Handling Missing Data in Statistical Analyses

Michael R. Elliott

Assistant Professor of Biostatistics, Department of Biostatistics, School of Public Health

Assistant Research Scientist, Survey Methodology Program

University of Michigan

mrelliot@umich.edu

http://www.sph.umich.edu/~mrelliot

Overview

4 October 2007

14:00-15:00: Overview of Missing Data

- Examples
- Missing Data Patterns
- Missing Data Mechanisms
  - MCAR
  - MAR
  - NMAR

15:30-16:30: Analyzing via Maximum-Likelihood Estimation

- EM Algorithm

Overview

5 October 2007

9:30-10:30: Multiple Imputation I

- Motivation
- Gibbs Sampling

10:45-11:45: Multiple Imputation II

- Sequential Imputation
- Congeniality
- Software

What is Missing Data?

- Some missing data is structural:
  - Voting preferences of ineligible voters
  - Time since most recent Pap smear for males
- Consider data to be missing if a value meaningful for analysis is somehow hidden.

Examples

- Sample surveys
  - Unit non-response
  - Item non-response
- Censoring in longitudinal studies
  - Study termination
  - Drop-out
- Design non-response
  - Subsampling difficult-to-reach respondents (American Community Survey)
- Latent variables
  - "Random effects" to induce correlation structures
  - Factor analysis

What are some of the problems that missing data cause?

- Loss of information, efficiency or power due to loss of data.
- Complication in data handling, computation and analysis due to irregularities in the data patterns and nonapplicability of standard software.
- Potentially serious bias due to systematic differences between the observed data and the unobserved data.

Notation

- Data matrix $Z = \left(Z^{obs}, Z^{mis} \right)$
- Associated missingness matrix R, where

  $R_{ij} = 1$ if $Z_{ij} \in Z^{obs}$

  $R_{ij} = 0$ if $Z_{ij} \in Z^{mis}$

- Assume that cases i are independent (may include multiple observations on the same subject).

Z

0	12.7	3
1	.	2
0	11.4	2
0	.	.
.	9.6	1
.	.	.

R

1	1	1
1	0	1
1	1	1
1	0	0
0	1	1
0	0	0

Missing Data Patterns

- Patterns concern the marginal distribution of *R*.
- Certain patterns may allow for simpler or more direct techniques to be applied:
  - "Monotone" missingness patterns may allow ML estimates (under a "missing at random" mechanism assumption) to be obtained without resorting to data augmentation or imputation.

Missing Data Mechanisms

- Mechanisms concern the conditional distribution of $R|Z$.
- Missing Completely at Random (MCAR):

  $P(R|Z) = P(R)$

  - Missingness independent of the value of the data.
  - Ex: Have dataset consisting of age, gender, and blood pressure measure.
  - Blood pressure measurement is missing due to equipment breakage.

Missing Data Mechanisms

- Mechanisms concern the conditional distribution of $R|Z$.
- Missing at Random (MAR):

  $P(R|Z) = P(R|Z^{obs})$

  - Conditional on observed data, missingness is random.
  - Ex: Men over 65 are more likely and men under 34 are less likely to refuse to have their blood pressure taken, where a) age and gender are observed for all subjects, and b) within the age-gender groups, missingness is unrelated to blood pressure.

Missing Data Mechanisms

- Mechanisms concern the conditional distribution of $R|Z$.
- Not Missing at Random (NMAR):

  $P(R|Z)$ depends on Z^{mis}

  - Missingness depends on unobserved data elements even <u>after</u> conditioning on observed data.
  - Ex: Within all age/gender categories, the probability of blood pressure being refused is positively associated with subject's blood pressure.
  - NMAR parameters usually not estimable from data.

Missing at Random

Let θ be the set of parameters that govern the data Z and ψ be the parameters that govern the missingness indicator R. Factor the joint distribution of the data and the missingness indicators and note that

$$f(Z,R\mid\theta,\psi)=f(Z\mid\theta)f(R\mid Z,\psi)$$

Thus

$$f(Z^{obs},R\mid\theta,\psi)=\int f(Z^{obs},Z^{mis}\mid\theta)f(R\mid Z^{obs},Z^{mis},\psi)\,dZ^{mis}$$

Under MAR:

1) $f(R\mid Z,\psi)=f(R\mid Z^{obs},\psi)$
2) θ and ψ are distinct.

Thus (Rubin 1976)

$$L(\theta,\psi\mid Z^{obs},R)\propto f(R\mid Z^{obs},\psi)f(Z^{obs}\mid\theta)\Rightarrow L(\theta\mid Z^{obs})\propto f(Z^{obs}\mid\theta)$$

Not Missing at Random

Selection model: $f(Z,R|\theta,\psi) = f(Z|\theta)f(R|Z,\psi)$

Pattern-mixture model: $f(Z,R|\theta,\psi) = f(Z|R,\theta)f(R|\psi)$

(Little 1993)

Previous MAR model is a selection model under $f(R|Z,\psi) = f(R|Z^{obs},\psi)$ and θ,ψ distinct. NMAR selection models posit $f(R|Z^{obs},Z^{mis},\psi)$ (usually not estimable from the data).

Pattern-mixture models can allow for (restricted) NMAR models to be fit from observed data.

An Illustrative Example

$$y_1 = \log(income),\ y_2 = \begin{cases} 0 \text{ if} < 4 \text{ year college} \\ 1 \text{ if } 4+ \text{ years college} \end{cases}$$

$$y_1 \mid y_2 = I \sim N(\mu_I, \sigma_I^2) \quad y_2 \sim BERNOULLI(p)$$

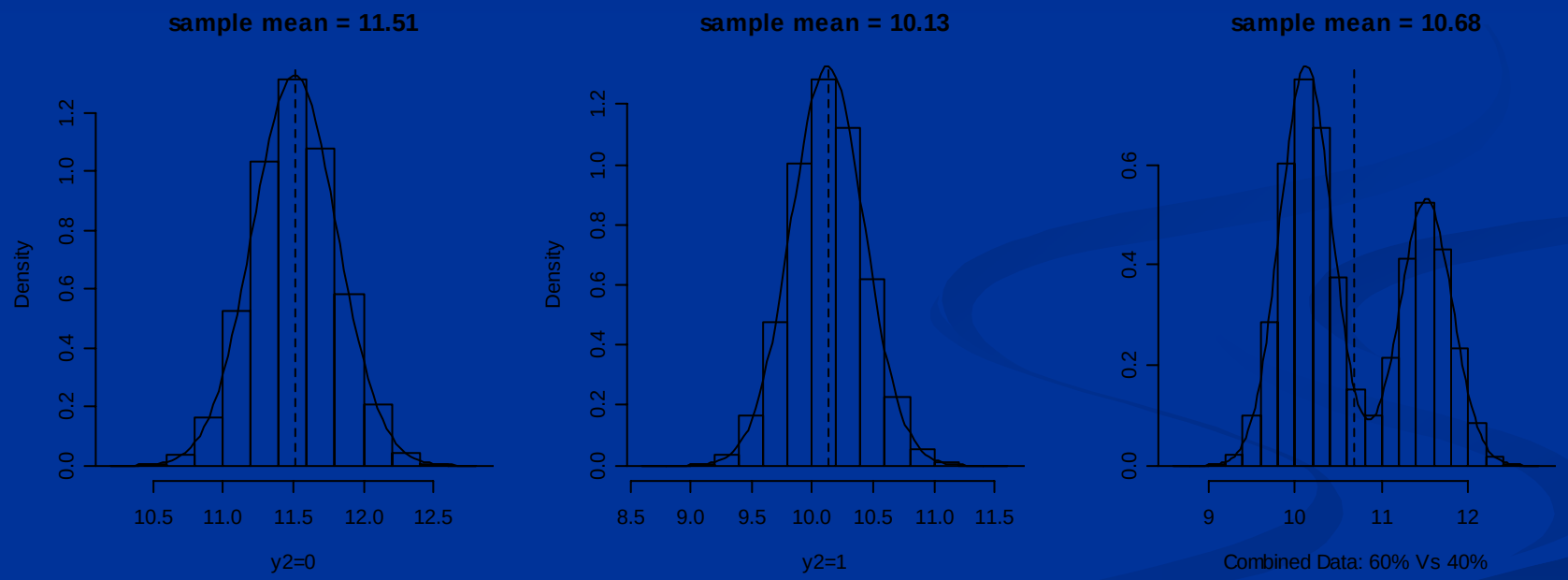
Assume that y_2 is fully observed but y_1 is subject to non-response. Suppose the missing data mechanism is given by

$$P(R=0) = \frac{e^{\beta_0+\beta_1 y_1+\beta_2 y_2}}{1+e^{\beta_0+\beta_1 y_1+\beta_2 y_2}}$$

Thus $\theta = (\mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$ and $\psi = (\beta_0, \beta_1, \beta_2)$

Complete Data

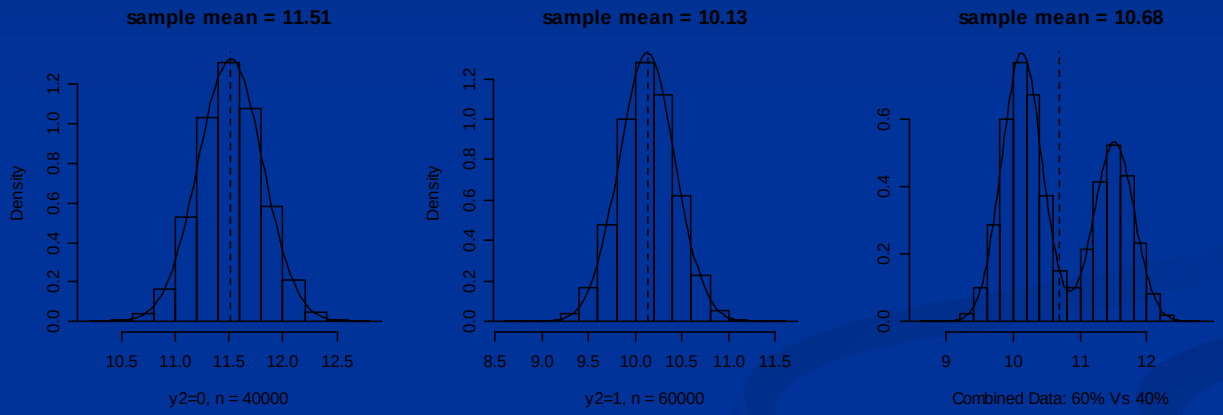
$$\mu_0 = \log(100000) = 11.51, \quad \mu_1 = \log(25000) = 10.13,$$
$$\sigma_0 = \sigma_1 = \log(1.35) = 0.30, \quad p = 0.4$$



