Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 14: missing data 1 -- overview

UNIVERSITY OF MICHIGAN

Missing data methods -- history

1. Before the EM algorithm (pre-1970's)
   - Ad-hoc adjustments (simple imputation)
   - ML for simple problems (Anderson 1957)
   - ML for complex problems too hard
2. ML era (1970's – mid 1980's)
   - Rubin formulates model for missing data mechanism, defines MAR (1976)
   - EM and extensions facilitate ML for complex problems
   - ML for more flexible models – beyond multivariate normal (see e.g. Little and Rubin 1987)

Missing data methods -- history

3. Bayes and Multiple Imputation (mid 1980's – present)
   - Tanner and Wong describes data augmentation for the multivariate normal problem (1984)
   - Rubin proposes MI, justified via Bayes (1977, 1987)
   - MCMC facilitates Bayes as an alternative to ML, with better small sample properties (see e.g. Little and Rubin 2002)
4. Robustness concerns (1990's – present)
   - Robins et al propose doubly robust methods for missing data
   - Robust Bayesian models, more attention to model checks

Finite population inference

- Bayesian modeling takes a predictive perspective on statistical inference – predict the non-sampled values
- Inference about parameters is intermediate step in predictive superpopulation model inference about finite population parameters
- Bayesian approach extends naturally to handle missing data
  - Predict nonsampled and missing values
- Missingness not under control of sampler, so missing not at random mechanisms complicate inferences

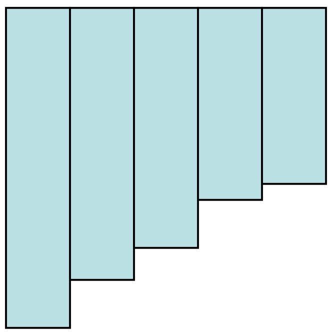
Likelihood methods with missing data

- Likelihood methods do not require rectangular data, hence apply directly to missing-data problems
- Statistical model + incomplete data □ Likelihood
- Approaches based on the likelihood:
  - ML estimates, large sample standard errors
  - Bayes: add priors, compute posterior distribution
  - Multiple imputation: multiple draws of missing values, apply MI combining rules
  - Flexible and general
  - Methods reflect added uncertainty from missing data

