Bayesian model averaging in **multiple imputation** under **informative sampling**

Gyuhyeong Goh

Department of Statistics
Kansas State University, Manhattan, KS

Joint work with
Jae Kwang Kim
Iowa State University
Korean Advanced Institute of Science & Technology

Introduction

- The analysis of survey data has received great attention.
- In practice, we frequently observe that some responses are missing.
- When missing values occur in survey data, it requires a special care to avoid the bias problem due to sampling design.
- In survey sampling, multiple imputation (MI) provides an effective way to handle missing data (Rubin, 1987; Kim and Yang, 2017)

Introduction

- The analysis of survey data has received great attention.
- In practice, we frequently observe that some responses are missing.
- When missing values occur in survey data, it requires a special care to avoid the bias problem due to sampling design.
- In survey sampling, multiple imputation (MI) provides an effective way to handle missing data (Rubin, 1987; Kim and Yang, 2017)

Introduction (cont.)

- In many cases, we may consider multiple candidate models for the data. Unfortunately, the existing MI procedure cannot account for the model uncertainty.
- Today, we will introduce a new multiple imputation method by combining multiple imputation with Bayesian model averaging (BMA).
- We restrict our attention to variable selection, which is the most common challenge in recent statistical problems.

Introduction (cont.)

- In many cases, we may consider multiple candidate models for the data. Unfortunately, the existing MI procedure cannot account for the model uncertainty.
- Today, we will introduce a new multiple imputation method by combining multiple imputation with Bayesian model averaging (BMA).
- We restrict our attention to variable selection, which is the most common challenge in recent statistical problems.

Basic setup

- $\mathcal{F}_N = \{(\boldsymbol{x}_i, y_i); i \in \mathcal{U}\}$: Finite population
  - a random sample from an infinite population with joint density $f(y|\boldsymbol{x};\boldsymbol{\theta})f(\boldsymbol{x})$.
  - y is a scalar response variable
  - $\boldsymbol{x}$ is a p-dim vector of possible explanatory variables.
  - $\mathcal{U} = \{1, 2, \ldots, N\}$: index set of finite population.

- $\boldsymbol{D}_n = \{(\boldsymbol{x}_i, y_i); i \in \mathcal{S}\}$: Sample
  - obtained by a probability sampling design from the finite population.
  - $\mathcal{S} = \{1, 2, \ldots, n\}$: index set of sample ($\mathcal{S} \subset \mathcal{U}$)

Basic setup

- $\mathcal{F}_N = \{(\boldsymbol{x}_i, y_i); i \in \mathcal{U}\}$: Finite population
  - a random sample from an infinite population with joint density $f(y|\boldsymbol{x};\boldsymbol{\theta})f(\boldsymbol{x})$.
  - y is a scalar response variable
  - $\boldsymbol{x}$ is a p-dim vector of possible explanatory variables.
  - $\mathcal{U} = \{1, 2, \ldots, N\}$: index set of finite population.
- $\boldsymbol{D}_n = \{(\boldsymbol{x}_i, y_i); i \in \mathcal{S}\}$: Sample
  - obtained by a probability sampling design from the finite population.
  - $\mathcal{S} = \{1, 2, \ldots, n\}$: index set of sample ($\mathcal{S} \subset \mathcal{U}$)

Assumption: MAR

- Suppose that y_i is subject to missingness, while $\boldsymbol{x}_i$ is fully observed.
  - Define
$$\delta_i = \begin{cases} 1 \text{ if } y_i \text{ is observed} \\ 0 \text{ if } y_i \text{ is missing} \end{cases}.$$
- We consider the missing mechanism to be (population) missing at random (MAR) in the sense that
$$\Pr(y_i \in B | \boldsymbol{x}_i, \delta_i = 1) = \Pr(y_i \in B | \boldsymbol{x}_i)$$
for any measurable set B and $i \in \mathcal{U}$.
- Let $\boldsymbol{y}_{\mathrm{obs}}$ and $\boldsymbol{y}_{\mathrm{mis}}$, respectively, be the observed part and the missing part in $\boldsymbol{y}_n = (y_1, \ldots, y_n)$

Assumption: Informative sampling design

- Define
$$I_i = \begin{cases} 1 \text{ if unit } i \text{ is sampled} \\ 0 \text{ o.w.} \end{cases}.$$

- We assume that the sampling process is informative, that is,
$$\Pr(y_i \in B | \boldsymbol{x}_i, I_i = 1) \neq \Pr(y_i \in B | \boldsymbol{x}_i)$$
for any measurable set B.

Parameter of interest

- We are mainly interested in estimating domains satisfying

$$\eta_{\boldsymbol{x}} = E\{g(Y)|\boldsymbol{x}\} = \int g(y)f(y|\boldsymbol{x};\boldsymbol{\theta})dy,$$

  where g is a known function.

- For example,

$$\begin{aligned} \eta_{\boldsymbol{x}} &= E(Y|\boldsymbol{x}), \\ \eta_{\boldsymbol{x}} &= \Pr(Y < 1|\boldsymbol{x}). \end{aligned}$$

Data structure

- Suppose that we have a sample of size $n = 10$ as follows:

$$\left[\boldsymbol{y}_n, \boldsymbol{X}_n, \boldsymbol{w}_n\right] = \left[\begin{array}{c|c|c} y_1 & \boldsymbol{x}_1 & w_1 \\ y_2 & \boldsymbol{x}_2 & w_2 \\ y_3 & \boldsymbol{x}_3 & w_3 \\ y_4 & \boldsymbol{x}_4 & w_4 \\ y_5 & \boldsymbol{x}_5 & w_5 \\ y_6 & \boldsymbol{x}_6 & w_6 \\ y_7 & \boldsymbol{x}_7 & w_7 \\ y_8 & \boldsymbol{x}_8 & w_8 \\ y_9 & \boldsymbol{x}_9 & w_9 \\ y_{10} & \boldsymbol{x}_{10} & w_{10} \end{array}\right],$$

where w_i is the sampling weight for the i^{th} observation.

Data structure

- Suppose that we have a sample of size $n = 10$ as follows:

$$\left[\boldsymbol{y}_n, \boldsymbol{X}_n, \boldsymbol{w}_n\right] = \left[\begin{array}{c|c|c} y_1 & \boldsymbol{x}_1 & w_1 \\ y_2 & \boldsymbol{x}_2 & w_2 \\ \times & \boldsymbol{x}_3 & w_3 \\ y_4 & \boldsymbol{x}_4 & w_4 \\ y_5 & \boldsymbol{x}_5 & w_5 \\ \times & \boldsymbol{x}_6 & w_6 \\ y_7 & \boldsymbol{x}_7 & w_7 \\ \times & \boldsymbol{x}_8 & w_8 \\ y_9 & \boldsymbol{x}_9 & w_9 \\ y_{10} & \boldsymbol{x}_{10} & w_{10} \end{array}\right],$$

where w_i is the sampling weight for the i^{th} observation.

