Multiple Bias Calibration under Nonignorable Nonresponse

Yumou Qiu [1]

Peking University

[1] Joint work with Jae-Kwang Kim and Seonghun Cho

Basic Setup

- We assume

$$(X_i, Y_i) \overset{iid}{\sim} F$$

for $i = 1, 2, \ldots, n$, for unknown distribution F.

- Missing response

$$\delta_i = \begin{cases} 1 & \text{observe } y_i \\ 0 & \text{otherwise.} \end{cases}$$

- MNAR: $\delta \not\perp Y \mid X$.

- propensity score (PS) function: $\pi(x, y) = P(\delta = 1 \mid X = x, Y = y)$.

Research Problem

- Estimating equation $E\{U(\theta; X_i, Y_i)\} = 0$.
- If $\pi(x, y)$ is known (survey sampling), then we can easily estimate θ by solving

$$\sum_{i=1}^{n} \frac{\delta_i}{\pi(x_i, y_i)} U(\theta; x_i, y_i) = 0, \tag{1}$$

- Unknown $\pi(x, y)$: many existing approaches use a "single" model for $\pi(x, y)$, say $\pi(x, y; \phi)$, and use the estimated $\hat{\pi}(x, y) = \pi(x, y; \hat{\phi})$ in (1). That is, use

$$\sum_{i=1}^{n} \frac{\delta_i}{\pi(x_i, y_i; \hat{\phi})} U(\theta; x_i, y_i) = 0. \tag{2}$$

The implicit assumption for (2) is that the model is correctly specified and $\hat{\phi}$ is consistent.

Proposal (Overview)

- Challenge for MNAR: models for $\pi(x, y; \phi)$ can not be validated based on observed data.
- Proposal: Instead of employing a single model for $\pi(x, y)$, we consider multiple models for $\pi(x, y)$, say $\pi_1(x, y; \phi_1), \cdots, \pi_K(x, y; \phi_K)$, and develop a statistical tool for incorporating the multiple PS models.
- Claim: As long as one of the K candidates models is correctly specified, the resulting estimator is consistent.
- This is similar in spirit to multiple robust estimation (Han and Wang, 2013; Han, 2014; Chen and Haziza, 2017) but we do not necessarily assume that the candidate PS models follow the missing at random (MAR) assumption of Rubin (1976).

Multiple robustness under MAR (Han, 2014)

- Empirical likelihood framework: $\max \sum_{i=1}^{n} \delta_i \log(p_i)$.
- Key: balance the estimated working PS $\pi(x, \hat{\phi}_k)$ on the observed and full samples for $k = 1, \ldots, K$:

$$\sum_{i=1}^{n} \delta_i p_i \pi(x_i, \hat{\phi}_k) = \frac{1}{n} \sum_{i=1}^{n} \pi(x_i, \hat{\phi}_k).$$

- Not applicable for MNAR: $\sum_{i=1}^{n} \pi(x_i, y_i, \hat{\phi}_k)$ is not available under MNAR.
- New approaches need to be proposed for multiple robustness under MNAR.

Contribution

- We have developed a unified framework for handling selection bias with multiple PS models.
- Make the overall estimation less dependent on the PS model assumptions.
- Approach 1: require a validation sample without missing from the same population to provide a pilot (initial) estimate.

  If the cost of collecting the voluntary sample is much lower than the cost of collecting a probability sample, the proposed method can be a cost-effective statistical tool for obtaining data.
- Approach 2: do not need a validation sample but could be computationally heavy.

Parametric likelihood under known PS

- Assume that F has a density $f(x, y; \theta)$, for some $\theta \in \Omega \subset \mathbb{R}^p$
- The observed likelihood is

$$L_{obs}(\theta) = \prod_{i=1}^{n} \{\pi(x_i, y_i) f(x_i, y_i; \theta)\}^{\delta_i} \times \{1 - W(\theta)\}^{n-n_1} \quad (3)$$

  where

$$W(\theta) = \iint \pi(x, y) f(x, y; \theta) dxdy.$$

- $n_1 = \sum_{i=1}^{n} \delta_i$ is the sample size of non-missing data.

