)를 통해 전술된 분야에서의 측정오차는 미미하거나 유의미하지 않음을 증명하고 있다. 비표집오차의 경우, 비표집오차 및 표본 이탈(attrition)에 대한 대응 방안으로 European Research Infrastructure Consortium(2021)은 가중치(calibrated weights) 부여를 제시하고 있다.

4. British Household Panel Survey

가. 개요

영국 British Household Panel Survey(이하 BHPS)는 영국 인구 전반의 변화, 변화의 원인 분석 및 예측을 핵심 목표로 1991년에 시작되었으며, 에식스대학교 사회경제연구소(Institute for Social and Economic Research at the University of Essex: ISER)와 영국 종단연구센터(UK Longitudinal Studies Centre)가 공동으로 주관하고 있

다. BHPS의 1차 연도 표본은 영국 우편번호 주소 파일(Postcode Address File)에서 추출하여 조사가 시도된 8,167가구 중 5,505가구로 구성되었다. 1차 연도 표본 가구에 소속된 16세 이상의 구성원 10,264명은 원표본구성원(Original Sample Members: OSM)으로 지칭되었으며, 동일한 원표본구성원을 지속 추적하되 분리된 가정을 형성하는 경우 새로운 세대의 모든 성인도 조사 대상에 포함하였다. 일시적 표본구성원(Temporary Sample Members: TSM)은 원표본구성원과 가구를 형성한 조사대상자를 지칭하며, 원표본구성원과 함께 거주하는 경우에만 조사대상에 포함되었다. 또한 지속 표본구성원(Permanent Sample Members: PSM)은 원표본구성원과 충분하고 지속적인 유대를 가진 것으로 판단되는 일시적 표본구성원을 지칭하며, 이 경우 지속 표본구성원으로 표본구성을 전환하여 조사를 수행하였다(Institute for Social and Economic Research, 2006).

나. 표본 및 조사 방식

〈표 2-5〉에 나타난 바와 같이 10,264명의 응답자를 대상으로 한 최초 조사 수행 이후 응답자 수는 전반적으로 감소 추세를 보이고 있다. 이는 BHPS가 원표본구성원을 지속 추적하는 종단면 조사의 방식을 유지하고 있기 때문이며, 원표본구성원의 탈락 발생 또는 무응답 사례가 원표본구성원의 자녀 출산, 신규 가구원 편입 등의 추가 표본 발생 사례보다 빈번하기 때문으로 판단된다. 그러나 통계의 대표성 제고를 위해 BHPS는 잉글랜드 이외 지역에서의 조사 표본을 점진적으로 확대하고 있다. 8차 조사까지 스코틀랜드와 웨일스는 전체 조사에서 각각 500가구 정도로 비교적 작은 비중을 차지하였으나, 각 지역의 독립적 분석 수행을 위해 9차 조

사부터 두 지역의 표본 가구 수를 2,000가구 이상으로 확대하였다. 또한, 2001년 11차 조사부터는 북아일랜드 지역 2,900가구를 추가하였고, 매년 약 1만 가구를 대상으로 영국 전역에서 조사가 수행되고 있다.

〈표 2-5〉 BHPS 조사 차수에 따른 응답자 수

(단위: 명)

구분	원표본구성원 (OSM)	지속 표본구성원 (PSM)	일시적 표본구성원 (TSM)	총계
1차	10,264	-	-	10,264
2차	9,351	10	484	9,845
3차	8,921	29	650	9,600
4차	8,609	77	795	9,481
5차	8,305	120	824	9,249
6차	8,315	179	944	9,438
7차	8,155	240	1,071	9,466
8차	7,992	291	1,032	9,315
9차	7,821	359	1,043	9,223
10차	7,600	334	1,018	8,952
11차	7,448	325	1,110	8,883
12차	7,299	310	1,161	8,770
13차	7,120	299	1,236	8,655

주: 2003년 수행된 13차 조사까지의 결과

자료: Institute for Social and Economic Research. (2006). Quality Profile: British Household Panel Survey Version 2.0. University of Essex.

조사는 1991년부터 현재까지 해마다 진행되고 있다. 1차 조사부터 8차 조사까지 모든 조사가 종이 설문지를 이용한 대면조사(Paper-Assisted Personal Interview: PAPI)로 진행되었으나, 1999년인 9차 연도 조사부터는 Initiative2(잉글랜드, 스코틀랜드, 웨일스 지역), Blaise(북아일랜드 지역) 소프트웨어를 활용한 컴퓨터 지원 대면조사(Computer-Assisted Personal Interview: CAPI) 방식으로 전환하여 조사 소요 시간 단축과 조사원으로부터 발생할 수 있는 측정오차 감소

를 도모하고 있다. 또한, 제13차 조사까지는 해마다 4월과 5월 중에 1,000명의 표본을 대상으로 예비 조사(pilot survey)를 실시하여 질문 내용 사전 검토를 진행하고 있으며, 본 조사는 9월부터 12월까지 4개월에 걸쳐 진행하고 있다(Taylor, Brice, Buck, & Prentice-Lane, 2018).

다. 조사 품질 향상 방안

전술된 바와 같이 BHPS는 9차 연도 조사부터 CAPI 방식으로 전환하여, 시간 및 비용 절감과 조사원으로부터 발생할 수 있는 측정오차 감소를 도모하고 있다. 또한, 응답의 누락을 방지하기 위해 모든 설문에 반드시 응답해야만 다음 설문을 진행할 수 있도록 설계하였다. CAPI 방식으로의 전환 이후에도 일정 기간 PAPI 방식조사를 중복으로 시행함으로써 전환 기간 중 조사의 정확성 제고를 도모하였다(Institute for Social and Economic Research, 2006).

BHPS는 응답자의 조사 참여를 독려하기 위해 모든 응답자에게 5파운드 상당의 인센티브 바우처(gift voucher)를 조사 이후 제공하였다. 6차 조사부터는 바우처 제공 금액을 7파운드로 상향 조정하였으며, 조사 수행 이전에 사례비를 제공하여 응답자의 적극적인 참여를 유도하고 있다. 표본 지역(sampling area)은 250개 이상의 지역으로 구분되는데, 〈표 2-6〉과 같이 각 표본 지역마다 1명의 조사원을 배치하여 조사의 전문성 제고를 도모하고 있다.

〈표 2-6〉 BHPS 차수별 조사원 수

(단위: 명, %)

구분	조사원 수	1차조사와 동일한 조사원 비율
1차	243	100
2차	237	59
3차	216	57
4차	217	50
5차	217	50
6차	212	44
7차	218	39
8차	228	39
9차	212	35
10차	212	32
11차	209	32
12차	259	29
13차	261	27

주: 2003년 수행된 13차 조사까지의 결과

자료: Institute for Social and Economic Research. (2006). Quality Profile: British Household Panel Survey Version 2.0. University of Essex.

조사원의 경험 및 능력과 관련하여, 모든 조사원은 조사 시작 전 조사의 목적을 설명하는 간단한 우편물을 모든 표본 주소에 발송하여야 하며, 이후 일주일 이내 응답자에게는 사전 연락을 취해야 한다. 이후 조사에 대한 보다 상세한 정보를 포함한 2차 우편이 발송되며, 조사대상자로부터의 질의에 대비하여 모든 조사 관련 내용을 명확히 숙지하도록 의무화하고 있다. 특히, 조사원들은 현장 조사기간 동안 감독관(supervisor)과 동행하여 2주간의 감독 및 평가를 받게 된다(Institute for Social and Economic Research, 2006).

라. 측정오차 처리 사례

BHPS의 경우, 사용자 지침(User Manual)은 측정오차에 대하여 언급하고 있지 않으나, Quality Profile은 BHPS 내 측정오차에 관한 연구 결과를 분야별로 상세히 보고하고 있다(Institute for Social and Economic Research, 2006). 고용 관련 통계에 대해서는 응답자가 과거의 사건이나 상황에 대해 정확히 인지하지 못하여 잘못된 정보를 제공하는 회고오차(recall error)가 측정오차 발생의 주요 원인으로 지목되고 있다. Institute for Social and Economic Research(2006)는 시간의 흐름에 따른 응답자의 상태 변화에 대한 측정오차 발생 또한 지적하고 있다. 직업에 대한 개별적 응답의 경우, 32.2%의 응답자가 1차 조사와 2차 조사에서 다른 직업 코드를 입력한 것으로 나타났으며, 소속 산업의 경우 28.9%의 응답자가 다른 산업 코드를 입력한 것으로 보고된 바 있다. 이에 대한 원인으로는 설문의 내용 변화, 컴퓨터를 이용한 자료 입력 여부, 조사원 변경, 조사원의 입력 오류 등이 지적되었다.

Institute for Social and Economic Research(2006)는 이전 조사에서 응답한 결과를 바탕으로 공통된 내용에 대해 설문을 대체하는 방식인 종속형 설문의 도입 및 활용을 측정오차 축소 방안 중 하나로 제시하고 있다. 실제로 BHPS는 16차 조사부터 고용 상태, 이전 직장, 소득 원천 등에 대한 설문에 종속형 설문을 적용하고 있으며, 이를 통해 시간의 흐름에 따른 응답자의 잘못된 정보 제공과 응답자의 응답 부담 증가 방지를 기대하고 있다.

제5절 소결

이 절에서는 해외의 주요 패널조사인 미국의 PSID, 미국의 SIPP, 유럽의 SHARE, 영국의 BHPS에 대한 조사 품질 향상 방안 및 측정오차 처리 방법 사례를 조사하였다.

비표집오차는 조사의 설계, 설문 진행, 데이터 처리 등 패널조사의 전 단계에서 발생할 수 있으며, 이러한 비표집오차의 발생은 패널조사의 정확성과 신뢰성에 영향을 미칠 수 있다. 비표집오차 중에서 측정오차는 조사 진행 과정에서 응답자나 조사원의 영향, 설문 내용의 모호성, 자료 입력 과정에서의 오류 등에 의해 발생하는 것으로 알려져 있다. 측정오차로 인한 편향을 줄이기 위해서는 사후적인 보정에 앞서 근본적으로 발생 가능성을 줄이는 것도 중요하다. 조사를 수행하는 조사원에 대한 적절한 관리, 조사의 설계 및 진행 과정에서 응답자의 적극적인 참여, 응답의 신뢰도를 높일 수 있는 방안에 대한 고려가 필요하다고 생각한다.

〈표 2-7〉에 정리된 바와 같이 PSID 등 해외 주요 패널조사에서는 조사의 설계 및 진행, 조사참여자의 응답, 데이터 처리 과정에서 발생할 수 있는 측정오차 방지를 위한 다양한 기술적, 시스템적 접근이 이루어지고 있음을 확인하였다. 가장 대표적인 기술적 접근으로는 컴퓨터를 이용한 전화조사(CATI) 또는 컴퓨터를 이용한 대면조사(CAPI) 방식의 도입을 꼽을 수 있다. 미국 PSID와 유럽 SHARE는 전화조사로 수집된 정보들을 조사원이 컴퓨터에 내장된 프로그램에 입력하는 CATI 방식의 도입을 통해 조사의 효율성과 입력된 데이터의 정확성 향상을 모색했다. SIPP와 BHPS에서도 역시 대면조사에 컴퓨터를 활용(CAPI)하여 설문 진행 시간과 데이터 처리 시간 단축을 도모하고 있다. 특히 BHPS는 CAPI 방식으로 전환하면서 PAPI 조사를 함께 시행하여, 오류 발생에 따른 조사 자료

의 정확성 저하를 방지하였다.

〈표 2-7〉 해외 주요 패널조사의 조사 품질 향상 및 측정오차 처리 방안

패널조사 (나라)	조사 품질 향상 및 측정오차 처리 방안
PSID (미국)	- CATI를 활용하여 조사 - 종속형 설문(DI) 도입 및 활용 - Event History Calendar(EHC) 활용
	- 사례비 지급 - 조사원 교육 - PSID-VS(Validation Survey) 실시
SIPP (미국)	- CAPI를 활용하여 조사 - 종속형 설문(DI) 도입 및 활용 - Event History Calendar(EHC) 도입 및 활용 - 스크린 설문 도입 및 활용
	- 비정기적인 VS(Validation Survey) 실시
SHARE (유럽)	- CATI를 활용하여 조사
	- 조사원 대상 설문조사(interviewer survey)
BHPS (영국)	- CAPI를 활용하여 조사 - 종속형 설문(DI) 도입 및 활용
	- 사례비 지급 - 조사원 감독 및 평가

자료: 필자 작성

PSID, SIPP, BHPS는 이전 조사에서 응답한 결과를 바탕으로 공통된 내용에 대한 설문을 대체하는 방식인 종속형 설문(Dependent Interviewing: DI)을 도입하여 활용하고 있다. 이는 설문의 불필요한 반복을 줄임으로써 전체적인 응답 시간을 줄이고, 응답자의 응답 부담(response burden)을 최소화하는 데 기여하고 있다. 특히 SIPP는 응답에 따라 미리 지정된 설문을 자동으로 통과하는 스크린 설문(income screener question) 도입을 통해 추가적인 응답 부담을 줄이고 있다. 한편, PSID와 SIPP에서 언급되고 있는 Event History Calendar(EHC)의

도입은 측정오차 중에서 응답자의 회고오차(recall error) 발생을 억제하는 데 도움을 줄 수 있다. 응답자는 과거의 사건이나 상황에 대해 정확히 인지하지 못하여 의도와 상관없이 왜곡된 응답을 할 수 있기 때문이다. PSID와 SIPP는 응답자의 회고를 요구하는 특정 설문에 관한 질문 시 관련된 정보를 함께 제공하여, 회고오차의 발생 가능성을 낮추기 위해 노력하고 있다.

컴퓨터를 활용한 기술적인 접근 외에도 응답자 또는 조사원이 발생시키는 오류를 방지하기 위한 다양한 조사관리 방안을 마련하고 있다. PSID와 BHPS는 응답자의 응답 부담 감소 및 적극적인 참여 유도를 위해 별도의 인센티브로 사례비를 제공하고 있다. 아울러 조사 시작 전에 별도의 우편물을 발송하여 조사 응답률 및 응답자의 설문 이해도를 높이기 위해 노력하고 있다. 조사원 오류의 발생을 최소화하기 위해 조사원에 대한 체계적인 교육 프로그램 또한 마련되어 있는 것으로 판단된다. PSID의 경우, 디지털 교육 프로그램을 구축하여 조사원의 이수를 요구하고 있으며, BHPS는 조사원에 대한 감독관의 감독 및 평가를 의무화하고 있다. SHARE는 정기적으로 조사원 대상 설문조사(interviewer survey)를 실시하여 조사 과정에서 조사원이 체감하는 애로사항을 파악하여 시스템 개선에 반영하고 있다.

전술된 바와 같이 패널조사에서는 측정오차의 발생 방지, 측정 및 제어가 용이하지 않은 편이다. 측정오차의 축소와 관련된 일원화된 해결 방안은 아직 제시된 바 없으나, 해외 주요 패널조사의 접근에서 시사점을 찾을 수 있다. 패널조사의 설계, 진행 및 데이터 처리 등의 각 단계에서 발생할 수 있는 측정오차의 다양한 원인을 파악해야 할 것이며, 세부 요인별 접근을 통한 다각적인 고려가 필요하다고 생각한다.

제3장

자기기입 소득에 대한 히핑 보정 방안

제3장 자기기입 소득에 대한 히핑 보정 방안

제1절 개요

히핑(heaping)이란 조사 자료에 있어 응답자가 자기기입 등 스스로 응답한 흡연 양, 음주 횟수 등 범주형 변수의 빈도가 특정 응답값의 배수에 집중되거나, 임금, 소득, 소비 등 연속형 변수에 대하여 관측된 분포가 특정 응답값의 배수에 지나치게 크게 집중되는 현상을 뜻한다.

히핑에 대한 측정오차는 히핑 현상이 발생한 특정 값의 배수에 응답한 값 모두 측정오차가 존재한다고 보기보다는 정확하게 응답한 경우와 측정오차가 포함된 응답 경우로, 2가지 응답 형태가 섞여 있는 일종의 혼합분포 형태로 해석할 수 있다. 이러한 혼합분포 형태의 측정오차는 추후 관심 변수를 추정할 때 편향을 발생시키고 분산을 증가시킨다. 특히 소득이 관심 변수인 경우, 특정 응답값에 집중된 분포는 소득 불평등 지수 등과 같은 양극화 지수를 계산할 때 왜곡된 값이 도출되는 원인이 될 수 있다.

히핑은 측정오차가 있는 경우 응답값을 활용한 모수 추정 문제와 유사하다고 볼 수 있다. 즉, 히핑을 보정하는 방법이란 불완전하게 관측된 응답값을 사용하여 히핑이 발생한 부분의 응답값을 대체하는 방법을 뜻한다. 히핑이 발생하는 응답을 모두 대체할지 일부만 대체할지는 히핑에 대한 혼합분포 가정을 통하여 조절할 수 있다.

히핑을 보정하게 되면 일반적으로 특정 응답값의 배수에 지나치게 집중된 분포가 주변으로 퍼지는 효과가 생겨 응답값의 분포가 부드러워진다.

이 연구의 목적은 한국복지패널조사 자료의 경상소득 변수에 대하여

히핑 현상이 있는지 살펴보는 것이다. 히핑이 존재한다면 그 현상을 보정할 수 있는 기존의 방법을 검토하고, 기존의 방법을 개선하는 새로운 방법론을 제안하여 실제 한국복지패널조사 자료에 적용하고자 한다.

제2절 소득 자료에 대한 히핑 분석

1. 소득에 대한 히핑 여부 분석

한국복지패널조사의 가구 자료 중에서 자기기입으로 작성된 경상소득에 대한 히핑 현상을 분석한다. 한국복지패널조사 자료에서 경상소득은 근로소득, 사업 및 부업소득, 재산소득, 사적 이전소득, 공적 이전소득의 합으로 계산된다. 한국복지패널조사 자료 1차부터 15차까지 가구 경상소득(이하 '경상소득')에 대하여 히핑 현상이 있는지 살펴본다. 7차에 추가된 신규패널은 따로 구분하여, 원패널과 신규 패널을 나누어 살펴보고자 한다.

히핑 현상은 조사된 소득 값의 분포가 특정 값의 배수에 특히 많이 치우쳐진 것을 의미한다. 먼저, 차수별 특정 응답값에 치우쳐진 분포가 있는지 소득 응답값별 비중을 분석하였다. 〈표 3-1〉을 보면 1차, 2차, 14차, 15차 경상소득의 응답값 비중 상위 10개에 대한 경상소득 값이 나타나 있다. 〈표 3-2〉는 7차에 새롭게 진입한 신규패널 표본에 대하여 7차, 8차, 14차, 15차 경상소득에 대한 응답값 비중 상위 10개에 대한 경상소득 값이 나타나 있다.

원표본 패널의 경우, 1차 연도와 2차 연도 응답 비중 상위 10개의 응답값이 5 또는 10의 배수 혹은 120의 배수임을 알 수 있다. 이러한 10의 배

수의 형태로 응답 비중이 높은 현상은 14차, 15차 연도에서는 찾아볼 수가 없다. 신규패널 표본의 경우, 7차 연도의 응답 비중 상위 10개의 응답값은 5의 배수 형태로 나타났으나 8차 연도 이후로는 특정 배수의 형태를 가지지 않았다.

〈표 3-1〉 원표본 1차, 2차, 14차, 15차 경상소득 응답 비중 상위 10개

(단위: 만 원)

1차 연도		2차 연도		14차 연도		15차 연도	
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%

자료: 한국보건사회연구원 (2006~2020). 1~15차 한국복지패널조사의 원자료를 분석함.

〈표 3-2〉 신규표본 7차, 8차, 14차, 15차 경상소득 응답 비중 상위 10개

(단위: 만 원)

7차 연도		8차 연도		14차 연도		15차 연도	
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%

자료: 한국보건사회연구원 (2006~2020). 1~15차 한국복지패널조사의 원자료를 분석함.

원표본의 경우 1차부터 15차까지 경상소득이 5 또는 120의 배수 형태로 나타나는 비율을 분석하였고, 신규표본의 경우 7차부터 15차까지 5의 배수 형태로 경상소득이 나타나는 비율을 분석하였다. 그 결과가 〈표 3-3〉과 〔그림 3-1〕에 나타나 있다.

원표본의 경우, 1차 연도에서 5의 배수인 경상소득의 비율은 약 38%이고, 2차 연도에서는 29%였다. 그 이후 차수에서는 약 20%로 나타나, 초반 차수에서 상당히 높은 비율을 가졌다. 120의 배수 비율도 1차 연도에서는 약 10%로 그 이후 차수보다 상당히 높게 나타나고 있다. 신규표본에서 5의 배수인 경상소득의 비율은 7차 연도에 약 27%로 그 이후 차수(약 20%)보다 다소 높게 나타남을 알 수 있다.

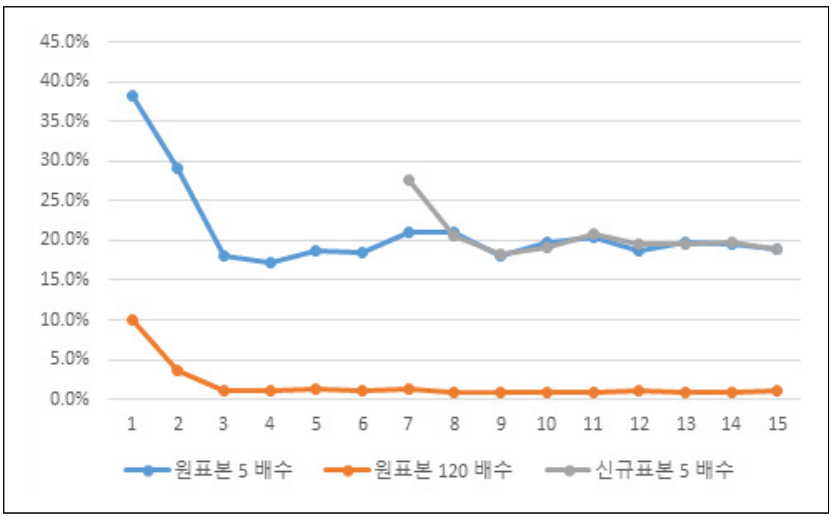
이로부터 1차 연도와 7차 연도에는 경상소득이 5 또는 120의 배수 형태로 응답한 경우가 다른 차수에 비해서 높음을 알 수 있다.

〈표 3-3〉 원표본과 신규표본에서 5 또는 120의 배수인 경상소득 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%

자료: 한국보건사회연구원 (2006~2020). 1~15차 한국복지패널조사의 원자료를 분석함.

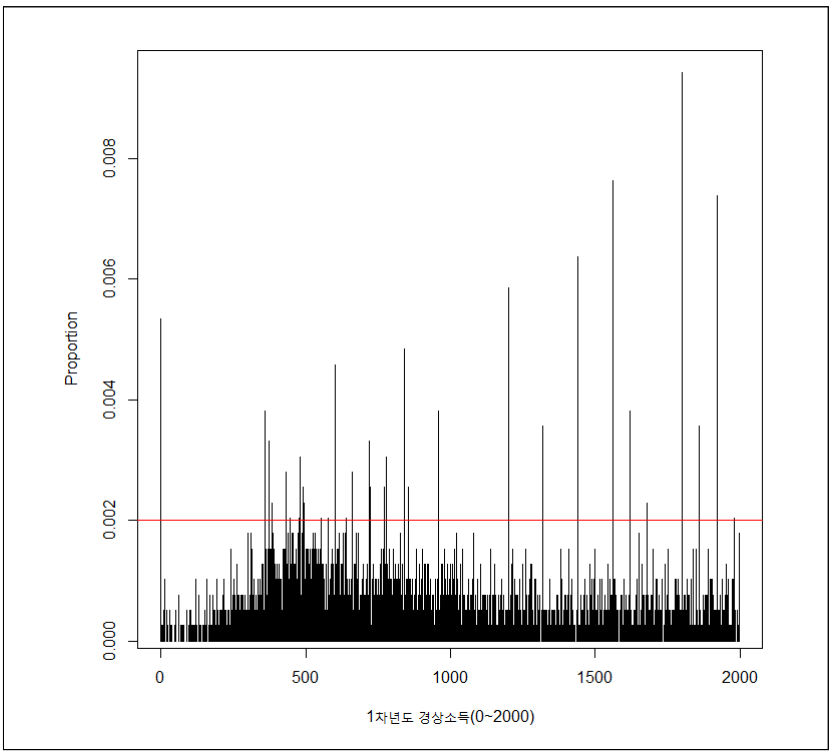
〔그림 3-1〕 원표본과 신규표본에서 5 또는 120의 배수인 경상소득 비율



자료: 한국보건사회연구원 (2006~2020). 1~15차 한국복지패널조사의 원자료를 분석함.