Estimation under informative sampling

- For the *complete data*, we can obtain the pseudo maximum likelihood estimator of $\boldsymbol{\theta}$ by solving

$$S_w(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}} w_i S(\boldsymbol{\theta} | \boldsymbol{x}_i, y_i) = \mathbf{0},$$

where $S(\boldsymbol{\theta}|\boldsymbol{x}, y) = \partial \log f(y|\boldsymbol{x}; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$.

- Under some regularity conditions, we have that

$$\left\{ \hat{V}(\hat{\theta}) \right\}^{-1/2} \left(\theta - \hat{\theta} \right) \xrightarrow{d} \mathcal{N}(0, \mathrm{I}),$$

where $\hat{V}(\hat{\theta})$ is a design-consistent estimator of $V(\hat{\theta})$.

- Since $\eta_{\boldsymbol{x}} = \eta_{\boldsymbol{x}}(\theta)$, we have that

$$\left\{ \eta_{\boldsymbol{x}}'(\hat{\theta})^{\mathrm{T}} \hat{V}(\hat{\theta}) \eta_{\boldsymbol{x}}'(\hat{\theta}) \right\}^{-1/2} \left(\eta_{\boldsymbol{x}} - \eta_{\boldsymbol{x}}(\hat{\theta}) \right) \xrightarrow{d} \mathcal{N}(0, 1).$$

Estimation under informative sampling

- For the *complete data*, we can obtain the pseudo maximum likelihood estimator of $\boldsymbol{\theta}$ by solving

$$S_w(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}} w_i S(\boldsymbol{\theta} | \boldsymbol{x}_i, y_i) = \mathbf{0},$$

  where $S(\boldsymbol{\theta}|\boldsymbol{x}, y) = \partial \log f(y|\boldsymbol{x}; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$.

- Under some regularity conditions, we have that

$$\left\{ \hat{V}(\hat{\boldsymbol{\theta}}) \right\}^{-1/2} \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right) \xrightarrow{d} \mathcal{N}(0, \mathrm{I}),$$

  where $\hat{V}(\hat{\boldsymbol{\theta}})$ is a design-consistent estimator of $V(\hat{\boldsymbol{\theta}})$.

- Since $\eta_{\boldsymbol{x}} = \eta_{\boldsymbol{x}}(\boldsymbol{\theta})$, we have that

$$\left\{ \eta'_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}})^{\mathrm{T}} \hat{V}(\hat{\boldsymbol{\theta}}) \eta'_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}) \right\}^{-1/2} \left(\eta_{\boldsymbol{x}} - \eta_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}) \right) \xrightarrow{d} \mathcal{N}(0, 1).$$

MI under informative sampling

The MI procedure can be implemented in the following three steps:

Step 1. Create R complete datasets, $\{\boldsymbol{D}_n^{(r)} = (\boldsymbol{y}_n^{(r)}, \boldsymbol{X}_n) : r = 1, \ldots, R\}$.

- Generate $\boldsymbol{y}_{\text{mis}}^{(r)} \overset{iid}{\sim} f(\boldsymbol{y}_{\text{mis}} | \boldsymbol{X}_n, \boldsymbol{y}_{\text{obs}})$.

Step 2. (i) Compute $\hat{\theta}^{(r)}$ and $\hat{V}(\hat{\theta}^{(r)})$ using each imputed dataset. (ii)Then compute

$$\hat{\eta}_x^{(r)} = \eta_x(\hat{\theta}^{(r)}) \quad \text{and} \quad \hat{V}_x^{(r)} = \left\{\eta_x'(\hat{\theta}^{(r)})\right\}^{\mathrm{T}} \hat{V}(\hat{\theta}^{(r)}) \left\{\eta_x'(\hat{\theta}^{(r)})\right\},$$

Step 3. Using Rubin (1987)'s formula, compute

$$\hat{\eta}_{\mathrm{MI}} = R^{-1} \sum_{r=1}^{R} \hat{\eta}_x^{(r)} \quad \text{and} \quad \hat{V}_{\mathrm{MI}} = W_R + \left(1 + R^{-1}\right) B_R,$$

where $W_R = R^{-1} \sum_{r=1}^{R} \hat{V}_x^{(r)}$ and $B_R = (R-1)^{-1} \sum_{r=1}^{R} (\hat{\eta}_x^{(r)} - \hat{\eta}_{\mathrm{MI}})^2$.

MI under informative sampling

The MI procedure can be implemented in the following three steps:

Step 1. Create R complete datasets, $\{\boldsymbol{D}_n^{(r)} = (\boldsymbol{y}_n^{(r)}, \boldsymbol{X}_n) : r = 1, \ldots, R\}$.

- Generate $\boldsymbol{y}_{\text{mis}}^{(r)} \overset{iid}{\sim} f(\boldsymbol{y}_{\text{mis}} | \boldsymbol{X}_n, \boldsymbol{y}_{\text{obs}})$.

Step 2. (i) Compute $\hat{\boldsymbol{\theta}}^{(r)}$ and $\hat{V}(\hat{\boldsymbol{\theta}}^{(r)})$ using each imputed dataset.
(ii)Then compute

$$\hat{\eta}_{\boldsymbol{x}}^{(r)} = \eta_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}^{(r)}) \quad \text{and} \quad \hat{V}_{\boldsymbol{x}}^{(r)} = \left\{\eta_{\boldsymbol{x}}'(\hat{\boldsymbol{\theta}}^{(r)})\right\}^{\mathrm{T}} \hat{V}(\hat{\boldsymbol{\theta}}^{(r)}) \left\{\eta_{\boldsymbol{x}}'(\hat{\boldsymbol{\theta}}^{(r)})\right\},$$

Step 3. Using Rubin (1987)'s formula, compute

$$\hat{\eta}_{\mathsf{MI}} = R^{-1} \sum_{r=1}^{R} \hat{\eta}_{\boldsymbol{x}}^{(r)} \quad \text{and} \quad \hat{V}_{\mathsf{MI}} = W_R + \left(1 + R^{-1}\right) B_R,$$

where $W_R = R^{-1} \sum_{r=1}^{R} \hat{V}_{\boldsymbol{x}}^{(r)}$ and $B_R = (R-1)^{-1} \sum_{r=1}^{R} (\hat{\eta}_{\boldsymbol{x}}^{(r)} - \hat{\eta}_{\mathsf{MI}})^2$.

