Statistical Analysis with Missing Data

Lecture for BDSI 2022

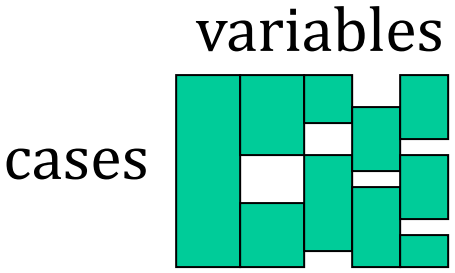
Dr. Peisong Han

Created based on Prof. Rod Little's lecture slides

Missing Data Problems

Missing data problems are very common in almost every field

- Nonresponse in sample surveys
- Noncompliance in clinical trials
- Two-stage design, etc.
- Dropout in longitudinal studies
- ……



Example: Longitudinal Data with Dropout (Hedeker and Gibbons, 1997)

- Randomized psychiatric trial
- 329 patients received drug therapies for schizophrenia; 108 patients received a placebo.
- Measurements at weeks 0, 1, 3, 6
- Missing data primarily due to dropout
- Outcome: severity of illness (1=normal, . . . , 7=extremely ill); treated as continuous

Sample Size of the Psychiatric Trial

		Time		
Group	0	1	3	6
Placebo ($n = 108$)	107	105	87	70
Drug ($n = 329$)	327	321	287	265

Note: The drug group combines three treatments.

Dropout Rates:

Placebo: 35%

Drug: 20%.

Bias when ignoring subjects with missing data: a simulation study

- True model:

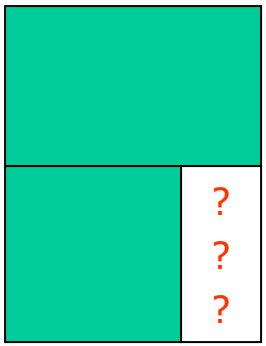
$$X \sim N(0,1)$$

$$\text{Logit}[Pr(E=1|X)]=0.5+X$$

$$\text{logit}[Pr(D=1|E,X)]=0.25+0.5X+1.1E$$

- Sample size: 500
- Number of Replicates: 5000

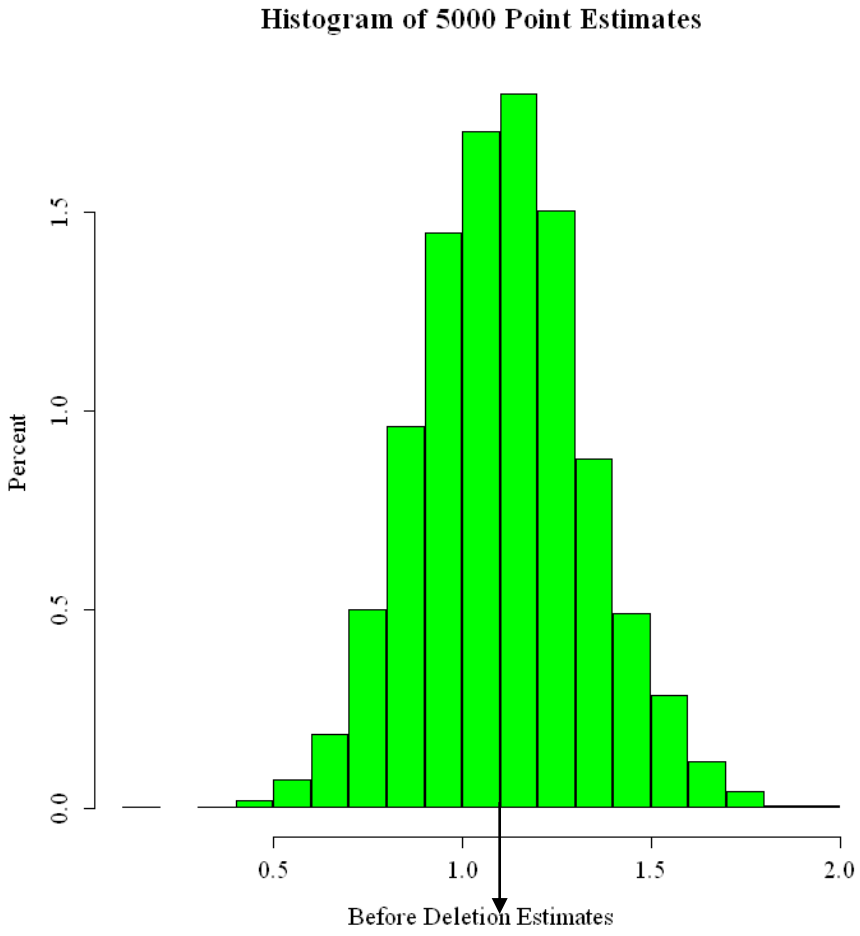
Missing-Data Mechanism



- D and E : completely observed
- X : sometimes missing
- Values of X in each cell are set to missing with the following underlying probabilities:

  D=0,E=0: p_{00}=0.19
  D=0,E=1: p_{01}=0.09
  D=1,E=0: p_{10}=0.015
  D=1,E=1: p_{11}=0.055

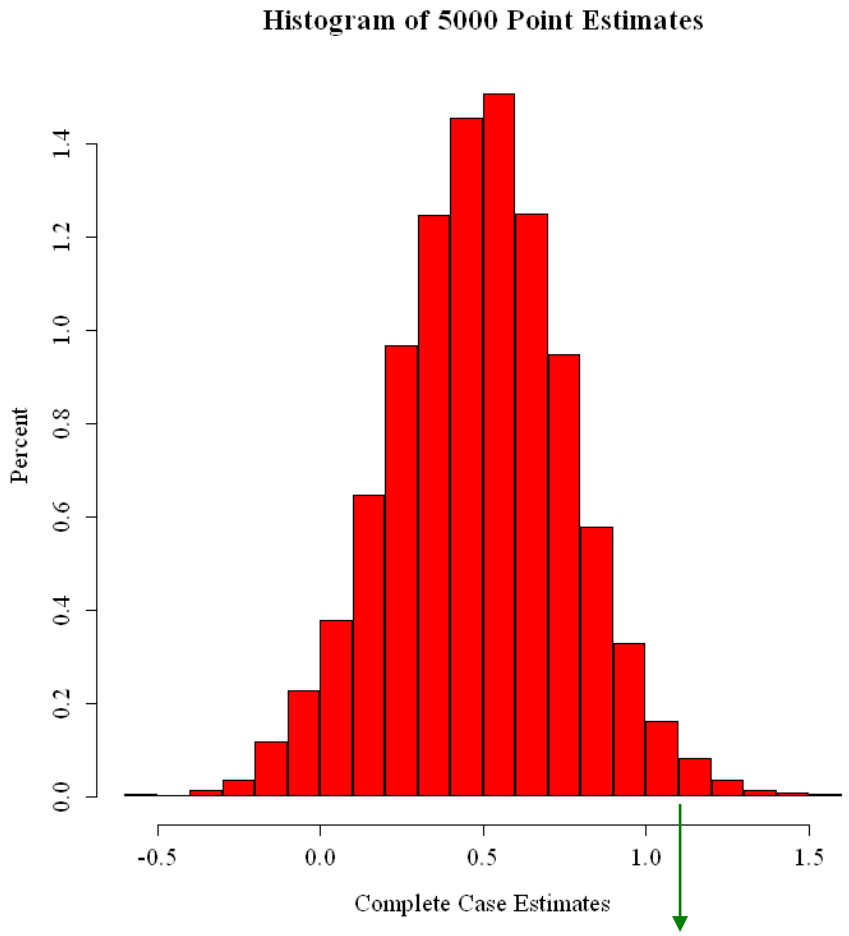
Before Deletion Estimates



- Histogram of 5000 estimates before deleting values of X
- logistic model

$$logit\ Pr(D=1|E,X)$$
$$=\beta_0+\beta_1 E+\beta_2 X$$

Complete-Case Estimates



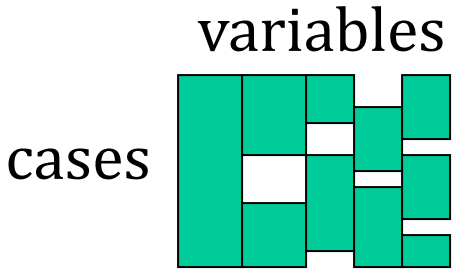
Histogram of complete- case analysis estimates

