발 간 등 록 번 호
11-1240245-000057-14

2023년 연구보고서

비확률표본을 위한 통계적 추론: 실증연구

2024. 4.

http://kostat.go.kr/sri

ISSN 2288-1166(Print)
ISSN 2733-4120(Online)

통계청
통계개발원

연구보고서 2023-11

비확률표본을 위한 통계적 추론: 실증연구

권순필 · 정희영 · 권영미 · 김승환

발간사

기업 경영, 개인의 일상에 이르기까지 합리적 의사결정의 근간인 통계에 대한 중요성이 점점 커지고 활용범위도 넓어지고 있으며, 특히 국가통계는 정책결정에 필수적으로 활용되면서 그 중요성이 더욱 증대되고 있습니다.

이러한 시대의 변화에 따라 통계청은 빅데이터의 활용, 조사자료와 행정자료 간의 연계 등과 같은 통계생산방식의 혁신을 통해서 응답자 부담은 최소화하면서 동시에 보다 정확하고 사용자 친화적인 통계를 만들고자 끊임없이 노력하고 있습니다.

통계개발원은 국가통계의 중추를 담당하는 통계청의 싱크탱크로써 전략적인 연구를 수행하고 있는 국내의 유일한 「국가통계 전문연구기관」 입니다. 2006년에 설립된 이래 기존의 조사통계를 보다 효율적으로 작성하기 위한 각종 기법과 관련된 통계방법론적 연구는 물론 데이터에 기반한 국가정책이 수립될 수 있도록 경제·사회현상에 대한 심층 분석 연구를 강화하고 있습니다.

또한 저출산·고령사회 현상 등으로 인해 대내외적으로 관심이 높아지고 있는 인구집단 및 인구동향에 관한 분석연구 및 인구동태 관련 방법론 연구를 밀도 깊게 수행하고 있습니다. 이러한 연구의 구체적인 결과를 중심으로 통계개발원은 「2023년도 연구보고서」를 발간하게 되었습니다.

이번「2023년도 연구보고서」에는 AI 통계분류 결과분석 및 실무활용성 제고방안 연구 등 데이터과학 연구, 2025년도 인구주택총조사 등 조사표 개선 연구, 경제·사회·환경 변화를 반영한 인구통계, 격자통계를 활용한 도시화 현상 분석 등 인구통계 연구, 인구감소지역과 생활밀접업종 관계 분석 등 경제통계 연구, 비확률표본을 위한 통계적 추론 등 국가통계 방법론 연구, 위성영상을 활용한 국토그린지표 개발 기초연구 등 SDG 지표 관련 연구 등을 수록하고 있습니다.

본 연구보고서는 통계개발원이 전년에 국가통계 개선·개발을 위해 수행한 연구과제로서 국가통계 생산자의 통계개발 및 개선에 유용한 자료로 활용되고 될 수 있기를 기대합니다. 앞으로도 통계개발원이 “국가통계 전문연구기관”으로서 대내외적으로 선도적인 역할을 할 수 있도록 독자 여러분의 지속적인 관심을 부탁드립니다.

통계개발원은 본 연구보고서가 데이터 이용자의 통계 활용에 도움이 되고, 통계 작성자의 통계 개발 및 개선에 유용한 자료로 활용될 수 있기를 기대합니다. 앞으로도 국가통계의 통계연구에 대한 독자 여러분의 지속적인 관심을 부탁드립니다. 아울러 실용적이고 품질 높은 연구 결과를 도출하기 위해 최선을 다한 연구진에게 따스한 감사를 전합니다.

2024년 4월

통계개발원장

목 차

요 약

다양한 출처의 비확률표본 증가와 자료처리를 위한 IT 기술의 발전은 확률표본 기반 통계 생산 패러다임의 변화를 요구하고 있다. 표본추출틀의 포함범위 감소, 무응답 및 조사비용의 증가, 코로나19와 같은 악화된 조사 환경 때문에 확률표본의 선택과 유지가 어렵기 때문이다. 그러나 유한모집단 추론을 위해서는 비확률표본의 선택편향, 과소포함, 미지의 추출확률 문제가 선결되어야 한다. 이를 위해서는 비확률표본과 고품질 참조확률표본의 통합(data integration) 및 두 자료를 연결하는 모형 식별이 필수적이다. 식별된 모형으로 확률표본의 비편향성을 빌려("borrow unbiasedness") 올 수 있기 때문이다.

본 연구는 비확률표본의 선택편향을 감소시키면서 미지의 추출확률 문제를 해결하기 위해 성향점수 가중치 방식(ipw), 캘리브레이션 가중치 방식(cal), 통대체 방식(reg), 이중강건 방식(dr)의 4가지 추정량을 검토한다. 각 추정량의 분산은 붓스트랩으로 추정한다.

모의실험을 위해 2021년 가계금융복지조사 공공용 마스터 자료를 모집단으로 사용한다. 관심값은 연간가구경상소득의 평균이며, 보조변수는 가구주의 인구특성과 가구원수이다. 확률표본과 비확률표본의 크기, 확률표본의 설계 등을 변화시키는 다양한 시나리오를 가정하고, 각 시나리오별로 비확률표본의 평균과 관련 신뢰구간을 추정한다.

4개의 평균 추정량 모두에서 상대편향 및 평균제곱오차가 감소한다. 추정량의 95% 신뢰구간의 모평균 포함확률은 80% 전후로 나타났다. 4개 추정량 중에는 이중강건 추정량과 캘리브레이션 추정량이 가장 안정적인 추정결과를 보여준다. 반면, 비확률표본을 단순임의표본인 것처럼 다루는 경우에는 심각한 선택편향 문제가 발생한다.

모의실험을 통해 가구의 연간경상소득처럼 변동이 큰 관심변수에 대해서도 적절한 보조변수를 사용한다면 안정적인 선택편향 보정이 가능하다는 것을 확인하였기 때문에 다양한 조사와 관심변수에 확대 적용이 가능할 것으로 기대한다.

주요 용어 : 비확률표본, 확률표본, 비편향성 빌려오기, 선택편향, 성향점수, 캘리브레이션, 통대체, 이중강건

제 1 장

서 론

제1절 연구배경

확률표본(probability sample)이란 잘 알려진 확률표집 이론에 따라 선정된 표본을 말한다. 확률표집은 고유한 이론적 체계 안에서 표본의 대표성과 결과의 객관성 확보가 가능하기 때문에 통계청, 국책연구기관 등 공식통계 생산기관에서 선호한다. 그러나 확률표본은 잘 정비된 표본추출틀과 정교한 표집설계, 표집설계에 의한 표본추출과 완전한 응답이 전제되므로 상당히 고비용이다.

최근에는 표본추출틀의 포함범위 감소 및 무응답 증가, 조사비용의 급격한 증가로 확률표본의 선택과 유지에 어려움이 발생하고 있다. 코로나19와 같은 예측하기 어려운 조사 환경의 변화까지 발생하며 확률표집 기반 조사는 시련의 시기를 맞고 있다.

비확률표본(nonprobability sample)이란 확률표본이 아닌 표본 혹은 데이터를 말하며, 미지의 생성 메커니즘 때문에 일반적으로는 목표모집단을 대표하지 않는다. 조사의 목적에 맞도록 설계(survey designed)되었지만 확률적으로 추출되지는 않은 할당표본, 편의표본, 웹표본 등이 전통적인 비확률표본이다. 그러나 근래에는 행정자료, 거래자료, 센서자료 및 인터넷자료와 같이 생성 메커니즘이 알려지지 않은 빅데이터들도 비확률표본으로 분류하고 있다. 한계에 봉착한 확률표본의 대안으로 저비용, 낮은 응답부담에 실시간으로 쏟아지는 대량의 비확률표본에 대한 활용 요구가 증가하고 있는데, 비확률표본이 모집단 추론을 위한 원천이 될 수 있을지 다시 관심이 집중되고 있다.

모든 확률표집은 확률표본 설계(sample design)와 근사적으로 설계에 비편향인 추정량의 올바른 조합, 견고한 수학적 이론에 기반한다. 이를 통해 비교적 작은 규모의 표본으로 대규모 유한모집단(finite population)에 대한 유효한 통계적 추론을 제공하고 있다(Castro-Martin 등, 2020).

미국여론조사협회의 2013년 “비확률표집에 대한 특별 보고서”[1]에 따르면, ⅰ)확률표집과 달리 모든 비확률표집을 적절하게 포함하는 단일 체계는 없다. ⅱ)확률조사든

1) Baker 등(2013), Report of the AAPOR task force on non-probability sampling.

비확률조사든 추론을 위해서는 모형화를 위한 가정에 어느 정도 의존해야 한다. ⅲ)비확률표본이 조사연구자 사이에서 더 폭넓게 수용되려면 품질을 평가하기 위한 보다 일관된 체계와 수반되는 일련의 측정이 있어야 한다.

비확률표본의 선택편향(selection bias), 과소포함(under coverage), 미지의 추출확률(unknown inclusion probability) 등의 문제 때문에 확률표집 이론을 직접 적용하는 것은 불가능하다. 그러나 비확률표본을 단순임의표본인 것처럼 다루는 경우 심각한 표본 선택편향이 발생한다. 선택편향 조정이 없으면 데이터가 클수록 더 확실하게 우리 자신을 속이게 되는 빅데이터의 역설[2]에 빠질 수밖에 없다.

비확률표본의 선택편향 조정을 위해서는 신뢰할 만한 다른 출처 자료의 활용이 필수적이며 다른 출처 자료의 대표적인 사례는 고품질 확률표본이다. 고품질 참조확률표본으로부터 얻어진 가장 최신의 정보를 비확률표본에 결합하는 것을 데이터 통합(data integration)이라고 볼 수 있다(Kim, 2022a). 이를 통해 비확률표본은 확률표본의 비편향성을 빌려와[3] 미지의 추출확률을 추정하게 된다.

통계개발원은 2022년에 “비확률표본을 위한 통계적 추론” 연구를 통해 유한모집단 추론에서 비확률표본이 재조명받게 된 이유와 선택편향 조정을 위해 비확률표본과 참조확률표본을 통합하는 모형기반 접근 등 방법론에 대한 종합적인 검토를 수행하였다. 어떤 조건하에서 비확률표본을 확률표본의 대안으로 사용할 수 있는지 등을 확인하였으며, 아직까지는 비확률표집을 적절하게 포함하는 단일 체계가 없기 때문에 경험적 연구에 의존할 수밖에 없다는 결론에 이르렀다.

본 연구는 통계청의 실제 조사자료를 활용해 모집단의 평균을 추정하는 모의실험을 수행했고 이를 통해 이론이 현실에서 어떻게 적용될 수 있는지 경험적인 결과를 보여준다.

제2절 연구내용

2장에서는 비확률표본 추정 방안을 소개한다. 이를 위해 필요한 설정과 가정을 기술하고, 성향점수 가중치 방식, 캘리브레이션 가중치 방식, 통대체 방식, 이중강건 방식에 관해 기술한다.

3장에서는 가계금융복지조사 자료를 이용한 모의실험을 통해 실증연구를 수행한다. 가계금융복지조사는 표본의 규모가 크면서 연속형, 이산형 등 다양한 형태의 변

2) Big data paradox: the bigger the data, the surer we fool ourselves. (Meng, 2018)

3) borrow unbiasedness from probability sample.

수를 가지고 있어 모의실험에 적합하다고 판단하였다. 모의실험을 위해 확률표본과 비확률표본의 상대적 크기 변화, 확률표본의 설계 변화 등 다양한 시나리오를 가정하였고, 시나리오별로 비확률표본의 평균과 관련 신뢰구간을 추정해 보았다. 이 같은 비확률표본을 위한 통계적 추론 과정을 통해 미국여론조사협회의 지적을 경험적으로 확인해 보려고 한다.

마지막으로 4장에서는 연구결과를 요약하고 시사점 및 결론을 제시한다.

본 연구는 조사목적에 맞게 설계된 비확률표본을 주요 관심 자료로 한다. 빅데이터와 같은 구조적 데이터는 추론 이전에 데이터 품질 이슈가 먼저 해소되어야 하기 때문이다.

제 2 장

추정방안

제1절 가정과 설정

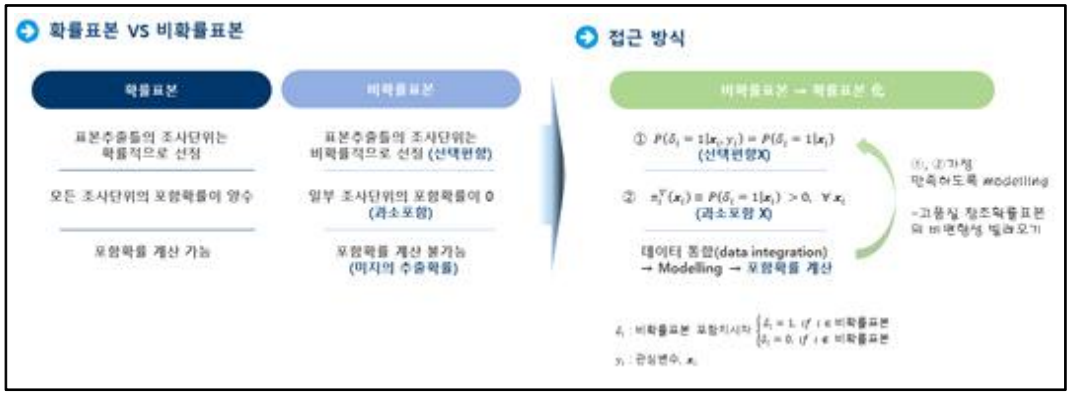
확률표본은 표본추출틀의 조사단위가 확률적으로 선정되며, 모든 조사단위의 포함확률은 양수여야 하고, 포함확률은 계산이 가능해야 한다. 반면, 비확률표본은 표본추출틀의 조사단위가 확률적으로 선정되지 않거나, 일부 조사단위의 포함확률이 0이거나, 포함확률 계산이 불가능한 경우를 말한다. 비확률표본의 이런 특성은 순서대로 선택편향, 과소포함, 미지의 추출확률로 바꿔서 표현할 수 있다.

본 연구에서는 비확률표본을 확률표본처럼 취급하여 확률표본 추론 체계를 사용하려고 한다(<그림 2-1>). 확률표집의 고유한 이론적 체계는 표집기법에 따라 이미 알려진 포함확률 π_k로부터 시작한다. 모집단 내 모든 단위의 π_k는 0보다 크고, 즉, $\pi_k > 0,\ k = 1,...,N$ 이다. π_k를 이용해 총계, 평균, 비율과 같은 모집단 특성에 대한 설계비편향(design unbiased) 추정량을 산출한다. 추론의 품질 지표로는 사전에 정의된 정확도와 함께 모집단 특성의 비편향 추정을 목표로 하여, 평균의 제곱과 분산의 합으로 구성된 평균제곱오차(Mean Square Error; MSE)를 이용한다. 이 같은 확률표본의 추론 체계를 사용하기 위해서는 비확률표본의 선택편향, 과소포함, 미지의 추출확률 문제가 해결되어야 한다.

