Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 16: missing data 3 – item nonresponse

UNIVERSITY OF MICHIGAN

Item nonresponse

- Item nonresponse generally has complex “swiss-cheese” pattern
- Weighting methods are possible when the data have a monotone pattern, but are very difficult to develop for a general pattern
- Two variants of Bayes for item nonresponse:
  - Compute posterior predictive distribution of population quantities, given the observed data
  - Multiple imputation of draws from predictive distribution of missing values
- By conditioning fully on all observed data, these methods weaken MAR assumption

Bayesian MCMC Computations

A convenient algorithmic approach for complex problems is to iterate between draws of the missing values and draws of the parameters:

$$(Y_{\text{mis}}^{(d,t+1)} \mid Y_{\text{obs}}, \theta^{(dt)}) \sim p(Y_{\text{mis}} \mid Y_{\text{obs}}, \theta^{(dt)})$$

$$(\theta^{(d,t+1)} \mid Y_{\text{obs}}, Y_{\text{mis}}^{(d,t+1)}) \sim p(\theta \mid Y_{\text{obs}}, Y_{\text{mis}}^{(d,t+1)})$$

As t tends to infinity, this sequence converges to a draw from the joint posterior distribution of (Y_{mis}, θ), as required.

- One of the first applications of the Gibbs' sampler (Tanner and Wong 1984)

  Unlike the related EM algorithm, yields full posterior distribution, not just an ML estimate.

- Draws $Y_{\text{mis}}^{(d,t)}$ of missing data can be used to create multiply-imputed data sets

Multiple imputation

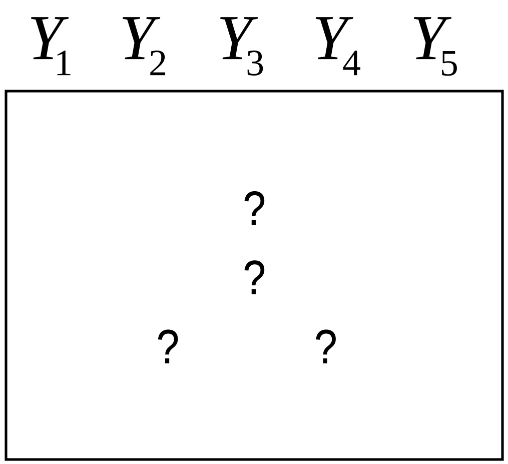
- Imputes *draws*, not means, from the predictive distribution of the missing values
- Creates $D > 1$ filled-in data sets with different values imputed
- Bayesian MI combining rules yield valid inferences under well-specified models – propagate imputation uncertainty, and averaging of estimates over MI data sets avoids the efficiency loss from imputing draws
- MI can also be used for non-MAR models, particularly for *sensitivity analyses*

Idea of Multiple Imputation

- Data matrix with missing values

Variables

Cases



$\hat{\mu}_1 =$ mean based on all cases

$\hat{\beta}_{51\cdot1234} = ?$

Impute to recover information in incomplete cases

Single Imputation

- Impute missing values with predictions

	Estimate (se^2)	
Dataset (l)	μ_1	$\beta_{51\cdot 1234}$
1	$12.6\ (3.6^2)$	$4.32\ (1.95^2)$

