

비선형 결과회귀모형에서 MNAR 무응답 처리를 위한 조각별 캘리브레이션 추정량¹⁾

정주원²⁾ · 신기일³⁾

요약

무응답 발생 체계에 따라 다양한 무응답 처리 방법이 연구되고 있다. 완전임의무응답인 MCAR(missing completely at random) 그리고 임의무응답인 MAR(missing at random)을 따르는 무응답은 그 처리가 쉽고, 정확한 추정 방법이 개발되었다. 반면 MNAR(missing not at random) 무응답의 적절한 처리에 관한 연구는 진행되고 있으며 특히 캘리브레이션 보정 방법에 관한 다양한 연구가 활발히 진행되고 있다. 캘리브레이션 추정량은 결과회귀모형(outcome regression model)이 선형인 경우에서 효과가 있는 것으로 알려져 있다. 이에 본 연구에서는 비선형 결과회귀모형에서 사용이 가능한 조각별 캘리브레이션 추정량(segmented calibration estimator)을 제안하며 모의실험을 통해 본 연구에서 제안한 추정량의 우수성을 확인한다.

주요 용어: 조각별 회귀, 일반화회귀 추정량, 응답확률모형

1. 서론

다수의 무응답이 조사통계에서 발생하고 있으며 무응답의 적절한 처리 방법은 무응답 발생 메커니즘에 따라 달라진다. 완전임의무응답인 MCAR(missing completely at random) 무응답은 응답확률이 동일하여 처리가 쉽고, 임의무응답인 MAR(missing at random) 무응답은 응답확률이 보조변수와 관계가 있어 적절한 보조변수를 사용하면 정확한 응답확률 추정이 가능하다. 반면 MNAR(missing not at random) 무응답의 응답확률은 보조변수와 관심변수의 함수이다. 따라서 응답확률 추정에 사용하는 모형에 관심변수가 독립변수로 포함되어야 하므로 직접적인 응답확률 추정은 매우 어렵다. 이에 다수의 MNAR 무응답 처리 방법은 응답확률을 직접 추정하지 않고, 우회적인 방법을 통하여 응답확률을 추정한다.

MNAR 무응답 처리 방법으로 이중강건추정량(doubly robust estimator) 또는 캘리브레이션 추정량(calibration estimator)을 확장하는 연구가 활발히 진행되고 있다. 이중강건추정량은 역확률가중추정량(inverse probability weighted estimator)을 확장한 증강역확률가중추정량(augmented inverse probability weighted estimator)이다. 이중강건추정량은 정확한 응답확률 추정이 이루어지거나 결과회귀모형(outcome regression

1) 이 연구는 2026년 한국외국어대학교 교내연구비 지원을 받아 수행되었음

2) 제1저자, 한국외국어대학교 통계학과, 석사과정, E-mail: 2001jjuwon@naver.com

3) 교신저자, 한국외국어대학교 통계학과, 교수, E-mail: keyshin@hufs.ac.kr

model)에서 얻어진 정확한 예측값이 사용되면 우수한 결과가 얻어지는 것으로 알려져 있다. Cao 등 (2009)과 Kim and Haziza (2014)는 이중강건추정량의 추론과 특징을 연구했으며 Uehara 등 (2023)은 이중강건추정량과 무응답 대체추정량을 확장한 준모수 방법추정량을 제안하였다.

이중강건추정량이 결과회귀모형에서 얻어진 정보, 즉 예측값을 사용한다면 캘리브레이션 추정량은 보조변수의 모집단 정보, 즉 보조변수 총합을 사용하여 응답확률의 정확성을 향상한다. 캘리브레이션 보정법은 사후층화 방법의 하나이며 설계 당시 주어진 설계 가중치를 알려진 보조변수 총합과 최종 응답 자료에서 추정된 보조변수 총합이 같아지도록 가중치를 보정하는 방법이다. χ^2 거리, Kullback-Leibler 거리 등 다양한 거리 함수가 사용될 수 있으며 자료 수가 많아지면 거리 함수에 무관하게 점근적으로 동일한 결과가 얻어지는 것으로 알려져 있다. 이와 관련한 내용은 Wu and Lu (2016) 그리고 Lim and Park (2024)에서 살펴보기 바란다.

캘리브레이션 추정량은 결과회귀모형(outcome regression model)이 선형인 경우에서 우수한 결과가 얻어지는 것으로 알려져 있다. 이는 사용하는 거리함수에 무관하게 캘리브레이션 추정량은 점근적으로 일반화회귀(GREG: generalized regression) 추정량과 일치하고 일반화회귀 추정량은 기본적으로 선형회귀 모형을 사용하기 때문이다. 따라서 결과회귀모형이 비선형인 경우에는 선형모형 가정에서 타당한 캘리브레이션 추정량을 사용하게 되면 좋지 않은 결과가 얻어질 수 있다. 이에 본 연구에서는 비선형 모형을 조각으로 나누어 선형화하고 여기서 만들어지는 캘리브레이션 제약식을 만족하도록 가중치를 보정하여 추정의 정확성을 향상하는 조각별 캘리브레이션 추정량을 제안한다.

본 논문의 구성은 다음과 같다. 2절에서는 기존에 제안된 캘리브레이션 방법을 살펴본다. 3절에서는 일반화회귀 추정량을 간단히 살펴보고, 본 연구에서 제안한 조각별 캘리브레이션 추정량을 설명한다. 4절에 모의실험 결과를 수록하였으며 5절에 결론이 있다.

2. 캘리브레이션 방법

캘리브레이션 방법은 알려진 보조자료의 모총합과 조사된 보조자료의 총합 추정값이 일치하도록 가중치를 조정하는 방법이다. 먼저 N 개의 유한모집단에서 n 개의 표본이 추출되었고 $n-r$ 개의 무응답이 발생하여 r 개의 최종 응답 자료가 얻어졌다고 가정하자. 또한 w_i 를 알려진 설계 가중치라 하고 모집단에 포함된 p 차원 보조변수 x_i 의 알려진 총합을 t_X 라 하자. 이제 주어진 설계 가중치 w_i 를 보정하여 얻어진 보정 가중치 w_i^F 는 다음의 제약식을 만족하도록 한다.

$$\sum_{i=1}^r w_i^F x_i = t_X \quad (2.1)$$

(2.1)을 만족하는 w_i^F 는 다수가 존재하므로 최적의 w_i^F 를 찾아 사용한다. 이를 위해 타당한 거리 함수를 사용하여 주어진 설계 가중치와의 거리를 최소화하면서 (2.1)을 만족하는 w_i^F 를 찾아낸다. 물론 사용하는 거리 함수에 따라 다른 w_i^F 결과가 얻어진다. 그러나 사용된 거리 함수와 무관하게 얻어진 최종 보정 가중치 w_i^F 는 $w_i^F = w_i \times g_i$ 형태로 얻어진다. 여기서 g_i 는 거리 함수와 (2.1) 제약식을 사용한 라그랑주 승수법 (Lagrange multiplier)을 통해 얻어진다. 이제 사용한 거리 함수에서 얻어진 최적의 g_i 는 다음과 같다.

