Handling missing data in matched case-control studies using multiple imputation

Shaun Seaman

MRC Biostatistics Unit, Cambridge, UK

Ruth Keogh

Department of Medical Statistics

London School of Hygiene and Tropical Medicine

International Biometric Conference 2016

Victoria, Canada

Outline

1. Matched case-control studies
2. Motivating example:
   - matched case-control study of fibre intake and colorectal cancer
3. Previous methods for handling missing data in matched case-control studies
4. Two methods using MI
   - MI using matching variables
   - MI using matched sets
5. Simulations
6. Illustration in motivating example
7. Concluding remarks

Matched case-control studies

Matched case-control studies

- Used to investigate associations between disease and putative risk factors
- Each case is individually matched to *M* controls based on matching variables
- Matching is used to control for confounding at the design stage
- The study is formed of matched sets

Types of matching variables

1. Matching on 'simple' variables:
   - sex, age, smoking status
2. Matching on 'complex' variables:
   - family, GP practice, neighbourhood

Matched case-control studies

- Used to investigate associations between disease and putative risk factors
- Each case is individually matched to *M* controls based on matching variables
- Matching is used to control for confounding at the design stage
- The study is formed of matched sets

Types of matching variables

1. Matching on 'simple' variables:
   - sex, age, smoking status
2. Matching on 'complex' variables:
   - family, GP practice, neighbourhood

Matched case-control studies: Data and notation

Set	Individual j	D	X^{cat}	X^{con}
1	1	1	x_{11}^{cat}	x_{11}^{con}
1	2	0	x_{12}^{cat}	x_{12}^{con}
.	.	.	.	.
.	.	.	.	.
1	M+1	0	$x_{1,M+1}^{\text{cat}}$	$x_{1,M+1}^{\text{con}}$
2	1	1	x_{21}^{cat}	x_{21}^{con}
2	2	0	x_{22}^{cat}	x_{22}^{con}
.	.	.	.	.
.	.	.	.	.
2	M+1	0	$x_{2,M+1}^{\text{cat}}$	$x_{2,M+1}^{\text{con}}$
3	1	1	x_{31}^{cat}	x_{31}^{con}
3	2	0	x_{32}^{cat}	x_{32}^{con}
.	.	.	.	.
.	.	.	.	.
3	M+1	0	$x_{3,M+1}^{\text{cat}}$	$x_{3,M+1}^{\text{con}}$

More generally we allow vector covariates: $\boldsymbol{X}^{\text{cat}}, \boldsymbol{X}^{\text{con}}$.

The matching variables are denoted $\boldsymbol{S}$

Matched case-control studies: Analysis

Logistic regression model

$$\Pr(D=1|\boldsymbol{X}^{\text{cat}},\boldsymbol{X}^{\text{con}},\boldsymbol{S}) = \frac{\exp\{\boldsymbol{\beta}_{\text{cat}}^{\mathsf{T}}\boldsymbol{X}^{\text{cat}} + \boldsymbol{\beta}_{\text{con}}^{\mathsf{T}}\boldsymbol{X}^{\text{con}} + q(\boldsymbol{S})\}}{1+\exp\{\boldsymbol{\beta}_{\text{cat}}^{\mathsf{T}}\boldsymbol{X}^{\text{cat}} + \boldsymbol{\beta}_{\text{con}}^{\mathsf{T}}\boldsymbol{X}^{\text{con}} + q(\boldsymbol{S})\}}$$

Conditional logistic regression

Set	Individual j	D	X^{cat}	X^{con}
i	1	1	x_{i1}^{cat}	x_{i1}^{con}
i	2	0	x_{i2}^{cat}	x_{i2}^{con}
.	.	.	.	.
.	.	.	.	.
i	M+1	0	$x_{i,M+1}^{\text{cat}}$	$x_{i,M+1}^{\text{con}}$

$$\frac{\exp\{\boldsymbol{\beta}_{\text{cat}}^{\mathsf{T}}\boldsymbol{x}_{i1}^{\text{cat}} + \boldsymbol{\beta}_{\text{con}}^{\mathsf{T}}\boldsymbol{x}_{i1}^{\text{con}}\}}{\sum_{j=1}^{M+1}\exp\{\boldsymbol{\beta}_{\text{cat}}^{\mathsf{T}}\boldsymbol{x}_{ij}^{\text{cat}} + \boldsymbol{\beta}_{\text{con}}^{\mathsf{T}}\boldsymbol{x}_{ij}^{\text{con}}\}}$$

Matched case-control studies: Analysis

Logistic regression model

$$\Pr(D=1|\boldsymbol{X}^{\text{cat}},\boldsymbol{X}^{\text{con}},\boldsymbol{S}) = \frac{\exp\{\boldsymbol{\beta}_{\text{cat}}^{\mathsf{T}}\boldsymbol{X}^{\text{cat}} + \boldsymbol{\beta}_{\text{con}}^{\mathsf{T}}\boldsymbol{X}^{\text{con}} + q(\boldsymbol{S})\}}{1+\exp\{\boldsymbol{\beta}_{\text{cat}}^{\mathsf{T}}\boldsymbol{X}^{\text{cat}} + \boldsymbol{\beta}_{\text{con}}^{\mathsf{T}}\boldsymbol{X}^{\text{con}} + q(\boldsymbol{S})\}}$$

Conditional logistic regression

Set	Individual j	D	X^{cat}	X^{con}
i	1	1	x_{i1}^{cat}	x_{i1}^{con}
i	2	0	x_{i2}^{cat}	x_{i2}^{con}
.	.	.	.	.
.	.	.	.	.
i	M+1	0	$x_{i,M+1}^{\text{cat}}$	$x_{i,M+1}^{\text{con}}$

$$\frac{\exp\{\boldsymbol{\beta}_{\text{cat}}^{\mathsf{T}}\boldsymbol{x}_{i1}^{\text{cat}} + \boldsymbol{\beta}_{\text{con}}^{\mathsf{T}}\boldsymbol{x}_{i1}^{\text{con}}\}}{\sum_{j=1}^{M+1}\exp\{\boldsymbol{\beta}_{\text{cat}}^{\mathsf{T}}\boldsymbol{x}_{ij}^{\text{cat}} + \boldsymbol{\beta}_{\text{con}}^{\mathsf{T}}\boldsymbol{x}_{ij}^{\text{con}}\}}$$