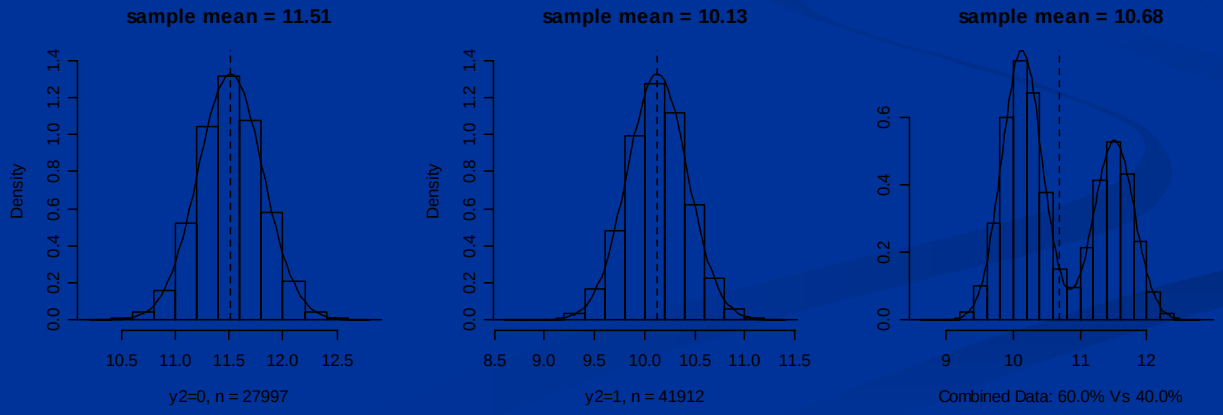
Missing Completely at Random

$$\beta_0 = -0.85,\ \beta_1 = \beta_2 = 0 \Rightarrow P(R=0) = \frac{e^{\beta_0}}{1+e^{\beta_0}} = 30\%$$

Complete



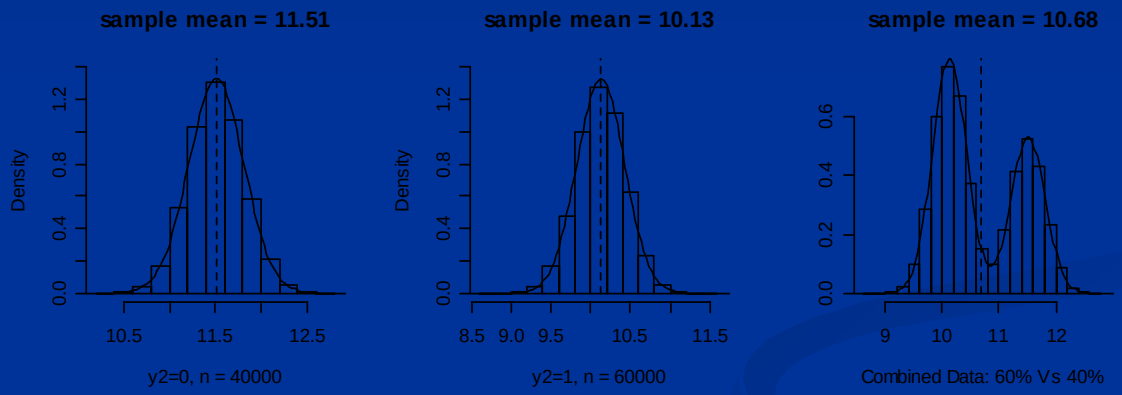
Observed



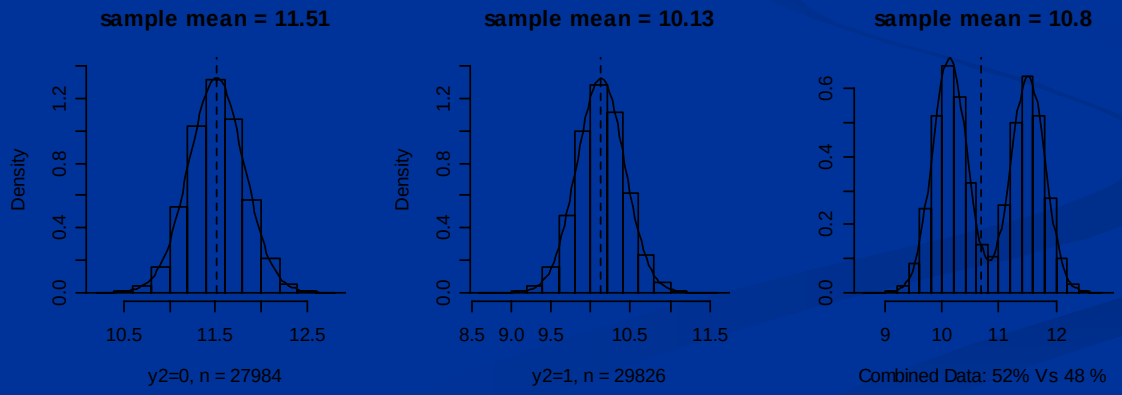
Missing at Random

Now Suppose $\beta_0 = -0.85$, $\beta_1 = 0$, $\beta_2 = 0.85$
$\Rightarrow P(R = 0|y_2 = 0) = 30\%$, $P(R = 0|y_2 = 1) = 50\%$

Complete



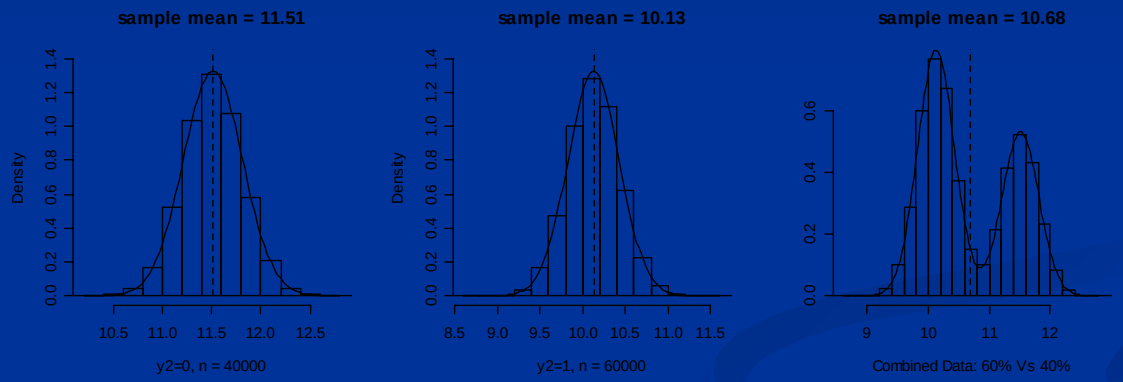
Observed



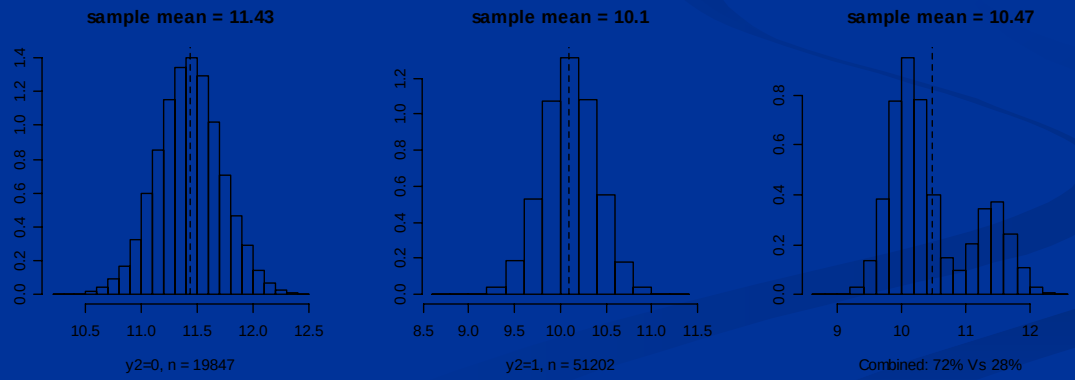
Not Missing at Random (Non-ignorable missing data)

$$\beta_0 = -23, \quad \beta_1 = 2, \quad \beta_2 = 0.85 \Rightarrow$$
$$\widehat{P}(R = 0|y_2 = 0) = 50.6\%, \ \widehat{P}(R = 0|y_2 = 1) = 14.5\%$$

Complete



Observed



Strategies for Analyzing Missing Data

- Complete-Case Analysis
- Analyze as Incomplete
- Imputation/Data Augmentation

Complete-Case Analysis

- Often OK if fraction of missing data is small (5-15%).
  - But, if item-level missingness is only a few percent for each item but is independent across predictor variables, a large number of cases can be discarded in a multivariate analysis.
- Can be biased if mechanism isn't MCAR, and is usually inefficient even if MCAR.

Analyze as Incomplete

- Obtain ML estimates (usually under either MCAR or MAR assumption).
- Development of algorithms may require certain missingness patterns (e.g., monotonicity).

Analyze as Incomplete

Reparameterize θ as ϕ and decompose log-likelihood (under MAR assumption)

$$l(\phi \mid Z^{obs}) = l_1(\phi_1 \mid Z^{obs}) + \cdots + l_J(\phi_J \mid Z^{obs})$$

where 1) $\phi_1, \ldots \phi_J$ are <u>distinct</u>

2) $l_j(\phi_j \mid Z^{obs})$ are either log-likelihood for complete-data or easier incomplete-data problems (Little and Rubin 2002).

- Maximize each $l_j(\phi_j \mid Z^{obs})$ → maximize $l(\phi \mid Z^{obs})$

Analyze as Incomplete

Obtain $\mathrm{Var}\left(\hat{\theta}\right)$ using inverse of information matrix and Delta method:

$$\mathrm{Var}\left(\hat{\theta}\right) = D\left(\hat{\theta}\right) I^{-1}\left(\hat{\phi}\middle| Z^{obs}\right) D^{T}\left(\hat{\theta}\right)$$

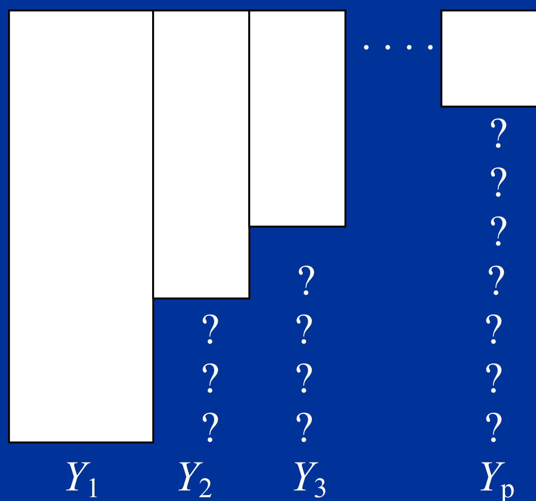
$$\left\{D_{jk}(\theta)\right\} = \left\{\frac{\partial \theta_j}{\partial \phi_k}\right\}$$

$$I^{-1}\left(\hat{\phi}\middle| Z^{obs}\right) = \begin{bmatrix} \left[-\frac{\partial^2}{\partial \phi_1^2} l_1(\phi_1 \mid Z^{obs})\right]^{-1} & & \\ & \ddots & \\ & & \left[-\frac{\partial^2}{\partial \phi_J^2} l_J(\phi_J \mid Z^{obs})\right]^{-1} \end{bmatrix}$$

Analyze as Incomplete

Example: Multivariate normal with monotone missing data pattern:

$$Y \sim MVN_p\left(\mu, \Sigma\right)$$



- Parameters of interest are $\theta = (\mu, \Sigma)$; the mean and covariance matrix for Y.
- Transform to

$$\phi = (\mu_{11}, \Sigma_{11}, \beta_{20\bullet 1}, \beta_{21\bullet 1}, \Sigma_{22\bullet 1}, \ldots,$$
$$\beta_{p0\bullet 1\ldots p-1}, \beta_{p1\bullet 1\ldots p-1}, \ldots, \beta_{p,p-1\bullet 1\ldots p-1}, \Sigma_{pp\bullet 1\ldots p-1})$$

the mean and variance of the fully-observed Y_1, the regression parameters of Y_2 on Y_1 and the residual covariance, the regression of Y_3 on Y_2 and Y_1, and so forth.

