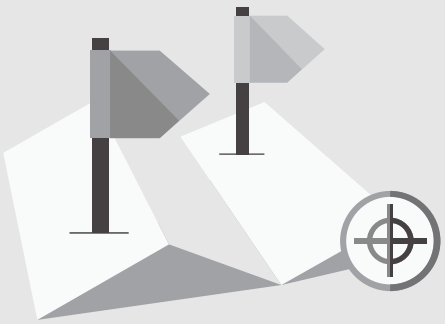
이달의 초점

보건복지분야 데이터 활용 현황과 과제



사회보장제도 근거 강화와 사회보장 행정데이터 구축: 사회보장 행정데이터 연계·활용을 위한 과제
이현주

빅데이터 정보시스템 활용 현황과 과제: 복지 사각지대 발굴 시스템을 중심으로
김은하

보건복지통계 현황과 발전 과제: 보건복지부 소관 국가승인통계를 중심으로
신정우·천미경·전예지·진재현

인구실태조사 사례와 과제: 세대 및 성에 관한 실태조사(GGS)를 중심으로
이소영

조사 자료의 측정오차 보정 및 관리 방안: 근로소득을 중심으로
이혜정

한국보건사회연구원
KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

조사 자료의 측정오차 보정 및 관리 방안: 근로소득을 중심으로[1)]

Measurement Error Correction and Management in Survey Data – Focusing on Earned Income

이혜정 | 한국보건사회연구원 부연구위원

실증분석 연구는 조사를 통해 수집한 자료를 기반으로 하여 의미 있는 연구 결과를 도출하는 것이다. 보통 조사 자료는 관심 모집단에서 일부 대상만을 추출하여 표본조사를 통해 구축한 것으로, 흔히 오차를 포함하고 있으므로 오차의 정도가 심한 자료를 사용하여 통계 분석을 한다면 심각하게 편향된 결과를 도출하게 된다.

이 글에서는 2017년 가계금융복지조사에서 가구주의 개인 근로소득 응답값에 측정오차가 포함되어 있는지를 살펴보고, 측정오차 보정을 통한 회귀계수 추정과 히핑 보정 방법을 사용하여 측정오차를 보정하였다. 또한 고품질 자료 생산을 위한 조사 관리 방안을 제안하였다.

1. 들어가며

실증분석 연구는 조사를 통해 수집한 자료(조사 자료)를 기반으로 하여 의미 있는 연구 결과를 도출하는 것이다. 보통 조사 자료는 관심 모집단에서 일부 대상만을 추출하여 표본조사를 통해 구축한 것이다. 이러한 조사 자료를 가지고 모집단 전체에 대해 추론하므로 추정값과 참값 간에 차이가 발생하기 마련이다. 추정값과 참값의 차이를 오차라고 정의한다. 오차는 다양한 원인과 상황에서 발생할 수 있다. 조사에서 발생 가능한 모든 오차를 총오차(total survey error)라고 하며, 크게 표집오차(sampling error)와 비표집오차(non-sampling error)로 구분할 수 있다. 표

1) 이 글은 이혜정, 신지영, 박승환, 지희정, 오미애. (2021). 조사 자료의 품질 검증 연구 – 측정오차를 중심으로(세종: 한국보건사회연구원)에 수록된 내용을 바탕으로 재구성하여 작성한 것이다.

DOI: 10.23062/2022.11.6

집오차는 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.

현실에 편재하는 통계 분석에서 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 가구 소득이 얼마나 되는지에 관한 질문에 응답자는 응답의 거부감, 세금 문제 때문에 솔직한 응답을 꺼리거나 과거를 기억하지 못하거나 설문을 잘못 이해하는 경우 등 다양한 이유로 정확하게 응답하지 않을 수 있다. 또한 조사원이 설문 진행을 거짓으로 하거나, 조사원이 잘못 이해하고 설문 응답을 받거나, 자료 입력 과정에서 단순 오류가 생기는 경우 등도 있다. 임상연구나 역학연구에서도 모호한 조사, 부정확한 도구 사용, 잘못 설계된 설문 문항, 생물학적 변동 등과 같이 다양한 원인에 의해 측정오차를 포함하고 있을 가능성이 큰 편이다(Guo, 2010). 이렇듯 조사 자료는 보통 오차를 포함하고 있으므로 오차의 정도가 심한 자료를 사용하여 통계 분석을 한다면 심각하게 편향된 결과를 도출하게 된다.

이 글에서는 조사 자료의 금액 변수에 대해 측정오차가 있는지 살펴보고 측정오차를 보정하였다. 이때 사용한 자료는 2017년 가계금융복지조사의 가구주 개인 근로소득 변수이다. 또한 고품질 자료 생산을 위한 조사 관리 방안을 제시하고자 한다.