1차 연도의 경상소득값이 5 또는 120의 배수 형태로 응답한 경우가 많다는 앞선 결과를 자세히 살펴보기 위하여 spike plot을 그려보았다. 해석을 용이하게 하기 위하여 1차 연도 경상소득 중 0 보다 크고 2,000보다 작은 응답값에 대하여 spike plot을 작성하였다. 〔그림 3-2〕를 보면 응답 비율이 약 0.2%를 넘어가면 특정 응답값에 분포가 집중되어 비율값이 튀는 것을 확인할 수 있다.

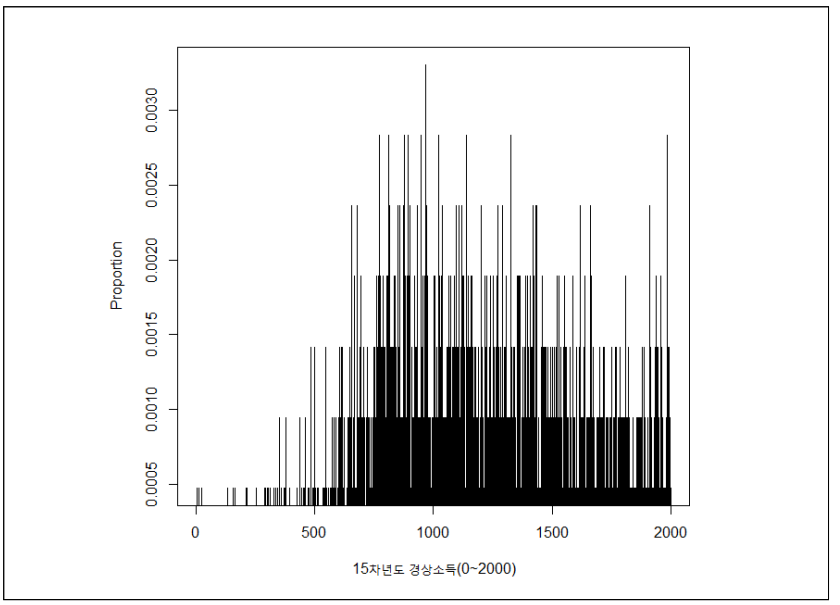
〔그림 3-2〕 원패널 1차 연도 경상소득(0~2,000)에 대한 spike plot



자료: 한국보건사회연구원 (2006). 1차 한국복지패널조사의 원자료를 분석함.

〔그림 3-3〕을 보면 15차 연도 경상소득에 대한 spike plot이 나타나 있는데 앞선 〈표 3-1〉에서 확인할 수 있듯이 특정 응답값에서 분포가 튀는 현상을 찾을 수는 없다.

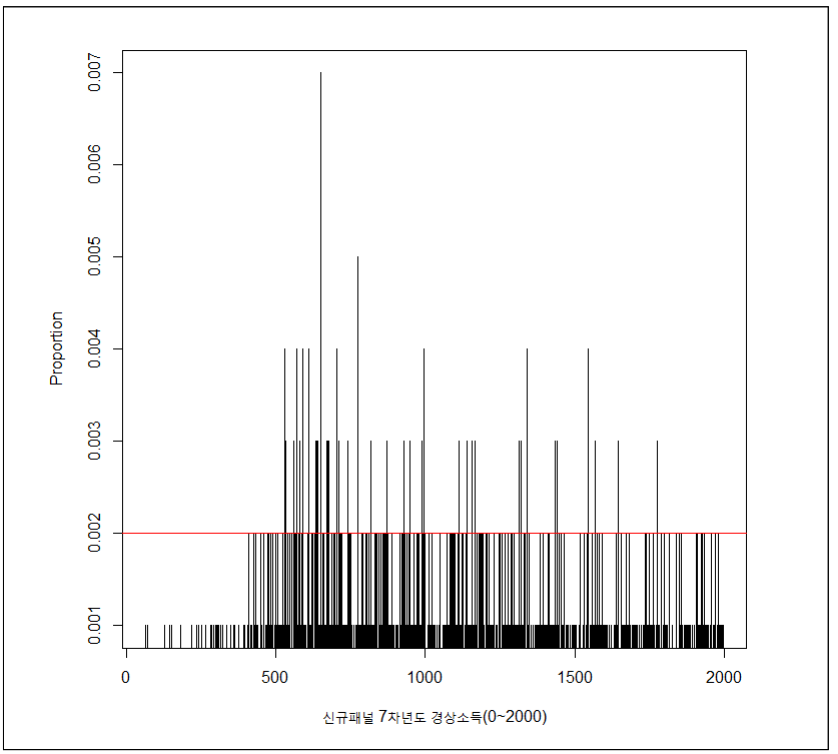
[그림 3-3] 원패널 15차 연도 경상소득(0~2,000)에 대한 spike plot



자료: 한국보건사회연구원 (2020). 15차 한국복지패널조사의 원자료를 분석함.

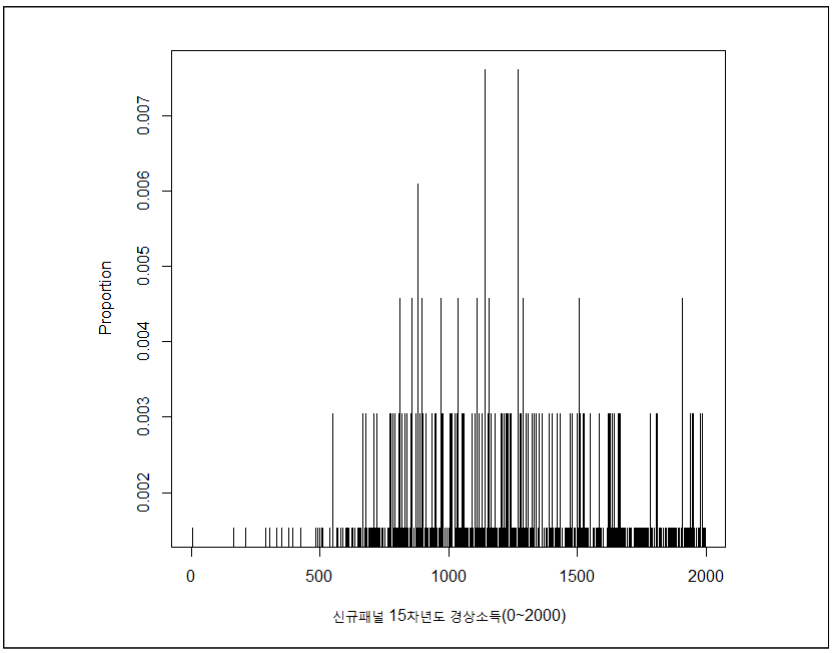
신규패널에 대해서도 7차 연도와 15차 연도의 경상소득에 대해서 spike plot을 작성하였다. 각각 [그림 3-4]와 [그림 3-5]에 그 결과가 나타나 있다. [그림 3-4]를 보면 응답 비율이 약 0.2%를 넘어가면 특정 응답값에 분포가 집중되어 비율 값이 튀는 것을 확인할 수 있다. [그림 3-5]에서는 앞선 〈표 3-2〉에서 확인할 수 있듯이 특정 응답값에서 분포가 튀는 현상을 찾을 수는 없다.

〔그림 3-4〕 신규패널 7차 연도 경상소득(0~2,000)에 대한 spike plot



자료: 한국보건사회연구원 (2012). 7차 한국복지패널조사의 원자료를 분석함.

〔그림 3-5〕 신규패널 15차 연도 경상소득(0~2,000)에 대한 spike plot



자료: 한국보건사회연구원 (2020). 15차 한국복지패널조사의 원자료를 분석함.

원표본과 신규표본의 1차 연도에서 응답 비율이 약 0.2%를 넘어가는 응답값에 대하여 각각 〈표 3-4〉 및 〈표 3-5〉에 경상소득 크기순으로 표시하였다. 〈표 3-4〉를 보면 원표본 1차 연도 경상소득 중에서 응답 비율이 특히 높은 응답값은 260, 600, 840 등으로 모두 120의 배수인 것을 확인할 수 있다. 〈표 3-5〉를 보면 신규표본에 대한 것으로, 경상소득의 끝자리가 9, 0, 5로 끝나는 응답값에서 비율이 높은 것으로 나타났다.

〈표 3-4〉 원패널 1차 연도 경상소득 중 응답 비율 높은 경상소득

(단위: 만 원)

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%

자료: 한국보건사회연구원 (2006). 1차 한국복지패널조사의 원자료를 분석함.

〈표 3-5〉 신규패널 7차 연도 경상소득 중 응답 비율 높은 경상소득

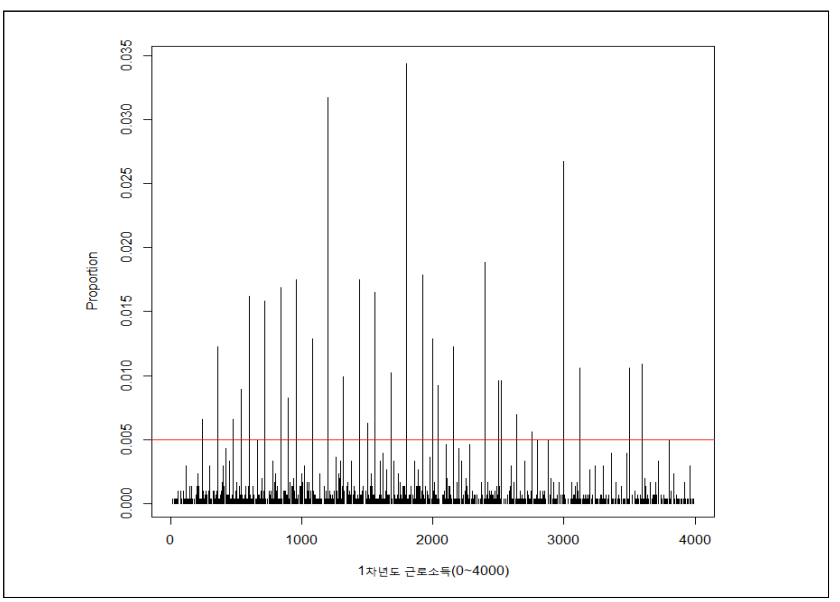
(단위: 만 원)

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%

자료: 한국보건사회연구원 (2012). 7차 한국복지패널조사의 원자료를 분석함.

앞에서 보았듯이 원표본의 경상소득은 1차 연도에서 120의 배수인 응답값에 응답이 집중되었다. 세부적으로 경상소득 계산 시 비중을 가장 많이 차지하는 근로소득에 대하여 히핑 존재 여부를 살펴 그 특성을 파악하고, 0 보다 큰 근로소득에 대하여 spike plot을 그려보았다. 근로소득의 3사분위수가 3,572이고 최대값이 30,000이므로 0에서 4,000 사이의 1차 연도 근로소득에 대하여 spike plot을 작성하였다. [그림 3-6]을 보면 각 값이 차지하는 비율이 약 0.5%보다 크면 분포의 쏠림 현상(spike)이 발생하는 것을 확인할 수 있다. 분포 비율이 0.5%보다 큰 근로소득의 값과 비율을 〈표 3-6〉에 정리하였다. 앞선 경상소득에서와 마찬가지로 근로소득은 240을 시작으로 120의 배수가 될 때마다 히핑 현상이 발생한다는 것을 알 수 있다.

[그림 3-6] 원패널 1차 연도 근로소득(0~4,000)에 대한 spike plot



자료: 한국보건사회연구원 (2006). 1차 한국복지패널조사의 원자료를 분석함.

〈표 3-6〉 원패널 1차 연도 근로소득 중 응답 비율 높은 근로소득

(단위: 만 원)

근로소득	비율	근로소득	비율
240	0.5%	2,040	0.7%
360	1.0%	2,160	1.0%
480	0.5%	2,400	1.5%
540	0.7%	2,500	0.8%
600	1.3%	2,520	0.8%
720	1.3%	2,640	0.6%
840	1.3%	3,000	2.1%
900	0.7%	3,120	0.8%
960	1.4%	3,500	0.8%
1,080	1.0%	3,600	0.9%
1,200	2.5%	4,000	1.1%
1,320	0.8%	4,200	0.6%
1,440	1.4%	4,500	0.6%
1,560	1.3%	4,800	0.6%
1,680	0.8%	5,000	1.1%
1,800	2.7%	6,000	1.1%
1,920	1.4%	7,000	0.8%
2,000	1.0%		

자료: 한국보건사회연구원 (2006). 1차 한국복지패널조사의 원자료를 분석함.

2. 히핑 여부에 대한 가구 특성 분석

앞의 히핑 분석 결과를 바탕으로 1차 연도 원표본의 경상소득에 대한 히핑을 120의 배수인 경우로 지정하였다. 히핑 존재 여부와 여러 설명변수 간의 관계를 살펴보았다. 설명변수는 지역구분, 균등화 소득에 따른 가구 구분, 가구원 수, 가구주 성별, 가구주 교육 수준이다. 〈표 3-7〉부터 〈표 3-12〉에 각각의 결과가 나타나 있다.

〈표 3-7〉 원표본: 지역별 히핑 존재 여부 빈도수와 비율

(단위: 가구)

	빈도수		비율	
	없음	있음	없음	있음
서울	1,171	164	87.7%	12.3%
인천/경기	1,386	183	88.3%	11.7%
부산/울산/경남	1,105	114	90.6%	9.4%
대구/경북	823	92	89.9%	10.1%
대전/충남	478	51	90.4%	9.6%
강원/충북	425	31	93.2%	6.8%
광주/전남북/제주	971	78	92.6%	7.4%
전체	6,359	713	89.9%	10.1%

자료: 한국보건사회연구원 (2006). 1차 한국복지패널조사의 원자료를 분석함.

〈표 3-8〉 원표본: 소득에 따른 가구 구분별 히핑 존재 여부 빈도수와 비율

(단위: 가구)

	빈도수		비율	
	없음	있음	없음	있음
일반 가구	3,279	515	86.4%	13.6%
저소득 가구	3,080	198	94.0%	6.0%

자료: 한국보건사회연구원 (2006). 1차 한국복지패널조사의 원자료를 분석함.

〈표 3-9〉 원표본: 히핑 존재 여부에 대한 경상소득의 차이

(단위: 만 원)

t-test	히핑 여부	
	없음	있음
경상소득 평균	2276.1	2755.1
t-statistic	-5.609	

자료: 한국보건사회연구원 (2006). 1차 한국복지패널조사의 원자료를 분석함.

〈표 3-10〉 원표본: 가구원 수별 히핑 존재 여부 빈도수와 비율

(단위: 가구)

	빈도수		비율	
	없음	있음	없음	있음
1인	1,370	151	90.1%	9.9%
2인	1,960	135	93.6%	6.4%
3인	1,193	164	87.9%	12.1%
4인	1,350	205	86.8%	13.2%
5인 이상	486	58	89.3%	10.7%

자료: 한국보건사회연구원 (2006). 1차 한국복지패널조사의 원자료를 분석함.

〈표 3-11〉 원표본: 가구주 성별에 따른 히핑 존재 여부 빈도수와 비율

(단위: 명)

	빈도수		비율	
	없음	있음	없음	있음
남자	4,798	553	89.7%	10.3%
여자	1,561	160	90.7%	9.3%

자료: 한국보건사회연구원 (2006). 1차 한국복지패널조사의 원자료를 분석함.

〈표 3-12〉 원표본: 가구주 교육 수준별 히핑 존재 여부 빈도수와 비율

(단위: 명)

	빈도수		비율	
	없음	있음	없음	있음
중졸 이하	3,052	174	94.6%	5.4%
고졸 이하	1,848	320	85.2%	14.8%
대학교 이상	1,459	219	86.9%	13.1%

자료: 한국보건사회연구원 (2006). 1차 한국복지패널조사의 원자료를 분석함.

서울, 인천, 경기 지역에서 히핑 발생 비율이 강원, 충북에 비하여 높게 나타났다. 일반 가구에서 히핑 발생 비율이 저소득 가구보다 높게 나타났다. 히핑 존재 여부를 기준으로 경상소득의 차이를 살펴본 결과, 히핑이 있는 경우 소득이 약 2,755만 원으로 히핑이 없는 경우에 비하여 통계적으로 유의한 차이가 있음을 알 수 있다. 1인 가구에 비하여 3인, 4인 가구

의 히핑 발생 비율이 높으나, 5인 이상의 가구에서는 1인 가구와 유사한 비율로 히핑이 발생한다. 가구주 성별에 따라서는 히핑 발생 비율의 차이가 크게 나타나지 않는다. 가구주 교육 수준은 중졸 이하와 그 이상의 교육과의 사이에 히핑 발생 비율 차이가 크게 나타난다.

〈표 3-13〉은 가구원 특성 분석을 통해 유의하게 나타난 설명변수를 사용한 포화모형에서 유의하지 않은 설명변수를 하나씩 제거하여 적합한 로지스틱 회귀분석의 최종 모형에 관한 결과를 나타내고 있다. 포화모형에서는 지역, 일반과 저소득 가구 구분, 가구원 수, 가구주 나이, 가구주 성별, 가구주 교육 수준을 설명변수로 사용하였다. 최종 모형의 설명변수는 포화모형에서 가구주 성별을 제외한 변수이다. 로지스틱 회귀모형에 대한 변수별 계수값, 표준오차와 유의확률이 〈표 3-13〉에 나타나 있다.

〈표 3-13〉 원표본: 히핑 존재 여부에 대한 로지스틱 회귀분석 결과

	계수(Coefficient)	표준오차(S.E)	유의확률(p-value)
(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
저소득 가구	-0.340	0.101	0.001
2인 가구	-0.574	0.129	0.000
3인 가구	-0.298	0.127	0.019
4인 가구	-0.361	0.124	0.003
5인 이상 가구	-0.504	0.170	0.003
가구주 나이	-0.039	0.004	0.000
가구주 학력 고졸 이하	0.357	0.122	0.003
가구주 학력 대학 이상	0.000	0.137	0.999

자료: 한국보건사회연구원 (2006). 1차 한국복지패널조사의 원자료를 분석함.

로지스틱 회귀모형 분석 결과를 보면, 다른 변수의 효과를 제거하였을 때 강원, 충북, 광주, 전남, 전북, 제주는 다른 지역에 비하여 히핑이 발생할 확률이 작게 나타난다. 저소득 가구는 일반 가구에 비하여 히핑 발생 확률이 낮다. 가구원 수가 증가할수록 히핑이 발생할 확률은 낮아진다. 가구주 나이가 많을수록 중졸 이하의 교육 수준일 때 히핑 발생 확률은 작게 나타난다.

1차 연도 신규표본의 경상소득에 대한 히핑을 소득이 5의 배수인 경우로 지정하고 히핑 존재 여부와 여러 설명변수 간의 관계를 살펴보았다. 설명변수는 지역구분, 균등화 소득에 따른 가구 구분, 가구원 수, 가구주 성별, 가구주 교육 수준이다. 〈표 3-14〉부터 〈표 3-19〉에 각각의 결과가 나타나 있다.

〈표 3-14〉 신규표본: 지역별 히핑 존재 여부 빈도수와 비율

(단위: 가구)

	빈도수		비율	
	없음	있음	없음	있음
서울	136	58	70.1%	29.9%
인천/경기	213	76	73.7%	26.3%
부산/울산/경남	227	110	67.4%	32.6%
대구/경북	180	76	70.3%	29.7%
대전/충남	150	45	76.9%	23.1%
강원/충북	132	53	71.4%	28.6%
광주/전남북/제주	264	80	76.7%	23.3%
전체	1,302	498	72.3%	27.7%

자료: 한국보건사회연구원 (2012). 7차 한국복지패널조사의 원자료를 분석함.

〈표 3-15〉 신규표본: 소득에 따른 가구 구분별 히핑 존재 여부 빈도수와 비율

(단위: 가구)

	빈도수		비율	
	없음	있음	없음	있음
일반 가구	589	265	69.0%	31.0%
저소득 가구	713	233	75.4%	24.6%

자료: 한국보건사회연구원 (2012). 7차 한국복지패널조사의 원자료를 분석함.

〈표 3-16〉 신규표본: 히핑 존재 여부에 대한 경상소득의 차이

(단위: 만 원)

t-test	히핑 여부	
	없음	있음
경상소득 평균	2488.803	2912.922
t-statistic	-3.074	

자료: 한국보건사회연구원 (2012). 7차 한국복지패널조사의 원자료를 분석함.

〈표 3-17〉 신규표본: 가구원 수별 히핑 존재 여부 빈도수와 비율

(단위: 가구)

	빈도수		비율	
	없음	있음	없음	있음
1인	421	115	78.5%	21.5%
2인	491	203	70.7%	29.3%
3인	180	86	67.7%	32.3%
4인	164	74	68.9%	31.1%
5인 이상	46	20	69.7%	30.3%

자료: 한국보건사회연구원 (2012). 7차 한국복지패널조사의 원자료를 분석함.

〈표 3-18〉 신규표본: 가구주 성별에 따른 히핑 존재 여부 빈도수와 비율

(단위: 명)

	빈도수		비율	
	없음	있음	없음	있음
남자	846	370	69.6%	30.4%
여자	456	128	78.1%	21.9%

자료: 한국보건사회연구원 (2012). 7차 한국복지패널조사의 원자료를 분석함.

〈표 3-19〉 신규표본: 가구주 교육 수준별 히핑 존재 여부 빈도수와 비율

(단위: 명)

	빈도수		비율	
	없음	있음	없음	있음
중졸 이하	751	252	74.9%	25.1%
고졸 이하	332	135	71.1%	28.9%
대학교 이상	219	111	66.4%	33.6%

자료: 한국보건사회연구원 (2012). 7차 한국복지패널조사의 원자료를 분석함.

부산, 울산, 경남 지역에서 히핑 발생 비율이 다른 지역에 비하여 높게 나타난다. 일반 가구에서 히핑 발생 비율이 저소득 가구보다 높게 나타난다. 히핑 존재 여부를 기준으로 경상소득의 차이를 살펴본 결과, 히핑이 있는 경우는 약 2,912만 원으로 히핑이 없는 경우에 비하여 통계적으로 유의한 차이가 있음을 알 수 있다. 1인 가구에 비하여 가구원 수가 증가할수록 히핑 발생 비율이 높으며 3인 가구에서 가장 높게 나타났다. 가구주 성별이 남자일 때 히핑 발생 비율이 높게 나타났다. 가구주 교육 수준이 높을수록 히핑 발생 비율이 높게 나타났다.

〈표 3-20〉은 가구원 특성 분석을 통하여 유의하게 나타난 설명변수를 사용한 포화 모형과, 그 포화모형에서 유의하지 않은 설명변수를 하나씩 제거하여 나온 적합한 로지스틱 회귀모형의 결과를 나타내고 있다. 최종 모형은 가구주 성별, 일반 및 저소득 가구 구분을 설명변수로 사용하였다. 로지스틱 회귀모형에 대한 변수별 계수값, 표준오차 및 유의확률이 〈표 3-20〉에 나타나 있다.

로지스틱 회귀모형 분석 결과를 보면, 다른 변수의 효과를 제거하였을 때 가구주 성별이 남자일수록, 일반 가구일수록 히핑 발생 확률이 높게 나타났다.

〈표 3-20〉 신규표본: 히핑 존재 여부에 대한 로지스틱 회귀분석 결과

모형	신규표본	계수(Coefficient)	표준오차(S.E)	유의확률(p-value)
포화 모형	상수항	-0.326	0.393	0.406
	인천/경기	-0.189	0.209	0.366
	부산/울산/경남	0.222	0.199	0.263
	대구/경북	0.066	0.212	0.755
	대전/충남	-0.325	0.234	0.166
	강원/충북	-0.020	0.230	0.931
	광주/전남북/제주	-0.267	0.206	0.196
	저소득 가구	-0.055	0.132	0.678
	2인 가구	0.222	0.168	0.187
	3인 가구	0.201	0.207	0.331
	4인 가구	0.032	0.231	0.890
	5인 이상 가구	0.055	0.323	0.865
	가구주 성별_여	-0.277	0.160	0.083
	가구주 나이	-0.011	0.005	0.032
	가구주 학력 고졸 이하	-0.067	0.150	0.657
	가구주 학력 대학 이상	0.062	0.177	0.727
최종 모형	상수항	-0.751	0.076	0.000
	가구주 성별_여	-0.365	0.126	0.004
	저소득 가구	-0.197	0.113	0.082

자료: 한국보건사회연구원 (2012). 7차 한국복지패널조사의 원자료를 분석함.

제3절 히핑 보정 방법 문헌 연구

이번 절에서는 히핑 보정 방법에 관해 연구한 3가지 대표 논문의 방법론을 검토하고, 이 중 일부 방법을 실제 한국복지패널조사의 자료에 구현하여 방법의 특징을 살펴본다.

