

Missing Data in Randomized Clinical Trials

John Barnard

Department of Quantitative Health Sciences

The Cleveland Clinic Foundation

What are the Data?

- Complete (intended) data: $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, where

Y_{obs} = observed data

Y_{mis} = missing data

- Missing data assumed to have true underlying values
Ex: Missing vital lung capacity after death
- Doesn't include missing potential outcomes
- R = response indicators
- Observed data: (Y_{obs}, R)

Missing Data Patterns

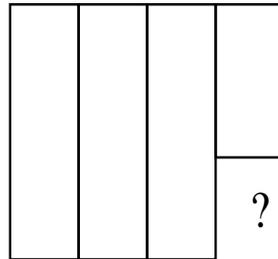
General Pattern

		X_1	X_2	X_p	
Cases units subjects	1	?			?		?=missing
	2					?	
	⋮						
	⋮	?	?			?	
	⋮				?	?	
	n						

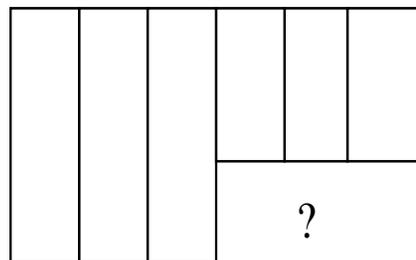
Missing Data Patterns

Special Patterns

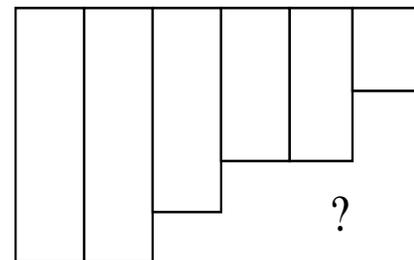
Univariate Missing



“Unit Nonresponse”



Monotone Missing Data



Example: IMPACT Multi-Centre Study

- Example from Tang et al. (2005)
- Multi-centre RCT to study effectiveness of a disease management programme for late-life depression
- 2 groups: usual care group and IMPACT group
- IMPACT group given access for 12 months to a depression care manager
- Total sample size of 1801
- Outcome measures taken at 3, 6, and 12 months
- Rich collection of covariates taken at baseline and a followup visits

Dropout in IMPACT RCT

Table I. Unit response pattern over waves*.

Baseline	Month-3	Month-6	Month-12	Overall	Usual care	Intervention
X	X	X	X	1433	688	745
X	X	X	M	91	52	39
X	X	M	X	22	18	4
X	X	M	M	78	41	37
X	M	X	X	31	19	12
X	M	X	M	15	10	5
X	M	M	X	8	4	4
X	M	M	M	123	63	60

*X: responded; M: missing the wave (missing, drop-out or dead).

Tale of Two Likelihoods

- *Full* likelihood – involves missing-data process

$$L_{\text{full}}(\theta, \phi \mid Y_{\text{obs}}, R) \propto \int f(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta) f(R \mid Y_{\text{obs}}, Y_{\text{mis}}, \phi) dY_{\text{mis}}$$

- *Partial* likelihood – ignores the missing-data process

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

- When can you safely use the much simpler partial likelihood?

Assumptions about Missingness

Assumptions about $f(R | Y_{\text{obs}}, Y_{\text{mis}}, \phi)$:

1. **Missing Completely at Random (MCAR)** – MD process does not depend on observed or missing values

$$f(R | Y_{\text{obs}}, Y_{\text{mis}}, \phi) = f(R | \phi)$$

2. **Missing at Random (MAR)** – MD process does not depend on missing values

$$f(R | Y_{\text{obs}}, Y_{\text{mis}}, \phi) = f(R | Y_{\text{obs}}, \phi)$$

3. **Not Missing at Random (NMAR)** – MD process can depend on observed and missing values

$$f(R | Y_{\text{obs}}, Y_{\text{mis}}, \phi) = f(R | Y_{\text{obs}}, Y_{\text{mis}}, \phi)$$

Ignorability

When can we ignore the missing-data process when making inferences?

- Sufficient conditions for **ignorability** of the MD process:
 1. MD process is MAR
 2. Parameters of the MD process (ϕ) and complete-data model (θ) are distinct
- If MAR holds but not parameter distinctness, ML based on L_{partial} is valid but not fully efficient
- MAR is the key condition for MD ignorability
 - Richer the observed data Y_{obs} , the more plausible the MAR assumption
 - MAR is nice \Rightarrow don't have to model the MD process
- NMAR assumption is often more plausible, but very difficult to justify specific choices because there is no evidence in the data against MAR

Handling Missing Data

- **Available/Complete Cases**
 - Naive: simple but often wrong
- **Summary Measures**
 - Common in longitudinal studies
- **Last Value Carried Forward**
 - Common in longitudinal studies
- **Single Imputation**
 - Fill in missing values
- **Multiple Imputation**
 - Principled
- **Bayesian/Likelihood Based**
 - Principled

Complete Case Analysis

Discard Incomplete Cases

Good :

- Easy

- Common sample base for comparisons

Bad:

- Loss of information in incomplete cases

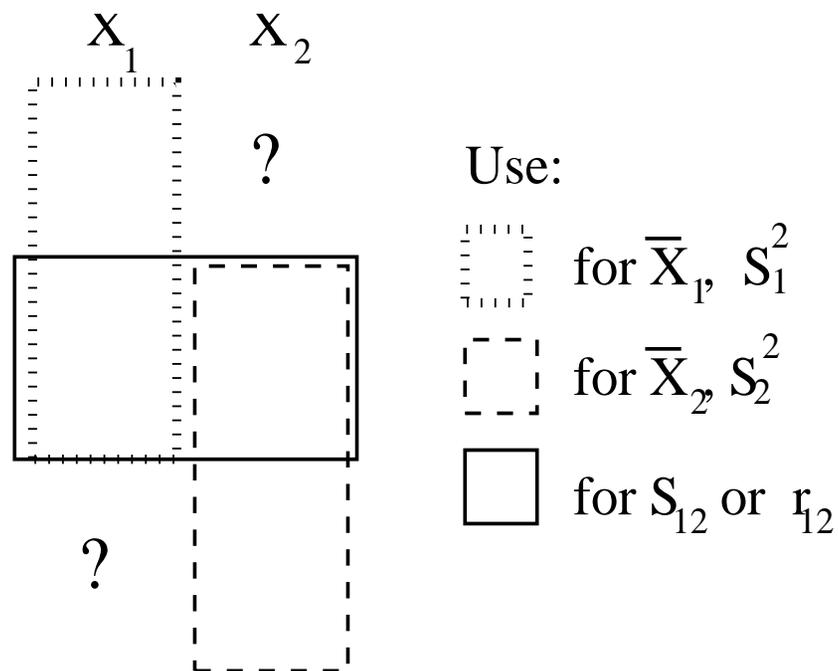
- Increased variance (smaller n)

- Bias (systematic differences between resp. and nonresp.)

- * Try to adjust with weights

Available-Case Analysis

- Use:
- Available cases for each mean, variance
 - Available pairs for correlation or covariance



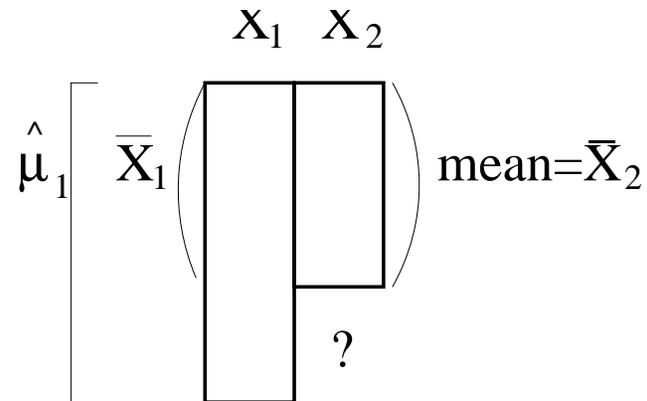
Available Case Analysis—Continued

- Good:
- Easy (but inference for some statistics non-trivial, e.g., slopes)
 - Makes apparently better use of available data

- Bad:
- Correlations outside (-1,1)
 - Correlation matrix not positive definite (slopes??)
 - Can be even worse than complete cases

X	1	2	3	
	✓	✓	?	→ $r_{12} = 1$
	?	✓	✓	→ $r_{23} = 1$
	✓	?	✓	→ $r_{13} = -1$

Example: Bivariate Monotone Missing Data



$$E(X_1) = \mu_1 \qquad E(X_2) = \mu_2$$

$$\text{Var}(X_1) = \text{Var}(X_2) = \sigma^2 \qquad \text{Corr}(X_1, X_2) = \rho$$

For inference about $\Delta = \mu_2 - \mu_1$:

Complete Cases: $\bar{X}_2 - \bar{X}_1$ (CC)

Available Cases: $\bar{X}_2 - \hat{\mu}_1$ (AC)

Example: Bivariate Monotone Missing Data

If Missingness of X_2 is Independent of X_1 and X_2
(Missing Completely at Random)

Then for Δ

- CC and AC unbiased
- $\text{Var}(\text{CC}) > \text{Var}(\text{AC})$ if $\rho < \frac{1}{2}$
 $\text{Var}(\text{CC}) < \text{Var}(\text{AC})$ if $\rho > \frac{1}{2}$

Example: Bivariate Monotone Missing Data

If Missingness of X_2 Depends on X_1
(Missing at Random (MAR))

- CC and AC biased
- Bias (CC) < Bias (AC) (usually)

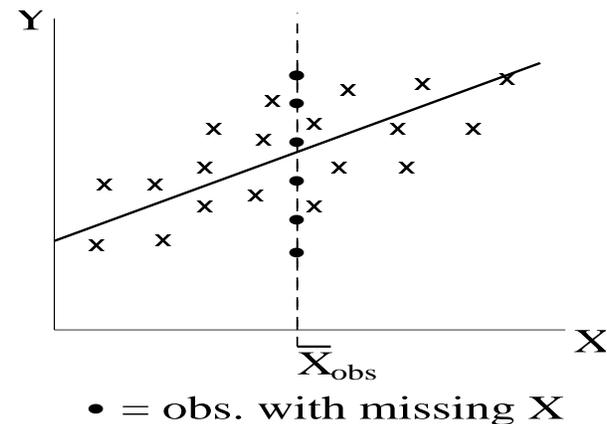
Linear Regression with Missing Predictor Variables