Delete subjects with missing X values

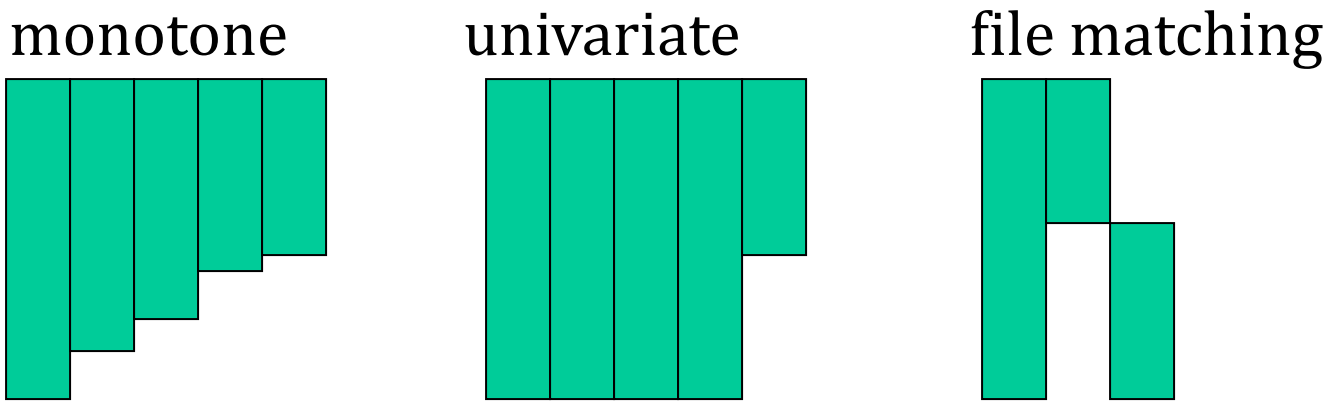
True value = 1.1, serious negative bias

Patterns of Missing Data

- General pattern



- Some special patterns



Missing Data Mechanisms

$\boldsymbol{Y}_{obs}$: Measurements observed.

$\boldsymbol{Y}_{mis}$: Measurements that should be available but are missing.

$\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$: Hypothetical complete data.

$\boldsymbol{R}$: Indicator of missingness.

- Missing Completely At Random (MCAR): if $\boldsymbol{R}$ is independent of both $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Y}_{mis}$.
- Missing At Random (MAR): if $\boldsymbol{R}$ is independent of $\boldsymbol{Y}_{mis}$, but dependent on $\boldsymbol{Y}_{obs}$, i.e., the probability of missingness only depends on $\boldsymbol{Y}_{obs}$, but not $\boldsymbol{Y}_{mis}$.
- Nonignorable (Informative) Missing (NMAR): if $\boldsymbol{R}$ is dependent on $\boldsymbol{Y}_{mis}$, i.e., the probability of missingness depends on both $\boldsymbol{Y}_{mis}$ and $\boldsymbol{Y}_{obs}$.

Some Examples of Missingness Mechanism

Example of MCAR:

Patients miss a scheduled visit because of bad weather or car out of service.

Example of MAR:

Older people may have a higher chance of dropping out of a study. (Suppose age is observable.)

Example of NMAR:

Subjects drop out because they have poor treatment outcomes or they die.

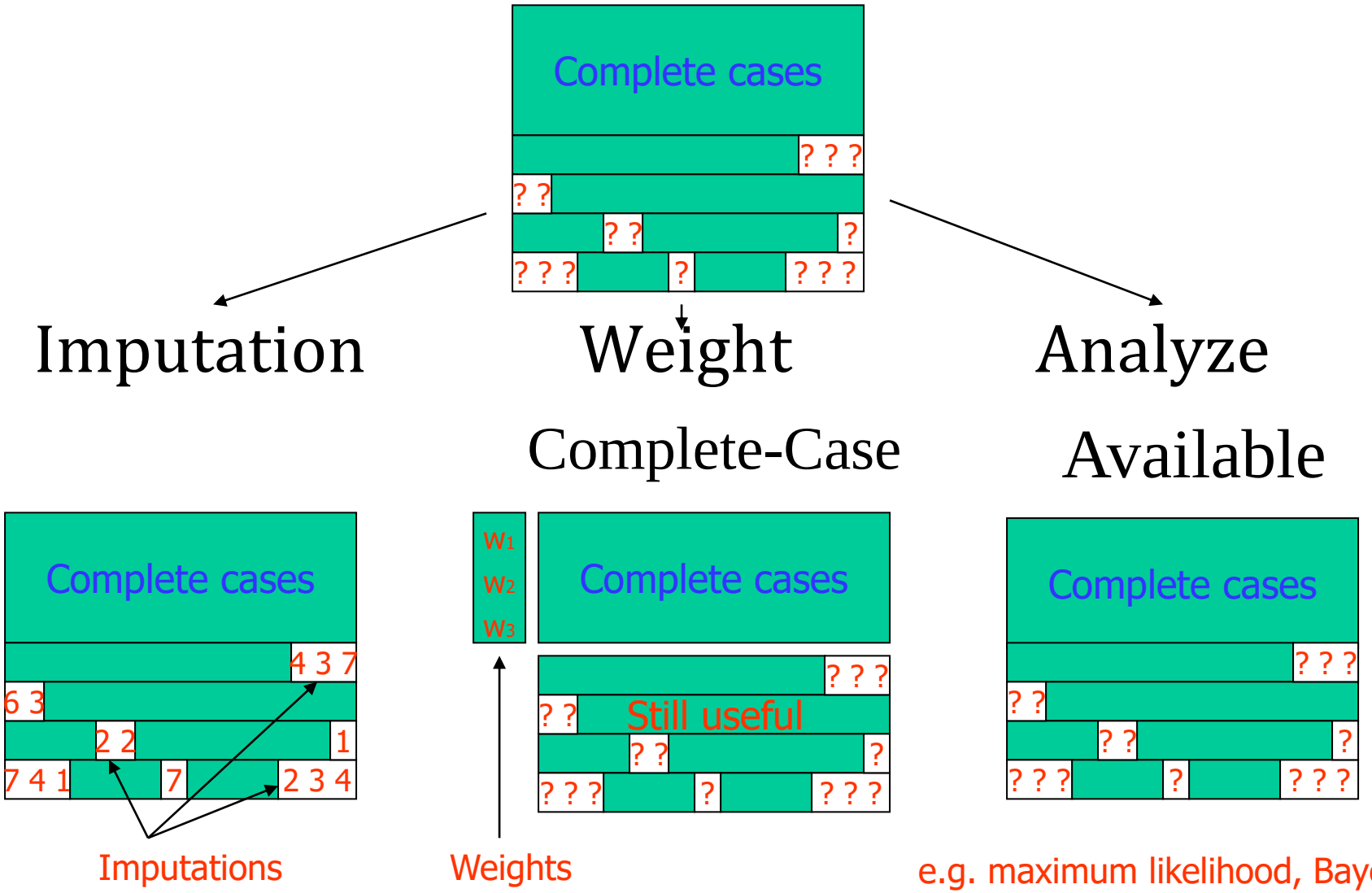
More Examples

- **MCAR**:
  - patients had their weight measured by flipping a coin.
- **MAR**:
  - patients with high blood pressure had their weight measured.
- **NMAR**:
  - overweight patients had their weight measured.

