Missing Data in Longitudinal Studies: A Review

Joe Schafer
Department of Statistics and
The Methodology Center
The Pennsylvania State University

Presented on November 9, 2005 at AAPS, Nashville. Supported by National Institute on Drug Abuse, P50-DA-10075.

Outline

1. Motivation
2. Basic theory
3. Efficient procedures
   - Linear mixed models
   - Generalized estimating equations (GEE)
   - Weighted estimating equations (WEE)
   - Multiple imputation (MI)
4. Modeling nonignorable missingness
   - 2×2 table
   - Random coefficient pattern-mixture model

Based on articles by Collins, Schafer and Kam (2001), Schafer and Graham (2002), Schafer (2003), Demirtas and Schafer (2003), Schafer and Kang (2005)

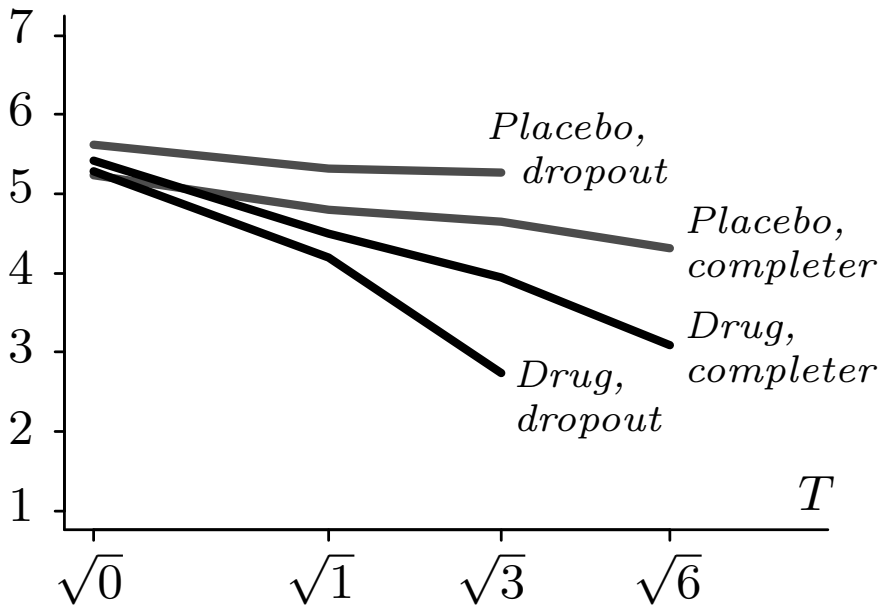
1. Motivation

Data example from Hedeker and Gibbons (1997)

A randomized psychiatric trial

- 312 patients received drug therapy for schizophrenia; 101 received placebo
- measurements at weeks 0, 1, 3, 6
- missing data primarily due to dropout
- outcome: severity of illness (1=normal, . . . , 7=extremely ill); treat as continuous

Plot of average response versus square root of week



A completers-only analysis would severely understate the treatment effect. We want a sensible procedure to analyze the incomplete data

- low bias
- high efficiency
- robust to assumptions about from population distribution and missing-data mechanisms

Resources on missing data in general

- Little and Rubin (2002) *Statistical Analysis with Missing Data, Second edition.* (New York: Wiley)
- Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data* (London: Chapman & Hall)
- Allison, P.D. (2001) *Missing Data* (Thousand Oaks: Sage)
- Schafer, J.L. and Graham, J.W. (2002) Missing data: our view of the state of the art. *Psychological Methods*

Resources on missing data in longitudinal studies

- Little, R.J. (1995) Modeling the dropout mechanism in repeated-measures studies. *JASA*
- Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data* (New York: Springer).
- 100+ articles in stat and biostat journals in last ten years

Mean imputation

$$\begin{aligned} y_{ij} &= \text{response for subject } i \text{ at occasion } j \\ r_{ij} &= 1 \text{ if } y_{ij} \text{ observed, 0 if missing} \end{aligned}$$

If y_{ij} is missing, we can replace it by

- the mean response for subject i

$$y_{i\cdot} = \frac{\sum_j r_{ij} y_{ij}}{\sum_j r_{ij}}$$

- the mean response for occasion j

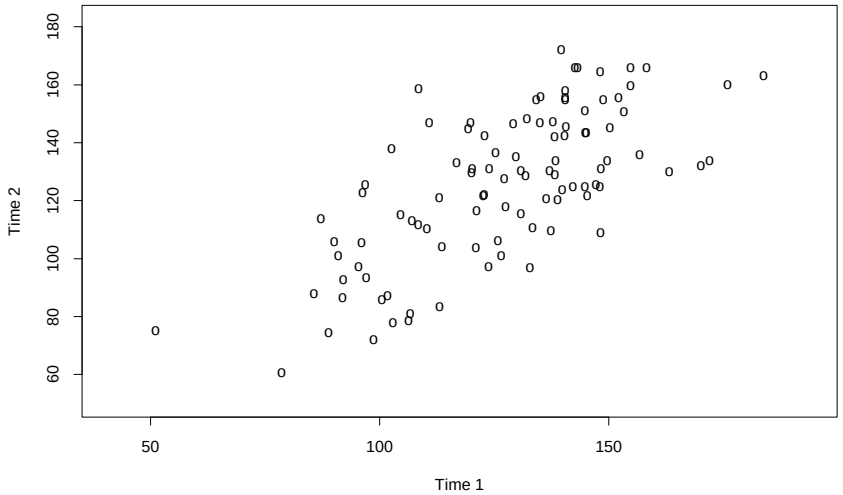
$$y_{\cdot j} = \frac{\sum_i r_{ij} y_{ij}}{\sum_i r_{ij}}$$

Both of these methods may seriously distort estimates and measures of uncertainty

