

# **Addressing Missing Data for Various Alternative Data Types**

Thomas R. Belin

UCLA Department of Biostatistics

# Effects of Nonresponse

Example reported in W.G. Cochran, *Sampling Techniques*, 3<sup>rd</sup> edition, 1977, ch. 13, where complete data from another source were available for one survey item:

	<u># of growers</u>	<u>% of pop'n</u>	<u>Ave. # of trees/grower</u>
1 <sup>st</sup> mailing responders	300	10	456
2 <sup>nd</sup> mailing responders	543	17	382
3 <sup>rd</sup> mailing responders	434	14	340
Nonresponders	1839	59	290

# Goals in incomplete-data settings

- **Accurately reflect available information**
  - Specifically, want to avoid bias in estimates of quantities of interest
  - Estimation could involve explicit model, or could involve procedure that corresponds to an implicit model
- **Accurately reflect uncertainty due to missingness**

# General strategies for analyzing data that are partially missing

- Complete-case/available-case analysis—drop cases that make analysis inconvenient
- Imputation procedures—fill in missing values, then analyze completed data sets using “complete-data” methods)
- Weighting procedures—modify “design weights” (i.e., inverse probabilities of selection from sampling plan) to account for probability of response
- Model-based approaches—develop model for partially missing data, base inferences on likelihood under that model

# Missing-data patterns

- General  
(e.g. item nonresponse)

	$X_1$	$X_2$	...	$X_p$
		?		
?				
				?
		?	?	

- Unit nonresponse  
(e.g., have some background data on all units, but some units don't respond to any question)

	$X_1$	$X_2$	...	$X_p$
	✓	✓	✓	✓
	✓	✓	?	?

## Missing-data patterns (cont'd)

- Monotone pattern  
(variables can be ordered such that one block of variables more observed than the next)

	$X_1$	$X_2$	...	$X_p$
	✓	✓	✓	✓
	✓	✓	✓	?
	✓	?	?	?

# Missing-data mechanism

Missing-data mechanism: Process by which some units observed, some units not observed

Does it matter to inferences? Yes!

Example (Little and Rubin): Sample of size 100 from Normal (mean=0, sd=1)

- Uncensored sample mean (n=100) = -0.03
- $P(\text{each case observed})=0.5 \Rightarrow m=52$ , sample mean = -0.11
- Pure censoring at 0  $\Rightarrow m=51$ , sample mean = -0.89
- Nonignorable censoring with  $P(y_i \text{ observed}) = \Phi(-2.05 y_i)$   
features  $m=53$ , sample mean = -0.81

# Missing-data mechanism with more than one variable, one subject to nonresponse

$X$  fully observed,  $Y = (Y_{\text{obs}}, Y_{\text{mis}})$

Possibilities:

- (1) Probability of response independent of  $X$  and  $Y$  (“missing completely at random” or MCAR)
- (2) Probability of response depends on  $X$  but not residually on  $Y$  (“missing at random” or MAR)
- (3) Probability of response depends on  $Y$  and possibly on  $X$  as well (“not missing at random” or NMAR)

$X$	$Y$
✓	✓
✓	?

# What can we discern from evidence in the data set at hand?

- May be evidence in the data set at hand to rule out (reject) MCAR
  - Example: Responders tend to have higher/lower average education by t-test
  - Example: Response more likely in one geographic area than another by chi-square test
- No evidence in data set at hand to rule out MAR (although there may be evidence from an external data source that bears on this question)

## Missing-data mechanisms: what is plausible?

- Cochran example: when human behavior is involved, MCAR must be viewed as an extremely special case that would often be violated in practice
- Missing data may be introduced by design (e.g., NAEP: measure some variables, don't measure others for reasons of cost, response burden), in which case MCAR would apply
- MAR is much more flexible than MCAR
- Cannot be too cavalier about assuming MAR, but anecdotal evidence shows that it often is plausible when conditioning on enough information (e.g., David, et al. 1986 *JASA*; Belin, et al. 1993 *JASA*; Rubin, Stern, Vehovar 1995 *JASA*)
- Sensitivity analysis or models drawing on external data might consider NMAR (e.g., Heitjan and Landis 1994 *JASA*)