선택편향은 Rosenbaum과 Rubin(1983)의 강한 무시가능성(ignorability) 혹은 무작위결측(Missing At Random; MAR) 가정으로 해소할 수 있다. 즉, 공변량 $\boldsymbol{x}$가 주어졌을 때, 관심값 y와 비확률표본에 포함 여부 δ가 서로 독립이기 때문에 y가 선택편향에 영향을 받지 않는다고 가정한다. 과소포함은 모든 단위 i가 비확률표본에 포함될 확률 π_i^V가 0보다 크다는 가정으로 해소한다. 이를 수식으로 표현하면 다음과 같다(Kim, 2022a).

- $P(\delta_i = 1|\boldsymbol{x}_i, y_i) = P(\delta_i = 1|\boldsymbol{x}_i)$
- $\pi_i^V(\boldsymbol{x}_i) \equiv P(\delta_i = 1|\boldsymbol{x}_i) > 0, \quad \forall \boldsymbol{x}_i$

여기서, δ_i는 비확률표본 지시변수로 비확률표본에 포함되면 1, 그렇지 않으면 0이다.



<그림 2-1> 비확률표본의 확률표본화를 통한 추론

비확률표본의 선택확률과 과소포함 문제를 가정으로 해결하고 나면 미지의 추출확률 문제만 남는다. 이는 통계적 모형을 통해 해소한다. 이렇게 되면 비확률표본의 추정 결과에 대해 확률표본처럼 MSE와 신뢰구간 같은 지표로 품질평가가 가능해진다. 이 같은 추론 과정의 수립이 현재까지는 미국여론조사협회의 2013년 보고에 대한 가장 일반적인 대응 방안이 될 수 있을 것이다.

물론 이 가정들은 상당히 강력하며, 검증이 거의 불가능하다. 그러나 조사를 통한 추론에서는 다양한 가정이 필수불가결하다. 확률기반 조사 역시 표본추출틀의 완비, 완전한 응답 등이 전제되며 과소포함, 무응답 등이 있는 경우 조사에서 관찰된 단위와 관찰되지 않은 단위를 연결하는 모형 식별(specify)이 필요한 것과 같다(Mercer 등, 2017).

본 연구에서는 미지의 추출확률을 계산하는 통계적 모형으로 성향점수모형, 예측모형, 이중강건모형, 캘리브레이션 방법을 적용하였고, 분산은 붓스트랩 방법(bootstrap method)으로 추정하였다. 이 과정에서 미지의 추출확률은 확률표본의 선택편향을 없애려는 방향으로 계산된다. 본 연구에서 비확률표본으로 추론한다는 것은 비확률표본이 모집단을 대표할 수 있도록 선택편향을 제거하고, 관련 신뢰구간을 제시한다는 것과 같다.

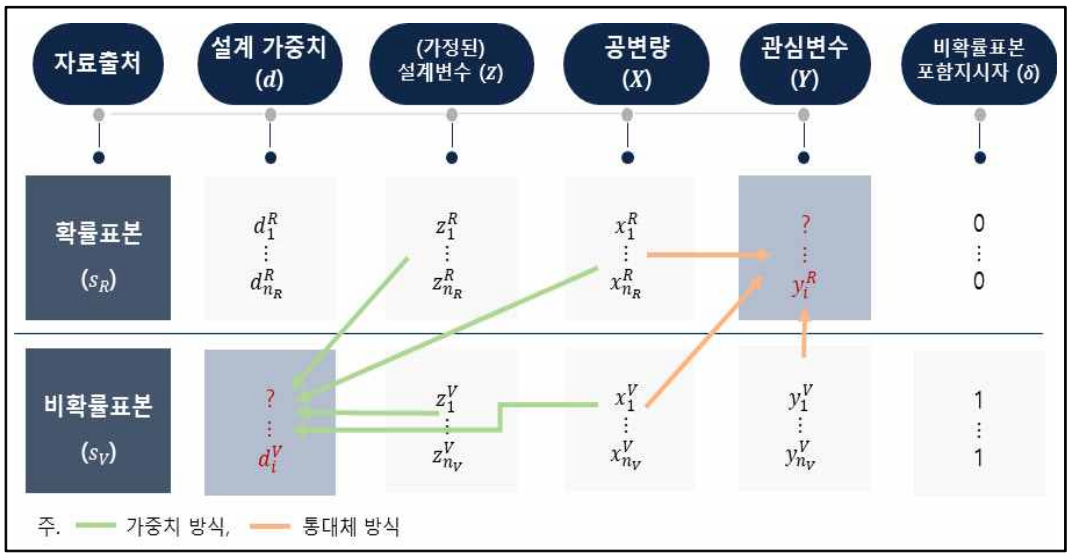
비확률표본의 추출확률을 계산하기 위해서는 보조변수 $\boldsymbol{x}$를 갖는 신뢰할 만한 참조확률표본과 비확률표본의 통합이 필수적이다. 확률표본의 비편향성에 기대 비확률표본의 선택편향을 조정할 것이기 때문이다.

본 연구는 관심변수는 비확률표본에서만 관측되고, 비확률표본과 공통의 공변량을 갖는 유용한 확률표본이 존재하는 상황에 대해서 수행된다. 확률표본과 비확률표본은 동일한 모집단을 대표하며, 두 표본은 서로 독립이고 자료 간 측정오차는 없다고 가정한다. 이는 두 표본에 중복되는 단위가 없어야 한다는 Valliant와 Dever(2011) 등의 연구보다는 완화된 조건이며, 더 현실적인 가정이기도 하다.

실험 환경을 수식으로 표기하면 다음과 같으며, 비확률표본의 선택편향 조정을 위한 확률표본과 비확률표본 간의 관계 및 추정 방식은 <그림 2-2>와 같이 표현될 수 있다.

- $U=\{1,2,...,N\}$: 크기 N 인 모집단
- y_i : 관심변수, $\boldsymbol{x}_i$: 보조변수(공변량), $\boldsymbol{z}_i$: 설계변수, $i=1,2,...,N$
- $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$: 관심변수의 유한모집단 평균(모평균)
- s_V : 크기 n_V로 $\{(\boldsymbol{x}_i, y_i),\ i \in s_V\}$ 자료를 갖는 비확률표본
- s_R : 크기 n_R로 $\{(\boldsymbol{x}_i, d_i^R),\ i \in s_R\}$ 자료를 갖는 확률표본으로, $d_i^R = 1/\pi_i^R$, 여기서 π_i^R 는 s_R의 포함확률
- δ_i : 비확률표본 지시변수 $\begin{cases}\delta_i = 1, & \text{if } i \in s_V \\ \delta_i = 0, & \text{if } i \notin s_V\end{cases}$, $i=1,2,...,N$

비확률표본 첨자는 자원자표본(Volunteer sample)을 나타내는 V로, 확률표본 첨자는 참조표본(Reference sample)을 나타내는 R로 표기하였다.



<그림 2-2> 비확률표본 추론 접근 방식의 도식적 표현

<그림 2-2>와 같은 설정에서 데이터 통합은 결국 결측 문제처럼 인식할 수 있다. 비확률표본의 미지의 가중치 d_i^V 혹은 확률표본의 관심변수 y_i^R을 결측으로 보는 것이다. 이와 같은 접근은 확률표집에서 무응답을 처리하는 방법과 유사하다.

<그림 2-2>에서 가중치 방식은 비확률표본 s_V의 미지의 추출확률 $\hat{\pi}_i^V$를 추정하여 이를 비확률표본 s_V의 포함확률인 것처럼 취급하는 성향점수 방식과 가중치 d_i^V를 직접 확률표본의 알려진 주변분포에 맞추는 캘리브레이션 방식이 있다. 통대체 방식은 비확률표본에서 식별된 모형으로 추정한 $\hat{y}_i^R$을 통째로 확률표본 s_R에 대체하여 확률표본의 관심값이 관측된 것처럼 추정하는 방식이다. 이중강건 방식은 성향점수 가중치 방식과 통대체 방식의 결합을 통해 각 방식의 모형 오식별(model misspecification)에 보다 강건하게 대응하는 추정 방식이다.

제2절 성향점수 가중치 방식(Propensity Score Weighting Method)

성향점수모형은 우리가 알지는 못하지만, 비확률표본이 실제로는 확률표집 메커니즘을 가진다는 가정하에, 확률표본을 참조하여 비확률표본의 의사포함확률(pseudo-inclusion probability)을 추정하는 유사확률화 접근(Quasi-randomization approach)의 일환이다. 성향점수모형을 이용하여 단위 i가 비확률표본에 포함될 의사포함확률 혹은 성향점수 π_i^V을 추정하고, 이를 이용하여 가중치를 산정한다.

성향점수모형은 $\pi_i^V = E(\delta_i|\boldsymbol{x},y_i) = P(\delta_i=1|\boldsymbol{x}_i,y_i),\ i=1,2,...,N$ 이다. 앞에서 가정한 강한 무시가능성에 따라 $\pi_i^V = P(\delta_i=1|\boldsymbol{x}_i),\ i\in s_V$가 되고 성향점수모형은 비확률표본 지시변수 δ_i를 보조정보 $\boldsymbol{x}_i$로 확률화하는 모형으로 정리된다.

성향점수모형의 $\boldsymbol{\theta}$는 다음의 로그우도함수 $l(\boldsymbol{\theta})$를 최대화하는 $\hat{\boldsymbol{\theta}}$으로 구할 수 있다.

$$
\begin{aligned}
l(\boldsymbol{\theta}) &= \sum_{i=1}^{N}\{\delta_i \log\pi_i + (1-\delta_i)\log(1-\pi_i)\} \\
&= \sum_{i\in s_V}\log\left\{\frac{\pi(\boldsymbol{x}_i,\boldsymbol{\theta})}{1-\pi(\boldsymbol{x}_i,\boldsymbol{\theta})}\right\} + \sum_{i=1}^{N}\log\{1-\pi(\boldsymbol{x}_i,\boldsymbol{\theta})\}
\end{aligned}
$$

유한모집단의 모든 단위에 대해 $\boldsymbol{x}_i$를 관찰하지 않기 때문에 $l(\boldsymbol{\theta})$는 실제로 사용할 수 없다. 여기서 $\boldsymbol{x}$ 정보가 포함된 참조확률표본 s_R이 필요하다. $l(\boldsymbol{\theta})$ 대신 의사로그우도함수(pseudo log-likelihood) 혹은 모집단로그우도함수(population log-likelihood) $l^*(\boldsymbol{\theta})$를 최대로 하는 $\hat{\boldsymbol{\theta}}$을 추정할 수 있다(Chen et al., 2020;, Kim, 2022a).

$$l^*(\boldsymbol{\theta}) = \sum_{i \in s_V} \log\left\{ \frac{\pi(\boldsymbol{x}_i,\boldsymbol{\theta})}{1-\pi(\boldsymbol{x}_i,\boldsymbol{\theta})} \right\} + \sum_{i \in s_R} d_i^R \log\left\{1-\pi(\boldsymbol{x}_i,\boldsymbol{\theta})\right\} \tag{2.1}$$

대표적인 성향점수모형인 로지스틱회귀모형(logistic regression model) $\pi_i = \pi(\boldsymbol{x}_i,\boldsymbol{\theta}) = \exp(\boldsymbol{x}_i^T\boldsymbol{\theta}) / \left\{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\theta})\right\}$을 식 (2.1)에 적용하면 식 (2.2)와 같이 정리된다.

$$l^*(\boldsymbol{\theta}) = \sum_{i \in s_V} \boldsymbol{x}_i^T\boldsymbol{\theta} - \sum_{i \in s_R} d_i^R \log\left\{1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\theta}\right)\right\} \tag{2.2}$$

식 (2.2)를 최대로 하기 위해 미분하면, 점수방정식 $U(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l^*(\boldsymbol{\theta}) = \sum_{i \in s_V} \boldsymbol{x}_i - \sum_{i \in s_R} d_i^R \pi(\boldsymbol{x}_i,\boldsymbol{\theta}) \boldsymbol{x}_i$가 되고 $U(\boldsymbol{\theta})$의 해는 다음의 Newton-Raphson 반복 절차로 찾을 수 있다.

- 초기값 $\boldsymbol{\theta}^{(0)} = \mathbf{0}$
- $\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + \left\{H(\boldsymbol{\theta}^{(m)})\right\}^{-1} U(\boldsymbol{\theta}^{(m)})$
- $H(\boldsymbol{\theta}) = \sum_{i \in s_R} d_i^R \pi(\boldsymbol{x}_i,\boldsymbol{\theta}) \{1-\pi(\boldsymbol{x}_i,\boldsymbol{\theta})\} \boldsymbol{x}_i \boldsymbol{x}_i^T$

성향점수 가중치 방식은 성향점수의 역수를 가중치로 사용하거나, 성향점수로 층을 구성하여 해당 층의 평균 성향점수의 역수를 해당 층의 가중치로 사용하는 방식이 대표적이다. 본 연구에서는 성향점수의 역수를 가중치로 사용하는 역확률가중(Inverse Probability Weighting; IPW) 방식을 적용한다. 즉, 추정된 성향점수가 표본의 포함확률과 같은 역할을 하게 된다.

성향점수 가중치를 이용한 평균의 IPW 추정량은 다음과 같다.

$$\hat{\mu}_{ipw} = \frac{1}{\hat{N}_V} \sum_{i \in s_V} \frac{1}{\hat{\pi}_i^V} y_i = \frac{1}{\hat{N}_V} \sum_{i \in s_V} \hat{d}_i^V y_i$$

여기서 $\hat{N}_V = \sum_{i \in s_V} \hat{d}_i^V$이다.

성향점수모형이 바르게 식별되면 $\hat{\mu}_{ipw}$는 일치추정량이다. 그러나 성향점수모형이 잘못 식별된 경우, 특히 특정 단위가 $\hat{\pi}_i^V$에서 매우 작은 값을 가질 때는 민감하다. 성향점수가 올바르게 식별된 경우에도 참조확률표본과 관련된 가중치를 사용하지 않는 IPW추정량은 편향되며, 성향점수가 관심변수와 관련성이 높으면 성향모형은 편향을 바로잡지 않는다(Valliant와 Dever, 2011).

제3절 캘리브레이션 가중치 방식(Calibration Weighting Method)

모집단에 대한 신뢰할 수 있는 보조정보가 설계 단계에서 사용되진 않았으나 외부로부터 주어진 경우에는 이를 이용하여 비편향추정량(일반적으로 Horvitz-Thompson; HT추정량)보다 효율적인 추정량을 정의할 수 있다.