Hypothetical example: systolic blood pressure measurements at two occasions

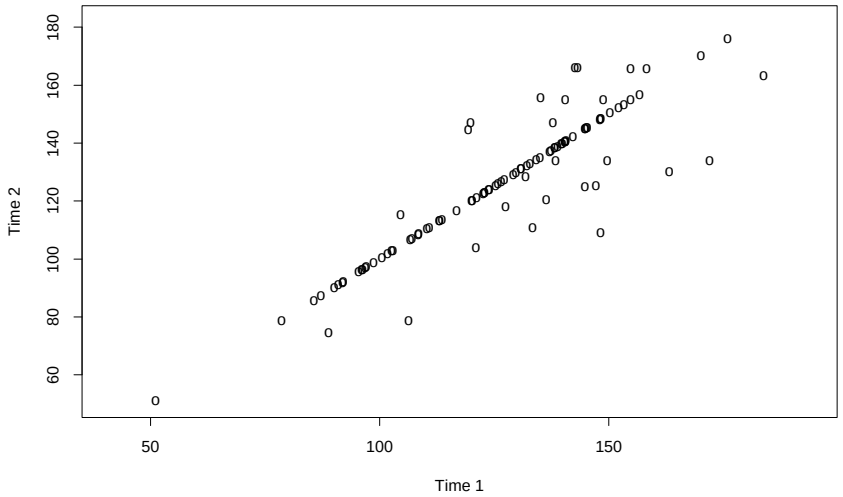
- Suppose $y_i = (y_{i1}, y_{i2})$ is bivariate normal with means $\mu_1 = \mu_2 = 125$, variances $\sigma_1 = \sigma_2 = 25$, $\rho = 0.6$ ($\beta_{2|1} = \beta_{1|2} = 0.6$)
- Sample $n = 100$ individuals; make y_{i2} missing with probability .5 independently of (y_{i1}, y_{i2}) (missing completely at random)
- Impute subject-means or occasion-means

Complete data



$$\bar{y}_2 = 126, \quad \hat{\sigma}_2 = 25.7, \quad r = .70, \quad \beta_{2|1} = .77, \quad \beta_{1|2} = .64$$

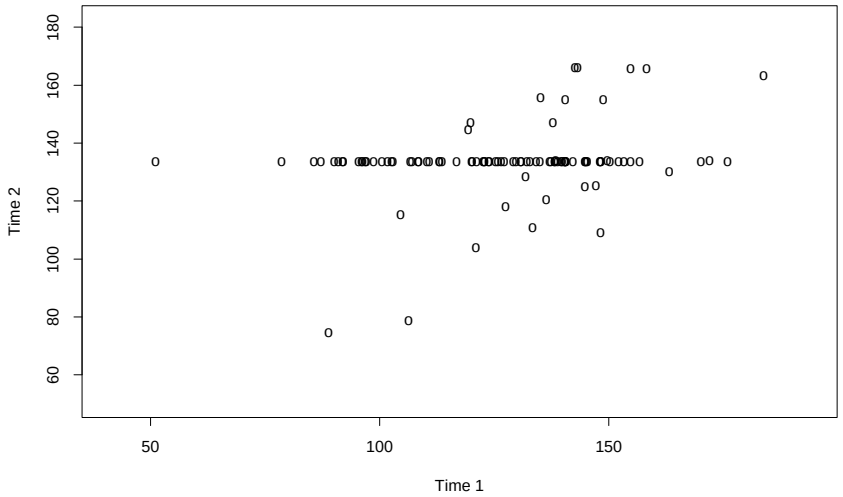
After imputing the subject mean



$$\bar{y}_2 = 125.8, \quad \hat{\sigma}_2 = 24.0, \quad r = 0.90$$

$$\beta_{2|1} = .93, \quad \beta_{1|2} = .88$$

After imputing the occasion mean



$$\bar{y}_2 = 133, \quad \hat{\sigma}_2 = 12.9, \quad r = 0.28$$

$$\beta_{2|1} = .15, \quad \beta_{1|2} = .50$$

Last observation carried forward

For attrition (dropout): If a subject drops out after occasion j,

$$\text{replace } y_{i,j+1}, y_{i,j+2}, \ldots \text{ by } y_{i,j}$$

- Equivalent to subject-mean imputation for dropout after first occasion
- Tends to understate differences in estimated time-trends between treatment and control groups (thought to be "conservative")
- Not necessarily "conservative," because standard errors are biased downward as well
- Especially bad for outcomes that have high variation within a subject

Comments on imputation in general

- Single-imputation strategies designed to precisely predict the missing values tend to distort estimates of population quantities
- The goal of the missing-data procedure is to draw accurate inferences about population quantities (e.g. mean change over time), not to accurately predict the missing values
- With imputation, the best way to achieve that goal is to preserve all aspects of the data distribution (means, trends, within- and between-subject variation, etc.)
- Ad hoc imputation methods inevitably preserve some aspects but distort others

Case-deletion methods

Often used in the past to produce balanced datasets for repeated-measures ANOVA

- Delete any subject with a missing value at any occasion
- Perhaps delete some complete subjects as well to balance the n's across treatment groups

Modern methods for analyzing longitudinal data (e.g. PROC MIXED) do not require balance, so case-deletion procedures have become less popular

A few comments on case deletion

- Not so bad for laboratory experiments, for which data are often nearly balanced
- In studies with human subjects (especially over longer periods of time), missed measurements and dropout are a more serious issue
- When completers and dropouts seem to follow different trajectories, analyzing only the completers may be very misleading
- For population inferences, it's nearly always better to analyze the data from all subjects whether they completed the study or not
  - less biased
  - more efficient

