# Likelihood-based methods with missing data

Roderick Little
University of Michigan

# A good missing-data method…

- Makes use of partial information on incomplete cases, for reduced bias, increased efficiency
- Is frequency valid ("calibrated") inferences under plausible model for missing data (e.g. confidence intervals have nominal coverage) E.g. Little (2011, 2012)
- Propagates missing-data uncertainty, both within and between imputation models
- Focus here on likelihood based approaches
  - Maximum Likelihood (ML)
  - Bayes/Multiple Imputation

# Broad historical overview

- <1970: ad hoc imputation, ML for simple problems
- 1970-1985: ML for harder problems – EM algorithm and extensions; and multiple imputation
- 1985-2000: Bayes via MCMC; multiple imputation; Inverse Probability Weighting (IPW) methods and extensions
- >2000: diagnostics, robust modeling
- Literature applies to surveys!

# Likelihood methods

- Statistical model + data $\Rightarrow$ Likelihood
- Three general likelihood-based approaches:
  - Large-sample maximum likelihood inference
  - Bayesian inference: posterior = prior x likelihood
  - Parametric multiple imputation – essentially a simulation approximation to Bayes
  - Parameter estimates or draws are used to (in effect) create predictions of non-sampled or missing values
- Likelihood does not require rectangular data, so likelihood methods can be applied to incomplete data – but first consider complete data

# Parametric Likelihood

- Data $Y$
- Statistical model yields probability density $f(Y \mid \theta)$

  for $Y$ with unknown parameters
- Likelihood function is then a function of $\theta$

$$L(\theta \mid Y) = \text{const} \times f(Y \mid \theta)$$

- Loglikelihood is often easier to work with:

$$\mathrm{l}(\theta \mid Y) = \log L(\theta \mid Y) = \text{const} + \log\{ f(Y \mid \theta)\}$$

Constants can depend on data but not on parameter $\theta$

# Example: Normal sample

- univariate iid normal sample

$$Y = (y_1, \ldots, y_n)$$

$$\theta = (\mu, \sigma^2)$$

$$f(Y \mid \mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mu\right)^2 \right)$$

$$\mathrm{l}(\mu, \sigma^2 \mid Y) = -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mu\right)^2$$

# Computing the ML estimate

- In regular problems, the ML estimate can be found by solving the <u>likelihood equation</u>

$$S(\theta \mid Y) \equiv \frac{\partial \log L(\theta \mid Y)}{\partial \theta} = 0$$

where *S* is the <u>score function.</u>

Explicit solutions for some models (normal regression, multinomial, …)

Iterative methods – e.g. Newton-Raphson, Scoring, EM algorithm -- required for other problems (logistic regression, repeated measures models, non-monotone missing data)

# Properties of ML estimates

- Under assumed model, ML estimate is:
  - Consistent
  - Asymptotically efficient
  - Asymptotically normal

  $(\hat{\theta} - \theta) \sim N(0, C)$

  Superpopulation models: $\hat{\theta}$ random, $\theta$ fixed

  Bayesian models: $\theta$ random, $\hat{\theta}$ fixed
- Model checks are important – model should fit the observed data

# Forms of precision matrix

- The precision of the ML estimate or posterior distribution is measured by $C^{-1}$. Some forms for this are:
  - Observed information (respects ancillarity)

$$C^{-1} = I(\hat{\theta} \mid Y) = -\left. \frac{\partial^2 \log L(\theta \mid Y)}{\partial \theta \, \partial \theta} \right|_{\theta=\hat{\theta}}$$

  - Expected information (inferior, may be simpler)

$$C^{-1} = J(\hat{\theta}) = E\left[ I(\theta \mid Y, \theta) \right]_{\theta=\hat{\theta}}$$

  - Sandwich/bootstrap. Some robustness features, provided ML estimate is consistent

# Finite population inference

- Modeling takes a predictive perspective on statistical inference – predict the non-sampled values
  - ML models for the sampling/nonresponse weights lie outside this perspective
- Inference about parameters is intermediate step in predictive superpopulation model inference about finite population parameters
- Dependence on model makes this approach unpopular in the survey sampling world (though it is pervasive in other areas of statistics)
  - However, the degree of model dependence varies, and is reduced by probability sampling – for example, the "design-based" and "model-dependent" answers are the same in some basic problems.

