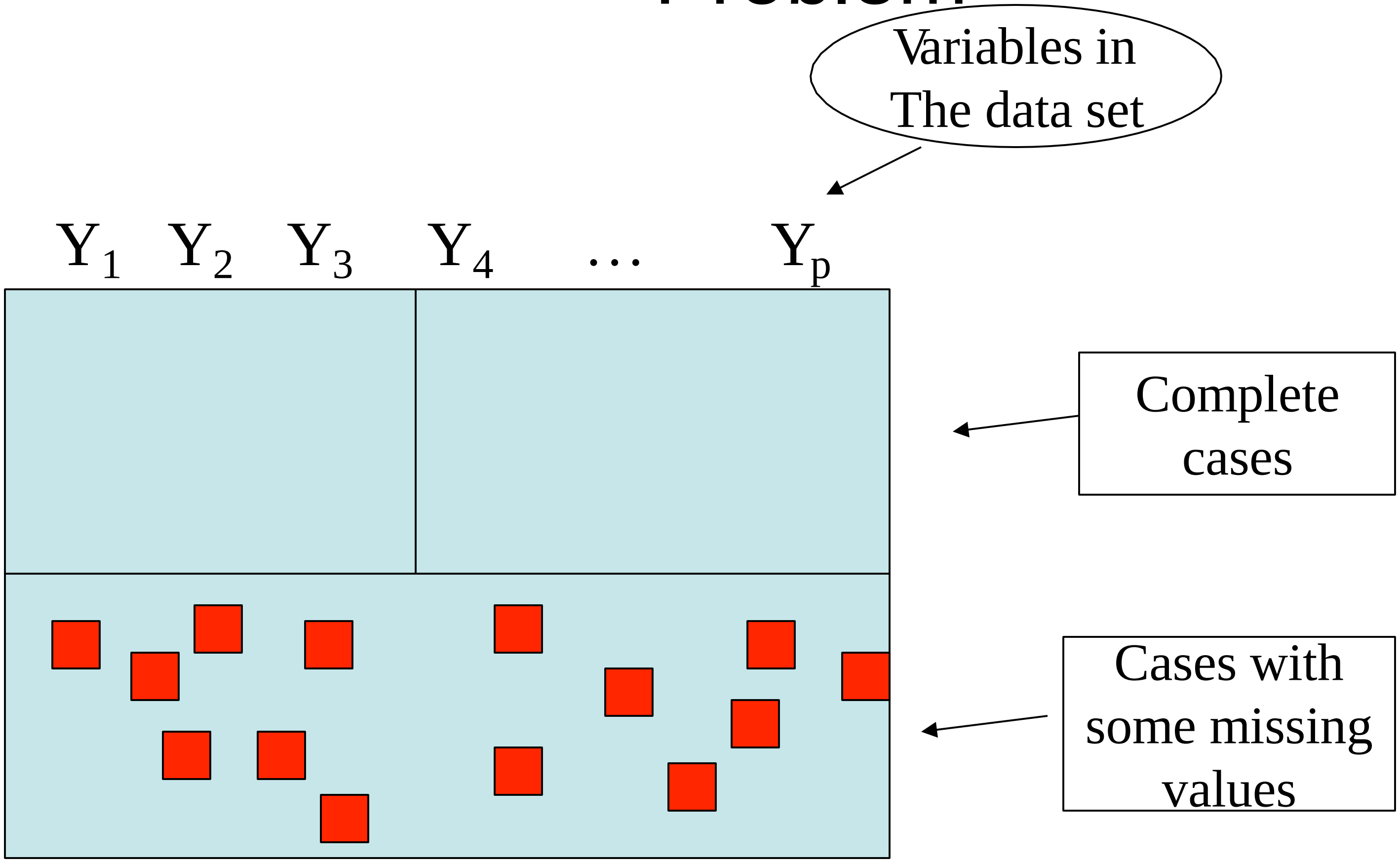
# Bayesian Inference for Surveys

Roderick Little and Trivellore Raghunathan

## Multiple Imputation using Sequential Regression/Chained Equations

UNIVERSITY OF MICHIGAN

# Problem



Variables in The data set

Complete cases

Cases with some missing values

$D_{obs}$ = Observed data: 🟦

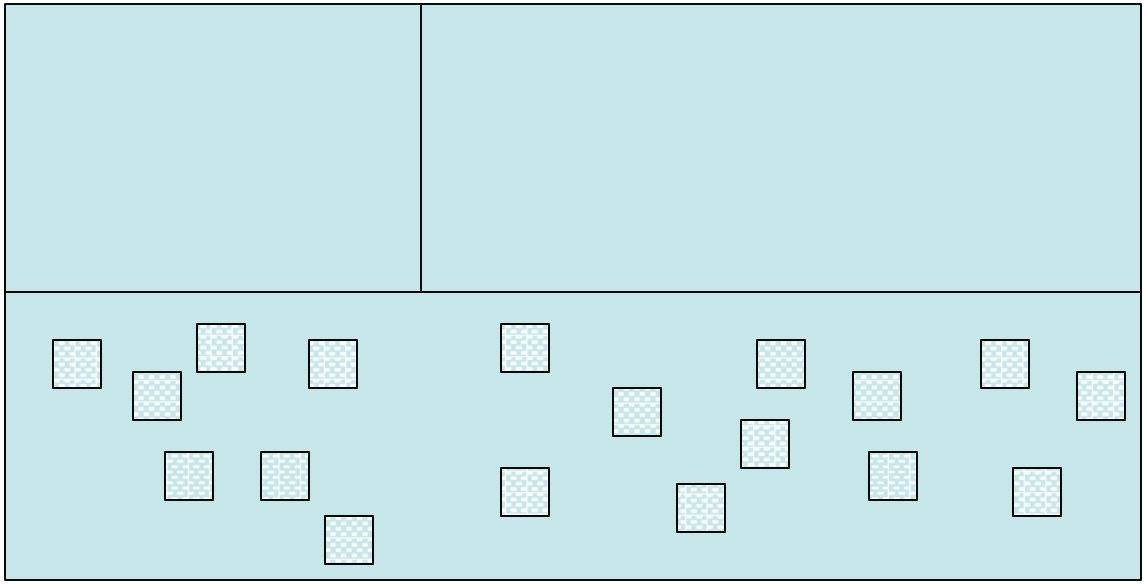
$D_{miss}$ = Missing data: 🟥

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

# 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



$$\text{"}\mathit{Ideal}\text{"}\ \mathit{imputations:}$$

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

Important issues:

Imputations are not real values

Uncertainties associated with imputes

# 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,….:

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

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

$$\vdots$$

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

$$\vdots$$

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

- Ability to specify individual regression model
- Types of variables
  - Continuous (Normal)
  - Categorical (Logistic or generalized logistic)
  - Count (Poisson)