캘리브레이션 가중치 방식은 주로 설계가중치가 있는 표본을 사후적으로 보정하는 데 이용되는 방식으로 비확률표본의 가중치 d_i^V를 모집단의 알려진 보조변수의 총계에 맞도록 직접 보정한다. 본 연구에서는 모집단 대신 신뢰할 수 있는 참조확률표본의 보조변수의 추정 총계에 맞도록 비확률표본의 가중치를 조정했다. 센서스와 같은 모집단 자료는 조사 오차는 없지만 대체로 시의성이 떨어지며 가용 보조변수가 제한적이기 때문이다.[4] 이 경우 캘리브레이션 가중치 $\hat{d}_{i,cal}^V$은 보정제약식

$$\sum_{i\in s_V}\hat{d}_{i,cal}^V\boldsymbol{x}_i=\sum_{i\in s_R}d_i^R\boldsymbol{x}_i\ \left(\text{혹은}=\sum_{i=1}^N\boldsymbol{x}_i\right)$$

을 만족하면서 가능한 원래의 초기가중치 d_i^V와 d_i^R 간 거리가 가장 가까운 수치로 정해진다. 이때 사용되는 거리함수 $F\left(d_i^V/d_i^R\right)$의 형태에 따라 추정량은 일반화회귀추정량(Generalized REGression; GREG), 레이킹비 추정량(Raking ratio) 등 다양한 형태로 나타난다(Devill과 Särndal, 1992). 캘리브레이션 가중치는 보정제약식을 만족하면서 가능한 원래의 초기가중치에 가깝게 설정되어야 하기 때문에 Newton-Raphson 같은 반복 알고리즘이 필요할 수도 있다. 표본의 크기가 충분히 커지면 모든 캘리브레이션 추정량들이 GREG와 근사적으로 동일하고, 분산 역시 그러하다.

---

4) 알려진 모집단 정보가 있는 상황과 비교하여 확률표본의 정보를 사용하면 예상할 수 있듯이 효율성 손실이 수반된다. 그러나 모집단 정보를 사용할 수 없을 때 실제적인 어려움을 해결할 수는 있다(Zhang, 2019).

본 연구에서는 캘리브레이션을 위해 보조변수가 범주형일 때 사용되는 레이킹비 추정량 방법을 적용한다. 레이킹비 추정량의 거리함수는 다음과 같다.

$$F(x)=\begin{cases} x\ln(x)-x+1, & \text{if } x\in(0,\infty) \\ -x+1, & \text{if } x=0 \end{cases}$$

레이킹비 가중은 반복비례적합(iterative proportional fitting) 혹은 림가중(Rim weighting)이라고도 부르며 실무에서 자주 사용되는 방법이다.[5] 실제 조사에서는 인구통계 등과 관련된 범주형 보조정보의 획득이 용이하고 방법의 이해와 적용이 쉽기 때문이다. 캘리브레이션 가중치를 이용한 평균 추정량은 다음과 같다.

$$\hat{\mu}_{cal}=\frac{1}{\hat{N}_V}\sum_{i\in s_V}\hat{d}_{i,cal}^{V}y_i$$

여기서 $\hat{N}_V=\sum_{i\in s_V}\hat{d}_{i,cal}^{V}$이며, 초기가중치는 $d_i^V=1$ 혹은 $d_i^V=N/n_V$으로 설정하였다.

캘리브레이션 추정량들은 거리함수가 일정 조건을 만족하는 경우 거리함수의 형태와 관계없이 점근적으로 설계일치성을 갖는다(Devill과 Särndal, 1992).

제4절 통대체 방식(Mass Imputation Method)

통대체 방식은 표본과 비표본(non-sample)이 모두 동일한 모형을 따른다는 초모집단모형(Super-population model) 가정하에, 비확률표본을 훈련 데이터로 사용하여 확률표본의 관심변수를 통째로 추정하는 방식이다. 예측모형(prediction model), 결과모형(outcome model) 접근법이라고도 한다(Chen 등, 2020; Elliott과 Valliant, 2017). 비확률표본이 목표모집단을 대표할 필요는 없지만 변수 간 관계는 표본과 비표본 간에 동일해야 한다.

유한모집단 $\{(\boldsymbol{x}_i,y_i),\ i\in U\}$가 모형 $y_i=m(\boldsymbol{x}_i)+\epsilon_i$, $i=1,2,...,N$의 무작위 표본이

5) 반복비례적합은 Deming과 Stephan(1940)이 표본설계 시 사용되지 않은 보조정보의 주변분포만 알고 있는 경우 표본조사 결과를 모집단 총계와 일치시키도록 제안한 방법이고, 레이킹비 추정량은 Devill과 Särndal(1992)이 제안하고 정의한 캘리브레이션 방법의 일종이다. 제안된 배경은 서로 다르지만, 결과적으로 두 가중치는 서로 일치한다.

라고 가정하면, 예측모형은 $m(\boldsymbol{x}_i)=E(y_i|\boldsymbol{x}_i)$이며, 오차항 ϵ_i는 $E(\epsilon_i)=0$, $V(\epsilon_i)=v(\boldsymbol{x}_i)\sigma^2$을 따른다. 무시가능 조건에 따라 $E(y_i|\boldsymbol{x}_i)=E(y_i|\boldsymbol{x}_i,\delta_i=1)$이 된다.

대표적인 예측모형은 회귀모형 $m(\boldsymbol{x}_i)=\boldsymbol{x}_i^T\boldsymbol{\beta}$ (즉, $\hat{y}_i=\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$), $v(\boldsymbol{x}_i)=1$이다. 모형 매개변수 $\hat{\boldsymbol{\beta}}$는 최소제곱추정(Least Squares Estimation; LSE), 최대우도추정(Maximum Likelihood Estimation; MLE) 등으로 추정할 수 있다.

회귀모형을 이용한 통대체 평균 추정량은 다음과 같다.

$$\hat{\mu}_{reg}=\frac{1}{\hat{N}_R}\sum_{i\in s_R}d_i^R\hat{y}_i$$

여기서 $\hat{N}_R=\sum_{i\in s_R}d_i^R$이다.

예측모형은 특정 관심변수 y에 대한 모형으로 서로 다른 관심변수에 대해서는 서로 다른 모형 설정이 필요하기 때문에 후속 분석에 어려움이 있다. 실제로 관심변수를 잘 설명하는 보조변수를 확보하여 모형을 식별하는 경우에는 다른 접근 방식들에 비해 분산과 편향이 작은 정확한 추정이 가능하며 모평균에 대해 일치추정량이 된다. 그러나 실제에서 설명력이 좋은 보조변수를 확보하는 것은 쉽지 않은 일이다.

제5절 이중강건 방식(Doubly Robust Method)

성향점수를 가중치의 역수로 이용하는 IPW 추정량은 잘못 식별된 성향모형에 민감하기 때문에 예측모형을 통합하여 추정량의 효율성과 견고성을 향상시킬 수 있다. 이를 증강역확률가중(Augmented IPW; AIPW) 추정량이라고 한다. AIPW 추정량은 성향모형과 예측모형 중 하나 이상의 모형이 올바르게 식별되면 일치추정량이기 때문에 이중으로 강건하다고 해서 이중강건(Doubly Robust; DR) 추정량이라고도 한다. 성향점수모형과 예측모형을 통합한 평균의 DR추정량은 다음과 같다.

$$\begin{aligned}\hat{\mu}_{DR}&=\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i\{y_i-m(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}})\}}{\pi(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}})}+\frac{1}{N}\sum_{i=1}^{N}m(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}}),\\&\approx\frac{1}{\hat{N}^V}\sum_{i\in s_V}\hat{d}_i^V\{y_i-m(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}})\}+\frac{1}{\hat{N}^R}\sum_{i\in s_R}d_i^R\hat{y}_i\end{aligned}$$

여기서 $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\theta}}$는 각각 예측모형과 성향점수모형의 일치추정량이다.

DR추정량은 잔차변수 $e_i = y_i - m(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})$를 통해 구성되기 때문에, 예측모형이 비확률표본에 잘 맞는 경우 일반적으로 분산이 더 작다. 그리고 성향점수모형이나 예측모형 중 하나만 일치추정량이면 $\hat{\mu}_{DR}$은 일치추정량이다. 그러나 추론 과정에서 필요한 많은 가정과 모형 타당성에 대한 검증이 어려워 실제 적용은 어려울 것으로 보인다.

가중치 방식은 관심결과에 상관없이 모든 후속 분석에 적용할 수 있는 단일 조정모형 생성이 가능하고, 분위수와 같은 다른 유한모집단 모수로 확장이 간단하며, 관심값이 바뀌어도 동일한 모형을 이용할 수 있다는 장점이 있다. 다만, 모형이 잘못 식별되거나 보조변수의 차원이 커지면 지나치게 큰 가중치나 1보다 작은 가중치를 갖게 될 수 있다. 캘리브레이션 가중치의 레이킹비 방식의 경우 보조변수의 주변분포에 대해 확률표본과 비확률표본 어느 쪽에도 0값이 존재하면 안 되기 때문에 상대적으로 가용한 보조변수의 차원에 제약이 있다.

이중강건 방식에서 보았듯이 두 모형 중 하나가 유지되는 한, 추정량의 일치성을 유지하는 방식으로 예측모형과 선택편향 모형을 모두 결합하는 추정량을 구성하는 것이 가능하다. 이런 추정 접근을 일반적으로 “Doubly Robust”라고 한다. 확률표집에서의 기존 GREG 역시 확률화 메커니즘이 실제로 알려져 있다는 점을 제외하고는 동일한 의미에서 이중으로 강력하다. 따라서 캘리브레이션 가중치 방식과 이중강건 방식의 결과를 비교해 보는 것도 흥미로울 것이다.

제 3 장

모의실험

제1절 설정

모의실험 자료는 2021년 가계금융복지조사 공공용 가구마스터 자료를 사용하였다. 가계금융복지조사 자료를 선택한 이유는 다양한 형태의 연속형과 이산형 변수를 포함하고 있으며 대규모 표본 규모를 가지고 있는 고품질 자료이기 때문이다.

2021년 가계금융복지조사는 18,187개 표본가구를 대상으로 가구원의 성별, 연령, 교육정도, 종사상지위 같은 인구사회적 변수와 가구의 자산, 소득, 부채, 지출과 같은 재무 관련 변수 등을 조사하였으며 공공용 가구마스터 자료는 그 중 총 160개 변수를 마이크로데이터 통합서비스[6]를 통해 제공하고 있다. 본 연구에서는 모의실험을 위해 <표 3-1>과 같은 변수들을 검토하였다.

<표 3-1> 모의실험 검토 변수

변수	내용
1. 가구주성별	남성, 여성
2. 가구주연령	19세~103세
3. 가구주교육정도	초졸이하, 중졸이하, 고졸이하, 대졸이상
4. 가구주혼인상태	미혼, 유배우자, 사별, 이혼
5. 가구주종사상지위	상용임금, 임시일용, 자영업자, 기타
6. 가구원수	1인, 2인, 3인, 4인, 5인이상
7. 수도권여부	수도권, 비수도권
8. 연간가구소비지출	105만원~2억220만원
9. 연간가구비소비지출	0만원~57,351만원
10. 연간가구경상소득	40만원~21억원

6) mdis.kostat.go.kr

관심추정량은 연간가구경상소득(혹은 소득)의 평균이며, 선택편향 조정 방안으로는 성향점수 가중치, 캘리브레이션 가중치, 통대체 방식, 이중강건 방식의 접근을 사용하였다. 분산 추정은 비선형 추정량의 분산 추정량을 구하는 데 있어 유용한 방법인 붓스트랩 추정 방식을 적용하였다.

붓스트랩 표본과 붓스트랩 추정량은 확률표본과 비확률표본으로부터 붓스트랩 표본을 각각 복원단순임의표집(SRSWR; Simple Random Sample With Replacement)으로 뽑고, 각 추정량의 원래 산출 방식과 동일한 절차를 사용하여 계산하였다.

추정량에 대한 평가는 몬테카를로(Monte Carlo; MC) 반복실험을 통해 평균 추정량의 상대편향(Relative Bias; RB), 평균제곱오차(Mean Squared Error; MSE), 95% 신뢰구간의 모평균 포함확률(Coverage Probability; CP)을 사용하였다.

- 상대편향 : $\%RB = \frac{1}{M}\sum_{m=1}^{M}\frac{\hat{\mu}^{(m)}-\mu}{\mu}\times 100$
- 평균제곱오차 : $MSE = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{\mu}^{(m)}-\mu\right)^2$
- 신뢰구간의 모평균 포함확률 : $\%CP = \frac{1}{M}\sum_{m=1}^{M} I\left(\mu \in CI^{(m)}\right)\times 100$

여기서 M은 반복실험 횟수, μ는 모평균, $\hat{\mu}^{(m)}$은 m번째 실험의 평균추정량, $CI^{(m)}$은 m번째 실험의 95% 신뢰구간 $\left[\hat{\mu}^{(m)}-1.96\times\sqrt{\hat{v}^{(m)}}\ ,\ \hat{\mu}^{(m)}+1.96\times\sqrt{\hat{v}^{(m)}}\right]$, $\hat{v}^{(m)}$은 m번째 실험의 분산추정량이다.

모의실험을 통해 경험적으로 얻고자 하는 점을 정리해 보면 다음과 같다.

- 비확률표본의 선택편향 조정 방안들은 실제로 편향을 감소시킬 것이다.
- 확률표본이든 비확률표본이든 규모가 클수록 추정의 효율이 개선될 것이다.
- 동일한 규모의 표본이라면 보다 정밀한 설계에 기반한 확률표본을 참조하는 것이 추정의 효율을 개선시킬 것이다.

이를 위해 모의실험 시나리오를 다음과 같이 수립하였다.

- 확률표본과 비확률표본의 크기를 각각 400, 600, 800, 1000으로 변화시켜 모의실험을 수행하여 $n_R \gg n_V$, $n_R \approx n_V$, $n_R \ll n_V$일 때 추정 효율을 검토한다.
- 확률표본의 설계를 단순임의표본(Simple Random Sample; SRS)과 크기비례확률표집(Probability Proportional to Size; PPS)으로 각각 설정한다. PPS에 사용한 크

기 변수는 비소비지출(exp2)로 최소가중치 대비 최대가중치가 50배를 넘지 않도록 하였다.

모의실험 시나리오는 확률표본 설계 2가지, 확률표본 크기 4가지, 비확률표본 크기 4가지 경우를 모두 고려하여 32개가 실행되었다.

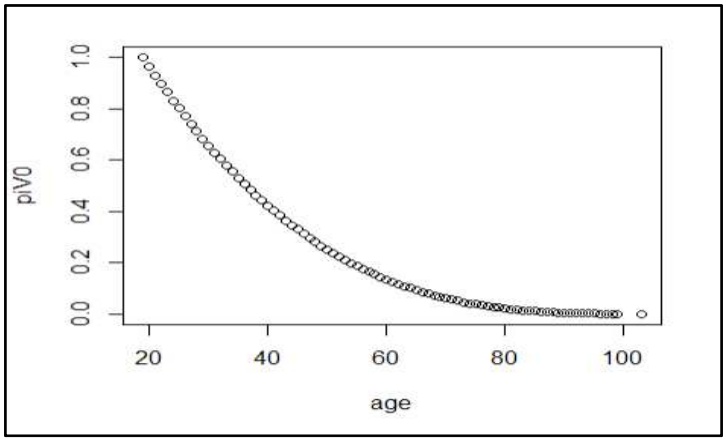
평균 추정량은 비확률표본을 SRS처럼 취급한 단순(naive) 추정량, 성향점수모형을 이용한 역확률가중(ipw) 추정량, 통대체 방식에서 회귀모형(reg) 추정량, 성향점수모형과 회귀모형을 결합한 이중강건(dr) 추정량, 레이킹비 추정량을 이용한 캘리브레이션(cal) 추정량 등 5개의 추정량을 고려하였다. 각 추정량의 형태는 다음과 같다.