2. 조사 자료의 금액 변수에 대한 측정오차 분석

가계금융복지조사는 통계청, 금융감독원 및 한국은행이 공동으로 수행하고 있다. 전국의 2만여 표본 가구를 대상으로 2010년부터 해마다 조사를 하고 있다(통계청, 2018). 조사 목적은 소득, 자산, 부채 등에 대한 규모, 구성 및 분포와 미시적 재무 건전성을 파악하여 사회 및 금융 관련 정책과 연구에 활용하는 것이다. 조사 대상은 가구 단위로 전국 동·읍·면에 거주하는 1인 이상의 표본 가구이고, 조사 단위는 1인 가구 및 혈연, 결혼, 입양 등으로 맺어져 생계를 함께하는 가족이다. 조사 항목은 금융과 복지 두 부문으로 구성되어 있는데, 금융 부문은 가구 구성, 자산 및 자산 운용, 부채 및 부채 상환 능력, 소득 및 지출 등이고 복지 부문은 가구 구성, 자산, 부채, 소득, 지출, 노후생활 등이다.

가계금융복지조사는 2017년 조사 자료에 한정하여 주요 금액 관련 설문 문항에 대해 응답값뿐만 아니라 행정보완값을 같이 제공하고 있다. 행정보완값은 행정 자료에서 얻은 것으로, 2017년 조사 자료에는 응답값과 행정보완값이 모두 포함되어 있어서 두 값을 직접 비교할 수 있고 이를 통해 측정오차 분석도 할 수 있어 적합하다고 생각한다. 분석 변수는 측정오차가 발생할 가능성

표 1. 근로소득 변수에 대한 조사 자료와 행정 자료의 개념 및 포괄 범위

조사 자료	행정 자료
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, 퇴직금(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

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

이 큰 편에 속하는 금액 변수인 가구주의 개인 근로소득(이하, 근로소득)으로 선정하였다. 근로소득은 조사 자료와 행정 자료의 개념과 포괄 범위가 '거의 일치'한다고 볼 수 있다(표 1).

근로소득 변수에 대하여 응답값과 행정보완값 간의 관계를 파악하여 측정오차가 있는지 살펴보았다. 분석 대상은 7,274명이며 종사상지위가 상용 임금근로자인 사람의 근로소득으로 한정하였다. 다른 종사상지위(임시/일용근로자, 고용원이 있는/없는 자영업자, 무급가족종사자, 기타 종사자)에 비해 상용 임금근로자는 행정보완 자료(국세청 자료)의 정확성이 높은 편이어서 참값이라고 가정할 수 있다고 판단하였다. 행정보완 자료도 측정오차를 포함할 수 있다는 한계가 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 가정한다(통계청, 2020, p. 74). 가구주로 한정한 이유는 가구원과 다르게 근로소득과 연관성이 높은 직업, 산업 등의 다양한 정보가 있어 활용할 수 있기 때문이다.

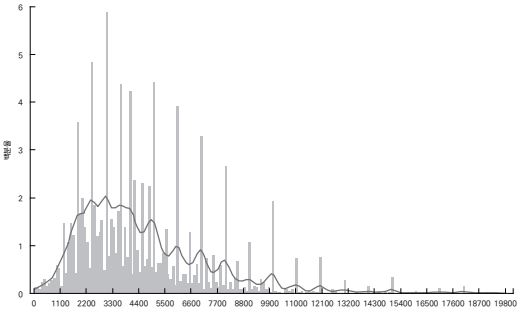
[그림 1]을 보면 (1) 응답값 분포는 (2) 행정보완값과는 다르게 일정한 값마다 꽤 높은 비율을 나타내는 특정 값들이 잦은 반면에, 행정보완값은 전체적으로 완만한 분포를 보이고 있다.[2] (3) 응답값과 행정보완값 간의 분포를 보면 $y = x$를 기준으로 위는 과대 보고, 아래는 과소 보고인 경우이다. 과대 보고는 낮은 근로소득에서 많이 나타나는 반면 과소 보고는 전반적으로 고르게 분포되어 있다.

〈표 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, 유의 수준 5%에서 응답값과 행정보완값은 통계적으로 유의한 차이가 있다고 할 수 있다.

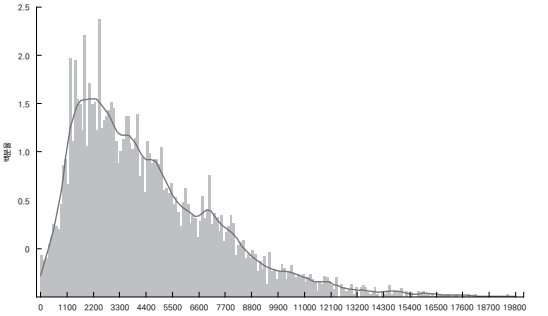
2) 조사 자료의 값은 '응답값'으로, 행정 자료의 값은 '행정보완값'이라고 정의한다.