  - Mixed or semi-continuous (Logistic/Normal)
  - Ordinal (ordered probit)
- 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

# Software

- Sequential regression imputations
  - R and Stata (MICE, ICE, MI)
  - Standalone (SRCWARE)
  - SAS (IveWare), PROC MI
- MI-Analysis
  - PROC MIANALYZE
  - IveWare (can handle complex sample survey)
  - SRCWARE
  - MICOMBINE/MITOOLS (STATA)
  - SUDAAN

# Software for Multiple Imputation Analysis

## For Creating Imputations

- SAS
  - PROC MI
  - IVEware
- Standalone
  - SRCware
- STATA
    - MI IMPUTE
    - IVEware
- R
    - MICE
    - IVEware
- SOLAS
- SPSS (Version 22)
  - IVEware

## For Analysis of multiply imputed data

- SAS
  - PROC MIANALYZE
  - IVEware
- Standalone
  - SRCware
- STATA
    - MI ESTIMATE
- SUDAAN
- R
- SPSS (Version 22)

# IveWare

- SAS, R, Stata, SPSS interface
  - A collection of C and Fortran routines
  - Handles linear (Continuous), logistic (Binary), multinomial logistic (categorical), Poisson (Count) and two-stage linear/logistic (Mixed or semi-continuous)
  - Predictive mean matching using Approximate Bayesian Bootstrap and 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