# Ignorable nonresponse

If missing-data mechanism is MCAR or MAR (and another much less restrictive assumption applies to the parameters of the data model and the parameters of the missing-data mechanism), then nonresponse is said to be “ignorable”

The term comes from the fact that in likelihood-based inference, both the data model and missing-data mechanism are important in general, but with MCAR or MAR missingness, inference can be based solely on a model for the data, ignoring the missing-data mechanism (thus making inference much simpler)

“Ignorability” is a relative assumption: missingness on income may be NMAR given only gender, but may be MAR given gender, age, occupation, region of the country

# Imputation

Fill in missing values, analyze completed data sets

Advantage:

- Rectangular data sets easier to analyze

Disadvantage:

- “Both seductive and dangerous” (Little and Rubin, *Statistical Analysis with Missing Data*, 1987, 2002) since it is possible to understate uncertainty due to missing values if not careful and to induce bias if imputing under the wrong model.

# Imputation illustration

- Unconditional mean imputation: impute “5”
- Conditional mean imputation (Buck’s method): impute “10”
- Hot-deck imputation: impute 2, 4, 6, or 8 (with unconditional, conditional variations)
- Regression imputation: impute  $10 \pm$  residual based on estimated residual variance
- Multiple imputation: repeat hot-deck or model-based (e.g., regression) imputation to create multiple completed data sets, combine separate analyses into overall inference

X	Y
1	2
2	4
3	6
4	8
5	?

# Multiple imputation inference

Consider  $m$  imputed data sets (e.g.,  $m = 5$ ). For some quantity of interest  $Q$  with squared standard error  $U$ , calculate  $Q_1, Q_2, \dots, Q_m$  and  $U_1, U_2, \dots, U_m$  (e.g., carry out 5 regression analyses, obtain point estimate, squared standard error of  $\beta_1$  from each)

Then calculate:

$$\bar{Q} = \sum_{i=1}^m Q_i \quad \bar{U} = \sum_{i=1}^m U_i \quad B = \frac{1}{m-1} \sum_{i=1}^m (Q_i - \bar{Q})^2$$
$$T = \bar{U} + \frac{m+1}{m} B$$

## Multiple imputation inference (cont'd)

Then significance tests, interval estimates can be based on

$$\frac{\bar{Q} - Q}{\sqrt{T}} \sim t_\nu \quad \text{where} \quad \nu = (m-1) \left(1 + \frac{1}{m+1} \frac{\bar{U}}{B}\right)^2$$

(approximate degrees of freedom analogous to calculation for comparison of normal means with unequal variances, i.e., using Satterthwaite approximation, with improvement available for small data sets: Barnard and Rubin 1999 *Biometrika*, see also Little and Rubin 2002)

Ratio of ( $B$  = between-imputation variance) to ( $T$  = between- + within-imputation variance) is known as the fraction of missing information (related to  $\nu$ , estimate:  $(1 + \frac{1}{m}) \frac{B}{T}$  )

# Multiple imputation as a paradigm

Logic: “Average over” uncertainty, don’t assume most likely scenario covers all plausible scenarios

Principle: Want nominal 95% intervals to cover targets of estimation 95% of the time

Rubin (*Multiple Imputation for Nonresponse in Surveys*, 1987) describes conditions for imputations to be “proper” (in broad terms, requires mean/variance of imputations to be consistent)

Simulation studies show that

- Proper imputations will yield close to nominal coverage
- Improvement over single imputation is meaningful
- Number of imputations can be modest—even 2 adequate for many purposes, some dependence on whether  $Q$  is univariate or multivariate and on fraction of missing information

# Imputation for normal continuous data

- Schafer (1997, *Analysis of Incomplete Multivariate Data*, CRC Press) outlines methods for multivariate normal model
- Connection to familiar imputation ideas: in multivariate normal model, conditional distribution of one variable (or set of variables) given others is a normal linear regression
- Software
  - Splus/R packages (NORM) as well as standalone Windows package at [www.stat.psu.edu/jls](http://www.stat.psu.edu/jls)
  - SAS PROC MI uses this approach