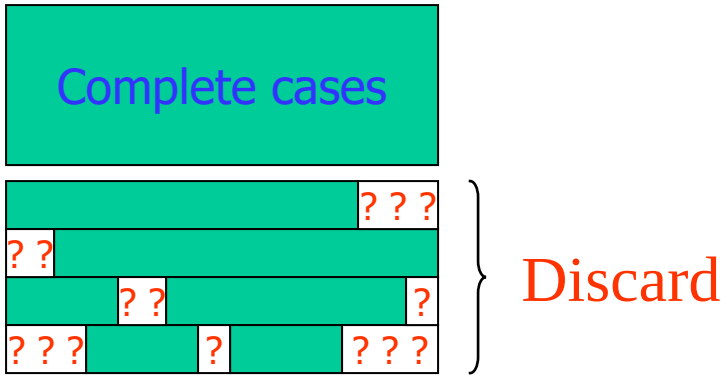
What Mechanism to Assume

- **MCAR**:
  - Simplest mechanism; strongest assumption; usually not the true mechanism in practice
- **NMAR**:
  - Most complex mechanism; weakest assumption; likely the true mechanism in practice
- **MAR**:
  - A mechanism between MCAR and NMAR; oftentimes a good approximation to the truth; easy to work with

General Strategies



Complete-Case Analysis



- Default analysis in statistical packages
- Simple and valid if MCAR
- Generally biased estimation

CC Analysis

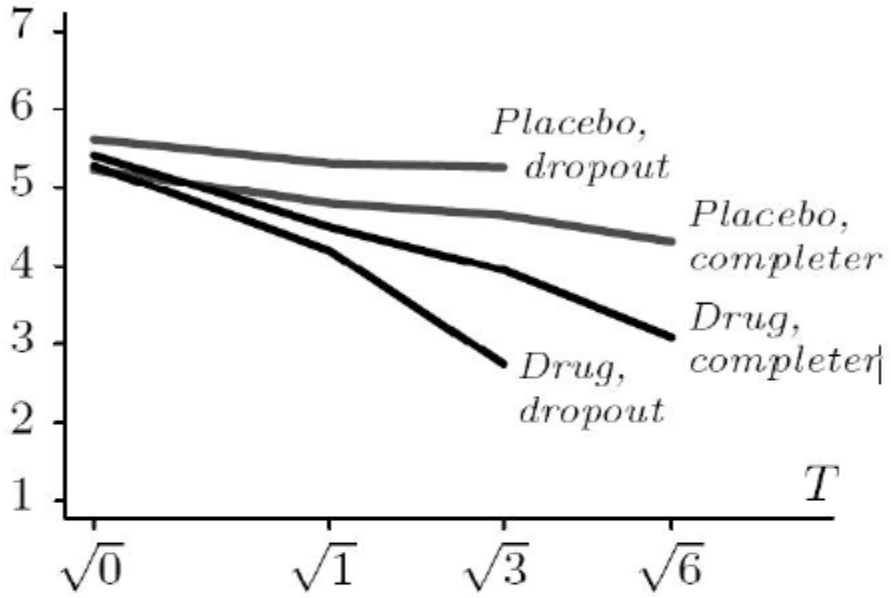
- Does not invent data

- Simple and may be good enough with small amounts of missing data
  - but defining "small" is problematic; depends on
    - fraction of incomplete cases
    - recorded information in these cases
    - parameter being estimated

Limitations of CC Analysis

- Loss of information in incomplete cases
  - Increased variance of estimates
  - Bias when complete cases differ systematically from incomplete cases
    - restriction to complete cases requires that the complete cases are representative of all the cases for the analysis in question, but this assumption is often questionable!

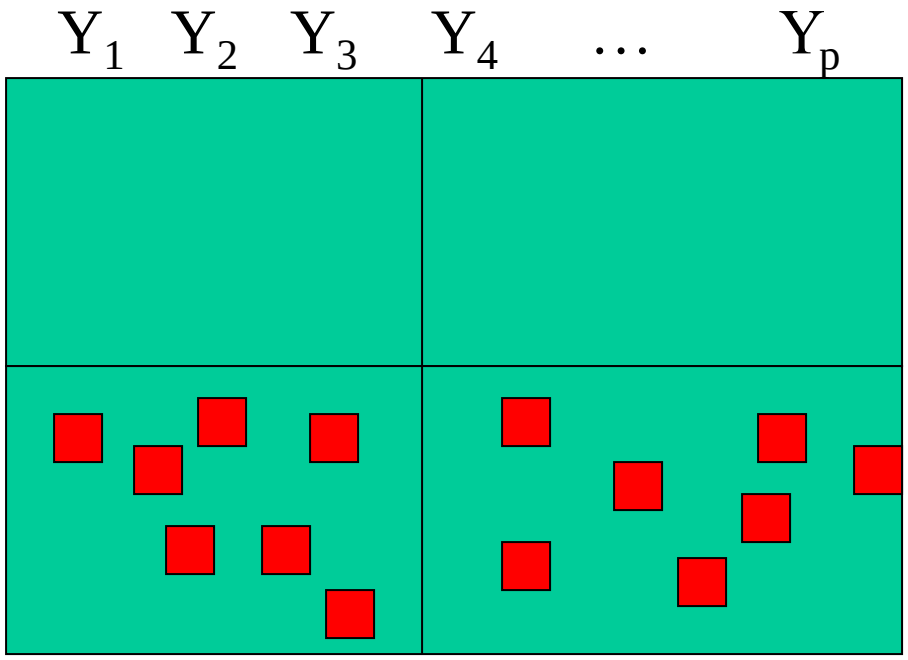
Figure: Average response versus square root of time (in weeks)



- In the treatment group, the subjects who dropped out had lower scores than the completers.
- In the control group, the subjects who dropped out had higher scores than the completers.
- A completer-only (complete case) analysis would severely understate the treatment effect.

Imputation and Multiple Imputation

Problem



Complete cases

Cases with some missing values

D_{obs} = Observed data:

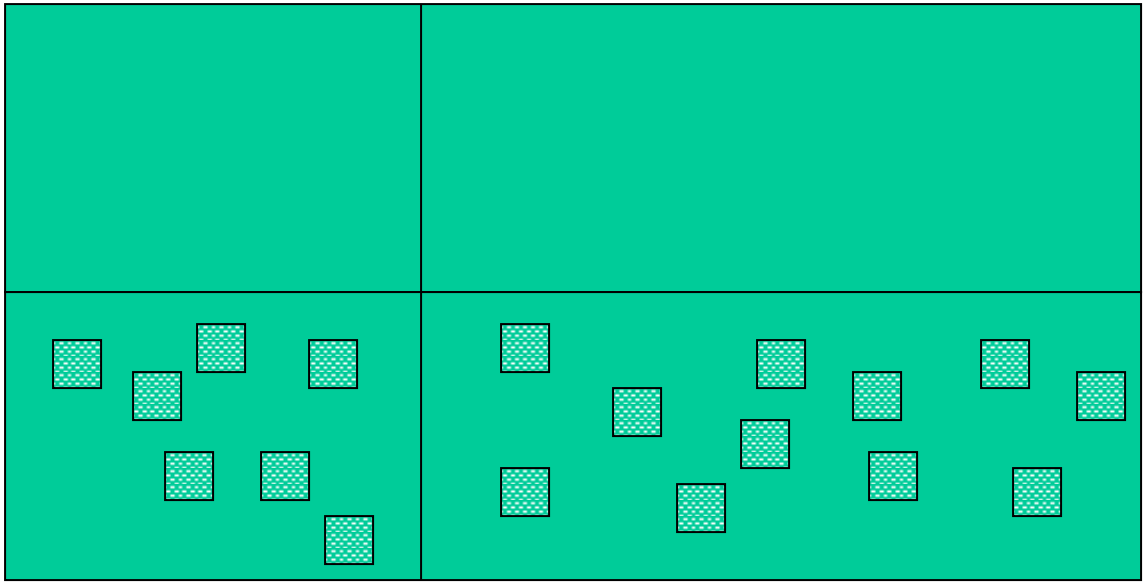
D_{miss} = Missing data:

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

Considerations behind Imputation

- Multiple users analyzing different subsets of variables
- Different skill levels dealing with incomplete data
- Software to perform complete data analysis is available
- Assume missing at random

Imputation



Imputation:

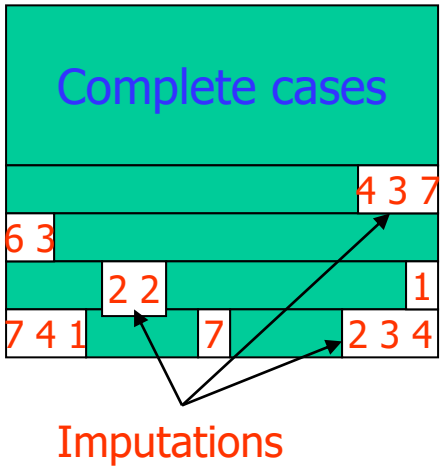
Draws from $\Pr(D_{miss} \mid D_{obs})$

Important issues:

Imputations are not real values

Uncertainties associated with imputes

Features of Imputation



Good

Rectangular File

Retains observed data

Handles missing data once

Exploits incomplete cases

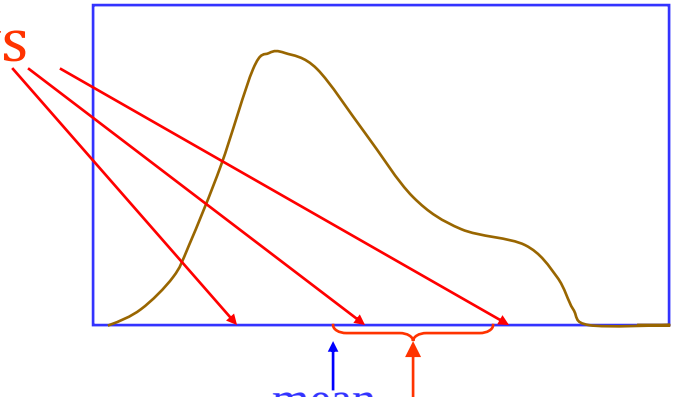
Bad

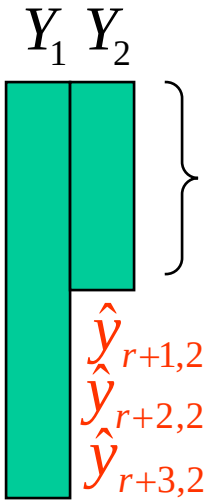
Naïve methods can be bad

Invents data –
Understates uncertainty

A Bivariate Example: Continuous Case

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





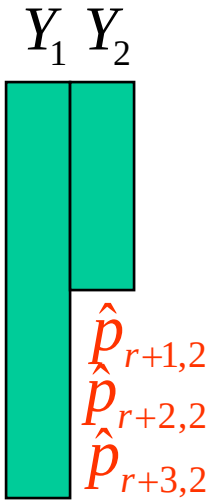
$$\hat{y}_{i2} = \hat{E}(y_{i2} \mid y_{i1}) + r_i$$

$r_i \sim N(0, s_{22\cdot1})$, $s_{22\cdot1}$ = resid variance, or

$r_i =$ residual from randomly selected complete case

A Bivariate Example: Binary Case

- For binary (0-1) data, impute 1 with probability = predicted prob of a one given observed covariates



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

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

Example: Should Imputations be conditional on all observed variables?

- Consumer Expenditure Survey (Bureau of Labor Statistics)

- Should the imputation of Income be conditional on the Expenditure variable?