Semiparametric model

- (X, Y): a vector of random variables satisfying

$$\mathbb{E}\{U(\theta; X, Y)\} = 0$$

for some function $U(\theta; x, y)$ with unknown parameter θ.

- That is, the model with distribution function F should satisfy

$$\mathbb{E}\{U(\theta; X, Y)\} \equiv \int U(\theta; x, y) dF(x, y) = 0 \tag{4}$$

for all θ, where F is completely unspecified other than the restriction in (4). Thus, it is a semiparametric model.

- For simplicity, we assume that the rank of $U(\theta; X, Y)$ is equal to the dimension of parameter space. (Just-identified model)

Empirical likelihood (Qin et al., 2002)

- Empirical likelihood (EL) method: discretizing F (a semiparametric approach)

- Let $\pi_i = \pi(x_i, y_i)$ and let S be the index set of the sample (with $\delta_i = 1$).

- The observed log-likelihood can be written as

$$\ell_{obs} = \ell_{obs}(p, W) = \sum_{i \in S} (\log \pi_i + \log p_i) + (n - n_1) \log(1 - W) \tag{5}$$

where $p_i \geqslant 0$,

$$\sum_{i \in S} p_i = 1, \tag{6}$$

$$\sum_{i \in S} p_i \pi_i = W \tag{7}$$

and

$$\sum_{i \in S} p_i U(\theta; x_i, y_i) = 0. \tag{8}$$

Heuristics for empirical likelihood

- Constraints (6)–(8) are due to

$$dF(x,y) = f(x,y) = \frac{f(x,y \mid \delta = 1)}{\pi(x,y)} P(\delta = 1)$$

- Classical EL weight: $f(x_i, y_i \mid \delta = 1) = n_1^{-1}$ for the observed sample
- $p_i = \{n\pi(x_i, y_i)\}^{-1}$

Remark

- Constraint (8) does not play any role in determining $\hat{p}_i$.
- For a fixed W, the maximizer of $\ell_{obs}(p, W)$ w.r.t. p_i is

$$\hat{p}_i = \hat{p}_i(W) = \frac{1}{n_1} \cdot \frac{1}{1 + \hat{\lambda}_1(\pi_i - W)}, \tag{9}$$

  where $\hat{\lambda}_1$ satisfies

$$\sum_{i \in S} \hat{p}_i \pi_i = W. \tag{10}$$

- For unknown W, we can obtain $\widehat{W}$ by finding the maximizer of the profile observed likelihood give by

$$\ell_{obs,p}(W) = \ell_{obs}\left(\hat{p}(W), W\right)$$

  where $\ell_{obs}\left(p, W\right)$ is defined in (5).

Maximum empirical likelihood estimator

- The final estimator of θ as the solution to

$$\sum_{i \in S} \hat{p}_i U(\theta; x_i, y_i) = 0,$$

where

$$\hat{p}_i = \hat{p}_i(\widehat{W})$$

is computed by (9) evaluated at $W = \widehat{W}$.

- The solution $\hat{\theta}$: maximum empirical likelihood estimator (MELE) of θ.

- Asymptotically, we can show that $\hat{\lambda}_1$ in (9) converges in probability to $\lambda_1^* = n/n_1$. Thus, we have

$$\hat{p}_i \approx \frac{1}{n_1} \cdot \frac{1}{1 + \lambda_1^*(\pi_i - \widehat{W})} \approx \frac{1}{n\pi_i}. \tag{11}$$

Understanding (7)

- Recall the constraint (7):

$$\sum_{i \in S} p_i \pi_i = W.$$

- Constraint (7) is another way of finding the sampling weight for removing the selection bias. The EL method provides an "implicit" weighting, while the Horvitz-Thompson method provides an explicit weighting with $w_i = 1/\pi_i$.

- Constraint (7) can be called the internal bias correction (IBC) condition (Firth and Bennett, 1998).