# Practical considerations

- Framework can easily accommodate fully-observed binary variables (which play the role of dummy variables in regression when used as predictors), and no adverse consequence associated with non-normality since not being imputed
- Typically desirable to include many predictors, both to improve precision of imputed values and to make MAR assumption more plausible, but number of covariance parameters goes up as the square of the number of variables in the model, implying practical limits on the number of variables for which parameters can be estimated well

## Software for multivariate normal imputation

SAS introduced PROC MI, available in Version 8.2 and higher, to produce multiple imputations under multivariate normal model, as well as PROC MI ANALYZE allowing users to combine results from analyses of multiply imputed data sets that used BY statements according to standard formulas

Incorporates alternative computing methods, but methods described by Schafer consistent with MCMC (Markov-Chain Monte Carlo) strategy

Includes an approach described by Schafer for using “ridge prior” that can stabilize estimation when number of variables is large (e.g., PRIOR=RIDGE=3)

# Alternative modeling strategies when dealing with unstable or inestimable parameters

If too many variables included in the model for covariance parameters to be estimated well (e.g., in case of warning message from PROC MI about lack of convergence), could:

- delete variables (?),
- Impute for subgroups of variables separately
  - Instead of partitioning variables into two groups (X,Y) and imputing separately, might partition variables into three groups (X,Y,Z) and impute separately for (X,Y) and (X,Z)
  - Two-group approach posits independence between X,Y; three-group (or multi-group) approach posits conditional independence between Y and Z given X
- Ridge prior

## Ridge prior

- Idea: if covariance matrix singular or close to singular, add a small positive element to the diagonal of the covariance matrix
- E.g., to add the information content of 3 observations to an imputation model attempting to accommodate 100 observations on 100 variables, use `PRIOR=RIDGE=3` option in model
- Song and Belin (2004 *Statistics in Medicine*) showed that ridge prior performed comparably to method specifically designed for factor-model imputation when factor model correct

## Methods for multinomial (categorical) data

- If data arise as multinomial sample (e.g., cross-classified contingency table), Schafer (1997) outlines methods for imputation either under saturated model (allowing all possible interactions) or under log-linear model positing existence of certain interactions—Splus/R software
- Unlike multivariate normal imputation, includes model-selection step (can carry out likelihood-ratio tests comparing nested reduced and expanded models)
- Note: interaction effects in multivariate normal setting could be represented in a back-door manner by fitting separate imputation models to distinct subgroups of cases, e.g., a model for treated and a model for control subjects)

## Combined normal and multinomial data: general location model

Schafer (1997) also outlines methods for mixed continuous and categorical data using a “general location model”

Continuous variables assumed multivariate normal within cells defined by categorical variables with (possibly) distinct means for the normal vector within each cell but with common covariance matrix across cells

Categorical variables governed by log-linear model constraints

Possible to allow for multivariate-regression constraints on means of continuous variables (i.e., means of continuous variables follow multivariate regression conditional on indicators for levels of categorical variables)

Spplus/R routines available (MIXED)

# Practical considerations from empirical evaluation

- Belin, Hu, Young, Grusky (1999 *Statistics in Medicine*) fit general location model in setting with 16 binary variables and 18 continuous variables
- Follow-up data were obtained on some initial nonrespondents, allowing for comparison of imputed values with later observed values
- Some discrepancies seen in continuously-scaled variables (particularly ordinal variables with skewed distributions)
- Substantial discrepancies seen in predictions of categorical variables
  - Assumption of common covariance matrix has side effects for prediction of binary variables—discriminant-analysis perspective suggests that model predictions would be too precise if reality features departures from common covariance assumption

# Robustness of Multivariate Normal Imputation for Incomplete Binary Data

- Bernaards, Belin, and Schafer (2007 *Statistics in Medicine*): Properties of methods for producing binary imputations from multivariate normal models?
- Candidate procedures:
  - Simple rounding: Round imputations from normal model to the nearer of 0 or 1
  - “Coin flipping”: View imputations between 0 and 1 from normal model as probabilities and take a corresponding Bernoulli draw (consistent with using  $E[Y|X]$  for imputation since  $E[Y|X] = P[Y=1|X]$  for binary  $Y$ )
  - Adaptive rounding: Round to 0 or 1, but instead of using 0.5 as a cut-point, use  $\bar{\omega} - \Phi^{-1}(\bar{\omega}) \cdot \sqrt{\bar{\omega}(1 - \bar{\omega})}$

## Related theoretical work

- Horton, Lipsitz, and Parzen (2003 *American Statistician*) show that producing binary imputations by rounding multivariate normal imputations produces predictable bias in estimates of proportions
- Present work: what are properties of rounding and other approximation methods
  - in finite samples?
  - for parameters other than proportions?

