Handling Missing Categorical Data in a Multi-level Structure

Dr Carole L. Birrell

NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

UNIVERSITY OF WOLLONGONG AUSTRALIA

Frontiers in Social Statistics Methodology
UTS, Sydney, 8th Feb 2017

Overview

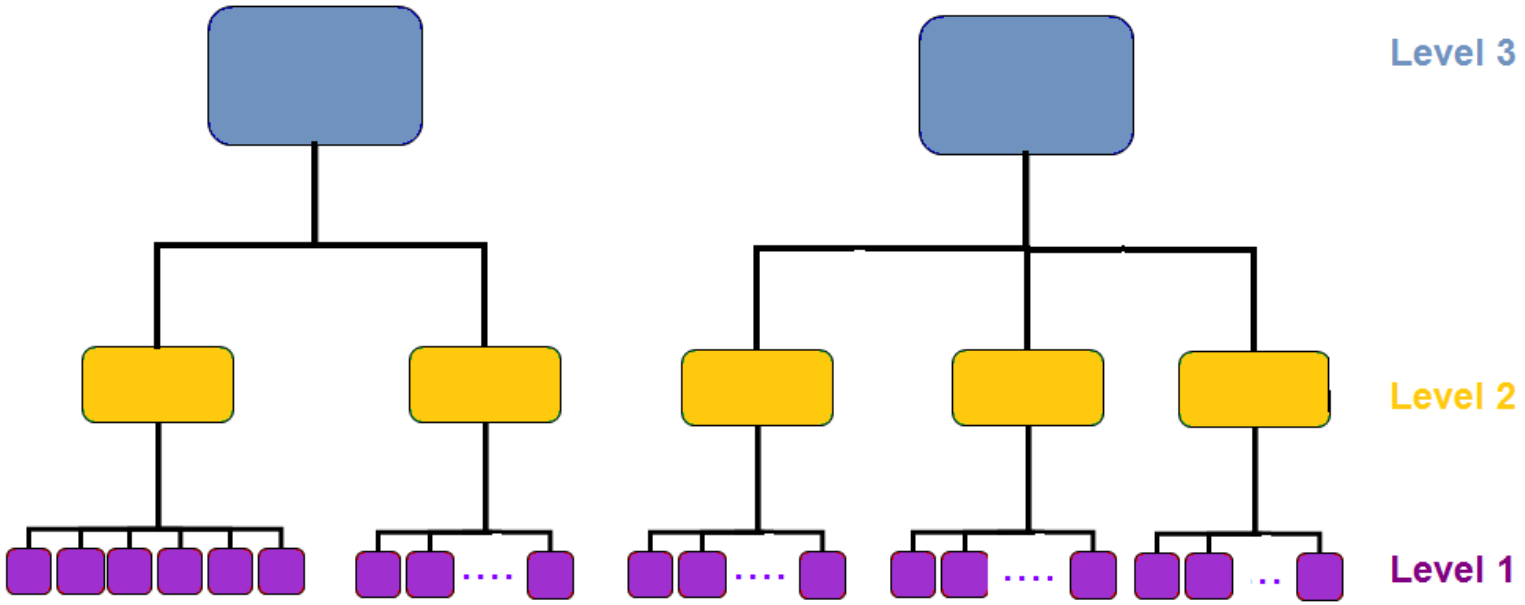
1. Introduction
2. Imputation
3. MI Approaches
4. Ex 1
5. Ex 2
6. Future Directions

Multilevel Data

It is a **hierarchical** data structure commonly found in social and health statistics data

Observations on individual units cannot be considered independent $\Rightarrow$ **correlated** within clusters

Missing data: could occur at any level and in any type of variable



Eg: Snapshot of data table

The structure or pattern of item non-response is often complex: Generalized pattern

SCH ID	Stud ID	AvStd Score	P1 SchEd	P1 NonSch Ed	P1 Occup	P2 SchEd	P2 NonSch Ed	P2 Occup	ATSI	CV_SCH
1	1	√	√	√	√	√	√	√	√	√
1	2	√	√	√	√	√	√	√	√	√
1	3	√	√	√	√	√	√	√	√	√
1	4	√	√	√	√	√	√	√	√	√
1	5	√	√	√	√	√	√	√	√	√
1	6	√	√		√	√	√	√	√	√
1	7	√		√	√	√			√	√
1	8	√	√	√			√	√	√	√
1	9		√		√		√	√	√	√
1	10	√	√		√	√	√			√

⋮

Missing data could be in categorical, ordinal, discrete or continuous variables.

Missing data mechanisms

Rubin (1976) defined **3 classes of missing data mechanisms**:

E.g. If X_1, X_2 are fully observed and X_3 has missing values:

1. **Missing Completely at Random (MCAR)**
   - missingness is completely unsystematic
   - missingness in X_3 does not depend on X_1, X_2 or X_3 itself
2. **Missing at Random (MAR)**
   - missingness is related to one or more other measured variables but not to the underlying values of itself
   - missingness in X_3 may depend on X_1 and/or X_2, but not on X_3
3. **Not Missing at Random (NMAR)**
   - missingness is related to missing values in that same variable, even after controlling for other variables
   - missingness in X_3 may depend on X_1 or X_2, but necessarily depends on X_3 itself

Missing data mechanisms

Rubin (1976) defined **3 classes of missing data mechanisms**:

E.g. If X_1, X_2 are fully observed and X_3 has missing values:

1. **Missing Completely at Random (MCAR)**
   - missingness is completely unsystematic
   - missingness in X_3 does not depend on X_1, X_2 or X_3 itself
2. **Missing at Random (MAR)**
   - missingness is related to one or more other measured variables but not to the underlying values of itself
   - missingness in X_3 may depend on X_1 and/or X_2, but not on X_3
3. **Not Missing at Random (NMAR)**
   - missingness is related to missing values in that same variable, even after controlling for other variables
   - missingness in X_3 may depend on X_1 or X_2, but necessarily depends on X_3 itself

Missing data mechanisms

Rubin (1976) defined **3 classes of missing data mechanisms**:

E.g. If X_1, X_2 are fully observed and X_3 has missing values:

1. **Missing Completely at Random (MCAR)**
   - missingness is completely unsystematic
   - missingness in X_3 does not depend on X_1, X_2 or X_3 itself
2. **Missing at Random (MAR)**
   - missingness is related to one or more other measured variables but not to the underlying values of itself
   - missingness in X_3 may depend on X_1 and/or X_2, but not on X_3
3. **Not Missing at Random (NMAR)**
   - missingness is related to missing values in that same variable, even after controlling for other variables
   - missingness in X_3 may depend on X_1 or X_2, but necessarily depends on X_3 itself

Missing data mechanisms

Rubin (1976) defined **3 classes of missing data mechanisms**:

E.g. If X_1, X_2 are fully observed and X_3 has missing values:

1. **Missing Completely at Random (MCAR)**
   - missingness is completely unsystematic
   - missingness in X_3 does not depend on X_1, X_2 or X_3 itself
2. **Missing at Random (MAR)**
   - missingness is related to one or more other measured variables but not to the underlying values of itself
   - missingness in X_3 may depend on X_1 and/or X_2, but not on X_3
3. **Not Missing at Random (NMAR)**
   - missingness is related to missing values in that same variable, even after controlling for other variables
   - missingness in X_3 may depend on X_1 or X_2, but necessarily depends on X_3 itself

Handling Missing Data - What to do?

Ignore them? **If missing data are not addressed**

- data set restricted to
  - complete cases - listwise deletion
  - available cases - pairwise deletion
- reduction in amount of data cases available for analysis
- requires MCAR assumption to produce unbiased estimates
- standard errors too large (listwise) or complicated (pairwise)
- valid inference under MAR not obtained unless
  - the only missing values are among the responses and
  - we have included the fully observed variables required for MAR to hold as covariates in the MOI.

Fill them in? How? Missing data can be imputed

- with a single value - single imputation
- with multiple values - multiple imputation

Imputations are not real values, the goal is to obtain valid inferences not to recover missing values (Rubin, 1987)

Handling Missing Data - What to do?

Ignore them? **If missing data are not addressed**

- data set restricted to
  - complete cases - listwise deletion
  - available cases - pairwise deletion
- reduction in amount of data cases available for analysis
- requires MCAR assumption to produce unbiased estimates
- standard errors too large (listwise) or complicated (pairwise)
- valid inference under MAR not obtained unless
  - the only missing values are among the responses and
  - we have included the fully observed variables required for MAR to hold as covariates in the MOI.

Fill them in? How? **Missing data can be imputed**

- with a single value - single imputation
- with multiple values - multiple imputation

Imputations are not real values: the **goal** *is to obtain* **valid inferences** *not to recover missing values (Rubin, 2011)*

Handling Missing Data - What to do?

Ignore them? **If missing data are not addressed**

- data set restricted to
  - complete cases - listwise deletion
  - available cases - pairwise deletion
- reduction in amount of data cases available for analysis
- requires MCAR assumption to produce unbiased estimates
- standard errors too large (listwise) or complicated (pairwise)
- valid inference under MAR not obtained unless
  - the only missing values are among the responses and
  - we have included the fully observed variables required for MAR to hold as covariates in the MOI.