$z_1, \cdots, z_n$은 조사된 소득 값 관측변수이고 $y_1, \cdots, y_n$은 관측할 수 없는 소득의 참값(잠재변수)이라고 하자. 소득 외 인구사회학적 변수는 설명변수로 $x_1, \cdots, x_n$라고 하자. 히핑이 발생하는 지점은 $H = \mathrm{h}_1, \cdots, \mathrm{h}_s$라고 하고, $s \in N$ 이며 N은 자연수의 집합이다.

1. 패널자료 품질개선 연구(Ⅲ)(홍민기 등, 2014)

이 연구는 비례대체(fractional imputation) 방법에 기초하여 주어진 자료 x_i, z_i를 통하여 참값 y_i를 생성하는 방법으로 히핑 자료를 보정하였다. 참값 y_i는 다음의 모형으로부터 생성한다.

$$f(y_i|x_i,z_i) \propto f_1(y_i|x_i)g(z_i|x_i,y_i)$$

여기서, $f_1(y_i|x_i)$에 대한 모형은 다음과 같이 가정하고

$$\log(y_i) = x_i\beta + \epsilon_i, \qquad \epsilon_i \sim N(0,\sigma^2)$$

$g(z_i|x_i,y_i)$에 대한 모형은 y_i가 주어졌을 때 x_i와 z_i는 조건부 독립이라는 가정하에서 다음의 모형을 가정하였다.

$$z_i \sim N(y_i, \frac{y_i}{c}), \quad c = 2 \text{ 또는 } 4$$

2. Modeling heaped count data(Cummings, Hardin, McLain, Hussey, Bennett, & Wingood, 2015)

이 연구는 주로 가산 자료(count data)에서 발생하는 히핑 자료에 대하여 혼합분포를 사용하여 통계적 추정을 수행한다. 히핑이 존재하는 지점의 응답값은 정확한 응답값이 아니며 히핑 지점을 기준으로 양측 절단(censored)되었다고 가정한다. 예를 들어 10의 배수 지점에서 히핑이 발

생한다면 10, 20 등으로 응답한 히핑 지점에서 응답값의 분포는 다음을 따른다고 가정한다.

$$P(Y \in (y-[10/2],\ y+[10/2])$$

여기서 $[x]$는 x를 넘지 않는 최대의 정수를 뜻한다. 히핑이 발생하지 않는 지점의 가산 자료에 대해서는 $P(y=y)$ 모형을 가정하며 이때 많이 사용하는 포아송(Poisson), 음이항(Negative Binomial) 분포 등을 가정한다.

이를 일반화하면 다음과 같은 수식으로 나타낼 수 있다. 히핑이 발생하지 않는 y에 대해서는 포아송 모형을 가정한다.

$$f(y_i;\mu_i)=\frac{\mu_i^{y_i}e^{-\mu_i}}{y_i!},\ \ y_i=0,1,2,...,\ \ \mu_i>0$$
$$\mu_i=\exp(x_i\beta)$$

$$y_{Li}=\max(0,y_i-\delta_i)$$
$$y_{Ri}=y_i+\delta_i$$
$$\delta_i=\max_{j=1,...,s}\{[h_j/2]\times I(y_i\in H)\},\qquad H=\{h_1,...,h_s\}$$

히핑 발생 지점에서 히핑이 발생하는 구간의 확률은 생존확률을 활용하여 표현할 수 있고, 양측 절단된 히핑 지점에서 가능도 함수는 생존확률을 이용해 표현할 수 있다.

$$L=\sum_{i=1}^{n}\log P(Y\in(y_{Li},y_{Ri})|\,Y\sim Poi),$$

$$P(Y \in (y_{Li}, y_{Ri}) | Y \sim Poi) = p_{1i} - p_{2i},$$

$$p_{1i} = p(Y > y_{Li} - 1 | Y \sim Poi) = \Gamma(y_{Li}, \mu_i)$$
$$p_{2i} = p(Y > y_{Ri} | Y \sim Poi) = \Gamma(y_{Ri} + 1, \mu_i),$$

$$\Gamma(y, \mu) = \frac{1}{\Gamma(y)} = \int_0^y t^{\mu - 1} e^{-t} dt$$

y에 대하여 GP(Generalized Poisson) 모형과 음이항 모형을 가정할 수도 있고 이에 대한 구현은 Stata(통계패키지)를 통해서 가능하다. 하지만 이러한 모형들은 가산 자료에만 적용 가능하므로, 이 연구의 한국복지패널조사 자료에는 적용할 수 없으며 추후 가산 자료에 대한 히핑 자료 보정 시 활용하도록 한다.

3. Kernel density estimation for heaped data(Marcus groß & Ulrich Rendtel, 2016)

이 연구는 커널 함수 추정법을 활용하여 베이지안 방법론을 적용한 히핑 자료 보정 방법을 다루고 있다. 히핑이 발생하는 지점을 $H = \mathrm{h}_1, \cdots, \mathrm{h}_\mathrm{s}$라 할 때 r_i는 i번째 응답값이 갖는 히핑 지점이며 이에 대응되는 확률을 $P = (p_1, \ldots, p_m)$라 하자. 이때 히핑은 응답값을 반올림을 통해서 이루어지며 반올림은 참값 y_i가 응답값 z_i를 기준으로 구간 $(z_i - h_i/2,\ z_i + h_i/2)$에 속할 때 이루어진다고 가정한다.

여기서 $K(\cdot)$는 주어진 커널 함수이며 h는 bandwidth로 추후 추정해야 할 대상이다.

응답값 Z에 대한 가능도 함수는 다음과 같이 표현할 수 있다.

$$L(Z|\ Y,R,\theta,) = \prod_{i=1}^{n}(\pi(z_i|y_i,r_i)\times\pi(r_i|P))\times\pi(Y|\theta),$$
$$\pi(z_i|y_i,r_i) = \begin{cases}1 & \text{if}\quad y_i\in(z_i-r_i/2,\ z_i+r_i/2)\\ 0 & else\end{cases},$$
$$\pi(r_i|P)\sim \text{multinomal}$$

y에 대한 분포 $\pi(Y|\theta)$는 비모수 분포를 가정하며 커널 분포 함수 추정법을 통하여 다음과 같이 추정할 수 있다.

$$\pi(Y|\theta=h) = \prod_{i=1}^{n}\widehat{f_h}(y_i),$$
$$\widehat{f_h}(y) = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{y-y_i}{h}\right)$$

따라서 응답값 Z가 주어졌을 때 Y, H에 대한 사후 분포는 다음과 같다. 여기서 P와 θ는 hyperprior이다.

$$\pi(Y,R,\theta,P|Z)\propto \pi(Z|\ Y,R)\times\pi(R|P)\times\pi(Y|\theta)\ \ (likelihodd)$$
$$\times\ \pi(P)\pi(\theta)\ \ (prior)$$

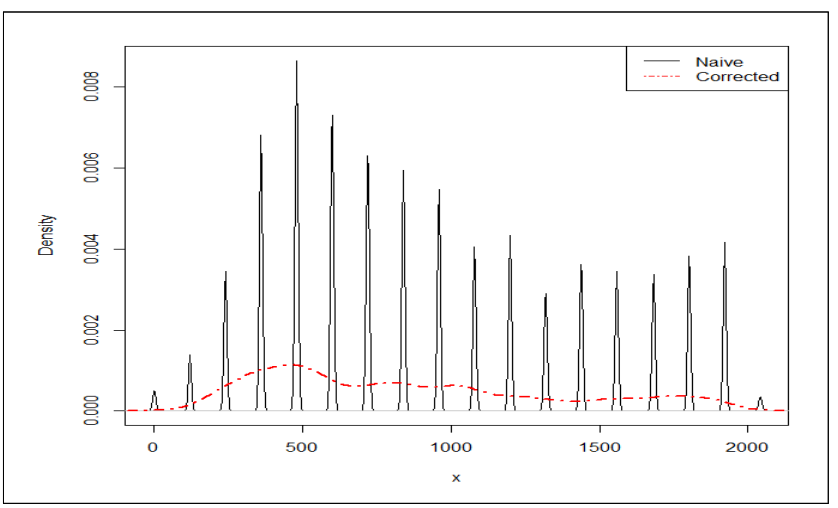
지금까지 살펴본 커널 함수 추정법을 활용한 히핑 보정 방법은 R 프로그램 내부에 패키지 형태로 구현되어 있다. ‘Kernelheaping’ 패키지 내부의 ‘dheaping’ 함수를 사용하여 앞선 방법을 구현할 수 있다. 1차 연도 경상소득에 대하여 히핑 발생 지점을 120의 배수로 하여 커널 히핑 보

정 방법을 적용하였다. 경상소득의 경우 최대값이 약 3만으로 방법의 특성상 많은 히핑 지점이 있을 시 베이지안 방법의 사후 분포 수렴에 문제가 생길 수 있으며 계산 시간이 너무 길어져 경상소득 0에서 2,000으로 제한하여 방법을 구현하였다.

〔그림 3-7〕은 히핑 발생 지점을 기준으로 커널 분포 함수를 보정 없이 추정한 결과와 보정 후 추정한 결과가 나타나 있다. 보정 전 히핑 지점에서는 120의 배수마다 분포가 집중되어 나타났지만, 보정 후 분포 함수 추정 결과는 모든 지점에서 부드러운 분포 함수의 곡선을 보이는 점을 확인할 수 있다.

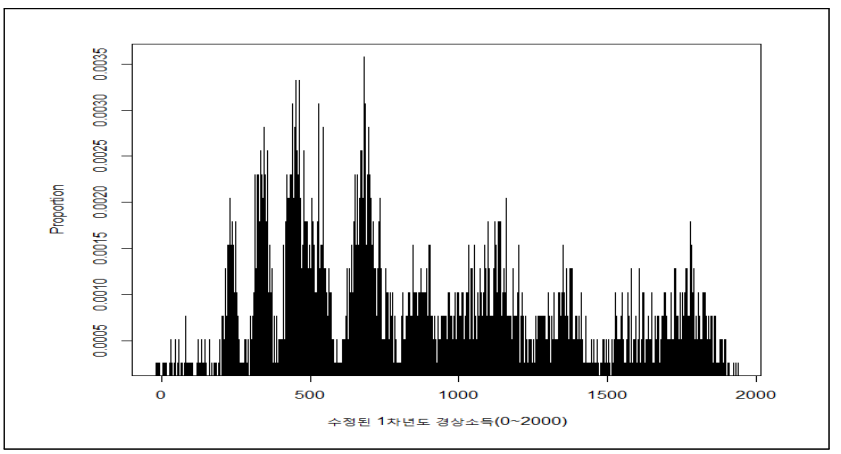
〔그림 3-8〕은 커널 함수를 통한 히핑 보정 결과에 대한 spike plot이다. 앞선 히핑 보정 전에 대한 〔그림 3-2〕에서는 특정 지점들에서 다른 지점에 비하여 분포가 튀는 현상을 확인할 수 있었으나, 보정된 spike plot에서는 훨씬 더 부드러운 분포임을 확인할 수 있다.

〔그림 3-7〕 커널 함수를 통한 히핑 보정 분포 결과



자료: 한국보건사회연구원 (2006~2020). 1~15차 한국복지패널조사의 원자료를 분석함.

[그림 3-8] 커널 함수를 통한 히핑 보정 spike plot



자료: 한국보건사회연구원 (2006~2020). 1~15차 한국복지패널조사의 원자료를 분석함.

제4절 확률 대체법을 통한 히핑 보정 방법

3절에서 소개한 홍민기 등(2014)에서는 한국노동패널조사 자료의 소득에 대해 히핑 보정을 시행하였다. 이때는 히핑 보정을 모든 경우에 대하여 실행하였는데, 이 연구에서는 히핑이 발생한 경우와 아닌 경우를 나누어 히핑 보정을 수행하는 방법을 고려하였다.

먼저 히핑을 나타내는 지시 변수 I_h를 생각하자.

$$I_h = \begin{cases} 1 & y_i \neq z_i \text{ (히핑이 발생한 경우)} \\ 0 & y_i = z_i \text{ (그 외)} \end{cases}$$

여기서 I_h에 관련한 확률값은 설명변수 x_i에 의존하는 어떤 함수로 표

현될 수 있다고 가정하자. 즉, 로지스틱 혹은 프로빗 같이 알려진 함수 $\pi(\cdot)$에 대하여

$$\Pr(I_h = 1 | x_{i,} y_i) = \pi(\phi_0 + x_i{}' \phi_1)$$

표현되고 (ϕ_0, ϕ_1)은 unknown 모수 값으로 볼 수 있다. 이 연구에서는 x_i에만 의존하게 모형을 세웠으나, y_i에도 의존하게 모형을 세울 수 있을 것이다. 히핑 발생 여부에 대한 확률 $\Pr(I_h = 1 | x_i)$에 대한 로지스틱 회귀모형 분석 추정에 관한 결과는 2절의 〈표 3-13〉에 나타나 있다.

또한 $\tilde{y}_i$를 y_i에서 측정오차가 추가되어 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.

이 같은 경우 히핑 현상을 보정한 보정 응답값은 대체(imputation) 방법을 통하여 생성할 수 있고 그 형태는 다음과 같다.

$$y^{**} = \Pr(I_h = 0 | x_i) y_i + \Pr(I_h = 1 | x_i) y_i^*$$

이때 y_i^*는 다음의 분포로부터 생성되며

$$y_i^* \sim f(y_i | x_i, z_i) \propto f_1(y_i | x_i) g(z_i | x_i, y_i),$$

$f_1(y_i | x_i)$에 대한 모형과 $g(z_i | x_i, y_i)$모형은 다음과 같다.

$$\log(y_i) = x_i\beta + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_1^2)$$

$$z_i \sim N(\gamma_0 + \gamma_1 y_i,\ \sigma_2^2) \text{ 또는 } \log(z_i) \sim N(\gamma_0 + \gamma_1 \log(y_i),\ \sigma_2^2)$$

모수 $(\beta,\ \sigma_1^2)$과 $(\gamma,\ \sigma_2^2)$은 EM 알고리즘을 통하여 추정할 수 있고 그 과정은 다음과 같다.

[Step 1] y_i 대신 z_i를 사용하여 $f(z_i|\, x_i;\, \beta,\, \sigma_1^2)$의 모수를 추정하고 이를 가지고 $(\beta,\ \sigma_1^2)$의 초기치를 구한다. 이로부터 $y_i^{*(j)} \sim f(y|\, x;\, \beta,\, \sigma_1^2), \quad j = 1,\ \cdots,\ M$을 생성한다. 생성된 $y_i^{*(j)}$를 사용하여 (γ, σ_2^2)의 초기치를 구한다.

[Step 2] 각 i에서 생성된 M개의 $y_i^{*(j)}$를 이용하여 fractional weights를 계산한다.

$$w_{ij}^{*} \propto f(y_i^{*(j)} \mid x_i; \beta, \sigma_1^2) f(z_i \mid y_i^{*(j)}; \gamma, \sigma_2^2),$$

$$\sum_{j=1}^{M} w_{ij}^{*} = 1$$

[Step 3] 현 단계에서의 각 모수의 추정치와 그에 따른 fractional weights를 이용하여 다음의 score function을 0으로 만드는 새로운 모수 추정치를 구한다.

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

1차 연도 한국복지패널조사 자료에서 경상소득에 대한 히핑 보정을 수행하였는데, 설명변수로는 지역구분, 균등화 소득에 따른 가구 구분, 가구원 수, 가구주 나이, 가구주 교육 정도를 사용하였다. 모수 추정 결과는 〈표 3-21〉과 같다.

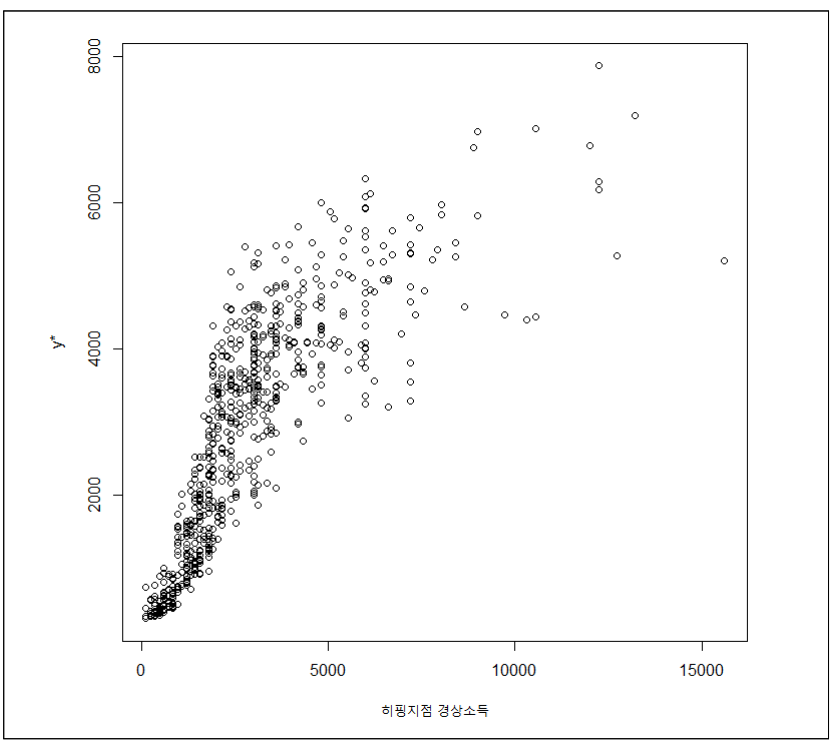
〈표 3-21〉 한국복지패널조사 자료 히핑 보정 모형 모수 추정 결과

모수	계수(Coefficient)	모수	계수(Coefficient)
상수항	7.560	γ_0	0.556
인천/경기	-0.053	γ_1	0.924
부산/울산/경남	-0.025	σ_2^2	0.204
대구/경북	-0.117		
대전/충남	-0.054		
강원/충북	-0.118		
광주/전남북/제주	-0.112		
저소득 가구	-1.253		
2인 가구	0.438		
3인 가구	0.711		
4인 가구	0.885		
5인 이상 가구	1.010		
가구주 나이	-0.003		
가구주 학력 고졸 이하	0.042		
가구주 학력 대학 이상	0.229		
σ_1^2	0.076		

자료: 한국보건사회연구원 (2006). 1차 한국복지패널조사의 원자료를 분석함.

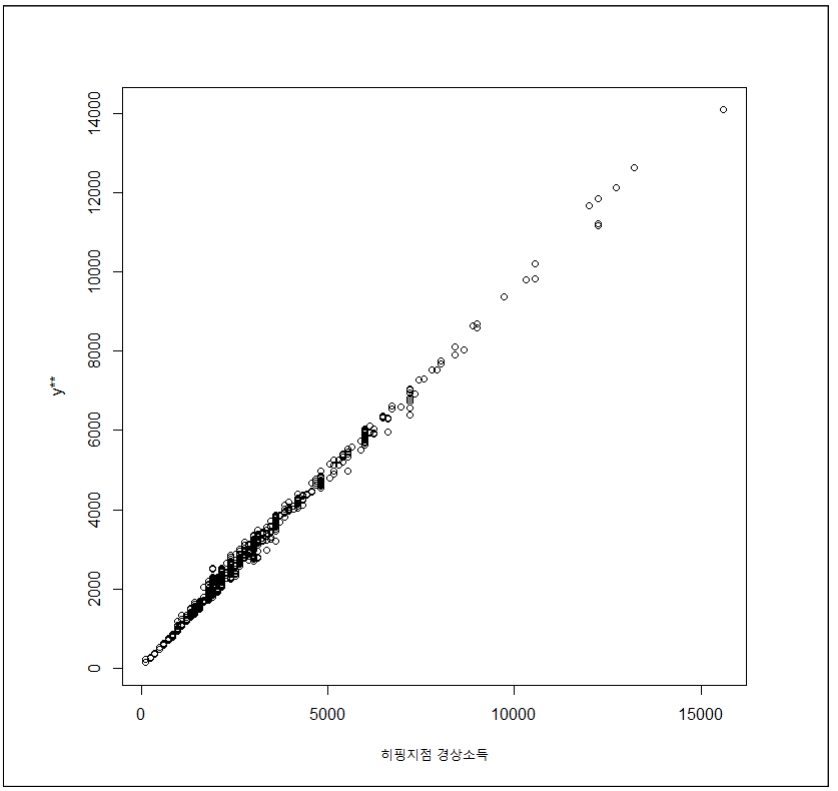
추정된 모수와 최종 fractional weights를 사용하여 얻어진 히핑 보정 대체값 y^*와 조사된 응답값과의 관계가 〔그림 3-9〕에 나타나 있다. 히핑 보정을 통해서 값들이 전체적으로 줄어들었지만, 경상소득이 증가함에 따라 분산이 커지는 모습을 볼 수 있다. 히핑 확률까지 고려한 히핑 보정 대체값 y^{**}와 조사된 응답값과의 관계는 〔그림 3-10〕에 나타나 있다. 히핑 보정을 통해서 값들이 전체적으로 줄어들었으며 그 조사된 응답값과의 차이도 크지 않다. 선행 연구에서 가정하듯이 히핑이 응답값 기준 일정 범위에서 발생한다면 조금 더 타당한 보정 방법이라고 생각된다.

〔그림 3-9〕 경상소득의 응답값과 대체값(y^*)의 산점도



자료: 한국보건사회연구원 (2006). 1차 한국복지패널조사의 원자료를 분석함.

〔그림 3-10〕 경상소득의 응답값과 히핑 확률 반영한 대체값(y^{**})의 산점도



자료: 한국보건사회연구원 (2006). 1차 한국복지패널조사의 원자료를 분석함.

균등화 소득에 의한 저소득 가구의 비율은 1차 연도에 46.3%로 계산된다. 이후 차수 연도에서는 저소득 가구의 비율이 41~43%로 나타나는 점으로 보아, 1차 연도의 높은 저소득 가구 비율이 히핑으로 인하여 나타나는 현상인지 살펴보고자 한다. 〈표 3-22〉에는 보정 전·후의 저소득 가구에 대한 평균 및 표준편차값이 나타나 있다. 저소득 가구의 경상소득이 증가하는 방향으로, 또 표준편차는 줄어드는 방향으로 보정이 실행되었음을 알 수 있다.

〈표 3-22〉 히핑 보정 전·후에 따른 저소득 가구의 경상소득 평균 및 표준편차

(단위: 만 원)

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

자료: 한국보건사회연구원 (2006). 1차 한국복지패널조사의 원자료를 분석함.

경상소득에 대하여 (중위값×0.6)의 값을 기준으로 하위소득 가구의 비율을 살펴보았다. 〈표 3-23〉은 차수별 기준 소득과 하위소득 가구의 비율에 대한 것이다. 마지막 행에 1차 연도에 히핑 보정된 경상소득을 통해서 계산된 결과를 보면, 기준값은 1,049만 4천 원이고 비율은 34.5%로 나타났다. 이는 보정 전 1차 연도(기준값: 1,032만 6천 원, 비율: 34.7%)와 비교했을 때 차이가 크지 않은 편이었다. 그러므로 히핑이 1차 연도 저소득 가구의 높은 비율의 큰 원인은 아니라고 생각한다.

〈표 3-23〉 (경상소득의 중위값×0.6) 기준에 따른 하위소득 가구 비율

(단위 : 만 원)

	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	32.4%
6차	1,549.2	33.4%
7차	1,440	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	33.0%
1차 (보정된 값)	1,049.4	34.5%

자료: 한국보건사회연구원 (2006~2020). 1~15차 한국복지패널조사의 원자료를 분석함.

제5절 소결

자기응답으로 조사된 자료에서 발생할 수 있는 히핑 현상에 대해 살펴보았다. 이때 관심 변수의 형태(type)는 연속형이다. 먼저 자료에 히핑 현상이 있는지 파악하기 위하여 관심 변수에 대한 응답 분포를 살펴보았다. 관심 변수값에 대한 상위 응답 비중이 특정 배수의 형태를 가지는지를 분석하여 spike plot을 통해 확인하였다.

이 장에서는 한국복지패널조사 자료 1차 연도부터 15차 연도를 활용하여 히핑 현상이 있는지 파악해 본 결과 주로 처음 응답한 1차 연도의 경상소득 및 근로소득에서 히핑 현상이 존재함을 알 수 있었다. 또한 경상소득에 대한 히핑 존재 여부와 가구 특성 변수 간의 관계에 대해 로지스틱 회귀모형 분석을 시행하였다. 강원, 충북, 광주, 전남, 전북, 제주는 다른 지역에 비하여 히핑이 발생할 확률이 낮게 나타났다. 저소득 가구는 일반 가구에 비하여 히핑 발생 확률이 낮았으며, 가구원 수가 증가할수록 히핑이 발생할 확률은 낮아졌다. 또한, 가구주 나이가 많을수록 중졸 이하의 교육 수준일 때 히핑 발생 확률은 낮게 나타남을 알 수 있었다.