- Uses Normal approximation for the posterior distribution of the parameters
- Sampling Importance Resampling to handle non-normal posterior
- Non-informative prior
- Single chain or multiple chain (starting with different seeds)
- Iterations and Multiples control the length of the chain
- Built-in diagnostics to assess imputed values
- Complex Survey Data Analysis
- Download: www.isr.umich.edu/src/smp/ive/dev

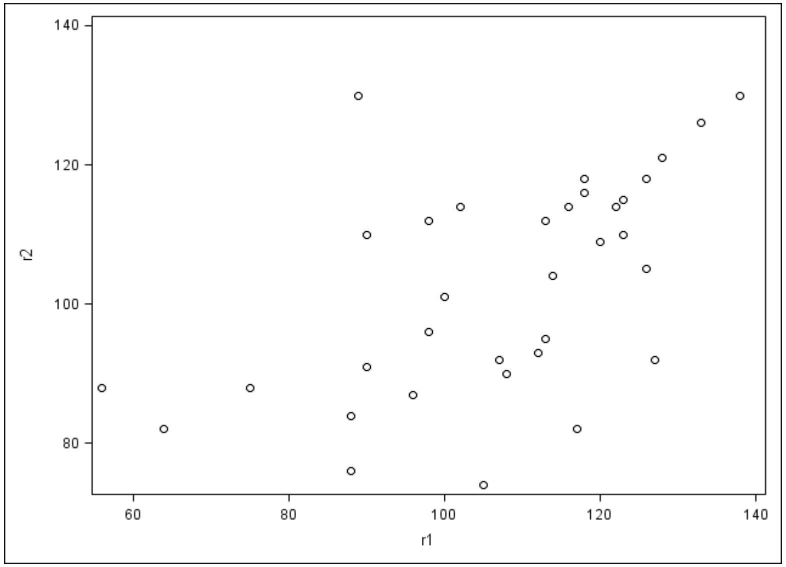
- Issues
  - Convergence
  - Several completed data statistics seem to converge to the same value regardless of seeds
  - Zhu and Raghunathan (JASA 2015) establish conditions for convergence
  - Good fitting models are needed to get results with desirable repeated sampling properties