Fill them in? How? **Missing data can be imputed**

- with a single value - single imputation
- with multiple values - multiple imputation

Imputations are not real values: the **goal** *is to obtain* **valid inferences** *not to recover missing values (Rubin, 2011)*

Single Imputation

- Imputes **one value** for each missing value:
  - Mean imputation: $\hat{Y}_i = \bar{Y}$
  - Regression imputation: eg $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$
  - Stochastic regression imputation: eg $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + r_i$ where $r_i \sim N(0, \hat{\sigma}^2)$
- **Advantages**
  - Result is one imputed data set - Standard complete data analysis methods can then be used.
  - Carried out once - convenient & easy
- **Disadvantages**
  - Requires MCAR (Mean Imp) or MAR (Regr Imp) to produce unbiased estimates
  - Sample size is overstated
  - Understates uncertainty i.e. SEs too small
  - Does not provide information to achieve valid inferences

Single Imputation

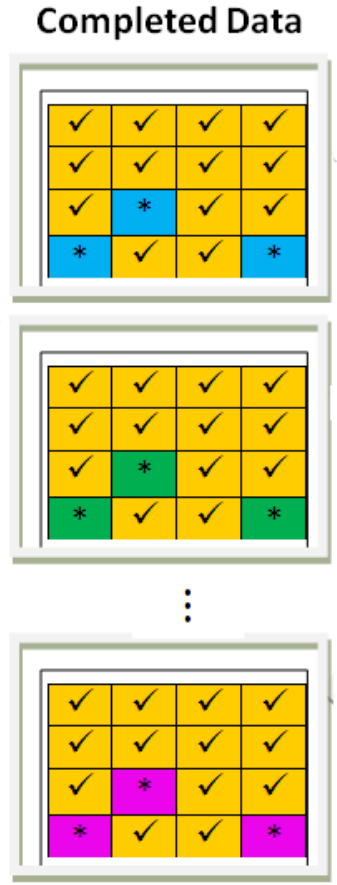
- Imputes **one value** for each missing value:
  - Mean imputation: $\hat{Y}_i = \bar{Y}$
  - Regression imputation: eg $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$
  - Stochastic regression imputation: eg $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + r_i$ where $r_i \sim N(0, \hat{\sigma}^2)$
- **Advantages**
  - Result is one imputed data set - Standard complete data analysis methods can then be used.
  - Carried out once - convenient & easy
- **Disadvantages**
  - Requires MCAR (Mean Imp) or MAR (Regr Imp) to produce unbiased estimates
  - Sample size is overstated
  - Understates uncertainty i.e. SEs too small
  - Does not provide information to achieve valid inferences

Single Imputation

- Imputes **one value** for each missing value:
  - Mean imputation: $\hat{Y}_i = \bar{Y}$
  - Regression imputation: eg $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$
  - Stochastic regression imputation: eg $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + r_i$ where $r_i \sim N(0, \hat{\sigma}^2)$
- **Advantages**
  - Result is one imputed data set - Standard complete data analysis methods can then be used.
  - Carried out once - convenient & easy
- **Disadvantages**
  - Requires MCAR (Mean Imp) or MAR (Regr Imp) to produce unbiased estimates
  - Sample size is overstated
  - Understates uncertainty i.e. SEs too small
  - Does not provide information to achieve valid inferences

Multiple Imputation

- Replaces each missing value with **several** acceptable values from a distribution of possibilities
- Generates a no. of different completed data sets
- Completed data sets used in complete case analysis
- Important goal is to provide statistically valid inferences

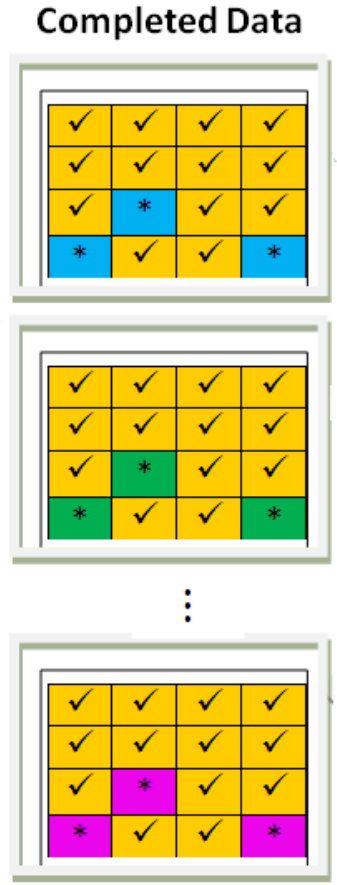


Assumptions

- Missing data are assumed MAR
- An imputation model is established relating unobserved to observed data

Multiple Imputation

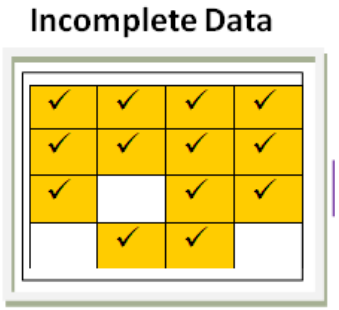
- Replaces each missing value with **several** acceptable values from a distribution of possibilities
- Generates a no. of different completed data sets
- Completed data sets used in complete case analysis
- Important goal is to provide statistically valid inferences



Assumptions

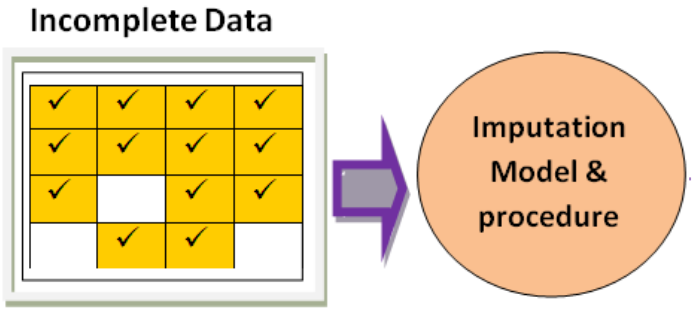
- Missing data are assumed MAR
- An imputation model is established relating unobserved to observed data

Overview of Multiple Imputation Process



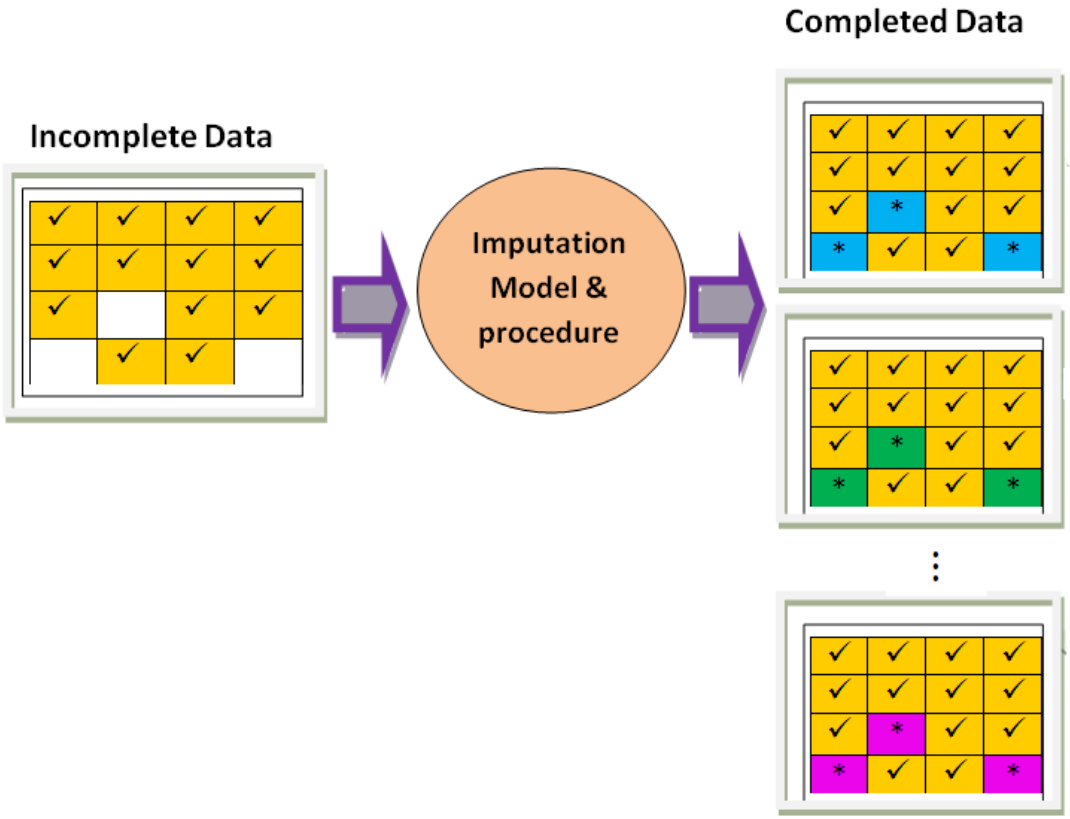
The set of variables in the imputation model should be larger and broader in scope than that required for ultimate analytic model applied to the completed data. This should assist in moving towards the MAR assumption and reduce any possible bias. (White et al, 2011)

