Multiple Imputation in Practice

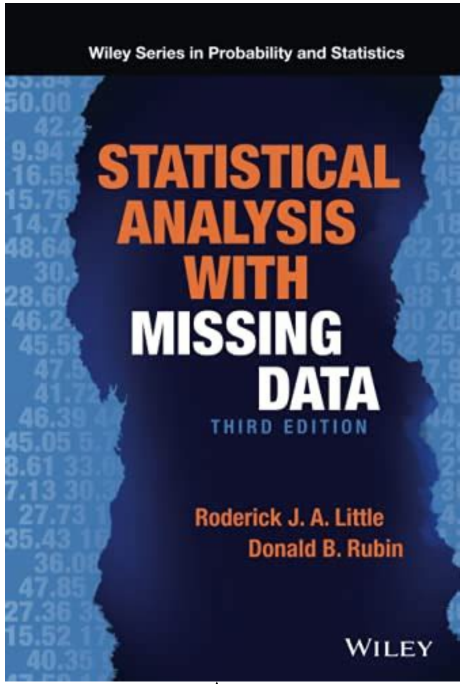
Trivellore Raghunathan

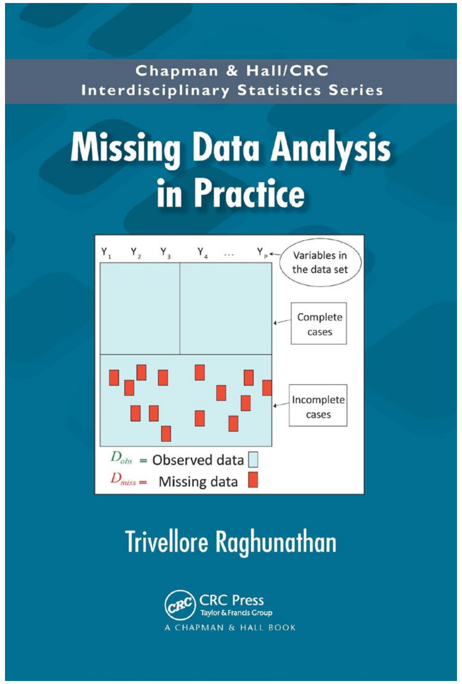
Part 1: Lecture Slides

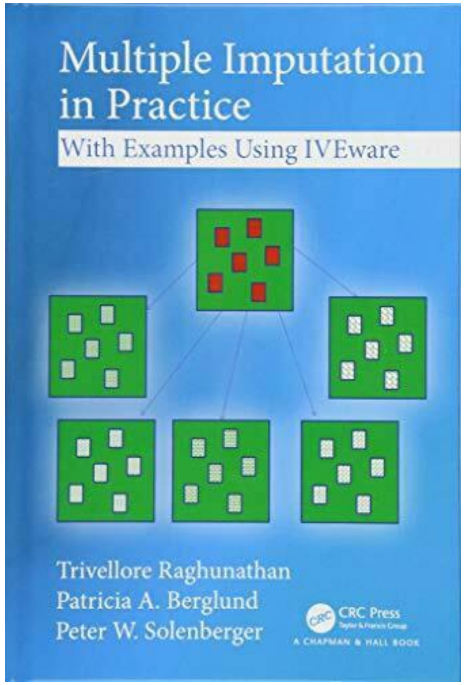
(Slides adapted from course slides jointly taught with Rod Little)

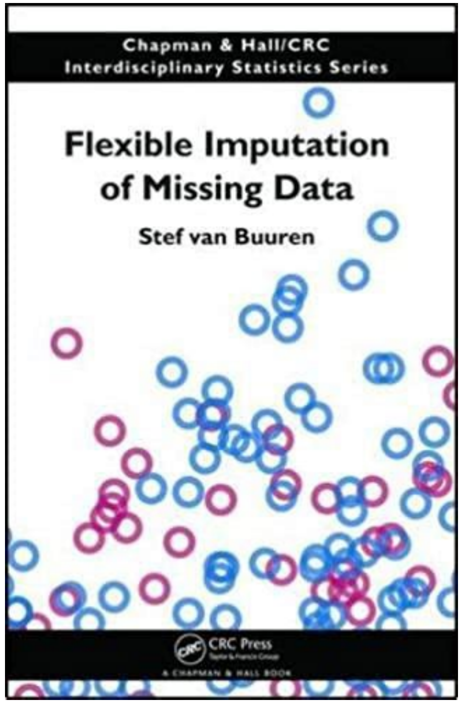
UNIVERSITY OF MICHIGAN

BOOKS









Practice oriented books with lots of examples and codes.

Comprehensive Book on Missing Data. Covers a wide variety of topics

Introduction

Example of missing data in surveys

- National Health and Nutrition Examination Survey (NHANES)
- Public Use Files subject to:
  - – Unit nonresponse
    - noncontact
    - refusal
  - – Item nonresponse
    - questionnaire interview complete, health examination missing
    - Individual items – “swiss cheese pattern”

Unit nonresponse

- Unit nonrespondents may differ from respondents, leading to
  - **nonignorable** missing data
  - biased estimates. A simple formula for means:

$$\bar{Y}_R - \bar{Y} = \quad \pi_{NR} \quad \times \quad (\bar{Y}_R - \bar{Y}_{NR})$$

Bias = NR rate * difference in R and NR means

NR = nonrespondent, R = respondent

Sampling unit nonrespondents

- One approach is to follow up a subsample of nonrespondents with special efforts:
  - abbreviated interview
  - monetary incentives
- Data collected can be
  - weighted to represent all nonrespondents
  - used to (multiply) impute other nonrespondents

Item Nonresponse

- Generally leads to general “swiss-cheese” pattern of missing data
- Weighting does not work well for this pattern
- Imputation is the usual approach
- Multiple imputation propagates imputation uncertainty

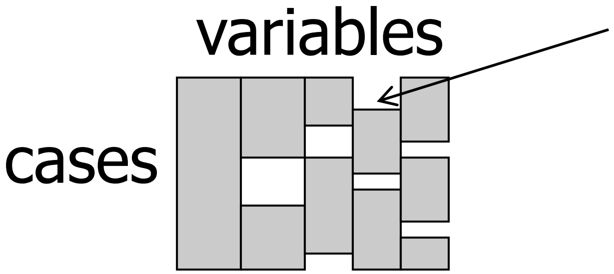
Example: Attrition in longitudinal surveys

- Longitudinal studies often have drop-outs
  – Move out of study catchment area
  – Participation becomes too onerous
  – Can be viewed as item nonresponse for concatenated file of the repeated surveys
- Common analyses have problems:
  – complete case analysis is biased if drop-outs differ
  – Naïve imputation (e.g. mean imputation, last observation carried forward) involves unrealistic assumptions
- We discuss better alternatives

Other problems formulated as missing data

- Finite population inference: nonsampled units are "missing"
- Response errors: true and measured variables, where true are missing or only observed for a calibration sample
- Disclosure limitation: replace some values by imputations to reduce disclosure risk
- Inference for causal effects

Missing data defined



If filled in, meaningful for analysis?