2.1 χ^2 거리 함수

$$D(w^F, w) = \sum_{i=1}^r \frac{(w_i^F - w_i)^2}{w_i} \quad (2.2)$$

χ^2 거리 함수와 제약식에서 얻어진 목적 함수 $Q(w_i^F, \lambda) = \sum_{i=1}^r \frac{(w_i^F - w_i)^2}{w_i} + 2\lambda_1^T (\sum_{i=1}^r w_i^F x_i - t_X)$ 를 사용한 라그랑주 승수법을 통해 g_i 가 얻어진다. 이제 χ^2 거리 함수를 이용한 $g_i = g_{1i}$ 라 하면 $g_{1i} = 1 + \lambda_1^T x_i$ 가 되고 여기서 $\lambda_1^T = (t_X - \hat{t}_{X,HT})^T (\sum_{i=1}^r w_i x_i x_i^T)^{-1}$ 가 얻어져 최종 보정 가중치 w_i^F 는 (2.3)으로 구해진다.

$$w_i^F = w_i \times [1 + (t_X - \hat{t}_{X,HT})^T (\sum_{i=1}^r w_i x_i x_i^T)^{-1} x_i] \quad (2.3)$$

여기서 $\hat{t}_{X,HT} = \sum_{i=1}^r w_i x_i$ 인 호르비츠-툼슨 총합 추정량이다.

2.2 Modified minimum entropy distance(Weighted backward Kullback-Leibler distance)

$$D(w^F, w) = ET(w^F, w) = \sum_{i=1}^r [w_i^F \log(\frac{w_i^F}{w_i}) - w_i^F + w_i] \quad (2.4)$$

χ^2 거리 함수에서와 유사하게 라그랑주 승수법을 사용하면 $g_{2i} = \exp(\lambda_2^T x_i)$ 가 구해지며 여기서 λ_2 는 Newton-Raphson 방법으로 추정한다.

2.3 Weighted forward Kullback-Leibler distance

$$D(w^F, w) = WKL(w^F, w) = - \sum_{i=1}^r [w_i \log(\frac{w_i^F}{w_i}) - w_i^F + w_i] \quad (2.5)$$

(2.5) 거리 함수를 이용하였을 때, $g_{3i} = \frac{1}{1 - \lambda_3^T x_i}$ 로 구해지며 λ_3 는 modified Newton-Raphson 방법으로 추정한다.

정의된 세 가지 거리 함수를 이용하여 얻어진 g_i 는 자료 수가 충분히 큰 경우, $g_{2i} = \exp(\lambda_2^T x_i) \approx 1 + \lambda_1^T x_i$ 이고 $g_{3i} = \frac{1}{1 - \lambda_3^T x_i} \approx 1 + \lambda_1^T x_i$ 가 되는 것이 알려졌다(Wu and Lu, 2016). 결국 자료의 수가 충분히 큰 경우에는 어느 거리 함수를 사용하여도 유사한 w_i^F 가 얻어지므로 추정 결과도 유사하게 된다.

3. 일반화회귀 추정량과 제안한 조각별 캘리브레이션 추정량

3.1 일반화회귀 추정량

p 차원 보조변수 $(X_1, X_2, \dots, X_p)^T = X$ 의 합 $t_X = (t_{X_1}, t_{X_2}, \dots, t_{X_p})^T$ 가 알려져 있고 r 개의 관심변수 y_i 의 표본 가중치 w_i 가 알려져 있다고 하자. 그러면 보조변수 x_i 를 이용하여 얻어진 일반화회귀(GREG: generalized regression) 추정량은 다음과 같다.

$$\hat{t}_{YGR} = \hat{t}_{Y,HT} + (t_X - \hat{t}_{X,HT})^T \hat{B} \quad (3.1)$$

여기서 $\hat{t}_{Y,HT} = \sum_{i=1}^r w_i y_i$ 이고 $\hat{t}_{X,HT} = \sum_{i=1}^r w_i x_i$ 인 호르비츠-톰슨 추정량이고 $\hat{B} = (\hat{B}_1, \hat{B}_2, \dots, \hat{B}_p)^T = \hat{B}(y)$ 이며 $\hat{B}$ 은 다음의 가중최소제곱추정량(weighted least squares estimator)으로 얻어진다.

$$\hat{B} = \left(\sum_{j=1}^r w_j x_j x_j^T \right)^{-1} \left(\sum_{j=1}^r w_j x_j y_j \right) \quad (3.2)$$

(3.1)에서 정의된 일반화회귀 추정량은 표본론에서 흔하게 사용되고 있으며 이 추정량의 특징은 Kim (2004)에서 확인할 수 있다. 이제 (2.3)에서 얻어진 가중치를 사용하여 구한 호르비츠-톰슨 추정량은 (3.1)이 된다. 즉 χ^2 거리를 사용한 캘리브레이션

추정량은 일반화회귀 추정량이 된다. 또한 2절에서 설명한 바와 같이 (2.4)와 (2.5)에서 정의된 거리를 사용하여 얻어진 캘리브레이션 추정량은 자료 수가 충분히 큰 경우는 근사적으로 일반화회귀 추정량이 된다. 결국 자료 수가 큰 경우에서 캘리브레이션 추정량은 선형 결과회귀모형을 가정하여 얻어진 추정량이므로 선형 결과회귀모형에서는 우수한 결과를 보인다. 그러나 선형이 아닌 비선형 결과회귀모형인 경우에는 좋지 않은 결과가 얻어질 수 있으므로 이를 해결할 수 있는 캘리브레이션 추정량의 개발이 필요하다.

3.2 조각별 캘리브레이션 추정량

기존의 캘리브레이션 방법은 최종 응답 자료를 이용하여 추정한 보조변수 총합이 알려진 모집단 총합과 일치하도록 하는 방법이다. 이 방법은 선형 자료모형에서 매우 효과적인 것으로 알려져 있는데 선형 결과회귀모형에서는 (2.1) 캘리브레이션 제약식이 효과적이기 때문이다. 반면 비선형 결과회귀모형에서 (2.1) 캘리브레이션 제약식을 사용하게 되면 얻어진 가중치를 사용한 캘리브레이션 추정량은 효과적이지 않을 수 있다. 이를 해결하기 위해 본 연구에서는 모집단을 조각 또는 세부 층으로 나눈 후 각 세부 층에서 추정한 보조변수 총합이 알려진 모집단 총합과 일치하도록 하는 방법을 제안한다. 즉 결과회귀모형을 조각별 결과회귀모형으로 확장하고, 각 조각 또는 세부 층별로 얻어진 캘리브레이션 제약식을 사용하는 조각별 캘리브레이션(segmented calibration) 추정량을 제안한다.