그림 1. 근로소득에 대한 응답값과 행정보완값의 분포

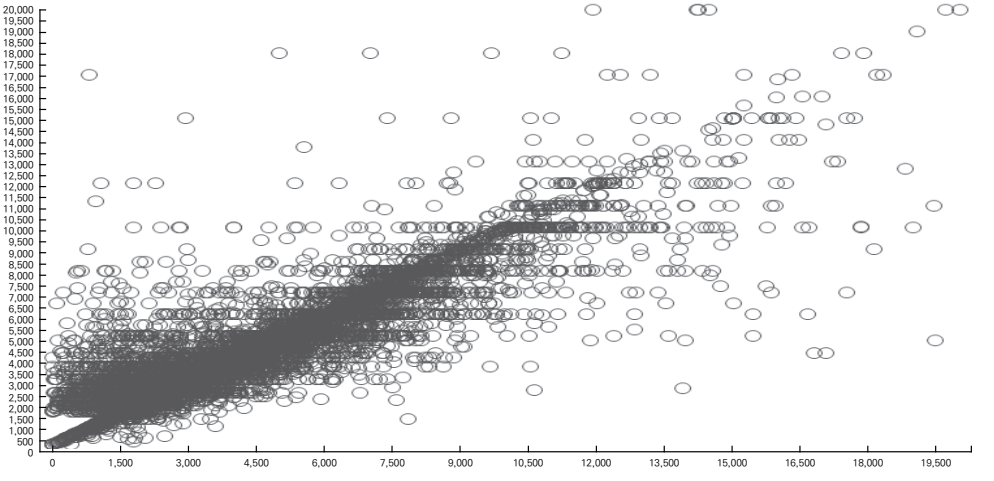
(1) 응답값



(2) 행정보완값



(3) 응답값과 행정보완값



주: 1) (1), (2)의 단위는 X축은 만 원, Y축은 %임.
2) (1)의 전체 수는 7,241명이고 (2)는 7,232명임(2억 원 미만인 경우만 시각화함).
3) (3) X축은 행정보완값이고, Y축은 응답값임(단위: 만 원).
4) (3)은 행정보완값과 응답값 모두 2억 원 미만인 경우에 대해서만 시각화함(7,229명).
자료: 이혜정, 신지영, 박승환, 지희정, 오미애. (2021). 조사 자료의 품질 검증 연구– 측정오차를 중심으로. p. 93. [그림 4-1].

표 2. 근로소득 응답값과 행정보완값의 대응 표본 t-검정 결과

(단위: 명, 만 원)

	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: 이혜정, 신지영, 박승환, 지희정, 오미애. (2021). 조사 자료의 품질 검증 연구– 측정오차를 중심으로. p. 94. 〈표 4-2〉.

표 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: 이혜정, 신지영, 박승환, 지희정, 오미애. (2021). 조사 자료의 품질 검증 연구- 측정오차를 중심으로. p. 95. 〈표 4-3〉.

응답값과 행정보완값 간 차이를 3개 집단으로 구분하였다(표 3). 응답값이 행정보완값보다 더 작은 경우가 48%였고, 응답값이 더 큰 경우는 43.2%였다. 응답값과 행정보완값이 같은 경우는 8.8%로 비율이 낮은 편이었다. 3개 집단별 분포를 보면 응답값이 행정보완값보다 작은 경우의 평균이 가장 크고, 다음으로 응답값이 행정보완값보다 큰 경우, 응답값과 행정보완값이 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 유의 수준 0.1%에서 응답값과 행정보완값 간에 통계적으로 유의한 차이가 있는 것으로 나타났다. 응답값이 행정보완값보다 큰 경우의 평균 차이는 1,024만 원이고 응답값의 표준편차가 행정보완값에 비해 약간 크나 차이가 크지 않은 편이었으며 유의 수준 0.1%에서 응답값과 행정보완값 간에 통계적으로 유의한 차이가 있는 것으로 나타났다.

3. 조사 자료의 금액 변수에 대한 측정오차 보정 방안

앞 장의 분석 결과를 통해 근로소득에는 측정오차가 존재함을 확인하여 이 장에서는 측정오차 보정을 하였다. 측정오차를 보정할 때 사용한 방법으로는 측정오차 보정을 통한 회귀계수 추정, 히핑 보정 방법이 있다. 이러한 방법에 관한 이론적 개념과 적용 방법은 이혜정, 신지영, 박승환, 지희정, 오미애(2021)를 참고하면 된다. 여기서는 근로소득에 적용한 결과를 중심으로 살펴보았다.

가. 측정오차 보정을 통한 회귀계수 추정 방안

1) 측정오차 구조[3)]

2개 처리 효과($X \in 0,1$)를 비교하는 확률화된 실험을 가정하면 X는 처리집단과 대조집단으로 정의한다. 또한 Y는 실제값(true trial endpoint)이고, Y^*는 Y의 오차를 포함한 값이며 연속형

이라고 하자. Y에 대한 선형회귀모형을 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 처리 효과는 보통최소제곱법(OLS: Ordinary Least Squares) 추정량이다.

측정오차의 구조는 4개(전형적인 측정오차, 이분산성 측정오차, 체계적인 측정오차, 차이가 있는 측정오차)로 구분할 수 있으며, 이 글에서는 차이가 있는 측정오차(differential measurement error)를 중심으로 살펴보았다.

2) 차이가 있는 측정오차 개념

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^2이며, 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 분산을 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.