MI under informative sampling

The MI procedure can be implemented in the following three steps:

Step 1. Create R complete datasets, $\{\boldsymbol{D}_n^{(r)} = (\boldsymbol{y}_n^{(r)}, \boldsymbol{X}_n) : r = 1, \ldots, R\}$.

- Generate $\boldsymbol{y}_{\mathrm{mis}}^{(r)} \overset{iid}{\sim} f(\boldsymbol{y}_{\mathrm{mis}} | \boldsymbol{X}_n, \boldsymbol{y}_{\mathrm{obs}})$.

Step 2. (i) Compute $\hat{\boldsymbol{\theta}}^{(r)}$ and $\hat{V}(\hat{\boldsymbol{\theta}}^{(r)})$ using each imputed dataset. (ii)Then compute

$$\hat{\eta}_{\boldsymbol{x}}^{(r)} = \eta_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}^{(r)}) \quad \text{and} \quad \hat{V}_{\boldsymbol{x}}^{(r)} = \left\{\eta_{\boldsymbol{x}}'(\hat{\boldsymbol{\theta}}^{(r)})\right\}^{\mathrm{T}} \hat{V}(\hat{\boldsymbol{\theta}}^{(r)}) \left\{\eta_{\boldsymbol{x}}'(\hat{\boldsymbol{\theta}}^{(r)})\right\},$$

Step 3. Using Rubin (1987)'s formula, compute

$$\hat{\eta}_{\mathrm{MI}} = R^{-1} \sum_{r=1}^{R} \hat{\eta}_{\boldsymbol{x}}^{(r)} \quad \text{and} \quad \hat{V}_{\mathrm{MI}} = W_R + \left(1 + R^{-1}\right) B_R,$$

where $W_R = R^{-1} \sum_{r=1}^{R} \hat{V}_{\boldsymbol{x}}^{(r)}$ and $B_R = (R-1)^{-1} \sum_{r=1}^{R} (\hat{\eta}_{\boldsymbol{x}}^{(r)} - \hat{\eta}_{\mathrm{MI}})^2$.

Remark in Step 1 of MI

- To generate $\boldsymbol{y}^*_{\mathrm{mis}}$ from $f(\boldsymbol{y}_{\mathrm{mis}}|\boldsymbol{X}_n, \boldsymbol{y}_{\mathrm{obs}})$, the following two-step procedure is commonly used.
  - (a) Generate $\boldsymbol{\theta}^*$ from $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{\mathrm{obs}}, \boldsymbol{X}_n)$.
  - (b) Generate y_i^* from $f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}^*)$ for each unit $i \in \{i \in \mathcal{S} : \delta_i = 0\}$.

- In step (a), the data augmentation algorithm of Tanner and Wong (1987) can be implemented by iterating following two steps until convergence:
  1. **Imputation:** For given θ^*, generate $\boldsymbol{y}^*_{\mathrm{mis}} = \{y_i^* : \delta_i = 0, i \in \mathcal{S}\}$ from

$$y_i^* \sim f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}^*).$$

  2. **Posterior sampling:** For given $\boldsymbol{D}_n^* = (\boldsymbol{X}_n, \boldsymbol{y}^*_n,)$, generate $\boldsymbol{\theta}^*$ from

$$\theta^* \sim \pi(\theta|\boldsymbol{D}_n^*) = \frac{f(\boldsymbol{D}_n^*|\theta)\pi(\theta)}{\int f(\boldsymbol{D}_n^*|\theta)\pi(\theta)d\theta}. \quad (1)$$

- caution! However, under informative sampling, determination of the complete-sample likelihood $f(\boldsymbol{D}_n|\theta)$ is very challenging.
  $(\because) \quad f(\boldsymbol{y}_n|\boldsymbol{X}_n, \theta) \neq \prod_{i\in\mathcal{S}} f(y_i|\boldsymbol{x}_i; \theta)$.

Remark in Step 1 of MI

- To generate $\boldsymbol{y}^*_{\rm mis}$ from $f(\boldsymbol{y}_{\rm mis}|\boldsymbol{X}_n, \boldsymbol{y}_{\rm obs})$, the following two-step procedure is commonly used.
  (a) Generate $\boldsymbol{\theta}^*$ from $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{\rm obs}, \boldsymbol{X}_n)$.
  (b) Generate y_i^* from $f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}^*)$ for each unit $i \in \{i \in \mathcal{S} : \delta_i = 0\}$.
- In step (a), the data augmentation algorithm of Tanner and Wong (1987) can be implemented by iterating following two steps until convergence:
  1. **Imputation:** For given $\boldsymbol{\theta}^*$, generate $\boldsymbol{y}^*_{\rm mis} = \{y_i^* : \delta_i = 0, i \in \mathcal{S}\}$ from

$$y_i^* \sim f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}^*).$$

  2. **Posterior sampling:** For given $\boldsymbol{D}_n^* = (\boldsymbol{X}_n, \boldsymbol{y}_n^*,)$, generate $\boldsymbol{\theta}^*$ from

$$\boldsymbol{\theta}^* \sim \pi(\boldsymbol{\theta}|\boldsymbol{D}_n^*) \quad = \quad \frac{f(\boldsymbol{D}_n^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\boldsymbol{D}_n^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{1}$$

- caution! However, under informative sampling, determination of the complete-sample likelihood $f(\boldsymbol{D}_n|\theta)$ is very challenging.
  $(\because) \quad f(\boldsymbol{y}_n|\boldsymbol{X}_n, \theta) \neq \prod_{i \in \mathcal{S}} f(y_i|\boldsymbol{x}_i; \theta)$.