Matched case-control studies: Missing data

Set	Individual j	D	X^{cat}	X^{con}
1	1	1	–	–
1	2	0	x_{12}^{cat}	x_{12}^{con}
.	.	.	.	.
.	.	.	.	.
1	M+1	0	$x_{1,M+1}^{\text{cat}}$	$x_{1,M+1}^{\text{con}}$
2	1	1	x_{21}^{cat}	–
2	2	0	x_{22}^{cat}	x_{22}^{con}
.	.	.	.	.
.	.	.	.	.
2	M+1	0	$x_{2,M+1}^{\text{cat}}$	$x_{2,M+1}^{\text{con}}$
3	1	1	x_{31}^{cat}	x_{31}^{con}
3	2	0	–	x_{32}^{con}
.	.	.	.	.
.	.	.	.	.
3	M+1	0	$x_{3,M+1}^{\text{cat}}$	$x_{3,M+1}^{\text{con}}$

Motivating example

- Matched case-control study nested within EPIC-Norfolk to study association between fibre intake and colorectal cancer

Explanatory variables

- Main exposure: fibre intake (g/day) from a 7-day diet diary
- Categorical potential confounders: smoking status (3 cats), education (4 cats), social class (6 cats), physical activity (4 cats), aspirin use (2 cats)
- Continuous potential confounders: height, weight, exact age, alcohol intake, folate intake, energy intake

Each case matched to 4 controls

sex, age (within 3 months), date of diary completion (within 3 months)

Motivating example

- Matched case-control study nested within EPIC-Norfolk to study association between fibre intake and colorectal cancer

Explanatory variables

- Main exposure: fibre intake (g/day) from a 7-day diet diary
- Categorical potential confounders: smoking status (3 cats), education (4 cats), social class (6 cats), physical activity (4 cats), aspirin use (2 cats)
- Continuous potential confounders: height, weight, exact age, alcohol intake, folate intake, energy intake

Each case matched to 4 controls

sex, age (within 3 months), date of diary completion (within 3 months)

Motivating example

- Matched case-control study nested within EPIC-Norfolk to study association between fibre intake and colorectal cancer

Explanatory variables

- Main exposure: fibre intake (g/day) from a 7-day diet diary
- Categorical potential confounders: smoking status (3 cats), education (4 cats), social class (6 cats), physical activity (4 cats), aspirin use (2 cats)
- Continuous potential confounders: height, weight, exact age, alcohol intake, folate intake, energy intake

Each case matched to 4 controls

sex, age (within 3 months), date of diary completion (within 3 months)

Motivating example: Missing data

- 318 cases, 1272 matched controls
- 328 individuals (20%) missing one or more adjustment variables
- Complete case analysis: uses only 240 matched sets
  - this is only 75% of matched sets
  - and 64% of individuals

Previous methods for handling missing data in matched case-control studies

- Lipsitz et al. (1998)
- Paik and Sacco (2000)
- Satten & Carroll (2000)
- Rathouz et al. (2002)
- Rathouz (2003)
- Paik (2004)
- Sinha et al. (2005)
- Sinha & Wang (2009)
- Gebregziabher & DeSantis (2010)
- Ahn et al. (2011)
- Liu et al. (2013)

Limitations of previous methods

- Assume only one partially observed covariate
- Assume partially observed covariates are collectively observed or missing on each individual
- Require parametric modelling of the matching variables
- Require bespoke computer code

Multiple imputation for matched case-control studies

Overview of Multiple imputation (MI)

1. Missing values are 'filled in' by sampling values from some appropriate distribution
2. This is performed K times to produce K imputed data sets
3. The analysis model is fitted in each imputed data set
4. Parameter and variance estimates are combined using 'Rubin's Rules'

We assume data are missing at random (MAR)

Overview of Multiple imputation (MI)

1. Missing values are 'filled in' by sampling values from some appropriate distribution
2. This is performed K times to produce K imputed data sets
3. The analysis model is fitted in each imputed data set
4. Parameter and variance estimates are combined using 'Rubin's Rules'

We assume data are missing at random (MAR)

Overview of Multiple imputation (MI)

1. Missing values are 'filled in' by sampling values from some appropriate distribution
2. This is performed K times to produce K imputed data sets
3. The analysis model is fitted in each imputed data set
4. Parameter and variance estimates are combined using 'Rubin's Rules'

We assume data are missing at random (MAR)

Overview of Multiple imputation (MI)

1. Missing values are 'filled in' by sampling values from some appropriate distribution
2. This is performed K times to produce K imputed data sets
3. The analysis model is fitted in each imputed data set
4. Parameter and variance estimates are combined using 'Rubin's Rules'

We assume data are missing at random (MAR)

Overview of Multiple imputation (MI)

1. Missing values are 'filled in' by sampling values from some appropriate distribution
2. This is performed K times to produce K imputed data sets
3. The analysis model is fitted in each imputed data set
4. Parameter and variance estimates are combined using 'Rubin's Rules'

We assume data are missing at random (MAR)

Advantages of using MI

- Many researchers familiar with the technique
- MI software readily available and easy to use
- Allows for multiple partially observed covariates without needing them to be collectively observed or missing
- Can incorporate information on auxiliary variables
- Reduces to conditional logistic regression when there are no missing data

Joint model MI *versus* Full conditional specification (FCS) MI

Joint model MI

- A Bayesian model is specified for the distribution of the partially observed variables given the fully observed variables

$$\boldsymbol{X}^{\text{cat}}, \boldsymbol{X}^{\text{con}} | D, \boldsymbol{S}$$

- Values for missing variables are sampled from their joint posterior predictive distribution

FCS MI

- A model is specified for the distribution of each partially missing variable conditional on all other variables

$$X^{\text{cat},k} | \boldsymbol{X}^{\text{cat},-k}, \boldsymbol{X}^{\text{con}}, D, \boldsymbol{S}$$

- FCS algorithm cycles through the imputation models until convergence is achieved

Joint model MI *versus* Full conditional specification (FCS) MI

Joint model MI

- A Bayesian model is specified for the distribution of the partially observed variables given the fully observed variables

$$\boldsymbol{X}^{\text{cat}}, \boldsymbol{X}^{\text{con}} | D, \boldsymbol{S}$$

- Values for missing variables are sampled from their joint posterior predictive distribution

FCS MI

- A model is specified for the distribution of each partially missing variable conditional on all other variables

$$X^{\text{cat},k} | \boldsymbol{X}^{\text{cat},-k}, \boldsymbol{X}^{\text{con}}, D, \boldsymbol{S}$$

- FCS algorithm cycles through the imputation models until convergence is achieved

‘Compatibility’ in MI

Imputation model

$$\boldsymbol{X}^{\text{cat}}, \boldsymbol{X}^{\text{con}} | D, \boldsymbol{S}$$

Analysis model: Conditional logistic regression

$$\frac{\exp\{\boldsymbol{\beta}_{\text{cat}}^{\mathsf{T}} \boldsymbol{x}_{i1}^{\text{cat}} + \boldsymbol{\beta}_{\text{con}}^{\mathsf{T}} \boldsymbol{x}_{i1}^{\text{con}}\}}{\sum_{j=1}^{M+1} \exp\{\boldsymbol{\beta}_{\text{cat}}^{\mathsf{T}} \boldsymbol{x}_{ij}^{\text{cat}} + \boldsymbol{\beta}_{\text{con}}^{\mathsf{T}} \boldsymbol{x}_{ij}^{\text{con}}\}}$$

Compatibility

- The imputation model and the analysis model are compatible if there exists a joint model for all variables which implies the imputation model and the analysis model as submodels.
- If the joint model and the analysis model are compatible, and the data are MAR, joint model MI gives consistent parameter and variance estimates.