Overview of Multiple Imputation Process



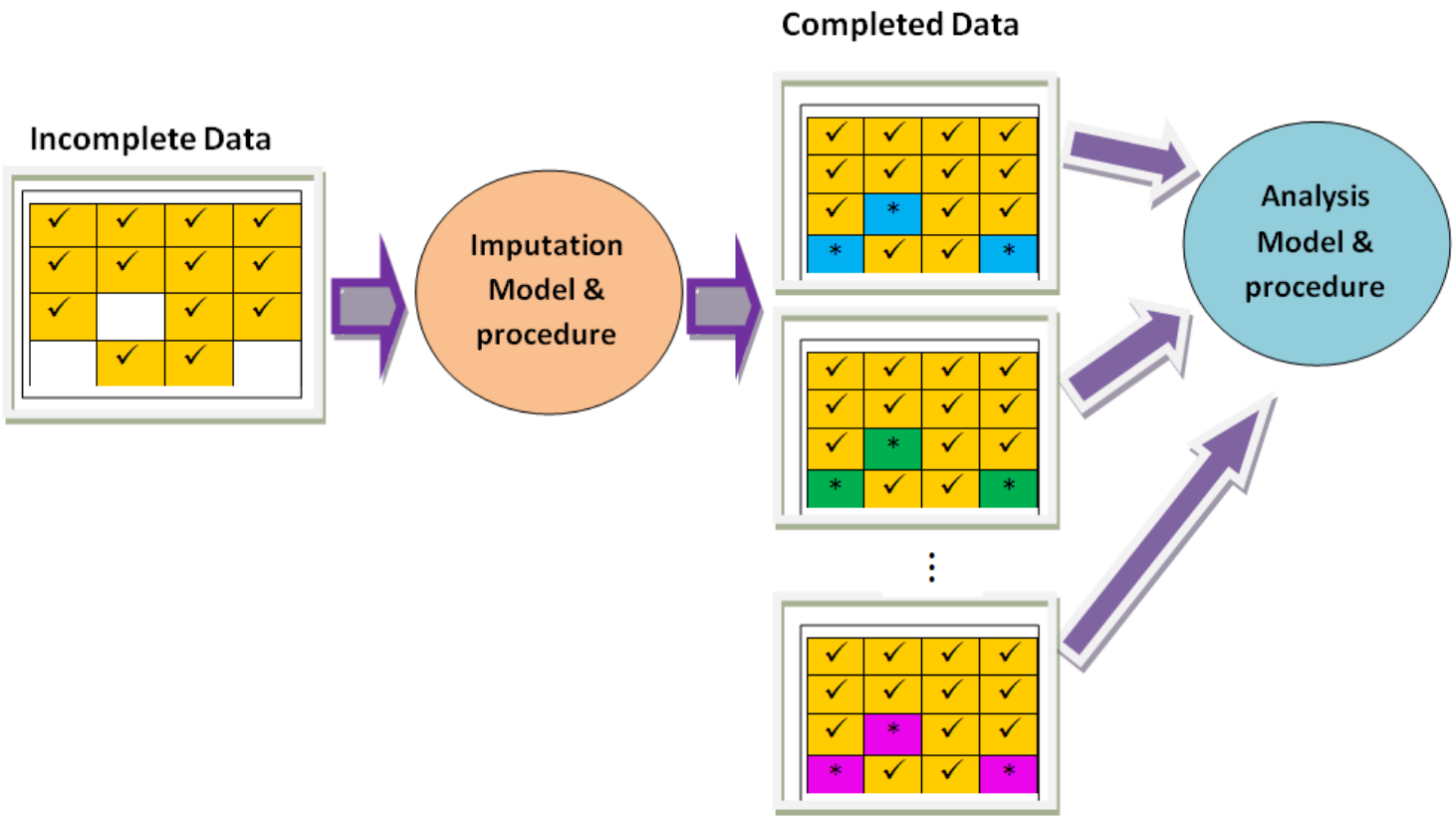
The set of variables in the imputation model should be larger and broader in scope than that required for ultimate analytic model applied to the completed data. This should assist in moving towards the MAR assumption and reduce any possible bias. (White et al, 2011)

Overview of Multiple Imputation Process



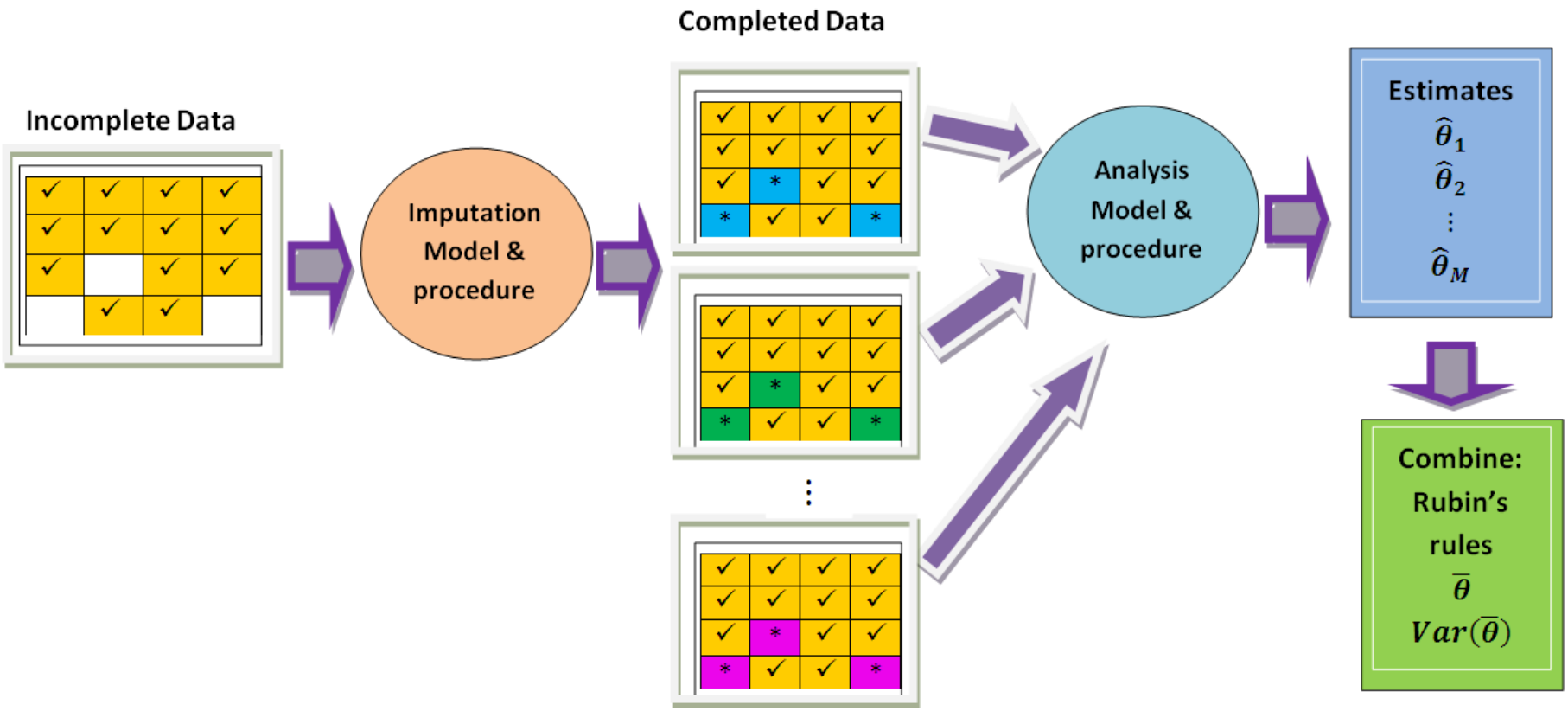
The set of variables in the imputation model should be larger and broader in scope than that required for ultimate analytic model applied to the completed data. This should assist in moving towards the MAR assumption and reduce any possible bias. (White et al, 2011)