N 개의 모집단 자료에 보조변수 자료 $(x_1, x_2, \dots, x_N)$ 가 있다고 가정하자. 그리고 n 개의 표본을 추출한 후, r 개의 최종 응답 자료가 얻어졌다고 가정하자. 이때 r 개의 최종 응답 자료의 설계 가중치 w_i 는 알려져 있다고 가정한다. 이제 모집단을 L 개의 세부 층으로 나누고, h 세부 층에는 n_h 개의 표본 개수와 r_h 개의 최종 응답 자료가 얻어졌다고 가정하자. 세부 층 구성 방법은 Kang and Shin (2025a, b)에서 사용한 방법과 동일한 방법을 사용한다. 즉 r 개의 최종 응답 자료 $(x_1, x_2, \dots, x_r)$ 에서 얻어진 분위수를 세부 층 경계로 하고, 세부 층 개수 L 은 $\frac{r}{L} \geq 10$ 이 되도록 결정한다. 그러면 사용된 세부 층별 캘리브레이션 제약식은 다음과 같다.

$$\sum_{i=1}^{r_h} w_{hi}^F x_{hi} = t_{h,X}, h = 1, \dots, L \quad (3.3)$$

만약 χ^2 거리가 사용된다면 (2.2)를 이용한 최종 가중치는 (3.3)을 만족하면서 h 세부 층별로 얻어진다. 즉 h 세부 층에 포함된 가중치는 다음과 같다.

$$w_{hi}^F = w_{hi} \times [1 + (t_{h,X} - \hat{t}_{h,X,HT})' (\sum_{i=1}^{r_h} w_{hi} x_{hi} x_{hi}')^{-1} x_{hi}], h = 1, \dots, L \quad (3.4)$$

이 가중치를 이용하여 얻어진 결과는 조각별 일반화회귀(segmented GREG) 추정량이 된다. 당연히 (2.4)와 (2.5)에서 정의된 거리 함수를 이용하여 가중치를 보정할 수 있다. 즉 세부 층별로 정의된 (3.3)을 제약식으로 그리고 (2.4) 또는 (2.5)를 거리 함수로 이용하면 세부 층별로 가중치를 구할 수 있으며 구해진 세부 층별 가중치를 이용하여 최종적으로 총합을 추정하게 된다.

4. 모의실험

모의실험에서는 최근 활발한 연구가 진행되는 캘리브레이션 추정량의 단점을 보완하기 위해 제안한 조각별 캘리브레이션 추정량(segmented calibration estimator)의 성능을 확인한다. 특히 비선형 결과회귀모형인 경우에서 제안한 조각별 캘리브레이션 추정량이 기존의 추정량에 비해 우수한 결과를 주는지 확인한다. 사용된 거리에 따라 다양한 캘리브레이션 추정량이 제안되었으나 응답 자료 수가 큰 경우에는 근사적으로 일치하는 것으로 알려져 있고 본 연구에서는 조각별 캘리브레이션 추정량이 비선형 결과회귀모형에서 효과적으로 사용될 수 있는지를 확인하는 것이 주된 목적이므로 모의실험에서는 χ^2 거리를 사용한 조각별 캘리브레이션 추정량, 즉 조각별 일반화회귀 추정량의 성능을 살펴본다. 모의실험 자료 생성은 흔히 가구 조사와 사업체 조사에서 발생하는 보조변수를 고려하기 위해 균일분포와 감마분포를 고려하였으며 응답확률 모형은 가장 자연스러운 모형인 로지스틱 응답확률모형과 선형 응답확률모형을 고려하였다. 다만 캘리브레이션 추정량은 복합표본설계(complex survey)에서 사용이 가능하므로 다수의 보조변수를 사용하는 것이 타당하지만 모의실험을 간단히 하기 위해 하나의 보조변수를 사용하였다.

4.1 자료 생성 과정

4.1.1 보조변수 x_i 생성

감마분포와 균일분포를 따르는 $N = 10,000$ 개의 보조변수 자료를 생성하였다. 먼저 $x_i^* \stackrel{iid}{\sim} \text{Gamma}(1, 50)$ 를 생성하였으며 최종적으로 $x_i = 100 + x_i^*$ 을 보조자료로 사용

하였다. 이때 $x_i \geq 300$ 인 자료는 삭제하였다. 또한 $x_i \stackrel{iid}{\sim} \text{Unif}(100, 300)$ 에서도 자료를

생성하였다. 따라서 모든 보조변수 x_i 는 분포에 무관하게 100에서 300 사이의 값을 갖는다. 이는 실제 자료 분석에서 흔히 자료는 양수이고 특히 사업체 조사의 대표적인 층인 100인 이상 300인 미만 층을 고려한 자료 생성 방법이다.

4.1.2 결과회귀모형(ORM: outcome regression model)

정규분포를 따르는 회귀모형과 로그-선형모형을 고려한 결과회귀모형을 사용하여 관심변수 y_i 를 생성하였다.

$$\text{ORM1: 단순 회귀모형: } y_i = \beta_{10} + \beta_{11}x_i + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} N(0, \sigma_1^2)$$

여기서 $\beta_{10} = 10$, $\beta_{11} = 5$, $\sigma_1^2 = 900$ 을 사용하였다.

$$\text{ORM2: } y_i = \beta_{20} + \beta_{21} + \beta_{22}(x_i - 130)^2 + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} N(0, \sigma_2^2)$$

여기서 $\beta_{20} = 10$, $\beta_{21} = 5$, $\beta_{22} = 0.05$, $\sigma_2^2 = 900$ 을 사용하였다.

$$\text{ORM3: } y_i \stackrel{iid}{\sim} LN(\mu_i, \sigma_3^2)$$

여기서 $\mu_i = \beta_{30} + \beta_{31}x_i$, $\beta_{30} = 0.01$, $\beta_{31} = 0.03$, $\sigma_3^2 = 0.01$ 을 사용하였다.

ORM1 모형은 선형모형으로 기존의 방법이 잘 적용될 것으로 예상되며, ORM2와 ORM3는 제안된 방법이 더 효과적일 것으로 예상된다.

4.1.3 응답확률모형(RPM: response probability model)과 무응답 자료 생성

응답확률모형(RPM)은 다음의 선형모형과 로지스틱 모형을 고려하였다.

$$\text{RPM1: } p_i = a_0 + a_1y_i$$

$$\text{RPM2: } p_i = a_0 + a_1y_i + a_2x_i$$

$$\text{RPM3: } \ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1y_i$$

$$\text{RPM4: } \ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1y_i + b_2x_i$$

RPM1과 RPM3은 보조변수 x_i 의 영향을 받지 않으며 RPM2와 RPM4는 보조변수 x_i 의 영향을 받는다. RPM1과 RPM3에서 a_0 , a_1 그리고 b_0 , b_1 은 $p_y^{\min} = 0.3$, $p_y^{\max} = 0.9$ 를 이용하여 결정하였다. 여기서 $p_y^{\min}$, $p_y^{\max}$ 는 y 가 최솟값과 최댓값일 때의 응답확률이며, 이는 y 가 증가할 때 응답확률도 증가하는 경우이다. 또한, RPM2와 RPM4에서 $a_2 = 0.01$, $b_2 = 0.01$ 을 사용하고 a_0 , a_1 그리고 b_0 , b_1 은 $p_y^{\min} = 0.3$, $p_y^{\max} = 0.9$ 를 이용

하여 결정하였다. 또한 y 가 증가할 때 응답확률이 감소하는 경우를 살펴보기 위해 $p_y^{\min} = 0.9$, $p_y^{\max} = 0.3$ 인 경우도 살펴보았다. 이 방법은 Kang and Shin (2025a)에서도 사용하였으며 이 방법에서 얻어진 응답확률을 이용하여 응답 자료 r 개와 무응답 자료 $n-r$ 개의 자료가 생성되었다.

4.1.4 총화

$n = 500$ 개 자료에서 얻어진 r 개의 최종 응답 자료를 이용하여 L 개의 세부 층으로 나누었다. 구성된 세부 층에 따라 얻어진 보조변수의 모집단 세부 층 총합 정보가 매우 다양할 수 있으므로 본 연구에서는 다음의 경우를 살펴보았다.