Remark in Step 1 of MI

- To generate $\boldsymbol{y}^*_{\mathrm{mis}}$ from $f(\boldsymbol{y}_{\mathrm{mis}}|\boldsymbol{X}_n, \boldsymbol{y}_{\mathrm{obs}})$, the following two-step procedure is commonly used.
  (a) Generate $\boldsymbol{\theta}^*$ from $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{\mathrm{obs}}, \boldsymbol{X}_n)$.
  (b) Generate y_i^* from $f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}^*)$ for each unit $i \in \{i \in \mathcal{S} : \delta_i = 0\}$.
- In step (a), the data augmentation algorithm of Tanner and Wong (1987) can be implemented by iterating following two steps until convergence:
  1. **Imputation:** For given $\boldsymbol{\theta}^*$, generate $\boldsymbol{y}^*_{\mathrm{mis}} = \{y_i^* : \delta_i = 0, i \in \mathcal{S}\}$ from
  $$y_i^* \sim f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}^*).$$
  2. **Posterior sampling:** For given $\boldsymbol{D}_n^* = (\boldsymbol{X}_n, \boldsymbol{y}_n^*,)$, generate $\boldsymbol{\theta}^*$ from
  $$\boldsymbol{\theta}^* \sim \pi(\boldsymbol{\theta}|\boldsymbol{D}_n^*) = \frac{f(\boldsymbol{D}_n^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\boldsymbol{D}_n^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \quad (1)$$
- caution! However, under informative sampling, determination of the complete-sample likelihood $f(\boldsymbol{D}_n|\boldsymbol{\theta})$ is very challenging.
  $(\because) \quad f(\boldsymbol{y}_n|\boldsymbol{X}_n, \boldsymbol{\theta}) \neq \prod_{i\in\mathcal{S}} f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta})$.

Remark in Step 1 of MI (cont.)

- To address the aforementioned issue, Kim and Yang (2017) proposed to replace $f(\boldsymbol{D}_n|\boldsymbol{\theta})$ by the sampling distribution of the pseudo likelihood estimator $g(\hat{\boldsymbol{\theta}}^*|\boldsymbol{\theta})$ as follows:
  1. **Imputation:** For given $\boldsymbol{\theta}^*$, generate $\boldsymbol{y}^*_{\text{mis}} = \{y_i^* : \delta_i = 0, i \in \mathcal{S}\}$ from
  $$y_i^* \sim f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}^*).$$
  2. new **Posterior sampling:** For given $\boldsymbol{D}_n^* = (\boldsymbol{X}_n, \boldsymbol{y}_n^*,)$, generate $\boldsymbol{\theta}^*$ from
  $$\boldsymbol{\theta}^* \sim \pi_g(\boldsymbol{\theta}|\boldsymbol{D}_n^*) \quad = \quad \frac{g(\hat{\boldsymbol{\theta}}^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int g(\hat{\boldsymbol{\theta}}^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{2}$$
  where $\hat{\boldsymbol{\theta}}^* = \hat{\theta}(\boldsymbol{D}_n^*)$ is the pseudo likelihood estimator.
- Recall
$$\left\{\hat{V}(\hat{\theta})\right\}^{-1/2}\left(\theta - \hat{\theta}\right) \xrightarrow{d} \mathcal{N}(0, \mathrm{I}).$$

Remark in Step 1 of MI (cont.)

- To address the aforementioned issue, Kim and Yang (2017) proposed to replace $f(\boldsymbol{D}_n|\boldsymbol{\theta})$ by the sampling distribution of the pseudo likelihood estimator $g(\hat{\boldsymbol{\theta}}^*|\boldsymbol{\theta})$ as follows:
  1. **Imputation:** For given $\boldsymbol{\theta}^*$, generate $\boldsymbol{y}^*_{\mathrm{mis}} = \{y_i^* : \delta_i = 0, i \in \mathcal{S}\}$ from
  $$y_i^* \sim f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}^*).$$
  2. new **Posterior sampling:** For given $\boldsymbol{D}_n^* = (\boldsymbol{X}_n, \boldsymbol{y}_n^*,)$, generate $\boldsymbol{\theta}^*$ from
  $$\boldsymbol{\theta}^* \sim \pi_g(\boldsymbol{\theta}|\boldsymbol{D}_n^*) \quad = \quad \frac{g(\hat{\boldsymbol{\theta}}^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int g(\hat{\boldsymbol{\theta}}^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{2}$$
  where $\hat{\boldsymbol{\theta}}^* = \hat{\boldsymbol{\theta}}(\boldsymbol{D}_n^*)$ is the pseudo likelihood estimator.
- Recall
$$\left\{\hat{V}(\hat{\boldsymbol{\theta}})\right\}^{-1/2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right) \xrightarrow{d} \mathcal{N}(0, \mathrm{I}).$$

Model uncertainty and model selection

- In practice, it is common to have several candidate models for given data.
- Let $\mathcal{M}$ be a set of candidate models under consideration.
- When the data generating model is unknown, it is common to perform model selection procedure to select a single best-fitting model from $\mathcal{M}$.
- Let $\hat{M}$ be the selected best model.
- Then, the MI procedure can be implemented under $\hat{M}$.
- The classical Rubin's formula does not account for the uncertainty associated with the selected model $\hat{M}$.
- Madigan and Raftery (1994) remarked that conditioning a single selected model ignores model uncertainty and so leads to underestimation of the uncertainty about the parameter of interest.

Model uncertainty and model selection

- In practice, it is common to have several candidate models for given data.
- Let $\mathcal{M}$ be a set of candidate models under consideration.
- When the data generating model is unknown, it is common to perform model selection procedure to select a single best-fitting model from $\mathcal{M}$.
- Let $\hat{M}$ be the selected best model.
- Then, the MI procedure can be implemented under $\hat{M}$.
- The classical Rubin's formula does not account for the uncertainty associated with the selected model $\hat{M}$.
- Madigan and Raftery (1994) remarked that conditioning a single selected model ignores model uncertainty and so leads to underestimation of the uncertainty about the parameter of interest.

Model uncertainty and model selection

- In practice, it is common to have several candidate models for given data.
- Let $\mathcal{M}$ be a set of candidate models under consideration.
- When the data generating model is unknown, it is common to perform model selection procedure to select a single best-fitting model from $\mathcal{M}$.
- Let $\hat{M}$ be the selected best model.
- Then, the MI procedure can be implemented under $\hat{M}$.
- The classical Rubin's formula does not account for the uncertainty associated with the selected model $\hat{M}$.
- Madigan and Raftery (1994) remarked that conditioning a single selected model ignores model uncertainty and so leads to underestimation of the uncertainty about the parameter of interest.

We need to take care of

Uncertainty about $\boldsymbol{y}_{\mathrm{mis}}$

Uncertainty about $\boldsymbol{\theta}$

Uncertainty about M

Existing method

$$\text{existing MI} \quad \begin{cases} \text{Uncertainty about } \boldsymbol{y}_{\text{mis}} \\ \text{Uncertainty about } \boldsymbol{\theta} \end{cases}$$

Uncertainty about M

Proposed method

$$\text{our new MI} \quad \begin{cases} \text{Uncertainty about } \boldsymbol{y}_{\text{mis}} \\ \text{Uncertainty about } \boldsymbol{\theta} \\ \text{Uncertainty about } M \end{cases}$$

MI under model uncertainty

- To account for the model uncertainty in the multiple imputation, we now propose to replace **Step 1** by a new step as follows:

  new **Step 1'.** Create R complete datasets, $\{\boldsymbol{D}_n^{(r)} = (\boldsymbol{y}_n^{(r)}, \boldsymbol{X}_n) : r = 1, \dots, R\}$, by repeating the following procedures independently R times:

  (a) Generate $M^{(r)}$ from $\Pr(M|\boldsymbol{y}_{\rm obs}, \boldsymbol{X}_n)$.