- The EL estimator of θ using W in (7) is more efficient than the EL estimator of θ using

$$\sum_{i \in S} p_i \pi_i = \frac{n_1}{n}.$$

Remark

- For $\theta = E(Y)$, if $E(Y \mid x) = x'\beta$ for some β, we can include

$$\sum_{i \in S} p_i x_i = \bar{x}_n, \tag{12}$$

where $\bar{x}_n = n^{-1} \sum_{i=1}^{n} x_i$, to the constraints in the EL method.

- Including (12) in addition to (6) and (7) in the EL optimization, we obtain

$$\hat{p}_i = \frac{1}{n_1} \frac{1}{1 + \hat{\lambda}_1(\pi_i - \hat{W}) + \hat{\boldsymbol{\lambda}}_2'(x_i - \bar{x}_n)}$$

where $\hat{\lambda}_1$ and $\hat{\boldsymbol{\lambda}}_2$ satisfy (7) and (12).

- Double robustness:

$$\sum_{i \in S} \hat{p}_i y_i = \frac{1}{N} \sum_{i=1}^{N} \left\{ \boldsymbol{x}_i'\hat{\beta} + \frac{\delta_i}{\pi_i} \left(y_i - \boldsymbol{x}_i'\hat{\beta} \right) \right\} + o_p(n^{-1/2}).$$

Correctly specified PS model

- We assume that π_i is unknown but the PS model

$$P(\delta = 1|X, Y) = \pi(X, Y; \phi)$$

is correctly specified. Note that we allow that the selection model is MNAR.

- $\tilde{\pi}(x; \phi) = P(\delta = 1 \mid x; \phi) = \int \pi(x, y; \phi) f(y \mid x) dy$.

Pfefferman-Sverchkov formula:

$$\frac{1}{\tilde{\pi}(x; \phi)} = E\left\{ \frac{1}{\pi(x, Y; \phi)} \middle| x, \delta = 1 \right\}. \tag{13}$$

- Model $\pi(X, Y) = \pi(X, Y; \phi)$ is identified if and only if the mapping

$$\phi \mapsto E\{O(x, Y; \phi) \mid \boldsymbol{x}, \delta = 1\}$$

is one-to-one, almost everywhere (Morikawa and Kim, 2021).

EL procedure under a correct PS model

1. Obtain the MLE $\hat{\phi}$ for the PS model $\pi(X, Y) = \pi(X, Y; \phi)$.

2. Using the estimate $\hat{\phi}$ obtained from Step 1, obtain $\hat{p}_i$ by finding the maximizer of

$$\ell(p, W) = \sum_{i \in S} \log(p_i) + (N - n) \log(1 - W)$$

subject to $\sum_{i \in S} p_i = 1$, the covariate balancing constraint in (12) and

$$\sum_{i \in S} p_i \pi(x_i, y_i; \hat{\phi}) = W.$$

3. Using $\hat{p}_i$ obtained from Step 2, find $\hat{\theta}$ by solving

$$\sum_{i \in S} \hat{p}_i U(\theta; x_i, y_i) = 0.$$

Theoretical result

- $\hat{\theta}$ is $n^{1/2}$-consistent.

- Influence function:

$$\sqrt{n}(\hat{\theta}-\theta_0) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tau_u^{-1}(\boldsymbol{\omega}_{i,1} - B_0\boldsymbol{\omega}_{i,2} + \boldsymbol{\omega}_{i,3}) + o_p(1),$$

  where $\tau_u = \{\partial U(\theta_0; X, Y)/\partial\theta\}$.