‘Compatibility’ in MI

Imputation model

$$\boldsymbol{X}^{\text{cat}}, \boldsymbol{X}^{\text{con}} | D, \boldsymbol{S}$$

Analysis model: Conditional logistic regression

$$\frac{\exp\{\boldsymbol{\beta}_{\text{cat}}^{\mathsf{T}} \boldsymbol{x}_{i1}^{\text{cat}} + \boldsymbol{\beta}_{\text{con}}^{\mathsf{T}} \boldsymbol{x}_{i1}^{\text{con}}\}}{\sum_{j=1}^{M+1} \exp\{\boldsymbol{\beta}_{\text{cat}}^{\mathsf{T}} \boldsymbol{x}_{ij}^{\text{cat}} + \boldsymbol{\beta}_{\text{con}}^{\mathsf{T}} \boldsymbol{x}_{ij}^{\text{con}}\}}$$

Compatibility

- The imputation model and the analysis model are compatible if there exists a joint model for all variables which implies the imputation model and the analysis model as submodels.
- If the joint model and the analysis model are compatible, and the data are MAR, joint model MI gives consistent parameter and variance estimates.

‘Compatibility’ in MI

Joint model MI

$$\boldsymbol{X}^{\text{cat}}, \boldsymbol{X}^{\text{con}} | D, \boldsymbol{S}$$

FCS MI

$$X^{\text{cat},k} | \boldsymbol{X}^{\text{cat},-k}, \boldsymbol{X}^{\text{con}}, D, \boldsymbol{S}$$

Result of Liu et al 2014:

- The set of conditional models, $\{\mathscr{M}_k\}$, is compatible with a joint model, $\mathscr{M}_{\text{joint}}$, if:
  - for each $\mathscr{M}_k$ and every possible set of parameter values for that model , $\exists$ a set of parameter values for the joint model $\mathscr{M}_{\text{joint}}$ such that $\mathscr{M}_k$ and $\mathscr{M}_{\text{joint}}$ imply the same distribution for the dependent variable of $\mathscr{M}_k$
- If this holds, the distribution of imputed data from FCS MI converges asymptotically to the posterior predictive distribution of the missing data under joint model MI

‘Compatibility’ in MI

Joint model MI

$$\boldsymbol{X}^{\text{cat}}, \boldsymbol{X}^{\text{con}} | D, \boldsymbol{S}$$

FCS MI

$$X^{\text{cat},k} | \boldsymbol{X}^{\text{cat},-k}, \boldsymbol{X}^{\text{con}}, D, \boldsymbol{S}$$

Result of Liu et al 2014:

- The set of conditional models, $\{\mathscr{M}_k\}$, is compatible with a joint model, $\mathscr{M}_{\text{joint}}$, if:
  - for each $\mathscr{M}_k$ and every possible set of parameter values for that model , $\exists$ a set of parameter values for the joint model $\mathscr{M}_{\text{joint}}$ such that $\mathscr{M}_k$ and $\mathscr{M}_{\text{joint}}$ imply the same distribution for the dependent variable of $\mathscr{M}_k$
- If this holds, the distribution of imputed data from FCS MI converges asymptotically to the posterior predictive distribution of the missing data under joint model MI

MI for matched case-control studies

1. MI using matching variables
2. MI using matched set

MI for matched case-control studies

1. MI using matching variables
2. MI using matched set

MI using matching variables

Basis for MI using matching variables

Multiply impute $\boldsymbol{X}^{\text{cat}}$ and $\boldsymbol{X}^{\text{con}}$ from their conditional distribution given $D, \boldsymbol{S}$

- We outline 3 ways of modelling the distribution of $\boldsymbol{X}^{\text{cat}}, \boldsymbol{X}^{\text{con}} | D, \boldsymbol{S}$
- The matching between cases and control is 'broken' at the imputation stage
- But the matching is restored at the analysis stage and conditional logistic regression is applied to each imputed data set

MI using matching variables

Basis for MI using matching variables

Multiply impute $\boldsymbol{X}^{\text{cat}}$ and $\boldsymbol{X}^{\text{con}}$ from their conditional distribution given $D, \boldsymbol{S}$

- We outline 3 ways of modelling the distribution of $\boldsymbol{X}^{\text{cat}}, \boldsymbol{X}^{\text{con}} | D, \boldsymbol{S}$
- The matching between cases and control is 'broken' at the imputation stage
- But the matching is restored at the analysis stage and conditional logistic regression is applied to each imputed data set

MI using matching variables: Method 1

Model for categorical variables

$$\Pr(\boldsymbol{X}^{\text{cat}} = \boldsymbol{x}^{\text{cat}} | \boldsymbol{S}, D) = \frac{\exp\{\boldsymbol{\gamma}_0 \boldsymbol{x}^{\text{cat}} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_{\text{cat}} \boldsymbol{x}^{\text{cat}} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_S \boldsymbol{S} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_D D\}}{\sum_{\boldsymbol{x}^{\text{cat}'}} \exp\{\boldsymbol{\gamma}_0 \boldsymbol{x}^{\text{cat}'} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_{\text{cat}} \boldsymbol{x}^{\text{cat}'} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_S \boldsymbol{S} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_D D\}}$$

Model for continuous variables

$$\boldsymbol{X}^{\text{con}} | \boldsymbol{X}^{\text{cat}}, \boldsymbol{S}, D \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D + \boldsymbol{\gamma} \boldsymbol{X}^{\text{cat}} + \boldsymbol{\delta} \boldsymbol{S}, \boldsymbol{\Sigma})$$

We have shown that this model is compatible with the analysis model

MI using matching variables: Method 1

Model for categorical variables

$$\Pr(\boldsymbol{X}^{\text{cat}} = \boldsymbol{x}^{\text{cat}} | \boldsymbol{S}, D) = \frac{\exp\{\boldsymbol{\gamma}_0 \boldsymbol{x}^{\text{cat}} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_{\text{cat}} \boldsymbol{x}^{\text{cat}} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_S \boldsymbol{S} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_D D\}}{\sum_{\boldsymbol{x}^{\text{cat}'}} \exp\{\boldsymbol{\gamma}_0 \boldsymbol{x}^{\text{cat}'} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_{\text{cat}} \boldsymbol{x}^{\text{cat}'} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_S \boldsymbol{S} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_D D\}}$$