Parameter of interest: slope (β) of the regression of Y on X

Y	X	I
		0
		0
		⋮
		0
	?	1
		⋮
		1

n_{obs} obs. values of
 n_{mis} miss. values of

impute \bar{X}_{obs}



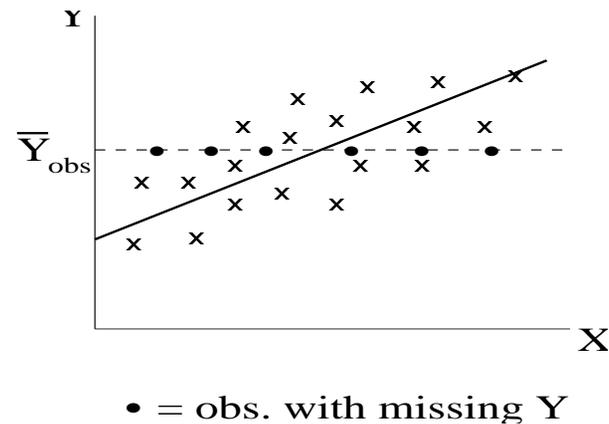
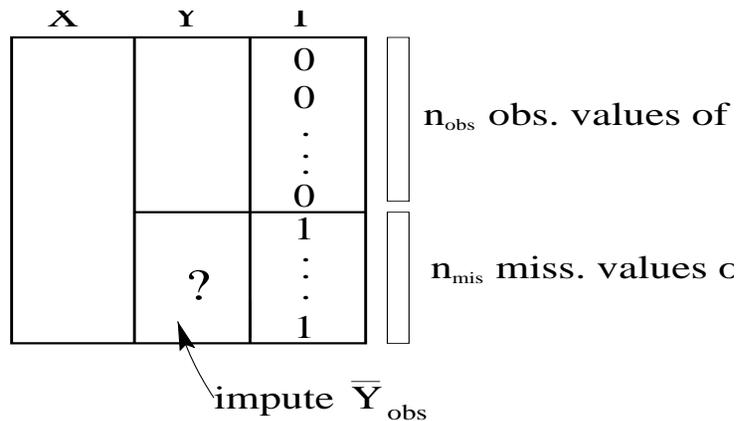
1. Impute \bar{X}_{obs} for all missing X 's (can impute all 0's or any other value)
2. Fit : $Y = \beta_0 + \beta_1 I + \beta_2 X + \varepsilon$

Results: $\hat{\beta}_2 = \hat{\beta}_{CC}$, $\hat{\beta}_1 = \frac{n}{n_{\text{obs}}}(\bar{Y}_{\text{mis}} - \bar{Y})$, $\hat{\beta}_0 = \hat{\beta}_{CC}$

$\hat{\beta}_2$ unbiased for β under MCAR, biased under MAR
 (Under MCAR, MLE more efficient)

Linear Regression with Missing Response Variable

Parameter of interest: slope (β) of the regression of Y on X



1. Impute \bar{Y}_{obs} for all missing Y 's
2. Fit : $Y = \beta_0 + \beta_1 I + \beta_2 X + \varepsilon$

Results: $\hat{\beta}_2 = \phi \hat{\beta}_{CC} + (1 - \phi) \hat{\beta}_{\text{zero}} = \phi \hat{\beta}_{CC}$

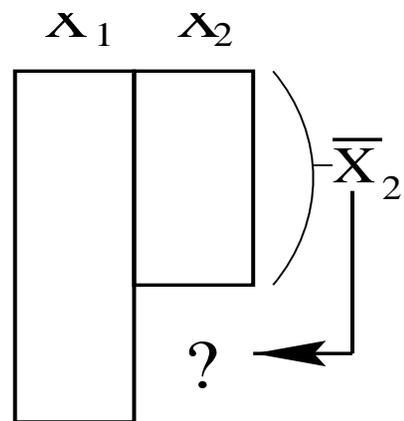
where $\phi = \frac{1}{\left(1 + \frac{SS_{X(\text{mis})}}{SS_{X(\text{obs})}}\right)} < 1$

$\hat{\beta}_2$ biased for β even under MCAR!

Imputation (single)

Impute Means

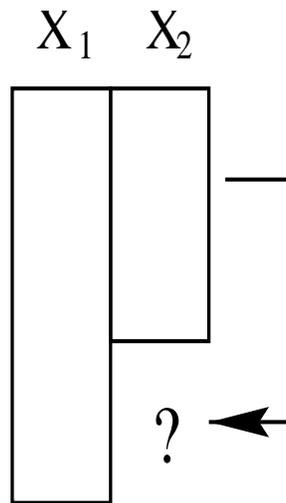
Unconditional



But $s_2^2 < \sigma_2^2$!