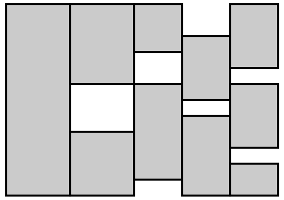
- Always assume missingness hides a meaningful value for analysis
- Examples:
  - Missing data from missed clinical visit(√)
  - Nonresponse in an election opinion poll (?)
  - In a longitudinal study of blood pressure medications:
    - losses to follow-up (√)
    - deaths (x)

Patterns of Missing Data

- Some methods work for a general pattern

variables

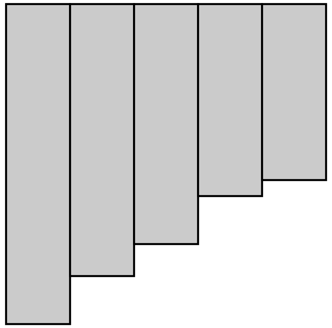
cases



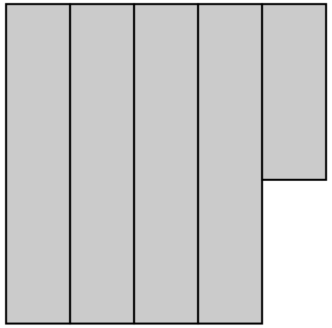
e.g. multiple imputation for item nonresponse

- Other methods apply only to special patterns

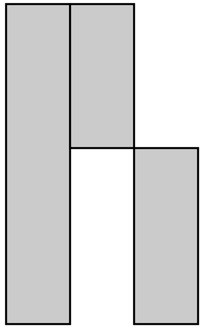
monotone



univariate



file matching



Pattern versus mechanism

- Pattern: Which values are missing?
- Mechanism: Why? Reasons related to the study variables?
  - Formalize by model for mechanism

    Y = data matrix, if no data were missing

    M = missing-data indicator matrix

    (i,j) th element indicates whether (i,j) th element of Y is missing (1) or observed (0)
  - Pattern concerns distribution of M
  - Mechanism concerns distribution of M given Y

More on mechanisms

- Data are:
  - missing completely at random (MCAR) if missingness independent of Y:

    $$p(M \mid Y) = p(M) \text{ for all } Y$$

  - missing at random (MAR) if missingness only depends on observed components Y_{obs} of Y:

    $$p(M \mid Y) = p(M \mid Y_{\text{obs}}) \text{ for all } Y$$

  - missing not at random (MNAR) if missingness depends on missing (as well as perhaps on observed) components of Y

MAR for univariate nonresponse

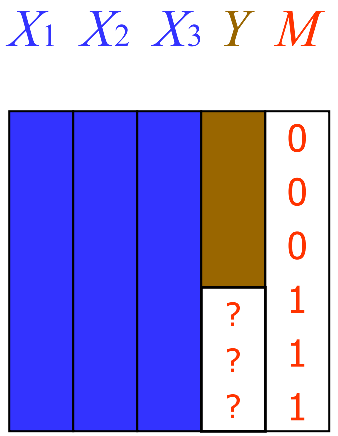
X_j = complete covariates

Y = incomplete variable

M = 1, Y missing

0, Y observed

$R = I - M$



MAR: missingness independent of Y given $X_1 \ldots X_k$

That is, M can depend on X's …

but not on Y given X's

MAR for monotone missing data

MAR if dropout depends on values recorded prior to drop-out

MNAR if dropout depends on values that are missing (that is, after drop-out)

Censoring by end of study: plausibly MCAR

Drop-out from panel study because moved to Arizona: MAR if reasons for moving (e.g. age) captured as covariates. Maybe MNAR for health survey