3) 차이가 있는 측정오차를 갖는 경우에 대한 보정 방법

Y^*는 무작위로 할당된 모든 사람에게서 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 본표본이다.

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

Y^*가 차이가 있는 측정오차를 갖는 경우에 대한 보정 방법을 살펴보았다. α_Y와 β_Y의 추정량은 다음과 같다.

$$\hat{\alpha}_Y = (\hat{\alpha}_{Y^*} - \hat{\theta}_{00})/\hat{\theta}_{10} \text{ 이고}$$

$$\hat{\beta}_Y = (\hat{\beta}_{Y^*} + \hat{\alpha}_{Y^*} - \hat{\theta}_{01})/\hat{\theta}_{11} - \hat{\alpha}_Y$$

여기서 $\hat{\theta}_{00}$, $\hat{\theta}_{10}$, $\hat{\theta}_{01}$, 그리고 $\hat{\theta}_{11}$은 보정 자료에서 보통최소제곱법을 사용하여 추정한다.

3) Nab, Groenwold, Welsing, & van Smeden(2019)을 참고하여 측정오차 구조, 개념 및 보정 방법을 정리하였다.

표 4. 차이가 있는 측정오차 모형에 대한 근로소득 회귀분석 결과

	계수	표준오차	표준오차(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: 계수 ( ) 안의 값은 p-value에 대한 유의성을 나타내며, (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 이혜정, 신지영, 박승환, 지희정, 오미애. (2021). 조사 자료의 품질 검증 연구- 측정오차를 중심으로. p. 110. 〈표 4-11〉.

$\hat{\theta}_{10}$과 $\widehat{\theta_{11}}$은 유한 추정값인 $\hat{\alpha}_Y$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, 부트스트랩 방법을 사용하여 구할 수 있다.

4) 근로소득에 대한 측정오차 보정 방법

근로소득의 측정오차 모형은 차이가 있는 측정오차 모형이며, 로그변환한 응답값에 대한 회귀분석 결과는 〈표 4〉와 같다. 설명 변수는 로그변환한 행정보완값, 가구주의 성별(기준 변수: 여성), 로그변환한 행정보완값과 가구주의 성별 교차항인데, 계수들이 모두 통계적으로 유의하게 나타났다. 이로써 θ_{01}=2.954이고 θ_{11}=0.652를 구할 수 있다.

다음은 측정오차 보정을 통한 근로소득 회귀분석 결과이다(표 5). 측정오차 모형에서 포함된 가구주 성별은 설명 변수이고 응답값은 종속 변수이다. 보정 전과 보정 후의 계수값 차이를 보면, 보정 후가 보정 전보다 상수항은 0.102 작고, 가구주 성별은 0.023 크게 나타났다. 그리고 보정 전과 보정 후 표준오차 차이를 보면, 보정 후가 보정 전보다 상수항은 0.021 컸으며, 가구

표 5. 측정오차 보정을 통한 근로소득 회귀분석 결과

		계수	표준오차	표준오차(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	-
행정보완 자료	상수항	7.632 (***)	0.025	-
	가구주 성별_남성	0.621 (***)	0.028	-

주: 계수 ( ) 안의 값은 p-value에 대한 유의성을 나타내며, (***)은 0.001, (**)은 0.05, (*)은 0.1에서 통계적으로 유의하다고 할 수 있음.
자료: 이혜정, 신지영, 박승환, 지희정, 오미애. (2021). 조사 자료의 품질 검증 연구- 측정오차를 중심으로. p. 111. 〈표 4-12〉.

표 6. 근로소득 응답값 개수 및 비율(상위 10개)

근로소득(만 원)	개수(개)	비율(%)
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: 이혜정, 신지영, 박승환, 지희정, 오미애. (2021). 조사 자료의 품질 검증 연구– 측정오차를 중심으로. p. 115. 〈표 4-17〉.

주 성별 역시 0.022 정도 크게 나타났다. 보정 후 계수들이 모두 통계적으로 유의하였다.

한편, 보정 후 계수값은 행정보완 자료를 종속변수로 한 계수값과 정확하게 일치한 일치 추정량이다. 앞에서 설명한 $\hat{\alpha}_Y$와 $\hat{\beta}_Y$의 식에 대입해 보면 보정 후 계수값과 일치함을 확인할 수 있다.

나. 히핑 보정 방안

히핑 현상은 응답자가 대략적인 값으로 응답하여 발생하는 측정오차의 한 가지 형태로 볼 수 있으며, 가계금융복지조사 자료의 근로소득에 대해 히핑 보정을 시행하였다.

우선, 근로소득의 분포가 특정 값의 배수에 많이 치우쳐 있는지 근로소득의 응답값에 대한 비율을 살펴보았다. 〈표 6〉은 응답값 비율이 높은 상위 10개를 나타낸다. 3,000만 원이 5.79%로 가장 비율이 높았으며, 다음은 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 7〉은 근로소득 응답값이 100의 배수와 120의 배수 형태를 보이는 비율이다. 100의 배수는 80.1%로 매우 높은 편이고, 120의 배수도 44.2%로 나타났다.

