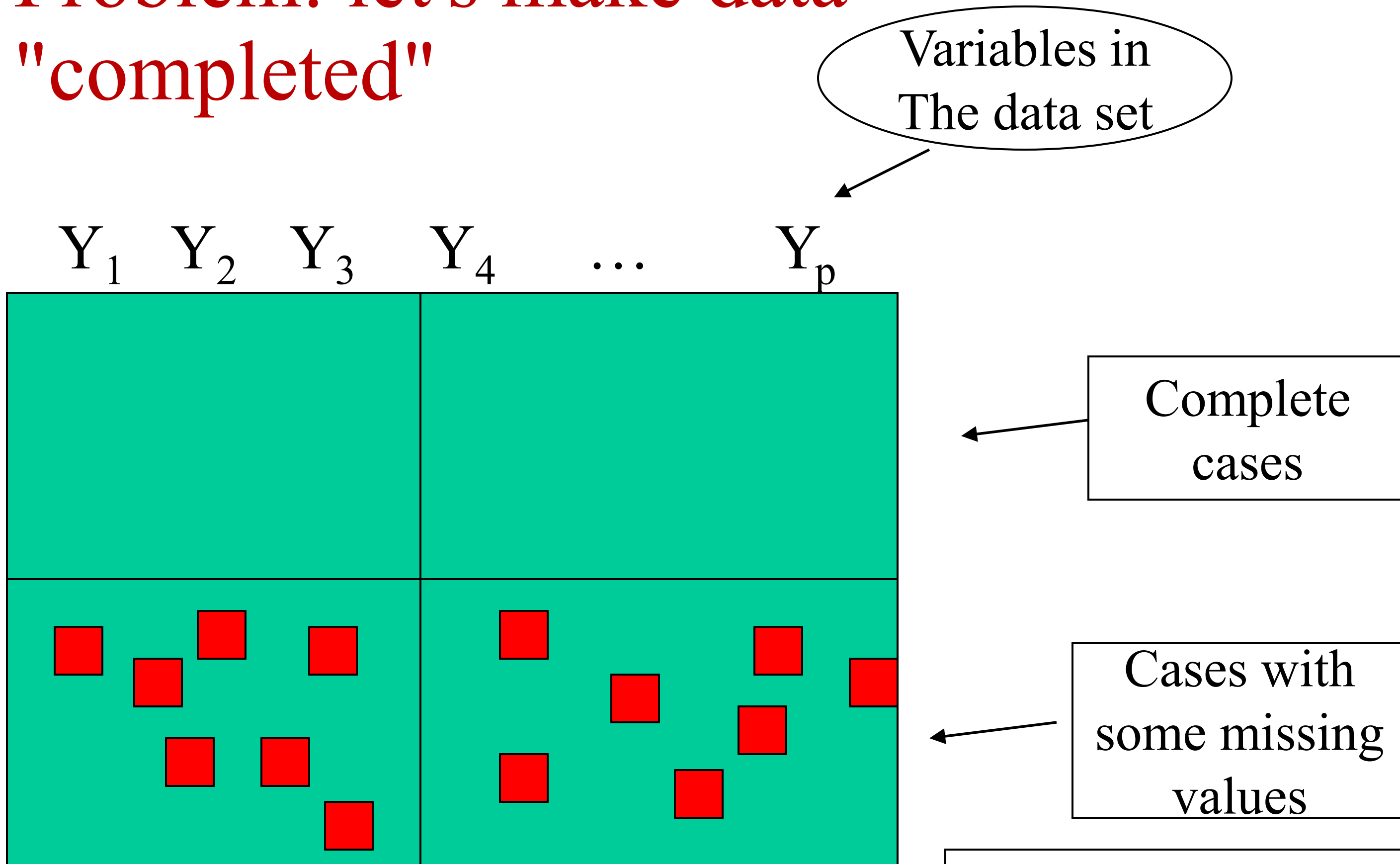
Statistical Analysis with Missing Data

Module 3. Imputation

University of Michigan

Problem: let's make data "completed"



D_{obs} = Observed data:

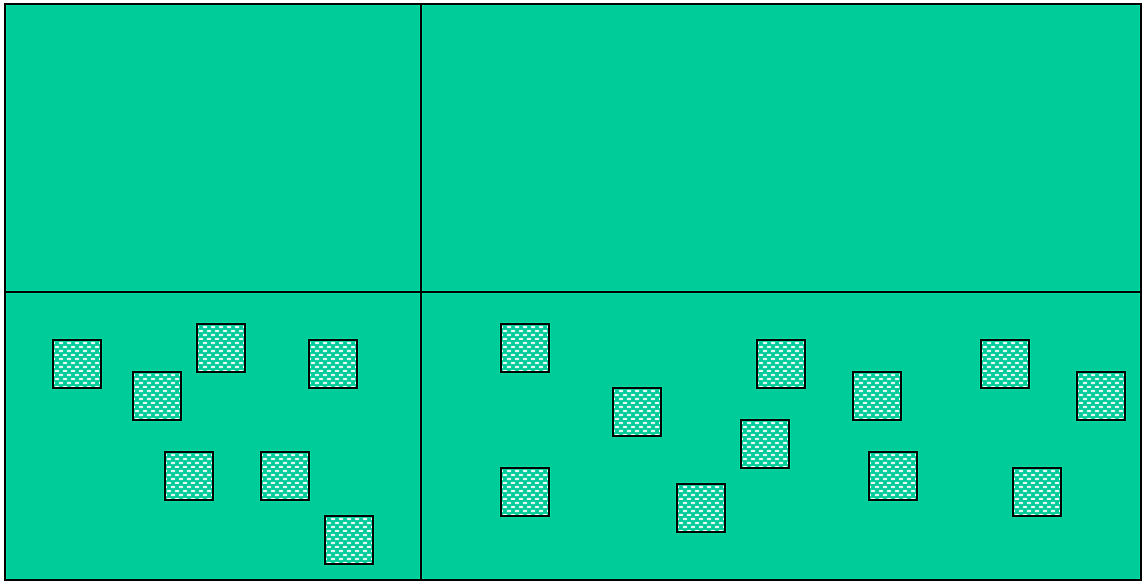
D_{mis} = Missing data:

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

Advantages of completing missing data

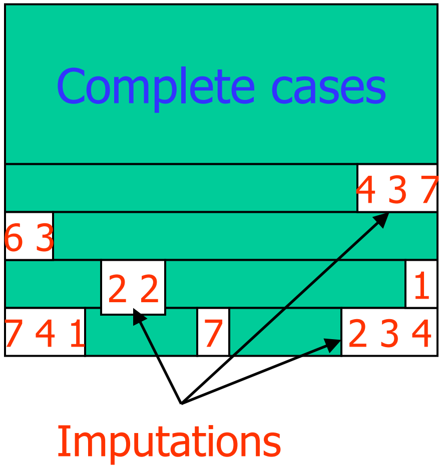
- Multiple users analyzing different subsets of variables
- Multiple analytical techniques
- Different skill levels dealing with incomplete data
- Analysis to be performed with complete data is known
- Software to perform complete data analysis is available
- Assume missing at random.
  - That is conditional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