# Finite population inference

- Models can and should reflect important design features such as stratification, weighting and clustering:
- Stratifying variables and weights are covariates in the model, from a prediction perspective
- Clustering: random effects
- Software: mixed models that allow fixed and random effects

# Bayes inference

- Given a prior distribution $p(\theta)$ for the parameters, inference can be based on the posterior distribution using Bayes' theorem:

$$p(\theta \mid Y) = \text{const.} \times p(\theta) \times L(\theta \mid Y)$$

- In surveys, weak priors are often preferred so that the inferences are data-driven
- For small samples, Bayes' inferences with weak priors based on the posterior distribution have better frequency properties than the large sample ML approximation, and provide credibility intervals that incorporate estimates of precision.

# Simulating Draws from Posterior Distribution

- With problems with high-dimensional $\theta$, it is often easier to draw values from the posterior distribution, and base inferences on these draws
- For example, if $(\theta_1^{(d)} : d = 1,\ldots, D)$

  is a set of draws from the posterior distribution for a scalar parameter $\theta_1$, then

  $\bar{\theta}_1 = D^{-1}\sum_{d=1}^{D}\theta_1^{(d)}$ approximates posterior mean

  $s_\theta^2 = (D-1)^{-1}\sum_{d=1}^{D}(\theta_1^{(d)} - \bar{\theta}_1)^2$ approximates posterior variance

  $(\bar{\theta}_1 \pm 1.96 s_\theta)$ or 2.5th to 97.5th percentiles of draws

  approximates 95% posterior credibility interval for $\theta$

These draws are a stepping stone to simulating the posterior predictive distribution of finite population quantities

# Likelihood methods with missing data

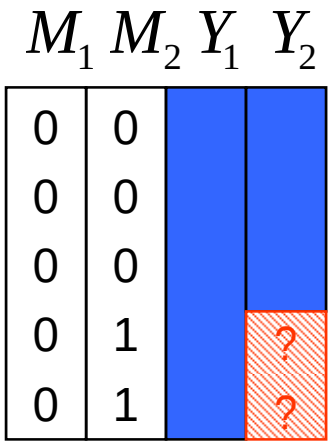
- Statistical model + incomplete data $\Rightarrow$ Likelihood
- Statistical models needed for:
  - data without missing values
  - missing-data mechanism
- Model for mechanism not needed if it is ignorable –MAR is the key condition
- With likelihood, proceed as before:
  - ML estimates, large sample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

# 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




complete-data model:

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

model for mechanism:

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

$\Phi =$ 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 $\theta$

then inference about $\theta$ 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):

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

  (B) Distinctness:

  $\theta$ and $\psi$ have distinct parameter spaces

  (Bayes: priors distributions are independent)

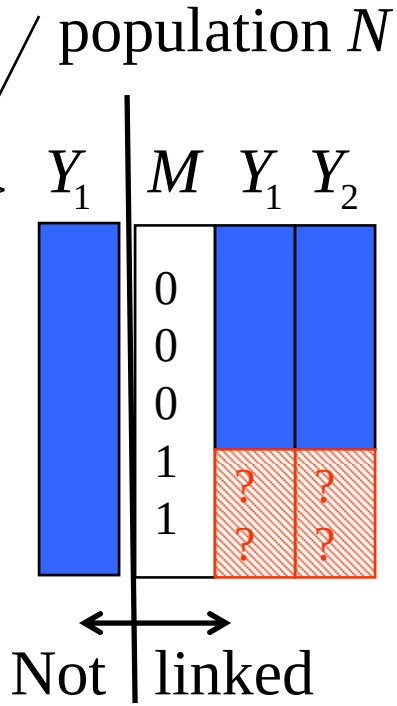
- If MAR holds but not distinctness, ML based on ignorable likelihood is valid but not fully efficient, so MAR is the key condition
- For frequentist inference, need MAR for all $M, Y_{mis}$