Y_1 Y_2 Y_3 Y_4 Y_5

2.1

4.5

24 1

Imputing best estimates biases slope - need to impute <u>draws</u>

SE of slope is too low – imputation error is not accounted for

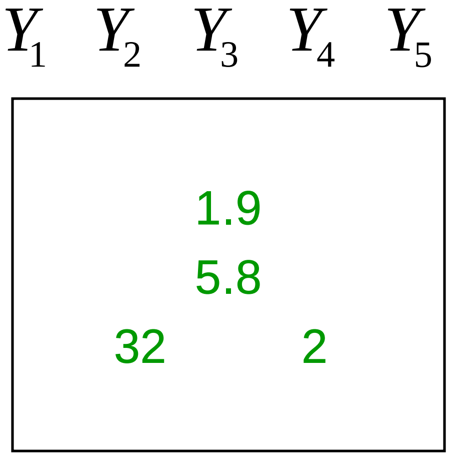
MI: repeat with other draws

Second imputed dataset

Y_1	Y_2	Y_3	Y_4	Y_5
		2.7		
		5.1		
	31		1	

Dataset (l)	Estimate (se^2) μ_1	$\beta_{51\cdot1234}$
1	12.6 (3.6^2)	4.32 (1.95^2)
2	12.6 (3.6^2)	4.15 (2.64^2)

Third imputed dataset



	Estimate (se^2)	
Dataset (l)	μ_1	$\beta_{51\cdot1234}$
1	12.6 (3.6^2)	4.32 (1.95^2)
2	12.6 (3.6^2)	4.15 (2.64^2)
3	12.6 (3.6^2)	4.86 (2.09^2)

Fourth imputed dataset

	Estimate (se^2)	
Dataset (l)	μ_1	$\beta_{51\cdot1234}$
1	12.6 (3.6^2)	4.32 (1.95^2)
2	12.6 (3.6^2)	4.15 (2.64^2)
3	12.6 (3.6^2)	4.86 (2.09^2)
4	12.6 (3.6^2)	3.98 (2.14^2)

Y_1	Y_2	Y_3	Y_4	Y_5
		2.5		
		3.9		
	18		1	

Fifth imputed dataset

Y_1 Y_2 Y_3 Y_4 Y_5

2.3
4.2
25 2

Dataset (l)	Estimate (se^2) μ_1	$\beta_{51\cdot1234}$
1	12.6 (3.6^2)	4.32 (1.95^2)
2	12.6 (3.6^2)	4.15 (2.64^2)
3	12.6 (3.6^2)	4.86 (2.09^2)
4	12.6 (3.6^2)	3.98 (2.14^2)
5	12.6 (3.6^2)	4.50 (2.47^2)
Mean	12.6 (3.6^2)	4.36 (2.27^2)
Var	0	0.339

MI combining rules

Simulation approximations of posterior mean, variance yield the ML combining rules:

$$E(\theta \mid Y_{\text{obs}}) \approx \frac{1}{D}\sum_{d=1}^{D} E(\theta \mid Y_{\text{obs}}, Y_{\text{mis}}^{(d)}) = \frac{1}{D}\sum_{d=1}^{D}\hat{\theta}_d$$

where $\hat{\theta}_d =$ is posterior mean from dth dataset

$$Var(\theta \mid Y_{\text{obs}}) \approx \frac{1}{D}\sum_{d=1}^{D} W_d + (1+1/D)\frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\theta}_d - \overline{\theta}_D\right)^2$$

where $W_d = \text{Var}(\theta \mid Y_{\text{obs}}, Y_{\text{mis}}^{(d)})$ is posterior variance from dth dataset

MI Inferences (M=5)

	$\bar{\theta}$	$\bar{W}$	B	$\sqrt{V} = \sqrt{\bar{W} + (1 + \frac{1}{D})B}$	$R = \frac{(1+1/D)B}{V}$
μ_1	12.6	3.6^2	0	3.6	0
$\beta_{51\cdot1234}$	4.36	2.27^2	0.339	2.36	0.073

$\bar{\theta} =$ MI estimate

$\sqrt{V} =$ MI standard error

$R =$ estimated fraction of missing information

Advantages of MI

- Imputation model can differ from analysis model
  - By including variables not included in final analysis
  - Promotes consistency of treatment of missing data across multiple analyses
  - MI combining rules can also be applied when the complete-data inference is not Bayesian (e.g. design-based survey inference).
  - Assumptions in imputation model are then confined to the imputations – with little missing data, simple methods suffice
- Public use data set users can be provided MI's, spared task of building imputation model
  - MI analysis of imputed data is easy, using complete-data methods (SAS PROC MIANALYZE)