  (b) Generate $\boldsymbol{\theta}^{(r)}$ from $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{\rm obs}, \boldsymbol{X}_n; M^{(r)})$.

  (c) Generate $y_i^{(r)}$ from $f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}^{(r)}, M^{(r)})$ for each unit $i \in \{i \in \mathcal{S} : \delta_i = 0\}$.

Data augmentation algorithm

- Using the data augmentation algorithm, **Step 1'**(a) and (b) can be simultaneously implemented by iterating following three steps until convergence:
  1. **Imputation:** For given $(\boldsymbol{\theta}^*, M^*)$, generate $\boldsymbol{y}^*_{\mathrm{mis}} = \{y_i^* : \delta_i = 0, i \in \mathcal{S}\}$ from
  $$y_i^* \sim f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}^*, M^*).$$
  2. **Model sampling:** For given $\boldsymbol{D}_n^* = (\boldsymbol{y}_n^*, \boldsymbol{X}_n)$, generate M^* from
  $$M^* \sim \Pr(M|\boldsymbol{D}_n^*) = \frac{\pi(M)\int f(\boldsymbol{D}_n^*|\boldsymbol{\theta}; M)\pi(\boldsymbol{\theta}|M)d\boldsymbol{\theta}}{\sum_{K\in\mathcal{M}}\pi(K)\int f(\boldsymbol{D}_n^*|\boldsymbol{\theta}; K)\pi(\boldsymbol{\theta}|K)d\boldsymbol{\theta}}. \quad (3)$$
  3. **Parameter sampling:** For given $(\boldsymbol{D}_n^*, M^*)$, generate $\boldsymbol{\theta}^*_{M^*}$ from
  $$\boldsymbol{\theta}^* \sim \pi(\boldsymbol{\theta}|\boldsymbol{D}_n^*, M^*) = \frac{f(\boldsymbol{D}_n^*|\boldsymbol{\theta}; M^*)\pi(\boldsymbol{\theta}|M^*)}{\int f(\boldsymbol{D}_n^*|\boldsymbol{\theta}; M^*)\pi(\boldsymbol{\theta}|M^*)d\boldsymbol{\theta}}. \quad (4)$$
- Under some regularity conditions, $(\theta^*, M^*) \xrightarrow{d} \Pr(\theta, M|\boldsymbol{y}_{\mathrm{obs}}, \boldsymbol{X}_n) = \pi(\theta|\boldsymbol{y}_{\mathrm{obs}}, \boldsymbol{X}_n; M)\Pr(M|\boldsymbol{y}_{\mathrm{obs}}, \boldsymbol{X}_n)$.

Data augmentation algorithm

- Using the data augmentation algorithm, **Step 1'**(a) and (b) can be simultaneously implemented by iterating following three steps until convergence:
  1. **Imputation:** For given $(\boldsymbol{\theta}^*, M^*)$, generate $\boldsymbol{y}^*_{\mathrm{mis}} = \{y_i^* : \delta_i = 0, i \in \mathcal{S}\}$ from
  $$y_i^* \sim f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}^*, M^*).$$
  2. **Model sampling:** For given $\boldsymbol{D}^*_n = (\boldsymbol{y}^*_n, \boldsymbol{X}_n)$, generate M^* from
  $$M^* \sim \Pr(M|\boldsymbol{D}^*_n) = \frac{\pi(M)\int f(\boldsymbol{D}^*_n|\boldsymbol{\theta}; M)\pi(\boldsymbol{\theta}|M)d\boldsymbol{\theta}}{\sum_{K\in\mathcal{M}}\pi(K)\int f(\boldsymbol{D}^*_n|\boldsymbol{\theta}; K)\pi(\boldsymbol{\theta}|K)d\boldsymbol{\theta}}. \quad (3)$$
  3. **Parameter sampling:** For given $(\boldsymbol{D}^*_n, M^*)$, generate $\boldsymbol{\theta}^*_{M^*}$ from
  $$\boldsymbol{\theta}^* \sim \pi(\boldsymbol{\theta}|\boldsymbol{D}^*_n, M^*) = \frac{f(\boldsymbol{D}^*_n|\boldsymbol{\theta}; M^*)\pi(\boldsymbol{\theta}|M^*)}{\int f(\boldsymbol{D}^*_n|\boldsymbol{\theta}; M^*)\pi(\boldsymbol{\theta}|M^*)d\boldsymbol{\theta}}. \quad (4)$$
- Under some regularity conditions, $(\boldsymbol{\theta}^*, M^*) \xrightarrow{d} \Pr(\theta, M|\boldsymbol{y}_{\mathrm{obs}}, \boldsymbol{X}_n) = \pi(\boldsymbol{\theta}|\boldsymbol{y}_{\mathrm{obs}}, \boldsymbol{X}_n; M)\Pr(M|\boldsymbol{y}_{\mathrm{obs}}, \boldsymbol{X}_n)$.

ABC approach

- Under informative sampling, determination of the likelihood $f(\boldsymbol{D}_n|\boldsymbol{\theta}; M)$ is very challenging.
- As in Kim and Yang (2017), we use the notion of Approximate Bayesian Computation (ABC).
- The key idea of ABC is to employ summary statistics computed by $\boldsymbol{D}_n$ as a substitute of the sample data (Blum, 2010).
- Motivated by Kim and Yang (2017), we propose to use the pseudo likelihood estimator under the full model.

ABC approach (cont.)

- Let M_{full} be the full model which contains all candidate models in $\mathcal{M}$.
- Let $\hat{\boldsymbol{\theta}}_{\text{full}}$ be the pseudo MLE of $\boldsymbol{\theta}$ under model M_{full}
- Let $g(\hat{\boldsymbol{\theta}}_{\text{full}}|\boldsymbol{\theta})$ be the sampling distribution of $\hat{\boldsymbol{\theta}}_{\text{full}}$.
- The partial posterior distribution of θ under model M is defined as

$$\pi_g(\theta|\boldsymbol{D}_n;M) = \frac{g(\hat{\theta}_{\text{full}}|\theta;M)\pi(\theta|M)}{\int g(\hat{\theta}_{\text{full}}|\theta;M)\pi(\theta|M)d\theta}.$$

- Similarly the partial posterior of model M is defined as

$$\Pr_g(M|\boldsymbol{D}_n) = \frac{\pi(M)\int g(\hat{\theta}_{\text{full}}|\theta;M)\pi(\theta|M)d\theta}{\sum_{K\in\mathcal{M}}\pi(K)\int g(\hat{\theta}_{\text{full}}|\theta;K)\pi(\theta|K)d\theta}.$$

ABC approach (cont.)