2. Basic theory

Basic notation

$$
\begin{aligned}
x_i &= \text{covariates for subject } i \\
&\quad \text{(assume completely observed)} \\
y_i &= \text{outcomes for subject } i \text{ at all occasions} \\
&\quad \text{(could be a vector or a matrix)}
\end{aligned}
$$

The data model

$$
\begin{aligned}
P(y_i|x_i, \theta) &= \text{some distribution} \\
\theta &= \text{population parameters of interest}
\end{aligned}
$$

For example, θ could be

- effects of covariates on response
- difference in mean response at final occasion

Notice that θ applies to the *entire population* of subjects

The missingness

$$r_i \quad = \quad \text{binary variables indicating whether each element of } y_i \text{ is observed or missing}$$

- In general, r_i is a matrix of 0's and 1's of the same size as y_i
- In special cases it can be reduced to a smaller set of variables
- If the only kind of missing data is dropout, then it can be reduced to a single number (time of last measurement)

The distribution of missingness (DOM)

$$P(r_i|x_i, y_i, \phi) = \text{some distribution}$$

First introduced by Rubin (1976, *Biometrika*); sometimes called the "missingness mechanism"

Why introduce the DOM?

- Not because we want to model it
- We typically do not have the information necessary to model it well
- Rubin's purpose in introducing the DOM was **to clarify the conditions under which it may be ignored**
- The conditions under which we may ignore the DOM vary depending on the mode of inference for θ (frequentist, likelihood, Bayesian)

Classification of DOM's

Based on Rubin (1976), Little and Rubin (1987), and Little (1995)

- **Missing completely at random (MCAR):** DOM does not depend on covariates or outcomes

$$P(r_i|x_i, y_i, \phi) = P(r_i|\phi)$$

- **Covariate-dependent (CD) missingness:** DOM may possibly depend on covariates but not outcomes

$$P(r_i|x_i, y_i, \phi) = P(r_i|x_i, \phi)$$

- **Missing at random (MAR):** DOM may depend on covariates and observed outcomes

$$P(r_i|x_i, y_i, \phi) = P(r_i|x_i, y_{i(obs)}, \phi)$$

  Note that MCAR $\subset$ CD $\subset$ MAR.

- **Missing not at random (MNAR):** Any violation of MAR; DOM still depends on $y_{i(mis)}$ even after any dependence on x_i and $y_{i(obs)}$ has been accounted for

Explanation

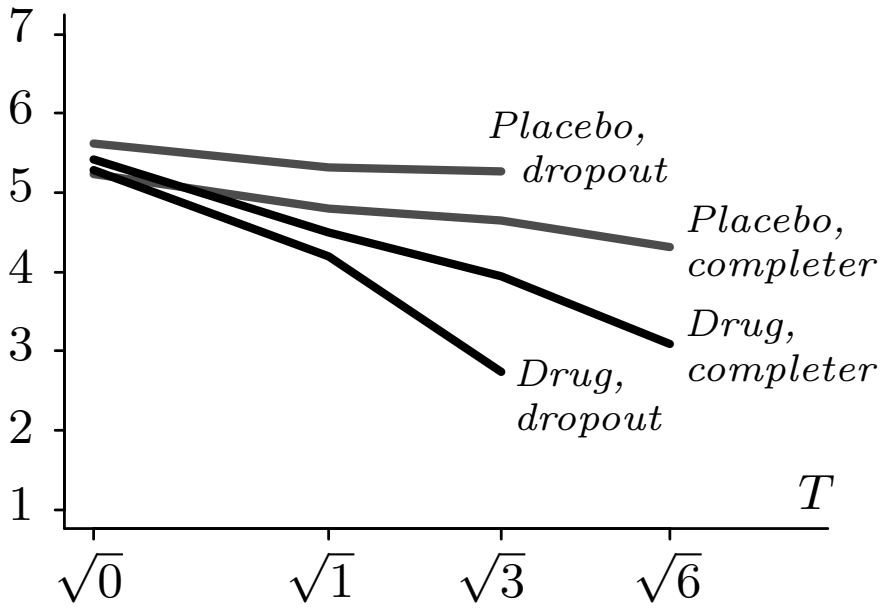
In the case of dropout,

- MCAR means that the probability of dropout is unrelated to any characteristics of the subject at all
- CD means that the probability of dropout may be related to covariates but is unrelated to outcomes at any time
- MAR means that the probability of dropout may be related to covariates and to *pre-dropout responses*
- MNAR means that probability of dropout is related to responses at the time of dropout and possibly afterward (the latter is often not unreasonable; see Little, 1995)

What can we tell from the data?

- Because we observe x_i, r_i, and $y_{i(obs)}$, it is often possible to reject MCAR and CD in favor of MAR
- It is never possible to reject MAR in favor of MNAR on the basis of the observed data, because we cannot see $y_{i(mis)}$
- Any procedure for "testing MAR" is making some other assumptions that are unverifiable

Example



Based on this plot, we may conclude:

- dropout is not MCAR, because it operates differently in the treatment and control groups
- dropout is not merely CD, because completers and dropouts follow different (pre-dropout) trajectories
- dropout could be MAR or MNAR; it's impossible to tell

Implications of MCAR, MAR

When may we ignore the DOM and not model it?

1. For **frequentist** statistical procedures, we may generally ignore the DOM only when the missing data are **MCAR** (e.g. GEE regression using PROC GENMOD)
2. For **likelihood/Bayes** procedures, we may generally ignore the DOM when the missing data are **MAR** (e.g. linear mixed models with PROC MIXED)

Because of result 2, we have the terminology

$$\begin{aligned} \text{MAR} &\iff \text{“ignorable”} \\ \text{MNAR} &\iff \text{“nonignorable”} \end{aligned}$$

but this is appropriate only for likelihood/Bayes procedures; see Rubin (1976) and Kenward and Molenberghs (1998)

When do we have to model the missingness?

