

# Multiple Imputation for Missing Data: Fully Conditional Specification compared with Multivariate Normal

Katherine Lee and John Carlin

Murdoch Childrens Research Institute &  
Department of Paediatrics, University of Melbourne

# Background

Two methods for multiple imputation now widely available in packages:

## **Multivariate normal imputation (MVNI)**

- Assumes a MVN distribution for all variables in the imputation model
  - Not always a realistic assumption e.g. binary / categorical variables

# Background

## Fully conditional specification (FCS)

- Specifies a (conditional) distribution for each variable with missing data in terms of all other variables
  - Flexible – e.g. can specify a logistic model for binary variables, ordinal logistic model for categorical variables
  - Can be cumbersome
  - Non-compatibility of conditionals

# Aims

- To compare multivariate normal imputation (MVNI) and fully conditional specification (FCS) as approaches to multiple imputation for dealing with missing data in a realistic simulation study
- Also compare the results from these multiple imputation (MI) methods to the standard complete case (CC) analysis

# Simulated dataset

- Synthetic population (Schafer, 2008)
  - 971,327 girls
  - based on data from the U.S. National Longitudinal Study of Adolescent Health
- Two waves of data:
  - **Wave I**: 14 covariates with primary focus on **DIET** (binary indicator for whether the participant reports dieting in the last 7 days)
  - **Wave II**: Primary outcome = **EMOTIONAL DISTRESS** (continuous 0-3)

# Simulated dataset

- Interested in the effect of diet on emotional distress at wave 2
- Potential confounders
  - emotional distress at wave 1 (continuous 0-3)
  - race (black / non-black Hispanic / other)
  - grade (ordinal 7-11)
  - self-rated overall health (ordinal 1-5)
  - self-rated physical fitness (ordinal 1-5)
- Datasets created from random draws of 1000 observations from synthetic population

# Analysis

- Linear regression model for emotional distress at wave 2:

$$l\text{dist}W2_i = \alpha + \beta_1 \text{diet}_i + \beta_2 l\text{dist}W1_i + \beta_3 \text{race}_{1i} \\ + \beta_4 \text{race}_{2i} + \beta_5 \text{grade}_i + \beta_6 \text{health}_i + \beta_7 \text{fitness}_i + e_i$$

Main focus is on the effect of diet,  $\beta_1$

“True values” estimated using full population

# Missing data mechanisms

**MDM 1:** Missing data in distress at wave 1

**MDM 2:** As for 1 PLUS overall health and physical fitness

**MDM 3:** As for 2 PLUS diet

# Missing data mechanisms

- For each variable, values set to be missing with probability determined by a logistic regression model:

$$\text{logit } Pr(\text{missing}) = a + b_1 \text{diet} + b_2 \text{race}_1 + b_3 \text{race}_2 \\ + b_4 \text{grade} + b_5 \text{ldistrW2}$$

where  $a=3$ ,  $b_1=1$ ,  $b_2=1$ ,  $b_3=1$ ,  $b_4=0.2$  and  $b_5=0.3$   
(Approximately 33% missing per variable)

# Analysis Methods

- For each simulated dataset & missing data mechanism estimate regression model using:
  - Complete Case analysis
  - MI using MVNI
    - NORM implemented in Stata
    - Log transformations of (skewed) categorical variables
    - Binary and categorical variable rounded into categories (standard or “adaptive” rounding)
  - MI using FCS
    - Iterative chained estimation (ICE) in Stata
    - Ordinal logistic regression for categorical and logistic regression for binary variables

# Analysis Methods

## MI methods

- Non-normality of distress data
  - Ignore and use raw values
  - Log transform
  - Log transform with an offset to give skewness of observed values = 0
$$u = \ln(\pm x - k) - \text{Inskew0 command}$$
  - Prediction matching (FCS only)
- 20 imputed datasets
  - Combined using Rubin's rules