이러한 형태를 자세히 살펴보기 위해서 분석 대상을 5,000만 원 미만으로 제한하였다. 히핑

표 7. 100 또는 120의 배수인 근로소득의 비율

100의 배수	120의 배수
80.1	44.2

자료: 이혜정, 신지영, 박승환, 지희정, 오미애. (2021). 조사 자료의 품질 검증 연구– 측정오차를 중심으로. p. 115. 〈표 4-18〉.

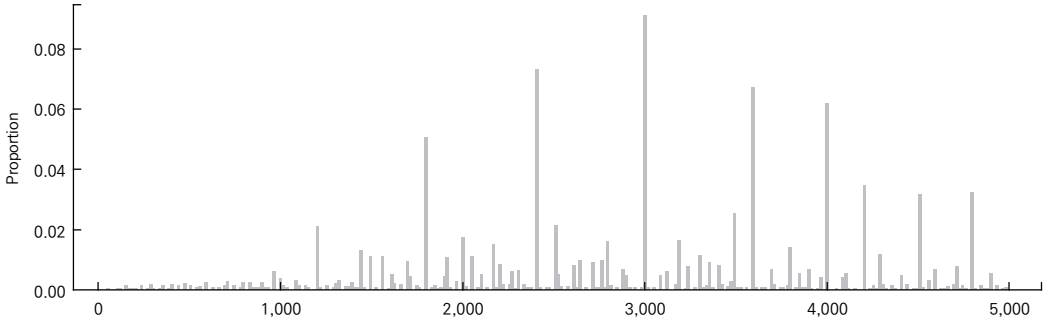
표 8. 5,000만 원 미만의 근로소득 응답값 개수 및 비율(상위 10개)

근로소득(만 원)	개수(개)	비율(%)
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

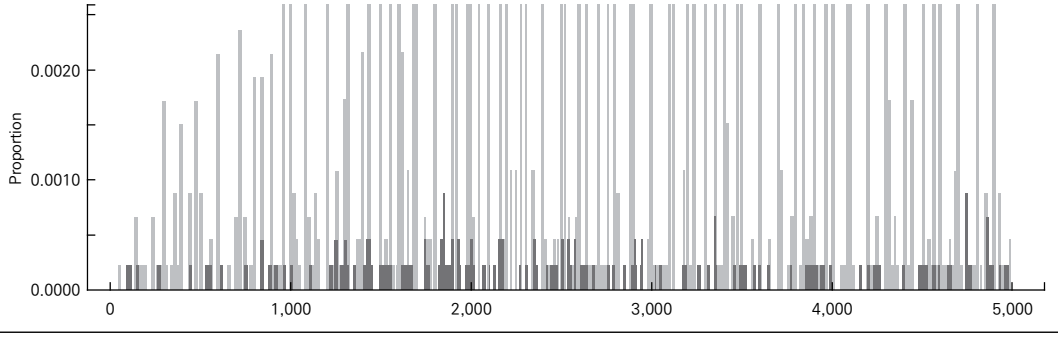
자료: 이혜정, 신지영, 박승환, 지희정, 오미애. (2021). 조사 자료의 품질 검증 연구– 측정오차를 중심으로. p. 116. 〈표 4-19〉.

그림 2. 근로소득에 대한 spike plot

(1) 전체



(2) 비율이 0.0025 이하인 경우



주: X축은 근로소득(단위: 만 원)이고, Y축은 응답값에 대한 비율(Proportion)임.
자료: 이혜정, 신지영, 박승환, 지희정, 오미애. (2021). 조사 자료의 품질 검증 연구– 측정오차를 중심으로. p. 117. [그림 4-4].

보정 시 많은 히핑 지점이 있는 경우 베이지안 방법의 사후분포가 수렴하지 않을 수 있기 때문이다. 〈표 8〉은 근로소득의 응답값 비율이 높은 상위 10개에 대한 것이다. 3,000만 원이 9.1%로 가장 비율이 높았으며, 다음은 2,400만 원(7.35%), 3,600만 원(6.74%), 4,000만 원(6.2%)이 차지하였다.

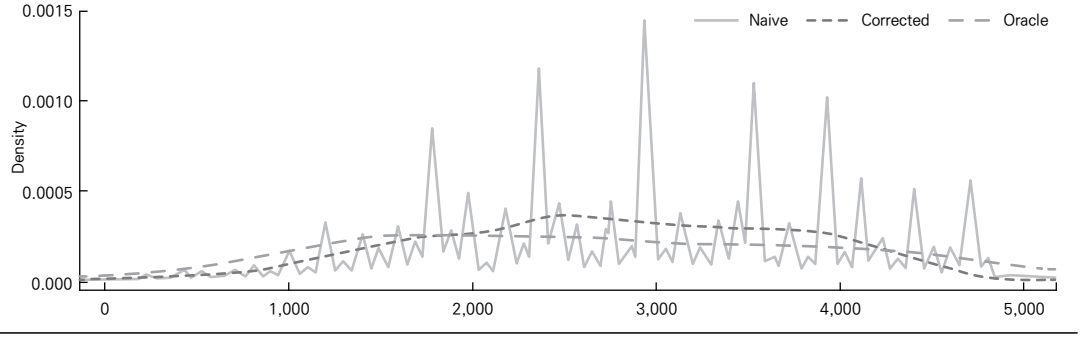
[그림 2]는 응답값에 대한 spike plot을 보여준다. 그림에서 (1) 전체를 보면, 특정 응답값에 분포가 집중되어 비율이 매우 높게 나타나는 곳이 꽤 많은 편이다. 특히 3,000만 원은 비율이 높아서 다른 값들의 분포를 자세히 보기 위하여 (2) 비율이 0.0025 이하인 경우도 살펴보았다.