If we are doing frequentist analyses, we will have to model r_i if the missingness is not MCAR

- Example: weighted estimating equations (Robins, Rotnitzky & Zhao, 1995)

If we are doing likelihood or Bayesian analysis, we will have to model r_i only if we believe that missingness is MNAR

- Examples: selection models; pattern-mixture models
- These can be difficult and very sensitive to misspecification
- I believe strong theory is needed for these to work well

Missing values that are out of scope

There is one additional situation when it is appropriate to not model the DOM: when the fact that an observation is missing causes it to leave the universe of interest

Example

Consider a study involving elderly persons trial where the outcomes in y_i are quality-of-life measures, and patients "drop out" because of death. In this case, we can regard death as ignorable.

- In reality, there are no missing data
- We may use missing-data procedures based on ML (e.g. PROC MIXED), understanding that we are estimating aspects of quality-of-life for live patients only
- population trajectory estimates the average QOL *for those who are alive at any given time*

For an interseting perspective on causal inference in the presence of death, see Frangakis and Rubin (2002), Zhang and Rubin (2003)

3. Efficient procedures

A. Linear mixed models

- Also known as multilevel models, linear mixed-effects models random-effects models, random-coefficient models, hierarchical linear models
- Implemented in HLM, PROC MIXED, S-PLUS, R, Stata, ...

Adopting the notation of Laird and Ware (1982), the model is

$$y_i = X_i\beta + Z_i b_i + \epsilon_i, \quad i = 1, \ldots, m$$

where

$$\begin{aligned} y_i &= (y_{i1}, y_{i2}, \ldots, y_{i,n_i})^T \\ b_i &\sim N_q(0, \psi) \\ \epsilon_i &\sim N_{n_i}(0, \sigma^2 V_i) \end{aligned} \tag{1}$$

$$
\begin{aligned}
\beta &= \text{fixed effects} \\
b_i &= \text{random effects for unit } i \\
\psi &= \text{between-unit covariance matrix} \\
\sigma^2 V_i &= \text{within-unit covariance matrix}
\end{aligned}
$$

- Handles unequal n_i's, time-varying covariates, unequally spaced responses
- Often we use $V_i = I$, but other structures—e.g., autoregressive—are useful, especially when n_i's are large
- measurement times are often incorporated into X_i, Z_i as polynomials
- Z_i contains a subset of the columns of X_i

An excellent treatment these models is the new book by Fitzmaurice, Laird and Ware (2004)

What about missing data?

1. When data are unbalanced by design, then ML or REML estimation is the right thing to do
2. If some responses for some subjects are missing, we may omit the missed occasions and apply ML or REML to the reduced data; this is appropriate if the missing responses are MAR
3. Note that important correlates of missingness need to be included in the model for MAR to be plausible

Notes about PROC MIXED

(This is not necessarily limited to PROC MIXED; other programs may behave in a similar fashion)

- PROC MIXED will automatically omit occasions with missing responses (which is good under MAR)
- PROC MIXED will also omit subjects or occasions with missing **covariates**, which implicitly assumes that these are MCAR (not so good)

B. Semiparametric regression using GEE

First introduced by Liang and Zeger (1986); see also Diggle, Liang and Zeger, (1994). Instead of attempting to model the within-subject covariance structure, treat it as a nuisance and simply model the mean response.

$$
\begin{aligned}
y_i &= (y_{i1}, y_{i2}, \ldots, y_{i,n_i})^T \\
y_{ij} &= \text{discrete or continuous response} \\
E(y_{ij}) &= \mu_{ij};\ \text{mean response} \\
g(\mu_i) &= X_i\beta \quad \text{link function} \\
\text{Cov}(y_i) &= \Delta_i^{1/2} R_i(\alpha) \Delta_i^{1/2}
\end{aligned}
$$

where R_i is a 'working correlation matrix' representing a guess at the true correlation structure.

Popular choices for R_i include

- identity (IEE method)
- exchangeable (compound symmetry)
- autoregressive
- unstructured (if subjects are measured at a small set of common time points)

Implemented in PROC GENMOD, Stata and elsewhere

Sandwich estimator

Provides a good estimate of $\text{Cov}(\hat{\beta})$ in large samples regardless of the true form of $\text{Cov}(y_i)$ (Huber, 1967; White, 1982; Liang & Zeger, 1986)

$$\left[X^T \hat{W} X\right]^{-1} \left[\sum_i X_i^T (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)^T X_i\right] \left[X^T \hat{W} X\right]^{-1}$$

What about missing data?

When elements of y_i are missing, we can omit the missed occasions. Liang and Zeger (1986) noted that

- if the working assumptions are wrong, GEE and sandwich estimators are consistent under MCAR
- if the working assumption is correct, the GEE estimator is consistent under MAR (but the sandwich estimator may not be—see Kenward and Molenberghs, 1998)

Comments

- The GEE and sandwich estimators are not motivated by likelihood or Bayesian principles; they are frequentist procedures, so in general they require MCAR to ignore the DOM
- These methods attempt to 'robustify' by relaxing assumptions on the data model, but there is no free lunch. By relaxing assumptions on the data model, they must impose stronger assumptions on the DOM

C. Weighted estimating equations

First proposed by Robins, Rotnitzky and Zhao (1994, 1995)

When the y_i's are observed with inverse-probabilities $w_i = \pi_i^{-1}$, we can remove bias in estimating β by solving weighted estimating equations

In WEE, we can throw away any subset of subject-occasions that is difficult to use because of missing responses and/or covariates, and reweight the rest to make them more representative.