- 단순(naive) 추정량 : $\hat{\mu}_{naive} = n_V^{-1} \sum_{i \in s_V} y_i$
- 역확률가중(ipw) 추정량 : $\hat{\mu}_{ipw} = \hat{N}_V^{-1} \sum_{i \in s_V} \frac{1}{\hat{\pi}_i^V} y_i = \hat{N}_V^{-1} \sum_{i \in s_V} \hat{d}_i^V y_i$,
- 회귀(reg) 추정량 : $\hat{\mu}_{reg} = \hat{N}_R^{-1} \sum_{i \in s_R} \frac{1}{\pi_i^R} \hat{y}_i = \hat{N}_R^{-1} \sum_{i \in s_R} d_i^R m(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})$
- 이중강건(dr) 추정량 : $\hat{\mu}_{dr} = \hat{N}_V^{-1} \sum_{i \in s_V} \frac{1}{\hat{\pi}_i^V} \{y_i - m(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})\} + \hat{N}_R^{-1} \sum_{i \in s_R} \frac{1}{\pi_i^R} m(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})$
  $= \hat{N}_V^{-1} \sum_{i \in s_V} \hat{d}_i^V \{y_i - m(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})\} + \hat{\mu}_{reg}$
- 캘리브레이션(cal) 추정량 : $\hat{\mu}_{cal} = \hat{N}_{V,cal}^{-1} \sum_{i \in s_V} \hat{d}_{i,cal}^V y_i$

여기서 $\hat{N}_R = \sum_{i \in s_R} \frac{1}{\pi_i^R} = \sum_{i \in s_R} d_i^R$, $\hat{N}_V = \sum_{i \in s_V} \frac{1}{\hat{\pi}_i^V} = \sum_{i \in s_V} \hat{d}_i^V$, $\hat{N}_{V,cal} = \sum_{i \in s_V} \hat{d}_{i,cal}^V$이다.

비확률표본은 Castro-Martin 등(2020)의 설정을 적용하여 <그림 3-1>과 같이 연령이 높아질수록 참여확률이 감소하도록, $\pi_i^V = (\text{최고령} - \text{연령})^3 / (\text{최고령} - \text{최저령})^3$ 확률로 포아송(poission) 표집하였다.

Castro-Martín 등(2020)은 가계금융복지조사와 유사한 스페인 생활실태조사(Living Conditions Survey; LCS)를 이용하여 모의실험을 수행하였다. 비확률표본에 대해 성향점수 가중치 접근으로 지니계수(gini coefficient), 빈곤율(poverty proportion), 사분위수 범위(interquartile range), 십분위수 범위(interdecile range) 등을 추정하였고, 이와 같은 편향을 조정한 접근이 그렇지 않은 접근에 비해 성공적임을 보였다.



<그림 3-1> 비확률표본 포함확률 메커니즘

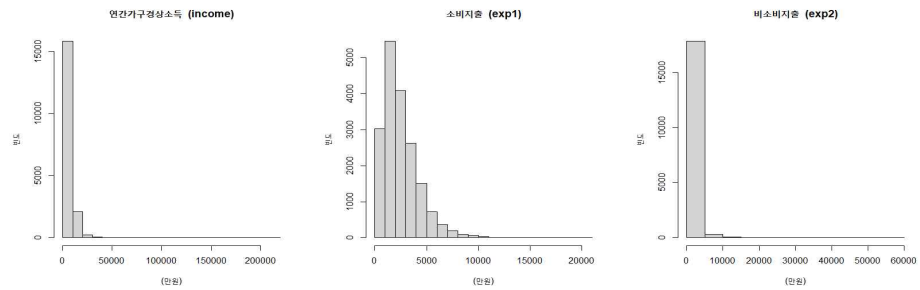
<그림 3-2>의 (a)는 2021년 가계금융복지조사의 18,187개 표본가구의 연간가구경상소득, 소비지출, 비소비지출의 분포이다. 원자료는 왜도와 첨도가 상당히 큰 자료로 이상치가 포함되어 있다. 원자료 역시 실제로는 표본자료이기 때문이다. 실제 모집단은 이와 같은 구조를 갖기 어렵기 때문에, 모의실험을 위해 연간가구경상소득은 3억원 이상, 소비지출은 8,000만원 이상, 비소비지출은 5,500만원 이상인 433가구를 이상치로 보고 제외하였다. (b)는 이상치를 제외한 17,754 가구 자료로 구성된 모의실험 모집단의 소득, 소비지출, 비소비지출 분포이다.

모의실험 모집단 자료의 주요 특성별 소득 분포를 보면 <표 3-2>, <그림 3-3>과 같다. 가구주는 남성, 60대 이상, 고졸이상, 유배우자, 상용임금자가 상대적으로 많으며, 1, 2인 이하 가구가 전체의 60% 정도를 차지한다. 지역적으로는 비수도권 지역에 많이 분포하였다. 소득 항목을 보면 가구주가 남자, 4~50대, 대졸이상, 유배우자, 상용임금자인 경우 평균 소득이 높았으며, 가구원수가 많을수록 평균 소득은 증가하였다. 지역적으로는 수도권 거주 가구의 평균 소득이 높은 것으로 나타났다. 소득과 연속형 자료인 소비지출, 비소비지출, 연령의 상관계수는 각각 0.74, 0.82, -0.33으로 나타났다.

확률표본과 비확률표본이 실제로 모집단을 얼마나 반영하는지 확인하기 위해 1,000개의 SRS, PPS, 비확률표본을 각각 10,000번 표집해서 <그림 3-4>, <표 3-3>과 같이 주요 특성의 표본 분포를 살펴보았다. SRS, PPS는 가중치를 사용한 결과이고, 비확률표본은 사용하지 않은 결과이다.

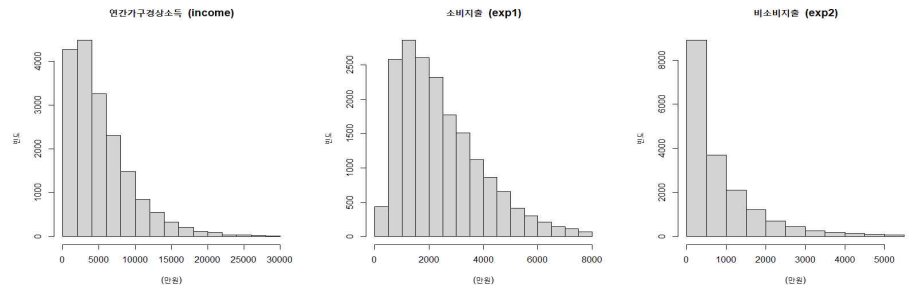
알려진 대로 확률표본은 모집단의 특성을 잘 대표하는 것으로 나타났다. 그러나 비확률표본은 모집단에 비해 가구주가 남성, 40대이하, 대졸이상, 미혼, 상용임금자가 많이 뽑혔으며, 여성, 60대이상, 중졸이하, 사별, 기타직 종사자가 상대적으로 적게

뽑혔다. 가구원수도 2인가구는 모집단보다 적게, 4인가구는 많게 뽑혀 연구를 위해 비확률표집에 가정한 선택편향 메커니즘이 잘 작동한 것으로 보인다. 후술하겠지만, 비확률표본은 모집단에 비해 평균소득이 높은 가구가 주로 뽑혔다. 결과적으로 비확률표본에 어떠한 조정도 하지 않는 경우 비확률표본은 모집단을 적절히 대표하지 않는 것을 알 수 있다.



평균	5,538	2,519	987
분산	31,073,507	2,898,641	3057739
왜도	6.98	1.62	9.30
첨도	160.62	7.71	168.21

(a) 원자료 18,187 가구



평균	5,314	2,437	834
분산	18,813,518	2,297,752	922,912
왜도	1.65	0.99	1.90
첨도	6.59	3.66	7.01

(b) 모의실험 모집단 자료 17,754 가구

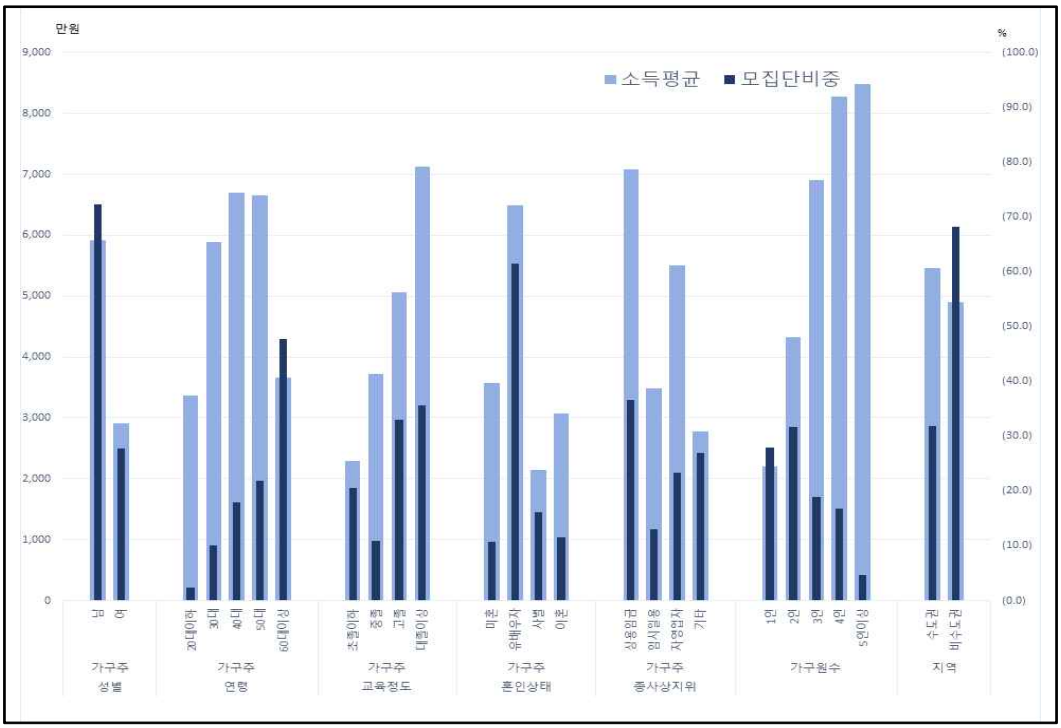
<그림 3-2> 모의실험 자료의 소득 분포

<표 3-2> 모의실험 모집단 자료의 주요 특성별 소득 분포

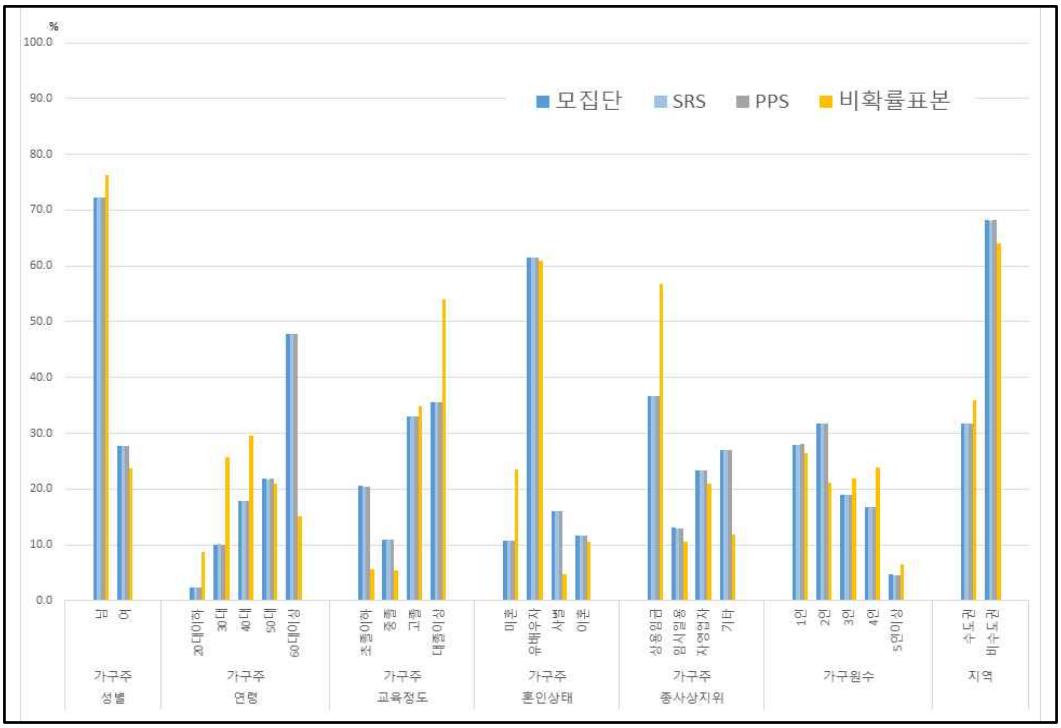
(단위 : %, 만원)

		가구수	(비중)	연간가구경상소득			
				최소	중위값	평균	최대
계		17,754	(100.0)	40	4,058	5,081	29,226
가구주 성별	남	12,830	(72.3)	40	5,053	5,912	29,226
	여	4,924	(27.7)	41	2,052	2,915	24,208
가구주 연령 범주	20대이하	425	(2.4)	70	2,980	3,370	16,071
	30대	1,782	(10.0)	52	5,288	5,895	25,534
	40대	3,178	(17.9)	40	6,157	6,703	27,767
	50대	3,883	(21.9)	41	5,724	6,660	29,226
	60대이상	8,486	(47.8)	50	2,592	3,665	29,103
가구주 교육 정도	초졸이하	3,644	(20.5)	50	1,622	2,295	22,337
	중졸	1,943	(10.9)	41	2,986	3,714	25,928
	고졸	5,850	(33.0)	41	4,328	5,064	25,333
	대졸이상	6,317	(35.6)	40	6,326	7,123	29,226
가구주 혼인 상태	미혼	1,912	(10.8)	40	3,054	3,570	18,270
	유배우자	10,926	(61.5)	122	5,674	6,492	29,226
	사별	2,860	(16.1)	50	1,364	2,140	24,208
	이혼	2,056	(11.6)	41	2,520	3,077	22,817
가구주 종사상 지위	상용임금	6,495	(36.6)	52	6,284	7,084	29,226
	임시일용	2,315	(13.0)	50	2,726	3,480	19,931
	자영업자	4,153	(23.4)	41	4,654	5,501	27,767
	기타	4,791	(27.0)	40	1,781	2,773	27,570
가구원 수	1인	4,958	(27.9)	40	1,576	2,208	24,208
	2인	5,634	(31.7)	122	3,554	4,331	25,826
	3인	3,358	(18.9)	168	6,252	6,912	29,226
	4인	2,984	(16.8)	279	7,416	8,277	29,103
	5인이상	820	(4.6)	965	7,789	8,477	25,534
지역	수도권	5,649	(31.8)	50	4,439	5,454	29,226
	비수도권	12,105	(68.2)	40	3,882	4,907	25,780