근로소득의 히핑 발생 지점을 100의 배수로 하여 커널 함수를 통한 히핑 보정 방법을 실시하였다. [그림 3]은 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 함수 형태를 보였다. 행정보완값의 분포 함수와 비교해 보면 약 2,000만 원 이하에서는 보정 후 분포 함수가 낮은 편이었고, 약 4,000만 원 정도까지는 높았다가 이후 다시 낮게 추정되었음을 알 수 있다. 행정보완값의 분포 함수와 차이가 있는 편이지만 응답값에 비해 비슷하게 추정되었다고 볼 수 있다.

[그림 4]는 히핑 발생 지점을 기준으로 커널 분포 함수를 보정한 spike plot을 보여 준다. [그림 3]과 다르게 보정 후 spike plot은 부드러운 분포의 형태로 나타났다.

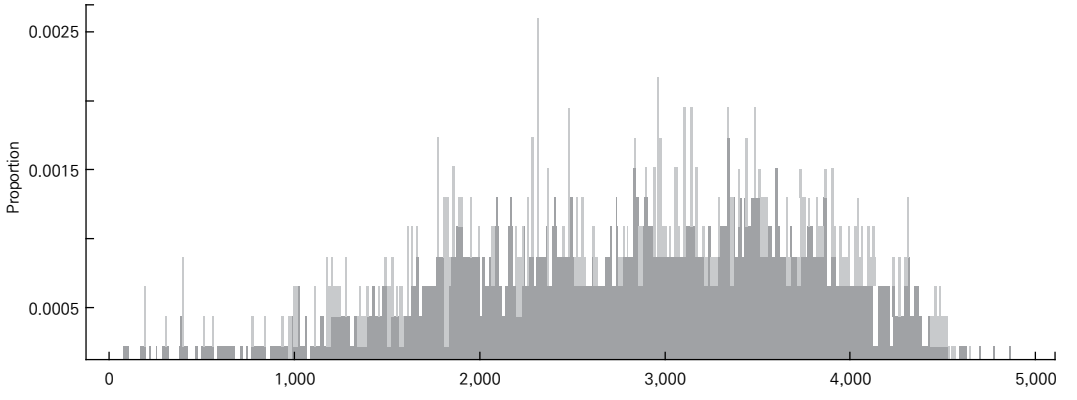
〈표 9〉는 히핑 보정 전후 행정보완값의 근로소득 평균과 표준편차에 대한 것이다. 보정 후 평균은 2,806만 2천 원으로 보정 전(응답값)에 비해 작았고, 행정보완값보다도 작았다. 단, 보정 후를 기준으로 행정보완값의 차이가 보정 전과의 차이보다 작은 것으로 나타났다. 보정 후

그림 3. 커널 함수를 통한 근로소득 히핑 보정 분포 결과



주: X축은 근로소득(단위: 만 원), Y축은 밀도(Density)임.
자료: 이혜정, 신지영, 박승환, 지희정, 오미애. (2021). 조사 자료의 품질 검증 연구– 측정오차를 중심으로. p. 118. [그림 4-5].

그림 4. 커널 함수를 통한 근로소득 히핑 보정 spike plot



주: X축은 근로소득(단위: 만 원)이고, Y축은 응답값에 대한 비율(Proportion)임.
자료: 이혜정, 신지영, 박승환, 지희정, 오미애. (2021). 조사 자료의 품질 검증 연구– 측정오차를 중심으로. p. 119. [그림 4-6].

표 9. 히핑 보정 전후, 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: 이혜정, 신지영, 박승환, 지희정, 오미애. (2021). 조사 자료의 품질 검증 연구– 측정오차를 중심으로. p. 120. 〈표 4-20〉.

표준편차는 933만 1천 원으로 보정 전과 행정보완값에 비해 작았다.

4. 측정오차를 줄이기 위한 조사 관리 방안

측정오차로 인한 편향을 줄이기 위해 여러 가지 사후 보정 방법을 살펴보고 실제로 적용해 보았다. 사후적인 보정도 중요하지만, 사후 보정에 앞서 근본적으로 조사 설계 및 진행, 조사 참여자의 응답, 데이터 처리 과정에서의 오차 발생 가능성을 낮추려는 노력이 필요하다. Kasprzyk(2005)가 측정오차의 원인을 네 가지, 즉 설문 내용, 자료 수집 방법, 조사원, 응답자로 구분하였는데, 이에 따른 조사 관리 방안을 〈표 10〉과 같이 제안하고자 한다.