# Evaluation framework using California Healthy Kids Survey (CHKS)

- CHKS target variables:
  - age (10-17)
  - Gender
  - gang membership
  - physical activity  
(# days/week with  $\geq 30$  minutes exercise)
  - carrot consumption  
(ordinal: 0, 1-3/week, 4-6/week, 1/day, 2/day, 3+/day)
- 206,802 respondents:  
198,262 complete, 8,540 with at least one missing item
- Viewed complete cases as finite population, imposed missing-data patterns consistent with incomplete cases

## Evaluation approach

- Sample sizes: 50, 500
- Missingness rate: 25% with  $\geq 1$  missing item, 50% with  $\geq 1$  missing item
- Imputation method:  
simple rounding, coin flipping, adaptive rounding
- Parameters of interest:
  - proportions (gender, gang membership)
  - odds ratio (gang membership vs. gender)
  - difference in mean physical activity by gang membership

# Evaluation framework

- 1,000 replicates for each approximation method
- Viewed finite-population quantity as target parameter, evaluated bias and coverage relative to finite-population target
- Margin of error of 1.4% for 95% coverage statistics, i.e., coverage not distinguishable from nominal level if observed coverage in the range (0.936, 0.964)

## Results: Gender proportion (50% missing scenario)

Population proportion: 0.459

<u>Method</u>	<u>n</u>	<u>Coverage (%)</u>
Simple rounding	50	95.4
Simple rounding	500	95.6
Coin flipping	50	95.4
Coin flipping	500	95.1
Adaptive rounding	50	95.3
Adaptive rounding	500	95.4

## Results: Gang membership proportion (50% missing scenario)

Population proportion: 0.092

<u>Method</u>	<u>n</u>	<u>Coverage (%)</u>
Simple rounding	50	89.7
Simple rounding	500	87.7
Coin flipping	50	97.0
Coin flipping	500	85.9
Adaptive rounding	50	90.5
Adaptive rounding	500	91.7

## Results: Odds ratio between gang membership and gender (50% missing scenario)

Population odds ratio: 0.119

(boys ~ 8x more likely than girls to be in gang)

<u>Method</u>	<u>n</u>	<u>Coverage (%)</u>
Simple rounding	50	96.4
Simple rounding	500	90.8
Coin flipping	50	95.4
Coin flipping	500	83.9
Adaptive rounding	50	97.5
Adaptive rounding	500	93.7

## Intuition regarding impact of coin-flipping

Why is large-sample coverage for coin-flipping so far below nominal level for OR but not for proportion?

Coin-flipping for separate variables is done independently.

Thus, in a setting with true odds ratio far from 1, coin-flipping will tend to bias odds ratio toward 1 instead of preserving observed pattern of association in data, resulting in degraded coverage.

(In a separate logistic regression analysis with less extreme odds ratios, all methods produced close-to-nominal coverage.)

## Results: Difference in mean physical activity days by gang membership (50% missing scenario)

Population mean difference: 0.088

<u>Method</u>	<u>n</u>	<u>Coverage (%)</u>
Simple rounding	50	94.1
Simple rounding	500	95.9
Coin flipping	50	96.7
Coin flipping	500	96.6
Adaptive rounding	50	93.9
Adaptive rounding	500	94.5

## Binary imputation from normal model: conclusions

- For proportions close to 0.5, imputing binary variables based on normal approximation performs well.
- For proportions far from 0.5, normal-approximation methods suffer some deficiency in coverage, with adaptive rounding not as far off as the other methods.
- For odds ratio involving two binary variables where missingness possible on both, coin-flipping shows deficiency in coverage compared to other methods, consistent with intuition about what would happen when binary draws performed independently for separate variables.
- Across the range of quantities estimated, adaptive rounding had the best performance