Overview of Multiple Imputation Process



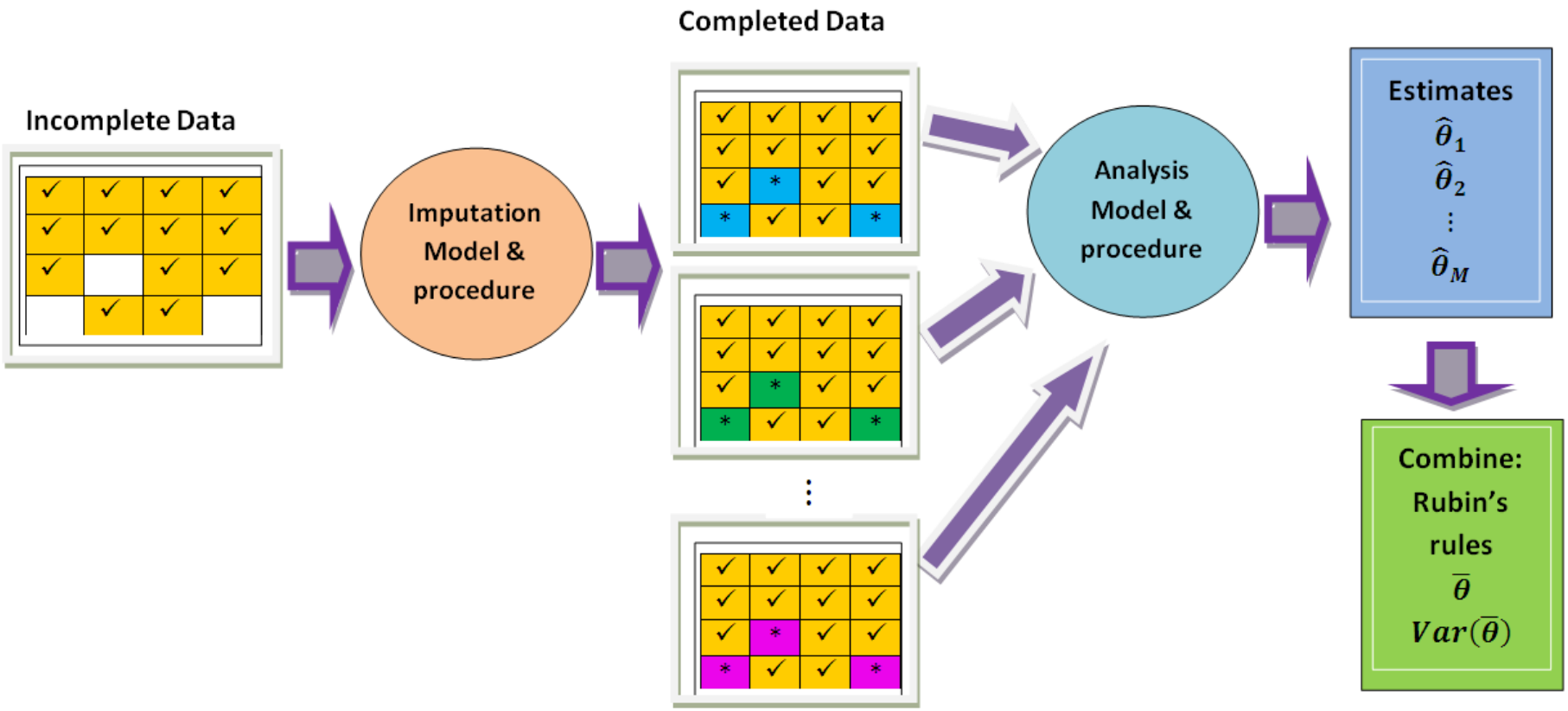
The set of variables in the imputation model should be larger and broader in scope than that required for ultimate analytic model applied to the completed data. This should assist in moving towards the MAR assumption and reduce any possible bias. (White et al, 2011)

Overview of Multiple Imputation Process



The set of variables in the imputation model should be larger and broader in scope than that required for ultimate analytic model applied to the completed data. This should assist in moving towards the MAR assumption and reduce any possible bias. (White et al, 2011)

Overview of Multiple Imputation Process



The set of variables in the imputation model should be larger and broader in scope than that required for ultimate analytic model applied to the completed data. This should assist in moving towards the MAR assumption and reduce any possible bias. (White et al, 2011)

From Imputed Data to Final Analysis

- Separate analysis carried out on each imputed data set $m = 1 \ldots M$.
- Estimates are combined using **Rubin's rules**:

$$\bar{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \qquad T = \bar{U} + \left(1 + \frac{1}{M}\right) B$$

where

$$\bar{U} = \frac{1}{M} \sum_{m=1}^{M} U_m = \frac{1}{M} \sum_{m=1}^{M} var(\hat{\theta}_m) \qquad \text{Within-imputation variance}$$

$$B = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \bar{\theta})^2 \qquad \text{Between-imputation variance}$$

and $\hat{\theta}_m$ is the estimate of θ from the mth completed data set

and U_m is the estimated variance of $\hat{\theta}_m$

This gives valid estimates of variance allowing for the fact that data has been imputed.

From Imputed Data to Final Analysis

- Separate analysis carried out on each imputed data set $m = 1 \ldots M$.
- Estimates are combined using **Rubin's rules**:

$$\bar{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \qquad T = \bar{U} + \left(1 + \frac{1}{M}\right) B$$

where

$$\bar{U} = \frac{1}{M} \sum_{m=1}^{M} U_m = \frac{1}{M} \sum_{m=1}^{M} var(\hat{\theta}_m) \qquad \text{Within-imputation variance}$$

$$B = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \bar{\theta})^2 \qquad \text{Between-imputation variance}$$

and $\hat{\theta}_m$ is the estimate of θ from the mth completed data set

and U_m is the estimated variance of $\hat{\theta}_m$

This gives **valid estimates of variance** allowing for the fact that data has been imputed.

Some History

In statistical literature, **imputation** means "filling in the data".

- 1957 - first mentioned in work of U.S. Census Bureau
- 1976 - **Multiple Imputation** - developed by Donald B. Rubin
  - as a solution to a practical problem in work missing income data on CPS
  - defined three missing data mechanisms (MCAR, MAR, NMAR)
  - the goal of MI is to provide statistically valid inferences
  - initially rejected as a crazy idea!
- 1983 - Dempster & Rubin stated *"Practical Implementation is still in the developmental state"* (p8)
- 1987 - **Rubin's rules** developed to combine data estimates
- 1996 - Rubin's paper: Multiple Imputation After 18+ Years - included a section on *Does a multiply-imputed data set take too much storage?"*

Some History

In statistical literature, **imputation** means "filling in the data".

- 1957 - first mentioned in work of U.S. Census Bureau
- 1976 - **Multiple Imputation** - developed by Donald B. Rubin
  - as a solution to a practical problem in work missing income data on CPS
  - defined three missing data mechanisms (MCAR, MAR, NMAR)
  - the goal of MI is to provide statistically valid inferences
  - initially rejected as a crazy idea!
- 1983 - Dempster & Rubin stated *"Practical Implementation is still in the developmental state"* (p8)
- 1987 - **Rubin's rules** developed to combine data estimates
- 1996 - Rubin's paper: Multiple Imputation After 18+ Years - included a section on *Does a multiply-imputed data set take too much storage?"*

More recently...

- 2000s
  - *"Multiple Imputation is now accepted as the best general method to deal with missing data in many fields"* (van Buuren, 2012)
  - Software has been developed - but took almost 3 decades after Dempster and Rubins's comment!
    - now being included in some but not all base statistical software
- 2010s
  - Most of the common statistical packages now have functions to carry out basic MI
  - More specialised functions are being developed eg multilevel
  - New function releases and existing function updates are currently appearing
- Useful resources: http://www.missingdata.org.uk/; www.multiple-imputation.com

MI Methods

Two widely used frameworks of model-based MI:

1. MI based on **Multivariate Normal Imputation (MVNI)**
   - assumes missingness is MAR
   - assumes all variables in the MI model jointly follow a MVN
     - MVN 1: Schafer & Yucel (2002)
     - MVN 2: Asparouhov & Muthén (2010)
   - implementation uses a Bayesian approach using the MCMC Gibbs Sampler algorithm

2. **Fully Conditional Specification (FCS) or MI with chained equations (MICE)**
   - also called Sequential Regression Imputation
   - random draws from conditional distribution specified for each variable with missing values conditional on all other variables in imputation model
   - does not restrict the conditional distributions to being normal
   - binary /ordinal/nominal variables: can use logistic regression (or ordered / multinomial)

MI Methods

Two widely used frameworks of model-based MI:

1. MI based on **Multivariate Normal Imputation (MVNI)**
   - assumes missingness is MAR
   - assumes all variables in the MI model jointly follow a MVN
     - MVN 1: Schafer & Yucel (2002)
     - MVN 2: Asparouhov & Muthén (2010)
   - implementation uses a Bayesian approach using the MCMC Gibbs Sampler algorithm

2. **Fully Conditional Specification (FCS) or MI with chained equations (MICE)**
   - also called Sequential Regression Imputation
   - random draws from conditional distribution specified for each variable with missing values conditional on all other variables in imputation model
   - does not restrict the conditional distributions to being normal
   - binary /ordinal/nominal variables: can use logistic regression (or ordered / multinomial)

MI Methods

Two widely used frameworks of model-based MI:

1. MI based on **Multivariate Normal Imputation (MVNI)**
   - assumes missingness is MAR
   - assumes all variables in the MI model jointly follow a MVN
     - MVN 1: Schafer & Yucel (2002)
     - MVN 2: Asparouhov & Muthén (2010)
   - implementation uses a Bayesian approach using the MCMC Gibbs Sampler algorithm

2. **Fully Conditional Specification (FCS) or MI with chained equations (MICE)**
   - also called Sequential Regression Imputation
   - random draws from conditional distribution specified for each variable with missing values conditional on all other variables in imputation model
   - does not restrict the conditional distributions to being normal
   - binary /ordinal/nominal variables: can use logistic regression (or ordered / multinomial)

Imputation Using Chained Equations (or FCS)

Fully Conditional Specification

- used when have large mixed set of continuous, nominal, ordinal variables
- variable by variable imputation is very flexible
  1. variable with smallest amount of missing data is regressed on variables in the imputation model with complete data, using an appropriate model (eg ordinal logistic)
     - imputations generated by stochastic draws from the predictive distribution defined by regression model
  2. variable with 2nd smallest amount of missing data is regressed on 1st imputed variable and all other variables in imputation model with complete data, and imputed values generated in same way
  3. algorithm continues until the variable with the most missing data is regressed on all other variables.