- Obtain μ_1, Σ_{11} for Y_1.
- Regress Y_2 on Y_1^{obs}, where Y_1^{obs} is the set of observations of Y_1 where Y_2 is also observed to obtain $\hat{\beta}_{20\bullet 1}, \hat{\beta}_{21\bullet 1}, \Sigma_{22\bullet 1}$.
- Regress Y_3 on Y_1^{obs}, Y_2^{obs} to obtain $\hat{\beta}_{30\bullet 12}, \hat{\beta}_{31\bullet 12}, \hat{\beta}_{32\bullet 12}, \Sigma_{33\bullet 12}$.
- Repeat through Y_p.
- Back transform the regression parameters to the MVN mean and variance parameters using the SWEEP operator (Beaton 1964, Little and Rubin 2002).

For a $p \times p$ symmetic matrix G, $k = 0,..., p$

$$SWP[k]G = H \rightarrow \begin{cases} h_{kk} = -1/g_{kk} \\ h_{jk} = h_{kj} = g_{jk}/g_{kk}, \\ h_{jl} = h_{lj} = g_{jl} - g_{jk}g_{kl}/g_{kk} \end{cases} \qquad RSW[k]G = H \rightarrow \begin{cases} h_{kk} = -1/g_{kk} \\ h_{jk} = h_{kj} = -g_{jk}/g_{kk}, \\ h_{jl} = h_{lj} = g_{jl} - g_{jk}g_{kl}/g_{kk} \end{cases}$$

Note that, for a 2x2 covariance matrix for Y_1 and Y_2, sweeping on row and column 1 yields a matrix whose off-diagonal elements are the regression coefficient of Y_1 when Y_1 is regressed on Y_2, and the lower-diagonal element is the residual variance.

$$A^1 = SWP[1]\begin{bmatrix} -1 & \hat{\mu}_1 \\ \hat{\mu}_1 & \hat{\Sigma}_{11} \end{bmatrix}$$

$$A^2 = SWP[2]\begin{bmatrix} A^1_{11} & A^1_{12} & \hat{\beta}_{20\bullet 1} \\ A^1_{21} & A^1_{22} & \hat{\beta}_{21\bullet 1} \\ \hat{\beta}_{20\bullet 1} & \hat{\beta}_{21\bullet 1} & \hat{\Sigma}_{22\bullet 1} \end{bmatrix}$$

$$\vdots$$

$$\begin{bmatrix} -1 & \hat{\boldsymbol{\mu}}^T \\ \hat{\boldsymbol{\mu}} & \hat{\Sigma} \end{bmatrix} = RSW[1,..,p-1]\begin{bmatrix} A^{p-1}_{11} & \cdots & A^{p-1}_{1p} & \hat{\beta}_{p0\bullet 1\ldots p-1} \\ \vdots & \ddots & \vdots & \vdots \\ A^{p-1}_{p1} & \cdots & A^{p-1}_{pp} & \hat{\beta}_{p,p-1\bullet 1\ldots p-1} \\ \hat{\beta}_{p0\bullet 1\ldots p-1} & \cdots & \hat{\beta}_{p,p-1\bullet 1\ldots p-1} & \hat{\Sigma}_{pp\bullet 1\ldots p-1} \end{bmatrix}$$

Bivariate Normal Example

Have n fully observed Y_1 and r fully observed Y_1 and Y_2.

$$f(y_{i1}, y_{i2} \mid \mu_1, \mu_2, \sigma_{11}^2, \sigma_{12}, \sigma_{22}^2) =$$
$$f(y_{i1} \mid \mu_1, \sigma_{11}^2) f(y_{i2} \mid y_{i2}, \beta_{20\bullet 1}, \beta_{21\bullet 1}, \sigma_{22\bullet 1}^2),$$
$$\beta_{20\bullet 1} = \mu_2 - \beta_{21\bullet 1}\mu_1,$$
$$\beta_{21\bullet 1} = \sigma_{12} / \sigma_{11}^2,$$
$$\sigma_{22\bullet 1}^2 = \sigma_{22}^2 - \sigma_{12}^2 / \sigma_{11}^2.$$

Bivariate Normal Example

From fully-observed data, have $\hat{\mu}_1 = n^{-1}\sum_{i=1}^{n} y_{i1}$, $\hat{\sigma}_{11}^2 = n^{-1}\sum_{i=1}^{n}(y_{i1} - \hat{\mu}_1)^2$.

Solving for μ_2, σ_{12}, and σ_{22}^2, we have

$\hat{\mu}_2 = \hat{\beta}_{20\bullet 1} + \hat{\beta}_{21\bullet 1}\hat{\mu}_1$

$\hat{\sigma}_{12} = \hat{\beta}_{21\bullet 1}\hat{\sigma}_{11}^2$

$\hat{\sigma}_{22}^2 = \hat{\sigma}_{22\bullet 1}^2 + \hat{\beta}_{21\bullet 1}^2\hat{\sigma}_{11}^2$

where $\hat{\beta}_{20\bullet 1}$ $\hat{\beta}_{21\bullet 1}$, and $\hat{\sigma}_{22\bullet 1}^2$ are obtained from the complete-case regression of Y_2 on Y_1.

(Note this is what is obtained by following the sweep algorithm on the previous page.)

Bivariate Normal Example

$$I^{-1}\left(\hat{\phi}\middle|Y^{obs}\right)=\begin{bmatrix} I^{-1}\left(\hat{\mu}_1,\hat{\sigma}_{11}^2\middle|Y^{obs}\right) & 0 \\ 0 & I^{-1}\left(\hat{\beta}_{20\bullet 1},\hat{\beta}_{21\bullet 1},\hat{\sigma}_{22\bullet 1}^2\middle|Y^{obs}\right)\end{bmatrix}$$

$$\text{Since } l\left(\hat{\phi}\middle|Y^{obs}\right)=-\frac{n}{2}\ln\sigma_{11}^2-\frac{\sum_{i=1}^{n}(y_{i1}-\mu_1)^2}{2\sigma_{11}^2}-\frac{r}{2}\ln\sigma_{22\bullet 1}^2-\frac{\sum_{i=1}^{r}(y_{i2}-\hat{\beta}_{20\bullet 1}-\hat{\beta}_{21\bullet 1}y_{i1})^2}{2\sigma_{22\bullet 1}^2},$$

$$I^{-1}\left(\hat{\mu}_1,\hat{\sigma}_{11}^2\middle|Y^{obs}\right)=\begin{bmatrix} n/\hat{\sigma}_{11}^2 & 0 \\ 0 & 2\hat{\sigma}_{11}^4/n\end{bmatrix},\quad I^{-1}\left(\hat{\beta}_{20\bullet 1},\hat{\beta}_{21\bullet 1},\hat{\sigma}_{22\bullet 1}^2\middle|Y^{obs}\right)=\begin{bmatrix} \hat{\sigma}_{22\bullet 1}^2(1+\bar{y}_1^{\,2}/s_{11}^2)/r & -\bar{y}_1\hat{\sigma}_{22\bullet 1}^2/(rs_{11}^2) & 0 \\ -\bar{y}_1\hat{\sigma}_{22\bullet 1}^2/(rs_{11}^2) & \hat{\sigma}_{22\bullet 1}^2/(rs_{11}^2) & 0 \\ 0 & 0 & 2\hat{\sigma}_{22\bullet 1}^4/r\end{bmatrix}$$

where $\bar{y}_j$, s_{ij} are based on the complete case data.

Thus

$$\hat{\text{Var}}\left(\hat{\mu}_2\right)=D\left(\hat{\mu}_2\right)I^{-1}\left(\hat{\phi}\middle|Y^{obs}\right)D^T\left(\hat{\mu}_2\right)=\hat{\sigma}_{22\bullet 1}^2\left[\frac{1}{r}+\frac{\hat{\rho}^2}{n(1-\hat{\rho}^2)}+\frac{(\bar{y}_1-\hat{\mu}_1)}{rs_{11}}\right],\ \hat{\rho}^2=\frac{\hat{\sigma}_{12}^2}{\hat{\sigma}_{11}^2\hat{\sigma}_{22}^2}$$

$$\text{since } D\left(\mu_2\right)=\left(\frac{\partial\mu_2}{\partial\mu_1},\frac{\partial\mu_2}{\partial\sigma_{11}^2},\frac{\partial\mu_2}{\partial\beta_{20\bullet 1}},\frac{\partial\mu_2}{\partial\beta_{21\bullet 1}},\frac{\partial\mu_2}{\partial\sigma_{22\bullet 1}^2}\right)=\left(\beta_{21\bullet 1},0,1,\mu_1,0\right)$$

Bivariate Normal Example

Note that, when missingness is MCAR, $(\bar{y}_1 - \hat{\mu}_1) \to 0$ as order $O(r^{-1})$, so

$$\hat{\mathrm{Var}}(\hat{\mu}_2) \approx \hat{\sigma}^2_{22\bullet 1}\left[\frac{1}{r} + \frac{\hat{\rho}^2}{n(1-\hat{\rho}^2)}\right] = \frac{\hat{\sigma}^2_{22}}{r}\left[1 - \hat{\rho}^2\frac{n-r}{n}\right]$$

- Thus, compared with a complete case analysis where the variance of $\bar{y}$ is given by σ^2_{22}/r, we see the factored likelihood method gives an estimator that is more efficient.
  - The increase in efficiency is a function of the correlation between Y1 and Y2, and the fraction of data that is missing.

Analyze as Incomplete

- ML estimates for other types of data (categorical, mixed categorical and normal) can be obtained as well.
- ML estimates for more general missing data patterns can be obtain via EM (expectation-maximization) algorithms.

EM algorithm

The EM algorithm is a "statistical" algorithm used to determine likelihood or posterior modes conditional on observed data. The algorithm capitalizes on the fact that parameter estimation would be easy if the data were not missing.