## Imputation using overlapping regression specifications (“incompatible Gibbs sampling”)

Alternatives exist for dealing with mixed data types using multivariate sequential regression specification (i.e., overlapping regression models)

IVEWare (Raghunathan, et al. 2002), available as SAS macros ([www.isr.umich.edu/src/smp/ive](http://www.isr.umich.edu/src/smp/ive))

Suppose  $X$  continuous,  $Y$  binary,  $Z$  Poisson count. Assume:

- $X | Y, Z$  is a linear regression
- $Y | X, Z$  is a logistic regression
- $Z | X, Y$  is a Poisson regression

(even if no joint distribution of  $(X, Y, Z)$  has these properties)

MICE (multiple imputation using chained equations), ICE (Stata routine) use similar strategy.

## Capabilities of IVEWare

- IMPUTE module accommodates continuous (normal linear regression), binary (logistic regression), categorical (polytomous logistic regression) count (Poisson regression), and semi-continuous data (logistic regression for any vs. none followed by linear regression to predict amount for those with any)
- Can restrict imputations to subpopulations (accommodate skip patterns), place bounds on imputed values, incorporate interactions into model specifications

## IVEWare Capabilities (cont'd)

- IMPUTE module (multivariate sequential regression)
- DESCRIBE module: descriptive summaries for means, proportions, contrasts
  - Multiple imputation for missing data
  - Taylor-series variance adjustments for complex sample designs
- Other modules allow a variety of approaches to address missing data in context of complex sample designs (including linear/logistic/Poisson/proportional hazards regression, PROC MIXED/GENMOD/NLIN/PROBIT, others)

# Multiple imputation using chained equations (MICE)

- Available through [www.multiple-imputation.com](http://www.multiple-imputation.com)
- Sequential regression specification with interface to R, Splus
- Built-in capabilities for models not quite as extensive as IVEWare (normal, binary, categorical)
- Also allows random sample from observed values (although if performing hot-deck, still want to condition on observed characteristics)

## Software for multiple imputation: other specialized approaches

Hot-deck imputation has the appealing feature that imputed values stay within range of existing data

Macros developed by Bob Bell (while at RAND) have been used by various projects at UCLA, RAND to impute using “predictive-mean matching” hot deck:

- For target variable to be imputed ( $Y$ ), regress  $Y$  on a set of predictor variables using available cases
- For all cases (both where  $Y$  missing and  $Y$  observed), use fitted regression to obtain predicted mean of  $Y$
- Use predicted means to define (e.g. 10) cells within which hot-deck imputation is performed

(Avoids problems with sparse cells when it is desired to control for a large number of characteristics.)

## Insights from empirical evaluation: hot-deck versus multivariate normal with variable transformations

- Tang, Song, Belin, and Unutzer (2005 *Statistics in Medicine*) use data from IMPACT study of late-life depression to simulate finite-population scenarios, comparing coverage of target quantities using hot-deck versus multivariate-normal model with adaptations from plausible data analysis (e.g., transformations to improve normality)
- Hot-deck produced close-to-nominal coverage throughout; normal model performed well for most variables but showed some deficiency in coverage (e.g. ~85-90% instead of 95%) for means of some skewed variables

## Modeling, programming considerations

- Hot-deck requires specification of predictors for predictive-mean-matching model for each variable with incomplete data
- Can give rise to cumbersome programming task in settings with dozens of variables
- Considerations in imputation strategies:
  - models with flexibility to fit data
  - “congeniality” among models used to impute different variables (joint models automatically provide)
  - desire to keep programming task manageable

Parameter-extended Metropolis-  
Hastings for multivariate probit  
(MVP), multinomial probit (MNP),  
multivariate multinomial probit  
(MMNP) models

## Multivariate probit (MVP) model (Chib and Greenberg 1998 *Biometrika*)

- Suppose  $Y_i$  is a  $p$ -vector of ordinal measures
- Model the ordinal data vector by assuming each of its elements is a discretized version of a latent continuous score
- $Y_i$  follows an MVP model if we assume the latent score vector  $Z_i \sim N_p(X_i\beta, \Phi)$  where  $\Phi$  is a  $p \times p$  correlation matrix
- Allows modeling of longitudinal or clustered binary data, ordinal data

Multinomial probit (MNP) model  
(McCulloch, Polson, Rossi 2000 *Journal  
of Econometrics*)