한편, 히핑 보정 방법에 대한 기존 문헌을 검토하였다. 가산 자료의 경우 포아송, 음이항 등의 모수 가정을 통해서 보정하는 방법, 커널 함수 추정 방법을 활용한 비모수 히핑 보정 방법 등을 가지고 모형 가정과 추정 과정에 대해 자세히 살펴보았다. 이 중에서 커널 함수 추정 방법을 통한 히핑 보정 방법을 1차 연도 한국복지패널조사 자료의 경상소득에 적용해 보았다. R 통계패키지에 'Kernelheaping' 패키지 내부의 'dheaping' 함수를 통해 히핑 보정 방법을 실시한 결과, 보정 전에는 다른 지점에 비해 특정 지점에서 분포가 튀는 현상을 가졌으나 보정 후에는 부드러운 분포를 가지는 것을 확인할 수 있었다.

마지막으로 홍민기 등(2014)에서 제안하였던 히핑 보정 방법을 확장하여 새로운 히핑 보정 방법을 제안하였다. 즉 히핑이 발생한 경우와 그렇지 않은 경우로 나누어 히핑 발생 확률을 추정(앞서 실시한 로지스틱 회귀모형의 분석 결과 활용)한 다음에, 대체를 통해 히핑 보정을 시행하여 한국복지패널조사 자료의 경상소득에 적용한 결과, 값들이 전체적으로 줄어들었으며 조사된 응답값과의 차이도 일정 구간에 속하였다. 선행연구에서 가정하듯이 히핑이 응답값 기준 일정 범위에서 발생한다면 조금 더 타당한 보정 방법이라고 생각한다. 일반 가구와 저소득 가구를 나누어 히핑 보정된 소득의 분포를 살펴보면 저소득 가구의 경상소득이 증가하는 방향으로, 또한 표준편차는 줄어드는 방향으로 보정되었음을 알 수 있었다. 경상소득에 대하여 (중위값×0.6)의 값을 기준으로 하위소득 가구의 비율도 살펴본 바, 1차 연도에서 히핑 보정된 경상소득에 대한 하위소득 가구의 비율은 보정 전과 비교했을 때 차이가 크지 않은 것으로 나타났다. 이를 통해 히핑이 1차 연도 저소득 가구의 높은 비율의 큰 원인이 아니라고 생각한다.

제4장

조사 자료의 금액 변수에 대한 측정오차 보정 방안

제 4 장 조사 자료의 금액 변수에 대한 측정오차 보정 방안

제1절 개요

비표집오차 중 하나인 측정오차는 자료를 수집하는 과정에서 발생하는 오차로, 응답값과 알려지지 않은 실제값 간의 차이로 발생한다. 이렇듯 보통 실제값을 알고 있지 않아서 표본조사 자료에서 측정오차를 파악하기는 쉽지 않은 편이다.

한편, 가계금융복지조사 자료의 경우 2017년에 한해서 주요 금액 관련 설문 문항에 대한 응답값뿐만 아니라 행정보완값을 제공하고 있다. 이러한 자료 특성을 활용한다면 표본조사 자료에서의 측정오차를 파악해 볼 수 있을 것이다.

이 장에서는 2017년 가계금융복지조사 자료에서 가구주의 개인 근로소득, 가구의 근로·자녀장려금, 가구원의 기초연금 변수에 대하여 응답값과 행정보완값 간의 관계를 파악하여 측정오차가 존재하는지 살펴보았다. 측정오차가 존재하면 여러 가지 방법을 사용하여 측정오차를 보정하였다. 그 방법으로는 측정오차 보정을 통한 회귀계수 추정, 히핑 보정, 비례핫덱대체 방법이 있는데, 이런 방법들에 관한 이론적 개념과 적용 방법을 검토하고, 가계금융복지조사 자료에 적용하여 결과를 살펴보았다.

제2절 가계금융복지조사 자료 현황 분석

통계청은 금융감독원 및 한국은행과 공동으로 가계금융복지조사를 수행하고 있다. 전국의 2만여 표본 가구를 대상으로 2010년부터 해마다 조사를 시행하고 있다(통계청, 2018). 조사목적은 소득, 자산, 부채 등에 대한 규모, 구성 및 분포와 미시적 재무 건전성을 파악하여 사회 및 금융 관련 정책과 연구에 활용하는 데 있다. 조사 대상은 가구 단위로 전국 동·읍·면에 거주하는 1인 이상의 표본 가구이고, 조사 단위는 1인 가구 및 혈연, 결혼, 입양 등으로 맺어져 생계를 함께 하는 가족이다. 조사항목은 금융과 복지 2부문으로 구성되어 있는데, 금융 부문은 가구 구성, 자산 및 자산운용, 부채 및 부채상환 능력, 소득 및 지출 등이고 복지 부문은 가구 구성, 자산, 부채, 소득, 지출, 노후생활 등이다.

그중에서 2017년 조사 자료는 응답값뿐만 아니라 행정보완값도 함께 제공하고 있다. 행정보완값은 행정 자료에서 얻은 자료로, 이 장에서 살펴볼 측정오차 분석을 하는데 적합한 자료라고 생각한다. 그래서 여러 변수 중에서 선정한 분석 변수는 측정오차가 발생할 가능성이 큰 편에 속하는 금액 변수인 가구주의 개인 근로소득(이하 '근로소득'), 가구의 근로자녀장려금(이하 '근로자녀장려금'), 가구원의 기초연금(이하 '기초연금')으로 선정하였다. 〈표 4-1〉은 변수에 대한 조사 자료와 행정 자료의 개념과 포괄범위에 대한 것이다. 근로소득은 조사 자료와 행정 자료의 개념과 포괄범위가 '거의 일치'로 볼 수 있고, 기초연금은 '일치'이다. 제대로 정리되어 있지는 않으나, 근로자녀장려금도 기초연금처럼 수혜 대상이 명확하게 정의되어 있으므로 '일치'라고 볼 수 있다.

〈표 4-1〉 분석 변수에 대한 조사 자료와 행정 자료의 개념 및 포괄범위

	조사 자료	행정 자료
근로소득	세금, 각종 부담금 등 세금 공제 전 소득 근로의 대가로 받은 일체의 현금·현물 보수 · 임금 및 수당(*) (*) 기본급, 근속급, 가족수당, 야근수당 등 · 상여금(명절휴가비·가계 지원비·정근수당·성과급 등) · 퇴직수당, 명예퇴직수당 포함 단, 퇴직급(퇴직일시금)은 제외 * 연말정산 환급금은 포함하지 않음	거의 일치 · 원천징수, 연말정산, 종합소득신고에 의한 소득 · 소득 하위층 미신고 소득 발생 가능 · 행정 자료는 현물성 보수 등은 미포함 · 자활급여 포함
기초연금	만 65세 이상의 수급 대상자에게 지급하는 현금 급여	일치

자료: 통계청. (2020). 가계금융복지조사에서의 조사자료와 행정자료의 통합방법 이해. p.27

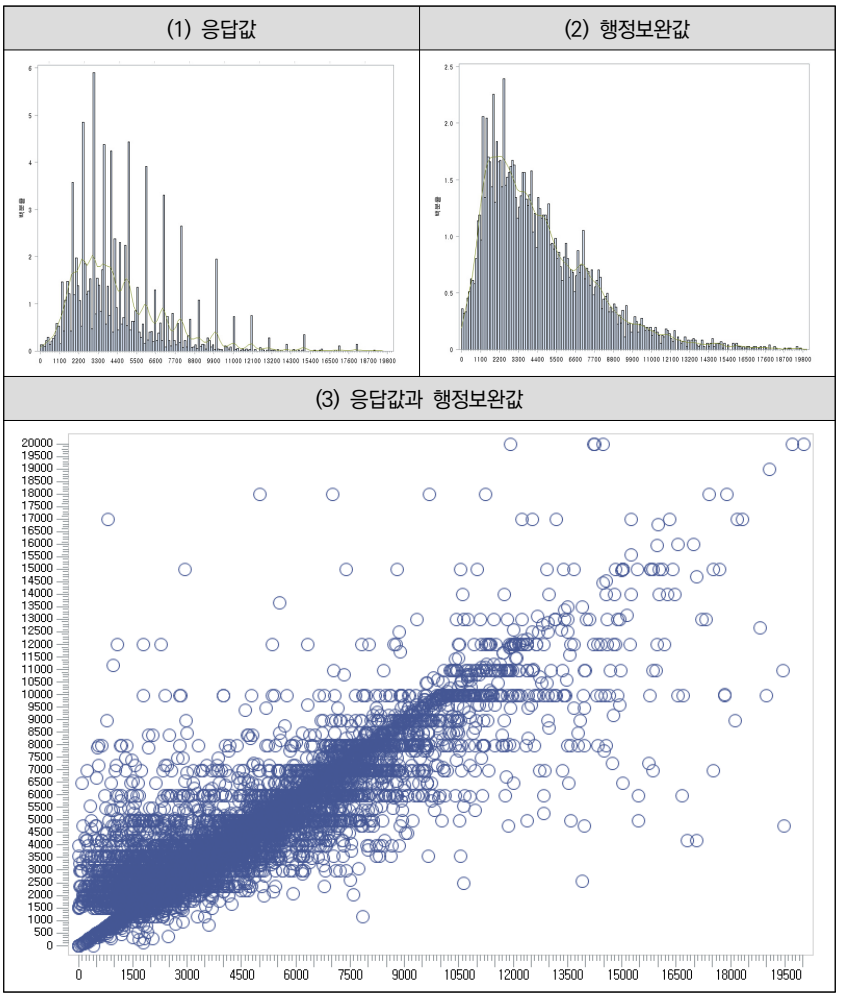
1. 근로소득

분석 대상은 7,274명이며 종사상지위가 상용 임금근로자인 근로소득으로 한정하였다. 다른 종사상지위(임시/일용근로자, 고용원이 있는/없는 자영업자, 무급가족종사자, 기타 종사자)에 비해 상용 임금근로자의 경우는 행정보완자료(국세청 자료)의 정확성이 높은 편이어서 참값이라고 가정할 수 있다고 판단하였다. 행정보완자료도 측정오차를 포함할 수 있다는 한계점을 가지고 있으나, 이러한 점들을 가능한 한 최소화하여 연구 대상 모집단으로 구성할 수 있기 때문이다. 또한 상용직을 중심으로 한 근로소득(일용소득 제외)은 국세청에 신고한 소득이 더 정확하다고 가정하고 있다(통계청, 2020, p.74). 또한 가구주로 한정한 이유는 가구원과 다르게 근로소득과 연관성이 높은 직업, 산업 등의 다양한 정보가 있어 활용할 수 있기 때문이다.

앞으로 조사 자료의 값은 '응답값'으로, 행정 자료의 값은 '행정보완값'이라고 정의한다.

〔그림 4-1〕을 보면 (1) 응답값 분포의 경우, (2) 행정보완값과는 다르게 일정한 값마다 꽤 높은 비율을 가지는 특정 값들이 잦은 반면에, 행정보완값은 전체적으로 완만한 분포를 나타내고 있다. (3) 응답값과 행정보완값 간의 분포를 보면 $y = x$를 기준으로 위는 과대 보고이고 아래는 과소 보고한 경우이다. 과대 보고의 경우 저 근로소득에서 많이 나타나고 있는 반면에 과소 보고는 전반적으로 고르게 분포되어 있다.

〔그림 4-1〕 근로소득에 대한 응답값과 행정보완값의 분포



주: 1) (1), (2)의 단위는 X축 만원이고, Y축은 %임.
2) (1)의 전체수는 7,241명이고 (2)는 7,232명임(2억 미만인 경우만 시각화).
3) (3) X축은 행정보완값이고, Y축은 응답값임(단위: 만 원).
4) (3)은 행정보완값과 응답값 모두 2억 미만인 경우에 대해서만 시각화함(7,229명).
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

다음 〈표 4-2〉를 보면 근로소득의 응답값 평균은 4,653만 1천 원으로 행정보완값(4,709만 7천 원)에 비해 56만 6천 원 적게 나타났다. 응답값과 행정보완값 간의 대응 표본 t-검정(paired t-test)을 시행한 결과, 유의수준 5%에서 응답값과 행정보완값은 통계적으로 유의한 차이가 있다고 할 수 있다.

〈표 4-2〉 근로소득 응답값과 행정보완값의 대응표본 t-검정 결과

	N	평균	표준편차	최솟값	최댓값	t Value
응답값	7,274	4,653.1	3,211.6	0	73,400	-2.09 (**)
행정보완값	7,274	4,709.7	4,380.6	0	158,935	

주: t Value ( ) 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

응답값과 행정보완값 간 차이를 3개 집단으로 구분하였다(〈표 4-3〉 참조). 응답값을 행정보완값 보다 더 작게 응답한 경우가 48%였고, 응답값을 더 크게 응답한 비율은 43.2%였다. 응답값과 행정보완값이 같은 경우는 8.8%로 낮은 편에 속하였다. 3개 집단별 분포를 보면 응답값이 행정보완값보다 작은 경우의 평균이 가장 큰 편이었고, 다음으로 응답값이 행정보완값보다 큰 경우, 같은 경우 순서로 나타났다.

3개 집단별 대응 표본 t-검정 결과는 다음과 같다. 응답값이 행정보완값보다 작은 경우, 평균 차이는 -1,040만 원이고 행정보완값의 표준편차가 응답값에 비해 매우 큰 편이며 유의수준 0.1%에서 응답값과 행정보완값 간 통계적으로 유의한 차이가 있는 것으로 나타났다. 응답값이 행정보완값보다 큰 경우의 평균 차이는 1,024만 원이고 응답값의 표준편차가 행정보완값에 비해 약간 크나 차이가 크지 않은 편이었다. 유의수준 0.1%에서 응답값과 행정보완값 간에 통계적으로 유의한 차이가 있는 것으로 나타났다.

〈표 4-3〉 근로소득 3개 집단별 대응 표본 t-검정 결과

	자료	N	평균	표준편차	최솟값	최댓값	t Value
응답값 〈 행정보완값	조사	3,493 (48.0%)	5,070.7	3,440.1	0	73,400	-22.77 (***)
	행정보완		6,110.3	5,313.7	140	158,935	
응답값 = 행정보완값	조사	637 (8.8%)	2,989.7	2,881.9	0	30,000	-
	행정보완						
응답값 〉 행정보완값	조사	3,144 (43.2%)	4,526.2	2,875.4	1,310	35,000	39.88 (***)
	행정보완		3,502.2	2,667.6	8	32,035	

주: t Value ( ) 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

3개 집단의 범주를 종속 변수로 하여 다항 로지스틱 회귀모형(multinomial logistic regression model) 분석을 시행하였다. 준거집단은 응답값과 행정보완값이 같은 경우(응답값=행정보완값)이고, 설명변수는 가구주의 연령, 성별, 학력, 배우자 유무, 가구원 수, 부채 유무, 입주 형태, 수도권 여부이다. 최적의 회귀모형을 선택하고자 후진제거법(backward elimination)을 활용하였으며, 〈표 4-4〉는 가구주의 학력과 수도권 여부가 제외된 최종 모형의 분석 결과이다.

(응답값=행정보완값) 경우를 준거집단으로 하여 (응답값〈행정보완값) 경우에 영향을 주는 요인으로는 부채 유무를 제외한 설명변수에서 모두 통계적으로 유의한 결과를 나타내고 있다. 가구주의 성별이 남성(기준변수: 여성)일수록 (응답값〈행정보완값)일 가능성이 (응답값=행정보완값)일 가능성보다 12.6% 증가한다고 볼 수 있고, 가구원 수가 한 단위 증가할수록 (응답값〈행정보완값)일 가능성이 8.2% 증가한다고 볼 수 있다.

(응답값=행정보완값) 경우를 준거집단으로 하여 (응답값〉행정보완값) 경우에 영향을 주는 요인으로는 가구원 수를 제외한 설명변수에서 모두 통계적으로 유의한 결과를 나타내고 있다. 가구주의 성별이 남성일수록

(응답값〉행정보완값)일 가능성이 (응답값=행정보완값)일 가능성보다 80.6% 증가한다고 볼 수 있고, 가구주의 연령이 한 단위 증가할수록 (응답값〉행정보완값)일 가능성이 4.7% 증가한다고 볼 수 있다.

〈표 4-4〉 근로소득에 대한 다항 로지스틱 회귀모형 분석 결과

변수	응답값 〈 행정보완값			응답값 〉 행정보완값		
	coefficient	SE	exp(b)	coefficient	SE	exp(b)
상수항	2.934(***)	0.272	-	3.433(***)	0.272	-
가구주의 성별_남	0.420(***)	0.126	1.522	0.591(***)	0.127	1.806
가구원 수	0.079(*)	0.048	1.082	-0.048	0.048	0.953
가구주의 연령	-0.038(***)	0.004	0.963	-0.049(***)	0.004	0.953
가구주의 배우자 유무_있음	0.500(***)	0.142	1.648	0.385(***)	0.143	1.470
입주 형태_기타	-0.810(***)	0.096	0.445	-0.624(***)	0.096	0.536
부채 유무_있음	-0.052	0.096	0.950	0.175(*)	0.097	1.191

주: t Value ( ) 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

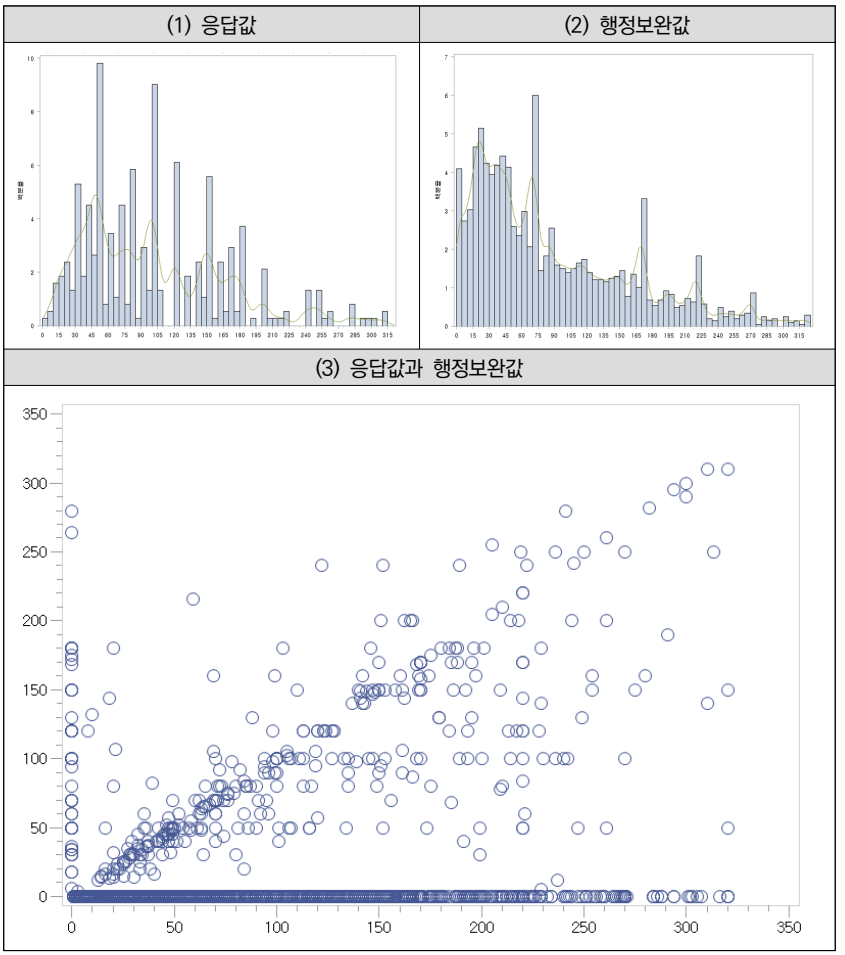
2. 근로자녀장려금

분석 대상은 응답값과 행정보완값이 모두 0원인 경우는 제외한 2,081가구이다.

응답값이 0원인 경우가 81.9%를 차지하여 매우 높게 나타나 제외하였음에도 불구하고 [그림 4-2]를 보면 (1) 응답값과 (2) 행정보완값 간의 분포 형태가 많이 다르게 나타났다. (3) 응답값과 행정보완값 간의 분포를 보면 행정보완값의 전 구간에서 응답값이 0원으로 나타나고 있다. 그 외 관측치는 주로 과소 보고 영역에 더 많이 분포되어 있다. 이러한 현상은 근로자녀장려금 수혜가 일 년에 한 번이기 때문에 응답자의 회고오차[1)]로

인한 측정오차라고 볼 수 있다.

[그림 4-2] 근로자녀장려금에 대한 응답값과 행정보완값의 분포



주: 1) (1), (2)의 단위는 X축 만원이고, Y축은 %임.
2) (1)은 0원 제외하고 시각화함(1,704가구, 81.9%가 0원이라고 응답).
3) (2), (3)의 전체수는 2,081가구임.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

1) 응답자가 과거 사건 또는 상황 등을 정확하게 응답하지 않는 것을 의미함.

근로자녀장려금의 응답값 평균은 18만 3천 원으로 행정보완값(87만 4천 원)에 비해 -69만 1천 원 적게 나타나 매우 큰 차이를 보였다. 대응표본 t-검정 결과를 보면 유의수준 5%에서 응답값과 행정보완값 간 통계적으로 유의한 차이가 있는 것으로 나타났다(〈표 4-5〉 참조).

〈표 4-5〉 근로자녀장려금 응답값과 행정보완값의 대응표본 t-검정 결과

	N	평균	표준편차	최솟값	최댓값	t Value
응답값	2,081	18.3	47.8	0	310	-42.6 (***)
행정보완값	2,081	87.4	71.6	0	320	

주: t Value ( ) 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

응답값과 행정보완값 간 차이를 3개 그룹으로 구분해 보면, 응답값을 행정보완값 보다 더 작게 응답한 경우가 91.5%였고 응답값을 더 크게 응답한 비율은 5.8%이었다. 응답값과 행정보완값이 같은 경우는 2.6%로 현저히 낮게 나타났다(〈표 4-6〉 참조).

3개 집단별 분포 결과를 보면, 응답값과 행정보완값이 같은 경우의 평균은 113만 3천 원이며, 응답값이 행정보완값보다 큰 경우는 응답값이 109만 3천 원으로 행정보완값의 약 2배 정도 큰 편이었다. 작은 경우는 응답값이 9만 8천 원으로 행정보완값 88만 6천 원 비해 현저히 낮게 나타났다. 이는 81.9%(1,704가구)가 0원이라고 한 응답에 기인한 것이다.

집단별 대응표본 t-검정을 실시한 결과, 응답값이 행정보완값보다 작은 경우의 평균 차이는 -78만 8천 원이고, 응답값의 표준편차가 행정보완값에 비해 절반 정도 작은 편이며, 유의수준 0.1%에서 응답값과 행정보완값 간에 통계적으로 유의한 차이가 있는 것으로 나타났다. 응답값이 행정보완값보다 큰 경우의 평균 차이는 53만 4천 원이고, 응답값의 표준

편차는 행정보완값과의 차이가 크지 않은 편이며 유의수준 0.1%에서 응답값과 행정보완값은 통계적으로 유의한 차이가 있는 것으로 나타났다.

〈표 4-6〉 근로자녀장려금 3개 집단별 대응 표본 t-검정 결과

	자료	N	평균	표준편차	최솟값	최댓값	t Value
응답값 〈 행정보완값	조사	1,905 (91.5%)	9.8	34.04	0	310	-50.94 (***)
	행정보완		88.6	71.24	1	320	
응답값 = 행정보완값	조사	55 (2.6%)	113.3	76.49	14	310	-
	행정보완						
응답값 〉 행정보완값	조사	121 (5.8%)	109.3	70.7	4	295	9.8 (***)
	행정보완		55.9	65.9	0	294	

주: t Value ( ) 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

3개 집단의 범주를 종속 변수로 하여 다항 로지스틱 회귀모형 분석을 실시하였다. 준거집단은 응답값과 행정보완값이 같은 경우(응답값=행정보완값)이고, 설명변수는 가구주의 연령, 성별, 학력, 종사상지위 형태, 가구원 수, 소득 5분위, 입주 형태이다. 최적의 회귀모형을 선택하고자 후진제거법을 활용하였으며, 〈표 4-7〉은 가구주의 성별, 연령, 가구원 수가 선택된 최종 모형의 분석 결과이다.

(응답값=행정보완값) 경우를 준거집단으로 하여 (응답값〈행정보완값) 경우에 영향을 주는 요인으로는 부채 유무를 제외한 설명변수에서 모두 통계적으로 유의한 결과를 나타내고 있다. 가구주의 성별이 남성(기준변수: 여성)일수록 (응답값〈행정보완값)일 가능성이 (응답값=행정보완값)일 가능성보다 96.9% 증가한다고 볼 수 있다. 가구주의 연령이 한 단위 증가할수록 (응답값〈행정보완값)일 가능성이 1.3% 증가한다고 볼 수 있다.