A non-monotone example

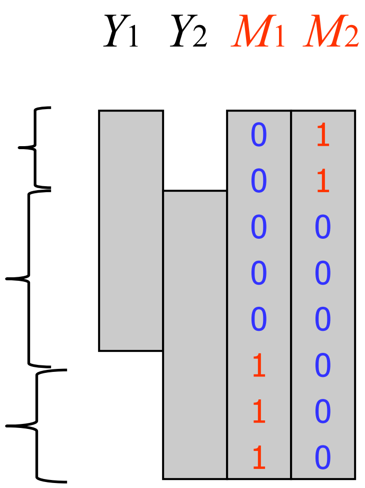
Mechanism is MAR if

$\Pr(Y_2 \text{ missing}) = g(Y_1)$

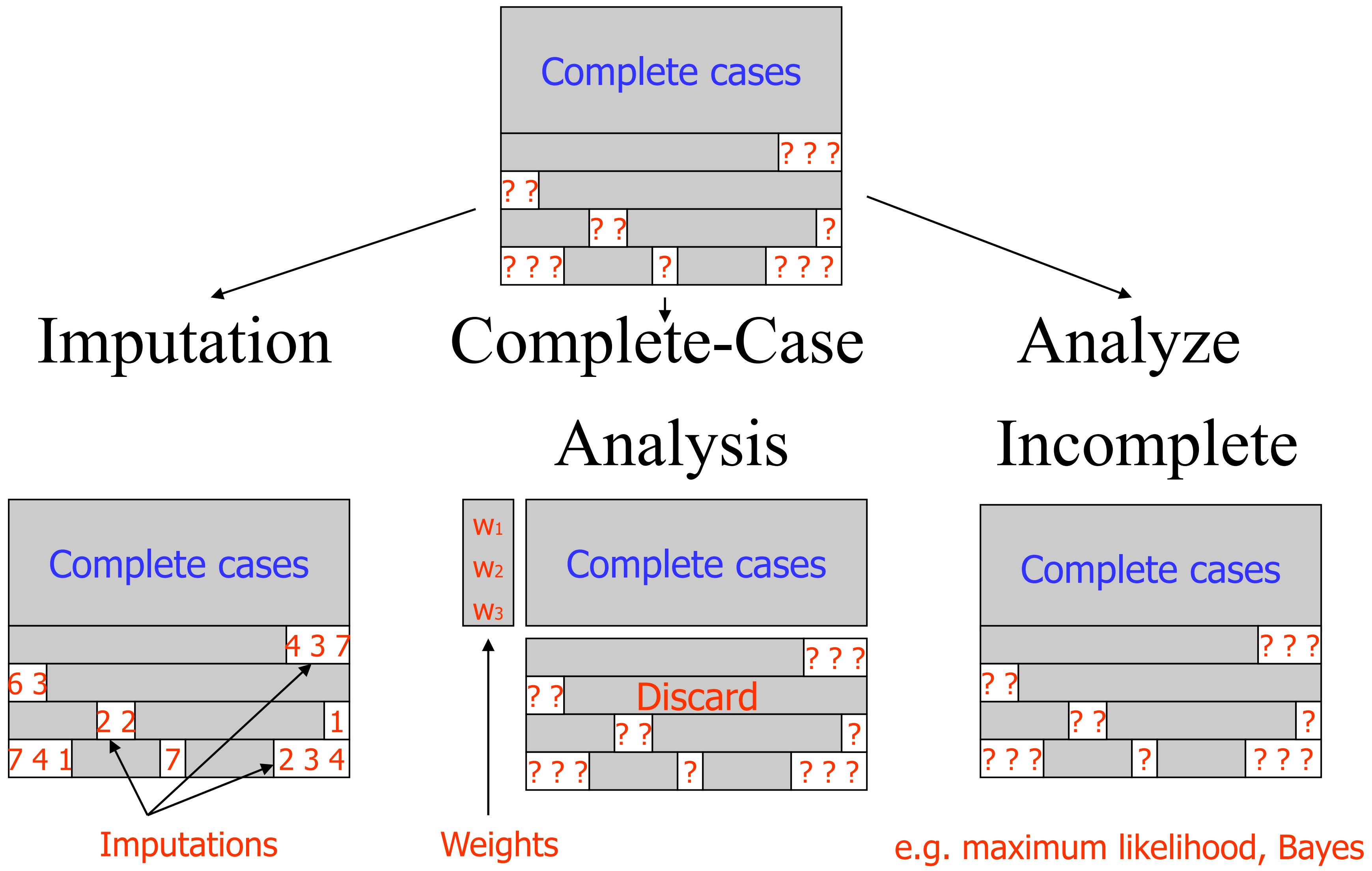
$\Pr(Y_1 \text{ missing}) = f(Y_2)$

$\Pr(\text{complete}) =$

$1 - f(Y_2) - g(Y_1)$



General Strategies

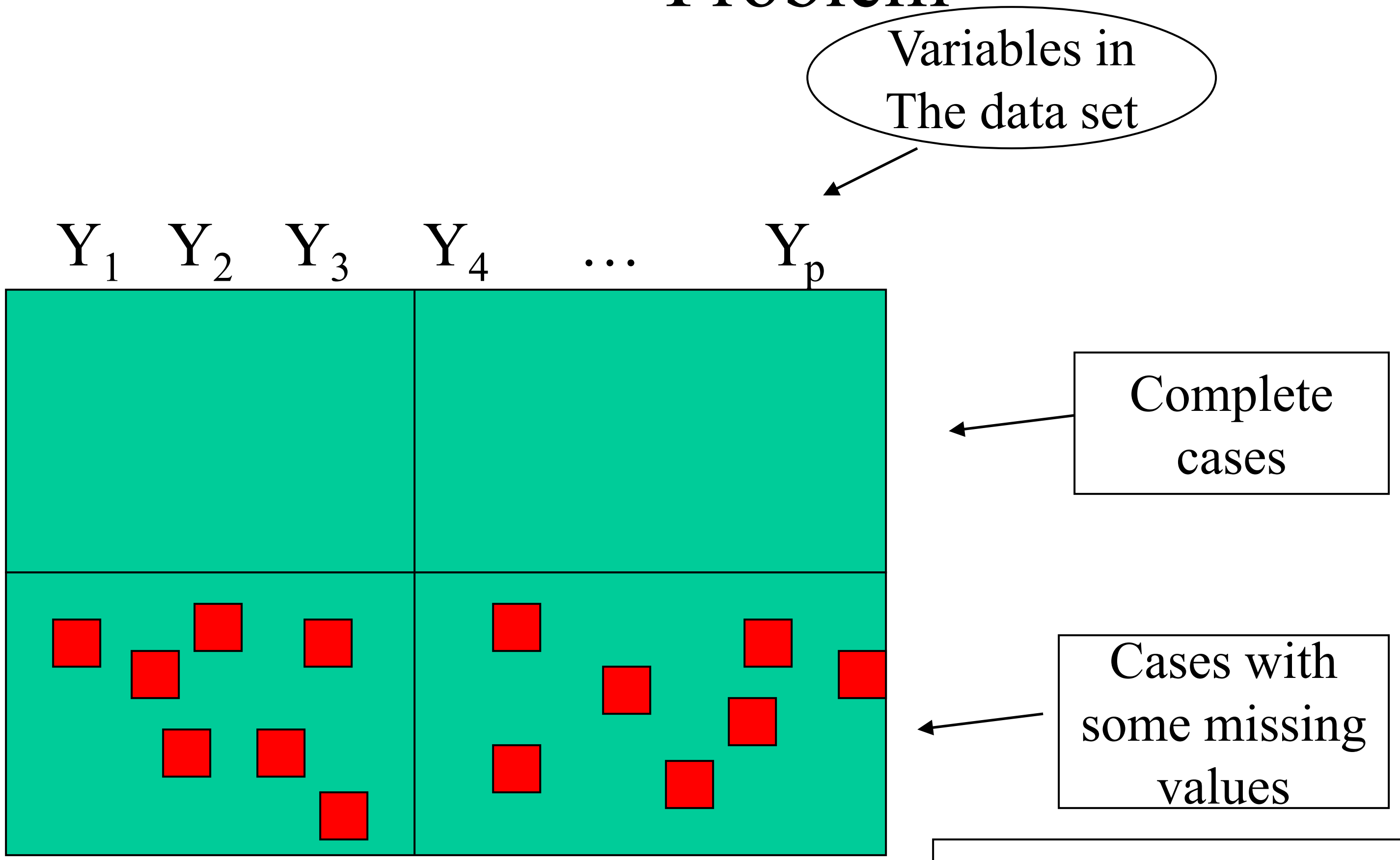


Properties of a good missing-data method

- Makes use of partial information on incomplete cases, for reduced bias, increased efficiency
- Frequency valid (“calibrated”) inferences under plausible model for missing data (e.g. confidence intervals have nominal coverage)
- Propagates missing-data uncertainty, both within and between imputation models
- Favor likelihood based approaches
  - Maximum Likelihood (ML) for large samples
  - Multiple Imputation/Bayes for small samples

Imputation Methods

Problem



D_{obs} = Observed data: ■

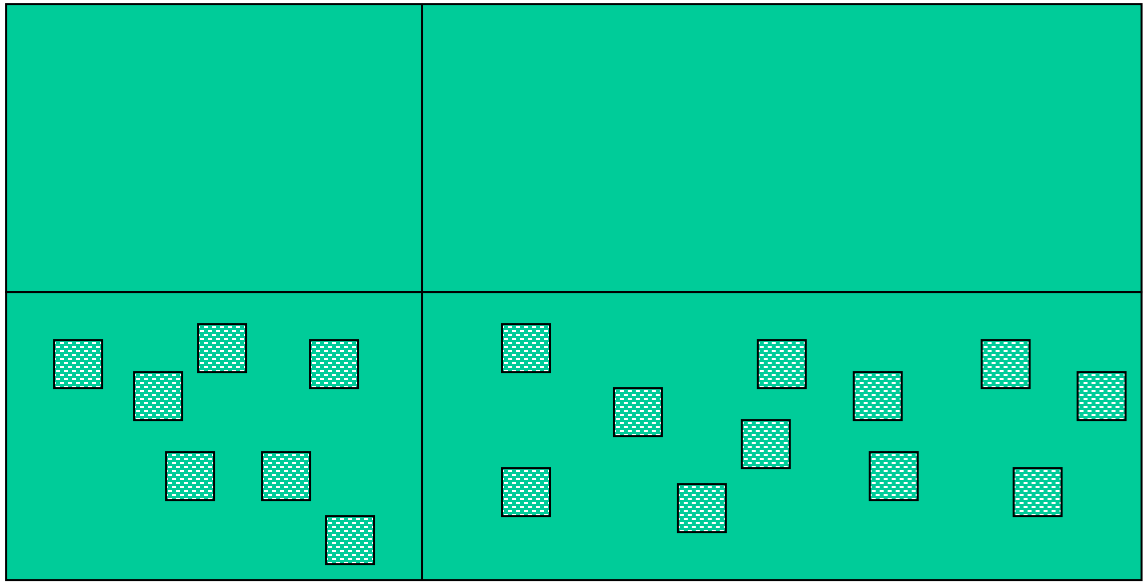
D_{mis} = Missing data: ■

Y: Discrete, continuous or semi-continuous as well as multivariate

Setting

- Multiple users analyzing different subsets of variables
- Multiple analytical techniques
- Different skill levels dealing with incomplete data
- Analysis to be performed with complete data is known
- Software to perform complete data analysis is available
- Assume missing at random.
  - That is conditional on the observed characteristics the residual differences between those with missing and those with no missing values are random

Imputation



Imputation:

Draws from $\Pr(D_{\text{mis}} \mid D_{\text{obs}})$

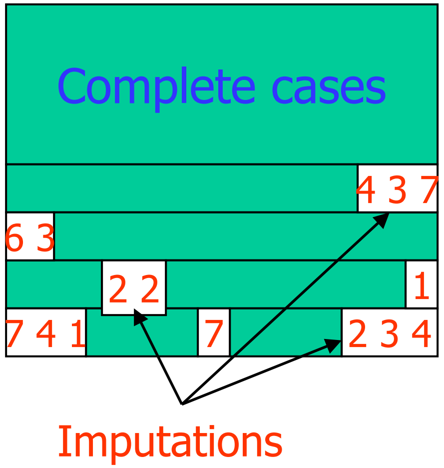
Important issues:

Goal is good inferences for parameters or population quantities, <u>not</u> best estimates of missing values