- Robins, Rotnitzky and Zhao (1994) discard subject-observations with missing covariates
- Robins, Rotnitzky and Zhao (1995) discard subject-observations with missing responses
- Rotnitzky and Robins (1997); Rotnitzky, Robins and Scharfstein (1998); and Scharfstein, Rotnitzky and Robins (1999) discard various sets of subject-occasions for which covariates and/or responses are missing

Many papers, basically the same idea

Estimating the weights

Let $r_{ij} = 0$ if we discard the jth occasion for subject i for any reason, and $r_{ij} = 1$ if we keep it

We must develop a model to estimate our propensity to discard,

$$w_{ij}^{-1} = \pi_{ij} = P(r_{ij} = 1|\text{something}),$$

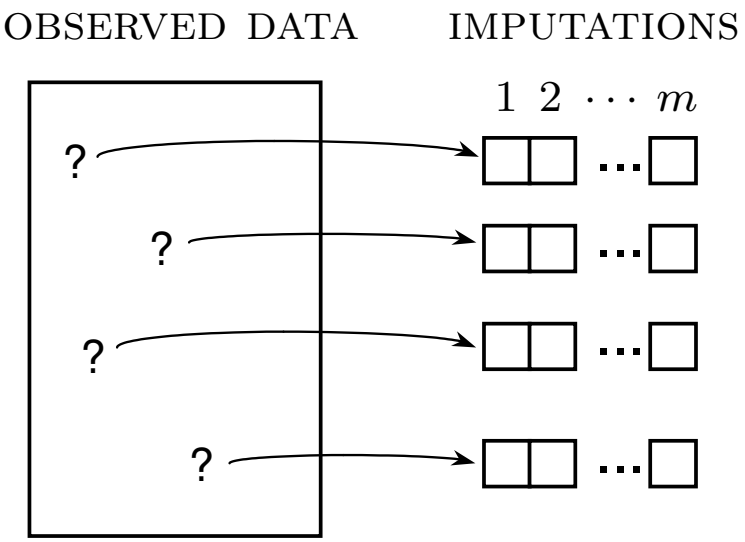
where "something" is some information we have that seems related to missingness: static covariates, time-varying covariates, baseline measures, pre-dropout responses, etc.

- If "something" is observed for every subject-occasion, then this is a straightforward logistic regression

WEE by itself can be inefficient. Improved estimators, called "doubly robust," increase the efficiency by incorporating additional information to predict the missing responses (Scharfstein, Rotnitzky & Robins, 1999; Davidian, Tsiatis & Leon, 2005)

D. Multiple imputation

A simulation-based approach to missing data



- Generate $m > 1$ plausible versions of Y_{mis}
- Analyze each of the m datasets by standard **complete-data** methods
- Combine the results

Rubin (1987) calls this the *repeated-imputation* inference method

Features of MI

- works with standard complete-data analysis methods
- one set of imputations may be used for many analyses
- can be highly efficient, even for very small m
  - The efficiency of an estimator based on m imputations is $(1 + \gamma/m)^{-1}$, where γ is the **rate of missing information**

Efficiency of multiple imputation (%)

	γ				
m	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

Rules for combining results

After obtaining m imputations of Y_{mis}, analyze the m completed datasets and combine the results

Rubin's (1987) rule for scalar estimands

$$\begin{aligned}
\hat{Q} &= \text{complete-data point estimate} \\
U &= \text{complete-data variance estimate} \\
\bar{Q} &= m^{-1}\textstyle\sum_{t=1}^{m}\hat{Q}^{(t)} \\
B &= (m-1)^{-1}\textstyle\sum_{t=1}^{m}(\hat{Q}^{(t)}-\bar{Q})^2 \\
\bar{U} &= m^{-1}\textstyle\sum_{t=1}^{m}U^{(t)} \\
T &= \bar{U}+(1+m^{-1})B
\end{aligned}$$

Interval estimate is $\bar{Q} \pm t_\nu \sqrt{T}$ where

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

Creating the imputations

Imputations should be drawn from a predictive distribution $P(Y_{mis} \mid Y_{obs})$ under an appropriate model (may need MCMC)

- Intuitively, we are averaging the complete-data posterior over the predictive distribution of the missing data

$$P(Q \mid Y_{obs}) = \int P(Q \mid Y)\, P(Y_{mis} \mid Y_{obs})\, dY_{mis}$$

- MI does not require the imputer and analyst to use the same model. It works
  - when the analyst's model is **congenial** to (i.e. can be embedded within) the imputer's model, and
  - in many **uncongenial** settings, e.g. where the analysis is not fully parametric (Meng, 1994; Rubin, 1996)

MI references

- Rubin, D.B. (1987) *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Meng, X.L. (1994) Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 10, 538–573.
- Rubin, D.B. (1996) Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, 91, 473–489.
- Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall
- Schafer, J.L. (1999) Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8, 3–15.

Software for imputation under a multivariate normal model

NORM (Schafer 1999): Free Windows program. Uses data augmentation.

Amelia (Gary King et al., 2001): Free program; also available as macros for GAUSS. Uses importance resampling rather than DA.

PROC MI (SAS version 8.2): New 'experimental' SAS procedure equivalent to NORM

S+ MissingData (S-PLUS Version 6): MI for Gaussian model; equivalent to NORM

Software for imputation under other models

CAT, MIX: Schafer's old S-PLUS functions for categorical data, mixed continuous and categorical data; now obsolete

PAN: Schafer's program for longitudinal data; under development

S+ MissingData (S-PLUS Version 6): MI for loglinear and conditional Gaussian models; replaces CAT and MIX

SOLAS (Statistical Solutions, Inc.): Multiple imputation by two methods

- Propensity-score method with approximate Bayesian bootstrap (Lavori, Dawson and Shera, 1995). This can be dangerous!
- Model-based method using a sequence of regression models. This is essentially equivalent to normal-based imputation in NORM, but it requires missingness pattern to be monotone.

