Consistent Variance Estimates for Multiple Imputation in R

James Reilly

University of Auckland

8 July 2009
1. Multiple imputation
2. MI bias and alternative approach
3. mitee R package
4. Summary and roadmap
Missing data is a common problem
- Many statistical methods require complete data
- Imputation methods fill in missing values
  - Standard methods can then be used on the imputed dataset
  - However this ignores uncertainty due to missing data
- Multiple imputation attempts to solve this problem
Multiple imputation

- Impute multiple times for each missing value
  - Should reflect uncertainty in imputation process (proper imputation)
  - Originally proposed for public-use datasets (Rubin, 1987)
    - Imputer and analyst are two different people
- Works when imputer and analyst share the same well-specified model
- Also a good approximation when close to this ideal
Multiple imputation issues

- Traditional MI can produce biased variance estimates for conflicting or misspecified models
  - E.g. if analyst allows for sample design, but imputer does not
- Concerns expressed by Fay (1991, 1996), Kim et al. (2006) and others
  - “MI is not generally recommended for public use data files.” — Kim et al. (2006)
Robins and Wang (2000) - MI using estimating equations
- Robust to model misspecification and disagreement
- Promising for public-use datasets
  - Especially mass imputation applications, e.g., statistical matching

- Estimating equations for imputer $\sum S_{obs}(\psi) = 0$ and analyst $\sum U(\beta) = 0$
- Impute from the fitted joint distribution, conditional on the observed data for that observation
- Asymptotic MI variance is $\Sigma = \tau^{-1}\Omega(\tau')^{-1}$, where ...
Estimating equations approach (continued)

\[ \hat{\beta} = -E \left\{ \frac{\partial \tilde{U}(\psi^*, \beta)}{\partial \beta'} \right\}_{\beta=\beta^*}, \quad \Omega = \Omega_1 + \Omega_2 + \Omega_3, \]
\[ \Omega_1 = E \left\{ \tilde{U}(\psi^*, \beta^*) \otimes^2 \right\}, \quad \Omega_2 = \kappa \Lambda \kappa', \]
\[ \Omega_3 = E \left\{ \kappa D(\psi^*) \tilde{U}(\psi^*, \beta^*)' + \{ D(\psi^*) \tilde{U}(\psi^*, \beta^*)' \}' \right\}, \]
\[ \kappa = E \left\{ U(\psi^*, \beta^*) S_{mis}(\psi^*)' \right\}, \quad \Lambda = E \left\{ D(\psi^*) \otimes^2 \right\}, \]
\[ S_{mis}(\psi^*) = \frac{\partial \log f(Y|Y_R, R; \psi)}{\partial \psi} \bigg|_{\psi=\psi^*}, \quad D(\psi^*) = I_{obs}^{-1} S_{obs}(\psi^*). \]
R package for Multiple Imputation Through Estimating Equations (mitee)

Implements Robins and Wang approach to MI
- Imputation using linear and logistic regression models
  - `eeimpute(formula, data, family='gaussian')`
  - Returns a multiply imputed dataset (a list of imputed data frames, including information about the imputation model)
- Analysis - linear model (and thus means, percentages) and logistic regression
  - `eeglm(formula, midata, family='gaussian')`
mtee example

```r
> head(nrs4)
     wine sex age work
 1     1   2   4   2
 2    NA   2   2   1
 3    NA   1   3   2
 4    NA   2   3   2
 5     1   2   4   1
 6     1   2   2   1

> nrs4mi <- eeimpute(wine ~ sex + age, nrs4,
                     family='''binomial''')
> eeglm(wine ~ work, nrs4mi, family='''binomial''')

$param
 [1]  1.1953369  -0.2597735

$vcov
 [,1] [,2]
[1,] 0.05362612 -0.03675407
[2,] -0.03675407  0.01621821

> # Traditional MI variances:  0.0677 and 0.0253.
> # Naive single imputation variances: 0.0378 and 0.0144.
```
Summary

- Traditional multiple imputation is useful, but fails in some circumstances
- Alternative estimating equations approach implemented in R
- Future work
  - Implement more imputation and analysis models
    - E.g. multivariate normal imputation
  - Integrate with King et al.’s Zelig system
  - Handle complex survey data
  - Imputation through chained equations
1. Multiple imputation

2. MI bias and alternative approach

3. mitee R package

4. Summary and roadmap