Imputations are not real values --

Uncertainty associated with imputes should be taken into account

Features of Imputation



Good

Rectangular File

Retains observed data

Handles missing data once

Exploits incomplete cases

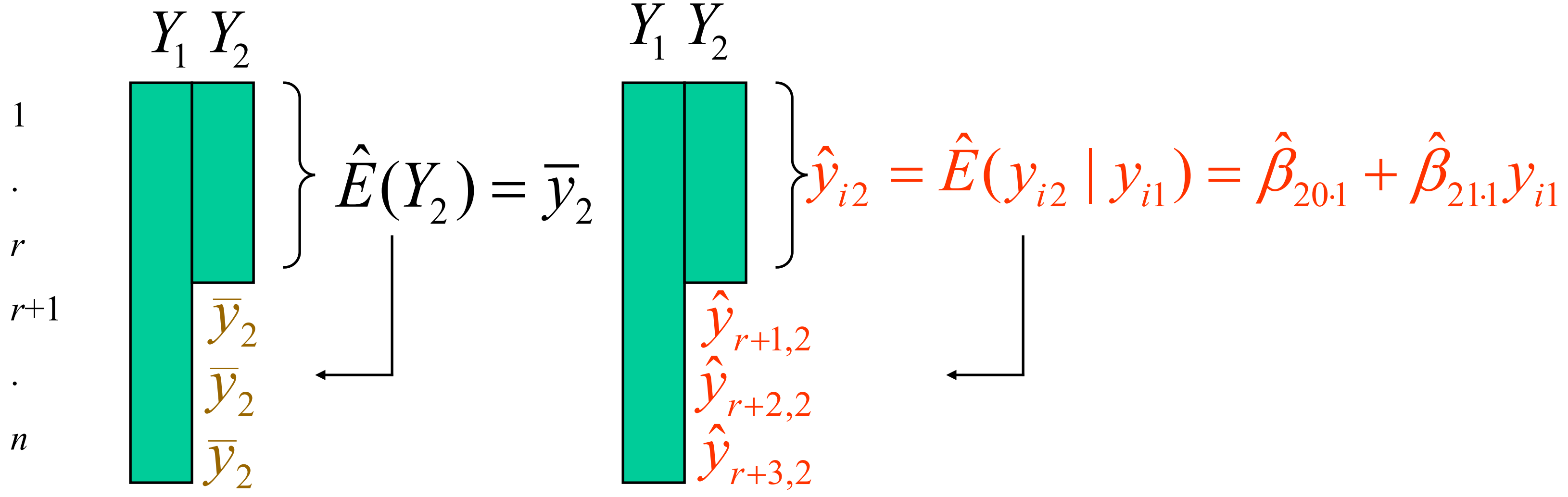
Bad

Naïve methods can be bad

Invents data –
Understates uncertainty

Imputing Means

Unconditional — Conditional on observed variables

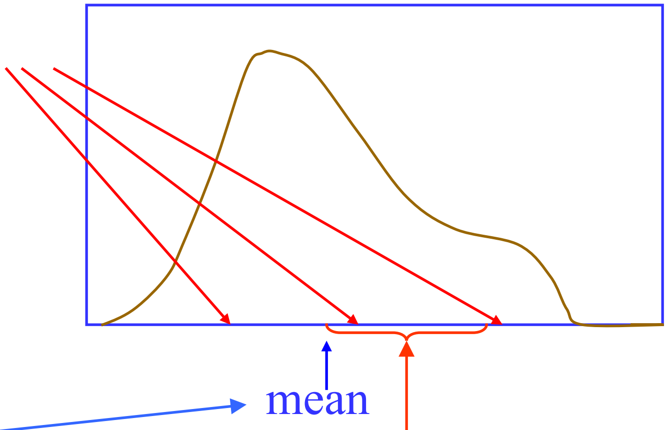


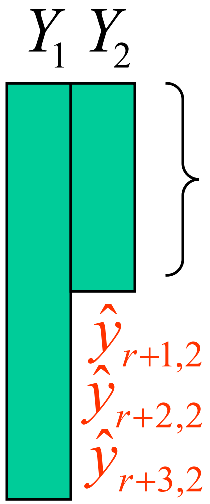
Properties of Mean Imputation

- Marginal distributions, associations estimated from filled-in data are distorted
- Standard errors of estimates from filled-in data are too small, since
  - Standard deviations are underestimated
  - “Sample size” is overstated
- Conditional better than unconditional mean, which can be worse than complete cases

Imputing Draws

- Imputations can be random draws from a predictive distribution for the missing values





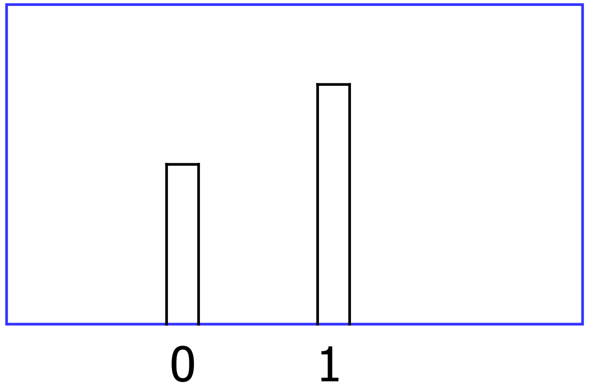
$$\hat{y}_{i2} = \hat{E}(y_{i2} \mid y_{i1}) + r_i$$

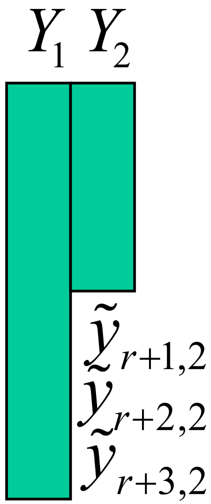
$$r_i \sim N(0, s_{22\cdot 1}),\ s_{22\cdot 1} = \text{resid variance, or}$$

$$r_i = \text{ residual from randomly selected complete case}$$

Imputing draws for binary data

- For binary (0-1) data, impute 1 with probability $\hat{p}_{i2}$ = predicted prob of a one, given observed covariates





$$\hat{p}_{i2} = \Pr(y_{i2} = 1 \mid y_{i1}) \text{(e.g. logistic regression)}$$

$$\tilde{y}_{i2} = \begin{cases} 1, \text{prob } \hat{p}_{i2} \\ 0, \text{prob } 1 - \hat{p}_{i2} \end{cases}$$

Properties of Imputed Draws

- Adds noise, less efficient than imputing means, but:
- No (or reduced) bias for estimating distributions
- More robust to nonlinear data transformations
- Conditional draws better than unconditional:
  - Improved efficiency
  - Preserves associations with conditioned variables
- Standard errors from filled-in data are improved, but still wrong:
  - Standard deviation is ok
  - "Sample size" overstated; multiple imputation fixes this

Impute conditional means or draws?

- Impute conditional means to get best estimates of missing values
  - Common in machine learning literature
- Impute conditional draws to get valid inference for parameters
  - Valid standard errors can be obtained by multiple imputation (discussed later)
  - The focus of this course

Missing covariates in regression

- What should imputes condition on?
  - Observed covariates and outcome, if imputing draws
  - Observed covariates only, if imputing means
- Imputing conditional means can be less efficient than complete case analysis, unless imputed cases are down-weighted
  - For details, see Little (1992)
- Standard errors from filled-in data are understated

Example : Should Imputations be conditional on all observed variables?