가구원 수가 한 단위 증가할수록 (응답값〈행정보완값)일 가능성이 33.1% 증가한다고 볼 수 있다.

(응답값=행정보완값) 경우를 준거집단으로 하여 (응답값〉행정보완값)인 경우에 영향을 주는 요인으로는 모든 설명변수가 통계적으로 유의한 결과를 나타내지 않았다.

〈표 4-7〉 근로자녀장려금에 대한 다항 로지스틱 회귀모형 분석 결과

변수	응답값 〈 행정보완값			응답값 〉 행정보완값		
	coefficient	SE	exp(b)	coefficient	SE	exp(b)
상수항	3.030(***)	0.895	-	1.726	1.058	-
가구주의 성별_남	0.678(**)	0.318	1.969	0.228	0.380	1.256
가구원 수	-0.403(***)	0.139	0.669	-0.055	0.166	0.947
가구주의 연령	0.026(**)	0.013	1.027	-0.019	0.015	0.981

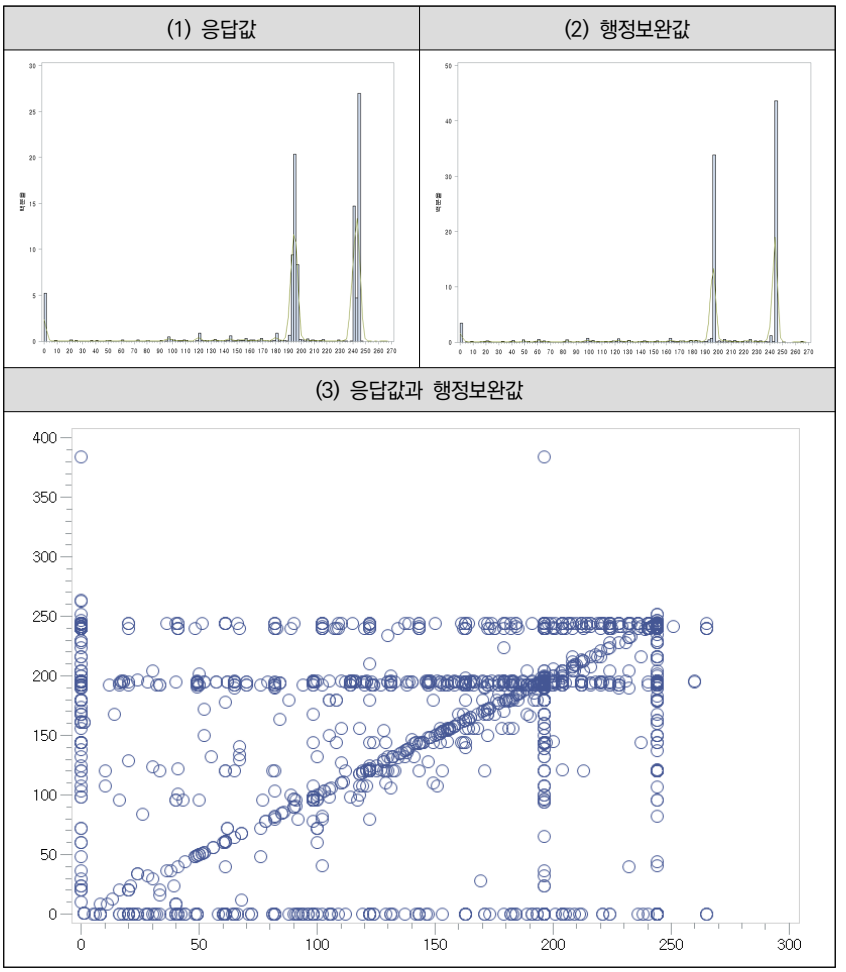
주: t Value ( ) 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

3. 기초연금

분석 대상은 응답값과 행정보완값이 모두 0원인 경우는 제외한 개인 응답자 5,896명이다.

〔그림 4-3〕을 보면 (1) 응답값과 (2) 행정보완값 간의 분포가 비슷한 편으로 보이나, 응답값은 비율이 높은 곳의 앞뒤로 일정한 비율을 가지고 있다. (3) 응답값과 행정보완값 간의 분포를 보면 모두 특정 값인 0원, 200만 원, 250만 원에 집중되어 있다. 그 외 관측치는 주로 과소 보고 영역에 더 많이 분포되어 있다. 근로자녀장려금과 비슷하게 응답자의 회고 오차가 영향을 미치고 있다고 생각한다.

〔그림 4-3〕 기초연금에 대한 응답값과 행정보완값의 분포



주: 1) (1), (2)의 단위는 X축 만원이고, Y축은 %임.
2) (3) X축은 행정보완값이고, Y축은 응답값임(단위: 만 원).
3) (1)~(3)의 전체수는 5,896명임.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

기초연금의 응답값 평균은 201만 3천 원으로 행정보완값(200만 5천 원)과 비슷하였다. 대응표본 t-검정 결과를 보면 유의수준 5%에서 응답값과 행정보완값 간에 통계적으로 유의한 차이가 없는 것으로 나타났다(〈표 4-8〉 참조).

〈표 4-8〉 기초연금 응답값과 행정보완값의 대응표본 t-검정 결과

	N	평균	표준편차	최솟값	최댓값	t Value
응답값	5,896	201.3	60.1	0	384	0.90
행정보완값	5,896	200.5	60.2	0	265	

주: t Value () 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

응답값과 행정보완값 간 차이를 3개 그룹으로 구분해 보면, 응답값을 행정보완값보다 더 작게 응답한 경우가 51.3%로 가장 높았고, 응답값을 더 크게 응답한 비율은 12.4%로 가장 낮은 편이었다. 응답값과 행정보완값이 같은 경우는 36.3%로 나타났다(〈표 4-9〉 참조).

3개 집단별 분포 결과를 보면, 응답값과 행정보완값이 같은 경우의 평균은 224만 6천 원이며, 응답값이 행정보완값보다 큰 경우는 203만 5천 원(응답값 기준)이고 작은 경우는 184만 2천 원(응답값 기준)으로 나타났다.

집단별 대응표본 t-검정을 시행한 바, 응답값이 행정보완값보다 작은 경우의 평균 차이는 -22만 8천 원이고, 응답값의 표준편차가 행정보완값에 비해 큰 편이며 유의수준 0.1%에서 응답값과 행정보완값 간에 통계적으로 유의한 차이가 있는 것으로 나타났다. 응답값이 행정보완값보다 큰 경우의 평균 차이는 100만 원이고, 행정보완값의 표준편차가 응답값에 비해 차이가 큰 편이며 유의수준 0.1%에서 응답값과 행정보완값 간에 통계적으로 유의한 차이가 있는 것으로 나타났다.

〈표 4-9〉 기초연금 3개 집단별 대응 표본 t-검정 결과

	자료	N	평균	표준편차	최솟값	최댓값	t Value
응답값 〈 행정보완값	조사	3,023 (51.3%)	184.2	69.76	0	244	-22.38 (***)
	행정보완		207.0	41.73	5	265	
응답값 = 행정보완값	조사	2,141 (36.3%)	224.6	36.74	1	244	-
	행정보완						
응답값 〉 행정보완값	조사	732 (12.4%)	203.5	46.64	10	384	32.81 (***)
	행정보완		103.5	82.23	0	244	

주: t Value ( ) 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

3개 집단의 범주를 종속 변수로 하여 다항 로지스틱 회귀모형 분석을 시행하였다. 준거집단은 응답값과 행정보완값이 같은 경우(응답값=행정보완값)이고, 설명변수는 가구원의 연령, 성별, 학력, 배우자 유무, 종사상지위이다. 최적의 회귀모형을 선택하고자 후진제거법을 활용하였으며, 〈표 4-10〉은 가구원의 학력과 종사상지위가 제외된 최종 모형의 분석 결과이다.

(응답값=행정보완값) 경우를 준거집단으로 하여 (응답값〈행정보완값) 경우에 영향을 주는 요인으로는 가구원의 연령을 제외한 설명변수에서 모두 통계적으로 유의한 결과를 나타내고 있다. 가구원의 성별이 남성(기준변수: 여성)일수록 (응답값〈행정보완값)일 가능성이 (응답값=행정보완값)일 가능성보다 56.6% 증가한다고 볼 수 있다. 가구원의 배우자가 있을수록 (응답값〈행정보완값)일 가능성이 307% 증가한다고 볼 수 있다.

(응답값=행정보완값) 경우를 준거집단으로 하여 (응답값〉행정보완값) 경우에 영향을 주는 요인으로는 가구원의 성별을 제외한 설명변수에서 모두 통계적으로 유의한 결과를 나타내고 있다. 가구원의 연령이 한 단위 증가할수록 (응답값〉행정보완값)일 가능성이 (응답값=행정보완값)일 가

능성보다 9.1% 증가한다고 볼 수 있다. 가구원의 배우자가 있을수록 (응답값〉행정보완값)일 가능성이 130.6% 증가한다고 볼 수 있다.

〈표 4-10〉 기초연금에 대한 다항 로지스틱 회귀모형 분석 결과

변수	응답값 〈 행정보완값			응답값 〉 행정보완값		
	coefficient	SE	exp(b)	coefficient	SE	exp(b)
상수항	0.320	0.430	-	5.643	0.667	-
가구원의 성별_남	-0.834(***)	0.079	0.434	-0.089	0.106	0.915
가구원의 연령	-0.004	0.006	0.996	-0.095(***)	0.010	0.909
가구원의 배우자 유무_있음	1.404(***)	0.081	4.070	0.836(***)	0.112	2.306

주: t Value ( ) 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

제3절 측정오차 보정을 통한 회귀계수 추정 방안

1. 측정오차 구조

2개 처리효과($X \in 0,1$)를 비교하는 확률화된 실험을 가정하면 X는 처리집단과 대조집단으로 정의한다. 또한 Y는 실제값(true trial end-point)이고, Y^*는 Y의 오차를 포함한 값이며 연속형이라고 하자. Y에 대한 선형회귀모형(linear regression model)을 다음과 같이 가정한다.

$$Y = \alpha_Y + \beta_Y X + \epsilon$$

여기서 $\epsilon \overset{iid}{\sim} N(0,\sigma^2)$이고, Y^*에 대한 모형의 가정으로 처리효과는 보통최소제곱법(Ordinary Least Squares : OLS) 추정량이다.

측정오차의 구조는 4개(전형적인 측정오차, 이분산성 측정오차, 체계적인 측정오차, 차이가 있는 측정오차)로 구분할 수 있다.

가. 전형적인 측정오차(classical measurement error)

Y^*가 Y의 비편향(unbiased) 변수(측정오차 모형 : $Y^* = Y + e$)이면 Y^*는 전형적인 측정오차가 있다고 하며 e는 평균이 0이고 분산이 τ^2이고, Y, X, ϵ와 독립이다. Y^*의 선형모형은 다음과 같다.

$$Y^* = \alpha_Y^* + \beta_Y^* X + \delta$$

여기서 $\beta_Y^* = \beta_Y$이고 잔차(δ)는 평균이 0이고 분산이 $\sigma_\delta^2 = \sigma^2 + \tau^2$이다. $\widehat{\beta_Y}$(β_Y 추정량)의 분산에 비해 $\widehat{\beta_Y^*}$(β_Y^* 추정량)가 더 큰 분산을 가진다. 따라서 전형적인 측정오차는 추정량이 편향을 초래하지 않으나, 표본크기에 따라 2종 오류(type-Ⅱ error)가 증가하게 된다.

나. 이분산성 측정오차(heteroscedastic measurement error)

전형적인 측정오차에서는 e의 분산이 동일하다고 가정하였다. 이 가정이 위배되면 이분산성 측정오차를 가진다고 할 수 있다. 이분산성 측정오차는 추정량이 편향을 초래하지는 않으나, $\widehat{\beta_Y^*}$의 분산 추정량이 유효하지

않을 것이다.

다. 체계적인 측정오차(systematic measurement error)

Y^*가 Y에 대해 체계적이면(측정오차 모형 : $Y^* = \theta_0 + \theta_1 Y + e$) Y^*는 체계적인 측정오차가 있다고 하며 e는 평균이 0이고 분산이 τ^2이고, Y, X, ϵ와 독립이다. $\theta_0 \neq 0$ 또는 $\theta_1 \neq 1$(모든 경우에서 $\theta_1 \neq 0$)이면 체계적인 측정오차라고 가정한다. 또한, e와 Y, X, ϵ는 독립이라고 가정한다. Y^*의 선형모형에서 $\beta_Y^* = \theta_1 \beta_Y$이고 잔차 δ는 평균이 0이고 분산이 $\sigma_\delta^2 = \theta_1^2 \sigma^2 + \tau^2$이다. θ_1에 영향을 받는 $\hat{\beta}_Y^*$의 분산은 $\hat{\beta}_Y$에 비해 더 크거나 작을 것이다. 따라서, 체계적인 측정오차가 있으면 2종 오류는 감소하거나 증가할 것이나, 1종 오류(type- I error)는 $\beta_Y \neq 0$ 이고 $\beta_Y^* \neq 0$이기 때문에 영향을 받지 않을 것이다. 즉, 귀무가설 검정은 체계적인 측정오차하에서는 여전히 유효하다고 볼 수 있다.

라. 차이가 있는 측정오차(differential measurement error)

Y^*가 X에 따라 다른 Y에 대해 체계적이면(측정오차 모형 : $Y^* = \theta_{00} + (\theta_{01} - \theta_{00})X + \theta_{10} Y + (\theta_{11} - \theta_{10})XY + e_X$) Y^*는 차이가 있는 측정오차가 있다고 하며, $X = 0, 1$에 대해 e_X는 평균이 0이고 분산이 τ_X^2이고, Y, ϵ와 독립이다. Y^*의 선형모형에서 $X = 0, 1$에 대해 $\beta_Y^* = \theta_{01} - \theta_{00} + (\theta_{11} - \theta_{10})\alpha_Y + \theta_{11}\beta_Y$이고 잔차 δ는 평균이 0이고 분산이 $\sigma_\delta^2 = [\theta_{10}^2 + (\theta_{11}^2 - \theta_{10}^2)X]\sigma^2 + \tau_X^2$이다. 잔차는 X에 따라 같지 않기

때문에 $\hat{\beta}_Y^*$의 분산 추정량은 유효하지 않으며 실제 분산(true variance)을 과소 추정할 것이다. $\hat{\beta}_Y^*$의 이분산성 일치 추정량(heteroscedastic consistent estimator)은 White 추정량으로 구할 수 있다. White 추정량이 $\hat{\beta}_Y^*$의 분산을 추정하기 위해 사용된다고 가정하면, 1종 오류는 유효하지 않으며 2종 오류도 차이가 있는 측정오차하에서는 감소하거나 증가할 것이다.

2. 측정오차 보정 방법

Y^*는 무작위로 할당된 모든 사람에서 측정한 값이고($i = 1, 2, \cdots, N$), Y와 Y^*는 크기가 더 작은 다른 집단에서 측정한 값을 가진다($j = 1, 2, \cdots, K,\ K < N$)고 가정한다.

연속형 변수에서 전형적인 측정오차의 보정은 표본 크기가 커질수록 감소하는 정확도(precision)를 보완하는 것이다. 예를 들면 새로운 표본(N^*)은 N/R을 계산한다. 여기서 R은 신뢰도(reliability coefficient)이고 N은 본 표본이다.

이분산성 측정오차의 보정은 $\hat{\beta}_{Y^*}$ 분산의 비편향 추정량을 구하기 위하여 회귀모형에서 이분산성 오차를 처리하는 것이 기본적인 이론이다.

Y^*가 체계적인 측정오차 및 차이가 있는 측정오차를 갖는 경우에 대한 보정 방법을 살펴보았다. 이처럼 측정오차를 고려하지 않는다면 회귀추정량은 편향을 가지게 되므로 측정오차에 대한 보정이 필요하다. 다음은 측정오차의 구조에 따라 외부보정(external calibration) 자료를 사용하여 측정오차를 보정하는 방법에 대해 정리하였다.

가. 체계적인 측정오차를 갖는 경우에 대한 보정 방법

α_Y와 β_Y의 추정량은 다음과 같다.

$$\hat{\alpha}_Y = (\widehat{\alpha}_{Y^*} - \hat{\theta}_0)/\hat{\theta}_1 \quad \text{이고} \quad \hat{\beta}_Y = \hat{\beta}_{Y^*}/\hat{\theta}_1$$

여기서 $\hat{\theta}_0$와 $\hat{\theta}_1$은 보정 자료에서 보통최소제곱 회귀를 적합하여 추정된 오차 모수(parameter)이다. $\hat{\theta}_1$은 유한 추정값(finite estimate)인 $\hat{\alpha}_Y$와 $\hat{\beta}_Y$를 위해 0이 아니라고 가정한다. α_Y와 β_Y의 추정량은 일치(consistent)한다. α_Y와 β_Y의 추정량에 대한 분산은 Delta 방법, Fieller 방법, Zero-Variance 방법, 붓스트랩(Bootstrap) 방법을 사용하여 구할 수 있다.

나. 차이가 있는 측정오차를 갖는 경우에 대한 보정 방법

α_Y와 β_Y의 추정량은 다음과 같다.

$$\hat{\alpha}_Y = (\widehat{\alpha}_{Y^*} - \hat{\theta}_{00})/\hat{\theta}_{10} \quad \text{이고} \quad \hat{\beta}_Y = (\hat{\beta}_{Y^*} + \hat{\alpha}_{Y^*} - \hat{\theta}_{01})/\hat{\theta}_{11} - \hat{\alpha}_Y$$

여기서 $\hat{\theta}_{00}$, $\hat{\theta}_{10}$, $\hat{\theta}_{01}$, 그리고 $\hat{\theta}_{11}$은 보정 자료에서 보통최소제곱법을 사용하여 추정한다. $\hat{\theta}_{10}$과 $\widehat{\theta_{11}}$은 유한 추정값(finite estimate)인 $\hat{\alpha}_Y$와 $\hat{\beta}_Y$를 위해 0이 아니라고 가정한다. α_Y와 β_Y의 추정량은 일치(consistent)한다. α_Y와 β_Y의 추정량에 대한 분산은 Delta 방법,

Zero-Variance 방법, 붓스트랩(Bootstrap) 방법을 사용하여 구할 수 있다.

3. mecor 패키지 활용

Nab, van Smeden, Keogh, and Groenwold(2021)는 통계패키지 R의 mecor 패키지를 사용하여 연속형 종속 변수 또는 공변량 변수에 대한 회귀모형에서 측정오차를 보정할 수 있도록 하였다. 측정오차 보정은 측정오차 모형과 측정오차 모형의 모수에 대한 정보가 필요하며, 이를 4가지 방법으로 구할 수 있다. 즉, 내부 타당성(internal validation) 연구, 반복(replicates) 연구, 보정(calibration) 연구, 그리고 외부 타당성(external validation) 연구이다. 연속형 종속 변수에 대한 회귀모형에서 측정오차를 보정할 때 회귀보정방법(regression calibration methods)과 최대우도방법(maximum likelihood method)을 사용한다. 보정된 추정량의 분산 추정은 붓스트랩을 사용하여 폐쇄형(closed form)으로 구한다.

다음에서는 mecor을 사용하여 가계금융복지조사 자료의 근로소득, 근로자녀장려금, 기초연금에 대해 측정오차를 보정한 회귀계수를 추정한다.

4. 측정오차 보정을 통한 회귀계수 추정

가. 근로소득

근로소득의 측정오차 모형은 차이가 있는 측정오차 모형이며, 로그변환한 응답값에 대한 회귀분석 결과는 〈표 4-11〉과 같다. 설명변수는 로

그변환한 행정보완값, 가구주의 성별(기준변수: 여성), 로그변환한 행정보완값과 가구주의 성별 교차항인데, 계수들이 모두 통계적으로 유의하게 나타났으며, 이를 통해 θ_{01}=2.954이고 θ_{11}=0.652를 구할 수 있다.

〈표 4-11〉 차이가 있는 측정오차 모형에 대한 근로소득 회귀분석 결과

	coefficient	SE	coefficient(기호)
상수항	3.372 (***)	0.114	θ_{00}
log(근로소득_행정보완)	0.572 (***)	0.015	θ_{10}
가구주 성별_남성	-0.418 (***)	0.125	$\theta_{01}-\theta_{00}$
log(근로소득_행정보완)× 가구주 성별_남성	0.080 (***)	0.016	$\theta_{11}-\theta_{10}$

주: coefficient ( ) 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

다음은 측정오차 보정을 통한 근로소득 회귀분석 결과이다(〈표 4-12〉 참조). 측정오차 모형에서 포함된 가구주 성별이 설명변수이고 응답값은 종속 변수이다. 보정 전과 보정 후의 계수값 차이를 보면, 상수항은 보정 후가 보정 전보다 -0.102 작았으며, 가구주 성별의 경우 0.023 크게 나타났다. 그리고 보정 전과 보정 후 표준오차 차이를 보면, 상수항은 보정 후가 보정 전보다 0.021 컸으며, 가구주 성별의 경우 역시 0.022 정도 크게 나타났다. 보정 후 계수들이 모두 통계적으로 유의하였다.

한편, 보정 후 계수값은 행정보완자료를 종속 변수로 한 계수값과 정확하게 일치한 일치추정량이다. 앞에서 설명한 $\hat{\alpha}_Y$와 $\hat{\beta}_Y$의 식에 대입해 보면 보정 후 계수값과 일치함을 확인할 수 있다.

〈표 4-12〉 측정오차 보정을 통한 근로소득 회귀분석 결과

		coefficient	SE	SE(zerovar)
보정 후	상수항	7.632 (***)	0.041	0.035
	가구주 성별_남성	0.621 (***)	0.044	0.037
보정 전	상수항	7.734 (***)	0.020	-
	가구주 성별_남성	0.598 (***)	0.022	-
행정보완자료	상수항	7.632 (***)	0.025	-
	가구주 성별_남성	0.621 (***)	0.028	-

주: coefficient () 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

나. 근로자녀장려금

근로자녀장려금의 측정오차 모형은 차이가 있는 측정오차 모형이며, 로그변환한 응답값에 대한 회귀분석 결과는 〈표 4-13〉과 같다. 설명변수는 로그변환한 행정보완값, 배우자 유무(기준변수: 없음), 로그변환한 행정보완값과 배우자 유무 교차항인데, 계수들이 모두 통계적으로 유의하게 나타났으며, 이를 통해 θ_{01}=0.518이고 θ_{11}=0.13을 구할 수 있다.

〈표 4-13〉 차이가 있는 측정오차 모형에 대한 근로자녀장려금 회귀분석 결과

	coefficient	SE	coefficient(기호)
상수항	-0.285 (***)	0.231	θ_{00}
log(근로소득_행정보완)	0.256 (***)	0.059	θ_{10}
배우자 유무_있음	1.200 (***)	0.287	$\theta_{01}-\theta_{00}$
log(근로소득_행정보완)× 배우자 유무_있음	-0.271 (***)	0.071	$\theta_{11}-\theta_{10}$

주: coefficient ( ) 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

다음은 측정오차 보정을 통한 근로자녀장려금 회귀분석 결과이다(〈표 4-14〉 참조). 측정오차 모형에서 포함된 배우자 유무가 설명변수이고 응답값은 종속 변수이다. 보정 전과 보정 후 계수값 차이를 보면, 상수항의 경우 보정 후가 보정 전보다 3.086 컸으며, 배우자 유무의 경우도 0.201 크게 나타났다. 그리고 보정 전과 보정 후 표준오차 차이를 보면, 상수항의 경우 보정 후가 보정 전보다 0.283 컸으며, 배우자 유무의 경우 역시 4.333 만큼 크게 나타났다. 보정 후 계수에서 상수항은 통계적으로 유의하였으나, 배우자 유무는 통계적으로 유의하지 않았는데 이는 표준오차가 큰 값을 가졌기 때문이다.

한편, 보정 후 계수값은 행정보완자료를 종속 변수로 한 계수값과 정확하게 일치한 일치추정량이다.

〈표 4-14〉 측정오차 보정을 통한 근로자녀장려금 회귀분석 결과

		coefficient	SE	SE(zerovar)
보정 후	상수항	3.762 (***)	0.347	0.239
	배우자 유무_있음	0.377	4.412	3.167
보정 전	상수항	0.676 (***)	0.064	-
	배우자 유무_있음	0.176 (**)	0.079	-
행정보완자료	상수항	3.762 (***)	0.043	-
	배우자 유무_있음	0.377 (***)	0.053	-

주: coefficient ( ) 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

다. 기초연금

기초연금의 측정오차 모형은 차이가 있는 측정오차 모형이며, 로그변환한 응답값에 대한 회귀분석 결과는 〈표 4-15〉와 같다. 설명변수는 로그변환한 행정보완값, 가구원 수(기준변수: 그외), 로그변환한 행정보완

값과 가구원 수 교차항인데, 계수들이 모두 통계적으로 유의하게 나타났으며, 이를 통해 θ_{01}=0.274이고 θ_{11}=0.053을 구할 수 있다.