- $U_i = U(\theta_0; x_i, y_i)$, $g_i = (1, (x_i - E(X))')'$, $h(x, y; \phi) = \partial\pi(x, y; \phi)/\partial\phi$

- $\boldsymbol{\omega}_{i,1} = \delta_i U_i/\pi_i$, $\boldsymbol{\omega}_{i,2} = \delta_i g_i/\pi_{i,0} - g_i$, and $\boldsymbol{\omega}_{i,3} = (1 - \delta_i/\pi_i)\boldsymbol{\alpha} b_i$,

$$\boldsymbol{\alpha} = E\left\{\frac{(U(\theta_0; X, Y) - B_0(X))h(X, Y; \phi_0)'}{\pi(X, Y; \phi_0)}\right\}E^{-1}\left\{\frac{b(X; \phi_0)h(X, Y; \phi_0)'}{\pi(X, Y; \phi_0)}\right\}.$$

Remark

- $\hat{\theta}$ is optimal within a class of projection estimators if ϕ and $E(X_i)$ are known.
- $\hat{\theta}$ is more efficient than the EL estimator without the covariate balancing constraints (Liu and Fan, 2023) under certain conditions.
- Under MAR and $E(U_i \mid X_i = x_i) = x_i'\beta$, $\hat{\theta}$ is the semiparametric efficient estimator.

Multiple PS models

- In practice, we do not know the true PS model.
- We can extend the EL method for multiple PS models. We consider K different PS models, $\pi_k(x, y; \phi_k)$ for $k = 1, \cdots, K$, and hope that the true model belong to one of the K candidate models. This is the basic idea of multiple robust estimation which was developed under MAR case only so far (Han and Wang, 2013; Han, 2014; Chen and Haziza, 2017).
- Unfortunately, we cannot use Pfefferman-Sverchkov formula in (13) because it is valid only under correctly specified PS model. (i.e., $\pi(x, y; \phi)$ is correct.)

Proposal–estimation of PS models

(1) Estimating equation approach: solving

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\delta_i}{\pi_k(x_i, y_i; \phi_k)} - 1\right\} b(x_i; \phi) = 0, \tag{14}$$

with respect to ϕ_k for $k = 1, \ldots, K$.

(2) Obtain $\widehat{p}$ and $\widehat{W}$ by finding the maximizer of the empirical log-likelihood

$$\ell(p, W) = \sum_{i=1}^{N} \delta_i \log p_i + (N - n)\log(1 - W) \quad (15)$$

subject to $\sum_{i=1}^{N} \delta_i p_i = 1$, the covariate balancing constraint in (12) and

$$\sum_{i=1}^{N} \delta_i p_i \pi_k(X_i, Y_i; \widehat{\phi}_k) = W \quad (16)$$

for all $k = 1, \ldots, K$.

(3) Using $\widehat{p}_i$ obtained from Step 2, find $\widehat{\theta}$ by solving

$$\sum_{i=1}^{N} \delta_i \widehat{p}_i U(\theta; X_i, Y_i) = 0. \quad (17)$$

A key condition for working PS models

Assume $\widehat{\phi}_k$ converges to ϕ_{k0} in probability, and ϕ_{k0} satisfies

$$E\pi_k(X, Y; \phi_{k0}) = P(\delta = 1)$$

- Automatically satisfied for the correct PS model
- May not be satisfied for incorrect PS models
- Automatically satisfied for logistical regression with intercept term under MAR, estimated by MLE
- Another challenge for MNAR

Multiple robustness of MCEL estimator $\widehat{\theta}$

- Constraint (16) is the bias-calibration condition for K different PS models. So, we can call it as the multiple bias calibration (MBC) condition.
- As long as one of the K different PS models is correctly specified, the proposed MCEL estimator $\widehat{\theta}$ is consistent and asymptotic normal.

Modified estimation equation for PS models

- One Approach to guarantee $E\pi_k(X, Y; \phi_{k0}) = P(\delta = 1)$
- Need outcome model for the conditional density $f(Y \mid X; \beta)$ of Y given X.
- $\tilde{\pi}(X; \phi, \beta) = E\{\pi(X, Y; \phi)|X\}$.
- Modified estimation equation: solving

$$\sum_{i=1}^{n} \mathbf{\Gamma}(\phi \mid \beta^*; \delta_i, x_i, y_i) = 0$$

where

$$\mathbf{\Gamma}(\phi \mid \beta; \delta, \boldsymbol{x}, y) = (\delta - \tilde{\pi}(x; \boldsymbol{\phi}, \beta), \{\delta\pi(x, y; \boldsymbol{\phi})^{-1} - 1\}b_{-1}(x; \boldsymbol{\phi})')'$$

