Three statistical issues on multiple imputation in complex survey sampling

Jae-kwang Kim

Iowa State University

January 26th, 2018

Three Issues on multiple imputation (MI)

1. Informative sampling design: We cannot simply ignore the sampling design features.
2. Congeniality and Self-efficiency (Meng, 1994): Statistical validity of MI is limited to a certain class of estimators
3. Statistical power in hypothesis testing

Issue One: Informative sampling design

Let $f(y \mid x)$ be the conditional distribution of y given x.

x is always observed but y is subject to missingness.

A sampling design is called noninformative (w.r.t f) if it satisfies

$$f(y \mid x, I = 1) = f(y \mid x) \tag{1}$$

where $I_i = 1$ if $i \in$ sample and $I_i = 0$ otherwise.

If (1) does not hold, then the sampling design is informative.

Missing At Random

Two versions of Missing At Random (MAR)

1. PMAR (Population Missing At Random)

$$Y \perp R \mid X$$

2. SMAR (Sample Missing At Random)

$$Y \perp R \mid (X, I)$$

R: response indicator function

Under noninformative sampling design, PMAR=SMAR

Imputation under informative sampling

Two approaches under informative sampling when PMAR holds.

1. Weighting approach: Use weighted score equation to estimate θ in $f(y \mid x; \theta)$. The imputed values are generated from $f(y \mid x, \hat{\theta})$.
2. Augmented model approach: Include w into model covariates to get the augmented model $f(y \mid x, w; \phi)$. The augmented model makes the sampling design noninformative in the sense that $f(y \mid x, w) = f(y \mid x, w, I = 1)$. The imputed values are generated from $f(y \mid x, w; \hat{\phi})$, where $\hat{\phi}$ is computed from unweighted score equation.

Imputation under informative sampling

- Weighting approach generates imputed values from $\hat{f}(y \mid x, R = 1)$. It is justified under PMAR.

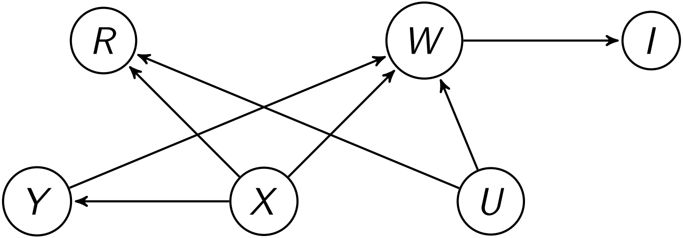
- The augmented model approach generates imputed values from $\hat{f}(y \mid x, w, I = 1, R = 1)$ and it is justified under SMAR.

- Under informative sampling, PMAR does not necessarily imply SMAR (see the next page).

- The classical multiple imputation approach is based on SMAR assumption.

Berg, Kim, and Skinner (2016; JSSAM)

Figure: A Directed Acyclic Graph (DAG) for a setup where PMAR holds but SMAR does not hold. Variable U is latent in the sense that it is never observed.



$f(y \mid x, R) = f(y \mid x)$ holds but $f(y \mid x, w, R) \neq f(y \mid x, w)$.

MI under informative sampling

- Under informative sampling, the sample distribution is different from the population distribution which follows from the marginal sample distribution,

$$f(y_i|x_i, I_i = 1) = \frac{P(I_i = 1|x_i, y_i)f(y_i|x_i)}{P(I_i = 1|x_i)}.$$

- Recall that the posterior distribution for multiple imputation is

$$p(\theta|X_n, Y_{\text{obs}}) = \frac{\int L_s(\theta|X_n, Y_n)\pi(\theta)dY_{\text{mis}}}{\int\int L_s(\theta|X_n, Y_n)\pi(\theta)dY_{\text{mis}}d\theta}.$$

- So, it is difficult to obtain the likelihood function $L_s(\theta|X_n, Y_n)$ directly from the population distribution.

New method (Kim and Yang, 2017; Biometrika)

- Under complete response, an approximate Bayesian inference can be based on

$$p_g(\theta|X_n, Y_n) = \frac{g(\hat{\theta}|\theta)\pi(\theta)}{\int g(\hat{\theta}|\theta)\pi(\theta)d\theta}, \tag{2}$$

  where g is the density for the sampling distribution of maximum pseudo likelihood estimator (PMLE) $\hat{\theta} = \hat{\theta}(X_n, Y_n)$, and $\pi(\theta)$ is a prior distribution of θ.

- The PMLE is obtained by

$$\hat{\theta} = \arg\max_{\theta} \sum_{i \in S} w_i \log f(y_i \mid x_i; \theta).$$

- The sampling distribution of PMLE is asymptotically normal.