Model for continuous variables

$$\boldsymbol{X}^{\text{con}} | \boldsymbol{X}^{\text{cat}}, \boldsymbol{S}, D \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D + \boldsymbol{\gamma} \boldsymbol{X}^{\text{cat}} + \boldsymbol{\delta} \boldsymbol{S}, \boldsymbol{\Sigma})$$

We have shown that this model is compatible with the analysis model

MI using matching variables: Method 1

$$\Pr(\boldsymbol{X}^{\text{cat}} = \boldsymbol{x}^{\text{cat}} | \boldsymbol{S}, D) = \frac{\exp\{\boldsymbol{\gamma}_0 \boldsymbol{x}^{\text{cat}} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_{\text{cat}} \boldsymbol{x}^{\text{cat}} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_S \boldsymbol{S} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_D D\}}{\sum_{\boldsymbol{x}^{\text{cat}'}} \exp\{\boldsymbol{\gamma}_0 \boldsymbol{x}^{\text{cat}'} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_{\text{cat}} \boldsymbol{x}^{\text{cat}'} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_S \boldsymbol{S} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_D D\}}$$

$$\boldsymbol{X}^{\text{con}} | \boldsymbol{X}^{\text{cat}}, \boldsymbol{S}, D \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D + \boldsymbol{\gamma} \boldsymbol{X}^{\text{cat}} + \boldsymbol{\delta} \boldsymbol{S}, \boldsymbol{\Sigma})$$

Bayesian modelling software can be used to impute missing $\boldsymbol{X}^{\text{cat}}$ and $\boldsymbol{X}^{\text{con}}$ from the posterior predictive distribution implied by the above joint model.

FCS MI

Uses a set of fully conditional models which is compatible with the joint model.

$X^{\text{con},k}$: linear regression on $\boldsymbol{X}^{\text{cat}}, \boldsymbol{X}^{\text{con},-k}, D, \boldsymbol{S}$

$X^{\text{cat},k}$: multinomial logistic regression on $\boldsymbol{X}^{\text{cat},-k}, \boldsymbol{X}^{\text{con},}, D, \boldsymbol{S}$

These are the default options in many MI packages

MI using matching variables: Method 1

$$\Pr(\boldsymbol{X}^{\text{cat}} = \boldsymbol{x}^{\text{cat}} | \boldsymbol{S}, D) = \frac{\exp\{\boldsymbol{\gamma}_0 \boldsymbol{x}^{\text{cat}} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_{\text{cat}} \boldsymbol{x}^{\text{cat}} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_S \boldsymbol{S} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_D D\}}{\sum_{\boldsymbol{x}^{\text{cat}'}} \exp\{\boldsymbol{\gamma}_0 \boldsymbol{x}^{\text{cat}'} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_{\text{cat}} \boldsymbol{x}^{\text{cat}'} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_S \boldsymbol{S} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_D D\}}$$

$$\boldsymbol{X}^{\text{con}} | \boldsymbol{X}^{\text{cat}}, \boldsymbol{S}, D \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D + \boldsymbol{\gamma} \boldsymbol{X}^{\text{cat}} + \boldsymbol{\delta} \boldsymbol{S}, \boldsymbol{\Sigma})$$

Bayesian modelling software can be used to impute missing $\boldsymbol{X}^{\text{cat}}$ and $\boldsymbol{X}^{\text{con}}$ from the posterior predictive distribution implied by the above joint model.

FCS MI

Uses a set of fully conditional models which is compatible with the joint model.

$X^{\text{con},k}$: linear regression on $\boldsymbol{X}^{\text{cat}}, \boldsymbol{X}^{\text{con},-k}, D, \boldsymbol{S}$

$X^{\text{cat},k}$: multinomial logistic regression on $\boldsymbol{X}^{\text{cat},-k}, \boldsymbol{X}^{\text{con},}, D, \boldsymbol{S}$

These are the default options in many MI packages

MI using matching variables: Method 1

$$\Pr(\boldsymbol{X}^{\text{cat}} = \boldsymbol{x}^{\text{cat}} | \boldsymbol{S}, D) = \frac{\exp\{\boldsymbol{\gamma}_0 \boldsymbol{x}^{\text{cat}} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_{\text{cat}} \boldsymbol{x}^{\text{cat}} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_S \boldsymbol{S} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_D D\}}{\sum_{\boldsymbol{x}^{\text{cat}'}} \exp\{\boldsymbol{\gamma}_0 \boldsymbol{x}^{\text{cat}'} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_{\text{cat}} \boldsymbol{x}^{\text{cat}'} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_S \boldsymbol{S} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_D D\}}$$

$$\boldsymbol{X}^{\text{con}} | \boldsymbol{X}^{\text{cat}}, \boldsymbol{S}, D \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D + \boldsymbol{\gamma} \boldsymbol{X}^{\text{cat}} + \boldsymbol{\delta} \boldsymbol{S}, \boldsymbol{\Sigma})$$

Bayesian modelling software can be used to impute missing $\boldsymbol{X}^{\text{cat}}$ and $\boldsymbol{X}^{\text{con}}$ from the posterior predictive distribution implied by the above joint model.

FCS MI

Uses a set of fully conditional models which is compatible with the joint model.

$X^{\text{con},k}$: linear regression on $\boldsymbol{X}^{\text{cat}}, \boldsymbol{X}^{\text{con},-k}, D, \boldsymbol{S}$

$X^{\text{cat},k}$: multinomial logistic regression on $\boldsymbol{X}^{\text{cat},-k}, \boldsymbol{X}^{\text{con},}, D, \boldsymbol{S}$

These are the default options in many MI packages

MI using matching variables: Method 2

- Uses a latent normal model
- $\boldsymbol{W}^{\text{cat}}$: set of latent variables, one for each element of $\boldsymbol{X}^{\text{cat}}$
- $X^{\text{cat},k} = 1$ if $W^{\text{cat},k} > 0$

Latent normal model MI

$$\boldsymbol{X}^{\text{con}}, \boldsymbol{W}^{\text{cat}} | \boldsymbol{S}, D \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D + \boldsymbol{\delta} S, \boldsymbol{\Sigma})$$

Implementation

- `jomo` package in R
- REALCOM-MI
- realcomImpute: interface between Stata and REALCOM-MI

MI using matching variables: Method 2

- Uses a latent normal model
- $\boldsymbol{W}^{\text{cat}}$: set of latent variables, one for each element of $\boldsymbol{X}^{\text{cat}}$
- $X^{\text{cat},k} = 1$ if $W^{\text{cat},k} > 0$

Latent normal model MI

$$\boldsymbol{X}^{\text{con}}, \boldsymbol{W}^{\text{cat}} | \boldsymbol{S}, D \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D + \boldsymbol{\delta} S, \boldsymbol{\Sigma})$$