Impute Means (continued)

Conditional on Observed Values in Case



Regress X_2 on X_1 from complete cases

Impute predictions $\hat{X}_{i2} = E(X_{i2}|X_{i1})$

But $s_{2.1}^2 < \sigma_{2.1}^2!$

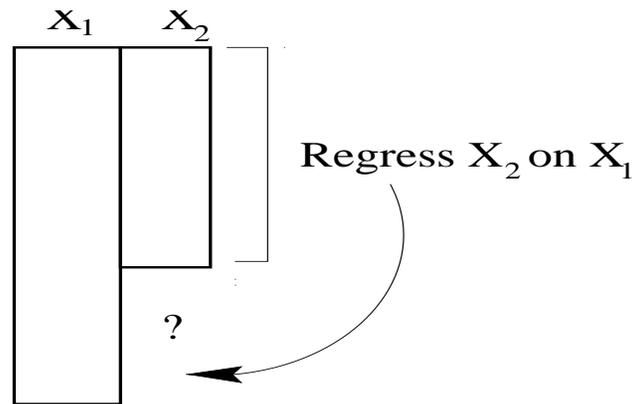
(residual variance of X_2 given X_1)

Notes on Mean Imputation

- Marginal distributions and associations distorted (no residual variance)
- Standard errors from filled-in data too small
 - no residual variance
 - n actually smaller
 - uncertainty of prediction
- Conditional better than unconditional, which is often worse than doing nothing (AC, CC)

Impute Values (not means) from a Distribution

Example 1



Impute: $E(X_{i2}|x_{i1}) + r_i$

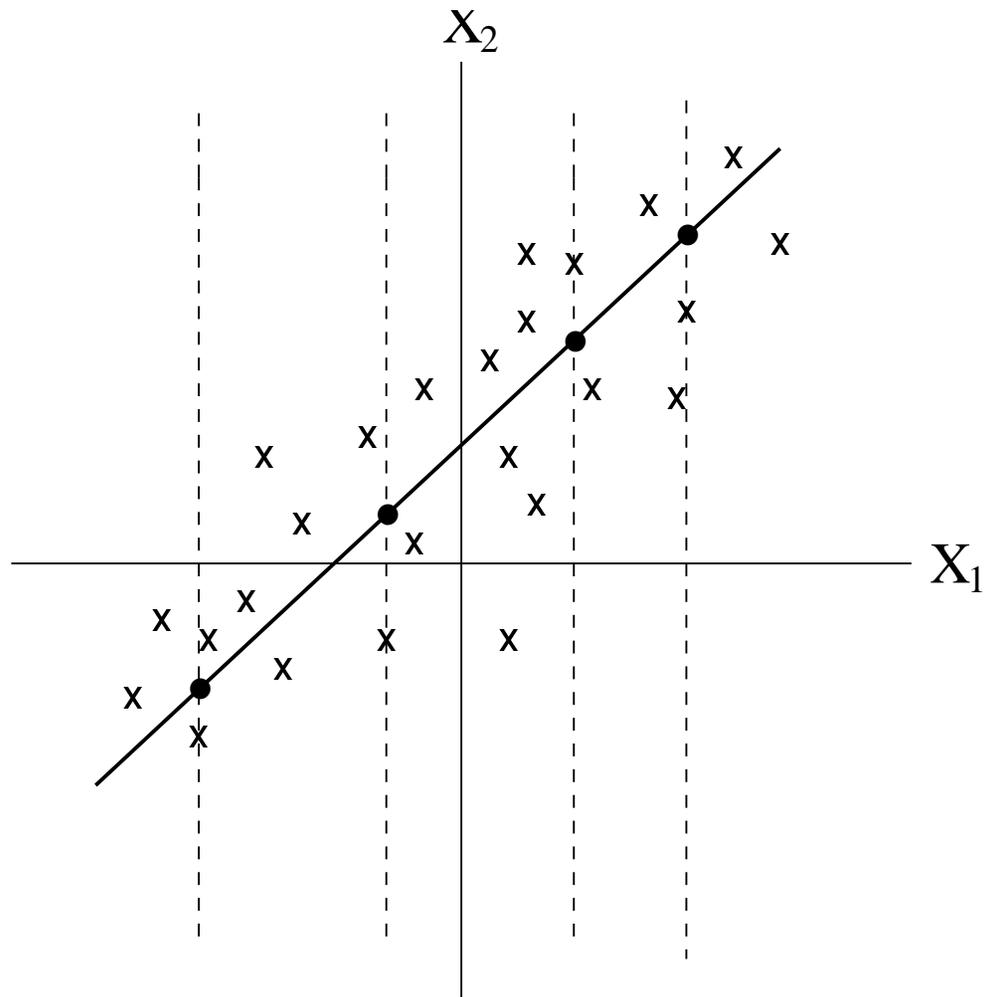
where

$r_i =$ residual from regression (empirical)

or

$r_i \sim N(0, s_{2.1}^2), s_{2.1}^2 =$ residual variance

Impute Values (not means) from a Distribution



Single Proper Imputation: Scalar Estimand

30% missing information is large in a typical study
Does **single proper imputation** work in this case?

Confidence Coverages for Scalar Estimand

(e.g., mean, regression coef., correlation coef.,
cell proportion, factor loading, etc.)

nominal	90%	95%	99%
actual	77%	85%	94%

Can lead to wrong conclusions about effects!
overly confident of results – underestimate **risks**

Single Proper Imputation: Multivariate Estimand

Significance Levels for Testing Ten Component Null Hypothesis

(e.g., 10 component regression coefficient,
10 component interaction in a contingency table)

nominal	1%	5%	10%
actual	25%	45%	57%

In fact \rightarrow 100% as number of components increase

Leads to **bad** conclusions about data!