- Let M_{full} be the full model which contains all candidate models in $\mathcal{M}$.
- Let $\hat{\boldsymbol{\theta}}_{\text{full}}$ be the pseudo MLE of $\boldsymbol{\theta}$ under model M_{full}
- Let $g(\hat{\boldsymbol{\theta}}_{\text{full}}|\boldsymbol{\theta})$ be the sampling distribution of $\hat{\boldsymbol{\theta}}_{\text{full}}$.
- The partial posterior distribution of $\boldsymbol{\theta}$ under model M is defined as

$$\pi_g(\boldsymbol{\theta}|\boldsymbol{D}_n; M) = \frac{g(\hat{\boldsymbol{\theta}}_{\text{full}}|\boldsymbol{\theta}; M)\pi(\boldsymbol{\theta}|M)}{\int g(\hat{\boldsymbol{\theta}}_{\text{full}}|\boldsymbol{\theta}; M)\pi(\boldsymbol{\theta}|M)d\boldsymbol{\theta}}.$$

- Similarly the partial posterior of model M is defined as

$$\Pr_g(M|\boldsymbol{D}_n) = \frac{\pi(M)\int g(\hat{\boldsymbol{\theta}}_{\text{full}}|\boldsymbol{\theta}; M)\pi(\boldsymbol{\theta}|M)d\boldsymbol{\theta}}{\sum_{K\in\mathcal{M}}\pi(K)\int g(\hat{\boldsymbol{\theta}}_{\text{full}}|\boldsymbol{\theta}; K)\pi(\boldsymbol{\theta}|K)d\boldsymbol{\theta}}.$$

Data augmentation algorithm using ABC

- We iterate following three steps until convergence:
  1. **Imputation:** For given $(\boldsymbol{\theta}^*, M^*)$, generate $\boldsymbol{y}^*_{\mathrm{mis}} = \{y_i^* : \delta_i = 0, i \in \mathcal{S}\}$ from
  $$y_i^* \sim f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}^*, M^*).$$
  2. new **Model sampling:** For given $\boldsymbol{D}_n^* = (\boldsymbol{y}_n^*, \boldsymbol{X}_n)$, generate M^* from
  $$M^* \sim \Pr_g(M|\boldsymbol{D}_n^*) = \frac{\pi(M)\int g(\hat{\boldsymbol{\theta}}^*_{\mathrm{full}}|\boldsymbol{\theta}; M)\pi(\boldsymbol{\theta}|M)d\boldsymbol{\theta}}{\sum_{K\in\mathcal{M}}\pi(K)\int g(\hat{\boldsymbol{\theta}}^*_{\mathrm{full}}|\boldsymbol{\theta}; K)\pi(\boldsymbol{\theta}|K)d\boldsymbol{\theta}}. \quad (5)$$
  3. new **Parameter sampling:** For given $(\boldsymbol{D}_n^*, M^*)$, generate $\boldsymbol{\theta}^*$ from
  $$\boldsymbol{\theta}^* \sim \pi_g(\boldsymbol{\theta}|\boldsymbol{D}_n^*, M^*) = \frac{g(\hat{\boldsymbol{\theta}}^*_{\mathrm{full}}|\boldsymbol{\theta}; M^*)\pi(\boldsymbol{\theta}|M^*)}{\int g(\hat{\boldsymbol{\theta}}^*_{\mathrm{full}}|\boldsymbol{\theta}; M^*)\pi(\boldsymbol{\theta}|M^*)d\boldsymbol{\theta}}. \quad (6)$$

Remarks on ABC approach

- Under some regularity condition, we still have

$$\left\{\hat{V}(\hat{\boldsymbol{\theta}}_{\text{full}})\right\}^{-1/2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{\text{full}}\right)\xrightarrow{d}\mathcal{N}(0,\mathrm{I}).$$

- Let $\hat{\theta}_M$ be the pseudo MLE under model M.
- We may fail to achieve the asymptotic normality of $\hat{\theta}_M$ for some $M \in \mathcal{M}$.

Remarks on ABC approach

- Under some regularity condition, we still have

$$\left\{\hat{V}(\hat{\boldsymbol{\theta}}_{\text{full}})\right\}^{-1/2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{\text{full}}\right)\xrightarrow{d}\mathcal{N}(0,\mathrm{I}).$$

- Let $\hat{\boldsymbol{\theta}}_M$ be the pseudo MLE under model M.
- We may fail to achieve the asymptotic normality of $\hat{\boldsymbol{\theta}}_M$ for some $M \in \mathcal{M}$.

Remarks on ABC approach (cont.)

- Define $m_M = \int g(\hat{\boldsymbol{\theta}}_{\text{full}}|\boldsymbol{\theta}; M)\pi(\boldsymbol{\theta}|M)d\boldsymbol{\theta}$.
- Define $\ell_M(\boldsymbol{\theta}) = \log\{g(\hat{\boldsymbol{\theta}}_{\text{full}}|\boldsymbol{\theta}; M)\pi(\boldsymbol{\theta}|M)\}$.
- Using the Laplace approximation as in Tierney and Kadane (1986), we can obtain

$$m_M = (2\pi)^{\frac{p_M}{2}} \left|-\tilde{\boldsymbol{H}}_M\right|^{-\frac{1}{2}} g(\hat{\boldsymbol{\theta}}_{\text{full}}|\tilde{\boldsymbol{\theta}}_M; M)\pi(\tilde{\boldsymbol{\theta}}_M|M)\left\{1 + O(n^{-1})\right\}, \tag{7}$$

where p_M is the number of free parameters under model M, $\tilde{\boldsymbol{\theta}}_M = \arg\max_{\boldsymbol{\theta}} \ell_M(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{H}}_M$ is the Hessian matrix of $\ell_M(\boldsymbol{\theta})$ at $\tilde{\boldsymbol{\theta}}_M$.

Remarks on ABC approach (cont.)