Imputation Using Chained Equations (or FCS)

Fully Conditional Specification

- used when have large mixed set of continuous, nominal, ordinal variables
- variable by variable imputation is very flexible
  1. variable with smallest amount of missing data is regressed on variables in the imputation model with complete data, using an appropriate model (eg ordinal logistic)
     - imputations generated by stochastic draws from the predictive distribution defined by regression model
  2. variable with 2nd smallest amount of missing data is regressed on 1st imputed variable and all other variables in imputation model with complete data, and imputed values generated in same way
  3. algorithm continues until the variable with the most missing data is regressed on all other variables.

Imputation Using Chained Equations (or FCS)

Fully Conditional Specification

- used when have large mixed set of continuous, nominal, ordinal variables
- variable by variable imputation is very flexible
  1. variable with smallest amount of missing data is regressed on variables in the imputation model with complete data, using an appropriate model (eg ordinal logistic)
     - imputations generated by stochastic draws from the predictive distribution defined by regression model
  2. variable with 2nd smallest amount of missing data is regressed on 1st imputed variable and all other variables in imputation model with complete data, and imputed values generated in same way
  3. algorithm continues until the variable with the most missing data is regressed on all other variables.

MVN vs FCS for Categorical missing data

Equivalence:

- Joint MVN imputation is equivalent to FCS with single level data and missingness only in continuous (normal) variables.

Categorical missing data

- Joint MVN imputation uses a **latent variable** formulation which can then be jointly modelled with the other continuous variables
- FCS imputation - GLM model chosen according to **variable type** conditional on other variables
  - Eg: logistic regression model for imputing missing data in binary variable

Latent Variables

Latent Variable Formulation

- Eg for a **Binary** variable Y_i: latent variable based on a cumulative probit model
  Define $Y_i^* \sim N(\beta, 1)$ such that $Y_i^* > 0 \Leftrightarrow Y_i = 1$
  $P(Y_i = 1) = P(Y_i^* > 0) = \Phi(\beta)$ where Φ is the CDF of standard Normal dist'n

$$Y_i = \begin{cases} 0 & \text{if } Y_i^* \leq \tau \\ 1 & \text{if } Y_i^* > \tau \end{cases}$$

- **Ordinal** with $K > 2$ categories: requires $K - 1$ threshold parameters based on a proportional probit model
- **Nominal** with K unordered categories: requires $K - 1$ latent difference scores based on a multinomial probit model

See ref: Carpenter and Kenward, (2013), Ch 4, 5

MI and Multilevel Data

Imputation and Substantive (MOI) models are said to be **congenial** if the Imputation model preserves associations in MOI.

Uncongenial Imputation models

1. **Richer**
   - MI model may contain more variables than MOI
     - this may be beneficial if suspect missingness to be MNAR
   - at least one congenial imputation model nested within it

2. **Poorer**
   - Imputation model contains fewer variables than MOI
   - Structure missing in imputation model that is in MOI
     -eg MOI is multilevel but MI model is single level
   - No congenial imputation model nested within it
   - Can lead to biases in the parameter estimates

See ref: Carpenter and Kenward, 2013

MVNI Model Specifications

Example for a random intercept MOI,

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_{ij} + u_{0j} + \varepsilon_{ij}$$

If x and y have missing values and z is completely observed, imputation models may be specified:

- **MVN 1 - Schafer & Yucel (2002)**- complete variables as predictors

$$\begin{matrix} x_{ij(mis)} \\ y_{ij(mis)} \end{matrix} \sim \text{MVN}\left(\begin{bmatrix} \beta_{0(x)} + \beta_{1(x)} z_{ij} + u_{0j(x)} \\ \beta_{0(y)} + \beta_{1(y)} z_{ij} + u_{0j(y)} \end{bmatrix}, \Sigma_\varepsilon \right)$$

- **MVN 2 - Asparouhov & Muthén (2010)** - treats all variables as outcomes

$$\begin{matrix} x_{ij(mis)} \\ y_{ij(mis)} \\ z_{ij} \end{matrix} \sim \text{MVN}\left(\begin{bmatrix} \beta_{0(x)} + u_{0j(x)} \\ \beta_{0(y)} + u_{0j(y)} \\ \beta_{0(z)} + u_{0j(z)} \end{bmatrix}, \Sigma_\varepsilon \right)$$

where Σ_ε is the level-1 residual covariance matrix

MVNI Model Specifications

Example for a random intercept MOI,

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_{ij} + u_{0j} + \varepsilon_{ij}$$

If x and y have missing values and z is completely observed, imputation models may be specified:

- **MVN 1 - Schafer & Yucel (2002)**- complete variables as predictors

$$\begin{matrix} x_{ij(mis)} \\ y_{ij(mis)} \end{matrix} \sim \text{MVN}\left(\begin{bmatrix} \beta_{0(x)} + \beta_{1(x)} z_{ij} + u_{0j(x)} \\ \beta_{0(y)} + \beta_{1(y)} z_{ij} + u_{0j(y)} \end{bmatrix}, \Sigma_\varepsilon \right)$$

- **MVN 2 - Asparouhov & Muthén (2010)** - treats all variables as outcomes

$$\begin{matrix} x_{ij(mis)} \\ y_{ij(mis)} \\ z_{ij} \end{matrix} \sim \text{MVN}\left(\begin{bmatrix} \beta_{0(x)} + u_{0j(x)} \\ \beta_{0(y)} + u_{0j(y)} \\ \beta_{0(z)} + u_{0j(z)} \end{bmatrix}, \Sigma_\varepsilon \right)$$

where Σ_ε is the level-1 residual covariance matrix

Multilevel MI

Limitations exist for each of the MI frameworks:

- **Joint model**
  - both MVNI appropriate for 2-level random intercept analysis
  - MVN1 (S-Y) may have random slopes for complete (auxiliary) variables
  - allows for incomplete categorical variables by way of latent normal variables
  - MVN2 (A-M) readily accommodates missing data at level-2
- **FCS (i.e. chained equations)**
  - joint distribution may not exist
  - single-level data MI works well
  - difficulties with convergence and coverage rates with simple multilevel MI
  - for imputing level-2 covariates - need to also include means $\bar{z}_j$ as well as z_{ij}

See references:

- Ch 9 of book *Multiple Imputation and its Application* by Carpenter and Kenward, 2013.
- A good review paper written by Enders, Mistler & Keller (2016)

Multilevel MI and Categorical variables

Common approaches to dealing with missing categorical data

- **Before specialised multilevel software available:**
  - impute separately within each cluster
  - fixed effects imputation
    - create fixed effects for cluster
    - preserves the between-cluster variation
    - i.e. use dummy codes for level-2 units in a single level MI
  - impute as if continuous and use a rounding scheme
  - include other level-2 covariates in the MI model as a proxy for the fixed effects for cluster
- **Utilising specialised software - MI model is also multilevel**
  - more parsimonious
  - preserves the between-cluster variation using random effects and level-2 covariance matrix
  - until very recently, software for imputing categorical variables in multilevel context has been limited

Summary of Multiple Imputation Software

Software	Imputed Variable Type				ML Structure
	Continuous	Binary	Ordinal	Nominal	2-level
Mplus	MVN1, MVN2, FCS	MVN1, MVN2, FCS	MVN1, MVN2, FCS	MVN1, MVN2, FCS	RI, RS(aux)
Stata:	mvn1, fcs	fcs	fcs	fcs	NA
MLwiN	MVN1				RI, RS(aux)
REALCOM*	MVN1	MVN1	MVN1	MVN1	RI, RS(aux)
SAS	mvn1, fcs	fcs	fcs	fcs	NA
`MMI_IMPUTE*`	MVN1				RI, RS(aux)
SPSS	fcs	fcs	fcs	fcs	NA
Stat-JR	MVN1	MVN1	MVN1	MVN1	RI, RS(aux)
R- MICE	FCS	fcs	fcs	fcs	RI, RS(aux)
R- PAN	MVN1				RI, RS(aux)
R- jomo / mitml	MVN1	MVN1	MVN1	MVN1	RI, RS(aux)
Blimp*	FCS	LFCS	LFCS	LFCS	RI, RS

Lowercase=Single level imputation only, RI=random Intercepts; RS=random slopes. Table prepared by Helani Kottage.