[1] 모집단 자료에 보조변수가 없고 일부 세부 층 총합만 알려진 경우

CASE 1: 모집단의 하위 50%의 총합 t_{50X} 와 전체 총합 t_{100X} 가 알려진 경우

CASE 2: 모집단의 하위 25%의 총합 t_{25X} , 하위 75%의 세부 층 총합 t_{75X} 그리고 전체 총합 t_{100X} 가 알려진 경우

[2] 모집단 자료에 보조변수가 있어 다수의 세부 층 구성이 가능한 경우

모집단에 보조변수 자료 $(x_1, x_2, \dots, x_N)$ 가 존재하면 Kang and Shin (2025a)에서 제안한 방법으로 L 개의 세부 층을 나눈 후 해당하는 h 세부 층의 t_{hX} 를 구할 수 있다. 본 모의실험에서는 L 값으로 5와 20을 고려하였으며, 이때 층 경계는 보조변수 자료의 분위수를 사용하였다.

4.2 비교에 사용된 평균 추정량

모의실험을 간단히 하기 위해 단순 랜덤추출법을 사용하고 설계 가중치 $w_i = w = \frac{N}{n}$ 을 사용한다. 본 연구의 주된 목적은 비선형 결과회귀모형을 조각별로 나누어 선형모형으로 만든 조각별 결과회귀모형 사용이 평균 추정에 효과적인지를 살펴보는 것이다. 또한 충분한 수의 자료가 사용된다면 거리에 무관하게 동일한 결과가 얻어지므로 본 모의실험에서는 2장에서 설명한 χ^2 거리를 사용한 캘리브레이션 추정량인 일반화 회귀 추정량을 고려한다. 이에 본 모의실험에서는 일반화회귀 추정량에 기초한 다음의 네 추정량의 성능을 비교한다.

4.2.1 단순 랜덤 추정량

얻어진 최종 응답 자료의 단순평균을 사용한다. 즉 다음의 추정량을 사용한다.

$$\hat{\bar{Y}}_{SRS} = \frac{1}{r} \sum_{i=1}^r y_i \quad (4.1)$$

여기서 r 은 최종 응답 자료의 개수를 의미한다.

4.2.2 MAR을 가정한 추정량

성향점수보정추정량 $\hat{\bar{Y}}_{MAR} = \frac{1}{N} \sum_{i=1}^n w_i \frac{R_i}{\hat{p}_i^{MAR}} y_i$ 을 사용하며 여기서 $\hat{p}_i^{MAR}$ 은 로지스틱 회귀모형에서 추정된 값을 사용한다. 로지스틱 회귀모형에서는 종속변수로 응답 유무 변수인 R_i 를 사용하며 이때 R_i 는 응답이면 '1' 무응답이면 '0' 값을 갖는 이항변수이다. 로지스틱 회귀모형의 독립변수로 보조변수 x_i 를 사용하며 추정된 $\hat{p}_i^{MAR}$ 은 $\sum_{i=1}^r \frac{1}{\hat{p}_i^{MAR}} = n$ 을 만족하도록 보정한다. 성향점수보정추정량에서 $w_i = \frac{N}{n}$ 을 사용하며 최종적으로 다음의 추정량을 사용한다.

$$\hat{\bar{Y}}_{MAR} = \frac{1}{n} \sum_{i=1}^n \frac{R_i}{\hat{p}_i^{MAR}} y_i \quad (4.2)$$

4.2.3 기존의 캘리브레이션 추정량

알려진 보조자료 모집단 총합으로 가중치를 보정한 기존의 캘리브레이션 추정량으로 정의는 다음과 같다.

$$\hat{\bar{Y}}_{CAL1} = \frac{1}{N} \sum_{i=1}^n w_i^F R_i y_i \quad (4.3)$$

여기서 $w_i^F = w_i \times [1 + (t_X - \hat{t}_{X,HT})' (\sum_i w_i x_i x_i')^{-1} x_i]$, $w_i = \frac{N}{n}$ 이고 r 은 최종 응답 자료 수이다.

4.2.4 조각별 캘리브레이션 추정량

알려진 보조자료의 세부 층 모집단 총합으로 가중치를 보정한 캘리브레이션 추정량으로 정의는 (4.4)와 같다.

$$\hat{Y}_{CAL2} = \frac{1}{N} \sum_{h=1}^L \sum_{i=1}^{n_h} w_{hi}^{F2} R_i y_i \quad (4.4)$$

여기서 $w_{hi}^{F2} = w_{hi} \times [1 + (t_{h,X} - \hat{t}_{h,X,HT})' (\sum_{i=1}^{r_h} w_{hi} x_{hi} x_{hi}')^{-1} x_{hi}]$, $h = 1, \dots, L$, $w_{hi} = \frac{N}{n}$

이고 n_h , r_h 는 각각 h 세부 층에 포함된 표본 수와 최종 응답 자료 수이다.

4.3 사용된 비교통계량

(4.1)에서 (4.4)에 정의된 추정량은 다음의 비교통계량: 편향(bias), 절대상대편향(ARB: absolute relative bias) 그리고 제곱근 MSE(RMSE: root mean squared error)를 이용하여 결과의 성능을 비교하였다. 각 통계량의 정의는 다음과 같다.

$$\begin{aligned} Bias &= \frac{1}{K} \sum_{k=1}^K (\hat{Y}_k - \overline{Y}_k) \\ ARB &= \frac{1}{K} \sum_{k=1}^K \frac{|\hat{Y}_k - \overline{Y}_k|}{\overline{Y}_k} \\ RMSE &= \sqrt{\frac{1}{K} \sum_{k=1}^K (\hat{Y}_k - \overline{Y}_k)^2} \end{aligned}$$

여기서 $K=1,000$ 을 사용하였으며 반복마다 새로운 모집단을 생성하여 통계량을 계산하였다. 이는 생성된 특정 모집단의 영향을 줄이기 위함이다. 이에 k 번째 반복 모집단의 참값을 $\overline{Y}_k$ 로 표시하였다.

4.4 모의실험 결과

모의실험 결과가 <표 4.1>에서 <표 4.6>에 수록되었으며 보조변수가 균일분포를 따르는 경우와 감마분포를 따르는 경우로 나누어 정리하였다. 각 표에서 $p_y^{\min} = 0.3$, $p_y^{\max} = 0.9$ 인 경우를 RP I라 표시하였고 $p_y^{\min} = 0.9$, $p_y^{\max} = 0.3$ 인 경우를 RP II라 표시하였다. 또한 r_1 은 RP I에서의 최종 응답 자료 개수이고, r_2 는 RP II에서의 최종 응답 자료 개수이다.