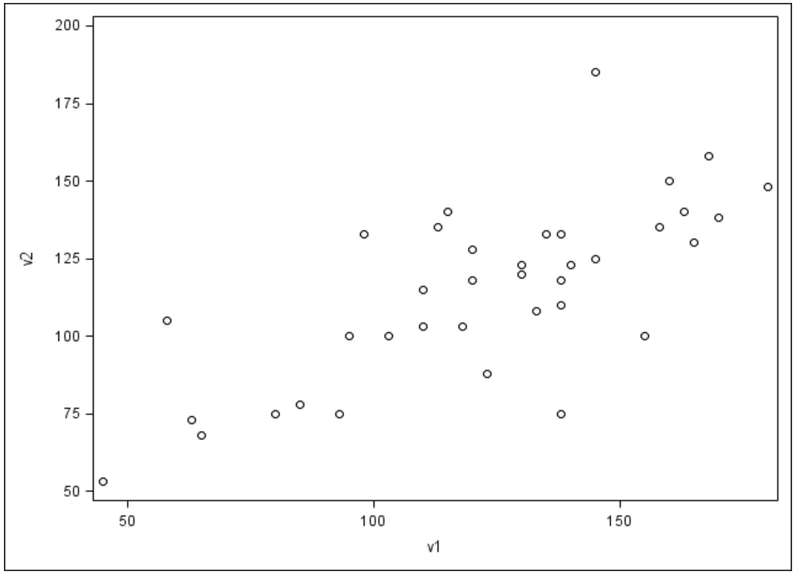
# St. Louis Risk Study

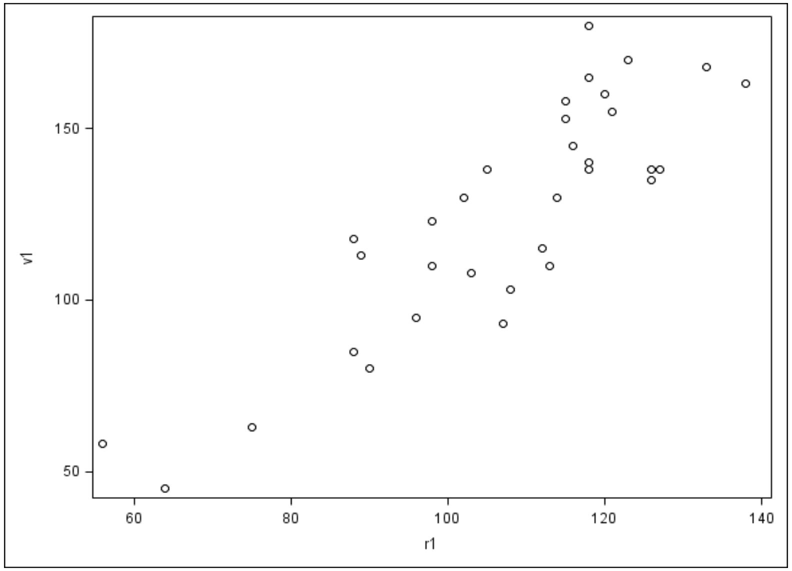
(Little and Rubin, 2002)

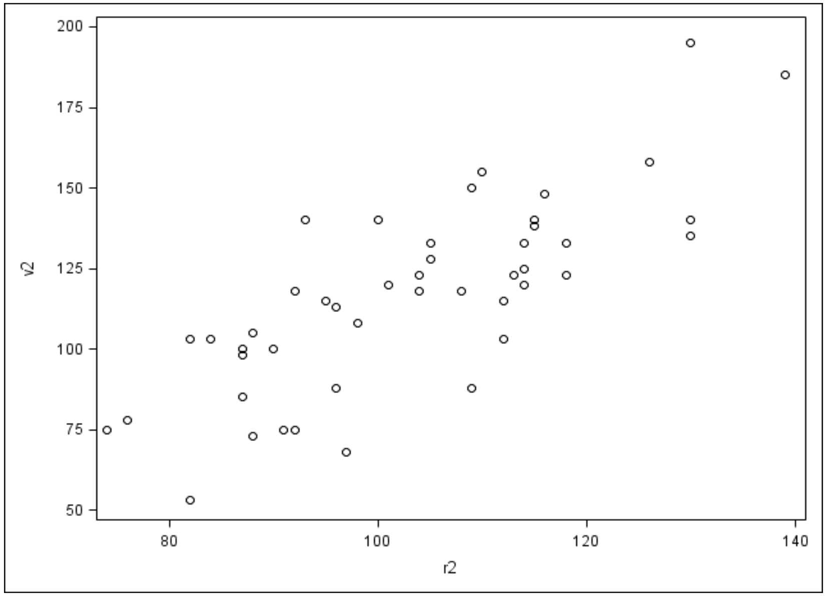
- A study was conducted to evaluate the effects parental psychological disorders on various aspects of the development of the children. Data from 69 families with two children were collected. Families were classified into risk group of the parent (G) with
  - G=1 normal or control group
  - G=2 Moderate risk group with one parent having some psychiatric illness
  - G=3 High risk group with one or more parent having schizophrenia or affective mental dis order

- Variables measured on Child 1
  - D1= Number of symptoms (1=Low, 2=High)
  - V1= Standardized verbal comprehension score
  - R1=Standardized Reading score
- Variables Measured on Child 2
  - D2, V2, R2
- G is always observed and other variables are missing with variety of different combinations


140
120
100
80
r2
60
80
100
120
140
r1


200
175
150
125
100
75
50
v2
50
100
150
v1


150
100
50
v1
60
80
100
120
140
r1


200
175
150
125
100
75
50
v2
80
100
120
140
r2

# St Louis Risk Study (Contd.)

- Sequential regression approach used to impute the missing values R1, R2, V1, V2 using normal linear regression model and D1, D2 using the logistic regression model
- Analysis
  - Regress R on G  and D , treating the Family ID as "cluster" or "Repeated" factor
  - Regress V on G and D, treating the Family ID as "cluster" or "Repeated" factor
  - Regress D on G, treating Family ID as "cluster" or "Repeated" factor