***Notes:**

- `REALCOM-IMPUTE`: is a macro for MLWin and Stata
- `MMI_IMPUTE`: is a macro for SAS multilevel MI
- `Blimp`: beta release - Jan2017 - imputes only
- 3-level: Mplus, STAT-JR

Example 1

- An educational data set was downloaded from the R package 'mlmRev' and modified according to the objectives of the current study (Bates, 2011)
- After all data manipulations, the data set comprised:
  - 1650 student records
  - within 22 schools
  - on 6 variables.

Ex 1: Description of data variables

Variable	Type	Description
school	Discrete	School ID
student	Discrete	Student ID within school
normexam	Continuous	Normalised Exam score of student at age 16
standLRT	Continuous	Student's standardised London reading test score at age 11
intake	Ordinal (3)	Student's verbal reasoning band at age 11, has 3 factors: bottom 25%, mid 50%, top 25%
sex	Binary	Gender of student: F / M

Ex 1: Substantive / Analysis Model of Interest (MOI)

A Random Intercept Model

- A linear mixed model with *normexam* as outcome variable
- Predictor variables are *standLRT*, *intake* and *sex*
  - *intake2* and *intake3* represents categories 'mid 50%' and 'top 25%' respectively
  - *sex2* represents 'male'.
- with random intercepts by **school**

$$
\begin{aligned}
normexam_{ij} &= \beta_0 + \beta_1 standLRT_{ij} + \beta_2 sex2_{ij} \\
&\quad \beta_3 intake2_{ij} + \beta_4 intake3_{ij} + u_j + e_{ij} \\
&\quad u_j \sim N(0, \sigma_u^2) \qquad e_{ij} \sim N(0, \sigma_e^2)
\end{aligned}
$$

This model was first fitted to the full data set.

Ex 1: Create Missing data

30% Missingness created under MAR in two variables:

- on **intake**: missingness is higher among males than females.

sex	Proportion Missing	No. missing
Male	0.6	297
Female	0.4	198
Total	1.0	495

- on **standLRT**: the lower the *normexam* the higher the missingness

normexam quartile	Proportion Missing	No. missing
min - Q1	0.35	173
Q1 - Q2	0.30	149
Q2 - Q3	0.20	99
Q3 - max	0.15	74
Total	1.00	495

Ex 1: Results - Complete Cases and Single level MI

	No imputation		Single level MI	
	Full Data	Complete	Ignore clusters	Sch -Fixed
	Stata	Stata	SPSS	Stata
Fixed Effects	Estimate	Estimate	Estimate	Estimate
	SE	*SE*	*SE*	*SE*
β_0 Intercept	0.315	0.469	0.395	0.411
	0.072	*0.081*	*0.080*	*0.080*
β_1 standLRT	0.400	0.381	0.359	0.351
	0.025	*0.036*	*0.031*	*0.035*
β_2 sex2	-0.160	-0.172	-0.164	-0.164
	0.036	*0.052*	*0.038*	*0.039*
β_3 intake2	-0.380	-0.493	-0.481	-0.490
	0.047	*0.068*	*0.057*	*0.057*
β_4 intake3	-0.727	-0.790	-0.815	-0.865
	0.082	*0.119*	*0.109*	*0.108*
Random Effects				
σ^2_u Sch Int.	0.075	0.061	0.077	0.081
	0.025	*0.024*	*0.027*	*0.027*
σ^2_e Residual	0.495	0.487	0.501	0.498
	0.017	*0.025*	*0.019*	*0.020*

Ex 1: Results cont. 2- level MI

	Full Data	2-level MI			
		MVN1	MVN2		LFCS
	Stata	Realcom	Realcom	Mplus	Blimp
Fixed Effects	Estimate	Estimate	Estimate	Estimate	Estimate
	SE	*SE*	*SE*	*SE*	*SE*
β_0 Intercept	0.315	0.427	0.415	0.398	0.578
	0.072	*0.076*	*0.076*	*0.081*	*0.094*
β_1 standLRT	0.400	0.370	0.372	0.366	0.356
	0.025	*0.028*	*0.031*	*0.030*	*0.031*
β_2 sex2	-0.160	-0.162	-0.162	-0.166	-0.166
	0.036	*0.038*	*0.036*	*0.039*	*0.037*
β_3 intake2	-0.380	-0.514	-0.496	-0.472	-0.495
	0.047	*0.053*	*0.055*	*0.060*	*0.056*
β_4 intake3	-0.727	-0.920	-0.903	-0.829	-0.867
	0.082	*0.091*	*0.098*	*0.104*	*0.106*
Random Effects					
σ^2_u Sch Int.	0.075	0.076	0.075	0.080	0.078
	0.025	*0.026*	*0.025*	*0.027*	*0.026*
σ^2_e Residual	0.495	0.481	0.484	0.496	0.495
	0.017	*0.018*	*0.019*	*0.019*	*0.020*

Summary for Example 1

Different approaches/software produce different estimates:

- **Complete cases**
  - bias in some estimates
  - SEs overestimated on fixed effects
  - student level variance, (σ_e^2), is overestimated
- **Single level imputation**
  - bias in some estimates
  - SEs on ordinal variable estimates overestimated
  - student level variance, (σ_e^2), is overestimated
- **2-level imputation**
  - fixed effects on ordinal variable overestimated (in absolute terms)
  - estimates and SEs of random effects approaches are similar to those for Full data

Example 2: Real Data Example

- **Fitting of statistical models which relate school performance**
  - measured by average of NAPLAN based scores (National Assessment Program for Literacy and Numeracy)
  - to overall parental background information for students
  - to be used for e.g. resource allocation
- **Data has hierarchical structure**
  - students within schools of varying sizes
- **Mixed variables types**
  - categorical data - both nominal and ordinal parent variables
  - continuous data - test score
- **Missing data pattern is complex**

Ex 2: Data structure & Decisions for MI

Parent characteristics

- Level of school education (4 categories - ordinal)
- Level of post-school education (4 categories - ordinal)
- Occupation group (5 categories - nominal)

1-Parent and 2-Parent families

- exploratory analysis showed different relationships between parent background variables if 1-P or 2-P
- allow for use of 2nd parent information in 2-Parent families
- Appreciable level of missing data items in parent variables

$\Rightarrow$ **separate into 2 data sets** and do MI separately for 1-P and 2-P families

Ex 2: Sample of 100 Schools

Individual Level Data

- 35959 students in 100 schools
- 30616 (85%) students in 2-Parent families (100 schools)
- 5343 (15%) students in 1-Parent families (99 schools)

Table: **Summary Statistics of School Size for Sample of 100 schools**

	min	max	median	mean	sd
School size	22	1179	315.5	359.6	277.2
No. test results in school	6	514	84.0	106.7	97.1

Ex 2: Building the Imputation Models

Impute missing data in

- Each of the parent variables (3 for 1-P or 6 for 2-P families)
- Student achievement Z-score (where applicable)

Other Auxiliary variables

- ATSI status (binary)
- Remoteness (Metropolitan, Provincial, Remote)
- Community census variables (SA1)

MI Model options

- Single-level imputation including Community census variables
- Two-level imputation using school as level-2 random intercept in MLM

Ex 2: Multiple Imputation Models and Methods

Single-level multiple imputation

- Include 10 Community census variables as predictors
- Method is FCS - Fully conditional specification
  - Software: Stata `mi impute chained`
  - involves different models for different variable types:
    - multinomial logistic regression (`mlogit`)
    - ordinal logistic regression (`ologit`) and
    - linear regression (`regress`)

Two-level multiple imputation

- using student as level-1 and school as level-2
- Method is MVNI- Multivariate Normal Imputation
  - involves latent normal variables for the categorical variables
  - Software: Stata with `REALCOM-Impute`

Ex 2: Model of interest (MOI)

We consider 2 options for the Model of Interest (MOI):