Imputation



Fill in the missing values with estimates

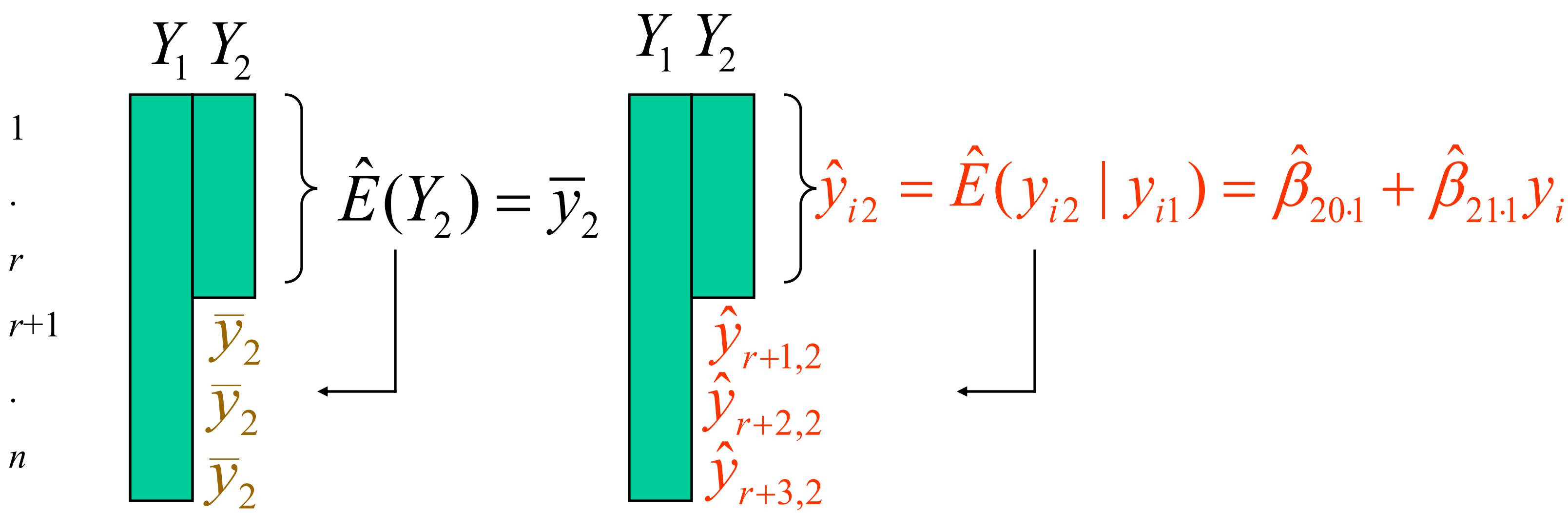
Features of Imputation



Good	Bad
Rectangular File	Naïve methods can be bad
Retains observed data	Invents data – Understates uncertainty
Handles missing data once	
Exploits incomplete cases	

Different ways for imputing data:
(a) Imputing Means

Unconditional — Conditional on observed variables

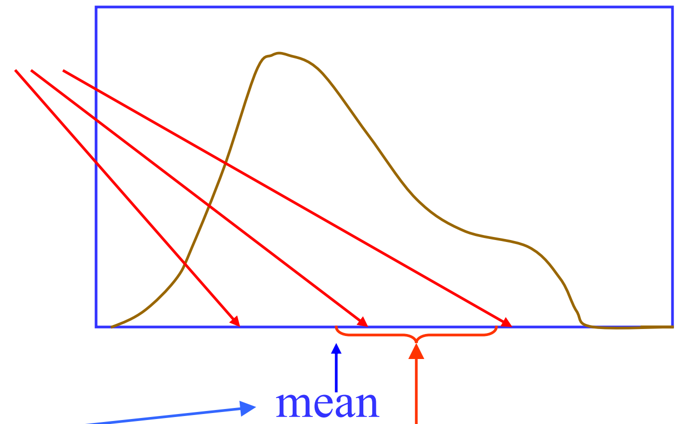


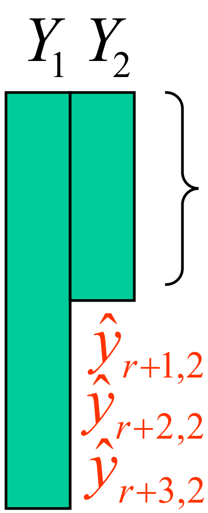
Properties of Mean Imputation

- 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

(b) Stochastic Imputation

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





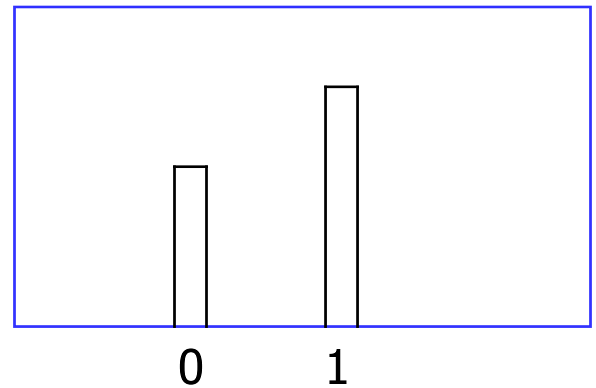
$$\hat{y}_{i2} = \hat{E}(y_{i2} \mid y_{i1}) + r_i$$