MI for parametric models

- Principled, MCMC methods for creating draws have predictable properties
- Parametric assumptions can be improved by usual data-analytic strategies, e.g. transformations
- Analysis of MI data sets can be based on less parametric methods if desired
- For monotone pattern, flexibility is achieved by *factoring* the joint distribution
- However, for general patterns, the requirement for a coherent joint distribution limits flexibility
  - E.g. multivariate normality assumes regressions are linear and additive

Sequential regression MI (SRMI)

- Sequential regression MI (IVEware, MICE) regresses each variable with missing values in succession on all the other variables, with missing values of regressors filled in from earlier steps
- Iterates until imputations appear “stable”
- For parametric model, sequential imputation is essentially a form of Gibbs’ sampler
- Flexibility allowed in regressions – e.g. logit links for binary variables, nonlinear terms
- Conditionals may be incoherent – do not correspond to well-specified joint d/n – but gain in flexibility outweighs this theoretical drawback

Example. Logistic regression simulation study

- True model:

$$\mathbf{X \sim N(0,1)}$$

$$\mathbf{Logit[Pr(E=1|X)]=0.5+X}$$

$$\mathbf{logit[Pr(D=1|E,X)]=0.25+0.5X+1.1E}$$

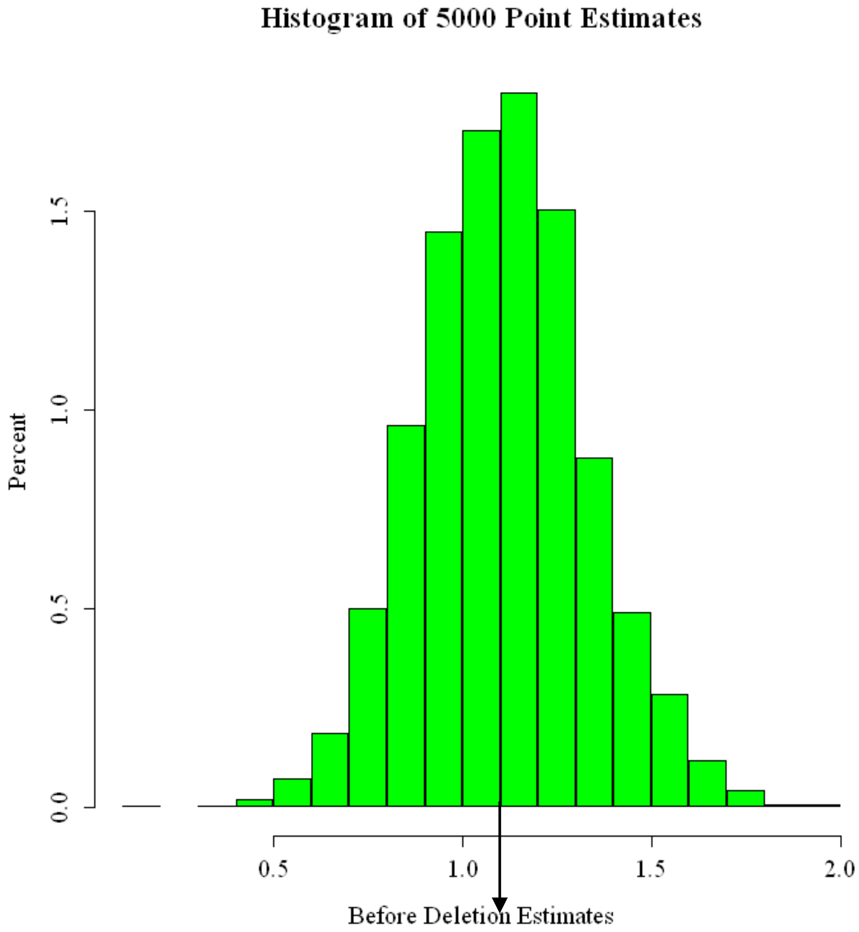
- Sample size: 500
- Number of Replicates: 5000
- Before Deletion Data Sets