Sketch of EM Theory

Factor the full distribution into its observed and missing components:

$$f\left(Y \mid \theta\right) = f\left(Y_{obs} \mid \theta\right) f\left(Y_{mis} \mid Y_{obs}, \theta\right)$$

Taking the log of both sides yields

$$l(\theta \mid y) = l\left(\theta \mid Y_{obs}\right) + l\left(Y_{mis} \mid Y_{obs}, \theta\right)$$

$$\Rightarrow l\left(\theta \mid Y_{obs}\right) = l(\theta \mid y) - l\left(Y_{mis} \mid Y_{obs}, \theta\right)$$

Sketch of EM Theory

Taking expectation with respect to the missing data given the observed data and a current estimate of $\theta = \theta(t)$ yields

$$l\left(\theta \mid Y_{obs}\right) = Q\left(\theta \mid \theta^{(t)}\right) - H\left(\theta \mid \theta^{(t)}\right)$$

where

$$Q\left(\theta \mid \theta^{(t)}\right) = E\left(l\left(\theta \mid Y_{obs}, Y_{mis}\right) \middle| Y_{obs}, \theta^{(t)}\right)$$

$$= \int l\left(\theta \mid Y_{obs}, Y_{mis}\right) f\left(Y_{mis} \middle| Y_{obs}, \theta^{(t)}\right) dY_{mis}$$

$$H\left(\theta \mid \theta^{(t)}\right) = E\left(l\left(Y_{mis} \mid Y_{obs}, \theta\right) \middle| Y_{obs}, \theta^{(t)}\right)$$

$$= \int \underbrace{\ln f\left(Y_{mis} \mid Y_{obs}, \theta\right)}_{f} \underbrace{f\left(Y_{mis} \middle| Y_{obs}, \theta^{(t)}\right)}_{g} dY_{mis}$$

Sketch of EM Theory

Note that, by Jensen's inequality,

$$\int \ln\left(\frac{f}{g}\right) g\,dx \le \ln \int \left(\frac{f}{g}\right) g\,dx = \ln \int f\,dx = \ln 1 = 0$$

$$= \int \ln\left(\frac{g}{g}\right) g\,dx$$

$$\Rightarrow \int \ln(f)\, g\,dx \le \int \ln(g)\, g\,dx$$

so that $H\left(\theta \mid \theta^{(t)}\right) \le H\left(\theta^{(t)} \mid \theta^{(t)}\right)$

Choosing $\theta(t+1)$ to maximize $Q(\theta \mid \theta(t))$ will increase the log-likelihood, since $l\left(\theta^{(t+1)} \mid Y_{obs}\right) - l\left(\theta^{(t)} \mid Y_{obs}\right) =$

$$\underbrace{Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) - Q\left(\theta^{(t)} \mid \theta^{(t)}\right)}_{>0} -$$

$$\underbrace{H\left(\theta^{(t+1)} \mid \theta^{(t)}\right) - H\left(\theta^{(t)} \mid \theta^{(t)}\right)}_{<0}$$

Sketch of EM Theory

Hence the EM algorithm works by:

1) Determining the expected value of the complete-data log-likelihood conditional on the observed data and the current value of the parameter estimates (E-step)

2) Maximizing the parameter estimates given the expected values of the complete data log likelihood ≡ expected value of the complete-data sufficient-statistics if the distribution is a member of the exponential family. (M-step)

EM example: Censoring

Suppose we have $y \sim \exp(\lambda), f(y \mid \lambda) = \frac{1}{\lambda} e^{-y/\lambda}$, but we only observe the first m values of y for y<c for a known constant c:

$$L(\lambda) = \prod_{i=1}^{m} f(y_i \mid R_i = 1) p(R_i = 1) \prod_{i=m+1}^{n} p(R_i = 0)$$

$$= \prod_{i=1}^{m} f(y_i \mid y_i < c) P(y_i < c) \prod_{i=m+1}^{n} P(y_i \geq c)$$

$$= \prod_{i=1}^{m} \frac{\lambda^{-1} e^{-y_i/\lambda}}{1 - e^{-c/\lambda}} \left(1 - e^{-c/\lambda}\right) \prod_{i=m+1}^{n} e^{-c/\lambda} = \lambda^{-m} e^{-\left(\sum_i y_i + (n-m)c\right)/\lambda}$$

EM example: Censoring

Then

$$l(\lambda) = -m \log \lambda - \frac{\sum_i y_i + (n-m)c}{\lambda}$$

$$\frac{\partial l}{\partial \lambda} = -\frac{m}{\lambda} + \frac{\sum_i y_i + (n-m)c}{\lambda^2}$$

$$\hat{\lambda} = \bar{y} + \left(\frac{n}{m} - 1\right)c$$

EM example: Censoring

Complete data log-likelihood: $l(\lambda) = -n\log\lambda - \sum_{i=1}^{n} y_i / \lambda$

Linear in $s = \sum_{i=1}^{n} y_i$

E-step: $E(s \mid Y^{obs}, \lambda) = \sum_{i=1}^{m} y_i + \sum_{i=m+1}^{n} E(y_i \mid y_i > c) = \sum_{i=1}^{m} y_i + (n-m)\underbrace{(c+\lambda)}$

memoryless property of the exp.

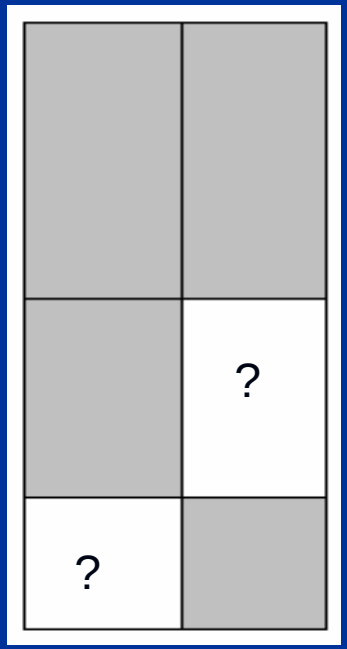
M-step: $\lambda^{(t+1)} = n^{-1}E(s \mid Y^{obs}, \lambda) = n^{-1}\left\{\sum_{i=1}^{m} y_i + (n-m)(c+\lambda^{(t)})\right\} \Rightarrow$

$$\hat{\lambda} = \frac{m}{n}\bar{y} + \left(1 - \frac{m}{n}\right)(c+\hat{\lambda}) \Rightarrow \hat{\lambda} = \bar{y} + \left(\frac{n}{m} - 1\right)c$$

Non-monotonic bivariate normal

Example: Multivariate normal with non-monotone missing data pattern:

$$Y \sim MVN_p(\mu, \Sigma)$$



EM for non-monotonic bivariate normal

The "complete data" log-likelihood is given by

$$l(\mu,\Sigma) = -n/2\log(2\pi|\Sigma|) - 1/2(y-\mu)^T\Sigma^{-1}(y-\mu)$$

$$= -n/2\log(2\pi) - n/2\ln(\sigma_{11}^2\sigma_{22}^2 - \sigma_{12}^2) -$$

$$\frac{\left[\sigma_{11}^2 s_{22} - 2\sigma_{12}s_{12} + \sigma_{22}^2 s_{11} + 2(\sigma_{12}\mu_1 - \sigma_{11}^2\mu_2)s_2 + 2(\sigma_{12}\mu_2 - \sigma_{22}^2\mu_1)s_1 + n(\mu_2^2\sigma_{11}^2 - 2\sigma_{12}\mu_1\mu_2 + \mu_1^2\sigma_{22}^2)\right]}{2(\sigma_{11}^2\sigma_{22}^2 - \sigma_{12}^2)}$$

which is clearly linear in the statistics $s_j = \sum_{i=1}^{n} y_{ij}$, $s_{jk} = \sum_{i=1}^{n} y_{ij}y_{ik}$ $j,k = 1,2.$

EM for non-monotonic bivariate normal

Assume WLOG that the first l observations are fully observed,
observations $l+1,\ldots,m$ are missing for Y2, and observations $m+1,\ldots,n$ are missing for Y1.

Then $E(s_1 \mid Y^{obs}) = \sum_{i=1}^{m} y_{i1} + \sum_{i=m+1}^{n} \hat{y}_{i1}$ $\quad E(s_2 \mid Y^{obs}) = \sum_{i=1}^{l} y_{i2} + \sum_{i=m+1}^{n} y_{i2} + \sum_{i=l+1}^{m} \hat{y}_{i2}$

$$E(s_{11} \mid Y^{obs}) = \sum_{i=1}^{m} y_{i1}^2 + \sum_{i=m+1}^{n} \tilde{y}_{i1}^2 \quad E(s_{22} \mid Y^{obs}) = \sum_{i=1}^{l} y_{i2}^2 + \sum_{i=m+1}^{n} y_{i2}^2 + \sum_{i=l+1}^{m} \tilde{y}_{i2}^2$$

$$E(s_{12} \mid Y^{obs}) = \sum_{i=1}^{l} y_{i1}y_{i2} + \sum_{i=l+1}^{m} y_{i1}\hat{y}_{i2} + \sum_{i=m+1}^{n} \hat{y}_{i1}y_{i2}$$

EM for non-monotonic bivariate normal

where $\hat{y}_{i1} = E(y_{i1} \mid y_{i2}, \mu, \Sigma) = \beta_{10\bullet 2} + \beta_{11\bullet 2} y_{i2}$, $\hat{y}_{i2} = E(y_{i2} \mid y_{i1}, \mu, \Sigma) = \beta_{20\bullet 1} + \beta_{21\bullet 1} y_{i1}$

$$\tilde{y}_{i1}^2 = E(\hat{y}_{i1}^2 \mid y_{i2}, \mu, \Sigma) = V(\hat{y}_{i1} \mid y_{i2}, \mu, \Sigma) + E^2(\hat{y}_{i1} \mid y_{i2}, \mu, \Sigma) = \sigma_{11\bullet 2}^2 + \hat{y}_{i1}^2$$

and similarly $\tilde{y}_{i2}^2 = \sigma_{22\bullet 1}^2 + \hat{y}_{i2}^2$

for

$$\beta_{10\bullet 2} = \mu_1 - \frac{\sigma_{12}}{\sigma_{22}^2}\mu_2, \quad \beta_{11\bullet 2} = \frac{\sigma_{12}}{\sigma_{22}^2}, \quad \sigma_{11\bullet 2}^2 = \sigma_1^2(1 - \sigma_{12}^2 / \sigma_1^2\sigma_2^2)$$

$$\beta_{20\bullet 1} = \mu_2 - \frac{\sigma_{12}}{\sigma_{11}^2}\mu_1, \quad \beta_{21\bullet 1} = \frac{\sigma_{12}}{\sigma_{11}^2}, \quad \sigma_{22\bullet 1}^2 = \sigma_2^2(1 - \sigma_{12}^2 / \sigma_1^2\sigma_2^2)$$

EM for non-monotonic bivariate normal

E-step: $s_1^{(t+1)} = \sum_{i=1}^{m} y_{i1} + \sum_{i=m+1}^{n} \hat{y}_{i1}^{(t)} \quad s_2^{(t+1)} = \sum_{i=1}^{l} y_{i2} + \sum_{i=m+1}^{n} y_{i2} + \sum_{i=l+1}^{m} \hat{y}_{i2}^{(t)}$

$$s_{11}^{(t+1)} = \sum_{i=1}^{m} y_{i1}^2 + \sum_{i=m+1}^{n} \tilde{y}_{i1}^{2(t+1)}, \quad s_{22}^{(t+1)} = \sum_{i=1}^{l} y_{i2}^2 + \sum_{i=m+1}^{n} y_{i2}^2 + \sum_{i=l+1}^{m} \tilde{y}_{i2}^{2\,(t+1)},$$

$$s_{12}^{(t+1)} = \sum_{i=1}^{l} y_{i1} y_{i2} + \sum_{i=l+1}^{m} y_{i1} \hat{y}_{i2}^{(t+1)} + \sum_{i=m+1}^{n} \hat{y}_{i1}^{(t+1)} \hat{y}_{i2}$$

where expectations are computed using current estimate $\mu^{(t)}, \Sigma^{(t)}$.