Implementation

- `jomo` package in R
- REALCOM-MI
- realcomImpute: interface between Stata and REALCOM-MI

MI using matching variables: Method 3

Method 2: Latent normal model MI

$$\boldsymbol{X}^{\text{con}}, \boldsymbol{W}^{\text{cat}} | \boldsymbol{S}, D \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D + \boldsymbol{\delta} S, \boldsymbol{\Sigma})$$

Method 3: Normal model MI

$$\boldsymbol{X}^{\text{con}}, \boldsymbol{X}^{\text{cat}} | \boldsymbol{S}, D \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D + \boldsymbol{\delta} S, \boldsymbol{\Sigma})$$

Imputed values of $\boldsymbol{X}^{\text{cat}}$ which are non-integer are handled using 'adaptive rounding'

Implementation

- `norm` package in R
- `mi mvn` in Stata

MI using matching variables: Method 3

Method 2: Latent normal model MI

$$\boldsymbol{X}^{\text{con}}, \boldsymbol{W}^{\text{cat}} | \boldsymbol{S}, D \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D + \boldsymbol{\delta} S, \boldsymbol{\Sigma})$$

Method 3: Normal model MI

$$\boldsymbol{X}^{\text{con}}, \boldsymbol{X}^{\text{cat}} | \boldsymbol{S}, D \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D + \boldsymbol{\delta} S, \boldsymbol{\Sigma})$$

Imputed values of $\boldsymbol{X}^{\text{cat}}$ which are non-integer are handled using 'adaptive rounding'

Implementation

- `norm` package in R
- `mi mvn` in Stata

MI using matching variables

Method 1: FCS MI

$X^{\text{con},k}$: linear regression on $\boldsymbol{X}^{\text{cat}}, \boldsymbol{X}^{\text{con},-k}, D, \boldsymbol{S}$

$X^{\text{cat},k}$: multinomial logistic regression on $\boldsymbol{X}^{\text{cat},-k}, \boldsymbol{X}^{\text{con},}, D, \boldsymbol{S}$

$$\Pr(\boldsymbol{X}^{\text{cat}} = \boldsymbol{x}^{\text{cat}} | \boldsymbol{S}, D) = \frac{\exp\{\boldsymbol{\gamma}_0 \boldsymbol{x}^{\text{cat}} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_{\text{cat}} \boldsymbol{x}^{\text{cat}} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_S \boldsymbol{S} + \boldsymbol{x}^{\text{cat}} \boldsymbol{\gamma}_D D\}}{\sum_{\boldsymbol{x}^{\text{cat}'}} \exp\{\boldsymbol{\gamma}_0 \boldsymbol{x}^{\text{cat}'} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_{\text{cat}} \boldsymbol{x}^{\text{cat}'} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_S \boldsymbol{S} + \boldsymbol{x}^{\text{cat}'} \boldsymbol{\gamma}_D D\}}$$

$$\boldsymbol{X}^{\text{con}} | \boldsymbol{X}^{\text{cat}}, \boldsymbol{S}, D \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D + \boldsymbol{\gamma} \boldsymbol{X}^{\text{cat}} + \boldsymbol{\delta} \boldsymbol{S}, \boldsymbol{\Sigma})$$

Method 2: Latent normal model MI

$$\boldsymbol{X}^{\text{con}}, \boldsymbol{W}^{\text{cat}} | \boldsymbol{S}, D \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D + \boldsymbol{\delta} S, \boldsymbol{\Sigma})$$

Method 3: Normal model MI

$$\boldsymbol{X}^{\text{con}}, \boldsymbol{X}^{\text{cat}} | \boldsymbol{S}, D \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D + \boldsymbol{\delta} S, \boldsymbol{\Sigma})$$

MI for matched case-control studies

1. MI using matching variables
2. MI using matched set

MI using matched set

Basis for MI using matched set

Multiply impute based on a model for

$$\boldsymbol{X}^{\text{set}} = (\boldsymbol{X}_1^{\text{cat}}, \boldsymbol{X}_1^{\text{con}}, \boldsymbol{X}_2^{\text{cat}}, \boldsymbol{X}_2^{\text{con}}, \ldots, \boldsymbol{X}_{M+1}^{\text{cat}}, \boldsymbol{X}_{M+1}^{\text{con}})$$

The imputation does not use the matching variables $\boldsymbol{S}$

Set	Individual j	D	X^{cat}	X^{con}
i	1	1	x_{i1}^{cat}	x_{i1}^{con}
i	2	0	x_{i2}^{cat}	x_{i2}^{con}
.	.	.	.	.
.	.	.	.	.
i	M+1	0	$x_{i,M+1}^{\text{cat}}$	$x_{i,M+1}^{\text{con}}$

- We outline 3 ways of modelling the distribution of $\boldsymbol{X}^{\text{set}}$
- The matching between cases and control is retained at both the imputation stage and the analysis stage

MI using matching variables vs MI using matched set

Basis for MI using matching variables

$$\boldsymbol{X}^{\text{cat}}, \boldsymbol{X}^{\text{con}} | D, \boldsymbol{S}$$

Basis for MI using matched set

$$\boldsymbol{X}^{\text{set}} = (\boldsymbol{X}_1^{\text{cat}}, \boldsymbol{X}_1^{\text{con}}, \boldsymbol{X}_2^{\text{cat}}, \boldsymbol{X}_2^{\text{con}}, \ldots, \boldsymbol{X}_{M+1}^{\text{cat}}, \boldsymbol{X}_{M+1}^{\text{con}})$$

Why use 'MI using matched set'?

- It may not be feasible/desired to specify effect of matching variables $\boldsymbol{S}$
- The analyst may not have information on $\boldsymbol{S}$
- The analysis model does not model the effect of $\boldsymbol{S}$

MI using matched set: Method 1

Model for categorical variables

$$\Pr(\boldsymbol{X}_1^{\text{cat}}, \ldots, \boldsymbol{X}_{M+1}^{\text{cat}}) \propto \exp\left\{ \sum_{j=1}^{M+1} \gamma_1 \boldsymbol{X}_j^{\text{cat}} + \sum_{j=1}^{M} \sum_{j'=2}^{M+1} \boldsymbol{X}_j^{\text{cat}} \gamma_2 \boldsymbol{X}_{j'}^{\text{cat}} + \boldsymbol{\tau} \boldsymbol{X}_1^{\text{cat}} \right\}$$

Model for continuous variables

$$\boldsymbol{X}_j^{\text{con}} | \boldsymbol{X}_1^{\text{cat}}, \ldots, \boldsymbol{X}_{M+1}^{\text{cat}}, u \sim N(\boldsymbol{\eta} + \boldsymbol{\xi}\, I(j=1) + \boldsymbol{\rho} \boldsymbol{X}_1^{\text{cat}} + \boldsymbol{\psi} \bar{\boldsymbol{X}}^{\text{cat}} + \boldsymbol{u}, \boldsymbol{\Lambda})$$