**Almost certainly reject adequate simple models
in favor of overly complex models**

Notes on Imputing Values from a Distribution

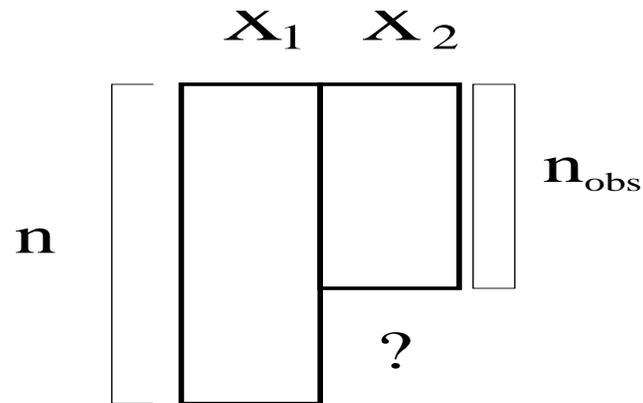
- Less efficient for estimates of means
- Better than mean imputation for distributions and associations
- Standard errors from filled-in data are still too small
 - n smaller
 - uncertainty of estimation

Multiple Imputation can fix these!

Summary of Some Naive Methods

- Methods can be useful if the amount of missing data is small (although good definition of small is difficult)
- Performance is unreliable
- Methods are ad hoc, and may need ad hoc adjustments
- Tests and confidence intervals are generally wrong (even asymptotically)
- Intuition can lead you astray – need principles

Principled Attacks: Bivariate Monotone Missing Data



- Maximum likelihood (assuming normality)
 - Factored likelihoods
 - EM
- Multiple imputation
 - Introduction to general approaches using this technique

Maximum Likelihood

- Model for (X_1, X_2) governed by parameters

$$(X_1, X_2) \sim N_2 \left([\mu_1, \mu_2], \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \right)$$

- Likelihood function of parameter, θ , given observed values (under MAR)

$$f(X_{\text{obs}} | \theta) = \prod_{i=1}^{n_{\text{obs}}} N_2(X_{i1}, X_{i2} | \theta) \prod_{i=n_{\text{obs}}+1}^n N_1(X_{i1} | \mu_1, \sigma_{11})$$

Maximum Likelihood

- Estimate θ by maximizing the likelihood (factored like. for special patterns):

$$f(X_{\text{obs}}|\theta) = \underbrace{\prod_{i=1}^n N_1(X_{i1}|\mu_1, \sigma_{11})}_{\text{available cases}} \underbrace{\prod_{i=1}^{n_{\text{obs}}} N_1(X_{i2}|\alpha_{2.1} + \beta_{2.1}X_{i1}, \sigma_{22.1})}_{\text{complete cases}}$$

distinct functions of θ

$$\hat{\mu}_1, \hat{\sigma}_{11}$$

mean, variance

available cases

$$\hat{\alpha}_{2.1}, \hat{\beta}_{2.1}, \hat{\sigma}_{22.1}$$

least squares

regression estimates

complete cases

- Transform to find MLE's of $\mu_2, \sigma_{12}, \sigma_{22}$:

$$\hat{\mu}_2 = \bar{X}_2 + \hat{\beta}_{2.1}(\hat{\mu}_1 - \bar{X}_1)$$

$$\hat{\sigma}_{12} = s_{12}(\hat{\sigma}_{11}/s_{11})$$

$$\hat{\sigma}_{22} = s_{22} + \hat{\beta}_{2.1}^2(\hat{\sigma}_{11} - s_{11})$$

$$\hat{\rho} = \hat{\sigma}_{12}/\sqrt{\hat{\sigma}_{11}\hat{\sigma}_{22}}$$

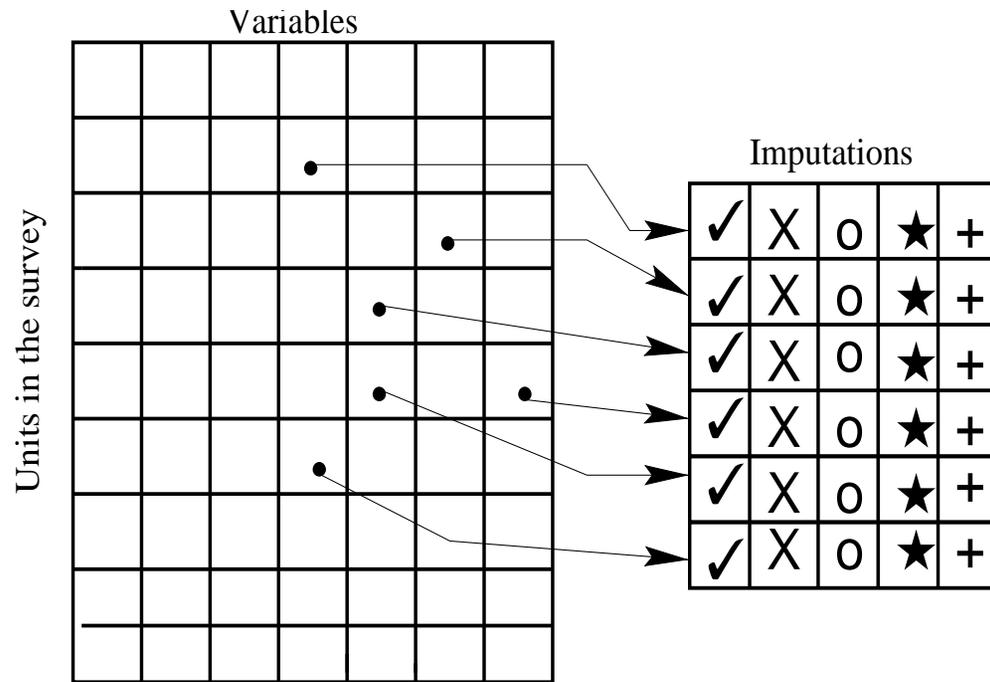
Advantages of Maximum Likelihood

Why are these estimates preferable?

- All estimates are in the parameter space
- They are consistent and efficient if underlying model is correct
E.g., $\hat{\mu}_2 - \hat{\mu}_1$ unbiased with smaller variance than $\bar{X}_2 - \bar{X}_1$ or $\bar{X}_2 - \hat{\mu}_1$
- This holds true under wider class of models
E.g., missingness can be MAR (missing at random) rather than more restrictive MCAR – missingness can depend on X_{i1} values
- SEs and tests based on large sample theory are available and valid
⇒ valid p -values and confidence intervals
- Provides guidance in harder problems without obvious **ad hoc** fixes
- Evidence suggests following principles works well even when model is not entirely correct

Multiple Imputation

- Impute $M \geq 2$ values for each missing value, typically, $M = 5$
- Retains advantages of single imp. and attains large sample optimality of ML

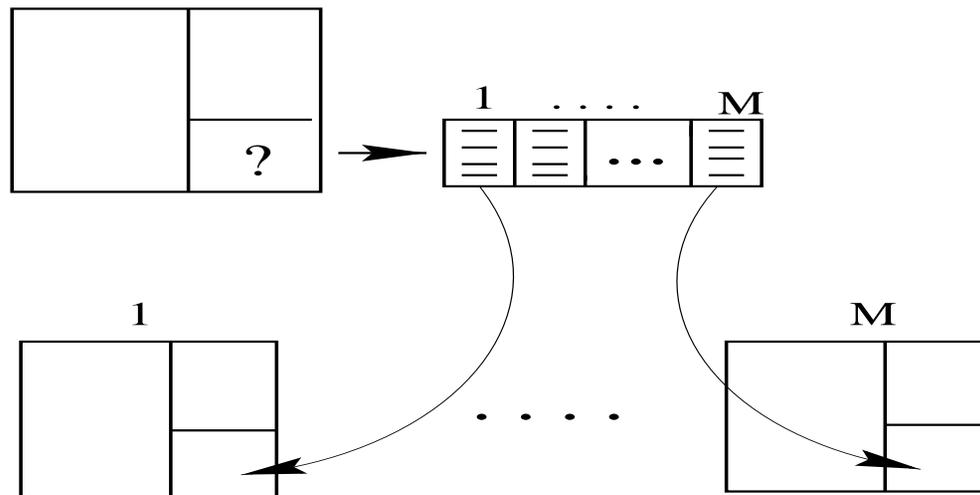


Completed data sets: ✓ X 0 ★ +

Multiple Imputation Overview

- Generate M independent imputations of the missing data under an imputation model
- Analyze each of the M completed data sets by standard complete-data methods
- Combine to get one answer across imputations
- Especially useful for database construction
(valuable data – many analyses – many potential users)

Creating Imputations in Bivariate Monotone Examples



Example 1:

$$\text{Let } \hat{X}_{i2} = \hat{\alpha}_{2.1} + \hat{\beta}_{2.1}X_{i1} \quad i = n_{\text{obs}} + 1, \dots, n$$

Then impute for $i = n_{\text{obs}} + 1, \dots, n$:

$$\tilde{X}_{i2}^{(\ell)} = \hat{X}_{i2} + r_i^{(\ell)} \quad \ell = 1, \dots, M$$

where $r_i^{(\ell)}$ are residuals drawn from complete case regressions

Creating Imputations in Bivariate Monotone Examples

Example 2: (better method)

Important adjustments when % missing info. \uparrow

$$\tilde{X}_{i2}^{(\ell)} = \hat{\alpha}_{2.1}^{(\ell)} + \hat{\beta}_{2.1}^{(\ell)} X_{i1} + r_i^{(\ell)}$$

(reflect sampling variability in L.S. regression)

Under normality

draw $\hat{\sigma}_{22.1}^{(\ell)}$ using $\chi_{n_{\text{obs}}-2}^2$

draw $\hat{\beta}_{2.1}^{(\ell)} \sim N(\hat{\beta}_{2.1}, \hat{\sigma}_{22.1}^{(\ell)} / (n_{\text{obs}} s_1^2))$

draw $\hat{\alpha}_{2.1}^{(\ell)} \sim N(\bar{X}_2 - \hat{\beta}_{2.1}^{(\ell)} \bar{X}_1, \hat{\sigma}_{22.1}^{(\ell)} / n_{\text{obs}})$