1-level (School) analysis:
Robust regression (RReg)

$$\bar{y}_g = \bar{x}_g^T \beta + \epsilon_g$$

- school level proportions $\bar{x}_g$ for each category of:
  - School Education
  - Non-School Educ.
  - Occupation

  for all parents in school g

2-level (MLM) analysis:
Random Intercept model

$$y_{ig} = x_{ig}^T \beta_I + \bar{x}_g^T \beta_C + u_g + \varepsilon_{ig}$$

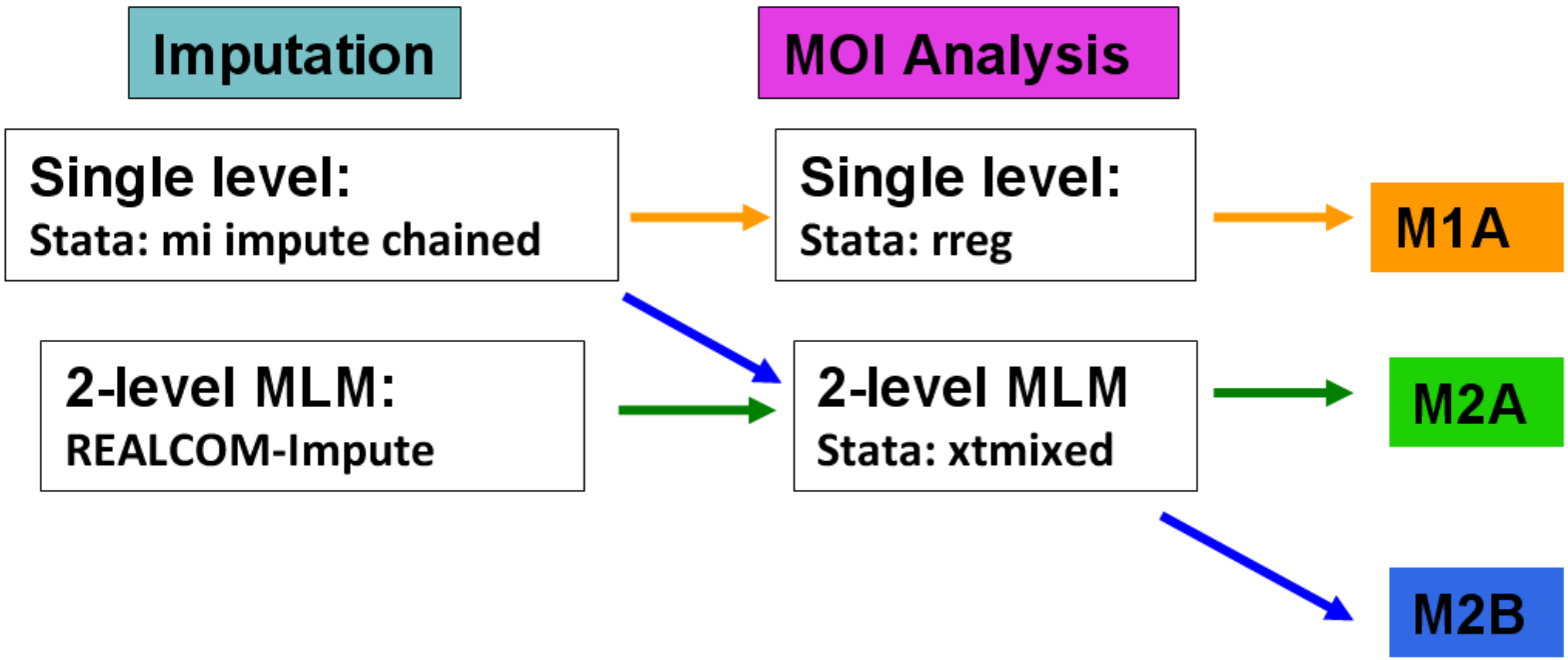
- student (i) in school (g)
- random intercept for school u_g
- fixed effects for parent background variables at individual level x_{ig}
- fixed effects for school level proportions $\bar{x}_g$

Quantity of interest is the set of fitted values at school level given the MOI

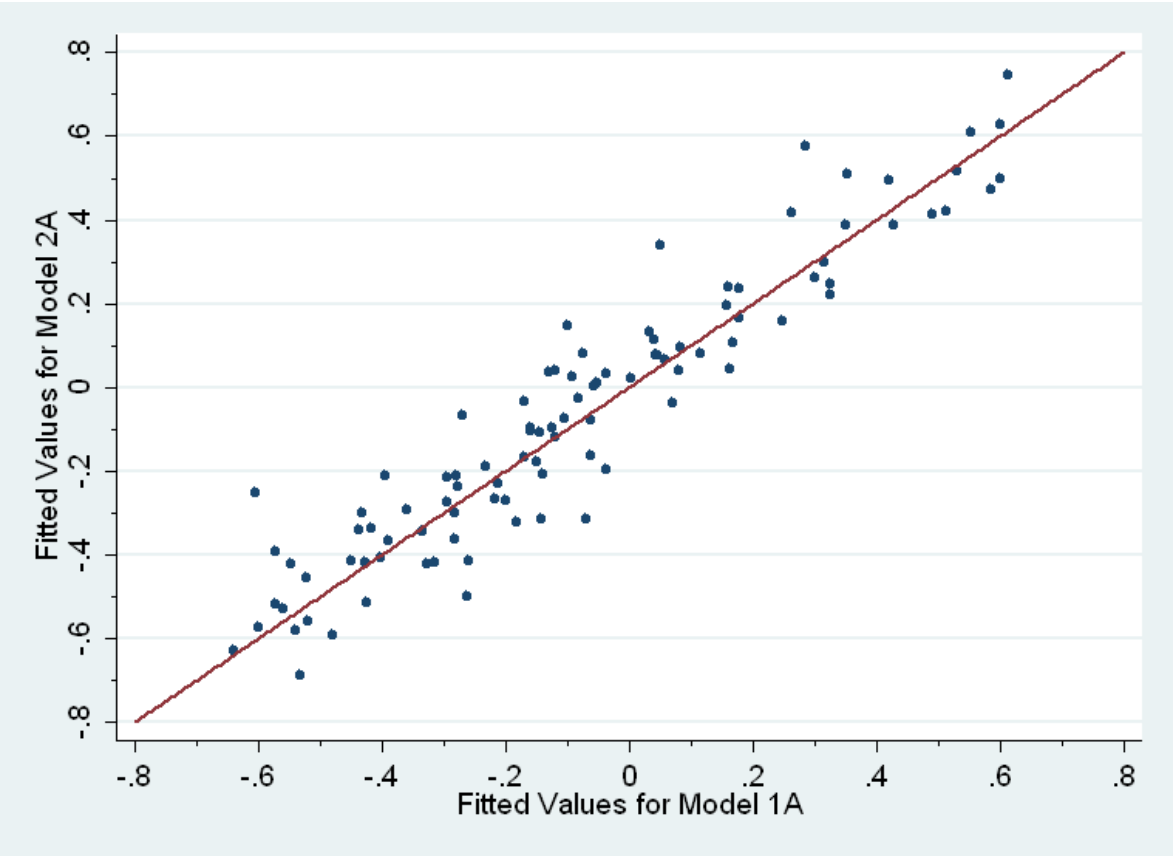
Ex 2: Imputation Models and MOI options

Consider **three possible combinations** of the Imputation model and the Model of Interest (MOI):