- Based on (7), the partial posterior probability of model M can be approximated as

$$\Pr_g(M|\boldsymbol{D}_n) \approx \{\hat{m}_M \pi(M)\} / \left\{ \sum_{K \in \mathcal{M}} \hat{m}_K \pi(K) \right\},$$

where $\hat{m}_M = (2\pi)^{\frac{p_M}{2}} \left| -\tilde{\boldsymbol{H}}_M \right|^{-\frac{1}{2}} g(\hat{\boldsymbol{\theta}}_{\text{full}} | \tilde{\boldsymbol{\theta}}_M; M) \pi(\tilde{\boldsymbol{\theta}}_M | M)$.

Theoretical result

Theorem

Under some regularity conditions, our proposed method yields

$$\begin{aligned} p \lim_{R \to \infty} \hat{\eta}_{MI} &= E_g(\eta_{\boldsymbol{x}} | \boldsymbol{X}_n, \boldsymbol{y}_{\mathrm{obs}}) \\ p \lim_{R \to \infty} \hat{V}_{MI} &= var_g(\eta_{\boldsymbol{x}} | \boldsymbol{X}_n, \boldsymbol{y}_{\mathrm{obs}}). \end{aligned}$$

Connection to BMA

Lemma

Under some regularity conditions, Kim and Yang (2017)'s method leads to

$$\begin{aligned} p\lim_{R\to\infty} \hat{\eta}_{MI}(M) &= E_g(\eta_{\boldsymbol{x}} | \boldsymbol{X}_n, \boldsymbol{y}_{\mathrm{obs}}; M); \\ p\lim_{R\to\infty} \hat{V}_{MI}(M) &= var_g(\eta_{\boldsymbol{x}} | \boldsymbol{X}_n, \boldsymbol{y}_{\mathrm{obs}}; M). \end{aligned}$$

- BMA incorporates model certainty into the multiple imputation estimator as follows:

$$\begin{aligned} \hat{\eta}_{\mathrm{BMA}} &= \sum_{M\in\mathcal{M}} \hat{\eta}_{\mathrm{MI}}(M)\mathrm{Pr}_g(M|\boldsymbol{X}_n, \boldsymbol{y}_{\mathrm{obs}}), \\ \hat{V}_{\mathrm{BMA}} &= \sum_{M\in\mathcal{M}} \left\{ \hat{V}_{\mathrm{MI}}(M) + \hat{\eta}_{\mathrm{MI}}(M)^2 \right\} \mathrm{Pr}_g(M|\boldsymbol{X}_n, \boldsymbol{y}_{\mathrm{obs}}) - \hat{\eta}^2_{\mathrm{BMA}}. \end{aligned}$$

Connection to BMA

Lemma

Under some regularity conditions, Kim and Yang (2017)'s method leads to

$$
\begin{aligned}
p \lim_{R\to\infty} \hat{\eta}_{MI}(M) &= E_g(\eta_{\boldsymbol{x}} | \boldsymbol{X}_n, \boldsymbol{y}_{\mathrm{obs}}; M); \\
p \lim_{R\to\infty} \hat{V}_{MI}(M) &= var_g(\eta_{\boldsymbol{x}} | \boldsymbol{X}_n, \boldsymbol{y}_{\mathrm{obs}}; M).
\end{aligned}
$$

- BMA incorporates model certainty into the multiple imputation estimator as follows:

$$
\begin{aligned}
\hat{\eta}_{\mathrm{BMA}} &= \sum_{M\in\mathcal{M}} \hat{\eta}_{\mathrm{MI}}(M) \mathrm{Pr}_g(M | \boldsymbol{X}_n, \boldsymbol{y}_{\mathrm{obs}}), \\
\hat{V}_{\mathrm{BMA}} &= \sum_{M\in\mathcal{M}} \left\{ \hat{V}_{\mathrm{MI}}(M) + \hat{\eta}_{\mathrm{MI}}(M)^2 \right\} \mathrm{Pr}_g(M | \boldsymbol{X}_n, \boldsymbol{y}_{\mathrm{obs}}) - \hat{\eta}^2_{\mathrm{BMA}}.
\end{aligned}
$$

Connection to BMA (cont.)

- It is straightforward to show that

$$
\begin{aligned}
p \lim_{R\to\infty} \hat{\eta}_{\text{BMA}} &= E_g \left\{ \mathrm{E}_g(\eta_{\boldsymbol{x}} | \boldsymbol{X}_n, \boldsymbol{y}_{\text{obs}}; M) | \boldsymbol{X}_n, \boldsymbol{y}_{\text{obs}} \right\} \\
&= \mathrm{E}_g(\eta_{\boldsymbol{x}} | \boldsymbol{X}_n, \boldsymbol{y}_{\text{obs}}) \\
&= p \lim_{R\to\infty} \hat{\eta}_{\text{MI}}.
\end{aligned}
$$

$$
\begin{aligned}
p \lim_{R\to\infty} \hat{V}_{\text{BMA}} &= E_g \left\{ \mathrm{var}_g(\eta_{\boldsymbol{x}} | \boldsymbol{X}_n, \boldsymbol{y}_{\text{obs}}; M) | \boldsymbol{X}_n, \boldsymbol{y}_{\text{obs}} \right\} \\
&\quad + E_g \left\{ \mathrm{E}_g(\eta_{\boldsymbol{x}} | \boldsymbol{X}_n, \boldsymbol{y}_{\text{obs}}; M)^2 | \boldsymbol{X}_n, \boldsymbol{y}_{\text{obs}} \right\} \\
&\quad - \left[E_g \left\{ \mathrm{E}_g(\eta_{\boldsymbol{x}} | \boldsymbol{X}_n, \boldsymbol{y}_{\text{obs}}; M) | \boldsymbol{X}_n, \boldsymbol{y}_{\text{obs}} \right\} \right]^2 \\
&= \mathrm{var}_g(\eta_{\boldsymbol{x}} | \boldsymbol{X}_n, \boldsymbol{y}_{\text{obs}}) \\
&= p \lim_{R\to\infty} \hat{V}_{\text{MI}}.
\end{aligned}
$$

Simulation study: continuous outcome

1. (Finite population) Generate a finite population of size $N = 20,000$ from

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon,$$

   where $x_1 \overset{iid}{\sim} \mathrm{N}(0,1)$, $x_2, \ldots, x_5 \overset{iid}{\sim} \mathrm{Ber}(0.5)$, $\epsilon \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$, and $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \sigma^2) = (-1, 1, 1, 0, 0, 0, 1)$.
2. (MAR) Generate the response indicator of y_i from $\delta_i \sim \mathrm{Bernoulli}(\psi_i)$, where

$$\mathrm{logit}(\psi_i) = 1 + 0.5 x_{1i} + 0.5 u_i,$$

   where $u_i \overset{iid}{\sim} \mathrm{N}(0,1)$. The average response rates are around 70%.
3. (Informative sampling) Draw a sample from the finite population using sampling indicator $I_i \sim \mathrm{Bernoulli}(\pi_i)$, where

$$\mathrm{logit}(1 - \pi_i) = 3.66 + 0.33 u_i - 0.1 y_i.$$

* The sample sizes range from 430 to 590.