Estimation with a Multiply-Imputed Data Set

Complete-data statistics for μ_2 : Average of all n X_{i2} observations = \bar{X}_2
Associated standard error = s_2/\sqrt{n}

M values across M completed data sets: $\bar{X}_2^{(\ell)}, SE^{(\ell)}, \ell = 1, \dots, M$

MI Estimate:

$$\hat{\mu}_2^* = \text{ave}_\ell \left[\bar{X}_2^{(\ell)} \right]$$

MI Standard Error:

$$SE^* = \left[U + \left(1 + \frac{1}{M}\right)B \right]^{1/2}$$

$$\text{Within Imputation Var.} \quad U = \text{ave}_\ell \left[SE^{(\ell)^2} \right]$$

$$\text{Between Imputation Var.} \quad B = \text{var}_\ell \left[\bar{X}_2^{(\ell)} \right]$$

Remarks about the Example and MI

- In the bivariate monotone example, as $M \rightarrow \infty$
 - $\hat{\mu}_2^* \rightarrow \hat{\mu}_2$, the MLE
 - $\text{SE}^* \rightarrow \text{Var}(\hat{\mu}_2 - \mu_2)^{1/2}$
- M small (2 or 3) gives valid inferences for modest fractions of missing data ($M = 5, 10$ most common and acceptable)
- **Creating a multiply-imputed data set in general is difficult, although no more difficult than creating a singly-imputed data set from which consistent estimation is possible**
- Imputations should be generated to reflect the uncertainty about the missing data (including uncertainty about the unknown model parameters)

Key Ideas of Multiple Imputation

MI doesn't create information, but represents observed information so that it can be extracted using standard complete-data methods of analysis

- Each missing value is replaced by M possible values = multiple imputations
- These represent distribution given observed data
- Multiple imputations “create” M completed data sets
- Each completed data set analyzed by standard complete-data software
- Outputs from M complete-data analyses combined to create final inference (e.g., average M regression coefficients)
- $M = 5$ is sufficient in most applications

MI vs. Maximum Likelihood

- **Flexibility:** Incorporation of auxiliary information
- **Consistency:** Same imputations for a variety of analyses
- **Robustness:** Model misspecification
- **Sensitivity:** Compare imputation models
- MI asymptotically equivalent to ML under the same model, but subasymptotically preferable (expectation vs. modal estimation)
- MI separates the tasks of handling the missing data and analyzing the complete(d) data
 - With MI have to worry about the congeniality of the imputation and analysis models

Multiple Imputation Theory: Bayesian

- Complete data $Y = (Y_{\text{obs}}, Y_{\text{mis}})$
- Quantity of interest is Q

$$P(Q | Y_{\text{obs}}) = \int \underbrace{P(Q | Y)}_{\text{correct posterior distn of } Q} \underbrace{P(Y_{\text{mis}} | Y_{\text{obs}})}_{\text{complete data posterior distn of } Q} dY_{\text{mis}}$$

correct posterior distn of Q
complete data posterior distn of Q
posterior predictive distn of missing values simulated by multiple imputations

$$E(Q | Y_{\text{obs}}) = E[E(Q | Y_{\text{com}}) | Y_{\text{obs}}] \approx \text{Ave}(\hat{Q})$$

$$\begin{aligned}
 V(Q | Y_{\text{obs}}) &= E[V(Q | Y_{\text{com}}) | Y_{\text{obs}}] + V[E(Q | Y_{\text{com}}) | Y_{\text{obs}}] \\
 &\approx \text{Ave}(U) + \frac{m+1}{m} V(\hat{Q})
 \end{aligned}$$

“Contest” Example

Objectives of the simulation study

- Compare fully Bayesian approach to the regression based method that ignores the uncertainty about the parameter estimates
- Investigate whether or not “expenditures” should be used in the “income” imputation especially when an econometric analysis involves regressing income on expenditure

Simulation Setup

- Population: complete reporters over several years
- 200 independent samples drawn each size 500 (approximately)
- An ignorable monotone missing data mechanism was imposed on each sample by BLS staff
- MI team was unaware of the exact nature of the mechanism
- All the variables used in the mechanism as well as some other variables were provided to the contractor without noting which were which
- 200 simulated data sets with missing values were shipped to the MI team
- $M = 5$ imputations created for each sample and sent to BLS for evaluation

Variables

- Variables with missing values
 - (R) ethnic: race/ethnicity variable with 4 categories
1=black 2=Hispanic 3=white 4=other
 - (Y) lincome: $\log(\text{income}+1)$
- Covariates with no missing values
 - X =(mortgag1, mortgag2, earnrel2, earnrel3, earnrel4, earnrel5, noearnr, occup2, occup3, occup4, occup5, occup6, educ2, educ3, educ4, family2, family3, family4, family5, perslt18, famsize, region2, region3, region4, urban1, homeon1, incweekq, inchrsq, ageref)
 - E =(lzalbev, lzapprl, lzcarnw, lzcarus, lzeduct, lzentan, lzfooda, lzfoodh, lzgasm, lzhltin, lzhouf, lzhouso, lzmedsr, lzmiscx, lzowndw, lzpercr, lzperln, lzpredg, lzpubtr, lzreadn, lzrentd, lztobac, lztrans, lztrntp, lzutils, lzothld)