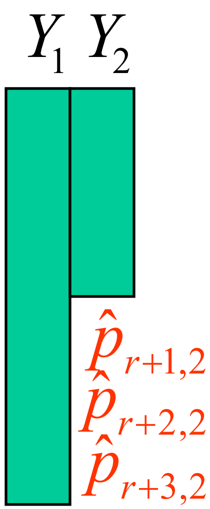
$r_i \sim N(0, s_{22\cdot1})$, $s_{22\cdot1}$ = resid variance, or

r_i = 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





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

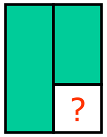
$$y_{i2} = \begin{cases} 1, \text{prob } \hat{p}_{i2} \\ 0, \text{prob } 1 - \hat{p}_{i2} \end{cases}$$

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

Example: bivariate MCAR data

Y_1 fully observed; Y_2 missing for fraction λ of cases;
MCAR mechanism



Large sample bias ~ E(estimate from filled-in data) - true value

Method	μ_2	σ_{22}	$\beta_{21\cdot1}$	Parameter $\beta_{12\cdot2}$
U Mean	0^*	$-\lambda\sigma_{22}$	$-\lambda\beta_{21\cdot1}$	0^*
U Draw	0	0	$-\lambda\beta_{21\cdot1}$	$-\lambda\beta_{12\cdot2}$
C Mean	0	$-\lambda(1-\rho^2)\sigma_{22}$	0^*	$\frac{\lambda(1-\rho^2)}{1-\lambda(1-\rho^2)}\beta_{12\cdot2}$
C Draw	0	0	0	0

* indicates that estimator is same as that from complete cases

(c) Imputing missing covariates in regression analysis

- 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 always understated for single imputation

Example 3: 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

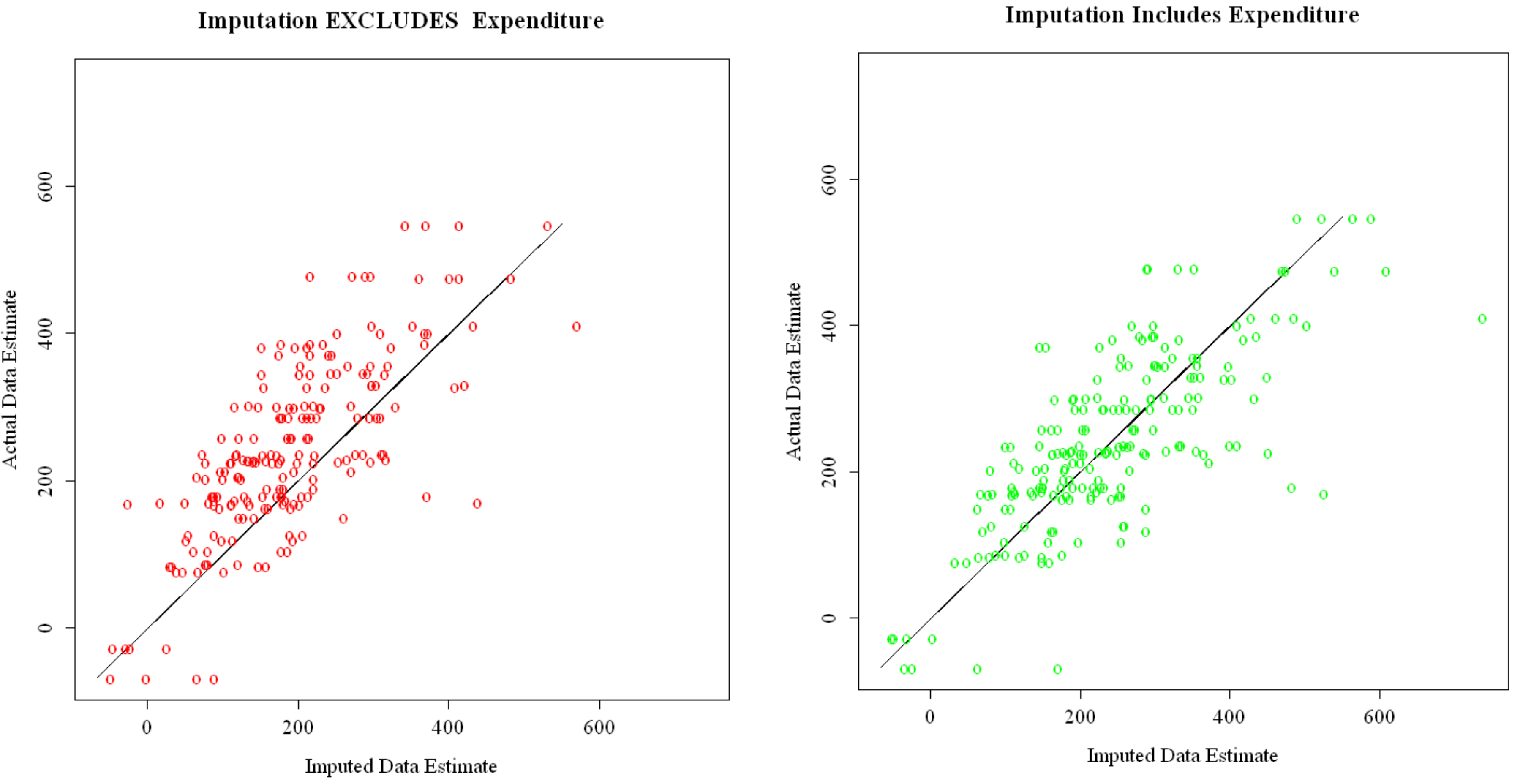
  Food-At-Home=β_0+β_1 *Income + covariates*