MICE: S-PLUS functions for approximate MI using "chained equations". Available from `http:\\www.multiple-imputation.com`

When do ML and MI agree?

When

- the user of ML and the imputer use the same input data, but
- the imputation model is **equivalent to** or **more general than** the analysis model,

then ML and ML yield **similar results**

- parameter estimates and SE's tend to be similar
- both are valid, assuming the model is correct

Example

Missing data in longitudinal study

Researcher A: Imputes m times with NORM (treating the responses at different occasions as different variables), fits the random-coefficients model m times to the imputed data, combines results

Researcher B: Fits model directly to incomplete data with PROC MIXED

The estimated coefficients will be similar, but A's standard errors may be slightly larger

- the imputation model assumes normality $y_i \sim N(X_i\beta, \Sigma)$
- the analysis model assumes normality with a patterned mean and covariance structure which is a special case of $N(\mu, \Sigma)$

When do ML and MI disagree?

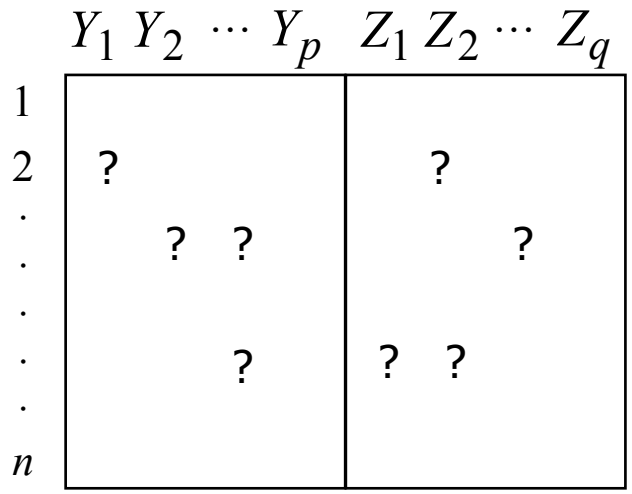
When the user of ML and the imputer use

- the same set of units
- a different set of variables

then the results from ML and MI could differ

Auxiliary variables

Suppose that we have variables of direct interest $Y_1, Y_2, \ldots, Y_p$ and additional variables $Z_1, Z_2, \ldots, Z_q$ that we might possibly want to include in the imputation model



- When is it beneficial to include the Z's?
- When may it be harmful?

Simulation study

Collins, Schafer and Kam (2001) investigated the effects of including three kinds of variables in the imputation model

Type A: correlated with outcomes and with missingness

Type B: correlated with outcomes only

Type C: not correlated with outcomes or missingness

General conclusions

- Omitting a correlate of missingness is usually not serious unless it is highly correlated with the outcome and the rate of missing values is high
- Including correlates of missingness in the imputation model can be **somewhat** helpful
- Including correlates of outcomes in the imputation model can be **very** helpful
- There is little danger in including too many variables
- An inclusive strategy is nearly always better than a restrictive one

4. Modeling Noningnorable Missingness

Sometimes it is not reasonable to ignore DOM (e.g. clinical trial where dropout is strongly related to outcome)

Selection models

$$P(Y_{com}, R \mid \theta, \phi) \;=\; P(Y_{com} \mid \theta)\, P(R \mid Y_{com}, \phi)$$

(e.g., Diggle and Kenward, 1994)

Pattern-mixture models

$$P(Y_{com}, R \mid \eta, \nu) \;=\; P(R \mid \eta)\, P(Y_{com} \mid R, \nu)$$

(Little, 1993). Note that parameters of interest are in

$$P(Y_{com} \mid \theta) \;=\; \sum_{R} P(R \mid \eta)\, P(Y_{com} \mid R, \nu)$$

Both require unverifiable assumptions!

MI under MNAR models

Current MI software was intended for use under MAR. But it can also produce imputations under a variety of MNAR models

- Include summaries of R in the model as "covariates"
- Becomes a pattern-mixture model
- Some aspects of the multivariate distribution might not be estimable, so we may need to
  - omit certain relationships
  - apply an informative prior distribution
- Many have already done this without realizing it
- Including summaries of R doesn't make the model more general; it could make it less general

EXAMPLE

Victimization status from the National Crime Survey (Schafer, 1997)

Victimized first period?	*Victimized second period?*		
	No	Yes	Missing
No	392	55	33
Yes	76	38	9
Missing	31	7	115

Source: Kadane (1985)

$$
\begin{aligned}
Y_1 &= \text{victim status, time 1} \\
Y_2 &= \text{victim status, time 2} \\
R_1 &= \text{missingness, time 1} \\
R_2 &= \text{missingness, time 2}
\end{aligned}
$$

Impute under loglinear model $(Y_1Y_2, R_1R_2, Y_1R_1, Y_2R_2)$ using loglinear routines in S+MissingData library (Schafer, 2003)