Simulation study: binary outcome

1. (Finite population) Generate a finite population of size $N = 20,000$ from
$$y_i \sim \text{Bin}(p_i),$$
where
$$p_i = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5)},$$
$x_1 \overset{iid}{\sim} \text{N}(0,1)$, $x_2, \ldots, x_p \overset{iid}{\sim} \text{Ber}(0.5)$, and
$\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (-1, 1, 1, 0, 0, 0)$.
2. (MAR) Generate the response indicator of y_i from $\delta_i \sim \text{Bernoulli}(\psi_i)$, where
$$\text{logit}(\psi_i) = 1 + 0.5x_{1i} + 0.5u_i, \tag{8}$$
where $u_i \overset{iid}{\sim} \text{N}(0,1)$. The average response rates are around 70%.
3. (Informative sampling) Draw a sample from the finite population using sampling indicator $I_i \sim \text{Bernoulli}(\pi_i)$, where
$$\text{logit}(1 - \pi_i) = 3.66 + 0.33u_i - 0.5y_i.$$
The sample sizes range from 590 to 750.

Simulation study: setup

- The sampling weights are defined by $w_i = 1/\pi_i$.
- We consider all possible $2^5 (= 32)$ models in $\mathcal{M}$.
- We define $\pi(\boldsymbol{\theta}) \propto 1$ and $\pi(M) = 1/32$ for all $M \in \mathcal{M}$.
- We define imputation size $R = 100$.
- We consider three domains:
  - 1) $\eta_{\boldsymbol{x}_1} = E\{Y|\boldsymbol{x} = (1,0,1,0,0,0)\}$,
  - 2) $\eta_{\boldsymbol{x}_2} = E\{Y|\boldsymbol{x} = (1,0,0,1,0,0)\}$,
  - 3) $\eta_{\boldsymbol{x}_3} = E\{Y|\boldsymbol{x} = (1,0,0,0,0,0)\}$,

    where $\boldsymbol{x} = (1, x_1, \ldots, x_5)$.

Simulation study: estimation methods

- Our BMA method is compared with the following four approaches:
  1. MI under the true model,
  2. MI under the full model,
  3. MI under the selected model by BIC,
  4. MI under the selected model by AIC,

  where we utilize Kim and Yang (2017) for MI and Lumley and Scott (2015) for BIC and AIC.

Simulation results: continuous outcome

Parameter	Method	95% CP	Var($\times 10^5$)	Bias($\times 10^2$)	MSE($\times 10^5$)
$\eta_{\boldsymbol{x}_1}$	TRUE	93.50	582.73	0.06	625.64
	BMA	95.30	814.62	0.31	767.20
	FULL	93.80	1490.09	0.41	1606.05
	BIC	89.70	628.68	0.37	930.27
	AIC	87.00	857.12	0.41	1384.23
$\eta_{\boldsymbol{x}_2}$	TRUE	93.70	646.60	0.24	747.63
	BMA	95.40	877.14	0.30	934.87
	FULL	92.10	1550.72	0.40	1932.66
	BIC	90.30	694.10	0.33	1111.80
	AIC	86.80	920.56	0.22	1697.60
$\eta_{\boldsymbol{x}_3}$	TRUE	93.70	646.60	0.24	747.63
	BMA	95.50	878.33	0.48	880.36
	FULL	93.90	1554.83	0.61	1692.14
	BIC	91.10	693.84	0.56	1027.40
	AIC	87.70	920.57	0.59	1498.78

Table: Result based on MC size =1,000

Simulation results: binary outcome

Parameter	Method	95% CP	Var($\times 10^5$)	Bias($\times 10^2$)	MSE($\times 10^5$)
$\eta_{\boldsymbol{x}_1}$	TRUE	94.70	141.00	-0.10	147.88
	BMA	95.80	200.81	-0.33	195.61
	FULL	93.00	356.78	-0.16	404.31
	BIC	91.30	147.83	-0.15	211.31
	AIC	88.00	202.73	-0.24	337.31
$\eta_{\boldsymbol{x}_2}$	TRUE	94.50	105.60	0.20	109.38
	BMA	95.30	149.53	0.50	156.82
	FULL	93.10	242.86	0.15	263.75
	BIC	91.50	109.87	0.24	156.09
	AIC	88.10	145.61	0.12	226.85
$\eta_{\boldsymbol{x}_3}$	TRUE	94.50	105.60	0.20	109.38
	BMA	95.60	149.49	0.51	153.71
	FULL	93.90	244.99	0.21	251.02
	BIC	90.80	109.91	0.25	157.86
	AIC	88.80	145.57	0.13	220.18

Table: Result based on MC size =1,000

data	method	%
Continuous	BIC	85
	AIC	35
Binary	BIC	89
	AIC	35

Table: % of selecting true model

Concluding remarks

- In this study, we assume that the number of parameters is relatively smaller than sample size n.
- If the full parameter space is high-dimensional, the asymptotic normality of $\hat{\boldsymbol{\theta}}_{\text{full}}$ may fail.
- To address this issue, we can employ the stochastic search variable selection (George & McCulloch, 1993) with the notion of spike and slab prior.

REFERENCES

Binder, D. A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review* **51,** 279–292.

Blum, Michael G. B. (2010) Approximate Bayesian Computation: A Nonparametric Perspective *J. Amer. Statist. Assoc.* **105,** 1178–1187.

George, EI & McCulloch, RE (1993), 'Variable selection via gibbs sampling', *Journal of the American Statistical Association*, **88**, 881–889.

Kim, J. K. and Yang, S. (2017). A note on multiple imputation under informative sampling *Biometrika* **104,** 221-228.

Lumley, T. and Scott, A. (2015). AIC and BIC for modeling with complex survey data *Journal of Survey Statistics and Methodology* **3,** 1–18.

Madigan, D. and Raftery, A. E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window *J. Amer. Statist. Assoc.* **89,** 1535–1546.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys *New York: Wiley.*

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation *Journal of the American statistical Association* **82,** 528–540.

Tierney, L. and Kadane, J. B. (1986) Accurate Approximations for Posterior Moments and Marginal Densities *J. Amer. Statist. Assoc.* **81,** 82–86.

Q/A

THANK YOU