4.4.1 보조변수가 균일분포를 따르는 경우

먼저 <표 4.1>은 결과회귀모형이 선형인 경우에서 얻어진 결과이다. 무응답을 처리하지 않은 $\hat{Y}_{SRS}$ 은 모든 응답확률모형 RPM에서 그리고 모든 응답확률 RP에서 매우 큰 값의 편향이 관찰되었고 ARB와 RMSE 또한 매우 커서 무응답 처리가 필요한 것을 확인할 수 있다. $\hat{Y}_{MAR}$ 는 MAR 무응답 가정에서 얻어진 응답확률을 사용한 성향점수보정추정량 결과로 $\hat{Y}_{SRS}$ 과 비교해 모든 비교통계량 기준으로 우수한 결과를 준다. 다음으로 기존의 캘리브레이션 추정량인 $\hat{Y}_{CAL1}$ 결과를 살펴보자. 먼저 $\hat{Y}_{MAR}$ 결과와 비교하면 편향 기준으로는 우수성을 확인할 수 없다. 그러나 ARB와 RMSE를 기준으로 살펴보면 $\hat{Y}_{CAL1}$ 의 결과가 매우 우수한 것을 확인할 수 있다. 따라서 결과회귀모형이 선형인 경우에는는 캘리브레이션 추정량 사용이 매우 타당하다. 이제 본 연구에서 제안한 추정량인 $\hat{Y}_{CAL2}$ 를 살펴보자. 먼저 모집단 하위 50%의 총합 t_{50X} 와 전체 총합 t_{100X} 가 알려진 경우인 $\hat{Y}_{CAL2}^{CASE1}$ 결과를 살펴보면 일부 편향 결과를 제외하면 $\hat{Y}_{CAL2}^{CASE1}$ 보다 기존 캘리브레이션 추정량인 $\hat{Y}_{CAL1}$ 이 우수한 결과를 주는 것을 확인할 수 있다. 또한 모집단의 하위 25% t_{25X} , 하위 75% t_{75X} , 전체 총합 t_{100X} 가 알려진 경우에서 얻어진 $\hat{Y}_{CAL2}^{CASE2}$ 결과는 모든 추정량에서 그리고 모든 응답확률모형 RPM과 모든 응답확률 RP에서 $\hat{Y}_{CAL2}^{CASE1}$ 결과보다 나쁜 것을 확인할 수 있다. 이러한 현상은 세부 층의 개수가 증가하면서 더욱 큰 차이를 보이며 나쁜 결과를 준다. 물론 세부 층의 개수가 큰 경우에서 얻어진 $\hat{Y}_{CAL2(L=20)}$ 결과를 포함한 $\hat{Y}_{CAL2}$ 결과와 $\hat{Y}_{MAR}$ 결과를 비교하면 $\hat{Y}_{CAL2}$ 가 우수한 결과를 준다. 결론적으로 결과회귀모형이 선형이라 판단되는 경우에는 기존의 $\hat{Y}_{CAL1}$ 추정량을 사용하는 것이 타당하다.

이제 결과회귀모형이 2차식인 <표 4.2> 결과를 살펴보자. $\hat{Y}_{SRS}$ 와 $\hat{Y}_{MAR}$ 결과는 <표 4.1>과 유사한 패턴을 보인다. 반면 $\hat{Y}_{CAL1}$ 은 모든 RPM과 RP에서 매우 좋지 않은 결과를 준다. 특히 $\hat{Y}_{CAL1}$ 의 편향은 매우 큰 값을 보이고 있으며 이에 따라 큰 ARB와 RMSE가 얻어져 $\hat{Y}_{MAR}$ 의 편향보다 나쁜 결과를 준다. 이는 잘못된 캘리브레이션 제약식을 사용하였기 때문으로 판단되며, 따라서 결과회귀모형이 2차식인 경우에는는 $\hat{Y}_{CAL1}$ 을 사용하는 것보다 $\hat{Y}_{MAR}$ 을 사용하는 것이 타당하다. $\hat{Y}_{CAL2}^{CASE1}$ 는 $\hat{Y}_{CAL1}$ 에 비해 모든 비교통계량에서 우수한 결과를 주지만 $\hat{Y}_{MAR}$ 가 더욱 우수한 결과를 주므로 $\hat{Y}_{MAR}$ 을 사용하는 것이 타당하다. $\hat{Y}_{CAL2}^{CASE2}$ 와 $\hat{Y}_{MAR}$ 를 비교하면 편향은 $\hat{Y}_{MAR}$ 가 우수하지만 $\hat{Y}_{CAL2}^{CASE2}$ 가 ARB와 RMSE 기준으로 우수하므로 $\hat{Y}_{CAL2}^{CASE2}$ 를 사용하는 것이

타당하다. 이제 세부 층의 수가 5와 20인 경우를 살펴보면 제안한 조각별 캘리브레이션 추정량은 매우 우수한 결과를 보인다. 특히 $\hat{Y}_{CAL2(L=20)}$ 결과를 살펴보면 모든 RPM과 RP에서 매우 우수한 결과를 주는 것을 확인할 수 있다.

<표 4.3>은 결과회귀모형이 이분산 모형인 로그-선형모형으로 <표 4.2>에서 얻어진 결과와 유사한 패턴의 결과를 보인다. 특히 기존의 $\hat{Y}_{CAL1}$ 은 큰 편향을 보이며 이에 따라 좋지 않은 결과를 보인다. 또한 RP II에서는 세부 층 개수가 3인 $\hat{Y}_{CAL2}^{CASE2}$ 가 $\hat{Y}_{MAR}$ 에 비해 나쁜 결과를 보인다. 따라서 조각별 캘리브레이션 추정량이 우수한 결과를 얻기 위해서는 다수의 세부 층 개수가 필요하다고 판단되며 본 모의실험에서는 최소 $L=5$ 이상에서 $\hat{Y}_{MAR}$ 보다 $\hat{Y}_{CAL2}$ 가 우수한 결과가 얻어진 것을 확인할 수 있다.

결론적으로 선형 결과회귀모형에서는 기존 캘리브레이션 추정량인 $\hat{Y}_{CAL1}$ 을 사용하는 것이 타당하며 비선형 결과회귀모형에서는 본 연구에서 제안한 조각별 캘리브레이션 추정량인 $\hat{Y}_{CAL2}$ 을 사용하는 것이 타당하다. 다만 세부 층 개수 결정은 결과회귀모형에 맞도록 결정해야 할 것으로 판단한다.