We have shown that this model is compatible with the analysis model

FCS MI

$X_j^{\text{con},k}$: linear regression on $\boldsymbol{X}_j^{\text{cat}}, \boldsymbol{X}_j^{\text{con},-k}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{cat}}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{con}}$

$X_j^{\text{cat},k}$: multinomial logistic reg on $\boldsymbol{X}_j^{\text{con}}, \boldsymbol{X}_j^{\text{cat},-k}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{cat}}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{con}}$

MI using matched set: Method 1

Model for categorical variables

$$\Pr(\boldsymbol{X}_1^{\text{cat}}, \ldots, \boldsymbol{X}_{M+1}^{\text{cat}}) \propto \exp\left\{ \sum_{j=1}^{M+1} \gamma_1 \boldsymbol{X}_j^{\text{cat}} + \sum_{j=1}^{M} \sum_{j'=2}^{M+1} \boldsymbol{X}_j^{\text{cat}} \gamma_2 \boldsymbol{X}_{j'}^{\text{cat}} + \boldsymbol{\tau} \boldsymbol{X}_1^{\text{cat}} \right\}$$

Model for continuous variables

$$\boldsymbol{X}_j^{\text{con}} | \boldsymbol{X}_1^{\text{cat}}, \ldots, \boldsymbol{X}_{M+1}^{\text{cat}}, u \sim N(\boldsymbol{\eta} + \boldsymbol{\xi} I(j=1) + \boldsymbol{\rho} \boldsymbol{X}_1^{\text{cat}} + \boldsymbol{\psi} \bar{\boldsymbol{X}}^{\text{cat}} + \boldsymbol{u}, \boldsymbol{\Lambda})$$

We have shown that this model is compatible with the analysis model

FCS MI

$X_j^{\text{con},k}$: linear regression on $\boldsymbol{X}_j^{\text{cat}}, \boldsymbol{X}_j^{\text{con},-k}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{cat}}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{con}}$

$X_j^{\text{cat},k}$: multinomial logistic reg on $\boldsymbol{X}_j^{\text{con}}, \boldsymbol{X}_j^{\text{cat},-k}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{cat}}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{con}}$

MI using matched set: Method 1

Model for categorical variables

$$\Pr(\boldsymbol{X}_1^{\text{cat}}, \ldots, \boldsymbol{X}_{M+1}^{\text{cat}}) \propto \exp\left\{ \sum_{j=1}^{M+1} \gamma_1 \boldsymbol{X}_j^{\text{cat}} + \sum_{j=1}^{M} \sum_{j'=2}^{M+1} \boldsymbol{X}_j^{\text{cat}} \gamma_2 \boldsymbol{X}_{j'}^{\text{cat}} + \boldsymbol{\tau} \boldsymbol{X}_1^{\text{cat}} \right\}$$

Model for continuous variables

$$\boldsymbol{X}_j^{\text{con}} | \boldsymbol{X}_1^{\text{cat}}, \ldots, \boldsymbol{X}_{M+1}^{\text{cat}}, u \sim N(\boldsymbol{\eta} + \boldsymbol{\xi}\, I(j=1) + \boldsymbol{\rho} \boldsymbol{X}_1^{\text{cat}} + \boldsymbol{\psi} \bar{\boldsymbol{X}}^{\text{cat}} + \boldsymbol{u}, \boldsymbol{\Lambda})$$

We have shown that this model is compatible with the analysis model

FCS MI

$X_j^{\text{con},k}$: linear regression on $\boldsymbol{X}_j^{\text{cat}}, \boldsymbol{X}_j^{\text{con},-k}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{cat}}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{con}}$

$X_j^{\text{cat},k}$: multinomial logistic reg on $\boldsymbol{X}_j^{\text{con}}, \boldsymbol{X}_j^{\text{cat},-k}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{cat}}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{con}}$

MI using matched set: Method 1

FCS MI

$X_j^{\text{con},k}$: linear regression on $\boldsymbol{X}_j^{\text{cat}}, \boldsymbol{X}_j^{\text{con},-k}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{cat}}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{con}}$

$X_j^{\text{cat},k}$: multinomial logistic reg on $\boldsymbol{X}_j^{\text{con}}, \boldsymbol{X}_j^{\text{cat},-k}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{cat}}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{con}}$

Set	Individual j	D	X^{cat}	X^{con}
i	1	1	x_{i1}^{cat}	x_{i1}^{con}
i	2	0	x_{i2}^{cat}	x_{i2}^{con}
.	.	.	.	.
.	.	.	.	.
i	M+1	0	$x_{i,M+1}^{\text{cat}}$	$x_{i,M+1}^{\text{con}}$

Set	X_1^{cat}	X_1^{con}	$\sum_{j \neq 1} X_j^{\text{cat}}$	$\sum_{j \neq 1} X_j^{\text{con}}$	X_2^{cat}	X_2^{con}	$\sum_{j \neq 2} X_j^{\text{cat}}$	$\sum_{j \neq 2} X_j^{\text{con}}$
i	x_{i1}^{cat}	x_{i1}^{con}	$\sum_{j \neq 1} x_{ij}^{\text{cat}}$	$\sum_{j \neq 1} x_{ij}^{\text{con}}$	x_{i2}^{cat}	x_{i2}^{con}	$\sum_{j \neq 2} x_{ij}^{\text{cat}}$	$\sum_{j \neq 2} x_{ij}^{\text{con}}$

Implementation: e.g. using `mice` in R, `mi impute` in Stata

MI using matched set: Method 1

FCS MI

$X_j^{\text{con},k}$: linear regression on $\boldsymbol{X}_j^{\text{cat}}, \boldsymbol{X}_j^{\text{con},-k}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{cat}}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{con}}$

$X_j^{\text{cat},k}$: multinomial logistic reg on $\boldsymbol{X}_j^{\text{con}}, \boldsymbol{X}_j^{\text{cat},-k}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{cat}}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{con}}$

Set	Individual j	D	X^{cat}	X^{con}
i	1	1	x_{i1}^{cat}	x_{i1}^{con}
i	2	0	x_{i2}^{cat}	x_{i2}^{con}
.	.	.	.	.
.	.	.	.	.
i	M+1	0	$x_{i,M+1}^{\text{cat}}$	$x_{i,M+1}^{\text{con}}$

Set	X_1^{cat}	X_1^{con}	$\sum_{j\neq 1} X_j^{\text{cat}}$	$\sum_{j\neq 1} X_j^{\text{con}}$	X_2^{cat}	X_2^{con}	$\sum_{j\neq 2} X_j^{\text{cat}}$	$\sum_{j\neq 2} X_j^{\text{con}}$
i	x_{i1}^{cat}	x_{i1}^{con}	$\sum_{j\neq 1} x_{ij}^{\text{cat}}$	$\sum_{j\neq 1} x_{ij}^{\text{con}}$	x_{i2}^{cat}	x_{i2}^{con}	$\sum_{j\neq 2} x_{ij}^{\text{cat}}$	$\sum_{j\neq 2} x_{ij}^{\text{con}}$