- If  $Y_i$  takes on a value in  $\{0, 1, 2, \dots, K\}$ , then define a  $K$ -dimensional latent “utility” vector  $Z_i \sim N_p(X_i\beta, \Phi)$  such that
$$Y_i = \begin{cases} 0 & \text{if all the utilities are less than 0, else} \\ k & \text{if the largest utility is } Z_{ik} \end{cases}$$
- Covariates  $X_i$  can be choice-specific or subject-specific
- Restrict  $\Phi_{11} = 1$  for identifiability

# Parameter-extended Metropolis-Hastings (PX-MH)

- Parameter expansion idea (Liu, Rubin, Wu 1998 *Biometrika*): Expanding the parameter space with an auxiliary variable (about which there is no information in the data) can speed convergence, facilitate computing (e.g., multivariate  $t$  models)
- MVP and MNP models difficult to fit due to constraints on covariance matrix—previous ideas (Chib and Greenberg 1998) include use of multivariate normal prior for correlation matrix in MVP model truncated to ensure positive-definiteness, Metropolis-Hastings to implement

# Parameter-extended Metropolis-Hastings (PX-MH) (cont'd)

- Idea: define diagonal matrix of artificial variance components  $D$ , and let  $\Sigma = D^{1/2}\Phi D^{1/2}$
- Assume a distribution for  $\Sigma$  (without restrictions) and account for Jacobian of transformation from  $\Sigma$  to  $(\Phi, D)$  in calculating acceptance probabilities in Metropolis-Hastings
- Conjugacy is lost after transformation—could use Wishart as easily as inverse-Wishart
- Can be used to fit MVP, MNP models

# Multivariate multinomial probit (MMNP) model

- PX-MH machinery facilitates extension of MNP model to multivariate version (MMNP)
- Geweke, Keane, Runkle (1997 *Journal of Econometrics*) extend MNP model to a multiperiod MNP model assuming AR(1) structure over time
- Boscardin and Weiss (2001 *SBSS Proceedings*) present modeling idea to allow departures from parametric covariance structures; PX-MH approach can be combined with Boscardin-Weiss strategy to allow departures from AR(1) (or from any other posited covariance structure)

## Potential Extensions

- Concept: embed imputation of other data types into normal modeling framework, using latent variables to accommodate ordinal, nominal, semi-continuous data
- Can model-based approaches be sufficiently flexible to accommodate large numbers of variables and arbitrary combinations of data types, yet be sufficiently structured to put exchangeability of units to good use?

# References

- Belin TR, Hu MY, Young AS, Grusky O. Performance of a general location model with an ignorable missing-data assumption in a multivariate mental health services study with incomplete data. *Statistics in Medicine*, 1999; 18:3123-3135.
- Bernaards C, Belin TR, Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, to appear.
- Boscardin WJ, Weiss RE. Models for the covariance matrix of multivariate longitudinal and repeated measures data, *Proceedings of the ASA Section on Bayesian Statistical Science*, 2001.
- Boscardin WJ, Zhang X. Modeling the covariance and correlation matrix of repeated measures, in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, A. Gelman and X.L. Meng, eds., 2004, New York: John Wiley.

## References (cont'd)

- Song J, Belin TR. Imputation for incomplete high-dimensional multivariate normal data using a common factor model. *Statistics in Medicine*, 2004; 23:2827-2843.
- Tang L, Song J, Belin TR, Unutzer J. A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine*, 2005; 24:2111-2128.
- Wang J, Belin TR. Handling incomplete high dimensional multivariate longitudinal data by multiple imputation using a longitudinal factor analysis model. *Proceedings of the ASA Section on Statistical Computing*, 2002; 3615-3620.
- Zhang X, Boscardin WJ, Belin TR. Sampling correlation matrices in Bayesian models with correlated latent variables. *Journal of Computational and Graphical Statistics*, to appear.
- Zhang X, Boscardin WJ, Belin TR. Multivariate extensions to multinomial probit models using parameter-extended Metropolis-Hastings. *Proceedings of the ASA Biometrics Section*, 2005; 169-176.

Thank you!

Best wishes to  
Naihua Duan  
Suzanne Slocum!