Missing-Data Mechanism

- D and E : completely observed
- X : sometimes missing
- Missing Data Probabilities:

D=0,E=0:	$p_{00}=0.19$
D=0,E=1:	$p_{01}=0.09$
D=1,E=0:	$p_{10}=0.015$
D=1,E=1:	$p_{11}=0.055$

Before Deletion Estimates

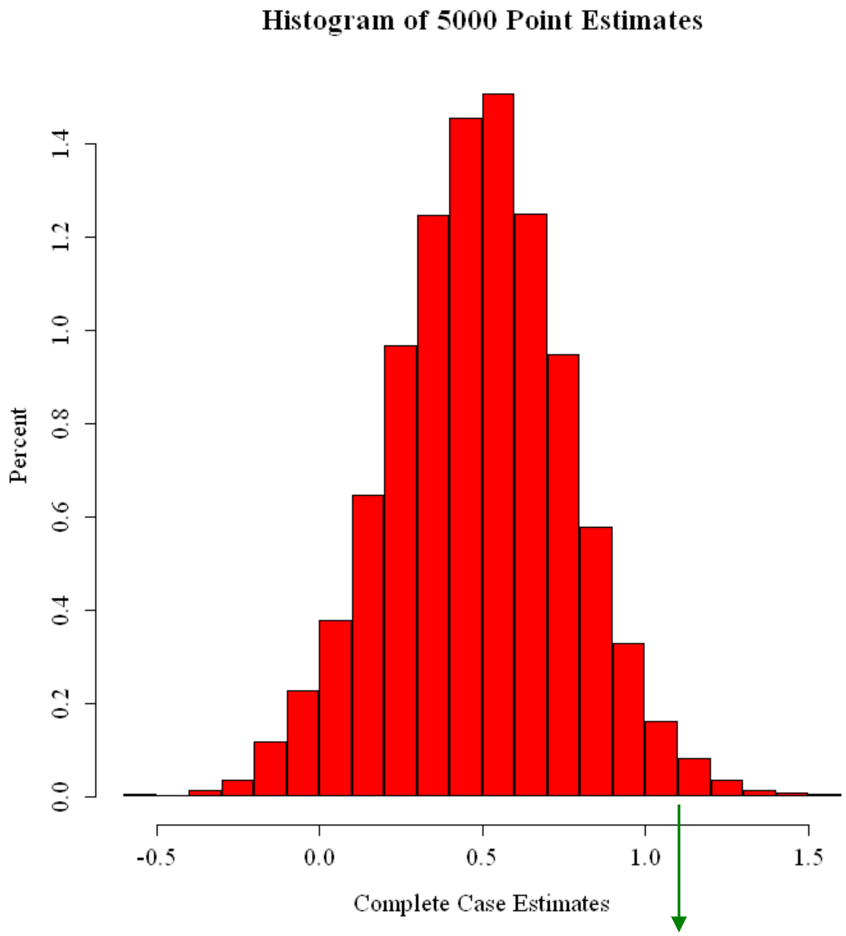


- Histogram of 5000 estimates before deleting values of X
- logistic model

logit Pr(D=1|E,X)

$$= \beta_0 + \beta_1 E + \beta_2 X$$

Complete-Case Estimates



Histogram of complete- case analysis estimates

Delete subjects with missing X values

True value = 1.1, serious negative bias

MI for logistic regression example

- The model for the data implies that for missing values of *X*:

$$(X_i \mid D_i = d, E_i = e, \mu_{ed}, \sigma^2) \sim N(\mu_{ed}, \sigma^2)$$

- Improper MI: substitute estimates of $\{\mu_{ed}\}, \sigma^2$
- Proper MI: Imputations are draws from the posterior predictive distribution
- Draw σ^2, then μ_{ed} and then missing X_i