M-step: $\hat{\mu}_1^{(t+1)} = s_1^{(t+1)} / n, \quad \hat{\mu}_2^{(t+1)} = s_2^{(t+1)} / n$

$$\hat{\sigma}_1^{2^{(t+1)}} = s_{11}^{(t+1)} / n - \left[\hat{\mu}_1^{(t+1)}\right]^2, \quad \hat{\sigma}_2^{2^{(t+1)}} = s_{22}^{(t+1)} / n - \left[\hat{\mu}_2^{(t+1)}\right]^2, \quad \hat{\sigma}_{12}^{(t+1)} = s_{12}^{(t+1)} / n - \hat{\mu}_1^{(t+1)} \hat{\mu}_2^{(t+1)}$$

Imputation/Data Augmentation

- If proper, a general solution of MCAR, MAR, and NMAR missingness mechanisms.
- Allows more readily for Bayesian approaches to be developed/incorporated.
- Implementations tend to be parametric and model-based.

Basic Principles of Imputation

- Condition on observed variables.
  - Especially if predictive.
  - Even if not in the primary analyses of interest.
- Impute jointly so as to preserve correlation structures in the data.
- Impute draws from the predictive distributions, not means.
- Impute multiply to allow assessment of imputation variance.

(Little 1988)

Improper Imputation

"The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and the situations where standard estimators applied to the real and imputed data have substantial biases."

Dempster and Rubin (Incomplete Data in Sample Surveys, volume 2, 1983)

- Unconditional mean imputation
- Conditional mean imputation
- Hot Deck imputation

Unconditional mean imputation

- Impute Z_{ij}^{mis} from $\bar{Z}_j^{(j)}$, the available case mean.
  - Introduces bias in estimates of mean unless mechanism is MCAR.
  - Underestimates (cov)variance by $\left(n^{(jk)}-1\right)/(n-1)$ where $n^{(jk)}$ is the sample size of the available cases common to Z_j and Z_k.
  - Can introduce correction factors for covariance, but MCAR requirement often not met.

Conditional mean imputation

- Assume $Y \sim MVN_p(\mu, \Sigma)$ and compute estimates of μ and Σ from the complete cases.
- Use estimates to impute Z_i^{mis} from Z_i^{obs}
- Mean estimates are unbiased under MAR mechanism.
- Variance of Z_j still underestimated by $\left(n^{(j)} - 1\right)\sigma^2_{jj \bullet i} \big/ (n-1)$
- Other problems: e.g., imputing conditional means tends to underestimate tail distributions.
- But if missing data proportion is small, an easy way to correct for bias.

Hot deck imputation

- Use Z_i^{obs} as a "donor" to impute Z_j^{mis}.
- Preserves distributional structure (marginal and joint) of the data.
- Unbiased under MCAR
  - MAR mechanism can be approximated by forming homogeneous adjustment cells (i.e., MCAR mechanism within adjustment cell) and carrying out imputations within cells, or by defining distance measure metric and imputing from "nearest neighbors".
- Still need to account for imputation in inference.
  - Multiple imputation can account for the uncertainty in the imputed values by imputing more than once.

Multiple Imputation

- Multiple imputation uses repeated imputations under a stochastic model to induce correct inference. More formally:

$$f\left(\theta \middle| Z^{obs}\right) = \int f\left(\theta \middle| Z^{obs}, Z^{mis}\right) f\left(Z^{mis} \middle| Z^{obs}\right) dZ^{mis}$$

where $f\left(Z^{mis} \middle| Z^{obs}\right)$ is a posterior predictive distribution under an MAR selection model. (Rubin 1987, Schafer 1997).

Multiple Imputation: General algorithm

- Let θ be a parameter (or function of parameters) of interest under the complete-data model, estimated by a statistic $\theta(Z)$.
- Obtain m imputations $Z_{(1)}^{mis},\ldots Z_{(m)}^{mis}$ from $f\left(Z^{mis}\middle|Z^{obs}\right)$.
- Multiple-imputation point estimate of θ is

$$\hat{\theta} = m^{-1}\sum_{t=1}^{m}\hat{\theta}\left(Z^{obs},Z_{(t)}^{mis}\right)$$

Multiple Imputation: General algorithm

- If θ is scalar, asymptotic variance of $\hat{\theta}$ is estimated by the sum of the within-imputation and between-imputation variance

$$T = U + (1 + m^{-1})B$$

where

$$U = m^{-1}\sum_{t=1}^{m} \hat{\mathrm{Var}}(\hat{\theta}_{(t)})$$

$$B = (m-1)^{-1}\sum_{i=1}^{m} (\hat{\theta} - \hat{\theta}_{(t)})^2$$

Inference based on $T^{1/2}(\hat{\theta} - \theta)^T \sim t_\nu$ where

$$\nu = (m-1)\left[1 + \frac{U}{(1+m^{-1})B}\right]^2$$

Missing Information

- The fraction of missing information about θ relative to the complete data model is given by

$$\lambda = \frac{(r+2)/(\nu+3)}{r+1}$$

where $r = \dfrac{U}{(1+m^{-1})B}$ is the relative increase in variance due to non-response.

Number of Imputations

- Often, m can be small (usually 3-10) to obtain stable inference about θ.
- Relative efficiency of a point estimate based on m imputations is given by $(1+\lambda/m)^{-1}$.If $\lambda = .5$ and $m = 5$, RE=0.95.

Heuristic Bayesian Justification

$$E\left[\theta \middle| Z^{obs}\right] = E_{Z^{mis}|Z^{obs}}\left[E\left[\theta \middle| Z\right] \middle| Z^{obs}\right]$$

$$\mathrm{Var}\left[\theta \middle| Z^{obs}\right] = E_{Z^{mis}|Z^{obs}}\left[\mathrm{Var}\left[\theta \middle| Z\right] \middle| Z^{obs}\right] + \mathrm{Var}_{Z^{mis}|Z^{obs}}\left[E\left[\theta \middle| Z\right] \middle| Z^{obs}\right]$$

- Adjust for finite m using t distribution and Satterwaite approximation.
- Rubin (1987) shows that these results approximate the observed-data posterior for θ based on m imputations.

Multiple Imputation: Multivariate parameters

- Alternative reference distribution for k-component θ. (Li, Raghunathan, and Rubin 1991; Schafer 1997, p.112-114)
- Let $\hat{\theta}$ and U be the multivariate analogs of the univariate case, and let

$$B = (m-1)^{-1}\sum_{i=1}^{m}(\hat{\theta}-\hat{\theta}_{(t)})(\hat{\theta}-\hat{\theta}_{(t)})^T$$

- More stable estimate of variance is given by

$$T = (1+r_1)U, \quad r_1 = \left(1+m^{-1}\right)\mathrm{tr}(BU^{-1})/k$$

and $D_1 = (\hat{\theta}-\theta_0)^T T^{-1}(\hat{\theta}-\theta_0)/k$ has an F_{k,v_1} distribution under $H_0 : \theta = \theta_0$, where

$$v_1 = 4 + (t-4)[1+(1-2t^{-1})r_1^{-1}]^2, \quad t = k(m-1).$$

Gibbs sampling

- Obtain draws from posterior distribution of $\theta \mid \text{data}$ for $\theta = (\theta_1, \ldots, \theta_q)^T$ by initializing θ at $\theta^{(0)}$, drawing $\theta_1^{(1)}$ from $\theta_1^{(1)} \mid \theta_2^{(0)}, \ldots, \theta_q^{(0)}, \text{data}$, $\theta_2^{(1)}$ from $\theta_2^{(1)} \mid \theta_1^{(1)}, \ldots, \theta_q^{(0)}, \text{data}$, and so forth. As $T \to \infty$, $\theta^{(T)} \sim \theta_1, \ldots, \theta_q \mid \text{data}$. (Gelfand and Smith 1990; Gelman and Rubin 1992)
- Data augmentation in a Bayesian framework is simple in principle: obtain a draw of $Z^{mis} \mid \theta, Z^{obs}$ then of $\theta \mid Z^{obs}, Z^{mis}$.

Gibbs sampling

- Obvious extension when utilizing Gibbs sampler.

$$\theta_1^{(1)} \left| \theta_2^{(0)}, \ldots, \theta_q^{(0)}, \text{data}, Z^{mis(0)} \right.$$

$$\vdots$$

$$\theta_q^{(1)} \left| \theta_1^{(1)}, \ldots, \theta_{q-1}^{(1)}, \text{data}, Z^{mis(0)} \right.$$

$$Z_1^{mis(0)} \left| \theta^{(1)}, \text{data}, Z_2^{mis(0)}, \ldots, Z_r^{mis(0)} \right.$$

$$\vdots$$

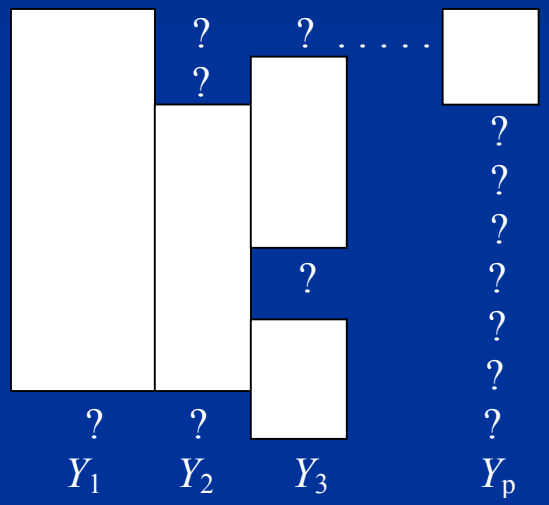
$$Z_r^{mis(0)} \left| \theta^{(1)}, \text{data}, Z_1^{mis(1)}, \ldots, Z_{r-1}^{mis(1)} \right.$$

Convergence, autocorrelation.