- Tobit Model

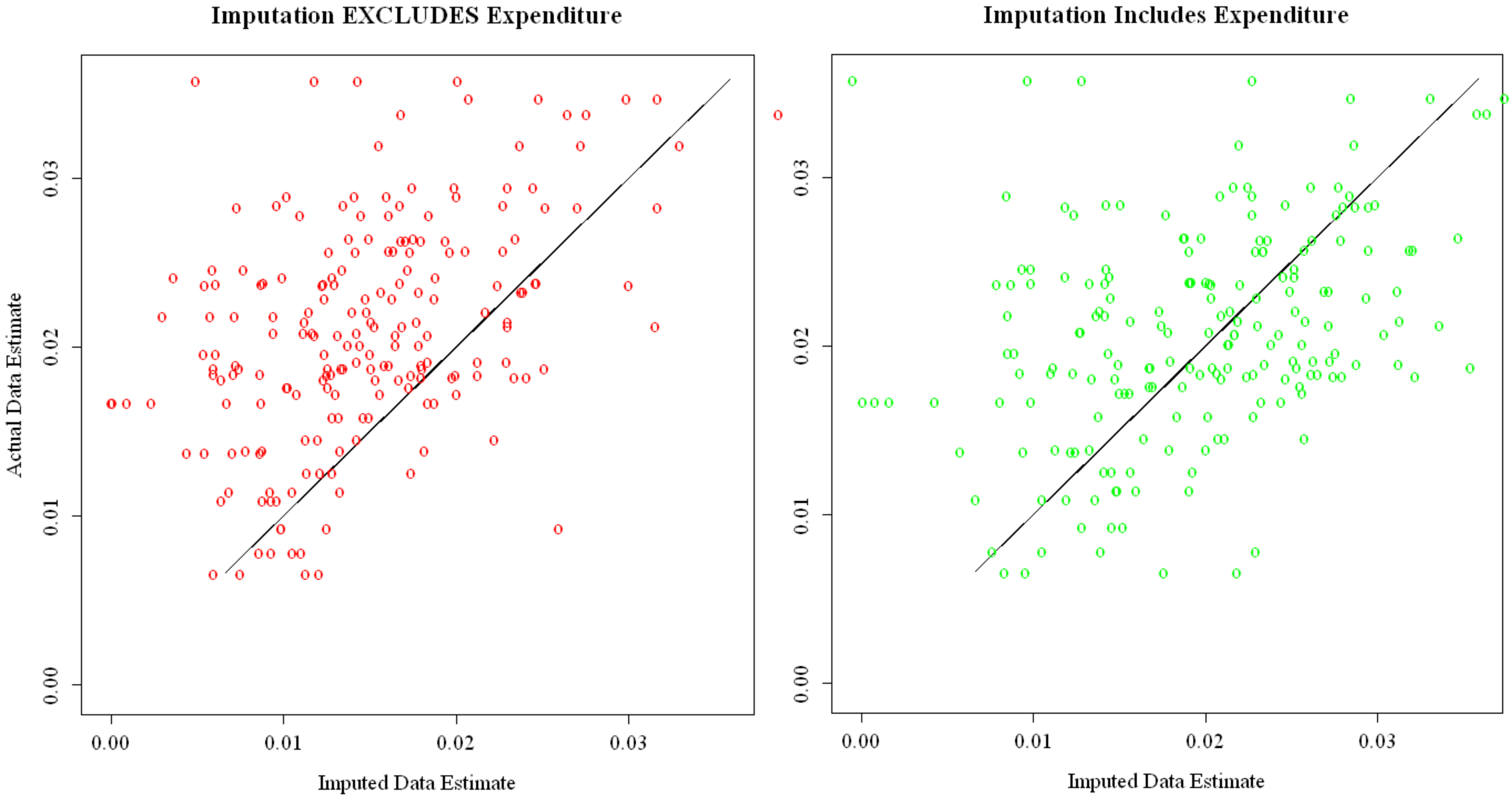
  Food-Away-Home= γ_0+γ_1 *Income + covariates*