- Consumer Expenditure Survey (Bureau of Labor Statistics)
- Should the imputation of Income be conditional on Expenditure variables?
- Substantive models of interest are relationship between income and expenditure

BLS Simulation Example

- BLS researchers:
  - created population by accumulating complete cases over several years
  - drew 200 random samples of size 500 each (Before deletion data sets)
  - created missing data on income in each data set
  - supplied 200 data sets along with 55 covariates to University of Michigan

BLS Example (Continued)

- UM did not know how Income values were deleted (except that some or all of 55 covariates were used in specifying missing data mechanism)
- UM created two sets of imputations

  Using Expenditure

  Not Using Expenditure

BLS Imputations

- Imputations were created by drawing values from the posterior predictive distribution of income under an explicit model
- One included expenditure as a conditioning variable and other did not
- Two sets of imputed data sets and actual data sets were analyzed by UM and BLS respectively.

BLS Models of Interest

- OLS model

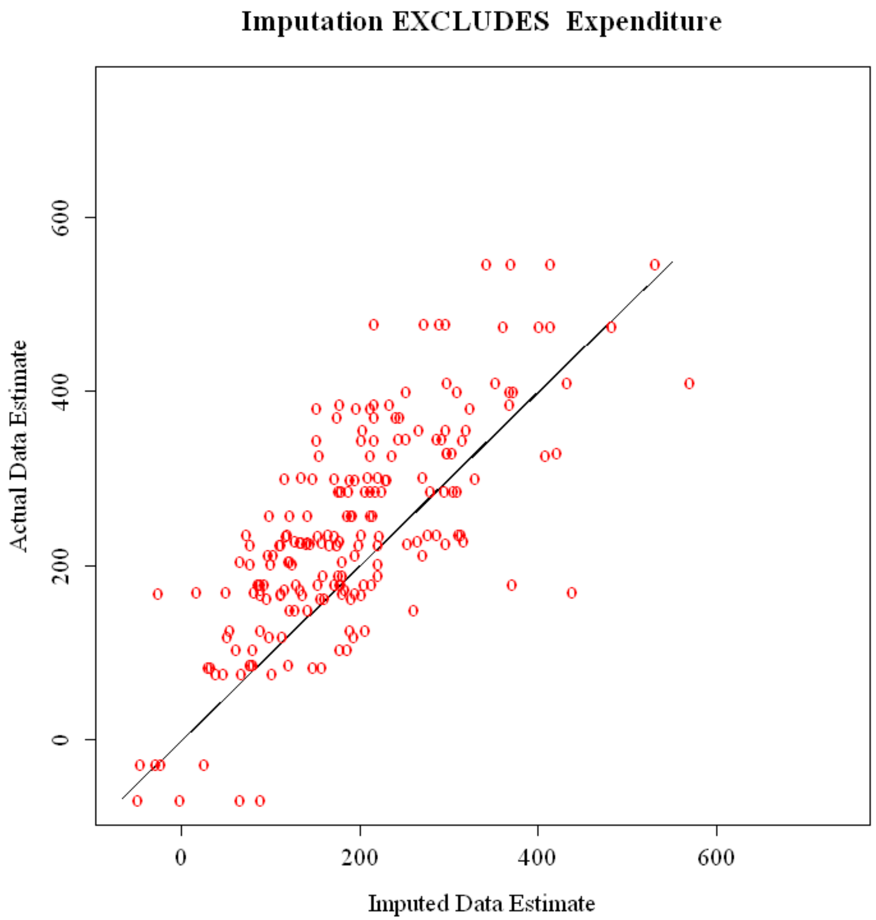
  Food-At-Home$=\beta_0+\beta_1$ *Income + covariates*