주. $r(\text{소득}, \text{소비지출}) = 0.74$, $r(\text{소득}, \text{비소비지출}) = 0.82$, $r(\text{소득}, \text{연령}) = -0.33$



<그림 3-3> 모의실험 모집단 자료의 주요 특성별 소득 분포



<그림 3-4> 주요 특성별 표본 분포

<표 3-3> 주요 특성별 표본 분포

(단위 : %)

		모집단	확률표본		비확률표본
			SRS	PPS	
계		100.0	100.0	100.0	100.0
가구주 성별	남	72.3	72.3	72.3	76.3
	여	27.7	27.7	27.7	23.7
가구주 연령 범주	20대이하	2.4	2.4	2.4	8.8
	30대	10.0	10.1	10.0	25.7
	40대	17.9	17.9	17.9	29.5
	50대	21.9	21.7	21.9	20.9
	60대이상	47.8	47.8	47.8	15.1
가구주 교육 정도	초졸이하	20.5	20.5	20.5	5.6
	중졸	10.9	10.9	10.9	5.5
	고졸	33.0	33.0	33.0	34.9
	대졸이상	35.6	35.6	35.6	54.0
가구주 혼인 상태	미혼	10.8	10.8	10.8	23.6
	유배우자	61.5	61.5	61.5	61.0
	사별	16.1	16.1	16.1	4.8
	이혼	11.6	11.6	11.6	10.6
가구주 종사상 지위	상용임금	36.6	36.6	36.6	56.7
	임시일용	13.0	13.0	13.0	10.5
	자영업자	23.4	23.4	23.4	21.0
	기타	27.0	27.0	27.0	11.9
가구원수	1인	27.9	27.9	28.0	26.4
	2인	31.7	31.7	31.7	21.2
	3인	18.9	18.9	18.9	21.9
	4인	16.8	16.8	16.8	23.8
	5인이상	4.6	4.6	4.6	6.6
지역	수도권	31.8	31.8	31.8	35.9
	비수도권	68.2	68.1	68.2	64.1

주. 1. 확률표본, 비확률표본 모두 크기가 1,000개인 표본을 10,000회 반복 추출하여 평균으로 집계
2. 확률표본은 가중치 사용, 비확률표본은 미사용

탐색적 자료 분석 결과[7] 편향 조정에 사용될 보조변수로 <표 3-4>처럼 가구주성별(2개 범주), 가구주연령(연속형 혹은 3개 범주), 가구주교육정도(3개 범주), 가구주종사상지위(2개 범주), 가구원수(5개 범주)를 선택하였다. 실제로 인구통계학적 범주형 변수들은 조사 시 상대적으로 수집이 용이하다. 회귀모형에 사용할 수 있는 설명력 높은 연속형 변수는 거의 없다고 가정하였다. 소비지출 같은 항목이 소득을 제외하고 조사에서 신뢰성 있게 수집될 수 있을지 의문이기 때문이다.

성향점수 가중치를 위해서는 로지스틱회귀모형 $\pi(\boldsymbol{x}_i,\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\theta})}$을 적용하였으며, 여기서 $\boldsymbol{x}$는 가구주의 성, 연령, 교육수준, 종사상지위 및 가구원수이다. 캘리브레이션 가중치를 위해서는 로지스틱회귀모형에 사용한 보조변수 중 연령만 범주형으로 사용하였고 나머지 변수는 동일하게 적용하였다. 통대체 방식을 위해서는 단순선형회귀모형 $\hat{y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + \epsilon_i$을 적용하였으며, 여기서 $E(\epsilon_i)=0$, $V(\epsilon_i)=\sigma^2$으로 가정하였다. 보조변수는 로지스틱회귀모형에 사용한 보조변수와 같다. 보조변수의 상세내용은 <표 3-1> 모의실험 검토 변수 내용에서 확인할 수 있다.

관심변수를 잘 설명하는 보조변수를 갖는 회귀모형은 다른 접근 방식들에 비해 분산과 편향이 작은 정확한 추정이 가능하다. 관심변수와 높은 상관관계를 갖는 보조변수를 확보할 수 있다면 회귀모형을 이용한 접근이 최선이 될 것이다. 그러나 실제에서는 쉽지 않은 일이기 때문에 본 연구에서는 획득이 용이한 보조변수를 사용하는 것에 초점을 맞춰 각 접근 방식의 적용을 검토하였다. 회귀모형은 모형 오식별에 따라 다른 방법에 비해 상대적으로 추정 성능이 떨어질 것으로 보이지만, 모형 오식별에 강건한 이중강건 추정량의 성능을 확인하는 것으로 그 효용을 찾을 수 있다.

붓스트랩 반복은 1,000회, MC 모의실험도 1,000회로 설정하였다. 각 방법은 많은 계산을 필요로 하는 Newton-Raphson 같은 수치해석적 방법들을 포함하고 있기 때문에 시나리오별로 상당한 프로그램 수행 시간이 소요되었다.

---

7) CHAID(CHi-square Automatic Interaction Detector) 이용

<표 3-4> 모의실험 검토 변수

변수 (변수명)	내용
1. 가구주성별 (gender)	남성(1), 여성(2)
2. 가구주연령 (age)	19세~103세
(age_g)	30대이하(1), 40대(2), 50대이상(3)
3. 가구주교육정도 (edu)	중졸이하(1), 고졸(2), 대졸이상(3)
5. 가구주종사상지위 (stt)	상용임금 · 자영업자(1), 임시일용 · 기타(2)
6. 가구원수 (ggw_no)	1인(1), 2인(2), 3인(3), 4인(4), 5인이상(5)
10. 연간가구경상소득 (income)	40만원~2억9,226만원

제2절 결과

실험결과는 <표 3-5>, <표 3-6>, <그림 3-5>, <그림 3-6>, <부그림 1>, <부그림 2>에 정리되어 있다. <표 3-5>, <그림 3-5>은 확률표본이 SRS일 때 각 접근 방법에 따라 평균을 추정하였고, <표 3-6>, <그림 3-6>은 확률표본이 PPS일 때의 결과이다. <그림 3-5>와 <그림 3-6>은 확률표본의 크기를 고정했을 때 비확률표본의 크기 변화에 따른 추정 결과를 볼 수 있도록 정리한 것이고, <부그림 1>과 <부그림 2>는 비확률표본의 크기를 고정했을 때 확률표본의 크기 변화에 따른 추정 결과를 볼 수 있도록 재정리하였다.

<그림 3-5>의 (a)는 평균의 상대편향인데, 비확률표본을 SRS인 것처럼 다룬 naive 추정량의 상대편향이 가장 크며, 선택편향을 조정한 ipw, reg, dr, cal추정량은 편향이 상당히 감소하였다. 그 중에서는 dr과 cal추정량이 ipw와 reg추정량에 비해 편향을 더 잘 감소시킨 것으로 나타났다. 본 실험에서는 확률표본이든 비확률표본이든 표본의 크기는 평균의 상대편향에 크게 영향을 미치지 않은 것으로 보인다.

<그림 3-5>의 (b)는 평균의 MSE인데, naive추정량의 MSE가 가장 크게 나타났고, 조정 추정량들은 MSE가 상당히 감소하였다. 상대편향과 마찬가지로 cal과 dr추정량의 MSE가 상대적으로 작게 나타난다. 조정 추정량들의 MSE는 확률표본이든 비확률표본이든 표본의 크기가 커질수록 작아지는데, 이는 분산이 표본의 크기에 영향을 받았기 때문으로 보인다. 표본의 크기가 클수록 추정의 효율은 높아진다고 볼 수 있다.

<그림 3-5>의 (c)는 모의실험마다 구해진 평균에 대한 신뢰구간이 모평균을 얼마나 포함하는지를 본 95% 포함확률 지표이다. naive추정량은 모평균을 거의 포함하지

못한다. dr과 cal추정량의 포함확률은 80%를 상회하였으며. ipw추정량은 70%를 상회하였다. 예상한 대로 reg추정량의 성능이 가장 좋지 않은 것으로 나타났다. 조정 추정량들은 확률표본이나 비확률표본의 크기가 커질수록 포함확률이 감소하는 경향을 보였는데, 이는 편향이 완전히 감소하지 않은 상태에서 표본의 크기가 큰 경우 신뢰구간이 잘못 추정된 값을 기준으로 상당히 작은 구간으로 지정되면서 오히려 모평균을 포함하지 못하기 때문이다. 또한 왜도 및 첨도가 큰 표본에 대해 정규근사화된 붓스트랩 신뢰구간을 사용한 점도 포함확률 지표가 95% 근처에 이르지 않은 것에 영향을 미친 것으로 보인다.

<그림 3-6>은 확률표본이 PPS일 때 추정 결과이다. (a)는 평균의 상대편향인데, naive추정량이 가장 큰 상대편향을 가졌으며, SRS의 경우와 같이 조정 추정량들의 편향은 상당히 감소하였다. 그중에서는 dr과 cal추정량이 ipw와 reg추정량에 비해 편향을 더 잘 감소시킨 것으로 나타났다. 실험에서는 확률표본이든 비확률표본이든 표본의 크기는 평균의 상대편향에 크게 영향을 미치지 않는 것으로 보인다. 확률표본이 SRS일 때와 비교해 큰 차이가 없다.

<그림 3-6>의 (b)를 보면, naive추정량의 MSE가 가장 크게 나타났고, 조정 추정량들은 MSE가 상당히 감소하였다. 특히 dr과 cal추정량의 MSE가 상대적으로 작게 나타난다. 조정 추정량들은 확률표본과 비확률표본의 크기가 커질수록 MSE가 보다 작아지는 경향을 보여 MSE 측면에서는 확률표본과 비확률표본 모두 자료의 크기가 클수록 유리한 것으로 보인다.

<그림 3-6>의 (c)를 보면 naive추정량은 모평균을 거의 포함하지 못하지만, ipw와 cal추정량은 포함확률이 85%를 상회하는 것으로 나타났다. 조정 추정량들은 모두 확률표본이나 비확률표본의 크기가 커질수록 포함확률이 감소하지만, 그 경향은 SRS에 비해 약하다. 모형 오식별된 reg추정량은 확률표본이나 비확률표본 중 하나라도 크기가 커질 경우 포함확률이 급격히 떨어지는 것으로 나타났다.

실험 결과, 비확률표본을 SRS처럼 취급한 것보다 선택편향 조정 방안을 적용했을 때 평균의 상대편향, 평균제곱오차, 추정 신뢰구간에 모평균 포함 정도가 모두 뚜렷한 개선을 보이는 것을 확인하였다. 그리고 표본의 크기가 클수록 추정의 효율은 개선된다는 것을 알 수 있었다. 그러나 실제 모수를 모르기 때문에 우리는 실제 상황에서는 MSE를 알 수 없다. 그런데도 모의실험처럼 편향이 있는 자료를 비편향 추정량인 것처럼 취급하여 신뢰구간을 제공할 경우에는 표본의 크기가 클수록 더 확실하게 잘못된 신뢰구간을 제시하게 된다. 즉, 데이터가 클수록 더 확실하게 우리 자신을 속이게 되는 빅데이터의 역설을 보여준다는 것을 확인할 수 있었다.

<표 3-5> 확률표본이 SRS일 때 비확률표본 평균 추정 결과

n_R		400			600		
n_V	추정량	%RB	MSE	%CP	%RB	MSE	%CP
400	$\hat{\mu}_{naive}$	14.79	602,644	1.0	14.95	613,127	1.6
	$\hat{\mu}_{ipw}$	4.67	139,610	80.2	4.75	129,690	81.1
	$\hat{\mu}_{reg}$	5.16	123,599	76.0	5.30	124,151	75.2
	$\hat{\mu}_{dr}$	2.74	79,440	83.5	2.92	74,791	82.5
	$\hat{\mu}_{cal}$	2.17	59,184	84.9	2.31	55,411	84.2
600	$\hat{\mu}_{naive}$	14.92	597,689	0.1	14.85	591,982	0.1
	$\hat{\mu}_{ipw}$	4.69	122,992	79.5	4.58	106,804	77.6
	$\hat{\mu}_{reg}$	5.45	119,642	68.1	5.20	108,204	67.9
	$\hat{\mu}_{dr}$	2.93	68,488	82.1	2.69	58,659	82.4
	$\hat{\mu}_{cal}$	2.58	54,733	81.9	2.33	46,816	82.5
800	$\hat{\mu}_{naive}$	14.93	592,182	-	14.88	588,438	-
	$\hat{\mu}_{ipw}$	4.60	108,707	78.5	4.61	99,814	74.8
	$\hat{\mu}_{reg}$	5.36	105,077	65.2	5.30	103,548	62.4
	$\hat{\mu}_{dr}$	2.77	54,963	81.5	2.68	51,479	80.0
	$\hat{\mu}_{cal}$	2.50	43,753	83.7	2.45	41,930	81.4
1000	$\hat{\mu}_{naive}$	14.90	586,305	-	14.78	577,734	-
	$\hat{\mu}_{ipw}$	4.55	102,444	77.0	4.64	95,328	72.2
	$\hat{\mu}_{reg}$	5.44	106,097	60.4	5.28	97,674	55.1
	$\hat{\mu}_{dr}$	2.77	53,680	79.4	2.66	46,179	80.1
	$\hat{\mu}_{cal}$	2.64	45,399	80.7	2.39	37,894	80.7

<표 3-5> 확률표본이 SRS일 때 비확률표본 평균 추정 결과 (계속)