New method of Kim and Yang (2017) (Cont'd)

- Under the existence of missing data, we generate parameters from

$$p_g(\theta|X_n, Y_{\text{obs}}) = \frac{\int g(\hat{\theta}|\theta)\pi(\theta)Y_{\text{mis}}}{\int\int g(\hat{\theta}|\theta)\pi(\theta)dY_{\text{mis}}d\theta}. \quad (3)$$

- To generate samples from (3), the following data augmentation can be used:
  - **I-Step**: Given $\theta^{(t-1)}$, draw $Y_{\text{mis}}^{*(t)} \sim f\left(Y_{\text{mis}}|X_n, Y_{\text{obs}}; \theta^{(t-1)}\right)$.
  - **P-step**: Given $Y_{\text{mis}}^{*(t)}$, draw

$$\theta^{(t)} \sim p_g\left(\theta|X_n, Y_n^{*(t)}\right) = \frac{g\left(\hat{\theta}^{*(t)}|\theta\right)\pi(\theta)}{\int g\left(\hat{\theta}^{*(t)}|\theta\right)\pi(\theta)d\theta},$$

where $\hat{\theta}^{*(t)} = \hat{\theta}\left(X_n, Y_n^{*(t)}\right)$ is PMLE calculated using the imputed values $Y_{\text{mis}}^{*(t)}$, and $Y_n^{(t)} = (Y_{\text{obs}}, Y_{\text{mis}}^{*(t)})$.

Simulation Study

- Superpopulation models (=models for the finite populations)
  1. Continuous outcome following a linear regression superpopulation model,
  $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$
  where $x_i \sim \text{Normal}(2, 1)$, $\epsilon \sim \text{Normal}(0, \sigma^2)$, and $(\beta_0, \beta_1, \sigma^2) = (-1.5, 0.5, 1.04)$.
  2. Binary outcome following a logistic regression superpopulation model,
  $$y_i \sim \text{Bernoulli}(p_i),$$
  where $p_i = \exp(\beta_0 + \beta_1 x_i)/1 + \exp(\beta_0 + \beta_1 x_i)$, $x_i \sim \text{Normal}(2, 1)$, and $(\beta_0, \beta_1) = (-1.5, 0.5)$.
- Finite populations of size $N = 50,000$ are independently generated from each superpopulation model.

Simulation Study

For each population,

- Missingness mechanism:
  $\delta_i \sim \text{Bernoulli}(\phi_i)$ with $\text{logit}(\phi_i) = -1 + 0.5x_i + 0.5u_i$
  where $u_i \sim \text{Normal}(2, 1)$, and u_i is independent of x_i and ϵ_i.

- Sampling mechanisim:
  Poisson sampling with $I_i \sim \text{Bernoulli}(\pi_i)$, where
  1. non-informative sampling:
     - both comes : $\text{logit}(1 - \pi_i) = 3 + 0.5x_i$,
  2. informative sampling:
     - continuous outcome: $\text{logit}(1 - \pi_i) = 3 + \frac{1}{3}u_i - 0.1y_i$
     - binary outcome : $\text{logit}(1 - \pi_i) = 3 + \frac{1}{3}u_i - 0.5y_i$.

Simulation Study

- Estimators for $\eta = N^{-1} \sum_{i=1}^{N} y_i$
  - Hajek estimator, assuming all observations are available.
  - Traditional MI estimator using augmented model $f(y|x, w)$ with imputation size 50
  - Kim & Yang's (KY) method for MI with imputation size 50
  - Posterior approach with the number of each MCMC simulation = 500.
- Assume flat prior distribution for both multiple imputation.
- $w_i = 1/\pi_i$.

Simulation Study : Results

Table: Simulation result under non-informative sampling design : bias, variance of the point estimator, and coverage of 95% confidence intervals based on 1,000 Monte Carlo samples.

Non-informative sampling design

	Method	Bias	Var (10^{-5})	Coverage (%)
Continuous outcome	Hajeck	0.00	167	95
	Traditional MI	0.00	213	95
	KY MI	0.00	212	95
Binary outcome	Hajeck	0.00	33	94
	Traditional MI	0.00	43	94
	KY MI	0.00	43	94

Simulation Study : Results

Table: Simulation result under informative sampling design : bias, variance of the point estimator, and coverage of 95% confidence intervals based on 1,000 Monte Carlo samples.