Four Kinds of Imputations Studied

1. (FB, include E): draws from the joint predictive distribution of (R, Y) given (X, E) (this model supported by MI theory)
 - $[R, Y|X, E] = [R|X, E][Y|R, X, E]$
 - $[R|X, E] \sim$ Multinomial Logit
 - $[Y|R, X, E]$ Normal linear regression model
 - Draw parameters from their posterior distribution and then draw multiple values (imputations) from the multinomial logit or normal linear regression model conditional on the drawn value of the parameters
2. (FB, exclude E): same as above except that E is not included as a covariate in the model for imputation
3. (Fixed Parameter, include E): here the approach is similar to 1 except that the parameters are not perturbed but are fixed at their maximum likelihood estimates
4. (Fixed Parameter, exclude E): same as (3) except that E is not included in the model for imputation

Complete-Data Analysis

$$\log(INCOME + 1) = \alpha_o + \alpha_1 \times \log(TEXP + 1) + \sum_j \alpha_j X_j$$

where

$TEXP$ = Total Expenditure

X_1 = Age of the reference person

X_2 = 1 if the reference person has completed High School but did not go to college, and 0 otherwise

X_3 = 1 if the reference person has attended some college but did not finish college, and 0 otherwise

X_4 = 1 if the reference person completed college, and 0 otherwise

X_2 , X_3 and X_4 represent education status with those with less than high school education serving as the reference category

X_5 = 1 if the reference person is Black, and 0 otherwise

X_6 = 1 if the reference person is Hispanic, and 0 otherwise

X_7 = 1 if the reference person is "Other", and 0 otherwise

X_5 , X_6 and X_7 represent ethnicity with White forming the reference category

Primary parameter of interest: α_1

Coverage Properties

- For each method computed the proportion of 200 nominal confidence intervals containing the true value of α_1

METHOD	90% NOMINAL	95% NOMINAL
FB,INCLUDE E	91.5	96.0
FB, EXCLUDE E	82.5	88.0
FIXED MLE, INCLUDE E	89.0	92.0
FIXED MLE, EXCLUDE E	80.0	86.5

- Paired t-test results comparing certain imputation methods
[FB, include E] VS [FIXED, include E]: t -statistic=8.841
[FB, include E] VS [FB, exclude E]: t -statistic=35.360

“Don’t Know” Survey Responses: the Slovenian Plebiscite

(Rubin, Stern, Vehovar JASA 1995)

- The Republic of Slovenia separated from Yugoslavia on Oct. 8, 1991
- One year before, Slovenians voted on independence through a plebiscite
- 88.5% of eligible voters voted in favor of independence
- To predict the results of the plebiscite, the Slovenian government inserted questions into the Slovenian Public Opinion (SPO) Survey that was conducted 4 weeks prior to the plebiscite

Slovenian Public Opinion Survey

- Questions on many aspects of Slovenian life, e.g. education, health, etc.
- Conducted 1-2 times each year for the past 30 years
- Face-to-face survey of approx. 2,000 voting-age Slovenians
- The independence questions were:
 1. Are you in favor of Slovenian independence?
 2. Are you in favor of Slovenia's secession from Yugoslavia?
 3. Will you attend the plebiscite?

Possible responses were YES, NO, and Don't Know (DK)
(in the plebiscite non-voters are treated as NO votes)

Secession	Attendance	Independence		
		Yes	No	DK
Yes	Yes	1,191	8	21
	No	8	0	4
	DK	107	3	9
No	Yes	158	68	29
	No	7	14	3
	DK	18	43	31
DK	Yes	90	2	109
	No	1	2	25
	DK	19	8	96

$n=2,074$

θ = proportion voting Yes and planning to attend

Simple estimates of θ :

method	$\hat{\theta}$	n
Complete Cases	.928	1,454
Available Cases	.929	1,549
Conservative (DK = No)	.694	2,074

Thinking more carefully about the DK responses

- In some surveys DK is a valid response
- **But** in the plebiscite, everyone votes either Yes or No
(either by voting in person or by staying home effectively voting No)
- We can treat DK's as missing data

We can view our data as partially cross classified:
(looking at questions 1 and 3)

		Independence		DK
		Yes	No	
Attend	Yes	1,439	78	159
	No	16	16	32
DK		144	54	136

Comparing Results

Assuming MAR we can model the 4 cells of interest as multinomial:

(i.e. assuming the probability of the occurrence of DK responses can depend on the observed answers to other questions, but given these, it does not depend on the missing value itself)

Estimate of θ using Multiple Imputation (or EM): $\hat{\theta} = .883$

Can also assume nonignorable nonresponse:

(i.e. can assume the probability of an observation's missingness depends on the unobserved value)

Plausible since eventual No voters might be more likely to say DK to avoid giving an unpopular answer.

Can use loglinear models to fit a plausible nonignorable model: $\hat{\theta} = .782$
(these models don't fit the data well)

Summary of Slovenian Plebiscite Data

- Need to treat DK's as missing data (DK's "hide" a Yes or No)
- Analyzing the data under the MAR assumption provides accurate predictions

Some MI References

Web

<http://www.multiple-imputation.com>

Books

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis With Missing Data*, New York: Wiley, 2nd edn.

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

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, New York: Chapman and Hall.