<표 4.1> 보조자료가 균일분포이고, 결과회귀모형이 선형모형(ORM1)인 경우

RPM	no of responses	EST	RP I			RP II		
			BIAS	ARB	RMSE	BIAS	ARB	RMSE
1	$r_1 = 300$ $r_2 = 300$	$\hat{Y}_{SRS}$	73.06	7.233	74.85	-73.85	7.311	75.58
		$\hat{Y}_{MAR}$	-0.129	0.998	12.79	0.221	0.996	12.67
		$\hat{Y}_{CAL1}$	-0.223	0.121	1.543	-0.457	0.144	1.817
		$\hat{Y}_{CAL2}^{CASE1}$	0.370	0.132	1.682	-0.738	0.149	1.884
		$\hat{Y}_{CAL2}^{CASE2}$	0.565	0.141	1.795	-0.818	0.152	1.912
		$\hat{Y}_{CAL2(L=5)}$	0.718	0.149	1.872	-0.865	0.154	1.942
		$\hat{Y}_{CAL2(L=20)}$	0.823	0.158	1.972	-0.880	0.159	2.002
2	$r_1 = 300$ $r_2 = 300$	$\hat{Y}_{SRS}$	76.96	7.619	78.64	-77.65	7.687	79.30
		$\hat{Y}_{MAR}$	-0.599	1.006	12.91	0.714	0.999	12.70
		$\hat{Y}_{CAL1}$	-0.486	0.126	1.594	-0.137	0.140	1.773
		$\hat{Y}_{CAL2}^{CASE1}$	0.102	0.129	1.642	-0.449	0.142	1.799
		$\hat{Y}_{CAL2}^{CASE2}$	0.285	0.136	1.732	-0.536	0.144	1.820
		$\hat{Y}_{CAL2(L=5)}$	0.440	0.141	1.784	-0.589	0.146	1.852
		$\hat{Y}_{CAL2(L=20)}$	0.565	0.149	1.887	-0.610	0.151	1.908
3	$r_1 = 322$ $r_2 = 322$	$\hat{Y}_{SRS}$	72.86	7.213	74.49	-73.64	7.289	75.18
		$\hat{Y}_{MAR}$	0.884	1.004	12.83	-0.791	0.993	12.68
		$\hat{Y}_{CAL1}$	-0.234	0.119	1.509	-0.387	0.137	1.735
		$\hat{Y}_{CAL2}^{CASE1}$	0.351	0.130	1.624	-0.707	0.143	1.811
		$\hat{Y}_{CAL2}^{CASE2}$	0.552	0.139	1.738	-0.806	0.146	1.845
		$\hat{Y}_{CAL2(L=5)}$	0.713	0.146	1.816	-0.861	0.149	1.880
		$\hat{Y}_{CAL2(L=20)}$	0.825	0.155	1.923	-0.872	0.154	1.925
4	$r_1 = 321$ $r_2 = 321$	$\hat{Y}_{SRS}$	73.59	7.286	75.21	-74.31	7.357	75.84
		$\hat{Y}_{MAR}$	0.846	1.006	12.86	-0.748	0.993	12.67
		$\hat{Y}_{CAL1}$	-0.284	0.119	1.514	-0.331	0.136	1.722
		$\hat{Y}_{CAL2}^{CASE1}$	0.296	0.128	1.611	-0.655	0.142	1.789
		$\hat{Y}_{CAL2}^{CASE2}$	0.497	0.137	1.718	-0.752	0.144	1.821
		$\hat{Y}_{CAL2(L=5)}$	0.656	0.144	1.794	-0.811	0.147	1.858
		$\hat{Y}_{CAL2(L=20)}$	0.767	0.152	1.892	-0.821	0.152	1.905

<표 4.2> 보조자료가 균일분포이고, 결과회귀모형이 2차 모형(ORM2)인 경우

RPM	no of responses	EST	RP I			RP II		
			BIAS	ARB	RMSE	BIAS	ARB	RMSE
1	$r_1 = 262,$ $r_2 = 337$	$\hat{\bar{Y}}_{SRS}$	225.4	15.85	230.1	-176.8	12.43	180.0
		$\hat{\bar{Y}}_{MAR}$	-4.015	1.682	30.30	-8.436	1.822	32.79
		$\hat{\bar{Y}}_{CAL1}$	117.6	8.270	118.9	-55.38	3.895	57.84
		$\hat{\bar{Y}}_{CAL2}^{CASE1}$	30.64	2.154	32.02	-31.22	2.197	33.81
		$\hat{\bar{Y}}_{CAL2}^{CASE2}$	13.36	0.946	14.78	-18.54	1.330	21.22
		$\hat{\bar{Y}}_{CAL2(L=5)}$	5.423	0.407	6.872	-8.925	0.675	11.38
		$\hat{\bar{Y}}_{CAL2(L=20)}$	1.048	0.140	2.456	-1.134	0.169	2.991
2	$r_1 = 274,$ $r_2 = 326$	$\hat{\bar{Y}}_{SRS}$	218.3	15.35	222.7	-185.1	13.01	188.2
		$\hat{\bar{Y}}_{MAR}$	-2.692	1.684	30.30	-3.843	1.792	32.17
		$\hat{\bar{Y}}_{CAL1}$	111.3	7.826	112.6	-55.72	3.918	58.28
		$\hat{\bar{Y}}_{CAL2}^{CASE1}$	29.01	2.040	30.39	-30.29	2.134	33.12
		$\hat{\bar{Y}}_{CAL2}^{CASE2}$	12.40	0.878	13.81	-17.80	1.285	20.71
		$\hat{\bar{Y}}_{CAL2(L=5)}$	4.873	0.374	6.340	-8.359	0.643	11.02
		$\hat{\bar{Y}}_{CAL2(L=20)}$	0.831	0.130	2.341	-0.937	0.170	3.032
3	$r_1 = 278,$ $r_2 = 357$	$\hat{\bar{Y}}_{SRS}$	232.7	16.37	236.8	-167.9	11.81	171.0
		$\hat{\bar{Y}}_{MAR}$	-4.957	1.699	30.62	-14.88	1.930	34.57
		$\hat{\bar{Y}}_{CAL1}$	115.3	8.107	116.5	-56.91	4.002	59.10
		$\hat{\bar{Y}}_{CAL2}^{CASE1}$	30.05	2.114	31.40	-34.02	2.393	36.17
		$\hat{\bar{Y}}_{CAL2}^{CASE2}$	13.11	0.928	14.46	-20.88	1.481	23.05
		$\hat{\bar{Y}}_{CAL2(L=5)}$	5.193	0.386	6.507	-10.39	0.753	12.44
		$\hat{\bar{Y}}_{CAL2(L=20)}$	1.033	0.134	2.397	-1.334	0.173	3.078
4	$r_1 = 281,$ $r_2 = 355$	$\hat{\bar{Y}}_{SRS}$	230.3	16.20	234.3	-170.1	11.96	173.2
		$\hat{\bar{Y}}_{MAR}$	-4.464	1.697	30.57	-14.22	1.915	34.32
		$\hat{\bar{Y}}_{CAL1}$	113.6	7.990	114.8	-57.37	4.035	59.55
		$\hat{\bar{Y}}_{CAL2}^{CASE1}$	29.69	2.088	31.03	-34.11	2.399	36.29
		$\hat{\bar{Y}}_{CAL2}^{CASE2}$	12.83	0.908	14.16	-20.84	1.479	23.06
		$\hat{\bar{Y}}_{CAL2(L=5)}$	4.991	0.373	6.310	-10.29	0.748	12.36
		$\hat{\bar{Y}}_{CAL2(L=20)}$	0.969	0.133	2.381	-1.332	0.171	3.043