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

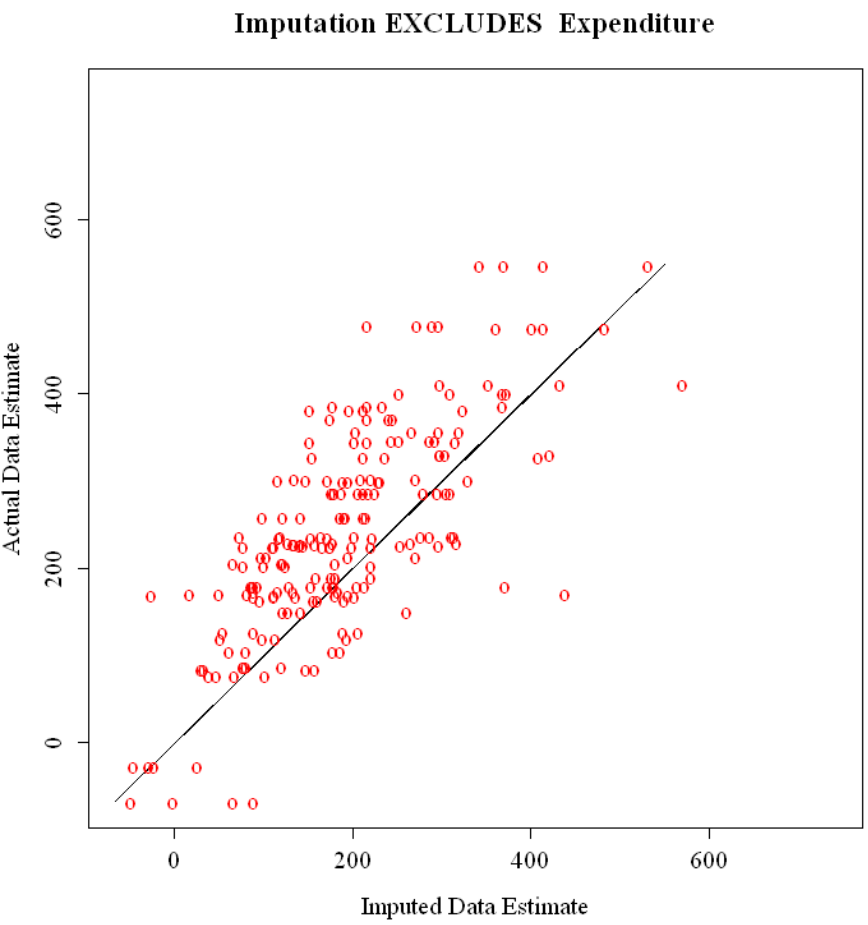
- OLS model

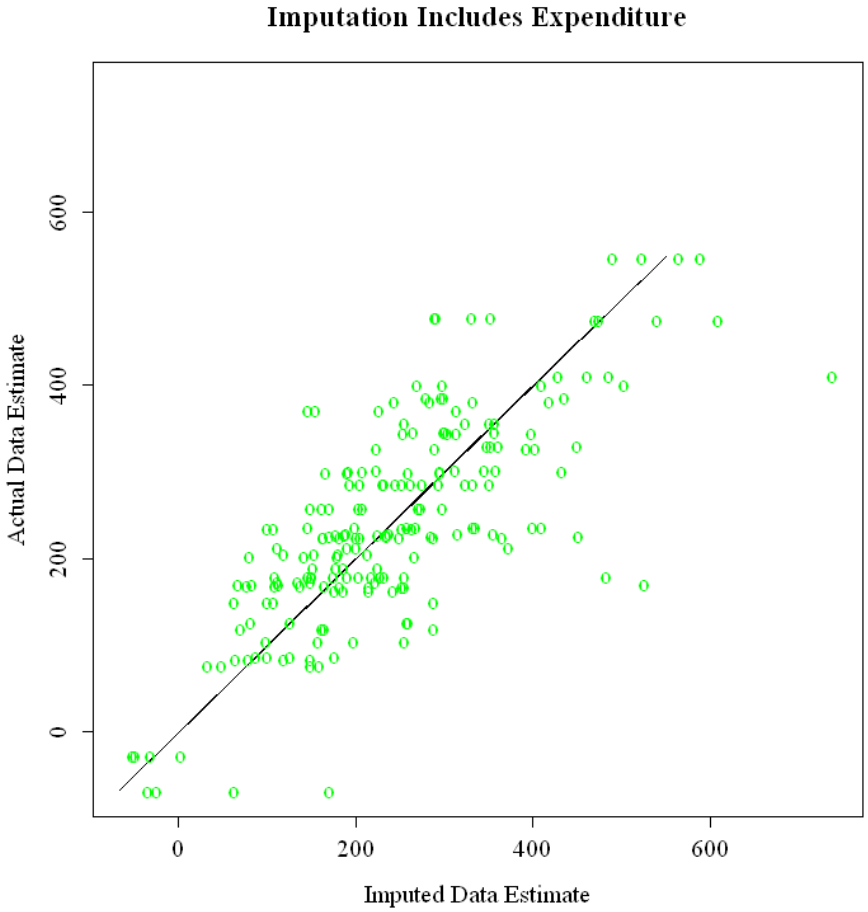
  $Food\text{-}At\text{-}Home = \beta_0 + \beta_1\, Income + covariates$

- Tobit Model

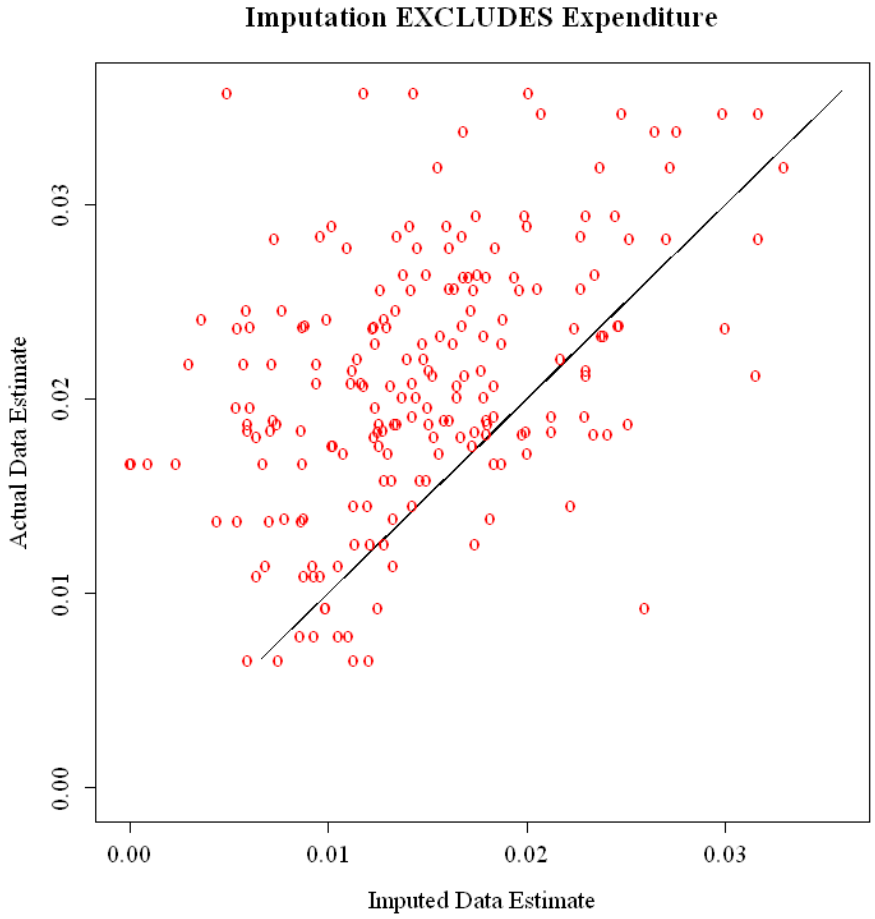
  $Food\text{-}Away\text{-}Home = \gamma_0 + \gamma_1\, Income + covariates$

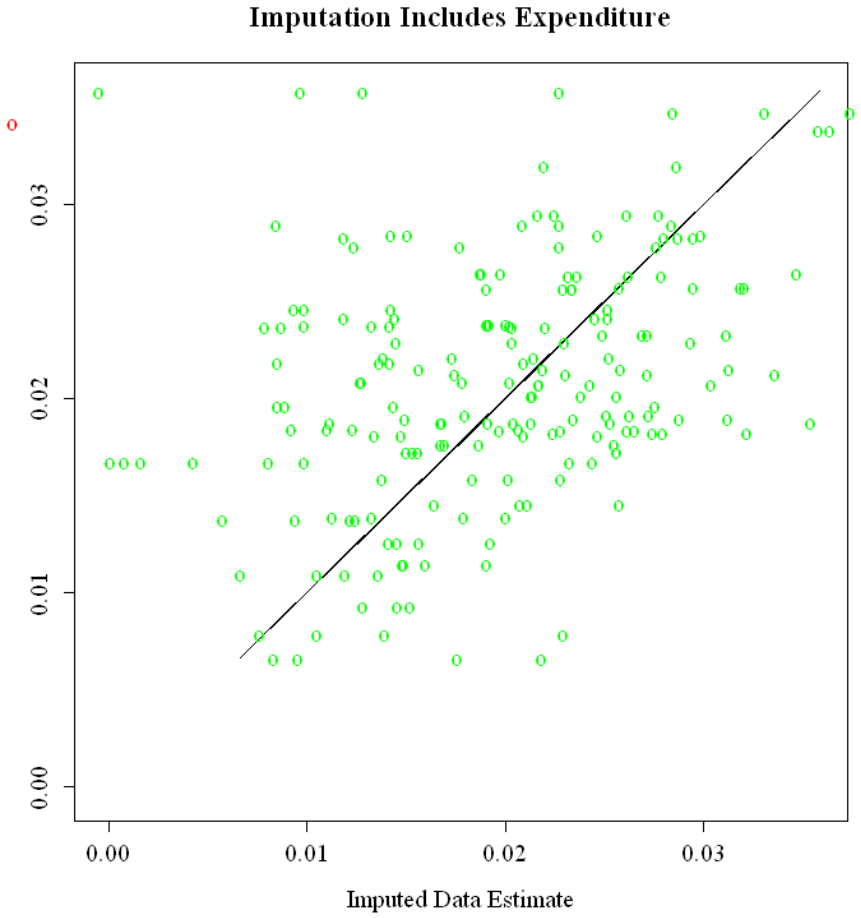
Estimated regression coefficients of income from undeleted and imputed data-sets: OLS Model





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





What should imputes condition on?