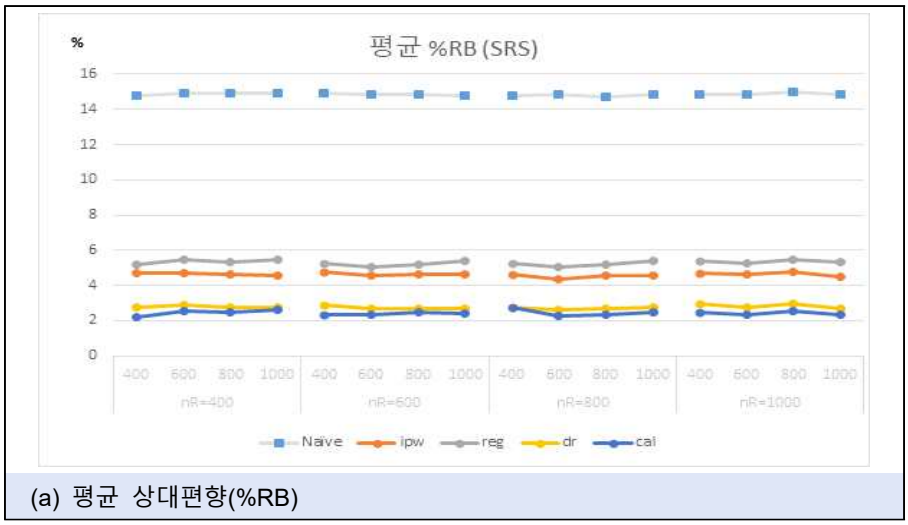
n_R		800			1000		
n_V	추정량	%RB	MSE	%CP	%RB	MSE	%CP
400	$\hat{\mu}_{naive}$	14.81	599,321	1.4	14.83	607,148	1.9
	$\hat{\mu}_{ipw}$	4.62	115,825	80.0	4.70	120,514	79.1
	$\hat{\mu}_{reg}$	5.23	118,257	74.5	5.40	126,731	71.9
	$\hat{\mu}_{dr}$	2.74	67,398	83.7	2.95	74,539	82.9
	$\hat{\mu}_{cal}$	2.39	52,400	84.8	2.49	56,792	82.6
600	$\hat{\mu}_{naive}$	14.83	589,879	-	14.87	594,942	-
	$\hat{\mu}_{ipw}$	4.35	93,838	78.5	4.62	99,294	75.9
	$\hat{\mu}_{reg}$	5.06	100,247	67.8	5.24	104,931	65.6
	$\hat{\mu}_{dr}$	2.60	52,069	81.9	2.77	54,818	82.2
	$\hat{\mu}_{cal}$	2.24	40,135	83.4	2.32	40,264	82.4
800	$\hat{\mu}_{naive}$	14.72	577,205	-	14.96	594,295	-
	$\hat{\mu}_{ipw}$	4.53	91,215	73.7	4.79	96,241	70.7
	$\hat{\mu}_{reg}$	5.20	97,053	60.7	5.49	105,282	55.5
	$\hat{\mu}_{dr}$	2.67	47,609	80.7	2.96	50,808	78.8
	$\hat{\mu}_{cal}$	2.31	37,416	80.6	2.56	39,894	79.6
1000	$\hat{\mu}_{naive}$	14.84	582,508	-	14.84	582,583	-
	$\hat{\mu}_{ipw}$	4.59	88,834	72.9	4.47	85,819	70.6
	$\hat{\mu}_{reg}$	5.40	99,206	53.6	5.30	95,420	51.7
	$\hat{\mu}_{dr}$	2.78	45,114	78.3	2.72	43,588	78.3
	$\hat{\mu}_{cal}$	2.47	36,289	80.0	2.34	33,236	79.5

<표 3-6> 확률표본이 PPS일 때 비확률표본 평균 추정 결과

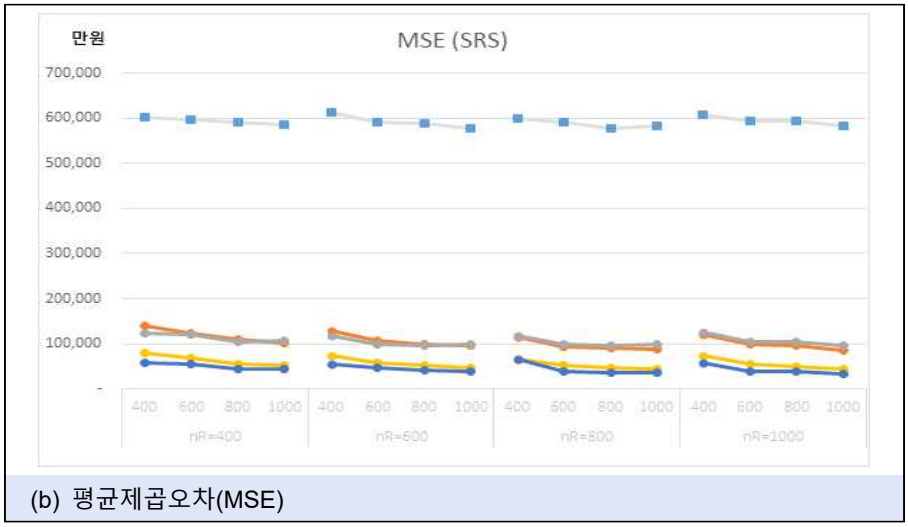
n_R		400			600		
n_V	추정량	%RB	MSE	%CP	%RB	MSE	%CP
400	$\hat{\mu}_{naive}$	14.93	611,862	1.3	14.78	598,996	1.5
	$\hat{\mu}_{ipw}$	5.48	178,289	89.2	5.44	176,062	87.1
	$\hat{\mu}_{reg}$	5.57	158,430	77.3	5.55	147,498	76.5
	$\hat{\mu}_{dr}$	2.94	106,196	85.0	3.11	99,183	84.1
	$\hat{\mu}_{cal}$	2.41	69,163	88.9	2.45	63,605	88.7
600	$\hat{\mu}_{naive}$	14.76	587,213	0.1	15.04	606,626	-
	$\hat{\mu}_{ipw}$	5.15	165,571	88.0	5.18	140,327	86.5
	$\hat{\mu}_{reg}$	5.29	133,673	75.6	5.60	135,485	71.2
	$\hat{\mu}_{dr}$	2.56	82,880	84.3	3.01	82,506	82.3
	$\hat{\mu}_{cal}$	2.41	58,719	86.6	2.57	52,411	88.5
800	$\hat{\mu}_{naive}$	14.79	581,704	-	14.85	586,696	-
	$\hat{\mu}_{ipw}$	5.47	158,057	87.8	5.44	144,737	86.9
	$\hat{\mu}_{reg}$	5.71	142,252	70.5	5.61	130,207	67.6
	$\hat{\mu}_{dr}$	2.95	87,767	80.6	3.00	78,143	80.1
	$\hat{\mu}_{cal}$	2.64	56,921	84.5	2.54	47,662	86.5
1000	$\hat{\mu}_{naive}$	14.84	581,595	-	14.80	580,234	-
	$\hat{\mu}_{ipw}$	5.25	151,242	89.3	4.97	129,116	87.4
	$\hat{\mu}_{reg}$	5.45	126,148	71.2	5.31	118,689	67.7
	$\hat{\mu}_{dr}$	2.61	72,615	82.8	2.65	70,181	80.2
	$\hat{\mu}_{cal}$	2.44	50,327	84.5	2.32	42,251	85.8

<표 3-6> 확률표본이 PPS일 때 비확률표본 평균 추정 결과 (계속)

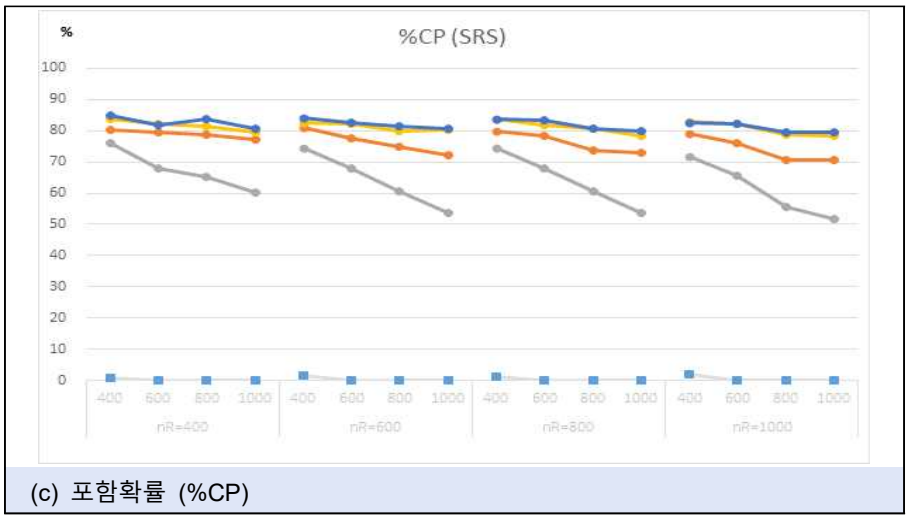
n_R		800			1000		
n_V	추정량	%RB	MSE	%CP	%RB	MSE	%CP
400	$\hat{\mu}_{naive}$	14.87	607,880	1.1	14.74	594,333	0.9
	$\hat{\mu}_{ipw}$	5.14	142,790	88.1	5.12	144,722	87.2
	$\hat{\mu}_{reg}$	5.17	130,362	77.6	5.32	130,790	75.1
	$\hat{\mu}_{dr}$	2.76	83,281	85.3	2.90	81,032	84.8
	$\hat{\mu}_{cal}$	2.31	56,568	89.9	2.21	54,032	90.1
600	$\hat{\mu}_{naive}$	14.92	596,971	0.1	14.86	592,906	0.1
	$\hat{\mu}_{ipw}$	5.06	121,969	87.9	5.14	125,748	85.8
	$\hat{\mu}_{reg}$	5.26	114,853	73.3	5.54	123,021	68.5
	$\hat{\mu}_{dr}$	2.68	64,019	84.5	2.97	68,798	82.2
	$\hat{\mu}_{cal}$	2.41	44,672	89.0	2.59	46,586	88.9
800	$\hat{\mu}_{naive}$	14.88	588,938	-	14.90	587,766	0.1
	$\hat{\mu}_{ipw}$	4.91	106,641	88.3	4.86	106,582	86.7
	$\hat{\mu}_{reg}$	5.21	107,360	72.0	5.18	102,127	68.9
	$\hat{\mu}_{dr}$	2.64	57,177	83.4	2.60	51,965	83.1
	$\hat{\mu}_{cal}$	2.37	41,203	87.5	2.28	37,050	87.9
1000	$\hat{\mu}_{naive}$	14.84	581,987	-	14.89	585,689	-
	$\hat{\mu}_{ipw}$	4.92	106,915	88.0	5.06	108,148	87.0
	$\hat{\mu}_{reg}$	5.42	110,274	63.5	5.43	107,855	61.9
	$\hat{\mu}_{dr}$	2.69	56,882	82.5	2.79	54,357	80.5
	$\hat{\mu}_{cal}$	2.58	41,180	84.8	2.50	36,188	86.7



(a) 평균 상대편향(%RB)

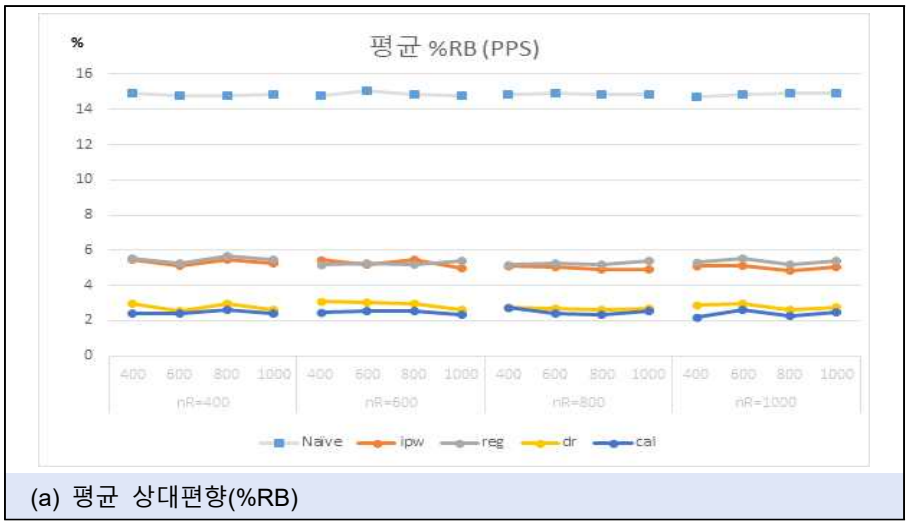


(b) 평균제곱오차(MSE)

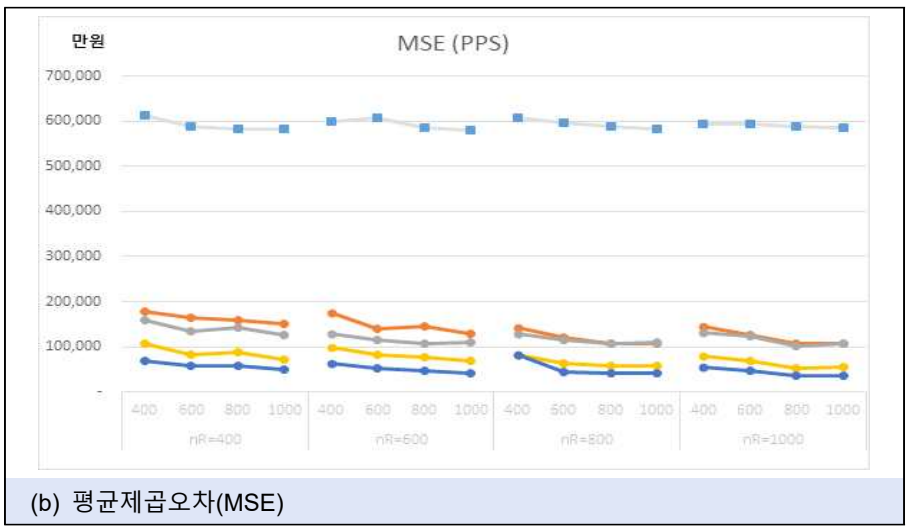


(c) 포함확률 (%CP)

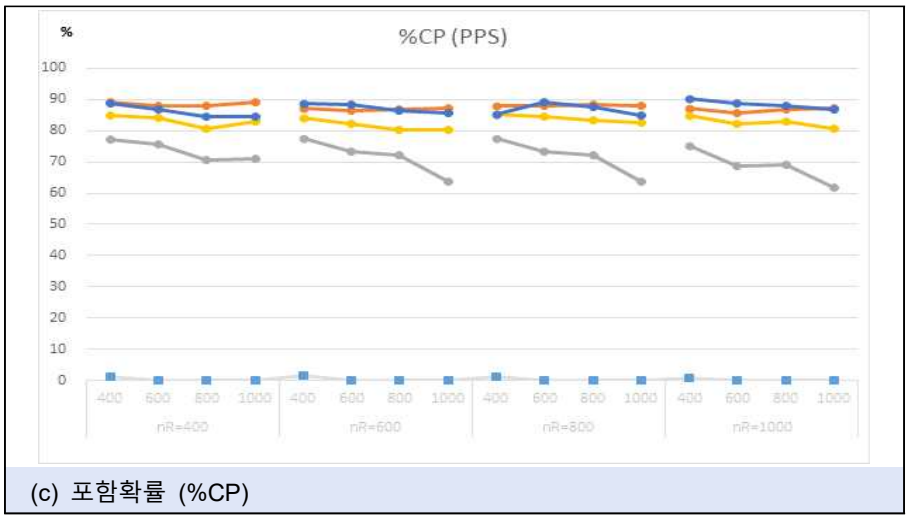
<그림 3-5> 확률표본이 SRS일 때 비확률표본 평균 추정 결과



(a) 평균 상대편향(%RB)



(b) 평균제곱오차(MSE)



(c) 포함확률 (%CP)

<그림 3-6> 확률표본이 PPS일 때 비확률표본 평균 추정 결과

제 4 장

결 론

제1절 요약

비확률표본은 미지의 생성 메커니즘 때문에 일반적으로는 목표모집단을 대표하지 않는다. 비확률표본을 SRS인 것처럼 다루는 경우 심각한 선택편향 문제에 부딪힐 수 있기 때문에 비확률표본을 확률표본의 대안으로 사용하기 위해서 선택편향을 줄이기 위한 조정 절차는 필수적임을 확인하였다. 선택편향 조정 절차는 비확률표본과 고품질 확률표본과의 데이터 통합을 필요로 하며 두 자료를 연결하는 모형에 의존할 수 밖에 없다.