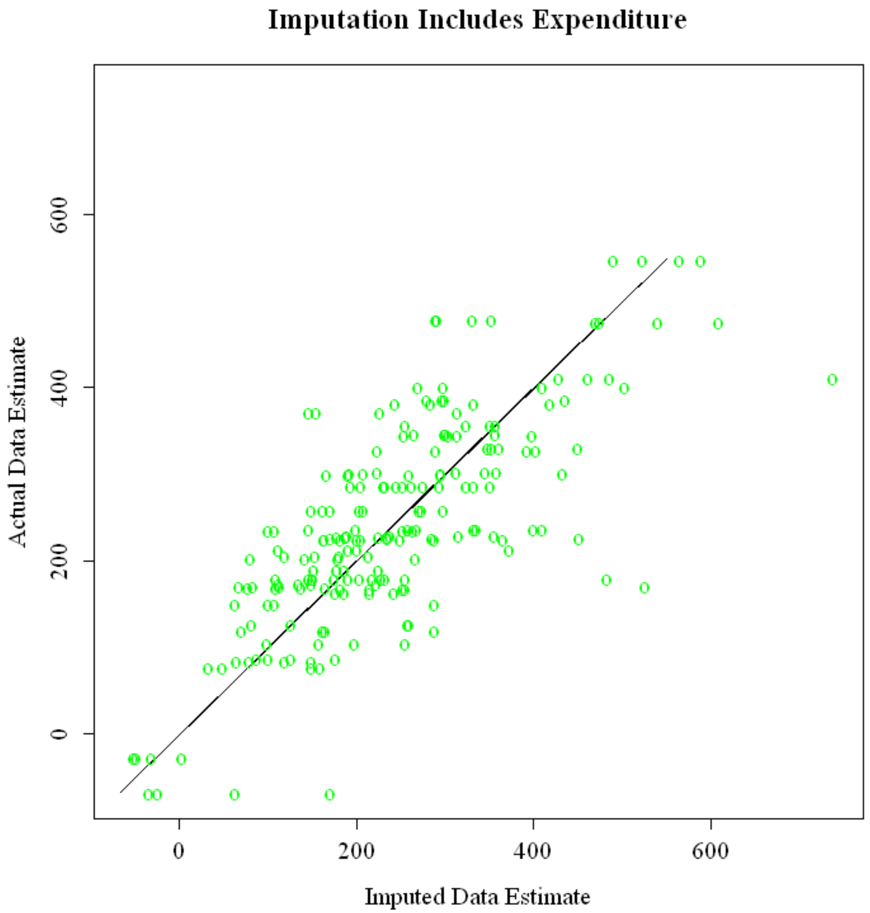
- Tobit Model

  Food-Away-Home$=\gamma_0+\gamma_1$ *Income + covariates*

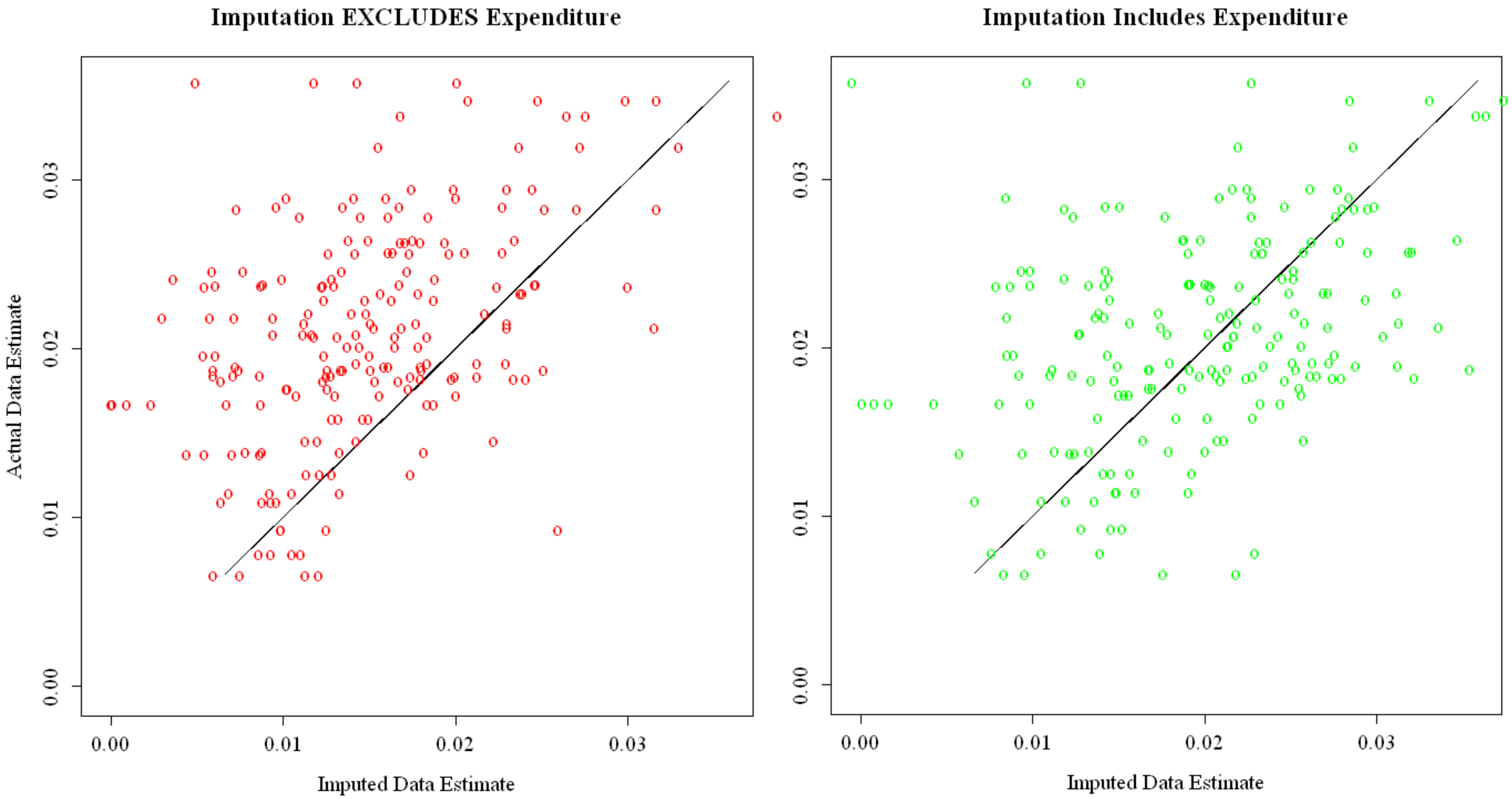
  Left Censored Values

Estimated regression coefficients of income from undeleted and imputed data-sets: OLS Model





Estimated regression coefficients of income from undeleted and imputed data-sets: Tobit Model



What should imputes condition on?

- In principle, all observed variables
  - Whether predictors or outcomes of final analysis model
  - May be impractical with a lot of variables
- Variable selection
  - Similar ideas to weighting adjustments apply
  - Priority to variables predictive of missing variable (and nonresponse)
  - Favor inclusion over exclusion (more later)

Creating the predictive distribution

All imputation methods assume a model for the predictive distribution of the missing values

- *Explicit*: predictive distribution based on a formal statistical model (e.g. multivariate normal); assumptions are explicit
- *Implicit*: focus is on an algorithm, but the algorithm implies an underlying model; assumptions are implicit

Implicit modeling procedure

- Hot deck imputation
- Classify respondents, nonrespondents into adjustment cells with similar observed values
  - impute values from random respondent in same cell
  - implicit model: regression of missing variables on variables forming cells, including all interactions

Current Population Survey Hot Deck

- Missing (Y): Earnings Variables
- Observed (X):
  - Age, Race, Sex, Family Relationship, Children, Marital Status, Occupation, Schooling, Full/Part time, Type of Residence, Income Recipiency Pattern
- Flexible matching:
  - Joint Classification by X yields giant matrix. If a match is not found, table is coarsened or collapsed in stages until a match is found

CPS Hot Deck (continued)

Good Features

- Imputes real values
- multivariate: associations preserved
- Conditions on *X*'s
- Assessments suggest method works quite well with large data sets

Bad Features

- Does not exploit previous earnings models
- Includes high order interactions at expense of main effects of omitted *X*'s
- Imputation uncertainty not included in standard errors

For comparison of CPS Hot Deck with stochastic regression imputation see David et al. (1986)

Other matching methods

- More generally, nonrespondents *j* can be matched to respondents *i* based on a <u>closeness metric *D*(*i*, *j*)</u>
  - Adjustment cell: $D(i,j) = \begin{cases} 0, & \text{if } i, j \text{ belong to same cell} \\ 1, & \text{if } i, j \text{ belong to different cells} \end{cases}$
  - Mahalanobis: $D(i,j) = (x_i - x_j)^T S_X^{-1} (x_i - x_j)$
  - Predictive Mean: $D(i,j) = (\hat{y}_i - \hat{y}_j)^T S_{Y \cdot X}^{-1} (\hat{y}_i - \hat{y}_j)$

    $\hat{y}_i = \text{ regression prediction of } Y \text{ given } X$

    $S_{Y \cdot X} = \text{ resid covariance matrix}$

Properties of matching methods

- Imputation error not propagated in standard errors from filled-in data
- One metric irrespective of outcome -- in contrast, models tailor adjustment to individual Y's
- Predictive mean metric better than Mahalanobis metric, since more targeted to Y's.
- Robust to model misspecification, but needs large samples: poor matches when sample is thin

See Little (1988 JBES) for more discussion

Practical Issues

- Hot deck imputation is limited
  - Variables have to be completely observed
  - Continuous variables have to be categorized
- Explicit Model is difficult
  - Large number of variables of different types
  - Restrictions
    - Question is valid only for certain subjects
    - Skip pattern
  - Bounds
    - Variables are bounded. *Example: Years smoked cannot exceed Age for current smokers and (Age-Years since Quit smoking) for former smokers. It can become more complex, if a question about teen age smoking was asked and age when started smoking was also asked*
    - Bracketed responses

Sequential Regression/Chained Equation/Flexible Conditional Specification Approach

Variables With Missing Values:

$$Y_1, Y_2, \cdots, Y_p$$

Variables With No Missing Values: U

Each step involves draws from the predictive distribution

Iteration 1:

$$Y_1 \mid U$$

$$Y_2 \mid Y_1^{(1)}, U$$

$$\vdots$$

$$Y_j \mid U, Y_1^{(1)}, \cdots, Y_{j-1}^{(1)}$$

$$\vdots$$

$$Y_p \mid U, Y_1^{(1)}, Y_2^{(1)}, \cdots, Y_{p-1}^{(1)}$$

Iteration t=2,3,….:

$$Y_1 \mid U, Y_2^{(t-1)}, \cdots, Y_p^{(t-1)}$$

$$Y_2 \mid U, Y_1^{(t)}, Y_3^{(t-1)}, \cdots, Y_p^{(t-1)}$$

$$\vdots$$

$$Y_j \mid U, Y_1^{(t)}, \cdots, Y_{j-1}^{(t)}, Y_{j+1}^{(t-1)}, \cdots, Y_p^{(t-1)}$$

$$\vdots$$

$$Y_p \mid U, Y_1^{(t)}, \cdots, Y_{p-1}^{(t)}$$

Flexible Features

- Ability to specify individual regression model
- Types of variables
  - Continuous (Normal, Tukey's gh distribution)
  - Categorical (Logistic or generalized logistic)
  - Count (Poisson)
  - Mixed or semi-continuous (Logistic/Normal)
  - Ordinal (ordered probit)
- Non parametric procedure using Approximate Bayesian Bootstrap
- Parametric or semi-parametric regression models
- Restrictions
  - Regression model is fitted only to the relevant subset
- Bounds
  - Draws from a truncated distribution from the corresponding regression model
- Models each conditional distribution. There is no guarantee that a joint distribution exists with these conditional distributions
- How many iterations?
  - Empirical studies show that nothing much changes after 5 or 6 iterations