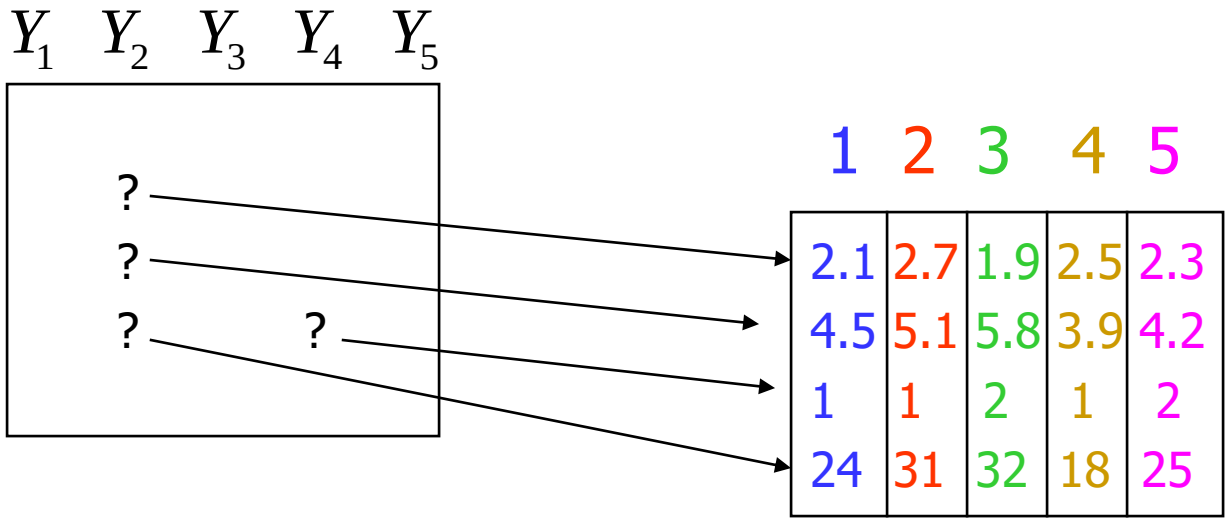
- In principle, all observed variables
  - Whether predictors or outcomes of final analysis model
  - May be impractical with a lot of variables
- Variable selection
  - Priority to variables predictive of missing variable (and nonresponse)
  - Favor inclusion over exclusion

Key Problem of Single Imputation

- Single imputations do not account for imputation uncertainty
  - bootstrapping the imputation method
  - multiple imputation

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 $\bar{\theta}_D = \frac{1}{D}\sum_{d=1}^{D} \hat{\theta}_d$

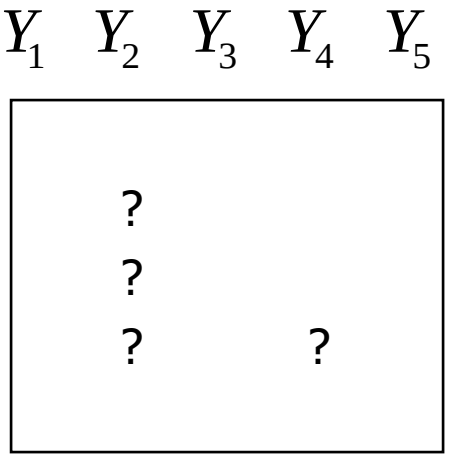
$W_d =$ estimate of variance of $\hat{\theta}_d$ from d th dataset

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

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

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

Example of Multiple Imputation



	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	$\sqrt{T_D} = \sqrt{\bar{W}_D + 6/5\, B_D}$
μ_1	12.6	3.6^2	0	3.6
$\beta_{53\cdot1234}$	4.36	2.27^2	0.339	2.36

Imputation for monotone patterns

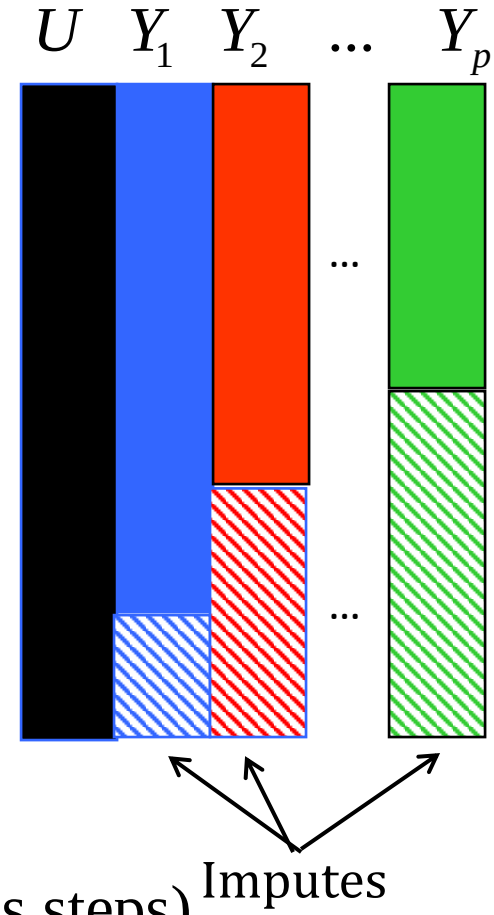
(a) regress Y_1 on U, impute missing values of Y_1

(b) regress Y_2 on Y_1 and U,
impute missing values of Y_2
(with imputes for missing Y_1 from (a))

...

(k) regress Y_p on $Y_1,...,Y_{p-1}$ and U,
impute missing values of Y_p
(with imputes for missing $Y_1 ... Y_{p-1}$ from previous steps)

E.g. SAS PROC MI



Chained Equations/ Sequential Regression approach

U = fully observed, $Y_1,...Y_p$ = incomplete, ordered least to most missing values. (Not necessarily a monotone pattern)

Iteration 1: Regress Y_1 on U, impute missing Y_1

Regress Y_2 on $U, Y_1,$ impute missing Y_2

...

Regress Y_p on $U, Y_1,...Y_{p-1}$, impute missing Y_p

Iteration 2...I: update imputed draws:

Regress Y_1 on $U, Y_2,...Y_p$ reimpute missing Y_1

Regress Y_2 on $U, Y_1, Y_3,...Y_p,$ reimpute missing Y_2

...

Regress Y_p on $U, Y_1,...Y_{p-1}$, reimpute missing Y_p

Chained Equations/ Sequential Regression approach

- Missing values are replaced by most recent imputes
- Regression is tailored to the type of variable:
  - Continuous (linear regression)
  - Binary (logistic regression)
  - Categorical (polytomous regression)
  - Count (Poisson regression)
- Regression diagnostics to fine tune each model
- Empirical studies show that nothing much changes after 5 or 6 iterations

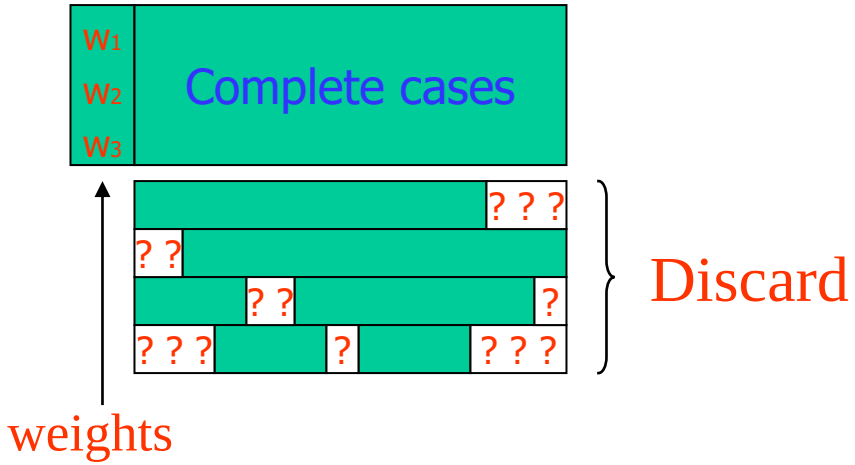
Chained Equation Approach

- Sacrifices a coherent joint distribution for the variables, and models a sequence of conditional distributions

- Can be a useful practical approach for
  - General patterns of missing data
  - Explicit modeling joint distribution is difficult
    - Multivariate normality does not model variables of different types well

Weighted Complete Case Analysis

Weighted CC Analysis



- Weight respondents differentially to reduce nonresponse bias – e.g. mean becomes weighted mean
- Most common weight: prop to 1/P(response)

