Consistent Variance Estimates for Multiple Imputation in R

James Reilly

University of Auckland

8 July 2009

Imputation

- 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)

Estimating equations approach to MI

- 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)

$$\hat{\tau} = -\mathsf{E}\left\{\frac{\partial \bar{U}(\psi^\star,\beta)}{\partial \beta'}\right\}_{\beta=\beta^\star},\ \Omega = \Omega_1 + \Omega_2 + \Omega_3,$$
$$\Omega_1 = \mathsf{E}\left\{\bar{U}(\psi^\star,\beta^\star)^{\otimes 2}\right\},\ \Omega_2 = \kappa\Lambda\kappa',$$
$$\Omega_3 = \mathsf{E}\left\{\kappa D(\psi^\star)\bar{U}(\psi^\star,\beta^*)' + \{D(\psi^\star)\bar{U}(\psi^\star,\beta^\star)'\}'\right\},$$
$$\kappa = \mathsf{E}\left\{U(\psi^\star,\beta^\star)S_{mis}(\psi^\star)'\right\},\ \Lambda = \mathsf{E}\left\{D(\psi^\star)^{\otimes 2}\right\},$$
$$S_{mis}(\psi^\star) = \frac{\partial \log f(Y|Y_R,R;\psi)}{\partial\psi}\Big|_{\psi=\psi^\star},\ D(\psi^\star) = I_{obs}^{-1}S_{obs}(\psi^\star).$$

mitee - R package

- 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'')`

mitee example

```
> 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 s̃ex + age, nrs4,
family=''binomial'')
> eeglm(wine w̃ork, 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