ETHNICITY RECORDING IN PRIMARY CARE

Multiple imputation of missing data in ethnicity recording using The Health Improvement Network database

Tra Pham
PhD Supervisors: Dr Irene Petersen, Prof James Carpenter, Dr Tim Morris

1Department of Primary Care & Population Health - UCL
2Department of Medical Statistics - LSHTM
3Hub for Trials Methodology Research, MRC Clinical Trials Unit - UCL

Research funded by the Farr Institute of Health Informatics, London
OUTLINE
· **Background**
  - Ethnicity recording in primary care
  - Aims of this project

· **Study I - Impute missing ethnicity in THIN**
  - Sample
  - Weighted multiple imputation
  - Comparison of missing data methods

· **Study II - Simulation study**
  - Motivation
  - Methods
  - Results

· **Weighted multiple imputation by chained equations**
  - Motivation
  - Syntax

· **Summary**
BACKGROUND
ETHNICITY RECORDING IN THE HEALTH IMPROVEMENT NETWORK

- THIN is one of UK’s largest primary care databases
  - Contains records 12+ million patients in over 550 practices
  - Broadly representative of the UK population
  - Offers many opportunities for research
- Complete and reliable ethnicity recording is important for health research
- High level of missing ethnicity in primary care
AIMS OF THIS PROJECT

1. Explore the method of weighted multiple imputation to deal with missing ethnicity data
2. Introduce a new Stata command to implement weighted multiple imputation in multivariate missing data settings
STUDY I - IMPUTE MISSING ETHNICITY IN THIN
7 millions+ individuals in THIN between 2000-2013
- Ethnicity information is extracted using Read codes
- 5 categories of ethnicity: **white**, **mixed**, **Asian**, **black**, **other**
- 60% of individuals had missing ethnicity status
HOW SHOULD MISSING ETHNICITY DATA BE HANDLED?

- Studies often handle missing ethnicity data by
  - Excluding ethnicity in the analysis
  - Excluding individuals with missing ethnicity from the analysis
  - Imputing missing ethnicity as white
- A popular alternative is to use multiple imputation (MI) ...

Create new values to replace missing values in dataset

Analyse each new dataset separately
HOW SHOULD MISSING ETHNICITY DATA BE HANDLED?

... However, MI does not give plausible estimates compared with the census

→ Census data can be use to weight MI such that the correct ethnicity distribution is recovered
• Deals with bias from non-response and under-represented groups in analysis of survey data

\[
\text{Weight}_{ps} = \frac{1}{(p_{\text{sample}}/p_{\text{census}})}
\]

• Hypothetical example:

<table>
<thead>
<tr>
<th>Ethnic Group</th>
<th>Census Proportion</th>
<th>Sample Proportion</th>
<th>Weight (_{ps})</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>0.6</td>
<td>0.8</td>
<td>0.75</td>
</tr>
<tr>
<td>Asian</td>
<td>0.2</td>
<td>0.1</td>
<td>2</td>
</tr>
<tr>
<td>Black</td>
<td>0.15</td>
<td>0.05</td>
<td>3</td>
</tr>
<tr>
<td>Other</td>
<td>0.05</td>
<td>0.05</td>
<td>1</td>
</tr>
</tbody>
</table>

• White ethnic group: \(p_{\text{weighted}} = 0.8 \times 0.75 = 0.6 = p_{\text{census}}\)
**Table 1:** Ethnic proportions in one imputed dataset using PS weights

<table>
<thead>
<tr>
<th>Ethnicity (%)</th>
<th>Observed</th>
<th>Imputed</th>
<th>Combined*</th>
<th>Census</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>67.1</td>
<td>63.2</td>
<td>65.7</td>
<td>59.8</td>
</tr>
<tr>
<td>Mixed</td>
<td>2.9</td>
<td>4.3</td>
<td>3.4</td>
<td>5.0</td>
</tr>
<tr>
<td>Asian</td>
<td>11.7</td>
<td>16.6</td>
<td>13.5</td>
<td>18.5</td>
</tr>
<tr>
<td>Black</td>
<td>11.7</td>
<td>13.2</td>
<td>12.2</td>
<td>13.3</td>
</tr>
<tr>
<td>Other</td>
<td>6.6</td>
<td>2.7</td>
<td>5.2</td>
<td>3.4</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

* Combined data = observed data + imputed data
Proportions required in the imputed data to get to the census proportions after imputation

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Observed</th>
<th>Imputed</th>
<th>Combined*</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>$p_{obs} \times N_{obs}$</td>
<td>$p_{req} \times N_{mis}$</td>
<td>$p_{census} \times N_{total}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Total</td>
<td>$N_{obs}$</td>
<td>$N_{mis}$</td>
<td>$N_{total}$</td>
</tr>
</tbody>
</table>

* Combined data = observed data + imputed data

\[
p_{req} = \frac{p_{census} \times N_{total} - p_{obs} \times N_{obs}}{N_{mis}}
\]

\[\rightarrow Weight_{MI} = 1/(p_{obs}/p_{req})\]
Table 2: Ethnic proportions in one imputed dataset using adjusted weights