첫째, 설문 구성을 할 때 사전·사후 정성 조사를 실시하여 설문 문구 수정, 사후 결과 해석 등에 활용하는 것이다. 현재 예산이 크거나 중요한 조사는 사전 조사 또는 예비 조사를 하여 반영하는 편이다. 그러나 지나치게 많은 설문 문항, 어려운 설문 문구, 정확한 응답을 꺼릴 만한 설문(소득, 자산, 부채 등), 응답자도 정확히 알지 못

표 10. 측정오차 관리 방안

구분	관리 방안
설문 내용	- 설문 구성을 할 때 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	- 조사원 교육 강화 - 조사원 교육 후 테스트 - 조사 초기 리뷰 강화로 조사원에 대한 피드백 제공 - 조사에 대한 조사원 태도를 사후에 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: 이혜정, 신지영, 박승환, 지희정, 오미애. (2021). 조사 자료의 품질 검증 연구– 측정오차를 중심으로. p. 145. 〈표 5-1〉.

하는 설문(가구원의 소득, 지출, 부동산의 현재 가격, 가구 전체의 생활비, 가구원 개인별 생활비 등) 등은 측정오차의 발생 원인으로 꼽을 수 있다. 이를 위해 과감한 설문 축소, 설문 문항을 응답자가 이해하기 쉽게 풀어 쓴 표현, 종속형 설문 및 Event History Calendar 도입 등을 고려해 볼 수 있다.

둘째, 자료 수집 방법에서의 해결 방안은 리뷰, 검증 등을 통한 사후 자료 확인 및 재조사이다. 리뷰, 검증 등으로 조사원에 의한 오류나 거짓 자료 수집 등을 확인할 수 있다. 보통 조사 과정에서는 조사원이 응답 내용을 1차 점검하고 그 조사 내용을 감독관이 다시 한번 리뷰(검토)하는 절차가 있다. 이렇게 조사를 완료한 자료에 대해서는 부분적으로 전화 검증 등을 한 후, 필요하다면 폐기 및 대체 혹은 재조사를 한다. 통계청의 경우 내검이라 하여 내부에서 조사 내용을 점검하고 확인하여 보완하는 절차가 있으며, 무응답에 대해서는 사후에 통계적으로 보정하는 절차로 진행하는 것으로 알려져 있다. 그런데 대부분의 조사에서는 리뷰 또는 검증에 대한 적정한 비용 산정이 이루어지지 않고 있으며, 조사에 충분한 시간을 반영하지 못하고 있다. 리뷰나 검증의 중요성을 강조하고 충분한 시간과 비용을 반영할 필요가 있다고 생각한다.

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

생각한다.

요즘에는 CATI, CAPI, CAWI 등[4]을 활용한 조사가 많은 편인데, 이러한 조사 도구를 사용하면 조사 과정에서 발생하는 모든 관련 정보인 Paradata를 수집할 수 있다. 예를 들어 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. 한국의료패널조사는 의료 이용 영수증을 수집하여 의료비를 확인하고 있다(김남순 외, 2018, p. 20). 한국재정패널조사에서도 근로소득세를 내거나 종합소득세 확정 신고를 한 사람에 대해서는 소득공제 현황과 근로소득 연말정산을 신청하기 위해 회사에 제출했던 영수증으로 소득을 확인(영수증 제출을 동의한 사람에 한정함)한다(한국조세재정연구원, 2020, p. 68). 에너지소비실태조사는 전년도 에너지소비량을 기억하지 못하는 가구 중 희망 가구에 한해 고객 번호를 조사하고 추후 공급사에 소비량(전력, 도시가스 등의 네트워크에너지)을 조회하여 정보를 얻는다(에너지경제연구원, 2018, p. 4). 장기적으로는 정확한 응답을 꺼리거나 응답자도 알 수 없는 소득이나 지출 등은 한국신용정보원의 전산 자료 또는 국세청과 같은 행정 자료를 다양하게 활용하여 연계해 볼 수 있다. 현재 개인정보 보호 문제로 한계가 있으나, 응답자의 동의를 얻은 후 자료를 활용하는 것 등과 같은 해결 방안을 모색할 필요가 있다. 이는 응답의 부담을 덜어 줄 뿐만 아니라 응답값에 대한 정확도를 향상하는 데도 기여할 것이다.

셋째, 조사원의 교육 강화이다. 다양한 실태조사에서 조사원이 설문을 그대로 읽어 준다면 응답자는 잘못 응답할 가능성이 커질 것이다. 이에 조사원을 대상으로 심층 교육을 시행하고 있다. 하지만 조사원 교육만으로는 이해도를 충분히 높이기 어려운 경우가 많다. 이를 감안하여 교육 후 테스트, 조사 초기 리뷰 강화로 조사원에 대한 피드백 제공, 조사에 대한 조사원 태도를 사후에 파악하여 피드백 제공 등을 할 수 있다.