<표 4.3> 보조자료가 균일분포이고, 결과회귀모형이 로그-선형모형(ORM3)인 경우

RPM	no of responses	EST	RP I			RP II		
			BIAS	ARB	RMSE	BIAS	ARB	RMSE
1	$r_1 = 198,$ $r_2 = 401$	$\hat{\bar{Y}}_{SRS}$	646.9	49.87	665.9	-316.2	24.38	324.9
		$\hat{\bar{Y}}_{MAR}$	108.7	9.237	142.5	-106.3	8.819	136.0
		$\hat{\bar{Y}}_{CAL1}$	676.2	52.14	686.6	-216.6	16.70	227.0
		$\hat{\bar{Y}}_{CAL2}^{CASE1}$	208.0	16.04	219.2	-174.0	13.42	184.0
		$\hat{\bar{Y}}_{CAL2}^{CASE2}$	94.15	7.291	103.7	-133.3	10.29	144.3
		$\hat{\bar{Y}}_{CAL2(L=5)}$	40.73	3.251	49.16	-85.16	6.670	96.57
		$\hat{\bar{Y}}_{CAL2(L=20)}$	10.09	1.095	17.43	-18.57	1.801	28.99
2	$r_1 = 231,$ $r_2 = 368$	$\hat{\bar{Y}}_{SRS}$	559.1	43.11	576.3	-347.9	26.82	356.3
		$\hat{\bar{Y}}_{MAR}$	65.50	6.940	110.3	-77.60	7.423	118.7
		$\hat{\bar{Y}}_{CAL1}$	563.4	43.44	573.4	-217.9	16.81	229.6
		$\hat{\bar{Y}}_{CAL2}^{CASE1}$	161.7	12.50	173.3	-165.1	12.75	177.4
		$\hat{\bar{Y}}_{CAL2}^{CASE2}$	73.75	5.771	84.86	-124.3	9.663	138.3
		$\hat{\bar{Y}}_{CAL2(L=5)}$	31.28	2.620	40.45	-77.86	6.202	92.31
		$\hat{\bar{Y}}_{CAL2(L=20)}$	6.541	0.948	15.26	-15.38	1.735	28.25
3	$r_1 = 205,$ $r_2 = 412$	$\hat{\bar{Y}}_{SRS}$	703.8	54.27	720.5	-279.5	21.54	289.3
		$\hat{\bar{Y}}_{MAR}$	103.8	8.882	137.1	-111.2	9.052	138.1
		$\hat{\bar{Y}}_{CAL1}$	704.3	54.30	713.4	-193.7	14.93	204.8
		$\hat{\bar{Y}}_{CAL2}^{CASE1}$	196.3	15.14	206.7	-161.7	12.47	171.5
		$\hat{\bar{Y}}_{CAL2}^{CASE2}$	89.24	6.911	98.07	-128.0	9.877	137.8
		$\hat{\bar{Y}}_{CAL2(L=5)}$	37.98	3.041	46.13	-84.46	6.569	94.10
		$\hat{\bar{Y}}_{CAL2(L=20)}$	9.076	1.018	16.39	-19.67	1.789	28.47
4	$r_1 = 213,$ $r_2 = 408$	$\hat{\bar{Y}}_{SRS}$	678.6	52.32	695.3	-288.5	22.24	298.1
		$\hat{\bar{Y}}_{MAR}$	93.62	8.294	129.3	-109.7	8.976	137.2
		$\hat{\bar{Y}}_{CAL1}$	674.4	55.00	683.7	-198.2	15.28	209.2
		$\hat{\bar{Y}}_{CAL2}^{CASE1}$	186.8	14.41	197.0	-163.9	12.64	173.7
		$\hat{\bar{Y}}_{CAL2}^{CASE2}$	84.59	6.568	93.73	-128.9	9.947	139.0
		$\hat{\bar{Y}}_{CAL2(L=5)}$	35.47	2.875	43.84	-84.90	6.611	94.84
		$\hat{\bar{Y}}_{CAL2(L=20)}$	8.475	0.982	15.73	-19.29	1.786	28.60

4.4.2 보조변수가 감마분포를 따르는 경우

<표 4.4>에서 <표 4.6>은 보조변수가 감마분포를 따르는 경우에서 얻어진 결과이다. <표 4.1>에서 <표 4.3>에서 확인하였듯이 <표 4.4>에서 <표 4.6>에서도 응답확률 모형인 RPM에 무관하게 유사한 패턴의 추정량 성능 결과가 얻어졌다. 이에 <표 4.4>에서 <표 4.6>에는 독립변수가 있는 로지스틱 응답확률모형인 RPM 4 결과만을 수록하였다. 이제 선형 결과회귀모형에서 얻어진 <표 4.4>를 살펴보면 <표 4.1>과 유사하게 $\hat{Y}_{CAL1}$ 이 가장 우수한 결과를 준다. 또한 본 연구에서 제안한 $\hat{Y}_{CAL2}$ 는 세부 층의 개수가 커질수록 모든 비교통계량에서 나빠지지만 $\hat{Y}_{MAR}$ 보다는 우수한 결과를 준다. 이러한 경향은 응답확률 RP I과 RP II에서 모두 얻어진다. 다음으로 2차 결과회귀모형인 <표 4.5> 결과를 살펴보면 $\hat{Y}_{CAL1}$ 보다 $\hat{Y}_{MAR}$ 가 우수한 결과를 보이므로 2차 결과회귀모형에서는 $\hat{Y}_{MAR}$ 를 사용하는 것이 타당하다. 그러나 모집단 자료에 보조변수 자료가 존재하고, 다수의 세부 층을 구성할 수 있다면 본 연구에서 제안한 추정량인 $\hat{Y}_{CAL2}$ 를 사용하는 것이 타당하다. 다만 보조변수가 균일분포인 경우보다 더 많은 수의 세부 층이 구성되어야 한다. 본 연구의 모의실험에 의하면 $L=5$ 이상에서 $\hat{Y}_{MAR}$ 보다 $\hat{Y}_{CAL2}$ 가 우수한 것을 확인할 수 있다. <표 4.6>은 로그-선형 결과회귀모형 결과로 $\hat{Y}_{CAL1}$ 보다 $\hat{Y}_{MAR}$ 가 우수하다. $\hat{Y}_{CAL2}$ 을 살펴보면, 특히 RP I의 $L=5$ 일 때 모든 비교통계량에서 $\hat{Y}_{MAR}$ 가 우수한 결과를 준다. 이에 $L=10$ 인 경우의 모의실험을 추가로 수행하였으며 결과는 BIAS=11.89, ARB=6.766 그리고 RMSE=18.16으로 얻어졌다. 따라서 조각별 캘리브레이션 추정량을 사용한다면 세부 층의 개수 $L=10$ 이상인 $\hat{Y}_{CAL2}$ 를 사용할 때 우수한 결과를 얻을 수 있을 것으로 판단한다.

<표 4.4> 보조자료가 Gamma(1,50)이고, 결과회귀모형이 선형모형(ORM1)인 경우

RPM	no of responses	EST	RP I			RP II		
			BIAS	ARB	RMSE	BIAS	ARB	RMSE
4	$r_1 = 248,$ $r_2 = 390$	$\hat{Y}_{SRS}$	54.40	7.338	56.55	-29.63	3.996	31.08
		$\hat{Y}_{MAR}$	0.679	1.057	9.835	-0.936	1.069	9.906
		$\hat{Y}_{CAL1}$	-0.113	0.188	1.761	-0.213	0.167	1.576
		$\hat{Y}_{CAL2}^{CASE1}$	0.814	0.222	2.073	-0.343	0.170	1.592
		$\hat{Y}_{CAL2}^{CASE2}$	0.973	0.231	2.157	-0.390	0.171	1.605
		$\hat{Y}_{CAL2(L=5)}$	1.078	0.238	2.231	-0.430	0.172	1.618
		$\hat{Y}_{CAL2(L=20)}$	1.099	0.247	2.321	-0.480	0.177	1.664

<표 4.5> 보조자료가 Gamma(1,50)이고, 결과회귀모형이 2차모형(ORM2)인 경우

RPM	no of responses	EST	RP I			RP II		
			BIAS	ARB	RMSE	BIAS	ARB	RMSE
4	$r_1 = 204,$ $r_2 = 419$	$\hat{\bar{Y}}_{SRS}$	111.7	13.27	117.0	-42.29	5.036	45.33
		$\hat{\bar{Y}}_{MAR}$	1.603	1.679	17.81	-5.664	1.763	18.52
		$\hat{\bar{Y}}_{CAL1}$	71.09	8.446	73.13	-14.31	1.735	16.27
		$\hat{\bar{Y}}_{CAL2}^{CASE1}$	30.79	3.658	32.99	-12.79	1.554	14.69
		$\hat{\bar{Y}}_{CAL2}^{CASE2}$	16.31	1.947	18.32	-10.93	1.341	12.85
		$\hat{\bar{Y}}_{CAL2(L=5)}$	7.167	0.905	9.144	-8.440	1.054	10.30
		$\hat{\bar{Y}}_{CAL2(L=20)}$	1.207	0.276	2.870	-3.080	0.463	4.812