Summary of imputation methods

- Imputations should:
  - condition on observed variables
  - be multivariate to preserve associations between missing variables
  - generally be draws rather than means
- Key problem: single imputations do not account for imputation uncertainty in se’s. Consider next two approaches to this problem
  - bootstrapping the imputation method
  - multiple imputation

Imputation Uncertainty

Accounting for Imputation Uncertainty

- Imputation “makes up” the missing data
  - treats imputed values as the truth
- For statistical inference (standard errors, P-Values, confidence intervals) need methods that account for imputation error
  - (A) redo imputations using sample reuse methods – bootstrap, jackknife
  - (B) Multiple imputation (Rubin 1987)

Bootstrapping: with complete data

- A bootstrap sample of a complete data set S with n observations is a sample of size n drawn with replacement from S
  - Operationally, assign weight w_i to unit i equal to number of times it is included in the bootstrap sample

$$w_1,\dots,w_n \sim \mathrm{MNOM}(n; 1/n,\dots,1/n)$$

Bootstrap distribution

- Let $\hat{\theta}^{(b)}$ be a consistent parameter estimate from the *b*th bootstrap data set
- Inference can be based on the bootstrap distribution generated by values of $\hat{\theta}^{(b)}$
- In particular the bootstrap estimate is

$$\hat{\theta}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}$$

with variance

$$\hat{V}_{\text{boot}} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}_{\text{boot}})^2$$

Bootstrapping with incomplete data

- For incomplete data:
  - bootstrap the complete and incomplete cases
  - impute bootstrapped data set
  - $\hat{\theta}^{(b)}$ = consistent estimate from *b*th data set, with values imputed; then as before:

$$\hat{\theta}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)} \qquad \hat{V}_{\text{boot}} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}_{\text{boot}})^2$$

* Bootstrap then impute, <u>not</u>

* Impute then bootstrap

Imputing the bootstrap sample

- Impute so that the estimate $\hat{\theta}_b$ from imputed data is consistent. In particular:

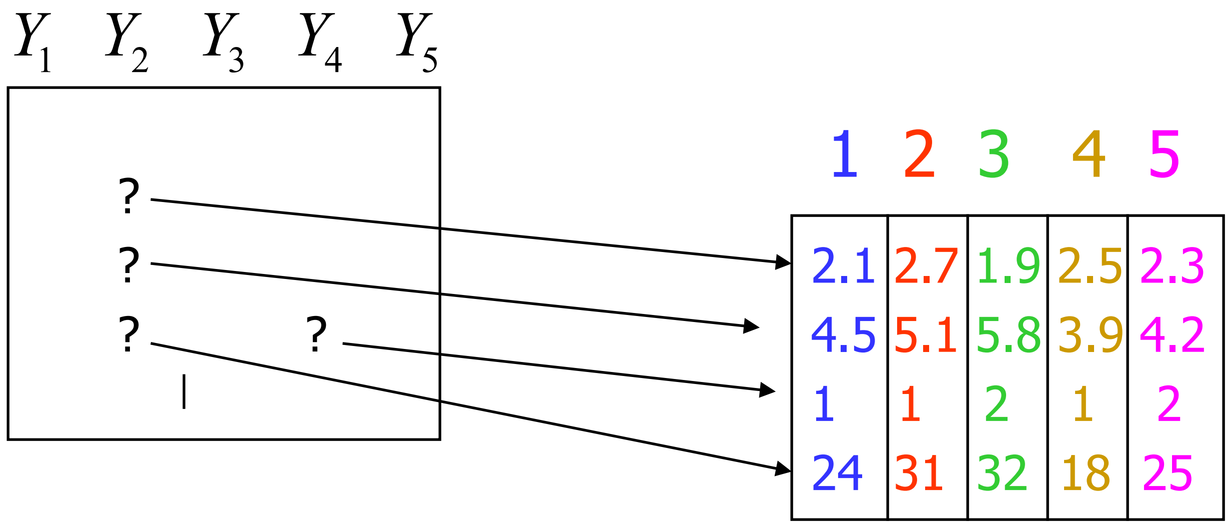
  - conditional mean ok for linear statistics
  - conditional draw ok for linear or nonlinear statistics; more general, but loss of efficiency

- Computationally intensive: imputations created for each bootstrap data set

  B=200, 1000 are typical numbers

Multiple Imputation

- Create *D* sets of imputations, each set a draw from the predictive distribution of the missing values
  - e.g. D=5



Multiple Imputation Inference

- D completed data sets (e.g. $D = 5$)
- Analyze each completed data set
- Combine results in easy way to produce multiple imputation inference
- Particularly useful for public use datasets
  – data provider creates imputes for multiple users, who can analyze data with complete-data methods

MI Inference for a Scalar Estimand

$\theta =$ estimand of interest

$\hat{\theta}_d =$ estimate from d th dataset $(d = 1, \ldots, D)$

The MI estimate of θ is $\bar{\theta}_D = \frac{1}{D}\sum_{d=1}^{D}\hat{\theta}_d$

$W_d =$ estimate of variance of $\hat{\theta}_d$ from d th dataset

The MI estimate of variance is $T_D = \bar{W}_D + (1 + 1/D)B_D$

$$\bar{W}_D = \frac{1}{D}\sum_{d=1}^{D} W_d = \text{Within-Imputation Variance}$$

$$B_D = \frac{1}{D-1}\sum_{d=1}^{D}(\hat{\theta}_d - \bar{\theta}_D)^2 = \text{Between-Imputation Variance}$$

Example of Multiple Imputation

- First imputed dataset

Y_1	Y_2	Y_3	Y_4	Y_5
	2.1			
	4.5			
	24		1	

	Estimate (se^2)	
Dataset (d)	μ_1	$\beta_{53 \cdot 1234}$
1	$12.6\ (3.6^2)$	$4.32\ (1.95^2)$

- Second imputed dataset

Y_1	Y_2	Y_3	Y_4	Y_5
	2.7			
	5.1			
	31		1	

	Estimate (se^2)	
Dataset (d)	μ_1	$\beta_{53\cdot1234}$
1	$12.6\ (3.6^2)$	$4.32\ (1.95^2)$
2	$12.6\ (3.6^2)$	$4.15\ (2.64^2)$

- Third imputed dataset

Y_1	Y_2	Y_3	Y_4	Y_5
	1.9			
	5.8			
	32		2	

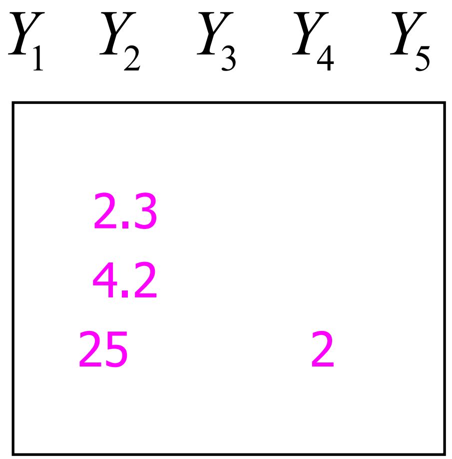
	Estimate (se^2)	
Dataset (d)	μ_1	$\beta_{53\cdot 1234}$
1	$12.6\ (3.6^2)$	$4.32\ (1.95^2)$
2	$12.6\ (3.6^2)$	$4.15\ (2.64^2)$
3	$12.6\ (3.6^2)$	$4.86\ (2.09^2)$

- Fourth imputed dataset

Y_1	Y_2	Y_3	Y_4	Y_5
	2.5			
	3.9			
	18		1	

	Estimate (se^2)	
Dataset (d)	μ_1	$\beta_{53 \cdot 1234}$
1	$12.6\ (3.6^2)$	$4.32\ (1.95^2)$
2	$12.6\ (3.6^2)$	$4.15\ (2.64^2)$
3	$12.6\ (3.6^2)$	$4.86\ (2.09^2)$
4	$12.6\ (3.6^2)$	$3.98\ (2.14^2)$

