Multiple imputation strategies for incomplete data in clustered longitudinal studies

Panteha Hayati Rezvan, Ph.D., phayatirezvan@mednet.ucla.edu
Thomas R. Belin, Ph.D., tblin@mednet.ucla.edu

Overview
- Objectives: assess the robustness of conclusions to proper handling of missing data
- Methods: contrast widely-used multiple imputation strategies
- Results: the findings from this research can provide methodological insights with enormous broad potential application to behavioural and biomedical research settings
- Future work: compare MI strategies with a recently developed multi-level imputation technique

Objectives

Missing data: General considerations
- Missing data are frequently encountered in HIV-related research
- Consequences of missing data
  - Can introduce bias
  - Loss of precision/power

Missing data: Modeling and Analysis Process
- Multiple imputation (MI) is a well-recognized technique for handling missing data
- Implementation of MI in standard statistical software typically assumes that data are ‘Missing at random’ (MAR)

Methods

Imputation strategies
- Multivariate Normal Imputation (MVNI)
- Fully Conditional-Specification (FCS)
  - Sequential regression models
  - Variable by variable
    - Linear regression for continuous
    - Logistic regression for binary

Recommendations for building imputation models
- Include all variables in the analysis model
- Include auxiliary variables
  - Predictors of incomplete variables
  - Predictors of missingness

Case study - Background
- Philani study examines a home-visiting prevention program delivered by neighbourhood mentor mothers
- Targeted mothers at risk for hazards alcohol use, HIV, TB and malnutrition as well as their children
- CRCT = 1238 pregnant women recruited
  - Standard Care (SC) - n=594
  - Philani Intervention Program (PIP) - n= 644

Case study - Research question
- Assess the effect of the PIP on children’s outcomes through the first five years of life.

Analysis model
- Linear mixed-effects model

Outcome
- Growth measure (waves 2-6)
  - Height-for-age z-score (HAZ)

Analysis variables
- HIV status, Neighborhood, Time-point, Interaction between time-point & neighborhood

Auxiliary variables
- Edinburgh Postnatal Depression Score, Any alcohol use, Married/lived with partner, Income above 2000 RAND, Food insecurity, Any violence, Maternal age*, Education*, Formal housing*

Number (%) of missing values

<table>
<thead>
<tr>
<th></th>
<th>Assessments</th>
<th>HIV status</th>
<th>HAZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pregnancy</td>
<td>156 (14.50)</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Birth</td>
<td>159 (14.78)</td>
<td>160 (14.87)</td>
<td></td>
</tr>
<tr>
<td>6 month</td>
<td>136 (12.64)</td>
<td>177 (16.45)</td>
<td></td>
</tr>
<tr>
<td>18 month</td>
<td>123 (11.43)</td>
<td>259 (24.07)</td>
<td></td>
</tr>
<tr>
<td>36 month</td>
<td>226 (21.00)</td>
<td>284 (26.39)</td>
<td></td>
</tr>
<tr>
<td>60 month</td>
<td>200 (18.59)</td>
<td>244 (22.68)</td>
<td></td>
</tr>
</tbody>
</table>

- 606/1076 (56.32) incomplete cases
- % of missing data in auxiliaries
  - Min (8.18), Max (20.82), p<0.3

Results

Imputation model failure
- Inclusive strategy did not converge initially
  - Collinearity among predictors

Strategies for handling lack of convergence
- Collapse HIV status
- Exclude “marital status” at wave 3
- Use augmentation when using MICE
- After modification, inclusion strategy converged
- Inclusive strategy - slightly higher point estimates/SE & wider CI

Future work
- Extension to covariates with missing data
- Comparison of the results with the Longitudinal Factor Imputation Method, a recently developed multi-level MI technique
- Linear mixed model for longitudinal associations
- Factor-analysis strategy for cross-sectional associations to keep the number of model parameters manageable

Acknowledgements
We would like to thank Dr. Mary Jane Rotheram for giving us the opportunity to use the Philani dataset. This work is supported by the T32 Postdoctoral Training Program for HIV Combination Prevention (T32MH109205).