Implementation: e.g. using `mice` in R, `mi impute` in Stata

MI using matched set: Method 1

FCS MI

$X_j^{\text{con},k}$: linear regression on $\boldsymbol{X}_j^{\text{cat}}, \boldsymbol{X}_j^{\text{con},-k}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{cat}}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{con}}$

$X_j^{\text{cat},k}$: multinomial logistic reg on $\boldsymbol{X}_j^{\text{con}}, \boldsymbol{X}_j^{\text{cat},-k}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{cat}}, \sum_{j' \neq j} \boldsymbol{X}_{j'}^{\text{con}}$

Set	Individual j	D	X^{cat}	X^{con}
i	1	1	x_{i1}^{cat}	x_{i1}^{con}
i	2	0	x_{i2}^{cat}	x_{i2}^{con}
.	.	.	.	.
.	.	.	.	.
i	M+1	0	$x_{i,M+1}^{\text{cat}}$	$x_{i,M+1}^{\text{con}}$

Set	X_1^{cat}	X_1^{con}	$\sum_{j \neq 1} X_j^{\text{cat}}$	$\sum_{j \neq 1} X_j^{\text{con}}$	X_2^{cat}	X_2^{con}	$\sum_{j \neq 2} X_j^{\text{cat}}$	$\sum_{j \neq 2} X_j^{\text{con}}$
i	x_{i1}^{cat}	x_{i1}^{con}	$\sum_{j \neq 1} x_{ij}^{\text{cat}}$	$\sum_{j \neq 1} x_{ij}^{\text{con}}$	x_{i2}^{cat}	x_{i2}^{con}	$\sum_{j \neq 2} x_{ij}^{\text{cat}}$	$\sum_{j \neq 2} x_{ij}^{\text{con}}$

Implementation: e.g. using `mice` in R, `mi impute` in Stata

MI using matched set: Methods 2 and 3

Method 2: Latent normal model MI

$$X_j^{\text{con}}, W_j^{\text{cat}} | D_1 = 1, D_2 = \cdots = D_{M+1} = 0, u \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D_j + u, \boldsymbol{\Sigma})$$

Method 3: Normal model MI

$$X_j^{\text{con}}, X_j^{\text{cat}} | D_1 = 1, D_2 = \cdots = D_{M+1} = 0, u \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D_j + u, \boldsymbol{\Sigma})$$

Implementation

- Latent normal model MI: `jomo` in R, REALCOM-MI
- Normal model MI: `pan` in R

MI using matched set: Methods 2 and 3

Method 2: Latent normal model MI

$$X_j^{\text{con}}, W_j^{\text{cat}} | D_1 = 1, D_2 = \cdots = D_{M+1} = 0, u \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D_j + u, \boldsymbol{\Sigma})$$

Method 3: Normal model MI

$$X_j^{\text{con}}, X_j^{\text{cat}} | D_1 = 1, D_2 = \cdots = D_{M+1} = 0, u \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi} D_j + u, \boldsymbol{\Sigma})$$

Implementation

- Latent normal model MI: `jomo` in R, REALCOM-MI
- Normal model MI: `pan` in R

Simulation study

Simulation study

Two matching variables: S^{cat}, S^{con}

$$\Pr(S^{\text{cat}} = 1 | D = 1) = 0.6, \quad S^{\text{con}} | S^{\text{cat}}, D = 1 \sim N(0, 1)$$

Three covariates: X^{cat}, X^{conA}, X^{conB}

$$\text{logit}\ \Pr(X^{\text{cat}} | S^{\text{cat}}, S^{\text{con}}, D) = -2.5 + 0.5 S^{\text{cat}} + 0.5 S^{\text{con}} + 0.75 D$$

$$X^{\text{conA}} | X^{\text{cat}}, S^{\text{cat}}, S^{\text{con}}, D \sim N(0.5 X^{\text{cat}} + 0.5 S^{\text{cat}} + 0.5 S^{\text{con}} + 0.5 D, 1)$$

True log ORs: $\beta_{\text{cat}} = 5/12, \beta_{\text{conA}} = \beta_{\text{conB}} = 1/3$

- 100 or 500 matched sets
- 1 control or 4 controls per case
- 10% or 25% missing data in X^{cat}, X^{conA}, X^{conB}. MCAR or MAR.
- 1000 simulations, 50 imputations

Simulation study

Two matching variables: S^{cat}, S^{con}

$$\Pr(S^{\text{cat}} = 1 | D = 1) = 0.6, \quad S^{\text{con}} | S^{\text{cat}}, D = 1 \sim N(0, 1)$$

Three covariates: X^{cat}, X^{conA}, X^{conB}

$$\text{logit}\ \Pr(X^{\text{cat}} | S^{\text{cat}}, S^{\text{con}}, D) = -2.5 + 0.5 S^{\text{cat}} + 0.5 S^{\text{con}} + 0.75 D$$

$$X^{\text{conA}} | X^{\text{cat}}, S^{\text{cat}}, S^{\text{con}}, D \sim N(0.5 X^{\text{cat}} + 0.5 S^{\text{cat}} + 0.5 S^{\text{con}} + 0.5 D, 1)$$

True log ORs: $\beta_{\text{cat}} = 5/12, \beta_{\text{conA}} = \beta_{\text{conB}} = 1/3$

- 100 or 500 matched sets
- 1 control or 4 controls per case
- 10% or 25% missing data in X^{cat}, X^{conA}, X^{conB}. MCAR or MAR.
- 1000 simulations, 50 imputations

Simulation study

Two matching variables: S^{cat}, S^{con}

$$\Pr(S^{\text{cat}} = 1 | D = 1) = 0.6, \quad S^{\text{con}} | S^{\text{cat}}, D = 1 \sim N(0, 1)$$

Three covariates: X^{cat}, X^{conA}, X^{conB}

$$\text{logit}\ \Pr(X^{\text{cat}} | S^{\text{cat}}, S^{\text{con}}, D) = -2.5 + 0.5 S^{\text{cat}} + 0.5 S^{\text{con}} + 0.75 D$$

$$X^{\text{conA}} | X^{\text{cat}}, S^{\text{cat}}, S^{\text{con}}, D \sim N(0.5 X^{\text{cat}} + 0.5 S^{\text{cat}} + 0.5 S^{\text{con}} + 0.5 D, 1)$$

True log ORs: $\beta_{\text{cat}} = 5/12, \beta_{\text{conA}} = \beta_{\text{conB}} = 1/3$