<표 4.6> 보조자료가 Gamma(1,50)이고, 결과회귀모형이 로그-선형모형(ORM3)인 경우

RPM	no of responses	EST	RP I			RP II		
			BIAS	ARB	RMSE	BIAS	ARB	RMSE
4	$r_1 = 169,$ $r_2 = 439$	$\hat{\bar{Y}}_{SRS}$	111.9	53.82	123.0	-30.94	15.44	36.35
		$\hat{\bar{Y}}_{MAR}$	22.25	13.18	34.44	-14.72	10.01	25.32
		$\hat{\bar{Y}}_{CAL1}$	169.4	81.29	178.0	-20.75	11.30	27.73
		$\hat{\bar{Y}}_{CAL2}^{CASE1}$	99.79	47.92	107.4	-20.94	11.04	27.01
		$\hat{\bar{Y}}_{CAL2}^{CASE2}$	66.27	31.93	73.33	-20.00	10.54	25.76
		$\hat{\bar{Y}}_{CAL2(L=5)}$	36.03	17.65	43.24	-18.40	9.615	23.47
		$\hat{\bar{Y}}_{CAL2(L=20)}$	3.555	3.171	8.873	-10.22	5.750	14.53

이상의 결과를 종합하면 선형 결과회귀모형에서는 기존의 캘리브레이션 추정량 $\hat{\bar{Y}}_{CAL1}$ 이 매우 효과적으로 사용될 수 있음을 확인하였다. 반면 비선형 결과회귀모형에서는 $\hat{\bar{Y}}_{CAL1}$ 보다 $\hat{\bar{Y}}_{MAR}$ 를 사용하는 것이 타당하지만 만약 충분한 수의 세부 층 모총합 정보가 있거나 모집단에 보조변수 자료가 포함되어 있어 충분한 수의 세부 층을 구성할 수 있다면 본 연구에서 제안한 $\hat{\bar{Y}}_{CAL2}$ 추정량 사용이 더욱 효과적으로 사용될 수 있음을 확인하였다.

5. 결론

MNAR 무응답을 적절히 처리하기 위한 다양한 방법이 연구되고 있다. MNAR 무응답의 응답확률은 관심변수의 함수이고, 무응답이 관심변수에서 발생하기 때문에 정확한 응답확률을 추정하는 것은 매우 어렵다. 이에 모집단 보조변수 정보를 이용하여 응답확률 추정의 정확성을 향상하는 방법을 연구하고 있다. 특히 결과회귀모형이 선형모형인 경우, 알려진 모집단 보조변수 총합에 추정된 보조변수 총합이 일치하도록 가중치를 보정하는 캘리브레이션 방법은 매우 효과적이다. 반면 본 연구에서도 확인하였지만, 비선형 결과회귀모형에 기존 캘리브레이션 추정량을 적용하면 정확한 평균 추정값을 얻기가 어렵다. 이에 본 연구에서는 조각별 캘리브레이션 추정량을 제안하였으며 특히 비선형 결과회귀모형에서는 매우 우수한 결과가 얻어짐을 확인하였다.

보조변수가 균일분포이고 결과회귀모형이 2차식인 경우의 모의실험에서는 3개로 구성된 세부 층을 사용한 경우에서도 조각별 캘리브레이션 추정량이 RMSE 기준으로 우수한 결과를 주는 것을 확인하였다. 그러나 실제 자료 분석에서는 결과회귀모형이 선형 모형인지 비선형 모형인지 확인하기 어려우며, 비선형 모형의 형태 또한 확인하기가 어렵다. 이제 비선형 모형이라 확인되어 제안된 조각별 캘리브레이션 추정량을 사용하기 위해서는 다수의 보조변수 세부 층 총합이 알려져 있어야 한다. 만약 모집단에 보조변수가 존재하여 다수의 세부 층 구성이 가능하다면 본 연구에서 제안한 조각별 캘리브레이션 추정량 사용은 우수한 결과를 줄 것으로 기대한다. 본 연구에서 제안한 조각별 캘리브레이션 추정량 사용을 위해서는 세부 층 개수 결정이 매우 중요하다. 만약 비선형 모형이 확실하다고 판단되는 경우는 모의실험 결과에 따라 10개 이상의 세부 층을 구성한다면 우수한 결과를 얻을 수 있다고 판단한다. 다만 최적 세부 층 개수 결정은 향후 추가 연구가 필요하다. 또한 향후 연구를 통해 비선형 결과회귀 모형의 모형식별이 가능하다면 보조변수의 변환 방법으로 비선형성 문제를 해결하여 추정의 정확성이 향상할 것으로 기대한다.

(2026년 4월 9일 접수, 2026년 5월 14일 수정, 2026년 5월 24일 채택)

참고문헌

- Kang, J. and Shin, K.-I. (2025a). A Study on Calibration Estimator to Improve the Accuracy of the Mean Estimation for Nonignorable Nonresponse in Infinite Population, *Survey Research*, 26(4), 117-142.
- Kang, J. and Shin, K.-I. (2025b). A post-stratification method to improve the accuracy of the naive propensity score adjusted estimator in MNAR nonresponse, *The Korean Journal of Applied Statistics*, 38(6), 775 - 792.
- Kim, K. S. (2004). General Regression Estimators in Survey Sampling. *Survey Research*, 5(2), 49-70.
- Lim, Y., and Park, M. (2024). Use of ridge calibration method in predicting election results. *Journal of the Korean Statistical Society*, 53(2), 479-494.
- Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3), 723-734.
- Kim, J.-K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling, *Statistica Sinica*, 24(1), 375-394.
- Uehara, M., Lee, D., and Kim, J. K. (2023). Statistical inference with semiparametric nonignorable nonresponse models. *Scandinavian Journal of Statistics*, 50(4), 1795-1817.
- Wu, C., and Lu, W. W. (2016). Calibration weighting methods for complex surveys. *International Statistical Review*, 84(1), 79-98.

Segmented calibration estimator for missing not at random nonresponse in nonlinear outcome regression model⁴⁾

Juwon Jeong⁵⁾ · Key-Il Shin⁶⁾

Abstract

Various methods to properly handle nonresponses have been developed. Nonresponse following MCAR (missing completely at random) or MAR (missing at random) mechanism is easy to handle, and accurate estimation methods are developed. Recently, research on appropriate handling of MNAR (missing not at random) nonresponse has been actively conducted. In particular, various calibration methods are being studied, and the calibration estimator is known to be effective when the outcome regression model is linear. In this study, we propose a segmented calibration estimator that can be used for nonlinear outcome regression models and perform the simulation studies to confirm the superiority of the proposed estimator.

Key Words: Segmented regression, generalized regression estimator, response probability model

4) This research was supported by Hankuk University of Foreign Studies research fund (2026)

5) First author, Graduate student, Department of Statistics, Hankuk University of Foreign Studies, E-mail: 2001jjuwon@naver.com

6) Corresponding author, Professor, Department of Statistics, Hankuk University of Foreign Studies, E-mail: keyshin@hufs.ac.kr