〈표 4-15〉 차이가 있는 측정오차 모형에 대한 기초연금 회귀분석 결과

	coefficient	SE	coefficient(기호)
상수항	4.124 (***)	0.090	θ_{00}
log(근로소득_행정보완)	0.164 (***)	0.018	θ_{10}
가구원 수_단독	0.872 (***)	0.184	$\theta_{01}-\theta_{00}$
log(근로소득_행정보완)× 가구원 수_단독	-0.102 (***)	0.035	$\theta_{11}-\theta_{10}$

주: coefficient ( ) 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

다음은 측정오차 보정을 통한 기초연금 회귀분석 결과이다(〈표 4-16〉 참조). 측정오차 모형에서 포함된 가구원 수가 설명변수이고 응답값은 종속 변수이다. 보정 전과 보정 후 계수값 차이를 보면, 상수항의 경우 보정 후가 보정 전보다 0.094 컸으며, 가구원 수의 경우 0.156 작게 나타났다. 그리고 보정 전과 보정 후 표준오차 차이를 보면, 상수항의 경우 보정 후가 보정 전보다 0.147 컸으며, 가구원 수의 경우 역시 0.58 정도 크게 나타났다. 보정 후 계수에서 상수항은 통계적으로 유의하였으나, 가구원 수는 통계적으로 유의하지 않았는데 이는 표준오차가 큰 값을 가졌기 때문이다.

한편, 보정 후 계수값은 행정보완자료를 종속 변수로 한 계수값과 정확하게 일치한 일치추정량이다.

〈표 4-16〉 측정오차 보정을 통한 기초연금 회귀분석 결과

		coefficient	SE	SE(zerovar)
보정 후	상수항	5.046 (***)	0.165	0.122
	가구원 수_단독	0.214	0.615	0.368
보정 전	상수항	4.952 (***)	0.018	-
	가구원 수_단독	0.370 (***)	0.035	-
행정보완자료	상수항	5.046 (***)	0.016	-
	가구원 수_단독	0.214 (***)	0.030	-

주: coefficient ( ) 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

제4절 히핑 보정 및 비례핫덱대체를 이용한 측정오차 보정 방안

1. 히핑 보정 방안 - 근로소득 사례

제3장에서 소개하였던 히핑 보정 방법을 Kernelheaping 패키지 내부의 dheaping 함수를 사용하여 가계금융복지조사 자료의 근로소득에 시행하였다.

우선 근로소득의 분포가 특정 값의 배수에 많이 치우쳐 있는지 근로소득의 응답값에 대한 비중을 살펴보았으며, 〈표 4-17〉은 응답값 비중이 높은 상위 10개를 나타내고 있다. 3,000만 원이 5.79%로 가장 많이 응답하였으며, 다음으로 2,400만 원(4.67%), 3,600만 원(4.29%), 5,000만 원(4.25%)이 차지하였다. 응답값의 형태는 100의 배수 또는 120의 배수로 볼 수 있다.

〈표 4-17〉 근로소득 응답값 개수 및 비중 - 상위 10개

근로소득(만 원)	개수(개)	비중(%)
1800	234	3.22
2400	340	4.67
3000	421	5.79
3600	312	4.29
4000	287	3.95
4200	160	2.20
5000	309	4.25
6000	281	3.86
7000	232	3.19
8000	185	2.54

자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

〈표 4-18〉은 근로소득 응답값이 100의 배수와 120의 배수 형태를 가지는 비율이다. 100의 배수는 80.1%로 매우 높은 편이고, 120의 배수도 44.2%로 나타났다.

〈표 4-18〉 100 또는 120의 배수인 근로소득의 비율

100의 배수	120의 배수
80.1	44.2

자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

이러한 형태를 자세히 살펴보기 위해서 분석 대상을 5,000만 원 미만으로 제한하였다. 히핑 보정 시 많은 히핑 지점이 있는 경우 베이지안 방법의 사후분포가 수렴하지 않을 수 있기 때문이다. 〈표 4-19〉는 근로소득의 응답값 비중이 높은 상위 10개에 대한 것이다. 3,000만 원이 9.1%로 가장 많이 응답하였으며, 다음으로 2,400만 원(7.35%), 3,600만 원(6.74%), 4,000만 원(6.2%)이 차지하였다.

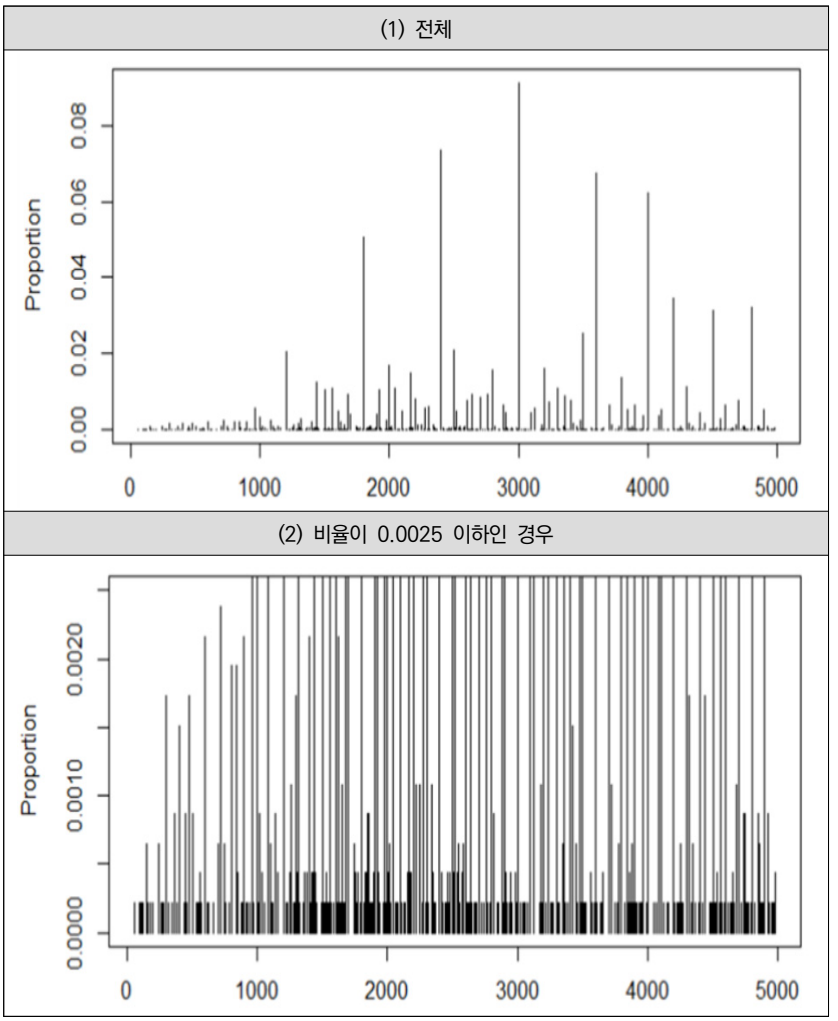
〈표 4-19〉 5,000만 원 미만의 근로소득 응답값 개수 및 비중 - 상위 10개

근로소득(만 원)	개수(개)	비중(%)
1800	234	5.06
2400	340	7.35
2500	97	2.10
3000	421	9.10
3500	118	2.55
3600	312	6.74
4000	287	6.20
4200	160	3.46
4500	144	3.11
4800	148	3.20

자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

〔그림 4-4〕는 응답값에 대한 spike plot을 보여주고 있다. 그림에서 (1) 전체를 보면, 특정 응답값에 분포가 집중되어 비중이 매우 높게 나타나는 곳이 꽤 많은 편으로 나타났다. 특히 3,000만 원은 비율이 높아서 다른 값들의 분포를 자세히 보기 위하여, 참고로 (2) 비율이 0.0025 이하인 경우도 살펴보았다.

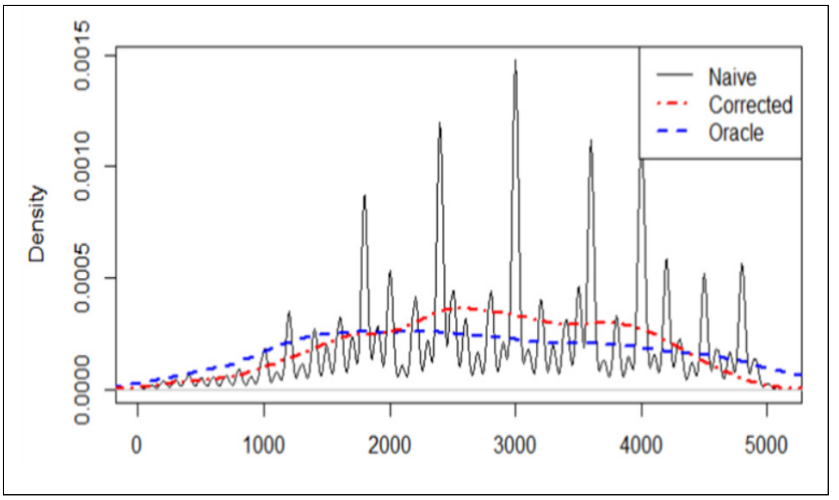
〔그림 4-4〕 근로소득에 대한 spike plot



주: X축은 근로소득(단위: 만 원)이고, Y축은 Proportion(응답값에 대한 비율 %)임.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

근로소득의 히핑 발생 지점을 100의 배수로 하여 커널 함수를 통한 히핑 보정 방법을 실시하였다. 〔그림 4-5〕는 히핑 발생 지점을 기준으로 커널 분포 함수를 보정 전(Naive), 보정 후(Corrected), 행정보완값(Oracle)에 대해 추정한 결과를 나타내고 있다. 보정 전 히핑 지점에는 100의 배수마다 분포가 집중되어 나타나 있는 반면에, 보정 후 분포 함수 추정 결과는 전반적으로 부드러운 분포 함수의 형태를 가졌다. 행정보완값의 분포 함수와 비교해 보면 약 2,000만 원 이하에서는 보정 후 분포 함수가 낮은 편이었고, 약 4,000만 원 정도까지는 높았다가 다시 낮게 추정되었음을 알 수 있다. 행정보완값의 분포 함수와 차이가 있는 편이지만 응답값에 비해 비슷하게 추정되었다고 볼 수 있다.

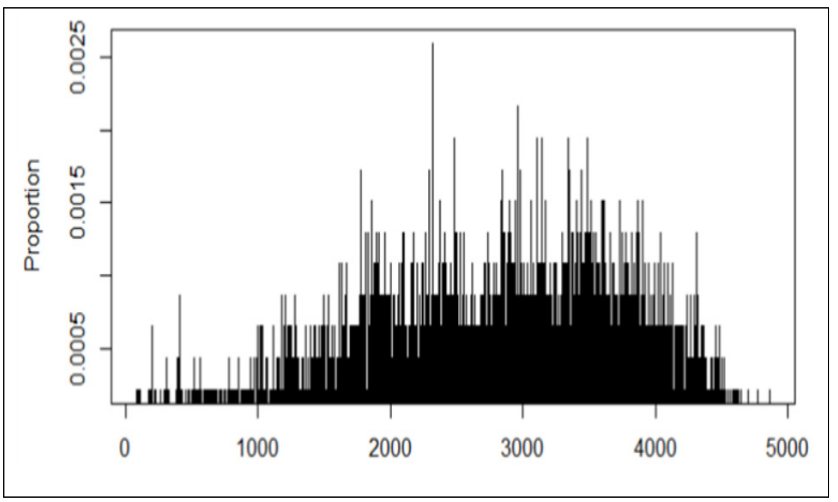
〔그림 4-5〕 커널 함수를 통한 근로소득 히핑 보정 분포 결과



주: X축은 근로소득(단위: 만 원), Y축은 Density(밀도)임.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

〔그림 4-6〕은 히핑 발생 지점을 기준으로 커널 분포 함수를 보정한 spike plot을 보여주고 있다. 〔그림 4-5〕와 다르게 보정 후 spike plot은 부드러운 분포의 형태로 나타났다.

〔그림 4-6〕 커널 함수를 통한 근로소득 히핑 보정 spike plot



주: X축은 근로소득(단위: 만 원)이고, Y축은 Proportion(응답값에 대한 비율 %)임.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

〈표 4-20〉은 히핑 보정 전·후 행정보완값의 근로소득 평균과 표준편차에 대한 것이다. 보정 후 평균은 2,806만 2천 원으로 보정 전(응답값)에 비해 작았고, 행정보완값보다도 작았으나 그 차이는 보정 전보다 작은 것으로 나타났다. 보정 후 표준편차는 933만 1천 원으로 보정 전과 행정보완값에 비해 작았다.

〈표 4-20〉 히핑 보정 전·후, 행정보완값의 근로소득 평균과 표준편차

	평균	표준편차
보정 후	2806.2	933.1
보정 전(응답값)	2940.5	1087.2
행정보완값	2858.9	1546.5

자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

2. 비례핫덱대체를 이용한 측정오차 보정 방안 – 근로자녀장려금 사례

가계금융복지조사 자료에서 근로자녀장려금의 경우 행정보완값에서는 0 초과의 값을 가지나, 응답값에서는 0원(받지 않음)이라고 81.9%가 응답하여 아주 높은 비율로 나타났다([그림 4-2] 참조). 0원을 무응답이라고 간주하여 적절한 값으로 대체하는 방법으로 측정오차를 보정하였다. 대체 방법은 여러 가지가 있으나 비례핫덱대체(Fractional Hot Deck Imputation: FHDI, 이하 'FHDI')방법을 활용하였고, 대체된 결과는 회귀대체(Regression Imputation) 방법으로 대체된 자료와도 함께 비교하여 비례핫덱대체 방법을 살펴본 후 실제 자료에 적용하였다.

가. 비례핫덱대체에 대한 이상샘플링 접근법 문헌 연구[2)]

핫덱대체(hot deck imputation)는 조사 자료에서 항목 무응답(item nonresponse)을 적절한 값으로 대체할 때 유용하게 활용되는 방법이다. 대체값은 같은 대체 칸(cell)에 있는 응답값으로 가져오며, 대체 칸은 대체 모형(model)을 계산하여 구할 수 있다. Kim and Fuller(2004)의 FHDI를 확장하여 사전에 대체 칸 정보가 필요하지 않은 경우이다. 제안

2) Im, Kim, and Fuller(2015)의 내용을 요약 발췌하였음.

된 FHDI 방법은 2단계로 실시되며, 2상 계통추출(two-phase systematic sampling)을 사용한 비모수 대체(nonparametric imputation) 접근 방법이라고 볼 수 있다.

1) 기본 설정 : 일변량 결측인 경우

N개의 유한모집단이 있다고 가정하자($U=(1, 2, \cdots, N)$). 확률추출메커니즘에 기반하여 선택된 표본 단위 집단을 A라고 정의하자. A는 보조정보 x를 사용하여 G개 그룹으로 나누며, x는 $(1, 2, \cdots, G)$에서 값을 가지며 $A=A_1 \cup \cdots \cup A_G$이다. y는 관심 변수이고, z는 $(1, 2, \cdots, H)$에서 값을 가지는 또 다른 범주형 변수이다. x와 z의 교차분류는 대체 칸을 구성하며 다음과 같이 가정한다.

$$y_i | (x_i = g, z_i = h) \sim ii(\mu_{gh}, \sigma^2_{gh}), \quad i \in U \qquad (1)$$

여기서, μ_{gh}와 $\sigma^2_{gh} > 0$이고 $\sim ii$는 독립이고 동일한 분포를 나타낸다. x_i는 항상 관측된 것이나, (y_i, z_i)는 결측을 가진다. (y_i, z_i)가 관측되면 $\delta_i = 1$이고, 아니면 $\delta_i = 0$이다. 단위 응답으로부터 A는 $A = A_M \cup A_R$인 $A_R = \{j \in A; \delta_j = 1\}$, $A_M = \{j \in A; \delta_j = 0\}$으로 다시 나눌 수 있다. 또한, A_g는 $A_{Rg} = \{j \in A_g; \delta_j = 1\}$, $A_{Mg} = \{j \in A_g; \delta_j = 0\}$로 나눈다. n_{Rg}와 n_{Mg}는 A_{Rg}와 A_{Mg}의 크기이다.

δ는 x가 주어진 경우 (y, z)에서 조건적으로 독립이라는 점에서 응답메커니즘은 MAR(Missing at random)이라고 가정한다. 즉,

$$f(y,z|x,\delta)=f(y,z|x) \qquad (2)$$

이다. MAR 조건 (2)는 수식 (1)이 응답단위에서 유지된다. 즉,

$$y_i|(x_i=g,z_i=h,\delta_i=1)\sim ii(\mu_{gh},\sigma_{gh}^2) \qquad (3)$$

무응답하에서 $Y_N=\sum_{i=1}^{N}y_i$의 핫덱 대체 추정량은 조건 (2)에 따라

$$f(y|x,\delta=1)=\sum_{h=1}^{H}P(z=h|x,\delta=1)f(y|x,z=h,\delta=1) \qquad (4)$$

이다. 수식 (4)는 유한혼합모형(finite mixture model) 형태를 가지며 $\pi_{h|g}=P(z=h|x=g,\delta=1)$라고 하자. (x,z)는 핫덱 대체를 위한 대체 칸이라고 정의하면 수식 (4)로부터

$$E(y_i|x_i=g,\delta_i=1)=\sum_{h=1}^{H}\pi_{h|g}E(y_i|x_i=g,z_i=h,\delta_i=1)$$

이다. 따라서 만약 $\pi_{h|g}$를 알고 있다면 아래 수식 (5)를 얻기 위해 $E(y_i|x_i=g,z_i=h)$를 추정에서 그 칸의 모든 응답자를 사용한다.

$$\widehat{Y_{FEFI}}=\sum_{g=1}^{G}\sum_{i\in A_g}w_i\left\{\delta_i y_i+(1-\delta_i)\sum_{h=1}^{H}\pi_{h|g}\widehat{\mu_{gh}}\right\} \qquad (5)$$

여기서 $\widehat{\mu_{gh}}=\dfrac{\sum_{j\in A} w_j\delta_j a_{jgh}y_j}{\sum_{j\in A} w_j\delta_j a_{jgh}}$이고, $x_j=g$와 $z_{j=h}$이면 $a_{jgh}=1$이고, 그 외는 $a_{jgh}=0$이다. 수식 (5)의 추정량은 대체 칸에 기증자(donors)로 모든 응답값을 사용하고, 완전 효율적인 비례대체(Fully Efficient Fractional Imputation: FEFI) 추정량이다(Kim & Fuller, 2004).

2) FHDI : 일변량 결측인 경우

확장된 FHDI는 수식 (4)에서 유한혼합모형이 주어졌을 때 대체값은 $\pi_{h|g}$에 비례하는 확률을 가진 대체 칸에서 가져온다. $\pi_{h|g}$는 모르므로 추정해야 한다.

제안한 FHDI는 층화를 위한 2단계 추출과 유사한 방법이다. 1단계에서 칸이 결정되고 $\pi_{h|g}$가 추정된다. 2단계에서 M개 기증자는 각 대체 칸에서 선택된다.

$\pi_{h|g}$는 추정되고, 각 집단 g에 대해 $\sum_{h=1}^{H}\widehat{\pi_{h|g}}=1$이다. 정의 $\pi_{h|g}=\Pr(z_i=h|x_i=g,\delta_i=1)$를 사용하여 $\hat{\pi}_{h|g}$의 추정량은

$$\widehat{\pi_{h|g}}=\frac{\sum_{j\in A} w_j\delta_j a_{jgh}}{\sum_{j\in A} w_j\delta_j a_{jg}} \qquad (6)$$

이다. 여기서 $a_{jg}=\sum_{h=1}^{H} a_{jgh}$이다. 따라서, 수식 (5)에서의 FEFI 추정량은

다음과 같이 다시 표현할 수 있다.

$$\hat{Y}_{FEFI} = \sum_{g=1}^{G} \sum_{i \in A_g} w_i \left\{ \delta_i y_i + (1-\delta_i) \sum_{h=1}^{H} \widehat{\pi_{h \mid g}} \widehat{\mu_{gh}} \right\}$$

$$= \sum_{g=1}^{G} \sum_{i \in A_g} w_i \left\{ \delta_i y_i + (1-\delta_i) \sum_{j \in A} w^*_{ij,\,FEFI} y_j \right\} \quad (7)$$

여기서 $w^*_{ij,\,FEFI} = \sum_{h=1}^{H} \widehat{\pi_{h|g}} \left\{ w_j \delta_j a_{jgh} / \sum_{l \in A} \delta_l w_l a_{lgh} \right\}$는 i번째 수령자(recipient)에 대한 j번째 기증자의 비례 가중치(fractional weights)이다. 각 수령자는 M개의 대체값이 선택되었고, Y_N의 2단계 비례대체(FI) 추정량은 다음과 같이 정의한다.

$$\hat{Y}_{FI} = \sum_{g=1}^{G} \sum_{i \in A_g} w_i \left\{ \delta_i y_i + (1-\delta_i) \sum_{h=1}^{H} \widehat{\pi_{h \mid g}} \overline{y_i^*} \right\}$$

$$= \sum_{g=1}^{G} \sum_{i \in A_g} w_i \left\{ \delta_i y_i + (1-\delta_i) \sum_{j=1}^{M} w^*_{ij} y_i^{*(j)} \right\} \quad (8)$$

여기서, $y_i^{*(j)}$는 y_i의 j번째 대체값이며 $\overline{y_i^*} = M^{-1} \sum_{j=1}^{M} y_i^{*(j)}$은 대체값의 평균이고, w^*_{ij}는 FI 추정량에 대한 비례 가중치이다.

FI 추정량은 FEFI 추정량으로 다음과 같이 표현할 수 있다.

$$\hat{Y}_{FI} = \hat{Y}_{FEFI} + (\hat{Y}_{FI} - \hat{Y}_{FEFI})$$

$$= \hat{Y}_{FEFI} + \sum_{g=1}^{G} \sum_{i \in A_{Mg}} w_i (\overline{y_i^*} - \widehat{\mu_g}) \quad (9)$$

여기서 $\widehat{\mu_g}=\sum_{j\in A_{Rg}} w_j y_j / \sum_{j\in A_{Rg}} w_j$이다. $\hat{Y}_{FEFI}$ 추정량은 기증자의 선택에 대한 분산이 없다.

계통적 크기비례확률(Probability Proportional to Size: PPS) 추출을 사용하여 FEFI의 추정량을 얻는 것이다. 즉, 수령자 $i\in A_{Mg}$는 비례가중치($w^*_{ij,\,FEFI}$)를 가지는 n_{Rg} FEFI 기증자가 있다. 따라서 y_i에 대한 M개의 대체값은 $w^*_{ij,\,FEFI}$에 비례하는 확률을 가지는 A_{Rg}에 있는 기증자를 체계적으로 선택하여 구할 수 있다. 이 절차의 효율성은 기증자를 선택하는 추출설계의 효율성에 의존한다. FHDI와 FEFI의 분산 추정은 잭나이프(jacknife) 방법으로 구한다.

이 방법은 다변량 결측인 경우로도 확장 가능하며 자세한 내용은 논문을 통해 확인할 수 있다.

나. FHDI 패키지 활용

Im, Cho, and Kim(2018)은 앞에서 소개한 FHDI 방법을 활용할 수 있도록 통계패키지 R의 FHDI 패키지를 개발하였다. 즉, 임의의 결측패턴을 가진 다변량 결측 자료를 FHDI 방법뿐만 아니라 FEFI 방법(Fuller and Kim, 2005)에 기반하여 적절한 값으로 대체하는 것이다. FHDI 패키지에는 3가지 주요 함수로, 대체 칸으로 사용할 수 있도록 연속 자료를 범주형 자료로 변환해 주는 함수, EM(Expectation Maximization) 알고리듬을 사용하여 대체 칸의 확률을 추정해 주는 함수, 결측값을 FHDI 방법에 기반하여 대체하고 분산 추정을 위한 반복된 비례 가중치 세트(set)를 제공해 주는 함수가 있다.

다음은 가계금융복지조사 자료의 근로자녀장려금에서 0원인 응답값을

FHDI 방법을 사용하여 대체를 실시하였다.

다. 비례핫덱대체를 이용한 측정오차 보정 방안

먼저 근로자녀장려금의 응답값이 0원에 대해 무응답으로 간주하기 위하여 결측(missing) 처리하였다. 결측 처리한 자료가 MAR 가정을 만족하는지 확인하기 위해서 로지스틱 분석을 시행하였다. 종속 변수는 근로자녀장려금이 0원인 경우(결측 처리) 1이고, 그 외는 0이며 설명변수는 가구원 수, 가구주의 배우자 유무, 가구주의 연령집단을 사용하였다. 모수 추정결과는 〈표 4-21〉에서 보듯이 모두 통계적으로 유의하게 나타났다. 이는 가구원 수, 배우자 유무, 연령집단은 근로자녀장려금의 무응답에 영향을 미친다고 볼 수 있다.