Predictive Distributions

$$\sigma^{2(\ell)} \sim WSS / \chi^2_{r-4},$$

WSS = residual sum of squares,

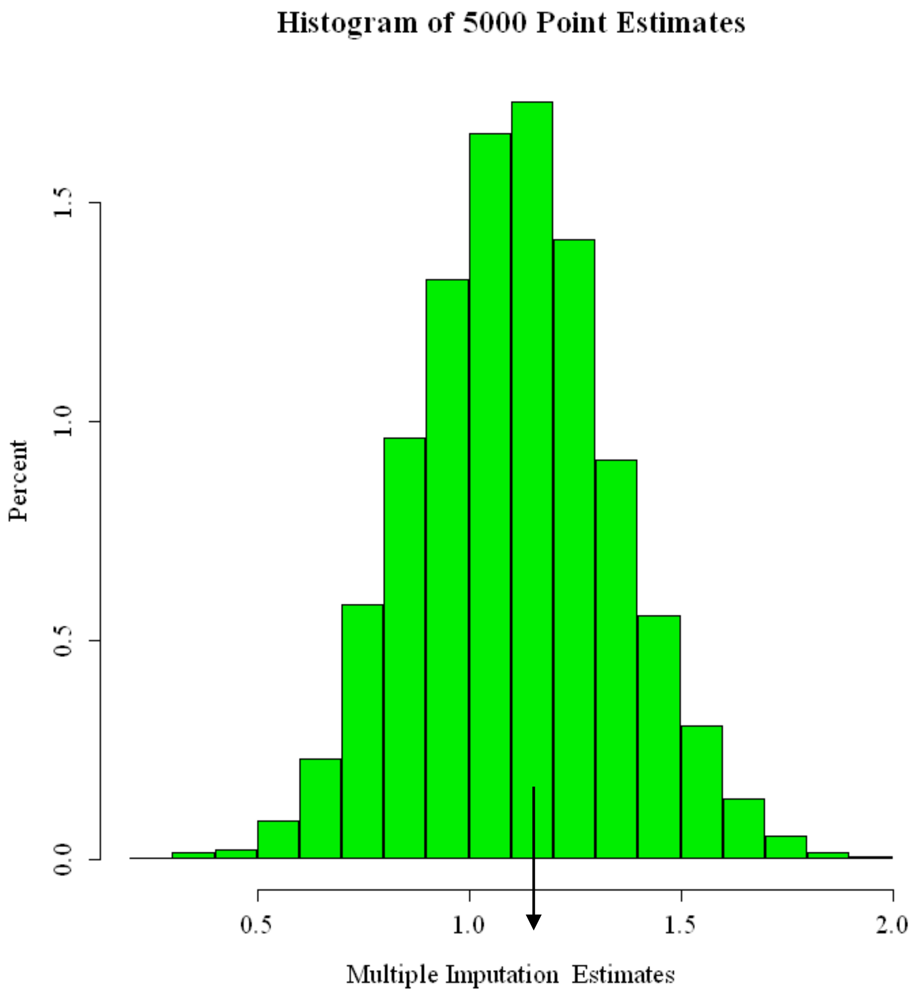
$r =$ number of complete cases

$$\mu_{ed}^{(\ell)} \sim N(\bar{x}_{ed}, \sigma^2 / r_{ed})$$

$\bar{x}_{ed}, r_{ed} =$ mean, complete cases in cell (e,d)

$$X_{edi}^{(\ell)} \sim N(\mu_{ed}^{(\ell)}, \sigma^{2(\ell)})$$

Histogram of Multiple Imputation Estimates



- 5 Imputations per missing value
- 5 completed Datasets
- Analyze each separately
- Combine using the formulae given earlier

Coverage and MSE of Various Methods

METHOD	COVERAGE (95% Nominal)	MSE
Complete-case	37.86	0.4456
Hot-Deck Single Imputation	90.28	0.0566
Multiple Imputation	94.56	0.0547
Before Deletion	***94.68***	***0.0494***

Bayesian Theory of MI (Rubin, 1987)