(Everywhere MAR, see Seaman et al. 2013)

# Auxiliary data can weaken MAR

$$D_{obs} = (D_{resp}, D_{aux})$$

$$D_{resp} = \{(y_{i1}, y_{i2}), i = 1, \ldots, m\}, D_{aux} = \{y^*_{j1}, j = 1, \ldots, n\}$$




$D_{aux}$ includes the respondent values of $Y_1$,

but we do not know which they are.

$$Y_1, Y_2 \sim_{ind} f(y_{i1}, y_{i2} \mid \theta)$$

$$\Pr(m_i = 1 \mid y_{i1}, y_{i2}, \phi) = g(y_{i1}, \phi)$$

Not MAR -- $y_{i1}$ missing for nonrespondents $i$

But... mechanism is ignorable, does not need to be modeled:

Marginal distribution of $Y_1$ estimated from $D_{aux}$

Conditional of $Y_2$ given $Y_1$ estimated from $D_{resp}$

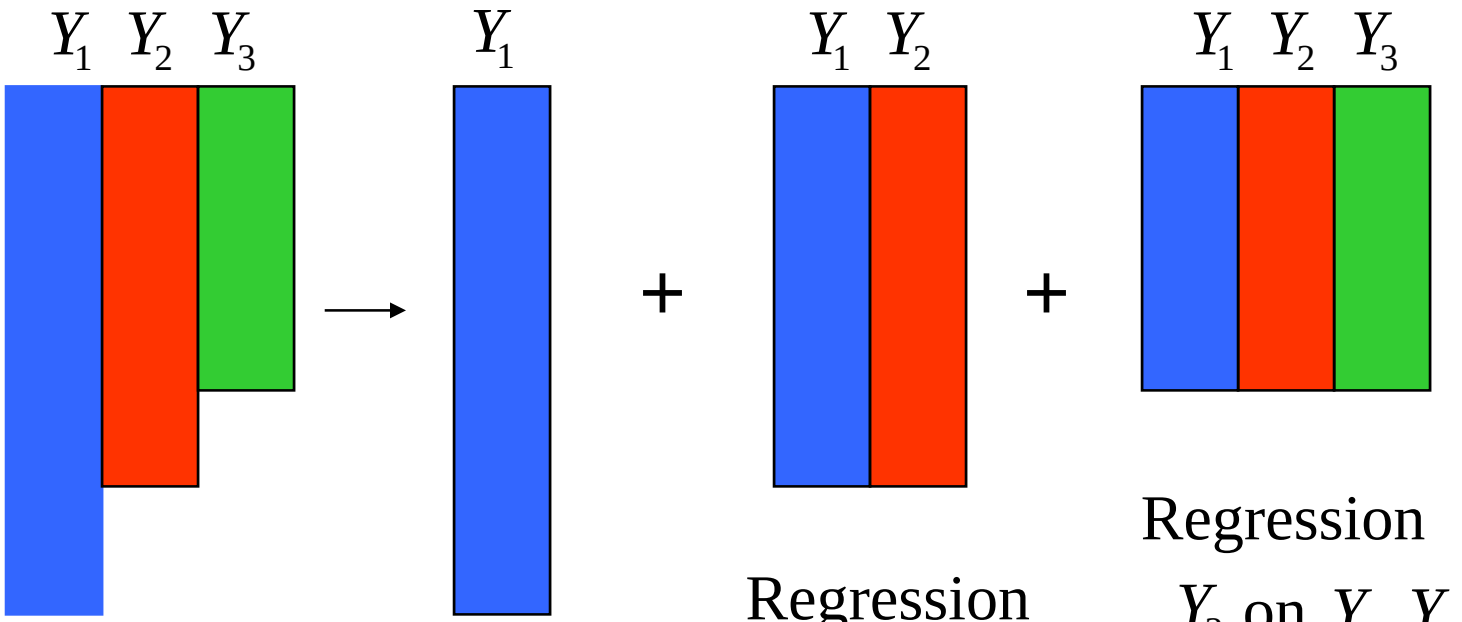
Little and Zangeneh (2014)