Augmented PS model

- Another Approach to guarantee $E\pi_k(X, Y; \phi_{k0}) = P(\delta = 1)$
- Need outcome model for the conditional density $f(Y \mid X; \beta)$ of Y given X.
- Augmented PS model $\pi_{aug}(x, y; \boldsymbol{\phi}_{aug})$ where $\phi'_{aug} = (\phi', a)$ and

$$\log \frac{\pi_{aug}(x, y; \phi_{aug})}{1 - \pi_{aug}(x, y; \phi_{aug})} = a + \log \frac{\pi(x, y; \phi)}{1 - \pi(x, y; \phi)}. \tag{18}$$

Let $\widehat{\boldsymbol{\phi}}$ be the solution to the estimating equation in (14). Given $\widehat{\phi}$, we obtain $\hat{a}$ by solving

$$\sum_{i=1}^{n} \tilde{\pi}_{aug}(\boldsymbol{x}_i; \widehat{\boldsymbol{\phi}}, a, \beta^*) = n_1.$$

Outcome model

- Both approaches require the outcome regression model $f(Y \mid X; \beta)$ and a pilot estimate β^* of β.
- This can be obtained if we have a validation sample without missingness.
- The validation sample only provides an initial estimate of β.
- The estimate of β can be refined by the proposed MCEL procedure.
- β can also be estimated if there is a non-response instrumental variable.

Outcome model–without validation sample or IV

- A numerical minimization approach
- Consider the initial estimate $\widetilde{\beta}$ as input and the MCEL estimate $\widehat{\beta}$ as output.
- $\widehat{\beta} = G(\widetilde{\beta})$ for a function $G(\cdot)$
- Q_n: difference between initial value $\widetilde{\beta}$ and updated estimate $\widehat{\beta}$ by MCEL.
- Minimize Q_n to obtain the pilot estimator β^*:

$$\beta^* = \arg\min_{\beta} Q_n(\beta) \quad \text{for} \quad Q_n(\beta) = \|G(\beta) - \beta\|^2.$$

Simulation setup

- **Outcome model**: $Y = 0.5 + X_1 + X_2 + \epsilon$
- **RM1** (MAR) :
$$\text{logit}(\pi_B(X, Y)) = 0.4 + 0.5X_1 + 0.5X_2. \tag{19}$$
- **RM2** (NMAR) :
$$\text{logit}(\pi_B(X, Y)) = 0.2 + 0.5X_1 + 0.5Y. \tag{20}$$
- **RM3** (NMAR) :
$$\text{logit}(\pi_B(X, Y)) = 0.2 + 0.5X_2 + 0.5Y. \tag{21}$$

Methods

- **HTA** : Horvitz-Thompson estimator on a validation sample with 10% of the sample size:

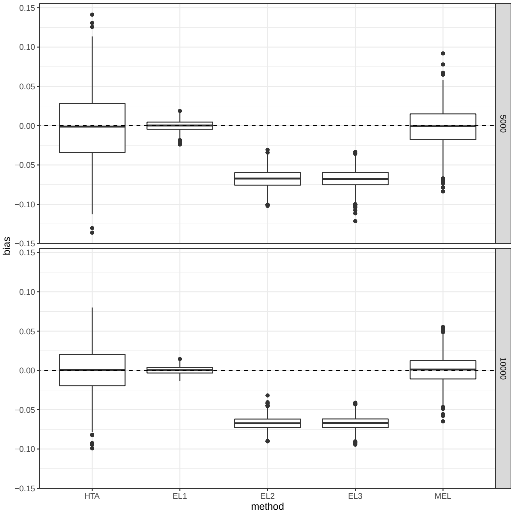
$$\hat{\mu}_{HTA} = n_A^{-1} \sum_{i \in A} y_i.$$

- **EL1**, **EL2**, **EL3** : For each method, we assume that the response model is