〈표 4-21〉 MAR 가정 확인

	coefficient	SE
상수항	0.778 (***)	0.045
가구원 수	-0.072 (***)	0.008
배우자 유무	0.086 (***)	0.020
연령집단	0.086 (***)	0.012

주: coefficient ( ) 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

다음은 근로자녀장려금의 결측 여부와 가구원 수, 배우자 유무, 연령집단에 대한 독립성 검정을 시행하였다. 〈표 4-22〉~〈표 4-24〉를 보면 근로자녀장려금의 결측 여부와 가구원 수는 유의수준 0.001에서, 배우자 유무는 유의수준 0.05에서, 연령집단은 유의수준 0.001에서 통계적으로 관련이 있다고 볼 수 있다.

〈표 4-22〉 결측 여부와 가구원 수에 대한 독립성 검정

(단위: 가구)

가구원 수	결측 여부	
	없음	있음
1인	9	287
2인	49	484
3인	103	406
4인	143	370
5인	73	157
chi-squared=137.21, p-value〈2.2e-16 (***)		

주: coefficient ( ) 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

〈표 4-23〉 결측 여부와 배우자 유무에 대한 독립성 검정

(단위: 가구)

가구주의 배우자 유무	결측 여부	
	없음	있음
없음	108	599
있음	269	1,105
chi-squared=5.824, p-value=0.01581 (**)		

주: coefficient ( ) 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

〈표 4-24〉 결측 여부와 연령집단에 대한 독립성 검정

(단위: 가구)

가구주의 연령집단	결측 여부	
	없음	있음
40세 미만	96	258
40세~55세 미만	228	539
55세 이상	53	907
chi-squared=191.7, p-value〈2.2e-16 (***)		

주: coefficient () 안의 값은 p-value에 대한 유의성을 나타내며 (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

앞에서 MAR 가정을 확인하여 근로자녀장려금은 MAR 가정하에서 무응답 대체를 시행하였다. FHDI 방법으로 대체하였고, 다른 대체 방법과의 비교를 위해서 회귀대체 방법도 추가로 시행하였다. FHDI 방법은 대체군을 형성할 때 행정보완값 활용 여부에 따라 2개의 대체값을 생성하였다. 참고로, 대체 방법에 대한 표기를 FHDI 방법에서 대체군 활용 변수가 4개인 경우 'FHDI_1', 3개인 경우 'FHDI_2', 회귀대체 방법은 '회귀대체'라고 정의하였다. 각 대체 방법에 대한 설명은 다음과 같다.

■ FHDI_1 방법

사용한 대체군 활용 변수 4개는 행정보완값, 가구원 수, 가구주의 배우자 유무, 가구주의 연령집단으로 구성하였다. 칸을 생성하는 방법은 'merging'이고 대체값은 20개를 생성하였다.

■ FHDI_2 방법

사용한 대체군 활용 변수 3개는 행정보완값을 제외한 나머지 변수인 가구원 수, 가구주의 배우자 유무, 가구주의 연령집단으로 구성하였다. FHDI_1과 동일한 조건으로 칸을 생성하는 방법은 'merging'이고 대체값(M)은 20개를 생성하였다.

■ 회귀대체 방법

종속 변수는 로그변환한 응답값이며, 설명변수는 로그변환한 행정보완값, 가구원 수, 가구주의 배우자 유무, 가구주의 연령집단으로 구성하여 회귀분석을 실시하였다. 추정된 회귀계수로 무응답에 대한 대체값을 생성하였다. 최종 대체값은 지수변환하여 본래 단위로 맞춰주었다.

〈표 4-25〉는 근로자녀장려금에 대한 대체 결과로, FHDI 방법의 대체군 활용이 4개인 경우(FHDI_1) 평균은 85만 9천원으로 나타났다. 3개의 대체 방법 중에서 대체값 평균과 행정보완값 평균과의 차이가 가장 작게 나타났다(FHDI_1: -1만 5천 원, FHDI_2: 8만 1천 원, 회귀대체: -8만 1천 원). 표준편차는 대체 방법 중에서 FHDI_1가 45만 1천 원으로 가장 크고, 다음으로 회귀대체(34만 5천 원)와 FHDI_2(30만 7천 원)가 차지하였다. 행정보완값의 표준편차(71만 6천 원)에 비해 꽤 작은 편이었다.

〈표 4-25〉 근로자녀장려금에 대한 대체 결과

(단위: 만 원)

	평균	표준편차	최소값	1사분위	중앙값	3사분위	최대값
행정 보완값	87.4	71.6	0.0	32.0	68.0	132.0	320.0
응답값	18.3	47.8	0.0	0.0	0.0	0.0	310.0
FHDI_1	85.9	45.1	4.0	60.5	67.5	130.5	310.0
FHDI_2	95.5	30.7	4.0	80.0	93.1	100.0	310.0
회귀 대체	79.3	34.5	4.0	58.3	76.2	91.2	310.0

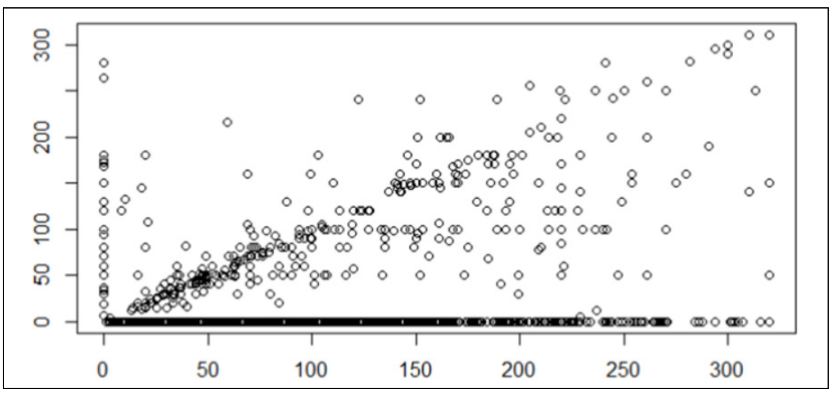
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

〔그림 4-7〕부터 〔그림 4-10〕은 응답값, 대체값과 행정보완값의 분포를 나타내고 있다. FHDI_1에 대한 대체값과 행정보완값의 분포를 보면 대체군 활용 변수로 행정보완값을 활용하였다. 그래서 행정보완값의 구간에 따라 몇 개의 집단으로 대체값이 분포되는 양상을 보였다(〔그림 4-8〕 참조). 이에 반해 FHDI_2에 대한 대체값과 행정보완값의 분포를 보면 100만 원 근처 값으로 대체값이 분포되어 있다(〔그림 4-9〕 참조). 회귀대체값과 행정보완값의 분포를 보면 행정보완값이 100만 원 이하인 경우 대체값은 50~100만 원을 가졌다. 행정보완값이 100만 원 이상인 경우에는 대체값이 주로 100만 원 근처 값으로 대체되어 행정보완값 보다 작은 값으로 대체된 것을 알 수 있다(〔그림 4-10〕 참조). 이러한 결과는 무

응답 비율이 81.9%로 매우 높은 편이어서 대체 시 사용할 수 있는 응답값이 많은 편이 아니었을 뿐만 아니라 행정보완값과 차이가 있기 때문으로 볼 수 있다.

[그림 4-7] 근로자녀장려금의 응답값과 행정보완값 분포

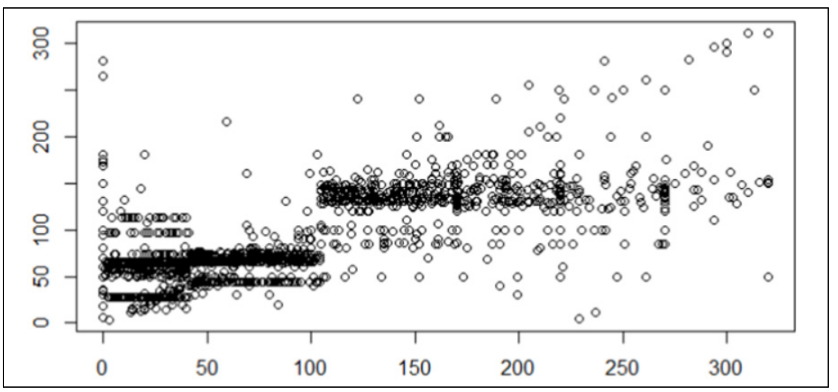
(단위: 만 원)



주: X축은 행정보완값이고 Y축은 응답값임.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

[그림 4-8] 근로자녀장려금의 FHDI_1에 대한 대체값과 행정보완값 분포

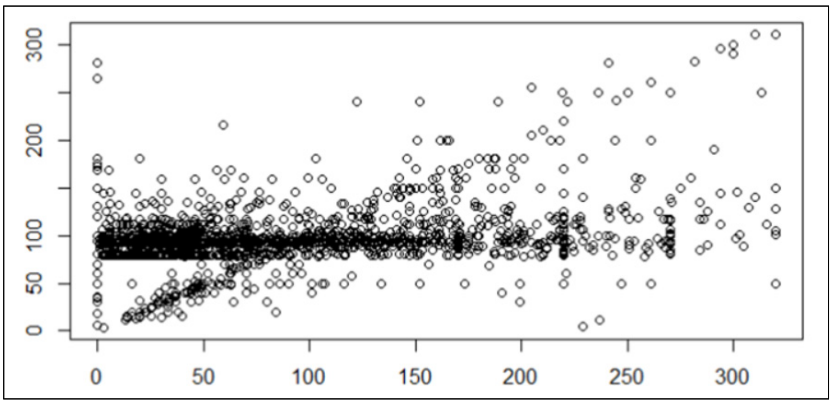
(단위: 만 원)



주: X축은 행정보완값이고 Y축은 FHDI_1에 대한 대체값임.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

〔그림 4-9〕 근로자녀장려금의 FHDI_2에 대한 대체값과 행정보완값 분포

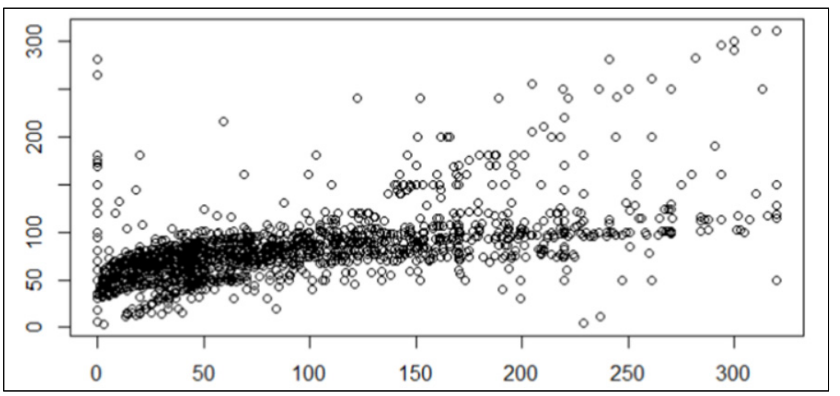
(단위: 만 원)



주: X축은 행정보완값이고 Y축은 FHDI_2에 대한 대체값임.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

〔그림 4-10〕 근로자녀장려금의 회귀대체값과 행정보완값 분포

(단위: 만 원)



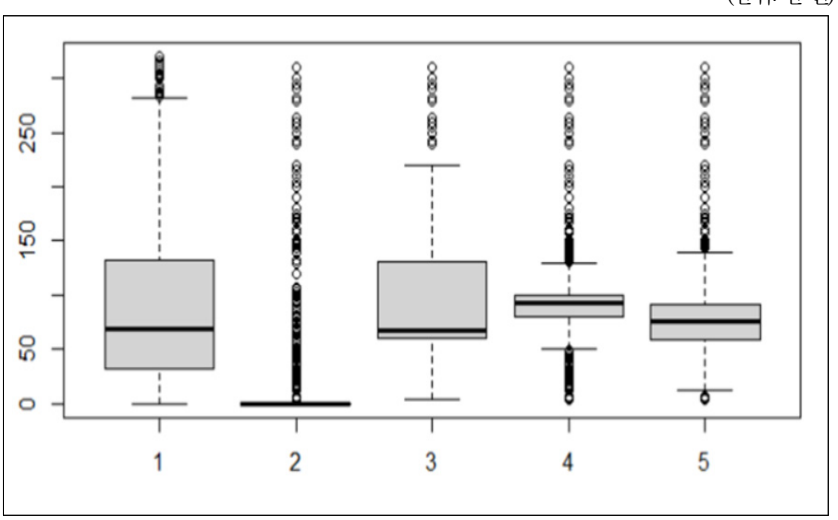
주: X축은 행정보완값이고 Y축은 FHDI_1에 대한 대체값임.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

〔그림 4-11〕은 상자그림(box plot)을 통해 대체 방법별로 대체된 자료의 분포를 살펴보았다. 행정보완값의 분포와 비교했을 때 대체 방법 중에서 FHDI_1이 유사한 편이었다. 그러나 1사분위수와 중앙값 이하가 매

우 짧게 나타났는데, 이는 해당 구간에 포함되는 자료가 없기 때문으로 볼 수 있다. 이는 〈표 4-25〉에서도 알 수 있듯이 1사분위수가 60만 5천 원으로 행정보완값(32만 원)의 2배 정도이었다. 3사분위수는 비슷한 값을 가졌다. FHDI_2는 대체 방법 중에서 구간도 가장 짧으며 사분위수 범위(Inter Quartile Range: IQR[3]) 또한 매우 짧고, 이상치도 많은 편으로 나타났다. 회귀대체는 FHDI_2에 비해 구간 및 사분위수 범위가 긴 편이나 이상치가 많은 편에 속하였다.

[그림 4-11] 근로자녀장려금 상자그림

(단위: 만 원)



주: X축의 범주는 1=행정보완값, 2=응답값, 3=FHDI_1, 4=FHDI_2, 5=회귀대체를 의미하고, Y축은 근로자녀장려금임.
자료: 통계청. (2018). 2017년 가계금융·복지조사의 원자료를 분석함.

3) 3사분위수와 1사분위수 간의 차이를 의미함.

제5절 소결

이 장에서는 2017년 가계금융복지조사 자료에서 가구주의 개인 근로소득, 가구의 근로·자녀장려금, 가구원의 기초연금에 대해 살펴보았다. 이렇듯 관심 변수의 형태(type)는 연속형이고, 2017년 가계금융복지조사 자료에는 응답값과 행정보완값을 모두 제공하고 있다는 점을 활용하였다. 먼저 응답값에 측정오차가 있는지를 파악하였고, 그런 다음 다양한 방법을 활용하여 측정오차를 보정하였다.

관심 변수에 대한 현황 분석을 통해 측정오차를 포함하고 있는지 살펴보았다. 그림(plot)을 통해 응답값과 행정보완값의 분포 파악, 대응 표본 t-검정 시행, 다항 로지스틱 회귀모형 등을 분석하였다. 또한 히핑 현상을 살펴볼 때는 관심 변수값에 대한 상위 응답 비중, 특정 배수의 형태 유무를 분석하여 spike plot을 통해 확인하는 과정을 가졌다. 측정오차 보정 방법으로는 측정오차 보정을 통한 회귀계수 추정, 커널 함수 추정법을 활용한 히핑 보정, 비례핫덱대체(FHDI) 방법을 활용한 보정 등이 해당하고, 이때 사용한 주요 분석 함수는 다음과 같다. 측정오차 보정을 통한 회귀계수 추정은 R 통계패키지에 'mecor' 패키지 내부의 'mecor' 함수를 사용하였다. 커널 함수 추정법을 활용한 히핑 보정은 R 통계패키지에 'Kernelheaping' 패키지 내부의 'dheaping' 함수를 활용하였다. 비례핫덱대체 방법을 활용한 보정은 R 통계패키지에 'FHDI' 패키지 내부의 'FHDI' 함수를 사용하였다.

측정오차 보정 결과는 다음과 같다. 근로소득, 근로자녀장려금과 기초연금의 경우 측정오차의 구조는 차이가 있는 측정오차(differential measurement error)였다. 차이가 있는 측정오차는 응답값이 관심 설명변수에 따라 행정보완값에 대해 체계적인 경우를 의미하며, 이에 대한 측

정오차 보정 방법을 사용하여 회귀계수를 추정하였다. 관심 설명변수의 회귀계수는 모두 행정보완값의 회귀계수와 정확하게 일치하는 것을 확인하였다. 그러나 대부분의 보통 자료에서는 참값을 알고 있는 경우가 드물어서 일반 연구자는 참값을 모르고 있을 가능성이 크므로 이러한 측면에서 봤을 때 유용한 방법이라고 볼 수 있다. 표준오차는 응답값에 따라 크기가 다르게 나타났는데, 이로 인해 근로자녀장려금과 기초연금은 보정 후 회귀계수의 유의성이 보정 전과 다른 결과를 나타냈다. 차이가 있는 측정오차 보정을 통한 회귀계수 추정 분석의 한계점은 설명변수 형태(type)를 이분형 범주만 사용 가능하다는 점, 측정오차에 영향을 미치는 설명변수에 대한 회귀모형 분석만 가능하여 더 많은 관심 설명변수의 영향을 파악할 수 없다는 점을 들 수 있다. 그러나 여러 설명변수가 함께 오차 없는 측정값과 연관되는 것을 현실적으로 모형화하기 어렵기 때문에, 본 연구와 같이 하나의 변수에 국한하는 경향이 있다고 볼 수 있다.

한편 측정오차 보정을 통한 회귀계수 추정 시 내부보정방법을 사용하였는데, 이때 2017년 가계금융복지조사 자료는 응답자마다 응답값과 행정보완값을 모두 가지고 있어서 측정오차 보정 시 모두 활용하였다. 보통 한정적인 예산과 실측 자료 수집의 어려움 등으로 이와 같은 형태의 자료를 사용하기는 쉽지 않은 편이다. 그러나 실측 자료(참값, 실측값 등)의 일부만 가지고도 측정오차를 보정할 수 있다. Nab, Groenwold, Welsing, and van Smeden(2019)의 논문에서는 보정 표본(calibration sample)의 크기, 측정오차가 있는 값과 없는 값 간의 상관관계 정도에 따라 측정오차의 보정 효과에 대한 모의실험을 통해 다음과 같은 결과를 도출하였다. 약한 상관관계(R^2=0.2)를 가지는 경우, 외부 보정 표본의 크기를 50개까지 늘리기 전까지는 측정오차의 보정 결과가 효과적이지 않았다. 그러나 보통 이상의 상관관계(R^2=0.5 또는 0.8)인

경우, 외부보정 표본의 크기가 15개로 적은 편이어도 회귀계수 보정 결과가 향상됨을 보였다. 이렇듯 측정오차를 판단할 수 있는 실측 자료(참값, 실측값 등)를 전체가 아닌 일부라도 구축하고 계속해서 측정오차 정도를 모니터링하는 방안을 고려해 볼 수 있다고 생각한다.

다음으로 근로소득에 대해 커널 함수 추정법을 활용하여 히핑 보정을 하였는데, 히핑 보정 후 결과는 전반적으로 부드러운 분포 함수 형태를 보였다. 히핑 보정 후 평균도 행정보완값과의 차이가 작았다. 반면에 히핑 보정 전 평균은 행정보완값과의 차이가 더 큰 편이었다. 그러나 히핑 보정 후 표준편차가 행정보완값 보다 작게 나타나, 내부보정 또는 외부보정 자료를 추가 정보로 사용한 히핑 보정을 고려해 볼 수 있을 것이다.

마지막으로 근로자녀장려금의 응답값이 0원인 경우에 대해 0원을 무응답으로 간주한 다음 비례핫덱대체 방법을 사용하여 적절한 값으로 대체하였다. 이때 근로자녀장려금의 무응답 비율은 81.9%로 아주 높은 편이었다. 대체 시 활용한 대체 방법은 3가지로 비례핫덱대체 방법에서 대체군 활용 변수가 3개, 4개인 경우와 회귀대체 방법이었다. 무응답 비율이 매우 높은 편이라는 점에서 대체 시 활용할 수 있는 자료(응답값 정보)가 적어 전반적으로 모든 대체 방법의 효과에 영향을 주었다. 그러나 대체 방법 중에서 비례핫덱대체 방법의 대체군 활용 변수가 4개인 경우의 대체 효과가 우수한 결과를 가지는 것으로 나타났다. 이는 무응답 변수와 연관성이 높은 행정보완값을 대체군 활용 변수로 사용하였고, 대체값을 여러 개 생성하여 최종 하나의 값으로 산출하여 대체값에 대한 변동 부분을 고려하였기 때문이라고 볼 수 있다. 회귀대체 방법도 행정보완값을 설명변수로 사용하였으나, 최적의 회귀모형 적합이 되지 않은 점을 원인으로 볼 수 있다. 대체 효과를 높이기 위해서는 다음과 같은 방법을 고려해 볼 수 있다. 현재 가계금융복지조사 자료에서는 근로자녀장려금을 하나

의 변수로 조사하여 근로, 자녀장려금을 구분하기 어려우나 이를 구분할 수 있고, 더 다양한 관련 정보(자녀 수, 소득 관련 변수)가 있다면 논리적 대체 방법을 활용하여 더욱 정교하게 대체할 수 있다고 생각한다. 또한, 근로자녀장려금과 같이 일 년에 한 번 지원하는 제도에 대한 설문 문항은 응답자의 회고오차가 발생할 가능성이 크므로 행정 자료를 연계하여 제공하는 방안을 고려해 볼 수 있다.

제5장

결론

제 5 장 결론

이 연구는 조사자료의 품질검증 연구로 표본조사 자료에서 발생하는 측정오차를 중심으로 살펴보았다. 이 장에서는 주요 연구 결과를 정리하고, 실제 조사자료에서 활용할 수 있는 방안과 측정오차를 포함한 관련 오차를 관리하는 방안에 대해 제안하고자 한다.

2장의 주요 연구 결과를 보면 해외 주요 패널조사(미국 PSID, SIPP, 유럽 SHARE, 영국 BHPS)의 조사 품질 향상 방안 및 측정오차 처리 방법 사례를 조사하였다. 측정오차로 인한 편향을 줄이기 위해 사후적인 보정에 앞서 근본적으로 조사 설계 및 진행, 조사참여자의 응답, 데이터 처리 과정에서의 발생 가능성을 낮추려는 노력으로 다양한 기술적, 시스템적 접근이 이루어지고 있었다. 기술적 접근으로는 컴퓨터를 이용한 전화조사(CATI) 또는 컴퓨터를 이용한 대면조사(CAPI), 이전 조사의 응답 결과를 바탕으로 공통된 내용에 대한 설문을 대체하는 방식인 종속형 설문(Dependent Interviewing: DI), 응답자의 회고오차 발생을 억제하는 데 도움을 주는 Event History Calendar(EHC) 등의 방식을 도입하고 있었다. 시스템적 접근으로는 응답자의 응답 부담을 감소하고 적극적인 참여를 유도하기 위해 별도의 인센티브(사례비) 제공, 조사응답률 및 응답자의 설문 이해도를 높이기 위해 조사 시작 전에 별도의 우편물을 발송하여 조사원에 대한 체계적인 교육 프로그램 구축하고 감독관의 감독·평가를 의무화하고 있었다. 또한 조사원이 체감하는 애로사항을 파악하여 시스템 개선에 반영하기 위하여 정기적인 조사원 대상 설문조사 등을 실

시하는 것으로 나타났다.

3장의 주요 연구 결과는 다음과 같다. 자기응답으로 조사된 자료에서 발생하는 히핑 현상에 대해 1~15차 연도 한국복지패널조사 자료의 경상소득을 중심으로 살펴보았다. 주로 처음 응답한 조사 차수, 1차 연도와 신규표본의 첫해 조사에서 히핑 현상의 비율이 다른 조사 차수에 비해 높았다. 경상소득에 대한 히핑 존재 여부와 가구 특성 변수 간의 관계를 로지스틱 회귀분석을 통하여 살펴본 결과, 강원, 충북, 광주, 전남, 전북, 제주는 다른 지역에 비하여, 저소득 가구는 일반 가구에 비하여, 가구원 수가 증가할수록, 가구주 나이가 많을수록 중졸 이하의 교육 수준일 때 히핑 발생 확률은 낮게 나타났다. 홍민기 등(2014)의 히핑 보정 방법을 확장한 새로운 방법으로 히핑 발생 확률을 반영한 대체를 통한 히핑 보정 방법을 제안하였다. 이 방법을 한국복지패널조사 자료의 경상소득에 적용한 결과, 히핑 보정을 통한 값들이 전체적으로 줄어들었으며 조사된 응답값과의 차이도 일정 구간에 속하였다. 선행연구에서처럼 히핑은 응답값 기준 일정 범위에서 발생한다는 가정하에서는 조금 더 타당한 보정 방법이라고 생각한다. 일반 가구와 저소득 가구를 나누어 히핑 보정된 소득의 분포를 살펴보면 저소득 가구의 경상소득이 증가하는 방향으로, 또 표준편차는 줄어드는 방향으로 보정이 실행되었음을 알 수 있었다. 한편, 경상소득에 대하여 (중위값×0.6)의 값을 기준으로 하위소득 가구의 비율은 히핑 보정 전 1차 연도와 크게 차이가 없는 것으로 나타나, 1차 연도 저소득 가구의 비율이 높은 부분에 있어서 히핑이 큰 원인은 아니라고 생각한다.