# 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

# 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

$Y = (Y_{obs}, Y_{mis})$, $Y_{obs}$ = observed data, $Y_{mis}$ = missing data

Complete-data posterior distribution,

if there were no missing values:

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

Posterior distribution given observed data:

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

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

(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 $\nu_{\text{com}}$ df, replace $\nu$ 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

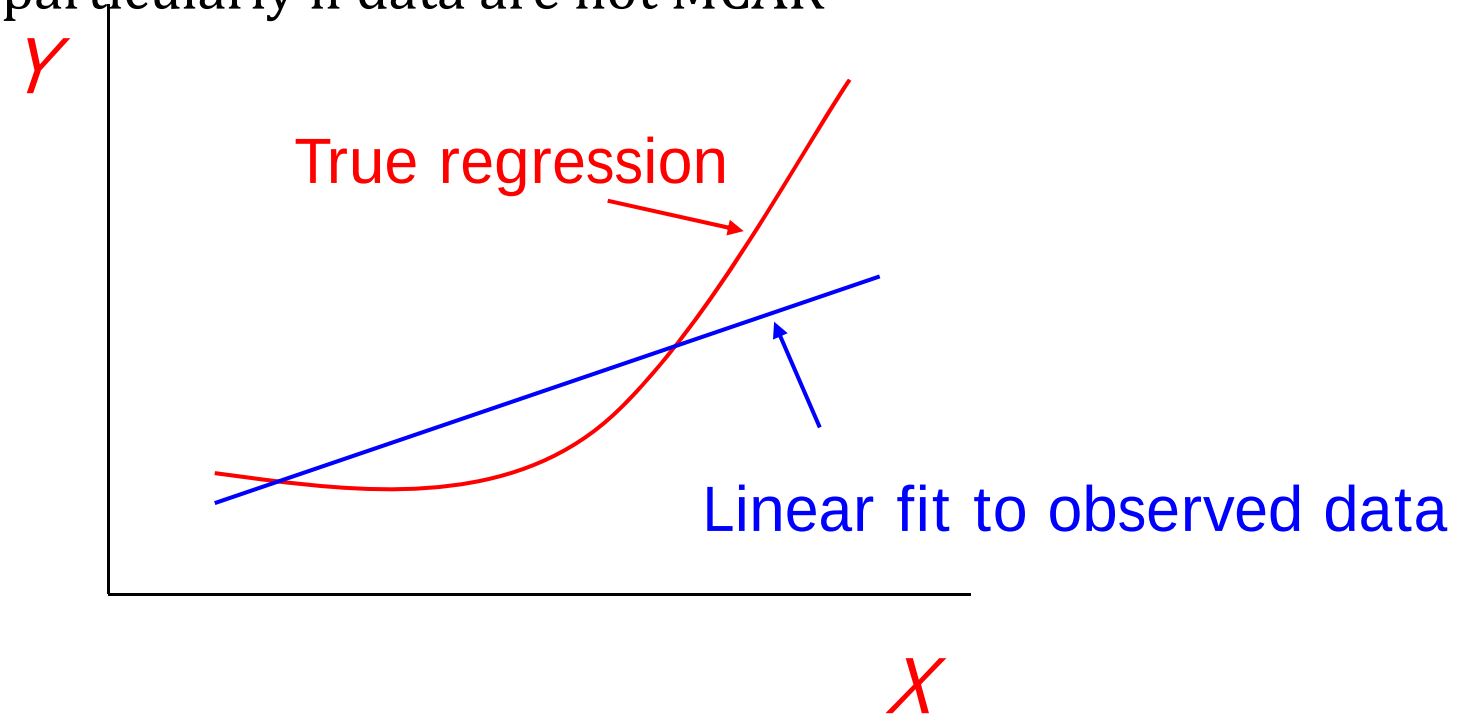
- It's model-based, and I don't want to make assumptions – but there is no assumption-free imputation method!
- Lack of congeniality between imputer model and analyst model
  - advice is to be inclusive of potential predictors, leading to at worst conservative inferences – parametric models allow main 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)