본 연구는 확률표본과 비확률표본 모두에서 공통의 보조변수가 관측되며, 비확률표본에서만 관심변수가 관측되는 환경을 설정하였다. 평균추정량으로 비확률표본을 단순임의표본인 것처럼 다룬 단순(naive) 추정량, 성향점수 가중치(ipw) 추정량, 회귀모형(reg) 추정량, 성향점수모형과 회귀모형을 결합한 이중강건(dr) 추정량, 캘리브레이션(cal) 추정량 등 5개 추정량을 고려하였다.

모의실험에는 2021년 가계금융복지조사 공공용 가구마스터 자료를 사용하였다. 관심값은 연간가구경상소득의 평균이며, 편향 조정을 위한 보조정보로 가구주의 성, 연령, 교육정도, 종사상지위와 가구원수를 사용하였다. 비확률표본과 확률표본의 크기, 확률표본의 설계 등을 변화시킨 총 32개의 시나리오에 대해 5개 추정량을 각각 산출하였다.

모의실험을 통해 경험적으로 얻고자 했던 점과 결과를 정리해 보면 다음과 같다. 첫 번째 가설 “비확률표본의 선택편향 조정 방안들은 실제로 편향을 감소시킬 것이다.”에 대한 결과는 “편향을 감소시킨다.”이다.

- 비확률표본을 단순임의표본인 것처럼 다루는 경우 심각한 선택편향이 발생할 수 있기 때문에 편향을 감소시키는 노력이 필요하며 편향을 조정한 4개 추정량은 모두 선택편향을 뚜렷이 감소시키는 것으로 나타났다. 그러나 편향을 완전히 없애지는 못했다.

- dr추정량은 예측모형(reg추정량)의 오식별에도 안정적인 추정 결과를 보여주며, cal추정량 역시 ipw나 reg추정량에 비해 안정적인 추정 결과를 보여준다. 이중강건 특성을 갖는 추정량의 사용이 실제에서도 유용할 것으로 보인다.

두 번째 가설 “확률표본이든 비확률표본이든 규모가 클수록 추정의 효율이 개선될 것이다.”에 대해서는 “추정의 효율은 개선된다. 그러나 이를 확인하기는 어렵다.”

- 표본의 크기에 따른 선택편향의 감소 경향은 보이지 않지만,
- 편향 조정된 평균 추정량의 MSE는 확률표본이나 비확률표본의 크기가 커질수록 뚜렷이 감소하는 경향을 보였다.
- 그러나, 95% 신뢰구간의 모평균 포함확률은 확률표본이나 비확률표본의 크기가 커질수록 감소하는 경향을 보인다. 그러나 이는 잔류 편향이 있는 상태에서 과소추정된 분산추정량으로 신뢰구간을 지정한 때문으로 보인다.
- 비확률표본의 추정에서 가장 큰 문제는 남겨진 편향의 존재이며, 더불어 안정적인 분산추정량 연구가 필요하다는 것을 의미한다. 이 문제가 해결되지 않은 상태에서 규모가 큰 표본을 MSE기반의 품질 지표로 해석하려 하면 오히려 잘못된 결론을 도출할 수 있다.

세 번째 가설 “동일한 규모의 표본이라면 보다 정밀한 설계에 기반한 확률표본을 참조하는 것이 추정의 효율을 개선시킬 것이다.”는 “일반화하기 어렵다”.

- 참조확률표본의 설계가 SRS인지 PPS인지는 추정에 크게 영향을 미치지 않는 것으로 보인다. 오히려 PPS일 때 상대편향과 MSE의 성능이 다소 떨어져 보이기도 한다. 이는 PPS의 가중치 변동에 따른 효과와 모형에 설계변수를 반영하지 않은 때문으로 보인다. 참조확률표본의 설계변수에 대한 정밀한 고려가 비확률표본 추론에 반영되지 않는다면 SRS를 참조하는 경우가 더 나을 수도 있다고 보인다.

본 연구의 실험 결과는 특별한 경우에 대한 것으로 모든 데이터로 일반화하여 해석할 수는 없다. 선택편향 조정이 오히려 좋지 않은 결과를 보여줄 수도 있기 때문이다. 그게 현재 비확률표본 추론이 가지고 있는 문제이다. 그러나, 이들 지표 간 관계를 통해 선택편향 조정이 없을 경우에는 데이터가 클수록 더 확실하게 우리 자신을 속이게 되는 빅데이터의 역설을 보여준다는 것은 확실히 확인할 수 있다.

고품질 확률표본이라면 확률표본의 크기나 설계는 추정에 크게 영향을 미치지 않는 것으로 보인다. 비확률표본의 크기는 평균 추정에서는 크게 영향을 미치지 않는 것으

로 보이나 MSE를 개선하는 것으로 나타났다. 따라서 비확률표본의 표본 크기는 추정에 중요한 고려 요인이 된다. 또한 비확률표본의 크기가 크다면 보조정보의 차원에 제약을 덜 받기 때문에 표본의 크기는 이런 점에서도 고려되어야 한다.

제2절 시사점 및 결론

본 연구에서는 비확률표본의 선택편향을 보정하기 위해 다양한 접근 방법을 검토하였다. 이를 위해 강력한 모형 가정을 전제하였지만, 실제로 모형과 가정이 타당하였는지 검증하기 위한 뚜렷한 방법이 있지는 않다. 이를 위해 활용한 보조변수와 추론 결과에 대한 평가 역시 아직은 미지의 영역이다.

그럼에도 선택편향 감소를 위해 어떤 접근 방식을 선택할 것인가에 대한 결정은 가용한 보조변수의 품질에 크게 의존하며, 확률표집의 설계 정보(복합설계, 포함확률의 불균등 여부 등), 변수의 특성(이산형, 연속형, 다변량 등)도 중요한 요인이 된다.

설명력이 좋은 보조변수를 확보했다면, 모형기반 방법인 reg나 dr추정량이 좋은 선택이 될 수 있다. 그러나 이 방법은 관심변수마다 모형이 필요하기 때문에 단일 가중치를 생산하지 못한다. 마이크로데이터를 제공하는 통계작성 기관에서는 채택하기 어려운 방법이다. 범주형 보조변수를 확보했다면, cal이나 ipw추정량을 선택할 수 있다. 본 연구에서 사용한 레이킹 방법과 같은 cal추정량은 적용이 간단하다는 장점이 있으나, 보조변수 형태 및 수에 제약을 받게 된다. 따라서 행정자료, 빅데이터 등 existed data에 적합할 것이다. ipw추정량은 상대적으로 보조변수 형태 및 수에 자유로우므로 survey designed 데이터에 적합한 것으로 보인다.

어떤 데이터를 확보했는지에 따라 각 방식의 장단점에 대한 균형 잡힌 고려가 필요하다. 결론적으로 비확률표본을 위한 통계적 추론의 성공 여부는 어떤 방법을 선택하느냐보다 좋은 공변량의 확보 여부에 있다.

모의실험을 통해 가구의 연간경상소득처럼 변동이 큰 관심변수에 대해서도 적절한 보조변수를 사용한다면 안정적인 선택편향 보정이 가능하다는 것을 확인하였기 때문에 다양한 조사와 관심변수에 확대 적용이 가능할 것으로 기대한다.

선택편향을 조정한 4개 추정량은 모두 선택편향을 뚜렷이 감소시키는 것으로 나타났지만, 편향을 완전히 없애지는 못했다. 잔존 편향이 있는 상태에서 전통적인 방식으로 구성된 신뢰구간은 잘못된 값에 중심이 있기 때문에 모수를 포함할 가능성 떨어진다. 따라서, 이를 실무적으로 적용하기에는 어려움이 있다.

잔존 편향을 어떻게 처리하고 해석할 지와 관련해서는 두 가지 검토사항이 남아

있다. 첫 번째로, 잔존 편향을 완전히 없앨 수 없다면 시간 경과에 따른 추세 분석을 통해 활용성을 검토할 수 있다. 예컨대 연간 조사의 경우, 비확률표본을 위한 단년도 추정에서는 편향이 남아 있지만, 매년 동일한 매커니즘으로 비확률표본이 생성된다는 가정하에 전년대비 증감 등의 방식으로 비확률표본 추정 결과의 활용이 가능한지 검토할 수 있다. 두 번째로는 Couper(2013)가 말한 것처럼 "빅데이터 또는 비확률 설문 조사에서 선택편향 또는 비포함의 위험을 정량화하는 다른 방법이 필요하다."

또한, 선택편향 감소를 위해 참조확률표본을 활용하는 접근 방식은 분산의 상당한 증가를 동반하기 때문에 편향감소와 분산증가 간 trade-off가 필요하다(김서영 등, 2010). 많은 연구가 분산보다는 편향을 조정하는 데 관심이 있지만, 표집방식과 적용 모형에 맞는 분산 추정량 연구도 필요하다. 본 연구에서는 붓스트랩 방식으로 분산을 추정하였는데, 이 방식은 비선형 추정량의 분산 추정량을 구하는 데 있어 실제에서 유용한 방법이지만 i.i.d 데이터가 아닌 복잡설계나 비확률표본을 위한 분산추정 방법 역시 일관된 프로세스가 없기 때문에 방법의 적용과 결과의 해석에 주의를 기울여야 한다. 계산에도 많은 시간이 소요되는 단점이 있다.

최근엔 검증할 수 없는 모형 가정들에 기반하지 않은 방법론 연구가 활발히 진행되고 있다. 강한 무시가능성, 혹은 무작위결측 가정이 필요하지 않은 방안들이다(최진웅과 임종호, 2023; Beppu 등, 2022; Kim, 2022b). 관련 연구의 진행에 따라 비확률표본의 통계적 추론이 보다 안정적으로 진행될 수 있을 것으로 기대된다.

공식통계 생산을 위한 확률조사가 사라질 것인가? 그렇지 않다. 확률표본은 여전히 매우 유용하다. 응답률이 감소하는 시대에도 확률표본의 정확도는 일반적으로 비확률표본보다 높다. 지속해서 감소하는 응답률이 확률표본 데이터의 품질을 훼손하기 때문에 비확률표본조사로 전환하는 것이 바람직하다는 주장에 대한 경험적 근거는 아직은 없다(Cornesse 등, 2020). 또한 통계기관에서 수집한 센서스 데이터 및 대규모 확률표본은 비확률표본의 통계적 분석을 위한 풍부한 정보의 원천 역할을 수행하고 있다. 관심변수에 대한 유용한 예측변수가 될 가능성이 있는 웨보그라픽(webographic) 같은 변수를 포함하도록 비확률표본 추론을 지원하는 확률조사를 설계할 수도 있을 것이다.

비확률표본을 이용한 통계 생산은 거스르기 어려운 시대적 요구로 보인다. 본 연구는 다출처 자료, 특히 확률표본과 비확률표본의 통합 및 관련 통계적 추론 방안을 검토하였다. 빅데이터 시대를 맞아 그동안의 확률표본에 기반한 현장조사 중심 통계 생산 패러다임의 변화를 시도한 것이다. 그러나 본 연구는 패러다임 변화의 일부분만을 일부 데이터 중심으로 살펴보았다. 비확률표본이 통계 생산의 주요 도구로 활용되기 위해서는 선택편향을 넘어서 측정오차, 처리오차, 무응답오차와 같은 비표본오차의 전 영역에 걸친 혁신과 대응이 필요하다.

참고문헌

가계금융복지조사(공공용). (2021). https://doi.org/10.23333/P.930001.001

권순필, 정미옥, 서수희, 임종호. (2022). **비확률표본을 위한 통계적 추론**, 통계개발원.

김서영, 안다영, 권순필, 이승희. (2010). **사회조사에서 자원자표본 인터넷조사 추정**, 통계개발원.

최진웅, 임종호. (2023). 데이터 결합을 통한 비확률표본 활용, 2023년 하계 통계학회 학술대회 발표자료.

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Denver, J.A., Gile, K.J. and Tourangeau, R. (2013). Report of the AAPOR task force on non-probability sampling. *J. Surv. Statist. Methodol., 1*, 90-143.

Beppu, K., Morikawa, K., and Im, J. (2022). "Imputation with verifiable identification condition for nonignorable missing outcomes", https://arxiv.org/pdf/2204.10508.pdf.

Castro-Martín et al. (2020). Inference from Non-Probability Surveys with Statistical Matching and Propensity Score Adjustment Using Modern Prediction Techniques, *Mathematics, June 2020, 8(6)*, 879.

Chen, Y., Li, P. and Wu, C. (2020). "Doubly robust inference with non-probability survey samples", *Journal of the Americal Statistical Assocation* 115, 2011-2021.

Couper, M. P. (2013). "Is the sky falling? New technology changing media, and the future of surveys". Surv. Res. Methods, 7, 145-156.

Cornesse, A., BLOM, A. G., Dutwin, D., Krosnick, J. A., et al. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research, *Journal of Survey Statistics and Methodology (2020) 8*, 4–36.

Deville, J.C. and Särndal, C.E. (1992). "Calibration estimators in survey sampling", *Journal of the Americal Statistical Assocation 87*, 376-382.

Elliot, M.. and Valliant, R. (2017). "Inference for nonprobability samples", *Statistical Science* 32, 249-264.

Kim, J.K. (2022a),."A gentle introduction to data integration in survey sampling", *The Survey Statistician 85*, 19-29.

Kim, J.K. (2022b). "Multiple Bias Calibration for Valid Statistical Inference with

Selection Bias", *in Proceedings of the Survey Research Methods Section, Joint Statistical Meetings,* American Statistical Association.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat., 12*, 685-726.

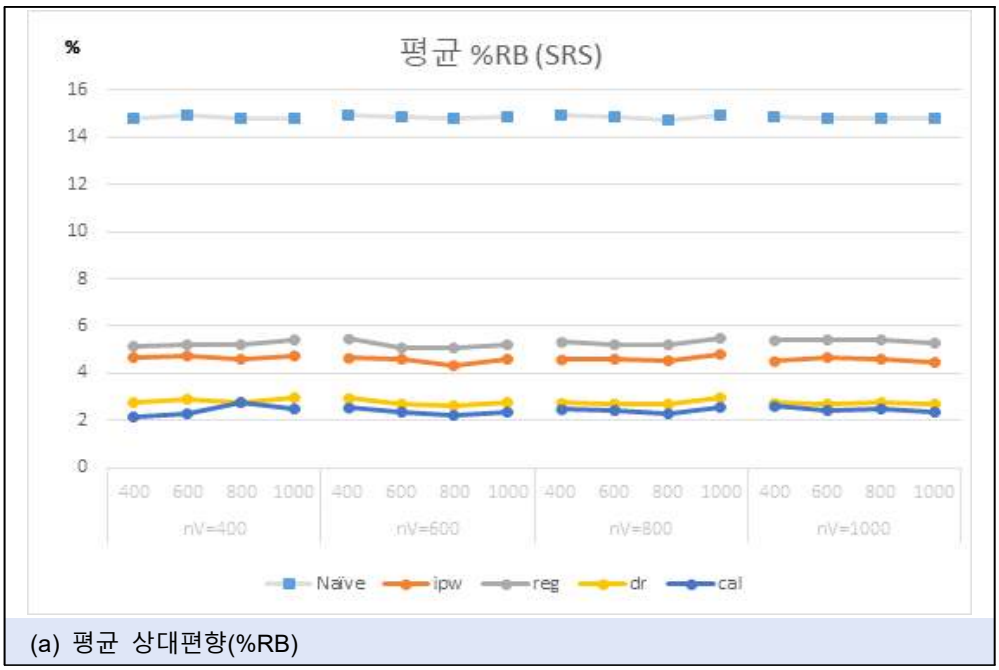
Mercer, A. W., Kreuter, F. and Stuart, E. A. (2017). Theory and practice in nonprobability surveys. Public Opin. *Q., 81*, 250-279.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika 70(1)*: 41-55.