Informative sampling design

	Method	Bias	Var (10^{-5})	Coverage (%)
Continuous outcome	Hajeck	0.00	114	95
	Traditional MI	0.04	138	84
	KY MI	0.00	152	95
Binary outcome	Hajeck	0.00	16	95
	Traditional MI	0.03	20	42
	KY MI	0.00	22	94

Issue Two: Class of estimators that MI works

Some history

- Rubin (1978, 1987) proposed MI as an imputation tool for general purpose estimation.
- Fay (1991, 1992) found that MI variance estimator is positively biased for domain estimation if the imputed values are obtained from a reduced model. It is essentially due to borrowing strength phenomenon.
- Meng (1994) gave a theory for the validity of MI. He showed that MI works only for a certain class of estimators and the class is called self-efficient estimator. Also, he argue that MI is still OK for other classes because the MI inference will be conservative.
- Kim, Brick, Fuller, and Kalton (2006) and Yang and Kim (2016) provide further insights on the self-efficient estimation.

Numerical illustration

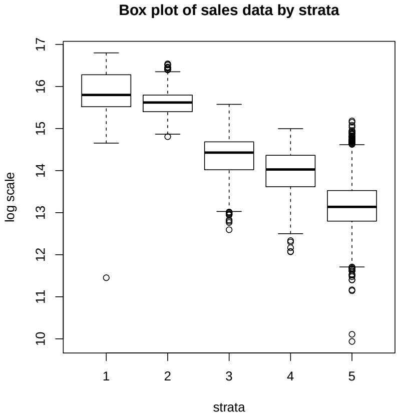
A pseudo finite population constructed from a single month data in Monthly Retail Trade Survey (MRTS) at US Bureau of Census

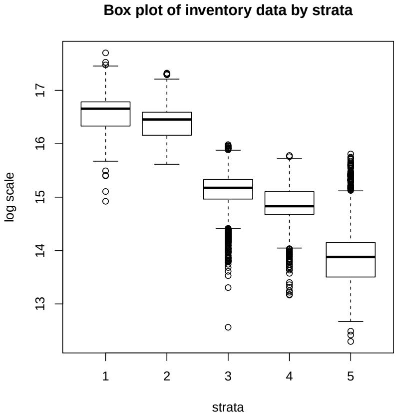
$N = 7,260$ retail business units in five strata

Three variables in the data

- h: stratum
- x_{hi}: inventory values
- y_{hi}: sales

Box plot of log sales and log inventory values by strata





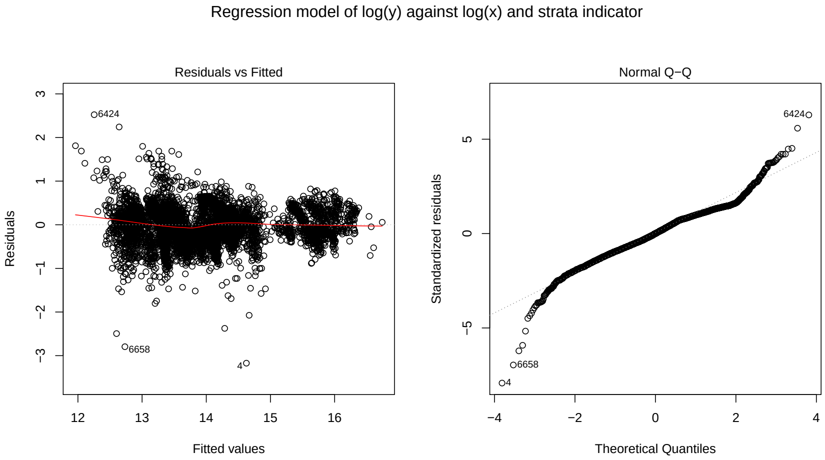
Imputation model

$$log(y_{hi}) = \beta_{0h} + \beta_1 \log(x_{hi}) + e_{hi}$$

where

$$e_{hi} \sim N(0, \sigma^2)$$

Residual plot and residual QQ plot



Stratified random sampling

Table: The sample allocation in stratified simple random sampling.

Strata	1	2	3	4	5
Strata size N_h	352	566	1963	2181	2198
Sample size n_h	28	32	46	46	48
Sampling weight	12.57	17.69	42.67	47.41	45.79

Response mechanism: PMAR

Variable x_{hi} is always observed and only y_{hi} is subject to missingness.
PMAR

$$R_{hi} \sim Bernoulli(\pi_{hi}), \ \pi_{hi} = 1/[1 + \exp\{4 - 0.3 \log(x_{hi})\}].$$

The overall response rate is about 0.6.

Simulation Study (Yang and Kim, 2017; Statistical Science)

Table 1 Monte Carlo bias and variance of the point estimators.

Parameter	Estimator	Bias	Variance	Std Var
	Complete sample	0.00	0.42	100
$\theta = E(Y)$	MI	0.00	0.59	134
	FI	0.00	0.58	133

Table 2 Monte Carlo relative bias of the variance estimator.