# Evaluation of Methods

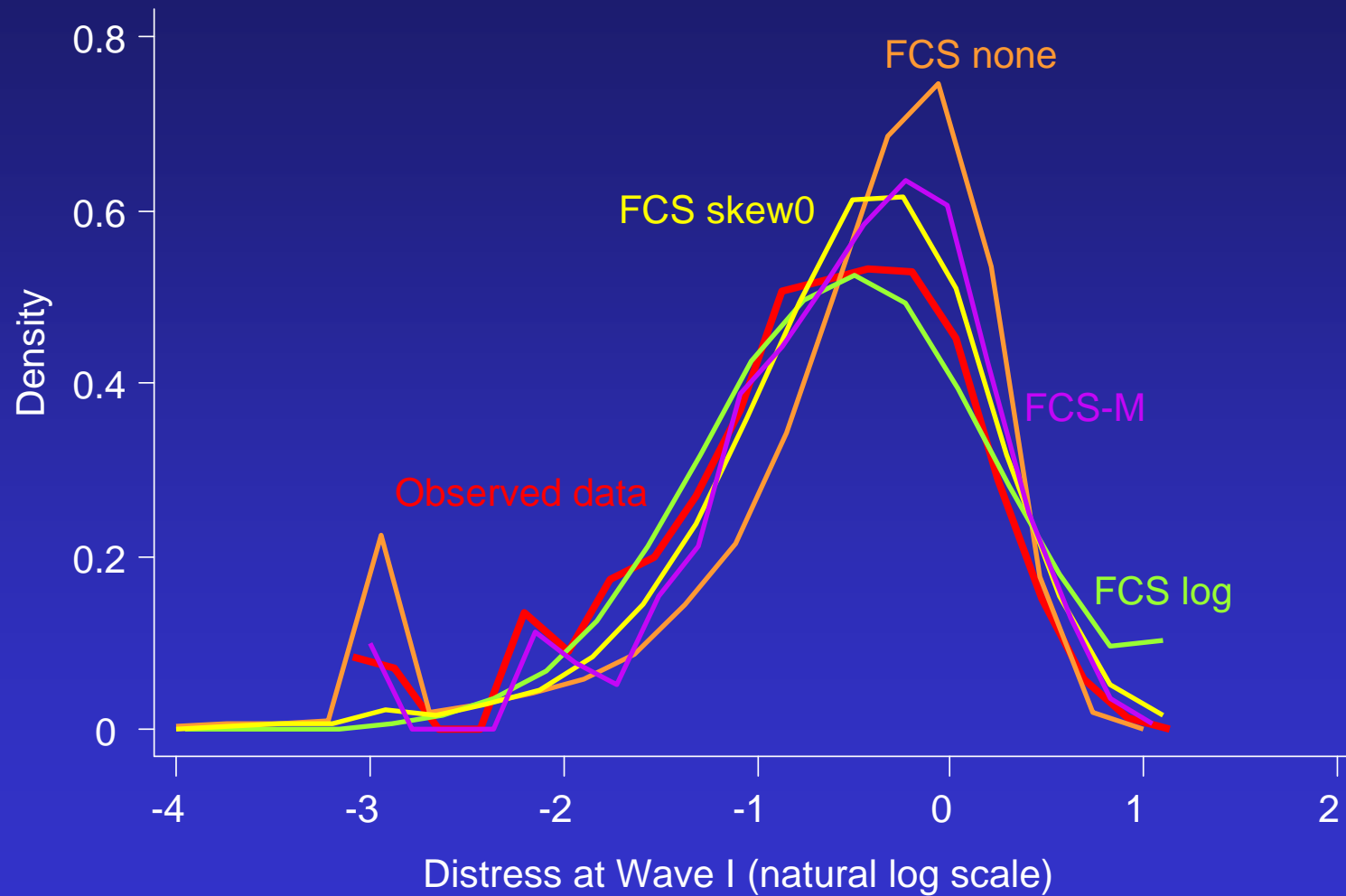
- Properties of regression coefficient estimates assessed using results from 1000 simulations
  - Bias of point estimate
  - Standard error (mean of estimates)
  - Coverage of estimated 95% CIs

# Results: MDM1 - FCS

(missing in distress W1 only)

		none	log	log-skew0
Diet ( $\beta_1 = -0.101$ )	Bias	0.012	0.007	0.004
	SE	0.069	0.069	0.068
	Coverage	0.955	0.958	0.948
Distress ( $\beta_2 = 0.554$ )	Bias	-0.086	-0.044	-0.012
	SE	0.044	0.039	0.042
	Coverage	0.484	0.769	0.942

# Imputed data for distress at Wave 1



# Results: MDM1

(missing in distress W1 only)

Diet ( $\beta_1 = -0.101$ )

	CC	FCS	FCS-M	MVNI
Bias	-0.039	0.004	0.003	0.003
SE	0.090	0.068	0.068	0.068
Coverage	0.923	0.948	0.955	0.946

# Results: MDM3

(missing distress, health, fitness and diet)

Diet ( $\beta_1 = -0.101$ )

	CC	FCS	FCS-M	MVNI	MVNI*
Bias	-0.143	-0.038	-0.032	-0.021	-0.019
SE	0.170	0.091	0.090	0.088	0.088
Coverage	0.838	0.907	0.919	0.927	0.959

\* Adaptive rounding for diet

# Conclusions

- No evidence that MVNI performed less well than the more flexible FCS approach
  - MVNI does well in the presence of both binary and categorical data
  - Under MDM3 MVNI resulted in less bias and better coverage than FCS
- Confirmed that CC analysis can lead to substantially biased results
  - However MI can introduce bias if imputations not carried out appropriately

# Conclusions

- Both FCS and MVNI were sensitive to non-normality in a highly predictive continuous covariate
  - log skew0 transformation dramatically improved the accuracy of estimation under both approaches
  - prediction matching also useful under FCS
- Adaptive rounding beneficial (MVNI)