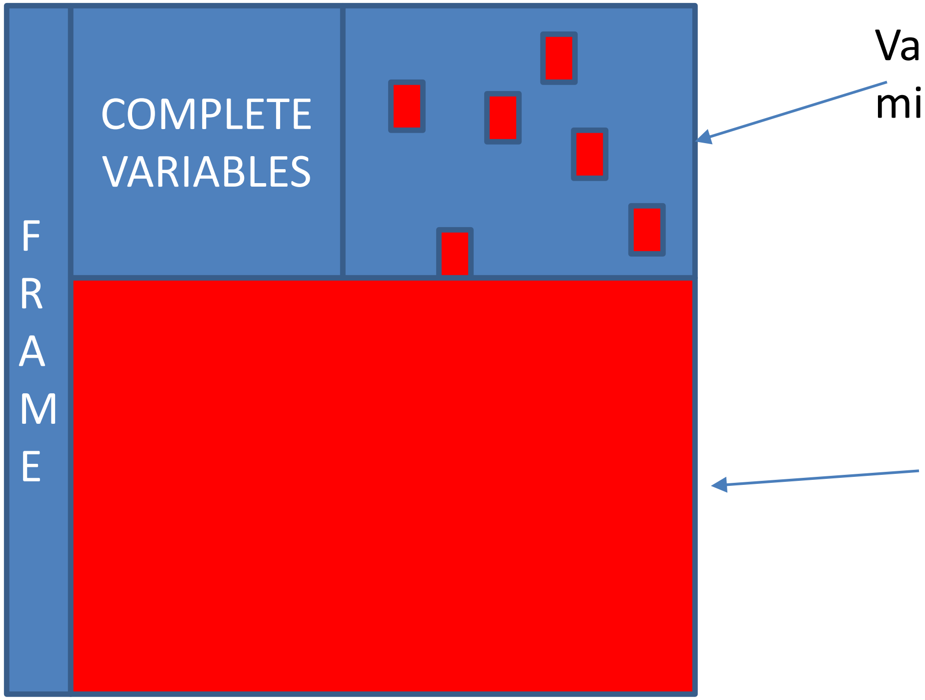
# Results

## Multiple Imputation Analysis

| Parameter | Reading | Verbal | Symptoms |
|---|---|---|---|
| Intercept | 114.23 (5.49) | 152.67 (15.50) | -0.32 (0.41) |
| Group 2 vs 1 | -9.85 (3.93) | -25.14 (13.56) | 1.05 (0.74) |
| Group 3 vs 1 | -9.69 (5.09) | -19.41 (11.03) | 0.47 (0.51) |
| Symptoms | -1.02 (3.35) | -10.86 (10.97) | |

# Prediction of the population

- Schematic Display



Variables missing values

Non-sampled subjects in the population

# Options

- Option 1: Impute the missing values in the sample and then predict the non-sampled portion of the population
- Option 2: Simultaneously impute all the missing values including the non-sampled portion of the population
- Model

$$\Pr(Y, I, M \mid Z) = \Pr(Y \mid Z)\Pr(I \mid Y, Z)\Pr(M \mid Y, I, Z)$$

$$= \Pr(Y \mid Z)\Pr(I \mid Z)\Pr(M \mid Y_{obs}, I, Z)$$

$$Y = \{Y_{obs}, Y_{mis}, Y_{exc}\}$$

- Option 1

$$\Pr(Y_{exc} \mid Y_{obs}, Y_{mis}) \Pr(Y_{mis} \mid Y_{obs})$$

  – Not all Y's to be predicted for the population

  – Simpler

- Option 2

$$\Pr(Y_{exc}, Y_{mis} \mid Y_{obs})$$

  – All Y's in the population are to be predicted

  – May be useful as a public-use file

# MI Applications

- 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 general location model.
- National Health Interview Survey 1997-Present
  - Multiple imputation of missing family income and personal earnings.
- Numerous applications in a variety of fields. Becoming a very common approach.

# Conclusion

- 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