- How large does *T* have to be to truly obtain draws from the posterior (convergence)?
- Monitor by starting chains from widely separated areas in the parameter space and see if converge (ratio of between-to-within chain variance → 0).
  - Generally ratio of the square root of the total variance to the within-chain variance < 1.1 is sufficient.
- If draws have high autocorrelation, *T* may need to be in the thousands to obtain convergence.
- Need imputations to be independent
  - Use widely separated draws from a single chain.
  - Use draws from independent chains.

Multivariate normal

- MV normal with non-monotonic missing data:
  $Y \sim MVN_p(\mu, \Sigma)$



- Assume a non-informative prior:
  $$p(\mu, \Sigma) \propto |\Sigma|^{-p/2}$$

Multivariate normal

- Initialize by imputing missing elements via conditional imputation and initialize $\mu_{(0)}$ by a draw from $N_p(\bar{y}_{(0)}, C)$, where C is the sample covariance matrix from $Y_{(0)} = (Y^{obs}, Y^{mis}_{(0)})$

1. Draw
$$\Sigma^{(1)} \left| \mu^{(0)}, y^{(0)} \sim Inv - Wishart_{n-1}(S), S = \sum_i \left(y_i^{(0)} - \mu^{(0)} \right) \left(y_i^{(0)} - \mu^{(0)} \right) \right.$$
2. Draw $\mu^{(1)} \left| \Sigma^{(1)}, y^{(0)} \sim N_p(\bar{y}^{(0)}, \Sigma^{(1)} / n) \right.$
3. Draw
$$y_{i1}^{mis\,(1)} \left| \mu^{(1)}, \Sigma^{(1)}, y_{i2}^{(0)}, \ldots, y_{ip}^{(0)} \sim N\left(\beta_{10.\text{rest}}^{(1)} + \beta_{11.\text{rest}}^{(1)} y_{2,\ldots,p}^{(0)}, \sigma_{1.\text{rest}}^{2(1)} \right) \right.$$
$$\beta_{10.\text{rest}}^{(1)} = \mu_1^{(1)} - \beta_{11.\text{rest}}^{(1)} \mu_{2,\ldots,p}^{(1)}$$
$$\beta_{11.\text{rest}}^{(1)} = \Sigma_{1,2,\ldots,p}^{(1)} \left[\Sigma_{2,\ldots,p,2,\ldots,p}^{(1)} \right]^{-1}$$
$$\sigma_{1.\text{rest}}^{2(1)} = \sigma_{11}^{2(1)} - \Sigma_{1,2,\ldots,p}^{(1)} \left[\Sigma_{2,\ldots,p,2,\ldots,p}^{(1)} \right]^{-1} \Sigma_{2,\ldots,p,1}^{(1)}$$

Multivariate normal

4. Repeat 3. for $y_2^{mis\,(1)},\ldots,y_p^{mis\,(1)}$.
5. Cycle through 1.-4.; imputations are taken after convergence and spaced far enough apart to eliminate correlation.

- Algorithms have been developed for a variety of data models, including multinomial and mixed normal and multinomial ("general location" model).

NMAR Multiple Imputation

- MAR selection model can be extended to NMAR selection model.
  - Raghunathan and Siscovick (1996) consider whether the risk of sudden cardiac death is related to the use of thiazide diuretics using a case control study, adjusting for a number of potential medical confounders and smoking status.
  - Use multiple imputation to impute missing covariates.
  - Conduct sensitivity analysis with respect to smoking status (17% missing) using non-ignorable imputation model that assumes smokers are more likely to have known smoking status than non-smokers.
  - Impute smoking with probability $\delta\phi$, where ϕ is the probability under ignorability and $0 \le \delta \le 1$.

Sequential Imputation

- In practice, some datasets (e.g., health surveys) may have dozens or even hundreds of variables with missing data.
  - Complex missing data patterns: missing data may be structural (years of smoking for non-smokers) or truly missing (years of smoking for smokers).
  - Many different types of variables (continuous, categorical, count).

- Raghunathan et al. (2001) proposed sequential imputation as a way to deal with these situations.

Sequential Imputation

- Sequential imputation proceeds by ordering variables $Y_1,...,Y_n$ in order of their fraction of missingness (lowest to highest). (Let X denote the set of variables that have no missing values.)
- Begin by filling in the missing data using some reasonable imputation technique (e.g., impute missing Y_1 conditional on X and parameters θ_1 estimated from the complete-case data, impute missing Y_2 conditional on X, Y_1, and parameters θ_2 estimated from the complete-case data, etc.)

Sequential Imputation

- Rather than drawing from the posterior distribution of θ (conditional on Y^{obs}, Y^{mis} and X), and then imputing Y^{mis} conditional on Y^{obs}, X, and θ, we draw from the posterior distribution of θ_1 given Y^{obs}, Y^{mis} and X , then impute Y_1^{mis} given θ_1 and Y^{obs}, Y^{mis}, X, draw from the posterior distribution of θ_2 given Y^{obs}, Y^{mis} and X, then impute Y_2^{mis} given $\theta_{2\ and}$ Y^{obs}, Y^{mis}, X, , and so forth.
- Could possible fail to preserve the joint distribution of the elements of Y because the conditional densities from which the draws are obtained are not compatible with any multivariate distribution of Y|X.
  - In practice this does not appear to be a major issue.

Sequential Imputation

Example: Cigarette use.

- Some subjects are missing smoking status (current smoker, past smoker, never smoker), and thus have no data entered on number of cigarettes currently smoked. Other subjects are known to be current smokers but have missing data on number of cigarettes smoked per day.
- Impute smoking status, and for those imputed to be current smokers as well as those known to be current smokers but whose daily cigarette use is missing, impute daily cigarette use.

Sequential Imputation

Let Y_1 represent smoking status.

$$Y_{i1} \sim MULTI(1, \pi_{i1}, \pi_{i2}, \pi_{i3}); \quad \log(\pi_{ij} / \pi_{i3}) = X_i \beta_j,\ j = 1,2,\ \pi_3 = (1 + \sum_j \exp(X_i \beta_j))^{-1}$$

Regress Y_1^{obs} and the most recent imputation of Y_1^{mis} on X to obtain the MLE of β=B, and the associated covariance matrix V.

Compute β^*=B+Tz, where T^TT=V and z~N(0,I).

Compute $P_{ij} = \exp(X_i \beta_j^*) / (1 + \sum_j \exp(X_i \beta_j^*)), \quad j = 1,2,$ and $P_{i3} = 1 - \sum_j P_{ij}$

For all missing elements, let $R_{i0} = 0, R_{i1} = P_{i1}, R_{i2} = P_{i1} + P_{i2}, R_{i3} = 1$. Draw $u_i \sim UNI[0,1]$.

Impute $Y_i^* = j$ where $R_{i,j-1} \le u_i < R_{i,j.}$

Sequential Imputation

Let Y_2 represent cigarettes use among current smokers.

$$Y_{i2} \sim POI(\lambda_i); \quad \log(\lambda_i) = X_i\beta$$

Regress Y_2^{obs} and the most recent imputation of Y_2^{mis} on X and Y_1^{obs} ,Y_1^{mis} to obtain the MLE of β=B, and the associated covariance matrix V.

Compute β^*=B+Tz, where T^TT=V and z~N(0,I).

Compute $\lambda_i^* = \exp(X_i\beta_j^*)$

For all missing elements, generate Y_{i2}^*~POI(λ_i^*).

Congeniality

- What happens is the imputer and analyst are different, or, more precisely, if the imputation model and the analytic model do not correspond?
- This situation was addressed in Meng (Statistical Science 1994), who coined the term "uncongeniality" to describe this situation.

Congeniality

- When the imputation is made under the correct model, inferences under an incorrect model will tend to be conservative.

- When the imputation model itself is incorrect, inferences may be conservative or anticonservative, depending on the nature of the model failure.

Congeniality

- Ex: X fully observed, Y partially observed, where

$$X \sim UNI(0,b)$$

$$Y = e^X + \varepsilon, \quad \varepsilon \sim N(0,1)$$

- Can show that

$$\beta = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} = \frac{12\left[e^b(b-2)+b+2\right]}{2b^3}$$

- As b-> 0, linear approximation improves; as b ->∞, linear approximation fails.
- Assume an MAR missingness mechanism for Y given by .

$$P(y \text{ observed} \mid x) = \frac{1}{x+1}$$

Probability of response ~80% for b=.5, declining to 55% for b=2

Congeniality

- Imputation under correct model

$$X \sim UNI(0, b)$$

$$Y = e^X + \varepsilon, \quad \varepsilon \sim N(0,1)$$

- Analysis under linear model

$$Y = \alpha + \beta X + \varepsilon$$

$$\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$$

- Uncongenial, but imputation model correct

Congeniality

- 100 simulations of samples of size 100 under different values of *b*:

b	True β	Mean $\hat{\beta}$	Coverage of Nominal 90% CI
.5	1.29	1.23	96
1.0	1.69	1.72	96
1.5	2.24	2.23	98
2.0	3.00	3.00	98

Congeniality

- Imputation and analysis under (incorrect) linear model
- 100 simulations of samples of size under different values of *b*:

b	True β	Mean $\hat{\beta}$	Coverage of Nominal 90% CI
.5	1.29	1.22	92
1.0	1.69	1.58	84
1.5	2.24	2.15	86
2.0	3.00	2.79	78

Congeniality

Because missing data is very common in population surveys, uncongeniality is often an issue:

- Imputation is made under Bayesian models that do not easily accommodate complex sample design considerations.
- Complete data analysis can then employ techniques (linearization, replication methods) that account for the complex sample design (clustering, unequal probability of selection) in the analysis.
- An open area for research.

Multiple Imputation with Hot Deck

- Rubin and Schenker (1986) suggest using a Bayesian bootstrap technique to obtain a proper imputation procedure in a hot-deck setting
  - Suppose each element in the population takes one of the values $d_1,...,d_K$ with probability $\theta_1,...,\theta_K$
  - If an improper prior of the form $p(\theta) = \prod_k \theta_k^{-1}$ is used, then

    $p(\theta \mid y) = \prod_k \theta_k^{n_k - 1}$ where

    n_k = number of times an observation y takes on the value q_k

    which is a Dirichlet distribution with parameters n_k

Multiple Imputation with Hot Deck

- Draw θ^* from the posterior distribution of $\theta_1,...,\theta_K$, then draw Y^{mis} from Y^{obs} with probability θ^* .
- Standard hot deck imputation (sampling each unique value with replacement with probability $q_k \rangle n$ does not account for the uncertainty in the empirical distribution function.
- Can extend by stratifying by adjustment cells.
- Complex sample design? Stratification, clustering, unequal probability of selection.