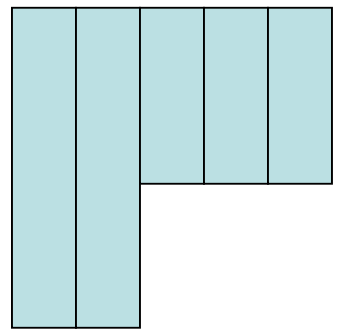
Patterns of Missing Data 1

- special patterns

monotone

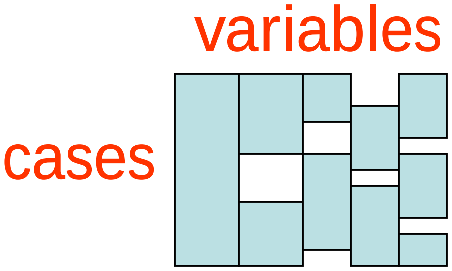


unit nonresponse



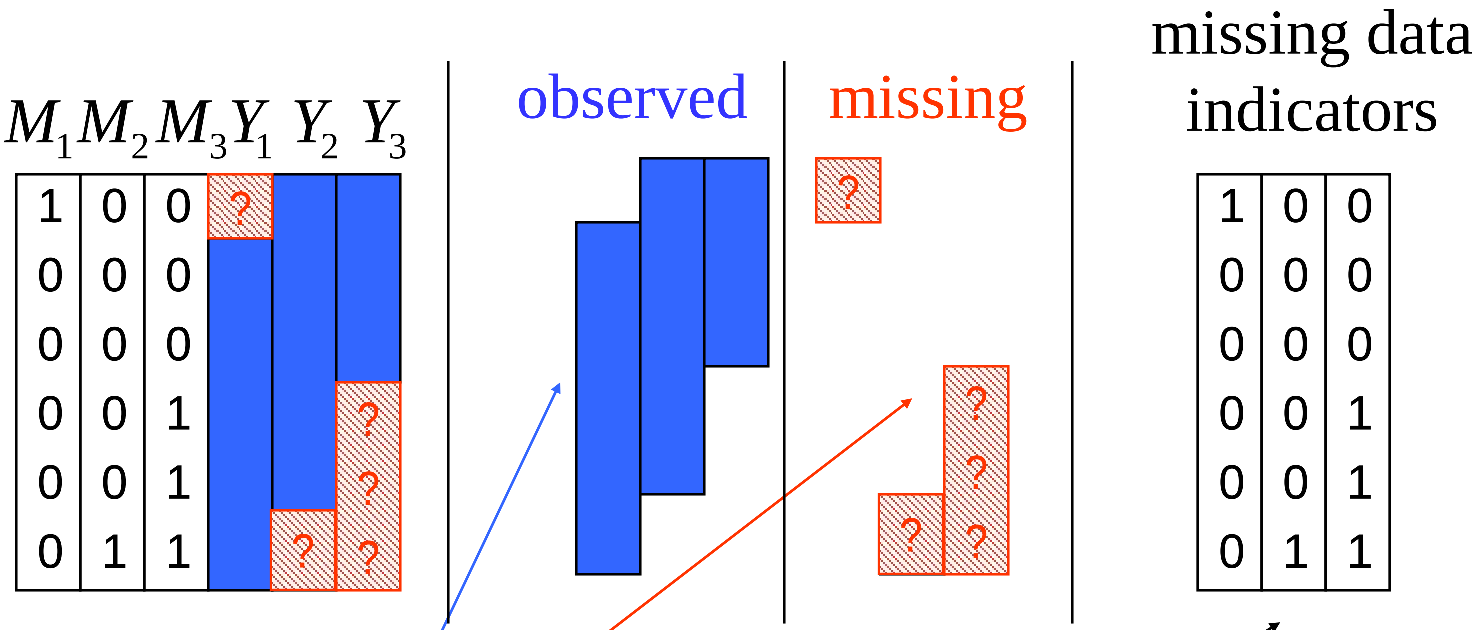
Patterns of Missing Data 2

- Item nonresponse: general pattern



Bayes can handle general patterns with relative ease

The Observed Data



$$Y = (y_{ij})_{n\times K} = (Y_{\text{obs}}, Y_{\text{mis}})$$

$$M = (m_{ij})_{n\times K}$$

$$m_{ij} = \begin{cases} 0, & y_{ij} \text{ observed} \\ 1, & y_{ij} \text{ missing} \end{cases}$$

Likelihood methods with missing data

- Statistical model + incomplete data → Likelihood
- Statistical models needed for:
  - data without missing values
  - missing-data mechanism
- Model for mechanism not needed if it is ignorable – missing at random (MAR) is the key condition
- With likelihood, proceed as before:
  - ML estimates, large sample standard errors
  - Bayes posterior distribution
  - Little and Rubin (2002, chapter 6)

Bivariate Monotone Data: Model for *Y* and *M*

$$f(Y,M|\theta,\psi) = f(Y|\theta)\times f(M|Y,\psi)$$

Complete-data model   model for mechanism

Example: bivariate normal monotone data

M_1	M_2	Y_1	Y_2
0	0		
0	0		
0	0		
0	1		?
0	1		?

complete-data model:

$$(y_{i1}, y_{i2}) \sim_{iid} N_2(\mu, \Sigma)$$

model for mechanism:

$$(m_{i2}|y_{i1}, y_{i2}) \sim_{ind} Bern\left[\Phi(\psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2})\right]$$

$\Phi =$ Normal cumulative distribution function

Two likelihoods

- *Full* likelihood - involves model for *M*

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

$$\Rightarrow L_{\text{full}}(\theta, \psi \mid Y_{\text{obs}}, M) = \text{const} \times f(Y_{\text{obs}}, M \mid \theta, \psi)$$

- Likelihood *ignoring the missing-data mechanism M*
  - simpler since it does not involve model for *M*

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

$$\Rightarrow L_{\text{ign}}(\theta \mid Y_{\text{obs}}) = \text{const} \times f(Y_{\text{obs}} \mid \theta)$$