마지막으로, 응답자에 대한 것이다. 현재 응답자 인센티브는 물질적 인센티브(선물, 사례비, 경품 등)가 대부분이다. 그러나 응답 유인 효과를 높일 만큼 충분한 사례가 이루어지지 않는다는 한계가 있다. 물질적 인센티브에 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 조사 취지 등을 홍보하는 방안이 있을 수 있다. 또한

4) CATI(Computer Assisted Telephone Interviewing)는 컴퓨터를 이용한 전화 조사 방법이고, CAPI(Computer Assisted Personal Interviewing)는 컴퓨터를 이용한 면접 조사 방법이며 CAWI(Computer Assisted Web Interviewing)는 컴퓨터를 이용한 인터넷(웹) 조사 방법을 의미함.

조사 참여자에 대한 제도적 인센티브 방안도 모색해 볼 수 있다. 예를 들면 세금 감면, 전기료 인하, 학생들에 대한 봉사활동 점수 제공 등이 있을 수 있다.

5. 나가며

2017년 가계금융복지조사 자료에서 가구주의 개인 근로소득에 대해 응답값에 측정오차를 포함하고 있는지 살펴보았고, 측정오차 보정을 통한 회귀계수 추정, 히핑 보정 방법을 사용하여 측정오차를 보정하였다.

근로소득에 대한 측정오차의 구조는 차이가 있는 측정오차였다. 차이가 있는 측정오차는 응답값이 관심 설명 변수에 따라 행정보완값에 대해 체계적인 경우를 의미하며, 이에 대한 측정오차 보정 방법을 사용하여 회귀계수를 추정한바, 관심 설명 변수의 회귀계수는 모두 행정보완값의 회귀계수와 정확하게 일치하는 것을 확인하였다. 그러나 대부분의 자료에서는 참값을 알고 있는 경우가 드물어서 일반 연구자는 참값을 모르고 있을 가능성이 크다는 측면에서 봤을 때 유용한 방법이라고 할 수 있다. 한편 응답값과 행정보완값 간 차이가 크고 연관성이 낮을수록 표준오차는 커지는 것으로 나타났다.

차이가 있는 측정오차 보정을 통한 회귀계수 추정 분석의 한계는 설명 변수 형태를 이분형 범주에만 사용 가능하다는 점, 측정오차에 영향을 미치는 설명 변수에 대한 회귀모형 분석만 가능하여 더 많은 관심 설명 변수의 영향을 파악할 수 없다는 점을 들 수 있다. 그러나 여러 설명 변수가 함께 오차 없는 측정값과 연관되는 것을 현실적으로 모형화하기 어렵기 때문에 이 글의 분석과 같이 하나의 변수에 국한하는 경향이 있다고 할 수 있다.

다음으로 근로소득에 대해 히핑 보정을 하였는데, 보정 후 결과는 전반적으로 부드러운 분포함수 형태를 보였다. 히핑 보정 후 평균도 행정보완값과의 차이가 작았다. 히핑 보정 전 평균이 행정보완값과의 차이가 더 큰 편이었다. 그러나 히핑 보정 후 표준편차가 행정보완값보다 작게 나타나므로, 내부보정 또는 외부보정 자료를 추가 정보로 사용하여 히핑 보정을 하는 것을 고려해 볼 수 있다.

관심 변수에 대한 측정오차의 포함 여부는 그림(plot)을 통한 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 비율, 특정 배수의 형태를 보이는지에 대한 분석, 그림을 통해 확인할 수 있다.

측정오차 보정 방법으로는 측정오차 보정을 통한 회귀계수 추정, 커널 함수 추정법을 활용한 히핑 보정 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 통계패

키지의 'mecor' 패키지 내부 'mecor' 함수를 사용하면 된다. 커널 함수 추정법을 활용한 히핑 보정은 R 통계패키지의 'Kernelheaping' 패키지 내부 'dheaping' 함수이다. 이 외에 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 있으므로 유용하다고 생각한다.

또한 측정오차로 인한 편향을 줄이기 위해 사후적인 보정에 앞서 근본적으로 조사 설계 및 진행, 조사 참여자의 응답, 자료 처리 과정에서의 오차 발생 가능성을 낮추려는 노력이 필요하다. 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 조사의 모든 단계에서 철저하게 관리되어야 할 것이다.

참고문헌

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

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

이혜정, 신지영, 박승환, 지희정, 오미애. (2021). **조사 자료의 품질 검증 연구- 측정오차를 중심으로.** 세종: 한국보건사회연구원.

통계청. (2018). **2017년 가계금융·복지조사.** 대전: 통계청

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

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

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

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

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.

Measurement Error Correction and Management in Survey Data: Focusing on Earned Income

Lee, Hyejung
(Korea Instiute for Health and Social Affairs)

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, which can occur from various causes and situations.

This study analyzes the measurement errors that occurr in sample survey data, and proposes two methods(measurement error correction in the linear regression models with continuous outcomes and kernel density estimation for heaped data) for correcting them. In addition, this study proposes a survey data management method that allows production of high-quality data.