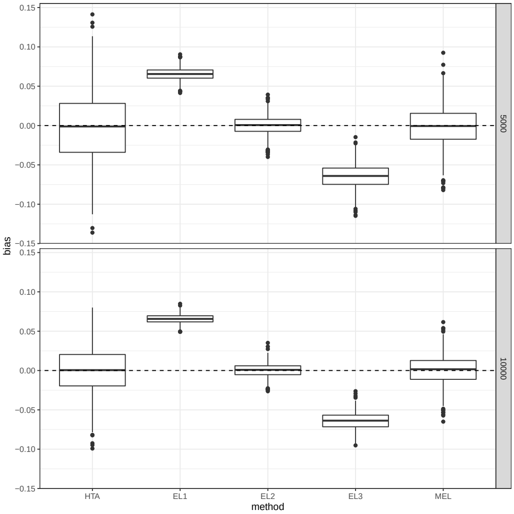
$$EL1 : \text{logit}(\pi(X, Y; \phi)) = \phi_0 + \phi_1 X_1 + \phi_2 X_2,$$
$$EL2 : \text{logit}(\pi(X, Y; \phi)) = \phi_0 + \phi_1 X_1 + \phi_2 Y,$$
$$EL3 : \text{logit}(\pi(X, Y; \phi)) = \phi_0 + \phi_1 X_2 + \phi_2 Y.$$

- **MEL**: use all three PS models

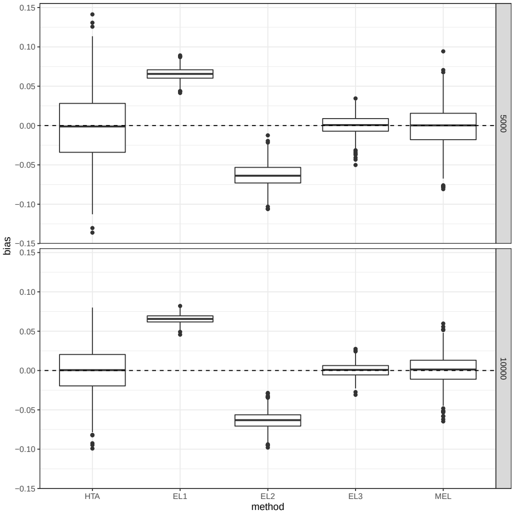
Box plot of $\widehat{\mu} - \mu_0$ under **RM1**



Box plot of $\widehat{\mu} - \mu_0$ under **RM2**



Box plot of $\widehat{\mu} - \mu_0$ under **RM3**



Discussion

- The EL-based method based on a single PS model is unbiased and efficient under the correct model setup, but it is severely biased when the PS model is not correctly specified.
- The proposed EL method incorporating the multiple PS models through MBC condition shows good performances for all scenarios considered, as the multiple candidate models contain the true PS model.
- Under the correct model, the EL method using a single PS model is more efficient than the MEL method using multiple PS models. It is because the MBC condition sacrifices the efficiency to obtain multiple robustness.
- The proposed estimator is more efficient than the estimator solely based on the validation sample.

Thank You

References I

Chen, S. and D. Haziza (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. Biometrika 104, 439–453.

Firth, D. and K. E. Bennett (1998). Robust models in probability sampling. Journal of the Royal Statistical Society: Series B 60, 3–21.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. Journal of the American Statistical Association 109(507), 1159–1173.

Han, P. and L. Wang (2013). Estimation with missing data: Beyond double robustness. Biometrika 100, 417–430.

Liu, Y. and Y. Fan (2023). Biased-sample empirical likelihood weighting for missing data problems: an alternative to inverse probability weighting. Journal of the Royal Statistical Society Series B: Statistical Methodology.

Morikawa, K. and J. K. Kim (2021). Semiparametric optimal estimation with nonignorable nonresponse data. The Annals of Statistics 49(5), 2991–3014.

References II

Qin, J., D. Leung, and J. Shao (2002). Estimation with survey data under non-ignorable nonresponse or informative sampling. Journal of the American Statistical Association 97, 193–200.

Rubin, D. B. (1976). Inference and missing data. Biometrika 63(3), 581–592.