Ignoring the missing-data mechanism

- Note that if:

$$L_{\text{full}}(\theta,\psi \mid Y_{\text{obs}}, M) = L(\psi \mid M, Y_{\text{obs}}) \times L_{\text{ign}}(\theta \mid Y_{\text{obs}})$$

where $L(\psi \mid M, Y_{\text{obs}})$ does not depend on θ

then inference about θ can be based on $L_{\text{ign}}(\theta \mid Y_{\text{obs}})$

- The missing-data mechanism is then called *ignorable* for likelihood inference

Ignoring the md mechanism continued

- Rubin (1976) showed that <u>sufficient</u> conditions for ignoring the missing-data mechanism are:

  (A) Missing at Random (MAR):

  $f(M \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) = f(M \mid Y_{\text{obs}}, \psi)$ for all Y_{mis}

  (B) Distinctness:

  θ and ψ have distinct parameter spaces

  (Bayes: priors distributions are independent)

- If MAR holds but not distinctness, ML based on ignorable likelihood is valid but not fully efficient, so MAR is the key condition
- For frequentist inference, need MAR for all M, Y_{mis}

(Everywhere MAR, see Seaman et al. 2013)

Two posterior distributions

$$p_{\text{complete}}(\theta,\psi \mid Y, M) = \pi(\theta,\psi) \times f(Y \mid \theta) \times f(M \mid Y, \psi)$$

Prior dn Complete-data model model for mechanism

- *Full* posterior distribution - involves model for *M*

$$p_{\text{full}}(\theta,\psi \mid Y_{\text{obs}}, M) \propto \pi(\theta,\psi) \times f(Y_{\text{obs}}, M \mid \theta,\psi)$$

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

- Posterior dn *ignoring the missing-data mechanism M* (simpler since it does not involve model for *M*)

$$p_{\text{ign}}(\theta \mid Y_{\text{obs}}) \propto \pi(\theta) \times f(Y_{\text{obs}} \mid \theta)$$

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

Ignoring the md mechanism continued

- Sufficient conditions for ignoring the missing-data mechanism and basing inference on $p_{\text{ign}}(\theta \mid Y_{\text{obs}})$ are:
- MAR: $f(M \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) = f(M \mid Y_{\text{obs}}, \psi)$ for all Y_{mis}
- Independent priors for parameters of dns of Y and M

$$\pi(\theta, \psi) = \pi_1(\theta) \times \pi_2(\psi)$$

- MAR is the key condition in practice
- Main challenges are choice of model, computation
- Missing Not at Random (MNAR) – missingness depends on missing data – harder problem