- Fifth imputed dataset



	Estimate (se^2)	
Dataset (d)	μ_1	$\beta_{53\cdot1234}$
1	$12.6\ (3.6^2)$	$4.32\ (1.95^2)$
2	$12.6\ (3.6^2)$	$4.15\ (2.64^2)$
3	$12.6\ (3.6^2)$	$4.86\ (2.09^2)$
4	$12.6\ (3.6^2)$	$3.98\ (2.14^2)$
5	$12.6\ (3.6^2)$	$4.50\ (2.47^2)$
Mean	$12.6\ (3.6^2)$	$4.36\ (2.27^2)$
Var	0	0.339

Summary of MI Inferences

	$\bar{\theta}_D$	$\bar{W}_D$	B_D	$\sqrt{T_D} = \sqrt{\bar{W}_D + \frac{6}{5} B_D}$	$\hat{\gamma}_D = 1.2 B_D / (1.2 B_D + \bar{W}_D)$
μ_1	12.6	3.6^2	0	3.6	0
$\beta_{53\cdot 1234}$	4.36	2.27^2	0.339	2.36	0.073

$$\hat{\gamma}_D = \frac{(1+1/D)B_D}{(1+1/D)B_D + \bar{W}_D} = \text{ estimated fraction of missing information}$$

- Confidence Interval $\hat{\theta}_D \pm t_{\nu,1-\alpha/2}\sqrt{T_D}$

$$\nu = (D-1)/\hat{\gamma}_D^2$$

MI for Complex Sample Design

MI for Complex Sample Design

- Survey Designs involve stratification, Clustering and Weighting
- Ignoring the survey design variables may introduce bias if they are correlated with substantive variables in the survey
- One option is to use the survey design variables as predictors
  - Impute each stratum separately or include strata as dummy variables
  - Use Weight variables as predictors
  - Use random effects to account for clustering
  - See, Yucel et al (2017, JSSAM), for an extension of Sequential Regression Approach

Option 2: "Uncomplex" and Impute

- Create synthetic populations by incorporating the complex sample features in a Non-parametric Bayes approach
- Impute the missing values in the synthetic populations using the regular SRMI (or any other procedure or even maximum likelihood)
- Combine estimates to form a single inference
- See Zhou et al (2016a JOS, 2016b JSSAM, 2016c Biometrics); Dong et al (2014a SM, 2014b SM)
- Implemented in IVEware (BBDESIGN module)

Overview

- Step 1: Bayesian Bootstrap to create nonsampled clusters within each stratum (Repeat S times)
- Step 2: Finite population weighted Bayesian Bootstrap to create nonsampled subjects within each cluster (Repeat B times for each draw in Step 1)
- Step 3: Multiply impute the missing values (L imputations) in each of the SxB synthetic populations
- A total SxBxL data sets are generated

Combining rules

- Point Estimate

$$\hat{\theta}_{sbl} = \text{ Estimate s} = 1,2,...,S; b = 1,2,...,B; l = 1,2,...,L$$

$$\overline{\theta}_{MI} = \sum_s \sum_b \sum_l \hat{\theta}_{sbl} / (SBL)$$

- Variance estimate

$$\hat{V}_{MI} = (1+S^{-1}) \sum_l (\overline{\theta}_{s++} - \overline{\theta}_{MI})^2 / (S-1)$$

$$\overline{\theta}_{s++} = \sum_b \sum_l \hat{\theta}_{sbl} / (BL)$$

- Confidence intervals

$$t - distribution : \nu = \min(c - H, S - 1)$$

$$c = \# clusters$$

$$H = \# strata$$

Choice of (S, B,L) and population sizes

- S may be fixed at c-H (complete data degrees of freedom)
- B and L are used mostly to obtain stable average estimate for each synthetic population
- For example (NHANES) (S=25, B=10 and L=10), (S=25, B=5, L=5) gave similar results for logistic and linear regression analyses
- Number of nonsampled clusters: C-c and the number of nonsampled elements within a cluster: N-n.
  - Approximation: $C = \infty; N = n / f; f = 0.01, 0.001$
  - Weights: Assume only combined final weight is available
- Computationally intensive (4,800 or 1,200 estimates).
- No variance estimate is needed from each synthetic data set
- See IVEware user manual for examples

Applications and Software

Applications of MI

- Survey of Consumer Finances, 1992
  - 5 multiply imputed data sets
- National Health and Nutritional Examination Survey
  - 5 multiply imputed data sets for a selected set of variables in NHANES-III. Uses multivariate normal model.
- National Health Interview Survey 1997-Present
  - Multiple imputation of missing family income and personal earnings (IVEware, PROC MI in SAS).
- Numerous applications in a variety of fields. Becoming a very common approach.

Software

- Sequential regression imputations
  - R and Stata (MICE, ICE, MI, IVEware)
  - Standalone (IVEware)
  - SAS (IVEware), PROC MI
  - SPSS (IVEware)
- MI-Analysis
  - PROC MIANALYZE
  - IVEware (can handle complex sample survey)
  - MI (Stata)
  - MITOOLS (R)
  - SUDAAN

IVEware

- A collection of SAS,R, SPSS, STATA, C and Fortran routines
- Handles linear (Continuous), logistic (Binary), multinomial logistic (categorical), Poisson (Count) and two-stage linear/logistic (Mixed or semi-continuous)
- semiparametric regression models
  - Response propensity and Predictive mean stratification
  - Tukey's *gh*-distribution
- Stepwise selection possible at each step to save computation time (use with caution and only if it is absolutely necessary)
- Add interaction terms
- Specify bounds
- Specify logical restrictions and skip patterns
- Imputation diagnostics

IVEware

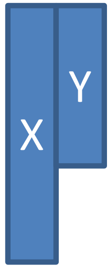
- Issues
  - Convergence
  - Several completed data statistics seem to converge to the same value regardless of seeds
  - Several articles establish conditions for convergence
  - Good fitting models are needed to get results with desirable repeated sampling properties

Model Diagnostics

- Good fitting regression models are key to obtain valid imputations
- Model building involves exploratory analysis, such scatter plots, histograms etc
- Residuals from the current model needs to be checked and refine the models, if necessary
- Model building tools, posterior predictive checks etc. are important component of imputation process
- Poorly fitting models can result in bias and may even be worse than the complete case analysis

Imputation Diagnostics

- Compare the distributions of the imputed and observed values
  - Under MCAR they should be similar
  - Not under MAR, even with the good fitting model
- Compare conditional distributions



$$\textit{Under}\ \text{MAR}$$

$$f(y_{obs} \mid x) = f(y_{imp} \mid x) \ \text{ or}$$

$$f(y_{obs} \mid e(x)) = f(y_{imp} \mid e(x))$$

$$e(x) = \Pr(y \text{ is missing})$$

  - Scatter plot of Y versus X, observed and imputed values different symbols or colors

Creating Multiple Imputations

- Multiple Imputations created within a single model take into account within-model uncertainty
- Multiple Imputations can also be created under alternative models, to account for imputation model uncertainty
- Imputations can be based on implicit or explicit models, as for single imputation

Summary of Imputation

- Sequential Regression/Chained Equation is a flexible approach for handling missing data with varying type of variables and complex structure
- Standard regression diagnostics can be used to fine tune the model to fit the observed data well
- Models can be parametric, semi-parametric or non-parametric
- Many software available to implement the method
- It is easy to program using a macro environment

Summary of Multiple Imputation

- Retains advantages of single imputation
  - Consistent analyses
  - Data collectors knowledge
  - Rectangular data sets
- Corrects disadvantages of single imputation
  - Reflects uncertainty in imputed values
  - Corrects inefficiency from imputing draws
    - estimates have high efficiency for modest D, e.g. 10