	MOI	
Imputation Model	Single	MLM
Single level	**M1A**	**M2B**
MLM	n/a	**M2A**

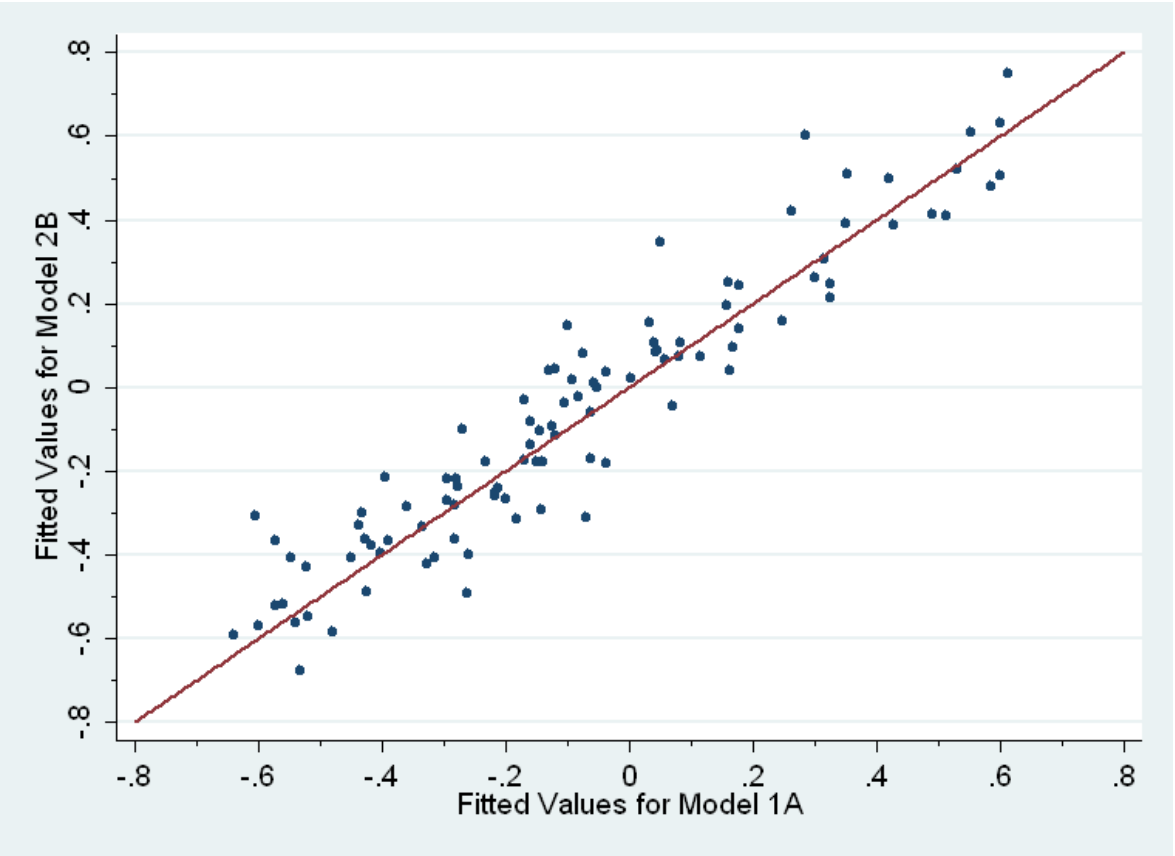


Ex 2 Results: Fitted values for M1A vs M2A



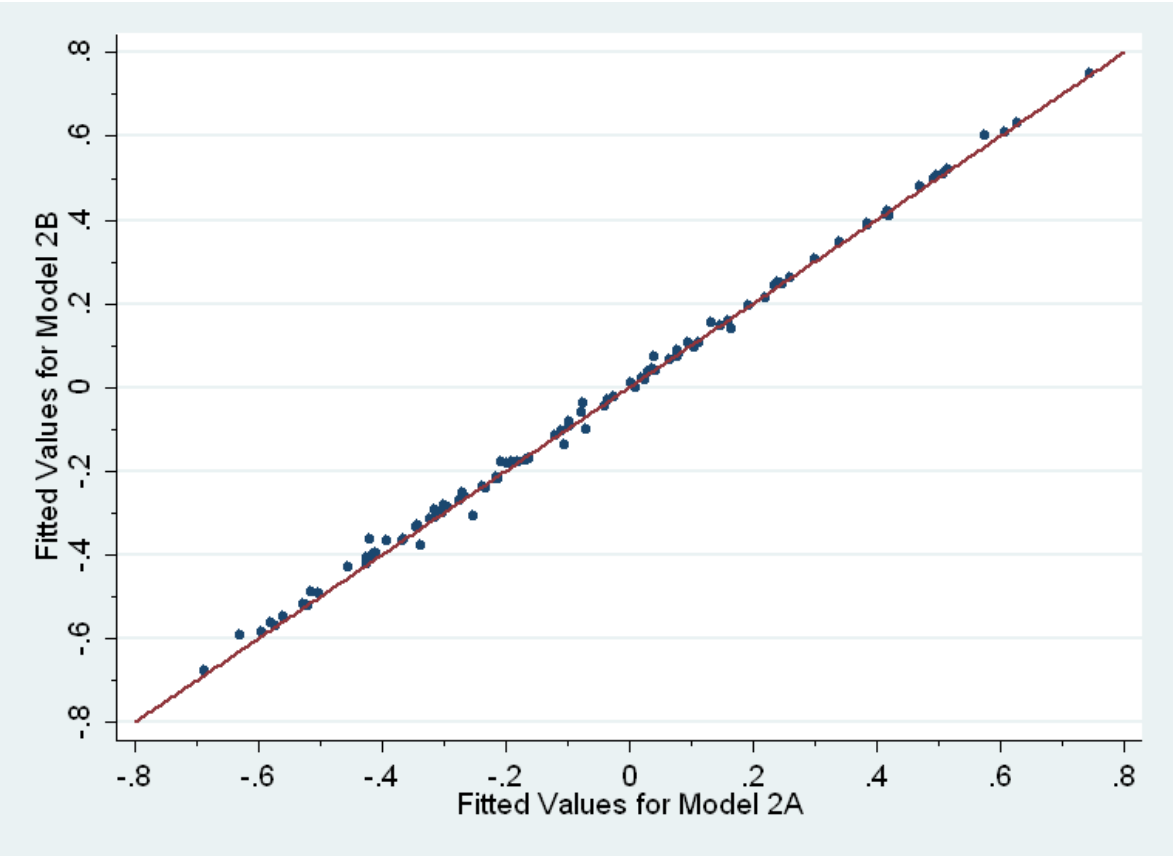
	Model of Interest (MOI) Analysis	
Imputation Model	Single	MLM
Single level	M1A	M2B
MLM	N/A	M2A

Ex 2 Results: Fitted values for M1A vs M2B



	Model of Interest (MOI) Analysis	
Imputation Model	Single	MLM
Single level	M1A	M2B
MLM	N/A	M2A

Ex 2 Results: Fitted values for M2A vs M2B



	Model of Interest (MOI) Analysis	
Imputation Model	Single	MLM
Single level	M1A	M2B
MLM	N/A	M2A

Summary for Example 2

In summary

- We are interested in the imputation of categorical variables at **individual level** in a hierarchical data structure
- We compare three combinations of **MI model** $\Longrightarrow$ **MOI** to obtain fitted values at **school level**
- The Single MI $\Longrightarrow$ Single level MOI (M1A) was a pragmatic method given time constraints for project
- The MLM MI $\Longrightarrow$ MLM MOI (M2A) is theoretically more sophisticated and appealing
- The Single MI $\Longrightarrow$ MLM MOI (M2B) may be useful alternative to running MLM MOI (current software running time is high)

Summary for Example 2 cont.

Results show that

- the MOI is important
- there is a slight bias in the Single MI $\Longrightarrow$ Single level MOI (M1A) when compared with the MLM MI $\Longrightarrow$ MLM MOI (M2A) and Single MI $\Longrightarrow$ MLM MOI (M2B)
- fitted values are similar for the Single MI $\Longrightarrow$ MLM MOI (M2B) and the MLM MI $\Longrightarrow$ MLM MOI (M2A)
- the use of the community variables with the Single MI is a good proxy for having a random intercept for school.

Emerging area of research

Handling of missing data in a multilevel structure is a growing area of research

- Multilevel MI more feasible - more modelling approaches
- Software is catching up - more available, very recent developments
- A growing number of missing data scenarios can be considered especially with mixed set of variables including categorical/ordinal

Further research

- Interest in topic has lead to new area of research for me
- UOW PhD student Helani Kottage is currently focussing on MI for categorical variables in multilevel structure
  - different approaches
  - optimal number of imputations
  - number of clusters & cluster sizes
  - use of covariates in MI model
  - diagnostics

Bibliography

- Asparouhov, T. & Muthén, B. (2010) *Multiple Imputation with Mplus* Available at http://www.statmodel.com/download/Imputations7.pdf .
- Bartlett, J. (2011) Using REALCOM Impute and Stata. Available at http://www.bristol.ac.uk/cmm/software/realcom/imputation.html .
- Bates, D. (2011) Examples from multilevel software comparative reviews. R Development Core Team.
- Carpenter, J. R., Goldstein, H., Kenward, M.G. (2011) REALCOM-IMPUTE Software for Mutlilevel Multiple Imputation with Mixed Response Types. *Journal of Statistical Software,* Vol. 45, Issue 5, p1-14.
- Carpenter, J. R., & Kenward, M.G. (2013) *Multiple Imputation and its Application.* Wiley: Statistics in Practice, Chichester, U.K.
- Dempster, A.P & Rubin, D.B., (1983) Introduction. In *Incomplete Data in Sample Surveys*, Vol. 2, p3-10, Academic Press, New York.
- Enders, C. & Mistler, S.A. & Keller, B.T. (2016) Multilevel Multiple Imputation: A Review and Evaluation of Joint Modeling and Chained Equations Imputation. *Psychological Methods, Vol 21, No. 2, p222-240.*
- Goldstein. H. & Carpenter, J. R., & Kenward, M.G., Levin, K.A. (2009) Multilevel models with multivariate mixed response types. *Statistical Modelling, Vol 9, No.3, p173197*
- Keller, B. T., & Enders, C. K. (2017). *Blimp Software Manual (Version Beta 6.7)*. Los Angeles, CA.
- Rubin, D.B., (1987) *Multiple Imputation for Nonresponse in Surveys*, Wiley: New York.
- Rubin, D.B., (2011) ABS Short Course *Imputation in Official Statistics*, held 13/11/2011.
- Schafer, J.L. & Yucel, R.M (2002) Computational strategies for multivariate linear mixed effects models with missing data. *Journal of Computational and Graphical Statistics, 11, p437-457*.
- van Buuren, S. (2012) *Flexible Imputation of Missing Data*, CRC Press, Boca Raton, U.S.A.
- Vermunt, J.K., van Ginkel, J.R., van derArk, L.A., Sijtsma, K. (2008) Multiple Imputation of Incomplete Categorical Data using Latent Class Analysis. *Sociological Methodology, Vol 38, p369-397.*
- White, I.R., Royston, P. & Wood, A.M. (2011) Multiple Imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine. Vol. 30, p377-399.*
- Yucel, R. M. (2011) State of the Multiple Imputation Software, *Journal of Statistical Software, Vol. 45, No.1 (Dec) p1-7.*

Finally . . .

So, all in all, there is

'MUCH ADO ABOUT NOTHING'

Thank you

Thank you: *Helani Kottage for her help with Ex 1 Results and Software Summary*

Acknowledgement: *The data for Ex 2 was provided by The NSW Department of Education and Communities and work was carried out in collaboration with David Steel.*