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Observed</th>
<th>Imputed</th>
<th>Combined*</th>
<th>Census</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>67.1</td>
<td>50.6</td>
<td>61.0</td>
<td>59.8</td>
</tr>
<tr>
<td>Mixed</td>
<td>2.9</td>
<td>7.4</td>
<td>4.6</td>
<td>5.0</td>
</tr>
<tr>
<td>Asian</td>
<td>11.7</td>
<td>26.6</td>
<td>17.3</td>
<td>18.5</td>
</tr>
<tr>
<td>Black</td>
<td>11.7</td>
<td>15.4</td>
<td>13.0</td>
<td>13.3</td>
</tr>
<tr>
<td>Other</td>
<td>6.6</td>
<td>0</td>
<td>4.1</td>
<td>3.4</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

* Combined data = observed data + imputed data
Figure 1: Ethnic proportions in one imputed dataset using adjusted weights
Predictors for imputation: year of birth, sex, deprivation score, indicator if first registration is between 2006-2013

**mi impute mlogit** command allows for probability weights specification:

```
mi impute mlogit varlist [pweights] [, mi_options ...]
```
Figure 2: Ethnic proportions under different missing data methods

1. Complete case analysis
2. Impute missing as white
3. Regular MI
4. Weighted MI

Census
STUDY II - SIMULATION STUDY
• So far the weighted MI gives plausible estimates of ethnicity proportions
• However, studies often focus on the association between ethnicity and outcome of interest
• How to deal with missing data when ethnicity is an exposure/covariate?
• Aim: evaluate the performance of weighted MI using a simulation study
DESIGN OF SIMULATION STUDY

DATA GENERATING MODEL → FULL DATA → MNAR MODEL FOR ETHNICITY → INCOMPLETE DATA → COMPLETE CASE → IMPUTE AS WHITE → REGULAR MI → WEIGHTED MI

MODEL OF INTEREST → COMPARE OUTCOMES
Motivated by the association between ethnicity and myocardial infarction

Statistical model: exponential survival model for the association between ethnicity and time to first diagnosis of myocardial infarction, adjusted for age and sex
1. Data generating mechanism:
   - Data are simulated that closely model a sample derived from THIN
   - Ethnicity is generated as a 3-category variable: white, Asian, black, using proportions from the census

2. MNAR mechanism:
   - Ethnicity is made MNAR depending on follow-up time, sex and ethnicity
· Define $\tilde{\beta} = \frac{1}{K} \sum_k \hat{\beta}_k \rightarrow$ Estimated bias $= \tilde{\beta} - \beta$
• Empirical standard error = \( \sqrt{\frac{1}{K-1} \sum_k (\hat{\beta}_k - \bar{\beta})^2} \)
\[ \text{Coverage} = \frac{1}{K} \sum_k 1( | \hat{\beta}_k - \beta | < z_{\alpha/2} s_k ) \]
1. Complete case analysis leads to biased results and loss in efficiency
2. Impute missing data as white biases the association between non-white groups and outcome
3. Regular MI does not correct for bias when ethnicity is MNAR
4. Weighted MI gives closest estimates to true values and substantial coverages
WEIGHTED MI BY CHAINED EQUATIONS
· Weighted MI only implemented in univariate missing data of ethnicity so far

· Multiple imputation of chained equations (MICE) for multivariate missing data problems
  - To get one set of imputed values, iterate over $t = 0, 1, \ldots, T$ and impute:

\[
\begin{align*}
Y_{1}^{t+1} & \text{ using } Y_{2}^{t}, Y_{3}^{t}, \ldots, Y_{k}^{t} \\
Y_{2}^{t+1} & \text{ using } Y_{1}^{t+1}, Y_{3}^{t}, \ldots, Y_{k}^{t} \\
\vdots & \\
Y_{k}^{t+1} & \text{ using } Y_{1}^{t+1}, Y_{2}^{t+1}, \ldots, Y_{k-1}^{t+1}
\end{align*}
\]

· `mi impute chained` only allows specification of global weights, which are applied to all equations
· Weighted imputation by chained equations - \textit{wice}

· Allows weights specification for one conditional models while no weights for others:

\[
\begin{align*}
Y_{t+1} & \text{ using } Y_t, Y_1^{t+1}, \ldots, Y_k^{t+1} \quad \text{\( pw_t \)} \\
Y_{t+1} & \text{ using } Y_{t+1}^{t+1}, Y^t_2, \ldots, Y^t_k \\
    & \quad : \\
Y_{t+1} & \text{ using } Y_{t+1}^{t+1}, Y_{t+1}^{t+1}, \ldots, Y_{k-1}^{t+1}
\end{align*}
\]
· Stata syntax:

\[ \text{wice (method}_1 \ \text{impvar}_1 [pweight}_1])\ldots (\text{method}_K \ \text{impvar}_K [pweight}_K]): \ \text{indvars [, options]} \]

· Current features:
  1. Univariate imputation methods: \texttt{regress}, \texttt{logit}, \texttt{ologit}, \texttt{mlogit}
  2. Global options: \texttt{add}, \texttt{cycles}, \texttt{seed}, \texttt{noisily}, \texttt{dryrun}
CONCLUSIONS

- Ad-hoc methods can lead to biased results and misleading conclusions
- Regular MI fails to produce valid inferences when data are MNAR
- Weighted MI is a simple solution and performs better than regular MI
- wice offers flexibility for weighted MI by chained equations