  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

Two special imputation algorithms

- Last observation carried forward (LOCF) imputation for repeated measures with drop-outs:
  - impute last recorded value
  - implicit model: values are unchanged after drop-out
- Hot deck imputation (see Andridge and Little 10)
  - 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

Matching methods for hot deck imputation

- Nonrespondents *j* can be matched to respondents *i* based on a closeness metric *D*(*i*, *j*)
  - 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: $D(i,j)=(\hat{y}_i-\hat{y}_j)^T S_{Y\cdot X}^{-1}(\hat{y}_i-\hat{y}_j)$

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

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

Alternative to LOCF in Longitudinal Data Analysis

- For repeated measures, the following Row +/* Col methods includes individual (row) and time (column) effects

$$y_{it} = \text{value for subject } i, \text{ time } t$$

$$\hat{y}_{it} = m + a_i + b_t + r_{kt} \quad (\text{Row} + \text{Col})$$

$$\hat{y}_{it} = m \times a_i \times b_t \times r_{kt} \quad (\text{Row} * \text{Col})$$

m = grand mean

a_i = row effect, from deviations of row i

b_t = column effect, from deviations of col t

r_{kt} = residual from matched respondent k

Example 1: Imputing Income in a Panel Survey

- Survey of Income and Program Participation (SIPP): panel survey of income, interviews every 4 months
- Over 1000 variables in each wave: full imputation is too hard
- In practice weighting is used for wave nonresponse
  - discards wave data, inefficient use of information
- Illustrate imputation methods on single variable, monthly wages and salary from primary job

SIPP Data extract

12 month md pattern	Sample size	Mean monthly WS (available data) Mean	SD
0000 0000 0000	10534	1344	984
0001 0000 0000	30	924	936
0100 0000 0000	429	1355	883
1000 0000 0000	22	1245	943
1001 0000 0000	413	1292	924
1010 0000 0000	408	1277	843
1111 0000 0000	321	1339	889
0000 0000 1111	124	1895	1435
0000 1111 0000	81	1827	1360
0000 1111 1111	60	2734	1895
1111 0000 1111	43	1426	738
1111 1111 0000	66	1541	932
Other	98	------	------

Three incomplete cases

		Month											
ID	Mean	1	2	3	4	5	6	7	8	9	10	11	12
1	98	*	167	*	167	80	80	80	100	85	85	85	50
11	1180	*	*	*	*	1400	1750	1400	1400	970	776	776	970
21	3680	3680	*	3680	3680	3680	*	*	*	*	*	*	*

Imputes from Cross-Sectional Hot Deck

		Month											
ID	Mean	1	2	3	4	5	6	7	8	9	10	11	12
1	98	208	167	208	167	80	80	80	100	85	85	85	50
11	1180	900	720	900	720	1400	1750	1400	1400	970	776	776	970
21	3680	3680		3680	3680	3680	3082	2465	2465	3082	1332	1332	1666