Parameter	Imputation	Relative bias (%)
$V(\hat{\theta})$	MI	18.4
	FI	2.7

Discussion

- Rubin's formula is based on the following decomposition:

$$V(\hat{\eta}_{MI}) = V(\hat{\eta}_n) + V(\hat{\eta}_{MI} - \hat{\eta}_n)$$

  where $\hat{\eta}_n$ is the complete-sample estimator of θ. Basically, U_m term estimates $V(\hat{\eta}_n)$ and $(1 + m^{-1})B_m$ term estimates $V(\hat{\eta}_{MI} - \hat{\eta}_n)$.

- For general case, we have

$$V(\hat{\eta}_{MI}) = V(\hat{\eta}_n) + V(\hat{\eta}_{MI} - \hat{\eta}_n) + 2Cov(\hat{\eta}_{MI} - \hat{\eta}_n, \hat{\eta}_n)$$

  and Rubin's variance estimator ignores the covariance term. Thus, a sufficient condition for the validity of unbiased variance estimator is

$$Cov(\hat{\eta}_{MI} - \hat{\eta}_n, \hat{\eta}_n) = 0.$$

- Meng (1994) called the condition congeniality of $\hat{\eta}_n$.
- Congeniality holds when $\hat{\eta}_n$ is the MLE of η (self-efficient estimator).

Discussion (Cont'd)

- For example, there are two estimators of $\eta = E(Y)$ when $\log(Y)$ follows from $N(\beta_0 + \beta_1 x, \sigma^2)$.
  1. Maximum likelihood method:
  $$\hat{\eta}_{MLE} = n^{-1} \sum_{i=1}^{n} \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i + 0.5\hat{\sigma}^2\}$$
  2. Method of moments:
  $$\hat{\eta}_{MME} = n^{-1} \sum_{i=1}^{n} y_i$$
- Asymptotically, $V(\hat{\eta}_{MME}) \geq V(\hat{\eta}_{MLE})$.

Discussion (Cont'd)

- When MI is applied to $\hat{\eta}_{MME}$, we have

$$\hat{\eta}_{MI} \cong n^{-1} \sum_{i=1}^{n} \left\{ R_i y_i + (1 - R_i) E(y_i \mid x_i; \hat{\theta}_{MLE}) \right\}$$

  where $\theta = (\beta_0, \beta_1, \sigma^2)$. Thus, MI estimator is a convex combination of MME and MLE.
- The MME of η does not satisfy the self-efficiency and Rubin's variance estimator applied to MME is upwardly biased.
- Rubin's variance estimator is essentially unbiased for MLE of η but MLE is rarely used in practice.

Reference: S. Yang and J.K. Kim (2016). "A Note on Multiple Imputation for Method of Moments Estimation", *Biometrika*, **103**, 244 – 251.

Issue Three: Statistical Power

- Some supporters of MI says that MI is still OK because it will provide conservative inference in most cases.
- How about statistical power in hypothesis testing?

Simulation Study (Kim and Yang 2014, SMJ)

- Bivariate data (x_i, y_i) of size $n = 100$ with

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 \left(x_i^2 - 1\right) + e_i \tag{4}$$

  where $(\beta_0, \beta_1, \beta_2) = (0, 0.9, 0.06)$, $x_i \sim N(0,1)$, $e_i \sim N(0, 0.16)$, and x_i and e_i are independent. The variable x_i is always observed but the probability that y_i responds is 0.5.
- The imputation model is

$$Y_i = \beta_0 + \beta_1 x_i + e_i.$$

  That is, imputer's model uses extra information of $\beta_2 = 0$.
- From the imputed data, we fit model (4) and computed power of a test $H_0 : \beta_2 = 0$ with 0.05 significant level.
- In addition, we also considered the Complete-Case (CC) method that simply uses the complete cases only for the regression analysis.

Simulation Study

Table 5 Simulation results for the Monte Carlo experiment based on 10,000 Monte Carlo samples.

Method	$E(\hat{\theta})$	$V(\hat{\theta})$	R.B. $(\hat{V})$	Power
MI	0.028	0.00056	1.81	0.044
CC	0.060	0.00234	-0.01	0.285

Table 5 shows that MI provides efficient point estimator than CC method but variance estimation is very conservative (more than 100% overestimation). Because of the serious positive bias of MI variance estimator, the statistical power of the test based on MI is actually lower than the CC method.

Conclusion

- We should understand the risks when MI is used in the production.
- MI has three main risks. Such risks should be clearly stated if we still want to use MI officially.
- Other options (such as fractional imputation) can also be considered.