For simplicity assume MAR -- MNAR also allowed

Model: $f(Y \mid \theta) \Rightarrow$ Likelihood $L(\theta \mid Y) \propto f(Y \mid \theta)$

Prior distribution: $\pi(\theta)$; md mechanism: MAR

$Y = \left(Y_{\text{obs}}, Y_{\text{mis}}\right), Y_{\text{obs}}$ = observed data, Y_{mis} = missing data

Complete-data posterior distribution,

if there were no missing values:

$$p(\theta \mid Y_{\text{obs}}, Y_{\text{mis}}) \propto \pi(\theta)\, L(\theta \mid Y_{\text{obs}}, Y_{\text{mis}})$$

Posterior distribution given observed data:

$$p(\theta \mid Y_{\text{obs}}) \propto \pi(\theta)\, L(\theta \mid Y_{\text{obs}})$$

Theory relates these two distributions ...

Relating the posteriors

- The posterior is related to the complete-data posterior by:

$$p(\theta \mid Y_{\text{obs}}) = \int p(\theta \mid Y_{\text{obs}}, Y_{\text{mis}})\, p(Y_{\text{mis}} \mid Y_{\text{obs}})\, dY_{\text{mis}}$$

$$\approx \frac{1}{D} \sum_{d=1}^{D} p(\theta \mid Y_{\text{obs}}, Y_{\text{mis}}^{(d)}), \text{ where } Y_{mis}^{(d)} \sim p(Y_{mis} \mid Y_{\text{obs}})$$

$Y_{\text{mis}}^{(d)}$ is a draw from the predictive distribution of the missing values

The accuracy of the approximation increases with D and the fraction of observed data

MI approximation to posterior mean

- Similar approximations yield MI combining rules:

$$\mathrm{E}(\theta \mid Y_{\text{obs}}) = \int \mathrm{E}(\theta \mid Y_{\text{obs}}, Y_{\text{mis}}) p(Y_{\text{mis}} \mid Y_{\text{obs}}) dY_{\text{mis}}$$

$$\approx \frac{1}{D}\sum_{d=1}^{D} \mathrm{E}(\theta \mid Y_{\text{obs}}, Y_{\text{mis}}^{(d)}) = \frac{1}{D}\sum_{d=1}^{D} \hat{\theta}_d,$$

where $\hat{\theta}_d =$ is posterior mean from *d*th imputed dataset

MI approximation to posterior variance

$$\mathrm{Var}(\theta \mid Y_{\mathrm{obs}}) = \mathrm{E}(\theta^2 \mid Y_{\mathrm{obs}}) - \left(\mathrm{E}(\theta \mid Y_{\mathrm{obs}})\right)^2$$

Apply above approx to $\mathrm{E}(\theta \mid Y_{\mathrm{obs}})$ and $\mathrm{E}(\theta^2 \mid Y_{\mathrm{obs}})$

Algebra then yields:

$$\mathrm{Var}(\theta \mid Y_{\mathrm{obs}}) \approx \bar{V} + B$$

$\bar{V} = \frac{1}{D}\sum_{d=1}^{D} V_d =$ within-imputation variance,

$V_d = \mathrm{Var}(\theta \mid Y_{\mathrm{obs}}, Y_{\mathrm{mis}}^{(d)})$ is posterior variance from dth dataset

$B = \frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\theta}_d - \bar{\theta}_D\right)^2$ = between-imputation variance

Refinements for small D

(A): $Var(\theta \mid Y_{\text{obs}}) \approx \bar{V} + \left(1+1/D\right)B$

(B) Replace normal reference distribution by t distribution with df

$$\nu = (D-1)\left(1+\frac{D}{D+1}\frac{\bar{V}}{B}\right)^2$$

(C) For normal sample with variance based on ν_{com} df, replace ν by

$$\nu^* = \left(\nu^{-1}+\hat{\nu}_{\text{obs}}^{-1}\right)^{-1}, \hat{\nu}_{\text{obs}} = (1-\hat{\gamma}_D)\left(\frac{\nu_{\text{com}}+1}{\nu_{\text{com}}+3}\right)\nu_{\text{com}}$$

$$\hat{\gamma}_D = \frac{\left(1+D^{-1}\right)B}{\bar{V}+\left(1+D^{-1}\right)B} = \text{estimated fraction of missing information}$$

Why MI for surveys?