Adjustment Cell method for estimating prob of response

- Group respondents and nonrespondents into adjustment cells with similar values on variables recorded for both:
- e.g. females aged 25-35 living in AA

100 in sample
- 80 respondents
- 20 nonrespondents

pr(response in cell) = 0.8

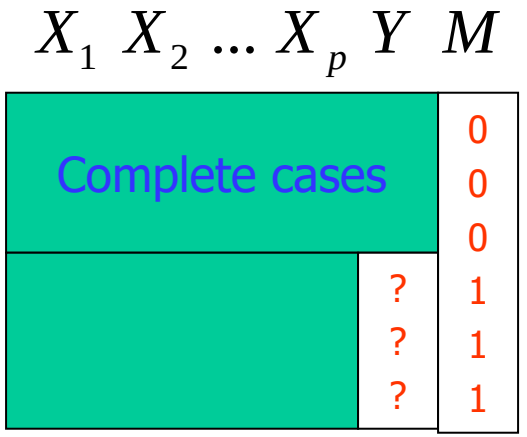
response weight = 1.25

- With extensive covariate information, can't cross-classify on all of them
- How do we choose which variables to use?

Response Propensity Score

- Regress *M* on *X* (probit or logistic), using respondent and nonrespondent data
  - p(M=0|X) is the response propensity score

- Weight respondents by inverse of propensity score
  - 1/p(M=0|X)



Propensity Score Weighting

- A widely used approach alternative to imputation
- Avoids modeling the data distribution
- A Fundamental concept in both missing data and causal inference
- May not be stable if some propensity scores are close to zero

Likelihood Methods

Likelihood methods

- Statistical model + data $\Rightarrow$ Likelihood
- Two general approaches based on likelihood
  - maximum likelihood inference for large samples
  - Bayesian inference for small samples:

    log(likelihood) + log(prior) = log(posterior)
- Methods use all available data
  - do not require rectangular data sets

Parametric Likelihood

- Data Y
- Statistical model yields probability density $f(Y \mid \theta)$

  for Y with unknown parameters θ
- Likelihood function is then the density as a function of θ

$$L(\theta \mid Y) = \text{const} \times f(Y \mid \theta)$$

- Loglikelihood is often easier to work with:

$$\mathrm{l}(\theta \mid Y) = \log L(\theta \mid Y) = \text{const} + \log\{ f(Y \mid \theta)\}$$

  Constants can depend on data but not on parameter θ

Example: Normal sample

- univariate iid normal sample

Data $Y = (y_1, ..., y_n)$

Parameters $\theta = (\mu, \sigma^2)$, μ = mean, σ^2 = variance

Normal density: $f(Y \mid \mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mu\right)^2 \right)$

Likelihood: $L(\mu, \sigma^2 \mid Y) = \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mu\right)^2 \right)$

Maximum Likelihood Estimate

- The maximum likelihood (ML) estimate $\hat{\theta}$ of θ maximizes the likelihood

$$L(\hat{\theta} \mid Y) \geq L(\theta \mid Y) \text{ for all } \theta$$

- The ML estimate is the

  "value of the parameter that makes the data most likely"

- The ML estimate is not always unique, but is for many regular problems given enough data

Computing the ML estimate

- In regular problems, the ML estimate can be found by solving the score equation

$$S(\theta \mid Y) \equiv \frac{\partial \log L(\theta \mid Y)}{\partial \theta} = 0$$

- Iterative methods – e.g. Newton-Raphson, Scoring, EM algorithm -- required for most problems

Properties of ML estimates

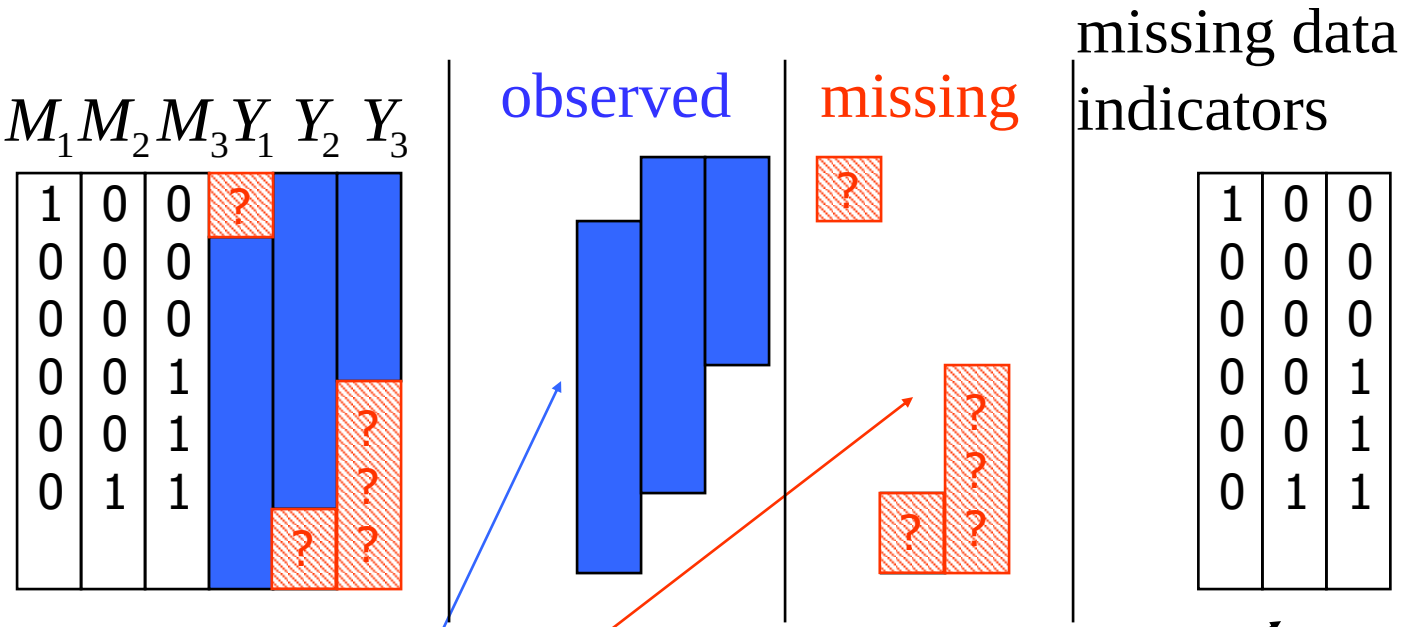
- Under assumed model, ML estimate is:
  - Consistent
  - Efficient for large samples
  - not necessarily the best for small samples
- ML estimate is transformation invariant
  - If $\hat{\theta}$ is the ML estimate of θ

    Then $\phi(\hat{\theta})$ is the ML estimate of $\phi(\theta)$

Likelihood methods with incomplete data

- Statistical models needed for:
  - data without missing values
  - missing-data mechanism
- Model for mechanism not needed if it is ignorable (to be defined later)
- With likelihood, proceed as before:
  - ML estimates, large sample standard errors
  - Bayes posterior distribution

The Observed Data



$$Y = (y_{ij})_{n \times K} = (Y_{\text{obs}}, Y_{\text{mis}})$$

$$M = (m_{ij})_{n' K}$$

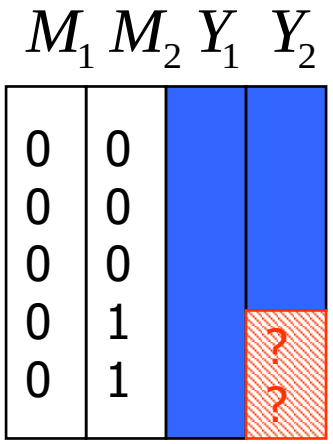
$$m_{ij} = \begin{cases} 0, & y_{ij} \text{ observed} \\ 1, & y_{ij} \text{ missing} \end{cases}$$