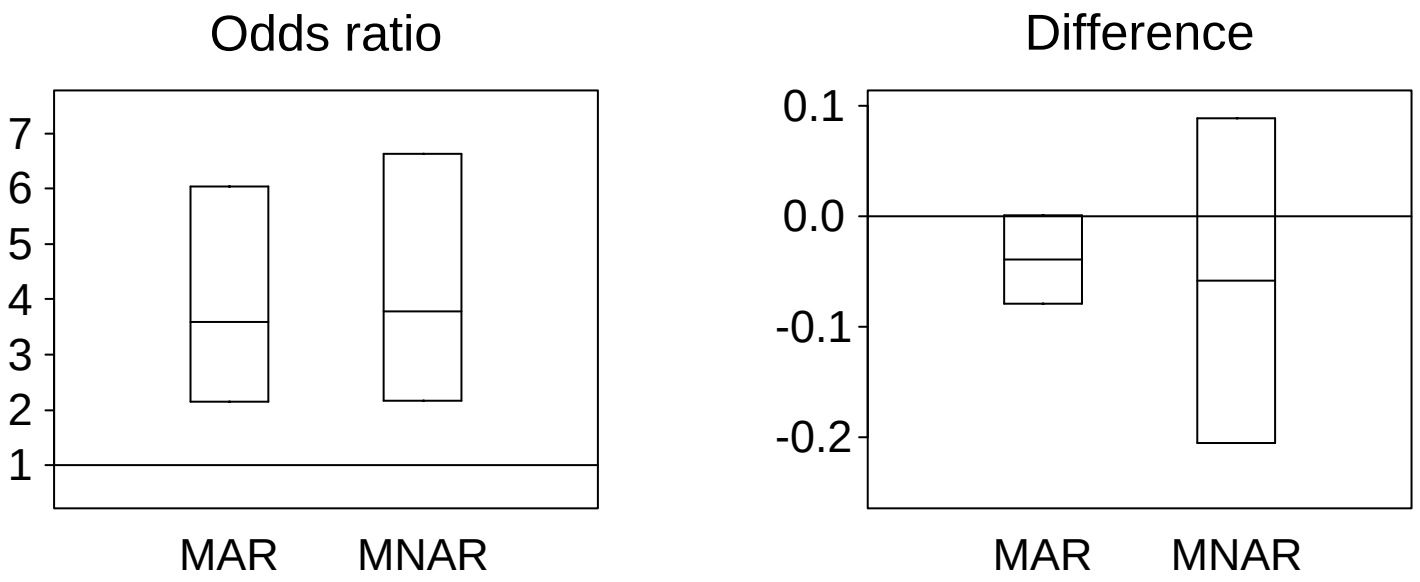
- Generate 10 imputations using Y_1, Y_2, R_1, R_2
- Analyze Y's, throw away R's

$$\pi_{ij} \;=\; P(Y_1 = i, Y_2 = j)$$

Produce estimates and intervals for

$$\begin{aligned} \alpha &= \left(\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}\right) \\ \delta &= \pi_{12} - \pi_{21} \end{aligned}$$

Point and 95% interval estimates for odds ratio and difference under MAR and MNAR imputation

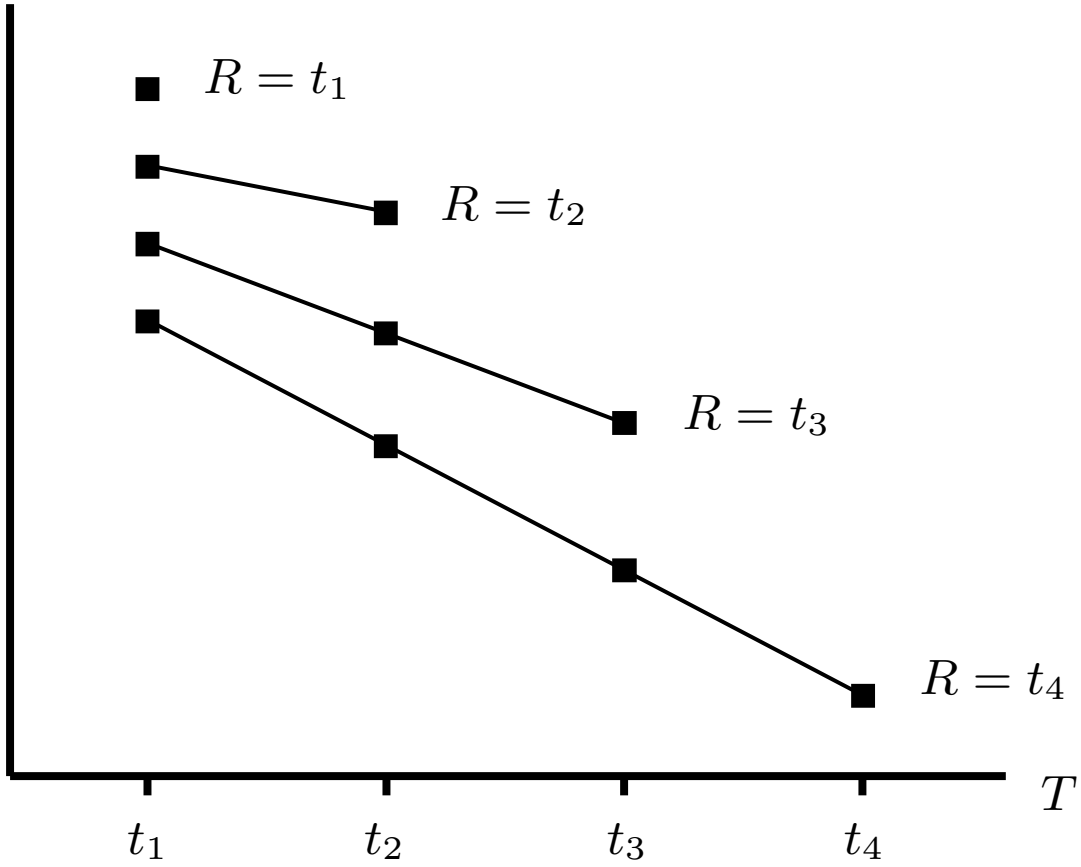


- Data contain almost no information about Y_1R_1 and Y_2R_2 associations
- Omitting Y_1R_1 and Y_2R_2 produces valid inferences for Y's under any DOM that is MAR
- Including Y_1R_1 and Y_2R_2 produces valid inferences only under this particular DOM
- Including these terms does not make the model "more general"
- Should we apply MNAR models simply because we can?
- Unless you have a particular theory telling you
  - that Y_1R_1 and Y_2R_2 do exist, and
  - what the likely size and direction of these associations are,

  it's probably better to omit them

Example: MI with pattern-mixture models

Longitudinal study with occasions $T = t_1, t_2, \ldots, t_k$; let R be the time of subject's last measurement