- Software is widely available (IVEware, MICE, etc.)
- MI based on Bayes for a joint model for the data has optimal asympotic properties under that model.
- Propagates imputation uncertainty in a way that is practical for public use files
- Flexible, using models that fully condition on observed data – makes MAR assumption "as weak as possible"
- Applies to general patterns – weighting methods do not generalize in a compelling way beyond monotone patterns

Why MI for surveys?

- Allows inclusion of auxiliary variables in the imputation model that are not in the final analysis
- “Design-based” methods can be applied to multiply-imputed data, with MI combining rules: model assumptions only used to create the imputations (where assumptions are inevitable).

Arguments against MI for surveys

- It’s model-based, and I don’t want to make assumptions – but there is no assumption-free imputation method!
- Lack of congeniality between imputer model and analyst model
  – advice is to be inclusive of potential predictors, leading to at worst conservative inferences – parametric models allow main effects to be prioritized over high order interactions
  – Congeniality problem also applies to other methods that falsely claim to be assumption free
  – Perfection is the enemy of the good – in simulation studies, MI tends to work well, because it is propagating imputation uncertainty

Arguments against MI for surveys

- Misspecified parametric models can lead to problems with the imputes – for example, imputing log-transformed data and then exponentiating can lead to wild imputations
- So, important to plot the imputations to check that they are plausible
- With large samples, chained equations with predictive mean matching hot deck has some attractions, since only actual values are imputed
- But hot deck methods are less effective in small samples where good matches are lacking (Andridge & Little, 2010)

Missing Not at Random Models

- Difficult problem, since information to fit non-MAR is limited and highly dependent on assumptions
- Sensitivity analysis is preferred approach – this form of analysis is not appealing to consumers of statistics, who want clear answers
- Selection vs Pattern-Mixture models
  - Prefer pattern-mixture factorization since it is simpler to explain and implement
  - Offsets, Proxy Pattern-mixture analysis
- Missing covariates in regression
  - Subsample Ignorable Likelihood

A simple pattern-mixture model

Giusti & Little (2011) extends this idea to a PM model for income nonresponse in a rotating panel survey:

* Two mechanisms (rotation MCAR, income nonresponse NMAR)
* Offset includes as a factor the residual sd, so smaller when good predictors are available
* Complex problem, but PM model is easy to interpret and fit

Readily implemented extension of chained equation MI to MNAR models

An Alternative: Proxy Pattern-Mixture Analysis

$$[y_i \mid x_i, r_{2i} = k] \sim G(\beta^{(k)} x_i, \tau^{2(k)})$$

$$\Pr(r_i = 1 \mid x_i, y_i) = g\left(y_i^*(\lambda)\right),\ y_i^*(\lambda) = \hat{y}(x_i) + \lambda y_i$$

$\hat{y}(x_i) =$ best predictor of y_i

MAR: $\lambda = 0$, MNAR: $\lambda \neq 0$

(Andridge and Little 2011)

(*) implies that [y_i indep $r_i \mid y_i^*(\lambda)$], which identifies the model

Interesting feature: $g()$ is arbitrary, unspecified

NMAR model that avoids specifying missing data mechanism

PPMA: Sensitivity analysis for different choices of λ

If x_i is a noisy measure of y_i, it may be plausible to assume $\lambda = \infty$

(West and Little, 2013)

Summary

- Bayesian approach to missing data meshes seamlessly with Bayesian approach to survey inference
  - Predict missing values as well as non-sampled values
- MAR key condition: MAR methods much easier if they can be justified
- Multiple imputation provides flexibility, allows design-based complete-data methods to be applied