4장에서는 2017년 가계금융복지조사 자료에서 가구주의 개인 근로소득, 가구의 근로자녀장려금, 가구원의 기초연금에 대해 응답값에 측정오차를 포함하고 있는지를 살펴보았고, 측정오차 보정을 통한 회귀계수 추정, 히핑 보정, 비례핫덱대체 방법을 사용하여 측정오차를 보정하였다.

근로소득, 근로자녀장려금과 기초연금에 대한 측정오차의 구조는 차이가 있는 측정오차(differential measurement error)였다. 차이가 있는 측정오차는 응답값이 관심 설명변수에 따라 행정보완값에 대해 체계적인 경우를 의미하며, 이에 대한 측정오차 보정 방법을 사용하여 회귀계수를 추정한 바 관심 설명변수의 회귀계수는 모두 행정보완값의 회귀계수와 정확하게 일치하는 것을 확인하였다. 그러나 대부분의 보통 자료에서는 참값을 알고 있는 경우가 드물어서 일반 연구자는 참값을 모르고 있을 가능성이 크다는 측면에서 봤을 때 유용한 방법이라고 볼 수 있다. 한편 응답값과 행정보완값 간의 차이가 크고 연관성이 낮을수록 표준오차는 커지는 것으로 나타났다. 이로 인해 근로자녀장려금과 기초연금은 보정 후 회귀계수의 유의성이 보정 전과 다른 결과를 가졌다. 차이가 있는 측정오차 보정을 통한 회귀계수 추정 분석의 한계는 설명변수 형태(type)를 이분형 범주만 사용 가능하다는 점, 측정오차에 영향을 미치는 설명변수에 대한 회귀모형 분석만 가능하여 더 많은 관심 설명변수의 영향을 파악할 수 없다는 점을 들 수 있다. 그러나 여러 설명변수가 함께 오차 없는 측정값과 연관되는 것을 현실적으로 모형화하기 어렵기 때문에, 본 연구와 같이 하나의 변수에 국한하는 경향이 있다고 볼 수 있다.

한편 측정오차 보정을 통한 회귀계수 추정 시 내부보정방법을 사용하였는데, 이때 가계금융복지조사 자료는 응답자가 응답값과 행정보완값을 모두 가지고 있어서 측정오차 보정 시 모두 활용하였다. 보통 한정적인 예산과 실측 자료 수집의 어려움 등으로 이와 같은 형태의 자료를 사용하기는 쉽지 않은 편이다. Nab, Groenwold, Welsing, and van Smeden(2019)의 논문에서는 보정 표본(calibration sample)의 크기, 측정오차가 있는 값과 없는 값 간의 상관관계 정도에 따른 측정오차의 보정 효과에 대한 모의실험을 통해 다음과 같은 결과를 도출하였다. 약한

상관관계(R^2=0.2)를 가지는 경우, 외부보정 표본의 크기를 50개까지 늘리기 전까지는 측정오차의 보정 결과가 효과적이지 않았다. 그러나 보통 이상의 상관관계(R^2=0.5 또는 0.8)인 경우, 외부보정 표본의 크기가 15개로 적은 편이어도 회귀계수 보정 결과가 향상됨을 보였다. 이렇듯 측정오차를 판단할 수 있는 실측 자료(참값, 실측값 등)를 전체가 아닌 일부라도 구축하고 계속 측정오차 정도를 모니터링하는 방안을 고려해 볼 수 있다고 생각한다.

만약 측정오차 보정의 근거로 사용할 수 있는 자료가 없으면 본 연구 결과에 근거하거나, 측정오차를 추정할 수 있는 다른 조사 차수의 자료에 근거하여 실측 자료가 없는 다른 조사 차수의 관심 변수에 대한 측정오차 보정을 고려해 볼 수 있다. 예를 들면 2017년 가계금융복지조사 자료에서 측정오차모형을 통해 얻은 추정치를 기반으로 하여 2020년 자료의 관심 변수에 대해 외부보정할 수 있을 것이다. 단, 이를 위해서는 2017년의 측정오차를 가지고 측정된 자료와 측정오차 없이 측정된 자료 간의 연관성이 2020년에서도 동일하게 유지되어야 하는 가정을 만족해야 하므로 주의가 필요하다.

다음으로 근로소득에 대해 히핑 보정을 하였는데, 보정 후 결과는 전반적으로 부드러운 분포 함수 형태를 보였다. 히핑 보정 후 평균도 행정보완값과의 차이가 작았다. 히핑 보정 전 평균은 행정보완값과의 차이가 더 큰 편이었다. 그러나 히핑 보정 후 표준편차가 행정보완값보다 작게 나타나므로, 내부보정 또는 외부보정 자료를 추가 정보로 사용하여 히핑 보정을 고려해 볼 수 있을 것이다.

마지막으로 근로자녀장려금의 응답값이 0원인 경우는 0원을 무응답으로 간주한 다음, 비례핫덱대체 방법을 사용하여 적절한 값으로 대체하였다. 근로자녀장려금의 무응답 비율은 81.9%로 아주 높은 편이었다. 활용

한 대체 방법은 3가지로 비례핫덱대체 방법에서 대체군 활용 변수가 3개, 4개인 경우와 회귀대체 방법이었다. 무응답 비율이 매우 높은 편이라는 점에서 대체 시 활용할 수 있는 자료(응답값 정보)가 충분하지 않아서 전반적으로 모든 대체 방법의 효과에 영향을 주었다. 그러나 대체 방법 중에서 비례핫덱대체 방법의 대체군 활용 변수가 4개인 경우의 대체 효과가 우수한 결과를 가지는 것으로 나타났다. 이는 무응답 변수와 연관성이 높은 행정보완값을 대체군 활용 변수로 사용하였고, 대체값을 여러 개 생성하여 최종 하나의 값으로 산출하여 대체값에 대한 변동 부분을 고려하였기 때문이라고 볼 수 있다. 회귀대체 방법도 행정보완값을 설명변수로 사용하였으나, 최적의 회귀모형 적합이 되지 않은 점이 원인이라고 생각한다.

앞서 실시한 방법은 해당 조사 차수에서 측정오차가 있는 관심 변수를 무응답으로 간주하고, 무응답을 대체할 때 행정보완값을 사용한 것이다. 추가적인 방법을 제안해 본다면, 해당 조사 차수의 측정오차모형에서 얻은 추정치를 기반으로 회귀모형을 적합하여 관심 변수의 예측값을 구하고, 이 예측값으로 대체해 주는 방법도 고려해 볼 수 있다. 이 방법이 가정을 만족한다면 다른 조사 차수에서 활용해 볼 수 있다.

실제 조사 자료에서 연속형 관심 변수에 대한 측정오차를 보정하는 방안은 먼저 관심 변수에 대해 측정오차(히핑 포함)를 가지는지 현황을 파악하고, 관심 변수값에 측정오차가 있으면 그 측정오차를 보정하는 것이다.

관심 변수에 대한 측정오차의 포함 여부는 그림(plot)을 통해 응답값과 실제값의 분포 파악, 대응 표본 t-검정 실시, 다항 로지스틱 회귀모형 분석 등을 통해 파악할 수 있다. 또한 히핑 현상을 살펴볼 때는 관심 변수값에 대한 상위 응답 비중, 특정 배수의 형태를 가지는지에 대한 분석, 그림을 통해 확인할 수 있다.

측정오차 보정 방법으로는 측정오차 보정을 통한 회귀계수 추정, 커널 함수 추정법을 활용한 히핑 보정, 비례핫덱대체(FHDI) 방법을 활용한 보정 등이 있다. 측정오차 보정을 통한 회귀계수 추정을 위해서는 실측 자료가 갖춰져 있어야 하지만, 해당하는 모든 응답값에 대한 자료를 가지고 있을 필요는 없다. 측정오차 보정을 통한 회귀계수 추정은 R 통계패키지에 'mecor' 패키지 내부의 'mecor' 함수를 사용하면 된다. 커널 함수 추정법을 활용한 히핑 보정은 R 통계패키지에 'Kernelheaping' 패키지 내부의 'dheaping' 함수이다. 비례핫덱대체 방법을 활용한 보정은 R 통계패키지에 'FHDI' 패키지 내부의 'FHDI' 함수이다. 이외에 적합한 다른 대체 방법을 활용하여 관심 변수에 대한 측정오차를 보정할 수 있다. 이렇듯 R 통계패키지에는 측정오차 보정을 위한 패키지가 있어 일반 연구자도 어렵지 않게 활용할 수 있어서 유용하다고 생각한다.

또한 측정오차로 인한 편향을 줄이기 위해 사후적인 보정에 앞서 근본적으로 조사 설계 및 진행, 조사참여자의 응답, 데이터 처리 과정에서의 발생 가능성을 낮추려는 노력이 필요하다. 마지막으로, Kasprzyk(2005)가 언급한 대로 측정오차의 원인을 4가지로 구분하여 관리 방안을 제안하고자 한다(〈표 5-1〉 참조).

〈표 5-1〉 측정오차 관리 방안

구분	관리 방안
설문 내용	- 설문 구성에 있어서 사전·사후 정성 조사: 설문 문구 수정 및 사후 결과 해석에 활용함 - 과감한 설문 축소 - 설문 문항을 응답자가 이해하기 쉽도록 최대한 쉽게 표현함 - 종속형 설문 및 Event History Calendar 도입
자료 수집 방법	- 리뷰 또는 검증 등을 통한 사후 자료 확인 및 재조사 - 일부의 실측 자료(참값, 실측값 등) 구축 및 활용 - Paradata를 활용하여 조사 과정 분석 - 관련 영수증 제출, 실제값 측정 등을 통한 자료 수집 - 행정 자료 등을 다양하게 활용하여 조사 자료와 연계
조사원	- 조사원의 교육 강화 - 조사원 교육 후 테스트 - 조사 초기 리뷰 강화로 조사원에 대한 피드백 제공 - 조사에 대한 조사원 태도를 사후 파악하여 피드백 제공 - 조사원의 처우 개선
응답자	- 금전적 인센티브의 현실화 필요성 - 정서적 인센티브 고취: 조사의 중요성을 강조하여 책임감이나, 국가 정책 수립에 관여한다는 효능감 등 고취 - 제도적 인센티브 : 세금 감면, 전기료 인하, 봉사활동 점수 제공 등

자료: 2장 내용 및 필자 작성

첫째, 설문 구성에 있어서 사전·사후 정성조사를 실시하여 설문 문구 수정, 사후 결과 해석 등에 활용하는 것이다. 현재 예산이 크거나 중요한 조사는 사전 조사 또는 예비조사를 실시하여 반영하고 있는 편이다. 그러나 지나치게 많은 설문 문항, 어려운 설문 문구, 정확한 응답을 꺼릴 만한 설문(소득, 자산, 부채 등), 응답자도 정확히 알지 못하는 설문(가구원의 소득, 지출, 부동산의 현재 가격, 가구 전체의 생활비, 가구원 개인별 생활비 등) 등의 원인을 꼽을 수 있다. 이를 위해 과감한 설문 축소, 설문 문항을 응답자가 이해하기 쉽게 풀어쓴 표현, 종속형 설문 및 Event History Calendar의 도입 등을 고려해 볼 수 있다.

둘째, 자료 수집 방법에서의 해결 방안은 리뷰, 검증 등을 통한 사후 자

료 확인 및 재조사이다. 리뷰, 검증 등으로 조사원에 의한 에러나 거짓 자료 수집 등을 확인할 수 있다. 보통 조사 과정에는 조사원이 응답 내용을 1차 점검하고, 그 조사 내용을 감독관이 다시 한번 리뷰(검토)하는 절차가 있다. 이렇게 조사를 완료한 데이터에 대해서는 부분적으로 전화 검증 등을 실시한 후, 필요시 폐기 및 대체, 혹은 재조사를 진행하고 있다. 통계청의 경우, 내검이라 하여 내부에서 조사 내용을 점검하고 확인하여 보완하는 절차를 가지며, 무응답의 경우 사후에 통계적으로 보정하는 절차로 진행하는 것으로 알려져 있다. 그런데 대부분 조사의 경우, 리뷰 또는 검증에 대한 적정한 비용 산정이 이루어지지 않고 있으며, 조사 일정에 충분한 시간을 반영하지 못하고 있는 편이다. 리뷰나 검증의 중요성을 강조하고, 충분한 시간과 비용의 반영이 필요하다고 생각한다.

또한, 수집된 조사자료의 관리 방안으로 조사자료 일부에 대한 실측 자료(참값, 실측값 등) 구축을 고려해 볼 수 있다. 일부 실측 자료만을 가지고 타당성 연구(validation study), 측정오차 보정 등에서 활용할 수 있다는 점에서 유용하다고 생각한다.

요즘에는 CATI, CAPI, CAWI 등을 활용한 조사가 많은 편인데, 이러한 조사 도구를 사용하면 조사 과정에서 발생하는 모든 관련 정보인 Paradata를 수집할 수 있다. 예를 들면 조사 참여 응답 시간, 문항별 응답 시간, 문항별 응답 경로 및 변경 횟수, 응답자 관련 특성, 조사원 관련 사항 등에 대한 것이다. 이렇듯 Paradata에서 수집된 정보는 측정오차와의 연관성이 높으므로, 적극적으로 활용할 필요가 있다고 생각한다.

한편, 영수증 제출, 실제값 측정 등을 활용한다면 응답값의 정확성을 높일 수 있다. 한국의료패널조사의 경우 의료 이용 영수증을 수집하여 의료비를 확인하고 있다(김남순 외, 2018, p.20). 한국재정패널조사의 경우도 근로소득세를 내거나 종합소득세 확정 신고를 한 사람에 대해서는

소득 공제 현황과 근로소득 연말정산을 신청하기 위해 회사에 제출했던 영수증으로 소득을 확인(영수증 제출을 동의한 사람에 한정함)하고 있다(한국조세재정연구원, 2020, p.68). 에너지소비실태조사는 전년도 에너지소비량을 기억하지 못하는 가구 중 희망 가구에 한해서 고객 번호를 조사하고 추후 공급사에 소비량(전력, 도시가스 등의 네트워크에너지)을 조회하여 얻는다(에너지경제연구원, 2018, p.4). 장기적으로는 정확한 응답을 꺼리거나 응답자도 알 수 없는 소득이나 지출 등은 한국신용정보원의 전산 자료 또는 국세청과 같은 행정자료를 다양하게 활용하여 연계해 볼 수 있다. 현재 개인정보보호 문제로 한계가 있으나, 응답자의 동의를 얻은 후 자료를 활용하는 등과 같은 해결 방안을 모색할 필요가 있다. 이는 응답의 부담을 덜어 줄 뿐만 아니라 응답값에 대한 정확도 향상에 기여할 것이다.

셋째, 조사원의 교육 강화이다. 다양한 실태조사 시 조사원이 설문을 그대로 읽어준다면, 응답자는 잘못 응답할 가능성이 커질 것이다. 이에 보통 조사원을 대상으로 심층 교육을 시행하고 있다. 하지만 조사원 교육만으로는 이해도를 충분히 높이기 어려운 경우가 많은 편이다. 이를 감안하여 교육 후 테스트, 조사 초기 리뷰 강화로 조사원에 대한 피드백(feedback) 제공, 조사에 대한 조사원 태도를 사후 파악하여 피드백 제공 등을 실시할 수 있다.

마지막으로, 응답자에 대한 것이다. 현재 응답자 인센티브는 물질적 인센티브(선물, 사례비, 경품 등)가 대부분이다. 그러나 응답 유인 효과를 크게 할 만큼 충분한 사례가 이루어지지 않는다는 한계가 있다. 물질적 인센티브가 한계가 있다면 정서적 인센티브도 중요하게 고려할 수 있다. 정서적 인센티브라고 한다면, 조사의 중요성을 강조하여 책임감을 느끼게 하거나, 국가 정책 수립에 관여한다는 효능감을 고취할 수 있도록 조

사 취지 등을 홍보하는 방안이 있을 수 있다. 또한, 조사참여자에 대한 제도적 인센티브 방안도 모색해 볼 수 있다. 예를 들면 세금 감면, 전기료 인하, 학생들에 대한 봉사활동 점수 제공 등이 있을 수 있다.

근본적으로 측정오차를 모두 해결하기는 어려우나, 응답자의 적극적인 참여, 조사원의 성실한 진행 및 조사원에 대한 적절한 관리, 연구자의 현실을 고려한 기획 등을 통해 상당 부분 해결이 가능할 것이므로 조사 모든 단계에서 철저하게 관리되어야 할 것이다.

참고문헌

김남순, 서제희, 정연, 오미애, 이정아, 정수경, …, 오하린. (2018). **2기 한국의료패널 구축·운영을 위한 기초 연구.** 한국보건사회연구원.

김준원, 신동균. (2016). 서베이 자료에 존재하는 측정오차의 문제점과 자료검증 연구의 필요성. **산업경제연구. 29**(1). 249-278

김정섭, 임경은. (2010). **총오차 축소를 위한 Paradata 수집방안**. 2010년 상반기 연구보고서Ⅱ, 62-110, 통계개발원.

박인호, 전경배, 주성제. (2017). 서베이통계의 비표집오차 축소방안 연구. **국민계정리뷰, 2017년 제1호**, 한국은행.

에너지경제연구원. (2018). **2018년 가구에너지 상설표본조사**.

이승희. (2010). 총조사 오차(Total Survey Error)의 패러다임으로 이해하는 표본조사. **통계연구, 15**(1), 44-74.

통계청. (2018). **2017년 가계금융·복지조사**.

통계청. (2020). **가계금융복지조사에서의 조사자료와 행정자료의 통합방법 이해**.

통계청. (2021). **국가통계 품질관리 매뉴얼 정기토예품질진단.**

통계청. (2018). **2017 가계금융·복지조사[데이터파일].** 통계청 MDIS, RAS.

한국보건사회연구원. (각연도). **2006~2020 한국복지패널조사[데이터파일].** https://www.koweps.re.kr에서 2021. 5. 9. 인출.

한국조세재정연구원. (2020). **12차년도 재정패널 조사 기초분석보고서**.

홍민기, 김재광, 한치록, 김기민. (2014). **패널자료 품질개선 연구(Ⅲ)**. 한국노동연구원.

Abowd, J. M., & Harrison Stinson, M. (2011). Estimating measurement error in SIPP annual job earnings: A comparison of Census Bureau survey and SSA administrative data. *US Census Bureau Center for Economic Studies Paper No. CES-WP-11-20.*

Biemer P. P., Groves R. M., Lyberg, L. E., Mathiowetz, N. A., & Sudman

S. (1991). *Measurement Errors in Surveys*. New York: John Wiley & Sons, INC.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality (Vol. 335)*. John Wiley & Sons.

Bingley, P., & Martinello, A. (2014). Measurement error in the Survey of Health, Ageing and Retirement in Europe: A validation study with administrative data for education level, income and employment. *Work Pap Ser*, 16-2014.

Cummings, T. H., Hardin, J. W., McLain, A. C., Hussey, J. R., Bennett, K. J., & Wingood, G. M. (2015). Modeling heaped count data. *The Stata Journal, 15*(2), 457-479.

Duncan, G. J., & Hill, D. H. (1985). An investigation of the extent and consequences of measurement error in labor-economic survey data. *Journal of Labor Economics, 3*(4), 508-532.

European Research Infrastructure Consortium. (2021). *SHARE Release Guide 1.0.0 of Wave 8*.

Fuller, W. A., & Kim, J. K. (2005). Hot deck imputation for the response model. *Survey Methodology, 31*(2), 139.

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public opinion quarterly, 74*(5), 849-879.

Groß, M., & Rendtel, U. (2016). Kernel density estimation for heaped data. *Journal of Survey Statistics and Methodology, 4*(3), 339-361.

Guo, Y. (2010). *Multiple Imputation for Measurement Error Correction Based on a Calibration Sample* (Doctoral dissertation).

Guo, Y., Little, R. J., & McConnell, D. S. (2012). On Using Summary Statistics From and External Calibration Sample to Correct for Covariate Measurement Error. *Epidemiology, 23*(1), 165-174.

Im, J., Cho, I. H., & Kim, J. K. (2018). FHDI: An R Package for Fractional Hot Deck Imputation. The *R Journal, 10*(1), 140.

Im, J., Kim, J. K., & Fuller, W. A. (2015). Two-phase sampling approach to fractional hot deck imputation. In *Proceedings of the Survey Research Methods Section* (pp. 1030-1043).

Institute for Social and Economic Research. (2006). *Quality Profile: British Household Panel Survey Version 2.0. University of Essex*.

Institute for Social Research. (2021). *PSID Main Interview User Manual*. University of Michigan.

Kalton, G., Kasprzyk, D., & McMillen, D. B. (1989). Non-sampling errors in panel surveys. In *Panel Surveys*. New York: John Wley and Sons.

Kasprzyk, D. (2005). *Measurement error in household surveys: sources and measurement*. Mathematica Policy Research.

Kim, J. K., & M. Hong. (2012). Imputation for statistical inference with coarse data. *The Canadian Journal of Statistics, 40*(3), 604-618.

Kim, J.K., & Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika, 91*(3), 559-578.

Marquis, K. H., & Moore, J.C. (1990). *Measurement errors in SIPP program reports*. The Survey of Income and Program Participation Working Paper No.113. Washington D.C.: US Department of Commerce Bureau of the Census.

McGonagle, K. A., Schoeni, R. F., Sastry, N., & Freedman, F. A. (2012). The Panel Study of Income Dynamics: overview, recent innovations, and potential for life course research. *Longitudinal and Life Course Studies, 3*(2), 268-284.

Moore, J. (2008). *Seam Bias in the 2004 SIPP Panel: Much Improved, but Much Bias Still Remains*. Washington D.C.: US Census

Bureau.

Nab, L., Groenwold, R. H., Welsing, P. M., & van Smeden, M. (2019). Measurement error in continuous endpoints in randomised trials: problems and solutions. *Statistics in medicine, 38*(27), 5182-5196.

Nab, L., van Smeden, M., Keogh, R. H., & Groenwold, R. H. (2021). mecor: An R package for measurement error correction in linear regression models with a continuous outcome. Computer Methods and Programs in Biomedicine, 106238.

National Research Council. (2009). *Reengineering the Survey of Income and Program Participation*. Washington DC: The National Academy Press.

Organisation for Economic Cooperation and Development. (2003). *Business Tendency Surveys: A Handbook*.

Pischke, J. (1995). Measurement Error and Earnings Dynamics: Some Estimates From the PSID Validation Study. *Journal of Business & Economic Statistics, 13*(3), 305-314.

Statistical Policy Office (2001). *Measuring and Reporting Sources of Error in Surveys*. Statistical Policy Working Paper No. 31. Washington D.C.: Office of Management and Budget.

Taylor, M. F., Brice, J., Buck, N., & Prentice-Lane, E. (2018). *British Household Panel Survey User Manual Volume A: Introduction, Technical Reports and Appendices*. University of Essex.

US Department of Commerce, Economic and Statistics Administration, & US Census Bureau. (2021). 2018 Survey of Income and Program Participation Users' Guide.

Zalsha, S. (2020). *Examining Multiple Imputation for Measurement Error Correction in Count Data with Excess Zeros*.

간행물 회원제 안내

회원에 대한 특전

- 본 연구원이 발행하는 판매용 보고서는 물론 「보건복지포럼」, 「국제사회보장리뷰」도 무료로 받아보실 수 있으며 일반 서점에서 구입할 수 없는 비매용 간행물은 실비로 제공합니다.
- 가입기간 중 회비가 인상되는 경우라도 추가 부담이 없습니다.

회원 종류

전체 간행물 회원	보건 분야 간행물 회원
120,000원	75,000원
사회 분야 간행물 회원	**정기 간행물 회원**
75,000원	35,000원

가입방법

문의처

- (30147) 세종특별자치시 시청대로 370 세종국책연구단지 사회정책동 1~5F
  간행물 담당자 (Tel: 044-287-8157)

KIHASA 도서 판매처

- 한국경제서적(총판) 02-737-7498
- 교보문고(광화문점) 1544-1900
- 영풍문고(종로점) 02-399-5600
- Yes24 http://www.yes24.com
- 알라딘 http://www.aladdin.co.kr