1332

Imputes from Row*Col Fit

Month

ID Mean 1 2 3 4 5 6 7 8 9 10 11 12

1 98 199 167 0 167 80 80 80 100 85 85 85 50

11 1180 1126 1676 1126 1126 1400 1750 1400 1400 970 776 776
970

21 3680 3680 3680 3680 3680 3804 3804 3804 3804 3814 3814 3814
3814

Results from five methods

Deviations from row means of average WS estimates from five imputation methods

	Month												
Method	1	2	3	4	5	6	7	8	9	10	11	12	Mean
Comp Cases	-87	-46	-42	-25	-2	-2	-2	-6	59	68	46	41	1344
Avail Cases	-40	-29	-6	7	-3	-4	-11	-14	30	39	21	10	1352
Normal ML	-82	-50	-45	-24	9	9	2	0	50	59	40	29	1365
Row*Col Fit	-81	-50	-36	-23	9	9	4	-1	44	60	36	25	1365
CS Hot Deck	0	-17	-18	6	-7	-9	-16	-17	24	34	17	6	1379

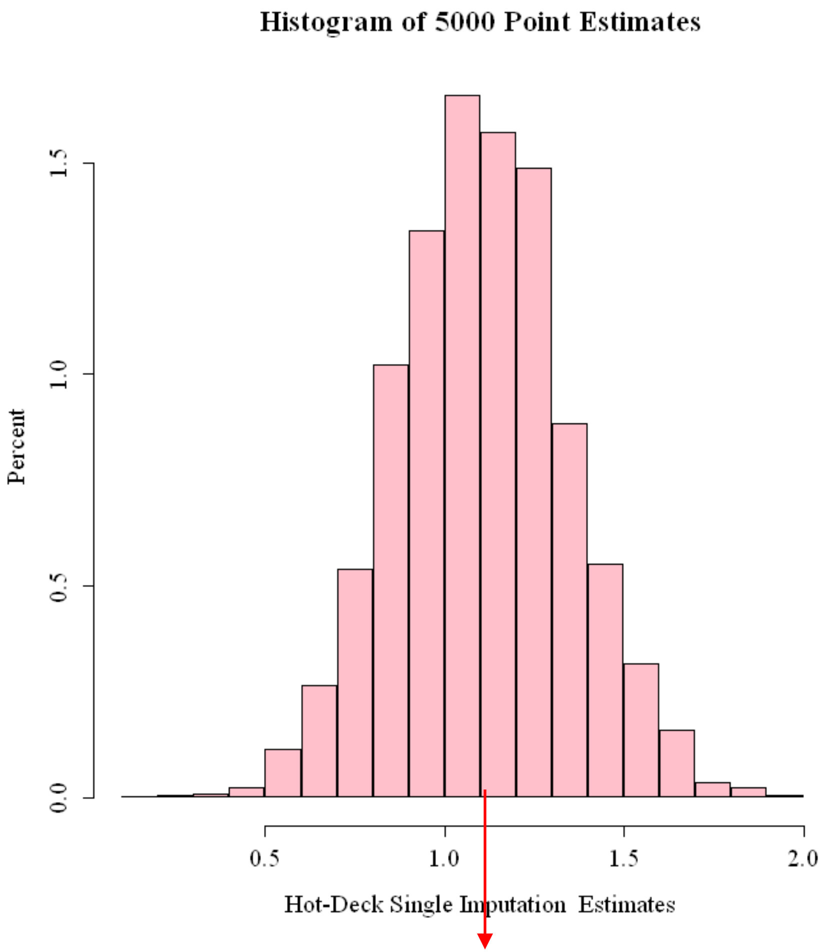
Normal ML and Row*Col results are similar -- both exploit available row and col information

These methods are more plausible and more similar to complete cases than others.

Example 2: Logistic simulation example

- Simple to create hot-deck imputations
- n_{ij} = Observed Sample size in cell $D=i,E=j$
- m_{ij} = Number of missing values
- Randomly draw m_{ij} values from n_{ij} observed values with replacement

Hot-deck Single Imputation Estimates



- Single Imputation
- Imputed Data Sets Analyzed as if Complete Data
- TRUE VALUE 1.1: estimates are unbiased

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