Suggests fitting a model with effects for R, T, and $R \times T$

Like a repeated-measures experiment with two crossed factors

R = between-subject factor

T = within-subject factor

	T					
R	t_1	t_2	t_3	$\cdots$	t_{k-1}	t_k
t_1	$\times$	—	—		—	—
t_2	$\times$	$\times$	—		—	—
t_3	$\times$	$\times$	$\times$		—	—
$\vdots$						
t_{k-1}	$\times$	$\times$	$\times$		$\times$	—
t_k	$\times$	$\times$	$\times$		$\times$	$\times$

$\times$ = responses observed

— = responses not observed

Predictions for any cell where $T > R$ are possible only by **extrapolation**, which requires strong modeling assumptions

Extrapolation strategies

Reviewed by Demirtas and Schafer (2003)

- **Polynomial surfaces:** Introduce polynomials $R^a T^b$ for $a = 0, 1, 2 \ldots$ and $b = 0, 1, 2, \ldots$
- **Coarse grouping by pattern:** Group together subjects with similar R-values (e.g. completers versus dropouts)
- **Pattern-specific extrapolation:** Extrapolate for each R-group using a mean response function estimated from that group alone
- **Polynomial-coefficient restrictions:** Borrow estimates of polynomial coefficients from
  - complete cases (CCPC)
  - available cases (ACPC)
  - neighboring cases (NCPC)

Different extrapolation strategies may have identical fit to the observed data

Application to schizophrenia data

			Treatment effect	
Method	Fixed effects	$2\times$loglik	est.	std.err.
Ignorable	4	-1595.6	-0.65	0.08
HedGib	8	-1569.5	-0.73	0.09
Rlin$\times$Tlin	8	-1573.8	-0.75	0.09
Full-Poly	19	-1548.9	-1.30	0.33
Red-Poly	16	-1549.1	-1.06	0.20
Patt-Spec	19	-1548.9	-0.78	0.12
CCPC	19	-1548.9	-0.99	0.15
NCPC	19	-1548.9	-1.22	0.36
ACPC	19	-1548.9	-0.95	0.17

The last six models give (nearly) the same fit to the observed data, but the range of the estimated treatment effects is about 6 times the SE's reported by Hedeker and Gibbons (1997)!

Concluding remarks

- MI with auxiliary variables is a good idea
- Assuming MAR seems okay in many circumstances
- MNAR models may be useful, but strong theory may be required
- If MNAR is anticipated, we should collect better information on reasons for/causes of missingness
- The ability to impute under a variety of models and analyze the imputed data under other models is an inherent strength of MI (Schafer, 2003)

A copy of this talk will be posted at

`http://www.stat.psu.edu/~jls/`

under "Presentations"

Allison, P.D. (2001) *Missing Data.* Thousand Oaks, CA: Sage.

Collins, L. M., Schafer, J. L. and Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods*, 6, 330–351.

Demirtas, H. and Schafer, J.L. (2003) On the performance of random-coefficient pattern-mixture models for non-ignorable dropout. *Statistics in Medicine*, 22, 2553–2575.

Davidian, M., Tsiatis, A.A. and Leon, S. (in press) Semiparametric estimation of treatment effect in a pretest-posttest study with missing data. *Statistical Science.*

Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994) *Analysis of longitudinal data.* Oxford: Clarendom Press.

Fitzmaurice, G., Laird, N.M. and Ware, J.H. (2004) *Applied Longitudinal Analysis.* New York: Wiley.

Frangakis, C.E. and Rubin, D.B. (2002), "Principal stratification in causal inference," *Biometrics*, 58, 21–29.

Hedeker, D. and Gibbons, R.D. (1997) Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, 2, 64–78.

Huber, P.J. (1967) The behavior of maximum likelihood estimates under non-standard conditions. In *Fifth Berkeley Symposium in Mathematical Statistics and Probability*, 221–233. Berkeley: University of California Press.

Kenward, M.G. and Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, 13, 236–47.

King, G., Honaker, J., Joseph, A., and Scheve, K. (2001) Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review*, **95**, 49–69.

Laird, N.M. and Ware, J.H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Liang, K.Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.

Little, R.J. (1995) Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112–1121.

Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data, Second edition*. New York: Wiley. Rubin, D.B. (1976) Inference and missing data. *Biometrika*, 63, 581–592.

Meng, X.L. (1994) Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, **10**, 538–573.

Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846–866.

Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106–121.

Rubin, D.B. (1976) Inference and missing data. *Biometrika*, 63, 581–592.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, New York.

Rubin, D.B. (1996) Multiple imputation after 18+ years. *Journal of the American Statistical Association*, **91**, 473–489.

Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall/CRC Press.

Schafer, J.L. (1999) Multiple imputation: a primer. *Statistical Methods in Medical Research*, **8**, 3–15.

Schafer, J.L. (2003). Multiple imputation in multivariate problems where the imputer's and analyst's models differ. *Statistica Neerlandica*, 57, 19–35.

Schafer, J.L. and Kang, J.D.Y. (2005) Comment on "Semiparametric estimation of treatment effect in a pretest-posttest study with missing data" by M. Davidian, A.A. Tsiatis and S. Leon, *Statistical Science*.

Scharfstein, D.O., Rotnitzky, A. and Robins, J.M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association*, 94, 1096–1146.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. New York: Springer.

White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.

Zhang, J. L. and Rubin, D. B. (2003), "Estimation of Causal Effects via Principal Stratification When Some Outcomes Are Truncated By Death." Journal of Educational and Behavioral Statistics, 28, 353-368.