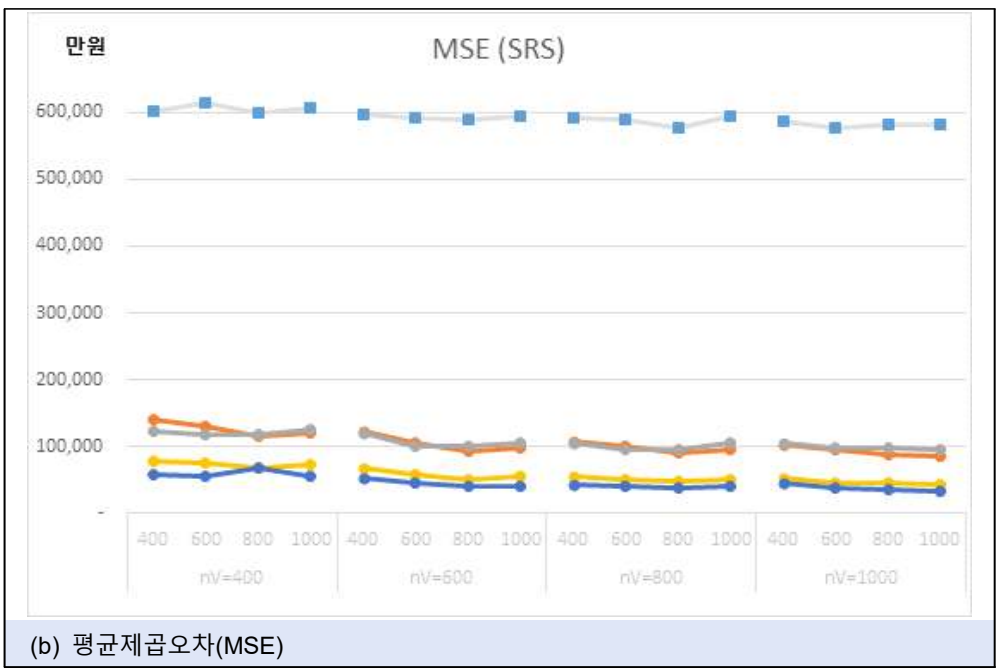
Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. Sociol. *Methods Res., 40*, 105-137.

Zhang, L.C. (2019). On valid descriptive inference from non-probability sample. *Statistical Theory and Related Fields, Taylor & Francis Journals, vol. 3(2)*, 103-113, July.

부 록

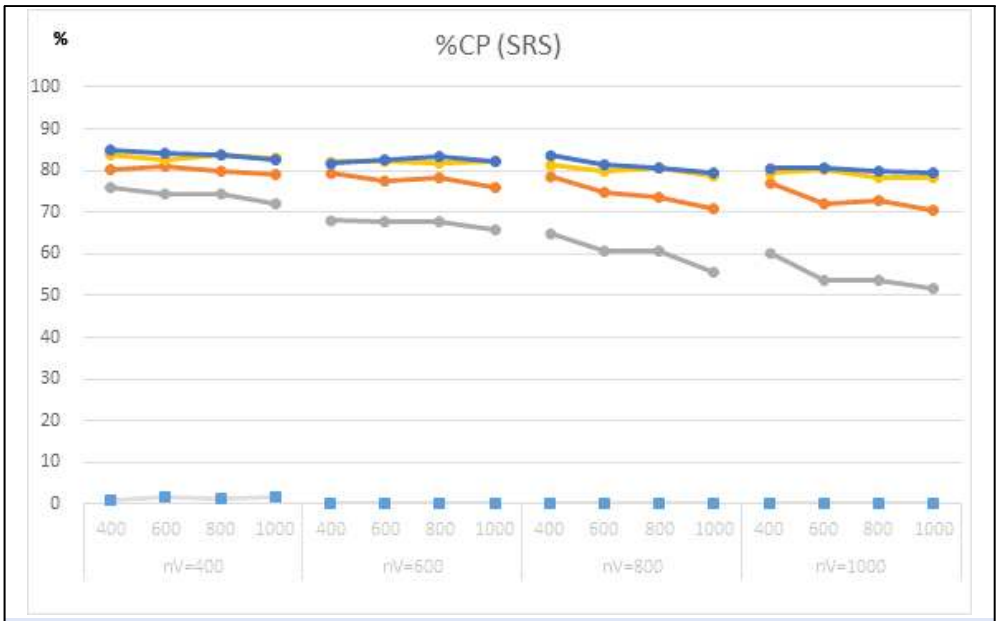


(a) 평균 상대편향(%RB)



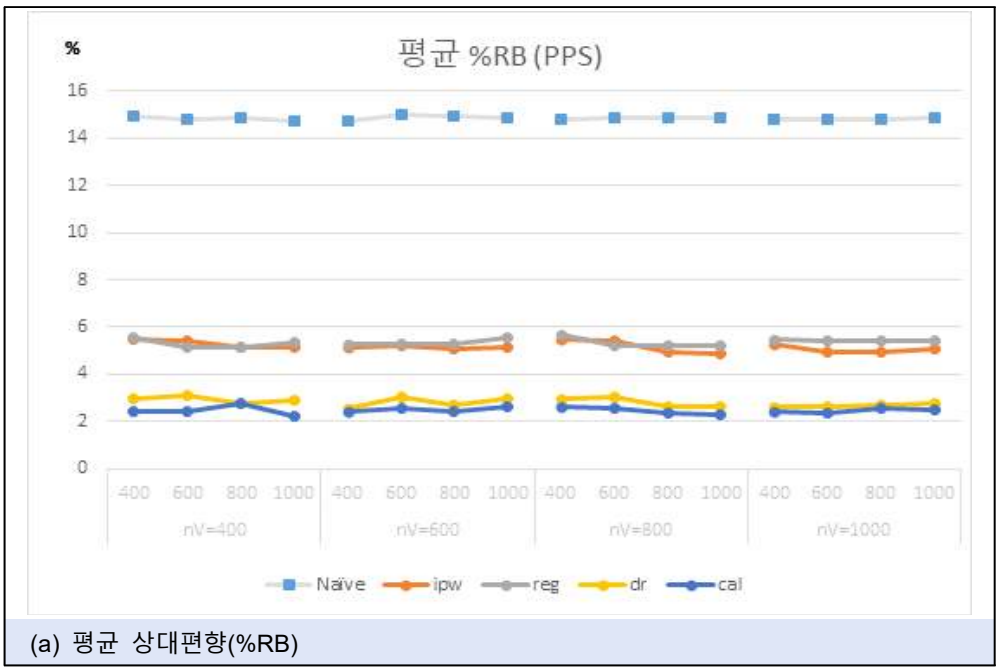
(b) 평균제곱오차(MSE)

<부그림 1> 확률표본이 SRS일 때 비확률표본 평균 추정 결과



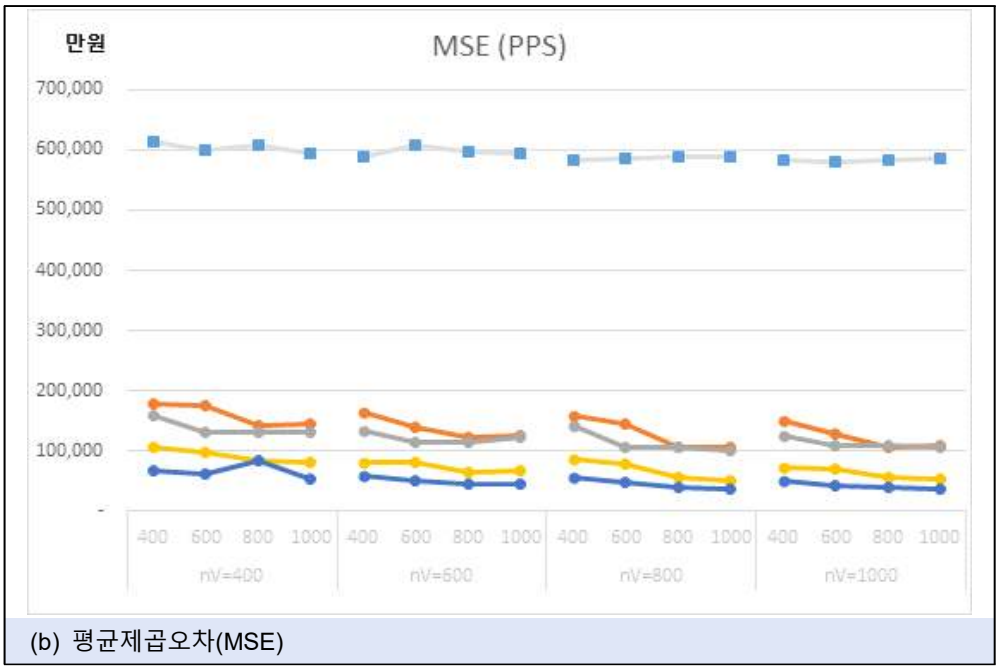
(c) 포함확률 (%CP)

<부그림 1> 확률표본이 SRS일 때 비확률표본 평균 추정 결과 (계속)

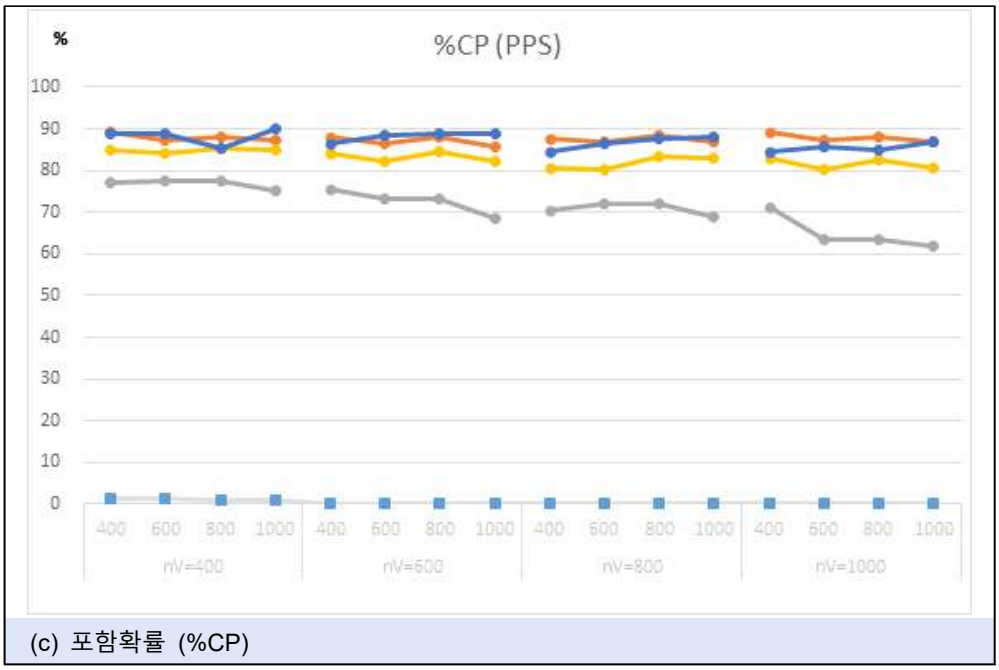


(a) 평균 상대편향(%RB)

<부그림 2> 확률표본이 PPS일 때 비확률표본 평균 추정 결과



(b) 평균제곱오차(MSE)



(c) 포함확률 (%CP)

<부그림 2> 확률표본이 PPS일 때 비확률표본 평균 추정 결과 (계속)

Abstract

Statistical Inference for Non-probability Sample: Simulation Study

Soonpil Kwon, Heeyoung Chung, Youngmi Kwon, Seunghwan Kim

An increase in non-probability samples from different sources and the development of IT for data processing require a change in the paradigm of statistical production based on probability samples. Because it's difficult to select and maintain a probability sample due to a decrease in the coverage of the sampling frame, an increase in non-responses and survey costs, and a worsening survey environment such as COVID-19. However, for the inference of finite populations, the problems of selection bias, under-coverage and unknown sampling probability of non-probability samples must be addressed. For this purpose, data integration of non-probability samples and high-quality reference probability samples, and the specification of a model connecting the two datasets are essential. It's possible to borrow the unbiasedness of probability samples.

This study examines four estimators of the propensity score weighting method(ipw), the calibration weighting method(cal), the mass imputation method(reg), the doubly robust method(dr). The variance of each estimator is estimated using bootstrapping.

For the simulation study, public master datasets of the 2021 Survey of Household Finances and Living Conditions are used as the population. The value of interest is the average annual current income of households, and the auxiliary variables are the demographic characteristics of the household head and the number of household members. We assume various scenarios and estimate the average of non-probability samples and the confidence interval for each scenario.

All four estimators show a drop in relative bias, mean square error. The probability of including the population mean in the 95% confidence interval is approximately 80% on average. Among the four estimators, the doubly robust estimator and the calibration estimator show the most stable estimation results. When a non-probability sample is treated as if it were a SRS, serious selection bias problems arise.

This simulation study shows that stable selection bias correction is possible if appropriate auxiliary variables are used even for variables of interest that fluctuate greatly, such as average annual current income of households. Therefore, it is expected that it will be possible to expand application to a variety of surveys and variables of interest.

Key words: non-probability sample, probability sample, borrow unbiasedness, selection bias, propensity score, calibration, mass imputation, doubly robust

연구진

- 권순필 (통계청 통계개발원 통계방법연구실 사무관)
- 정희영 (통계청 통계개발원 통계방법연구실 주무관)
- 권영미 (통계청 통계개발원 통계방법연구실 주무관)
- 김승환 (통계청 통계개발원 통계방법연구실 연구지원)

* 연구진의 소속 및 직급은 연구과제 완료 시 기준임을 알려드립니다.

연구보고서 2023-11

비확률표본을 위한 통계적 추론: 실증연구

인 쇄	2024년 4월
발 행	2024년 4월
발 행 인	박상영
발 행 처	통계청 통계개발원
	35220 대전광역시 서구 한밭대로 713
	TEL.(042)366-7100 Fax.(042)366-7123
홈페이지	http://sri.kostat.go.kr
ISSN(Online)	2733-4120

통계청
통계개발원