An Example of MI in Practice: Multiple Imputation in the Presence of Outliers (Elliott and Stettler 2006)

- To ascertain the prevalence of pediatric obesity in medically underserved areas, the Healthy For Life Survey obtained data from a probability sample of children using Health Resource and Service Administration (HSRA) supported Community Health Centers at least once during calendar year 2001 (Stettler et al. 2005).

- Compute body-mass index (BMI) and Box-Cox transform as a function of age and gender; if BMI ``z-score'' exceeds 95th percentile of reference population, child is classified as obese.

An Example of EM in Practice: Multiple Imputation in the Presence of Outliers

- Abstract height and weight during last visit to the health clinic in 2001.
- One-fourth of height data missing.
  - Height measured only sporadically; less likely to be observed among older children and children seen more frequently at the clinic.
- Use multiple imputation to reduce bias and inefficiency associated with a complete-case analysis.
  - Potentially problematic: data overdispersed and included incorrectly recorded or abstracted elements.
  - Failure to account for abstraction errors may cause insufficient standardization between centers to be interpreted as unequal risk for pediatric obesity.

An Example of EM in Practice: Multiple Imputation in the Presence of Outliers

- Standardization in multi-center studies is expensive; propose analytic alternative to outlier correction when extensive training impossible.
  - Treat outliers as belonging to an unknown "latent class" of high-variance subject
  - Impute latent class along with height data
  - Drop "outlier" class before complete-data analysis
  - Can extend "latent variance class" to account for overdispersion in height/weight data
  - Allows for uncertainty in whether or not a subject is an outlier to be carried through the inference.

Accounting for the Complex Sample Design

- Include design variables in mean model
- Consider association between posterior distribution of latent class membership and probability of selection
- Use standard design-based analyses at the complete-data stage of analysis to further enhance robustness.
- Use of MI to compute obesity estimates relies more heavily on the empirical distribution of the data than a fully model-based approach.

"Complete Data" mixture model

$$Z_i \mid C_i = k \sim N_q(\mu_i, \Sigma_k)$$

$$C_i \sim MULTI(1, p_1, \ldots, p_K)$$

where Z_i is a q-dimensional outcome of interest, $\mu_{ij} = x_i^T \beta_j$, $j = 1, \ldots, q$, $\mid \Sigma_1 \mid < \cdots < \mid \Sigma_K \mid$.

- Mean of each subject depends on p covariates x_i, and a covariance given by his or her latent variance class membership given by C_i.
- Class K is the "clerical error" class with the largest variability.
  - Assume that responses with clerical errors have the same mean but larger variability than other responses.

"Complete Data" mixture model: priors

$$p(\beta) \sim N(0, V_\beta)$$

$$p(\Sigma_k) \sim INV - WISHART(2, S_k), \;\; k = 1, \ldots, K$$

$$p(p_1, \ldots, p_K) \sim DIRICHLET(1, \ldots, 1)$$

Missing Data

- C_i are missing for all subjects

- Allow some components of Z_i to be missing under missing at random (MAR) assumption (Rubin 1978): conditional on the observed elements of Z_i , the missingness status of the elements of Z_i is unrelated to their value.

Model Estimation

Gibbs sampler data augmentation algorithm (impute missing elements of Z_i and the completely unobserved C_i at each step of the algorithm).

Multiple Imputation

Take m independent draws of Z^{comp} given by replacing the missing elements of Z with their imputed values, analyze using standard complete data procedures, and combine (Rubin 1987):

$$\hat{Q} = m^{-1} \sum_{t=1}^{m} Q\left(Z^{comp(t)}\right).$$

where

$$V^{1/2}(\hat{Q} - Q) \sim t_\nu$$

Multiple Imputation

for

$$
\begin{aligned}
V &= U + (1 + m^{-1})B \\
U &= m^{-1} \sum_{t=1}^{m} \widehat{\mathrm{Var}} \left(Q \left(\boldsymbol{Z}^{comp(t)} \right) \right) \\
B &= (m-1)^{-1} \sum_{t=1}^{m} (\hat{Q} - Q \left(\boldsymbol{Z}^{comp(t)} \right))^2 \\
\nu &= (m-1) \left[1 + \frac{U}{(1 + m^{-1})B} \right]^2
\end{aligned}
$$

Delete subjects assigned to the Kth latent class when computing $Q(\boldsymbol{Z}^{comp(t)})$.

Application to the Healthy for Life Project

- Probability sample of children aged 2-11 served at one of 141 HRSA-supported Community Health Centers in the eastern United States and Puerto Rico during calendar year 2001.

- Stratified sample of 30 centers, with second-stage sample of approximately 100 children/center stratified by age (2-5 vs. 6-11).

- Inverse probability-of-selection case weights were post-stratified to known age group-region (US mainland urban, suburban, and rural, Puerto Rico (PR) urban and non-urban, and New York City Chinatown) totals.

Application to the Healthy for Life Project

•Dropped 373 cases because of unknown age, gender, or both height and weight information; additional 3 cases dropped because of unknown weight information (to simplify analysis). 2,474 cases remained, of which 606 were missing height data.

•Improve normality approximation via ``z-score'' or Box-Cox transformation (Weiss et al. 2004)

Modeling Healthy for Life Project

- x_i consists of age group-by-center dummy variables, to accommodate within-center correlation systematic association between BMI and the probability of selection.
- Restrict $\rho_k = \frac{\sigma_{12k}}{\sigma_{11k}\sigma_{22k}} \equiv \rho$ for $k = 1, \ldots, K-1$.
- Assume

$$V_\beta = 1000 I_2$$

$$p(\log \sigma_{jjk}) \overset{ind}{\sim} N(0, 4)\ j = 1, 2, k = 1, \ldots, K-1$$