Model for Y and M

$$f(Y, M|\theta, \psi) = f(Y|\theta) \times f(M|Y, \psi)$$

Complete-data model model for mechanism

Example: bivariate normal monotone data



complete-data model:

$$(y_{i1}, y_{i2}) \sim_{iid} N_2(\mu, \Sigma)$$

model for mechanism:

$$(m_{i2} \mid y_{i1}, y_{i2}) \sim_{ind} Bern\left(\Phi(\gamma_0 + \gamma_1 y_{i1} + \gamma_2 y_{i2})\right)$$

$\Phi =$ Normal cumulative distribution function

The likelihood

- Likelihood should involve model for *M*

$$f(Y_{\text{obs}}, M \mid \theta, \psi) = \int f(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta) f(M \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) dY_{\text{mis}}$$

$$\Rightarrow L_{\text{full}}(\theta, \psi \mid Y_{\text{obs}}, M) = const \times f(Y_{\text{obs}}, M \mid \theta, \psi)$$

Likelihood when *ignoring the missing-data mechanism M*

simpler since it does not involve model for *M*

works (only) when the mechanism is ignorable

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

$$\Rightarrow L_{\text{ign}}(\theta \mid Y_{\text{obs}}) = const \times f(Y_{\text{obs}} \mid \theta)$$

Ignoring the md mechanism

- It is easy to show that sufficient conditions for ignoring the missing-data mechanism are:

(A) Missing at Random (MAR):

$$f(M \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) = f(M \mid Y_{\text{obs}}, \psi) \text{ for all } Y_{\text{mis}}$$

(B) Distinctness:

θ and ψ have distinct parameter spaces

(Bayes: priors distributions are independent)

Proof: $$f(Y_{obs}, M \mid \theta, \psi) = \int f(Y_{obs}, Y_{mis} \mid \theta) f(M \mid Y_{obs}, Y_{mis}, \psi) dY_{mis}$$

$$= \overset{(MAR)}{=} \int f(Y_{obs}, Y_{mis} \mid \theta) f(M \mid Y_{obs}, \psi) dY_{mis}$$

$$= f(M \mid Y_{obs}, \psi) \times \int f(Y_{obs}, Y_{mis} \mid \theta) dY_{mis}$$

$$= f(M \mid Y_{obs}, \psi) \times f(Y_{obs} \mid \theta)$$

- If MAR holds but not distinctness, ML based on ignorable likelihood is valid but not fully efficient
- So MAR is the key condition

Bayes: add prior distributions

$$p_{\text{complete}}(\theta,\psi \mid Y, M) = \pi(\theta,\psi) \times f(Y \mid \theta) \times f(M \mid Y,\psi)$$

Prior dn Complete-data model model for mechanism

$$p_{\text{full}}(\theta,\psi \mid Y_{\text{obs}}, M) \propto \pi(\theta,\psi) \times f(Y_{\text{obs}}, M \mid \theta,\psi)$$

- *Full* posterior dn - involves model for *M*

$$f(Y_{\text{obs}}, M \mid \theta,\psi) = \int f(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta) f(M \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) dY_{\text{mis}}$$

Posterior dn *ignoring the missing-data mechanism M*
(simpler since it does not involve model for *M*)

$$p_{\text{ign}}(\theta \mid Y_{\text{obs}}) \propto \pi(\theta) \times f(Y_{\text{obs}} \mid \theta)$$

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

Summary

- Likelihood Inference Ignoring the Missing Data Mechanism is valid if
  - Model for *Y* is correctly specified
  - Data are MAR
  - Fully efficient if distinctness condition holds

Some Discussion and Take Home Messages

A Few Notes

- Any missing data method involves modeling assumptions
- Collect and use relevant covariate information
  - Covariates related to missingness and main outcomes
- Sensitivity analyses for nonignorable missing data

Two Major Approaches

- Likelihood approach
  - EM algorithm (Dempster et al. 1977; Wu 1983; Meng and Rubin 1993; McLachlan and Krishnan 2008)
  - multiple imputation method (Rubin 1978, 1987, 1996)
- Semiparametric approach
  - inverse probability weighting (IPW) method (Horvitz and Thompson 1952)
  - augmented IPW (AIPW) method (Robins, Rotnitzky and colleagues 1994, 1995; Tsiatis 2006)

Comparison of the Two Approaches

Likelihood approach

- specifies the joint distribution
- higher estimation efficiency if the distribution is correctly specified
- hard to deal with complex distribution
- sensitive to model misspecification

Semiparametric approach

- specifies the marginal feature of interest
- less efficient due to semiparametric specification of the model
- relatively easy to deal with complex distribution
- robust against model misspecification

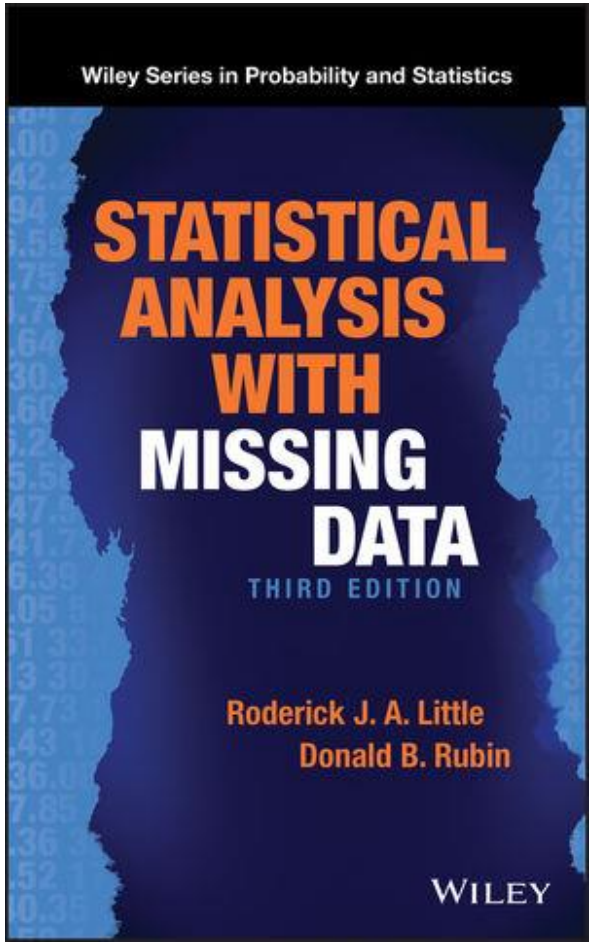
Missing data methods -- history

1. Before the EM algorithm (pre-1970's)
   - Ad-hoc adjustments (simple imputation)
   - ML for simple problems (Anderson 1957)
   - ML for complex problems too hard
2. ML era (1970's – mid 1980's)
   - Rubin formulates model for missing data mechanism, defines MAR (1976)
   - EM and extensions facilitate ML for complex problems
   - ML for more flexible models – beyond multivariate normal (see e.g. Little and Rubin 1987)

Missing data methods -- history

3. Bayes and Multiple Imputation (mid 1980's – present)
   - Rubin proposes MI, justified via Bayes (1977, 1987)
   - MCMC facilitates Bayes as an alternative to ML, with better small sample properties (see e.g. Little and Rubin 2019)
4. Robustness concerns (1990's – present)
   - Robins et al propose doubly robust methods for missing data based on *semiparametric approach*
   - Robust Bayesian models, more attention to model checks

A Great Textbook on Missing Data



Summary

- Missing data problems are widely seen
- Many methods dealing with missing data
  - Parametric, semiparametric, nonparametric
- We covered some basic concepts and methods
- Hopefully a useful introduction