Remarks

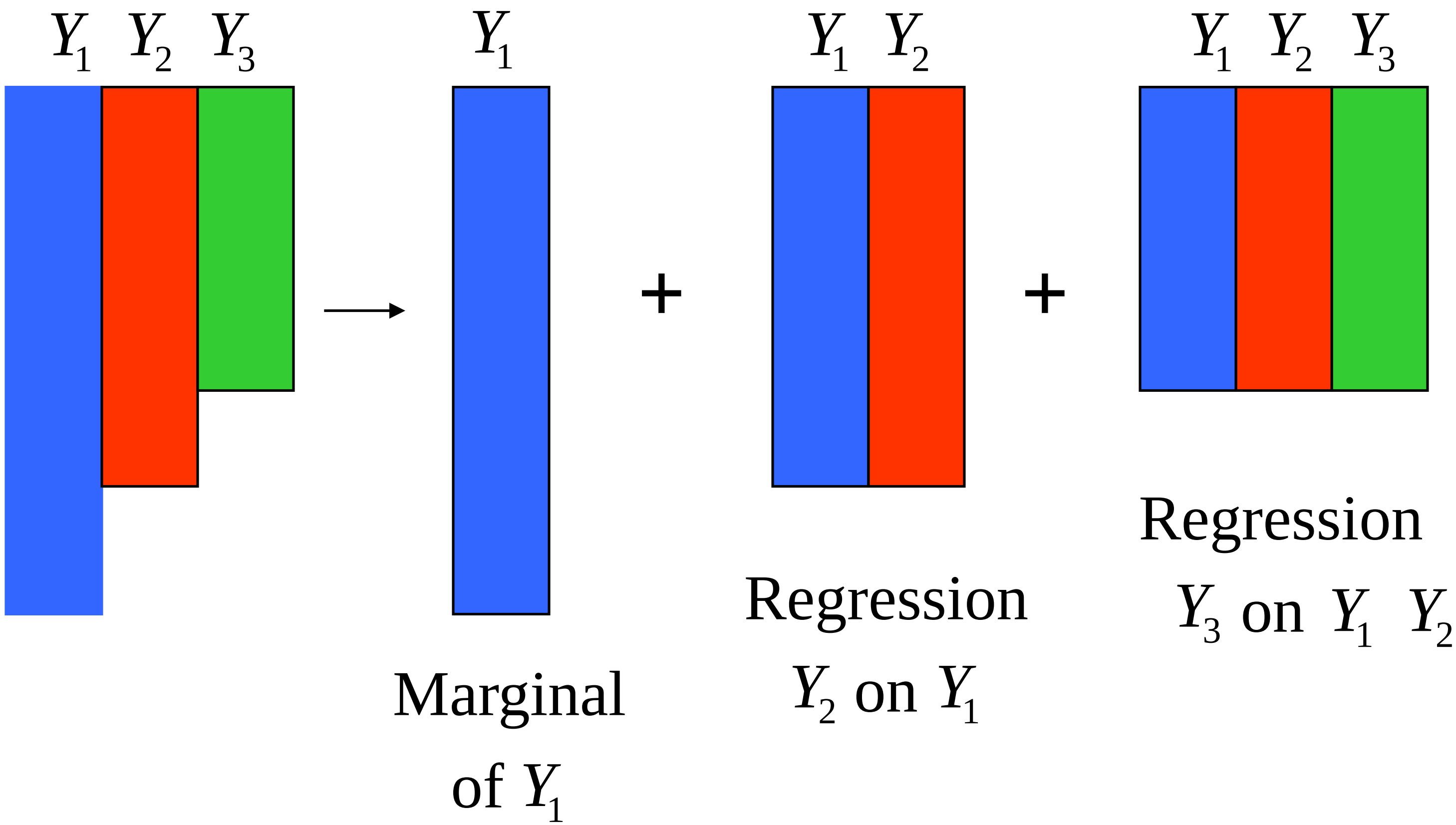
- When data are assumed to be MNAR. One has to specify the relationship between *M*, the response indicator, and the unobserved portion of substantive data, y_{mis}
  - This assumption cannot be verified form the observed data in hand
  - Some external information is needed to specify this relationship.
  - The MNAR is more or less a subjective opinion and as such inferences are highly sensitive to assumption made about the relationship between *M* and y_{mis}
- MAR conditional on rich set of covariates may be the most reasonable approach rather than making some empirically unverifiable assumption
- Since missing data may be a problem, attempts should be made to collect a rich set of correlates of missing outcomes
- Use designs to reduce nonresponse bias (multiple matrix sampling)

Computational tools

- Tools for monotone patterns
  - Maximum likelihood based on factored likelihood
  - Draws from Bayesian posterior distribution based on factored posterior distributions
- Joint distribution is factored into sequence of conditional distributions
- ML/Bayes is then a set of complete data problems (at least under MAR)

Monotone Data, Three blocks

Regress current on more observed variables using available cases; e.g. 3 variables:



ML: computational tools for general patterns

- ML usually requires iterative algorithms – general optimization methods like Newton Raphson and scoring, the EM algorithm and extensions (ECM, ECME, PXEM, etc.), or combinations
- Software – mixed model software like PROC MIXED, NLMIXED can handle missing values in outcomes, under MAR assumption
- This does not handle missing data in predictors
- MI has some advantages over this approach, as discussed later

Bayes: computational tools for general patterns

- Iterative algorithms are usually needed
- Bayes based on Gibbs' sampler (which also provides multiple imputations of missing values)
- Gibbs' is essentially a stochastic version of ECM algorithm, yielding draws from posterior distribution of the parameters
- These draws from an intermediate step for creating multiple imputations from the posterior predictive distribution of the missing values
- Chained equation MI: logic of the Gibbs' sampler, with flexible modeling of sequence of conditional distributions. Trades rigor for practical flexibility

Conclusion

- Bayesian prediction extends naturally to handle unit and item nonresponse
- MAR: missingness depends only on observed variables – key assumption for many methods