# Making MI's under MAR more robust

- Aim to reduce sensitivity of parametric MI's to model misspecification, particularly when data are not MCAR
- Hot deck methods like predictive mean matching
- Weaken regression assumptions of parametric MI's are potentially sensitive to model misspecification, particularly if data are not MCAR



# Penalized Spline of Propensity Prediction (PSPP)

- PSPP (Little & An 2004, Zhang & Little 2009, 2011).
- Regression imputation that is
  - Non-parametric (spline) on the propensity to respond
  - Parametric on other covariates
- Exploits the key property of the propensity score that conditional on the propensity score and assuming missing at random, missingness of *Y* does not depend on other covariates
- This property leads to a form of double robustness.

# PSPP method

Estimate: $Y^*$=logit (Pr(M=0|$X_1$,…,$X_p$ ))

Impute using the regression model:

$$(Y \mid Y^*, X_1, ..., X_p; \beta) \sim$$

$$N(s(Y^*) + g(Y^*, X_2, ..., X_p; \beta), \sigma^2)$$

- Nonparametric part
- Need to be correctly specified
- We choose penalized spline

- Parametric part
- Misspecification does not lead to bias
- Increases precision
- $X_1$ excluded to prevent multi-collinearity

# 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 $\lambda$

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

(West and Little, 2013)

# Conclusion

- Likelihood-based methods are flexible, and place the emphasis on the underlying imputation model rather than estimation formulae
- Model-based multiple imputation methods are attractive: practical, make use of all the data, and propagate imputation error – more in Joe Schafer's talk

# References


Andridge, R.H. & Little, R. J. (2010). A Review of Hot Deck Imputation for Survey Nonresponse. *International Statistical Review,* 78, 1, 40-64.

_____ & Little, R.J. (2011). Proxy Pattern-Mixture Analysis for Survey Nonresponse. *Journal of Official Statistics*, 27, 2, 153-180.

Little, R.J. (2011). Calibrated Bayes, for Statistics in General, and Missing Data in Particular (with Discussion and Rejoinder). *Statistical Science* 26, 2, 162-186.

____ (2012). Calibrated Bayes: an Alternative Inferential Paradigm for Official Statistics (with discussion and rejoinder). *Journal of Official Statistics*, 28, 3, 309-372.

____ & An, H. (2004). Robust Likelihood-Based Analysis of Multivariate Data with Missing Values. *Statistica Sinica*, 14, 949-968.

____ & Rubin, D. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley.

_____ & Zangeneh, S. (2014). Partially Missing At Random And Ignorable Inferences For Parameter Subsets With Missing Data. Submitted.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika* **63**, 581-592.

_____ (1987). *Multiple imputation for Nonresponse in Surveys.* Wiley

Seaman, S., Galati, J., Jackson, D. and Carlin, J. (2013). What Is Meant by "Missing at Random?" *Statistical Science*, 28, 2, 257–268.

Zhang, G. & Little, R. J. (2009). Extensions of the Penalized Spline of Propensity Prediction Method of Imputation. *Biometrics*, 65, 3, 911-918..

Zhang, G. & Little, R. J. (2011). A Comparative Study of Doubly-Robust Estimators of the Mean with Missing Data. *Journal of Statistical Computation and Simulation*, 81, 12, 2039-2058.

Giusti, C. & Little, R.J. (2011). A Sensitivity Analysis of Nonignorable Nonresponse to Income in a Survey with a Rotating Panel Design. *Journal of Official Statistics*, 27, 2, 211-229.

West, B. and Little, R.J. (2013). Nonresponse Adjustment Based on Auxiliary Variables Subject to Error. *Applied Statistics*, 62, 2, 213–231.