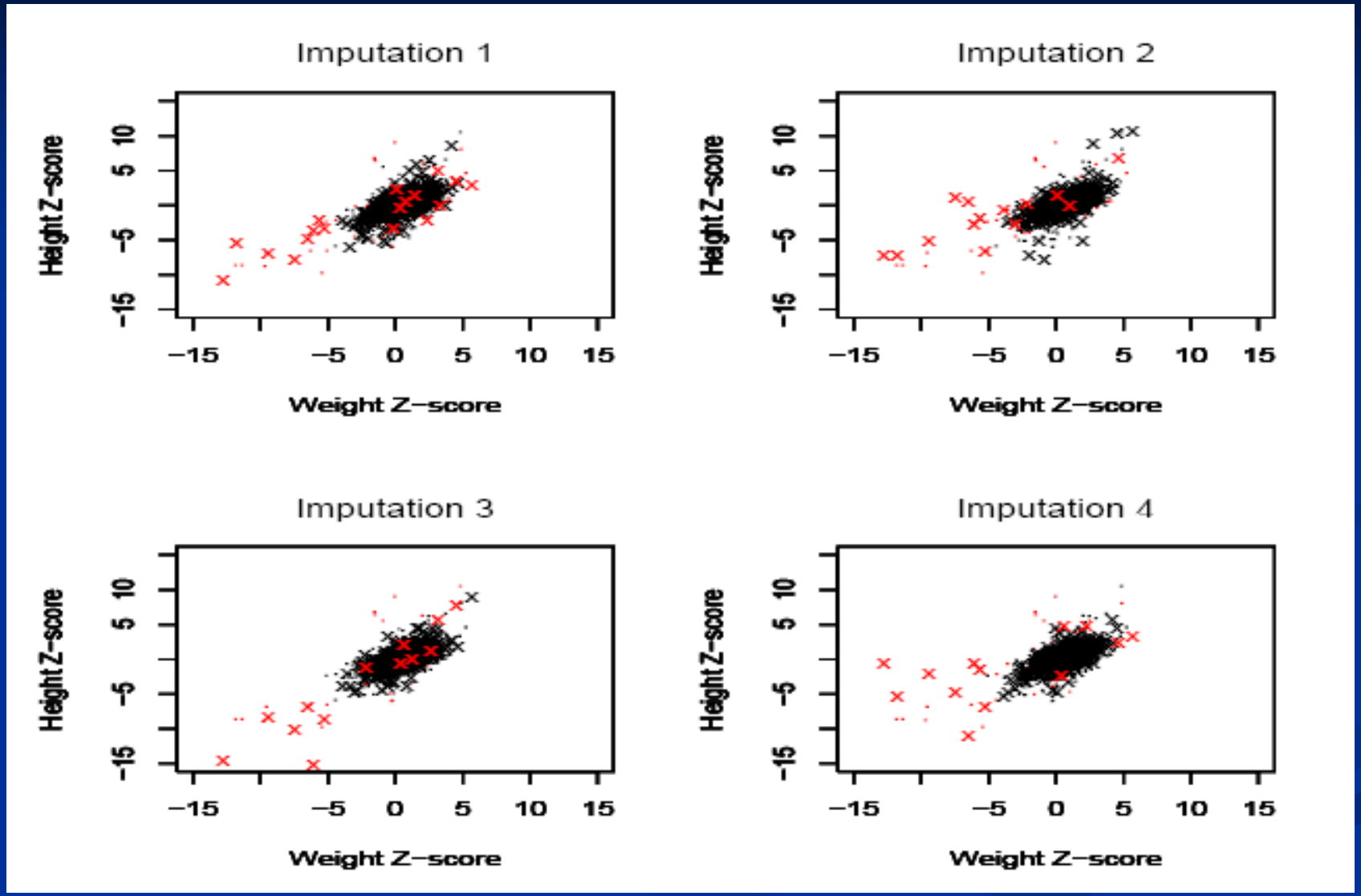
$$p(\rho) \sim U(-1, 1)$$

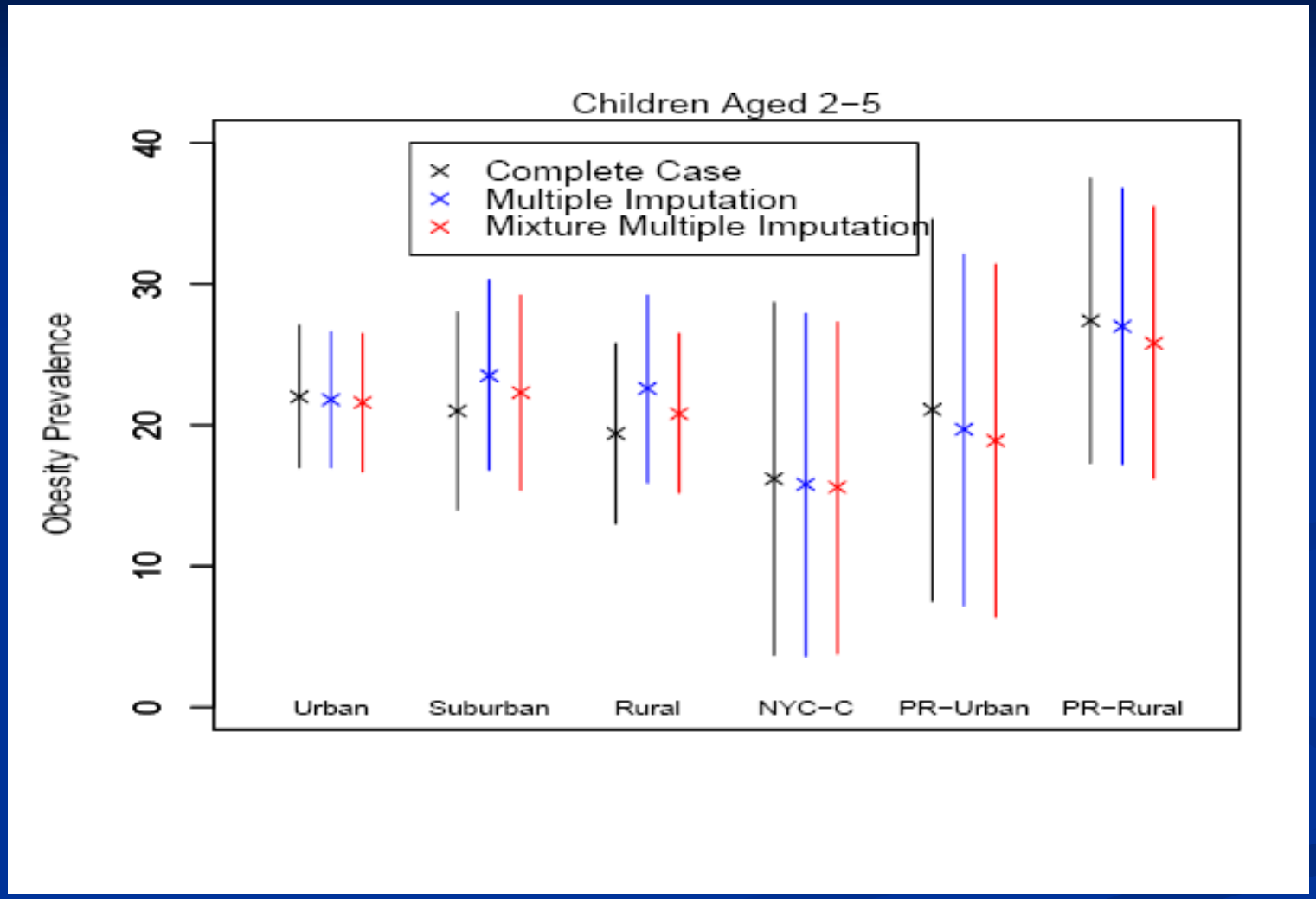
$$S_K = 5 I_2$$

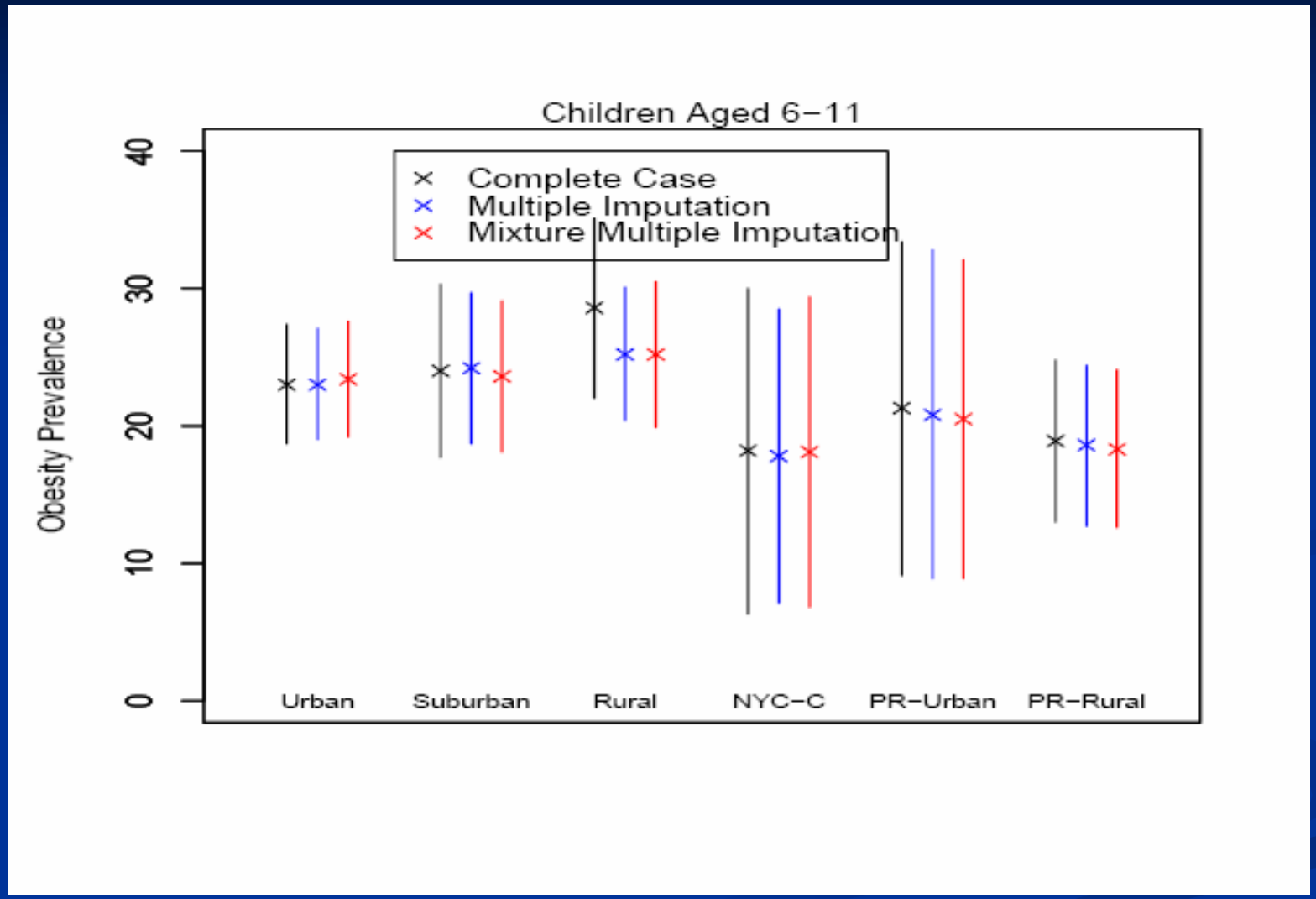
Choosing the number of classes

- Both AIC and BIC suggest that the 3-class model provides the best fit to the data.

	p_k	σ^2_{11k}	σ^2_{22k}	ρ_k
k=1	$.912_{.873,.936}$	$1.43_{1.35,1.55}$	$1.14_{1.04,1.24}$	$.70_{.67,.72}$
k=2	$.072_{.049,.106}$	$3.88_{2.40,6.07}$	$12.34_{7.01,18.83}$	$.70_{.67,.72}$
k=3	$.015_{.007,.029}$	$37.48_{21.14,83.88}$	$29.23_{15.23,64.03}$	$.92_{.63,.98}$


Imputation 1
Imputation 2
Imputation 3
Imputation 4
Height Z−score
Weight Z−score
−15
−5
0
5
10
15


Children Aged 2−5
Complete Case
Multiple Imputation
Mixture Multiple Imputation
Obesity Prevalence
0
10
20
30
40
Urban
Suburban
Rural
NYC−C
PR−Urban
PR−Rural


Children Aged 6−11
Complete Case
Multiple Imputation
Mixture Multiple Imputation
Obesity Prevalence
0
10
20
30
40
Urban
Suburban
Rural
NYC−C
PR−Urban
PR−Rural

Effect of Outliers

- If height data is missing and an older child incorrectly noted as younger, the resulting weight z-score would be extremely large, likely yielding a large BMI after height imputation, and potentially classifying a non-obese child as obese; the reverse is true if a younger child is incorrectly noted as older.
- Since children are more likely than not to be non-obese, the net effect of age transcription errors should be to inflate obesity rates among younger children, and deflate to a much lesser degree obesity rates among older children.
- Analysis of 2.5% and 97.5% quantiles suggested that younger children tended to have large BMI outliers and older children tended to have small BMI outliers, consistent with clerical errors in age.
- Overall impact modest.

Software

- Horton and Lipsitz (2001) provide an overview of currently available software for multiple imputation.
  - SOLAS 3.0 (http://www.statsol.ie/solas/solas.htm)
    - Generally designed for regression models
    - Utilizes both predictive distribution and predictive mean matching to obtain imputations, a "hot-deck"-like MI procedure (Rubin 1987).
    - Earlier versions did not preserve correlation structure.
  - SAS V 9.1
    - PROC MI.
    - Easy to combine results.
    - Little control over model; generally requires MVN assumption.
  - Joe Schafer's free software for multiple imputation (http://www.stat.psu.edu/~jls/misoftwa.html#top)
    - Multivariate normal
    - Categorical
    - Mixed normal and categorical (general location)
    - Clustered multivariate normal.

References

- Barnard, J and Meng, X-L (1999) Applications of multiple imputation in medical studies: From AIDS to NHANES, *Statistical Methods in Medical Research*, 8, 17-36
- Dempster, AP, Laird, NM and Rubin, DB (1977) Maximum likelihood from incomplete data via the EM algorithm *Journal of the Royal Statistical Society, Series B: Methodological*, 39, 1-22
- Elliott, M.R., Stettler, N. (2006) Using a mixture model for multiple imputation in the presence of outliers: the Healthy for Life project. *Applied Statistics*, 56, 63-78.
- Horton, NJ and Lipsitz, SR (2001) Multiple imputation in practice: Comparison of software packages for regression models with missing variables T*he American Statistician*, 55, 244-254
- Little, RJA (1988) Missing-data adjustments in large surveys, *Journal of Business & Economic Statistics*, 6, 287-296
- Little, RJA, Rubin, DB (2002) *Statistical Analysis with Missing Data*, Wiley, Hoboken, New Jersey

References

- Meng, X-L (1994) Multiple-imputation inferences with uncongenial sources of input *Statistical Science*, 9, 538-558
- Raghunathan, TE and Siscovick, DS (1996) A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives A*pplied Statistics*, 45, 335-352
- Raghunathan, TE, Lepkowski, JM, van Hoewyk, J and Solenberger, P (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models S*urvey Methodology*, 27, 85-95

References

- Rubin, D (1987) *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons (New York; Chichester)
- Rubin, D, Schenker, N (1986) Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse *Journal of the American Statistical Association*, 81, 366-374.
- Schafer, JL (1997) Analysis of incomplete multivariate data *Chapman & Hall Ltd (London; New York)*
- Stettler, N., Elliott, M. R., Kallan, M. Auerbach, S. B. and Kumanyika, S. K. (2005) High prevalence of pediatric overweight in medically underserved areas. *Pediatrics*, **116**, 381–388.