- 100 or 500 matched sets
- 1 control or 4 controls per case
- 10% or 25% missing data in X^{cat}, X^{conA}, X^{conB}. MCAR or MAR.
- 1000 simulations, 50 imputations

Simulation study

Two matching variables: S^{cat}, S^{con}

$$\Pr(S^{\text{cat}} = 1 | D = 1) = 0.6, \quad S^{\text{con}} | S^{\text{cat}}, D = 1 \sim N(0, 1)$$

Three covariates: X^{cat}, X^{conA}, X^{conB}

$$\text{logit}\ \Pr(X^{\text{cat}} | S^{\text{cat}}, S^{\text{con}}, D) = -2.5 + 0.5 S^{\text{cat}} + 0.5 S^{\text{con}} + 0.75 D$$

$$X^{\text{conA}} | X^{\text{cat}}, S^{\text{cat}}, S^{\text{con}}, D \sim N(0.5 X^{\text{cat}} + 0.5 S^{\text{cat}} + 0.5 S^{\text{con}} + 0.5 D, 1)$$

True log ORs: $\beta_{\text{cat}} = 5/12, \beta_{\text{conA}} = \beta_{\text{conB}} = 1/3$

- 100 or 500 matched sets
- 1 control or 4 controls per case
- 10% or 25% missing data in X^{cat}, X^{conA}, X^{conB}. MCAR or MAR.
- 1000 simulations, 50 imputations

Simulation study

Two matching variables: S^{cat}, S^{con}

$$\Pr(S^{\text{cat}} = 1 | D = 1) = 0.6, \quad S^{\text{con}} | S^{\text{cat}}, D = 1 \sim N(0, 1)$$

Three covariates: X^{cat}, X^{conA}, X^{conB}

$$\text{logit}\ \Pr(X^{\text{cat}} | S^{\text{cat}}, S^{\text{con}}, D) = -2.5 + 0.5 S^{\text{cat}} + 0.5 S^{\text{con}} + 0.75 D$$

$$X^{\text{conA}} | X^{\text{cat}}, S^{\text{cat}}, S^{\text{con}}, D \sim N(0.5 X^{\text{cat}} + 0.5 S^{\text{cat}} + 0.5 S^{\text{con}} + 0.5 D, 1)$$

True log ORs: $\beta_{\text{cat}} = 5/12, \beta_{\text{conA}} = \beta_{\text{conB}} = 1/3$

- 100 or 500 matched sets
- 1 control or 4 controls per case
- 10% or 25% missing data in X^{cat}, X^{conA}, X^{conB}. MCAR or MAR.
- 1000 simulations, 50 imputations

Simulation study results

	X^{cat}			X^{conA}		
	LOR	SE	estSE	LOR	SE	estSE
Complete data	0.426	0.213	0.206	0.336	0.078	0.082
Complete cases	0.449	0.379	0.377	0.341	0.144	0.149
MI using matching variables						
Method 1: FCS	0.431	0.240	0.241	0.336	0.090	0.096
Method 2: Latent norm	0.446	0.238	0.241	0.322	0.085	0.095
Method 3: Normal	0.386	0.215	0.235	0.338	0.090	0.095
MI using matched set						
Method 1: FCS	0.430	0.247	0.243	0.335	0.094	0.097
Method 2: Latent norm	0.455	0.249	0.247	0.300	0.085	0.095
Method 3: Normal	0.407	0.238	0.251	0.350	0.098	0.101

LOR = mean estimated log OR

SE = empirical standard error

empSE = mean estimated standard error

Overview of simulation results

- All MI methods appear to work well
- MI using matching variables more efficient than MI using matched set
- FCS MI (Method 1) nearly always gave the least biased estimates
- MI using matching variables
  - latent normal MI and normal MI more efficient
- MI using matched set
  - FCS MI slightly better than latent normal and normal MI when 4:1 matching
  - no method obviously best or worst when 1:1 matching

Illustration

Motivating example

- Matched case-control study nested within EPIC-Norfolk to study association between fibre intake and colorectal cancer

Explanatory variables

- Main exposure: fibre intake (g/day) from a 7-day diet diary
- Categorical potential confounders: smoking status (3 cats), education (4 cats), social class (6 cats), physical activity (4 cats), aspirin use (2 cats)
- Continuous potential confounders: height, weight, exact age, alcohol intake, folate intake, energy intake

Each case matched to 4 controls

sex, age (within 3 months), date of diary completion (within 3 months)

Motivating example: results

Method	LOR	SE	p-value
Complete cases	−0.196	0.126	0.121
MI using matching variables			
Method 1: FCS	−0.176	0.104	0.090
Method 2: Latent normal	−0.176	0.104	0.089
Method 3: Normal	−0.177	0.104	0.088
MI using matched set			
Method 1: FCS	−0.175	0.104	0.092
Method 2: Latent normal	−0.174	0.104	0.094
Method 3: Normal	−0.181	0.104	0.082

Log odds ratio is for six-gram per day increase in fibre intake, conditional on the confounders.

Conclusions

- MI is a simple and versatile solution to problem of missing data in matched case-control studies.
- Proposed two overall approaches:
  - MI using matched set, MI using matching variables
- Three sub-methods:
  - FCS MI, Latent Normal MI, Normal MI
- FCS MI uses imputation model that is compatible with analysis model.
- The other methods use imputation models that are incompatible with analysis model. These use joint model MI.
- All methods can be applied in standard software.

Conclusions

- MI is a simple and versatile solution to problem of missing data in matched case-control studies.
- Proposed two overall approaches:
  - MI using matched set, MI using matching variables
- Three sub-methods:
  - FCS MI, Latent Normal MI, Normal MI
- FCS MI uses imputation model that is compatible with analysis model.
- The other methods use imputation models that are incompatible with analysis model. These use joint model MI.
- All methods can be applied in standard software.

Conclusions

- MI is a simple and versatile solution to problem of missing data in matched case-control studies.
- Proposed two overall approaches:
  - MI using matched set, MI using matching variables
- Three sub-methods:
  - FCS MI, Latent Normal MI, Normal MI
- FCS MI uses imputation model that is compatible with analysis model.
- The other methods use imputation models that are incompatible with analysis model. These use joint model MI.
- All methods can be applied in standard software.

Conclusions

- MI is a simple and versatile solution to problem of missing data in matched case-control studies.
- Proposed two overall approaches:
  - MI using matched set, MI using matching variables
- Three sub-methods:
  - FCS MI, Latent Normal MI, Normal MI
- FCS MI uses imputation model that is compatible with analysis model.
- The other methods use imputation models that are incompatible with analysis model. These use joint model MI.
- All methods can be applied in standard software.

Reference

Seaman, S.R. and Keogh, R.H. Handling missing data in matched case-control studies using multiple imputation. Biometrics 2015; 71(4): 1150-1159.