Multiple Imputation and Complex Survey Data

James Reilly
Dept of Statistics
University of Auckland

Workshop on Missing Data
Toronto, August 2004

Overview

- Review multiple imputation (MI) and complex surveys
- Multiple imputation issues
  - Standard combining rules can cause problems
  - Estimating equation approach – pros and cons
- Generalization to complex surveys
- Comparison of results
- Applications

Multiple Imputation

- Multiple imputation (or repeated imputation)
  - Widely used method for dealing with missing data
    - Impute for each missing value several times, based on a statistical model
    - Combine resulting imputed datasets to estimate variances
      - Naïve treatment of single imputation underestimates variances
  - Developed by Rubin (1987)
    - Standard combining rule gives variance as

$$T_D = \overline{W}_D + \frac{D}{D+1} B_D$$

    - Requires just the analysis results for each imputed datasets (parameter estimates $\hat{\theta}_d$ and naïve variances W_d)

Multiple Imputation Issues

- Standard combining rules give biased variance estimates in some situations
  - See Fay (1991, 1996), Robins and Wang (2000), Nielsen (2003)
- Problems can arise when the imputer's or analyst's models are misspecified
  - I.e. when they differ from each other, or from reality
- Variances are commonly overestimated, but can also be underestimated

Survey Analysis in Practice

- Relationships between variables are often analysed crudely
  - Two-way cross-tabulations are ubiquitous
    - E.g. wine consumption by gender and other demographics
- Underlying statistical models are simple
  - Parameter estimates may be biased when reality is more complex
    - E.g. If wine consumption truly explained by age and gender, a model based only on gender may give biased estimates of the true gender effect
  - But unbiased variance estimates are still desired

Simulation Results – i.i.d. Data

- Survey example (from Reilly 2003)
  - i.i.d. data generated from survey's joint distribution of wine consumption (43% missing), gender, age and working status
  - Logistic regression models for wine consumption
    - Imputer used gender and age, analyst used gender only
    - Imputed values generated using parameter values drawn from the asymptotic distribution of the MLE (Little & Rubin 2002, p216)
    - Variance of gender parameter (m=5, n=1000):

Simulation Variance	Av. Variance Estimate	Relative RMSE
0.0377	0.0282	32%

- Schafer (1999), Barnard & Meng (1999) and SAS Institute (2004) advise the imputer to use all analysis variables
  - Above example shows this can still give poor results
  - A more nuanced discussion of the need for correctly specified models is given by Little and Rubin (2002)

Estimating Equation Approach

- Robins and Wang (2000) developed a new MI approach based on estimating equations
  - Gives asymptotically unbiased variance estimates
    - Even when imputer's and analyst's models are misspecified
    - But can be more variable than standard MI estimator
- Example from Robins and Wang (2000)
    - Regression through origin with heteroscedastic errors
    - Variance of slope of regression line (for m=5, n=100):

Method	Simulation Variance	Average Variance Estimate	Relative RMSE
Robins & Wang	0.0197	0.0188	36%
Standard MI (Rubin)	0.0197	0.0137	32%

EE Method and Sample Surveys

- Original estimating equation method has two disadvantages for routine survey use
  - Variance estimates cannot be calculated using just the imputed data
    - Requires information from the imputer's model as well
  - Assumes i.i.d. data
- In practice, surveys often require complex sample designs and estimators
  - E.g. cluster samples, weighting, stratification
- Methods for i.i.d. data will then typically underestimate sampling variance

Generalization to Complex Surveys

- Have adjusted Robins and Wang's formulae to account for complex sample designs, including
  - Cluster samples
  - Inverse probability weights
  - Stratification
  - Finite population correction
- Formulae on following slide ignores stratification and finite population correction for simplicity

Formulae for Complex Surveys

$$\hat{\Sigma}=(\hat{\tau})^{-1}\hat{\Omega}(\hat{\tau}')^{-1},\ \hat{\tau}=-w_{\cdot\cdot}^{-1}\sum_{k=1}^{c}\sum_{i=1}^{c_k}w_{ik}\frac{\partial\bar{U}^{ik}(\hat{\beta})}{\partial\beta'},\ \hat{\Omega}=\hat{\Omega}_1+\hat{\Omega}_2+\hat{\Omega}_3,$$

$$\hat{\Omega}_1=\frac{n}{(c-1)w_{\cdot\cdot}}\sum_{k=1}^{c}w_{\cdot k}\left(\sum_{i=1}^{c_k}\bar{U}^{ik}\right)^{\otimes 2},\ \hat{\Omega}_2=\hat{\kappa}\hat{\Lambda}\hat{\kappa}',\ \hat{\Omega}_3=\frac{n}{(c-1)w_{\cdot\cdot}}\sum_{k=1}^{c}w_{\cdot k}\left(\hat{\kappa}\bar{D}^{k}\bar{\bar{U}}^{k\prime}+\left\{\hat{\kappa}\bar{D}^{k}\bar{\bar{U}}^{k\prime}\right\}'\right),$$

$$\hat{\kappa}=\frac{1}{mw_{\cdot\cdot}}\sum_{j=1}^{m}\sum_{k=1}^{c}\sum_{i=1}^{c_k}w_{ik}U^{ikj}(\hat{\psi},\hat{\beta})[S_{mis}^{ikj}(\hat{\psi})]',\ \hat{\Lambda}=\frac{n}{(c-1)w_{\cdot\cdot}}\sum_{k=1}^{c}w_{\cdot k}\bar{D}^{k\otimes 2},\ \bar{D}^{k}=\frac{1}{w_{\cdot k}}\sum_{i=1}^{c_k}w_{ik}\hat{D}^{ik},$$

$$\bar{\bar{U}}^{k}=\frac{1}{w_{\cdot k}}\sum_{i=1}^{c_k}w_{ik}\bar{U}^{ik},\ \hat{D}^{ik}=-\left[\frac{1}{w_{\cdot\cdot}}\sum_{k=1}^{c}\sum_{i=1}^{c_k}w_{ik}\frac{\partial S_{obs}^{ik}(\hat{\psi})}{\partial\psi'}\right]^{-1}S_{obs}^{ik}(\hat{\psi}),$$

$$S_{mis}^{ikj}=\frac{\partial\log f\left(Y^{ikj}(\hat{\psi})\middle|Y_R^{ik},R^{ik};\psi\right)}{\partial\psi}\Big|_{\psi=\hat{\psi}},\ \bar{U}^{ik}=\bar{U}^{ik}(\hat{\psi},\beta)=\frac{1}{m}\sum_{j=1}^{m}U^{ikj}(\hat{\psi},\beta),$$

$$U^{ikj}(\hat{\psi},\beta)=u\left\{Y^{ikj}(\hat{\psi}),\beta\right\},\ w_{\cdot k}=\sum_{i=1}^{c_k}w_{ik}\ \text{and}\ w_{\cdot\cdot}=\sum_{k=1}^{c}\sum_{i=1}^{c_k}w_{ik}$$

Comparison of Methods

- Simulation results for clustering and weighting
- Logistic regression models
  - Imputer using age and gender; analyst using gender only

Method	Simulation Variance	Average Variance Estimate	Relative RMSE
Extended Robins & Wang	0.0444	0.0416	10%
Standard MI (Rubin)	0.0445	0.0319	33%

  - Same imputation model; analyst using working status

Method	Simulation Variance	Average Variance Estimate	Relative RMSE
Extended Robins & Wang	0.0184	0.0170	12%
Standard MI (Rubin)	0.0184	0.0320	86%

Comparison of Methods 2

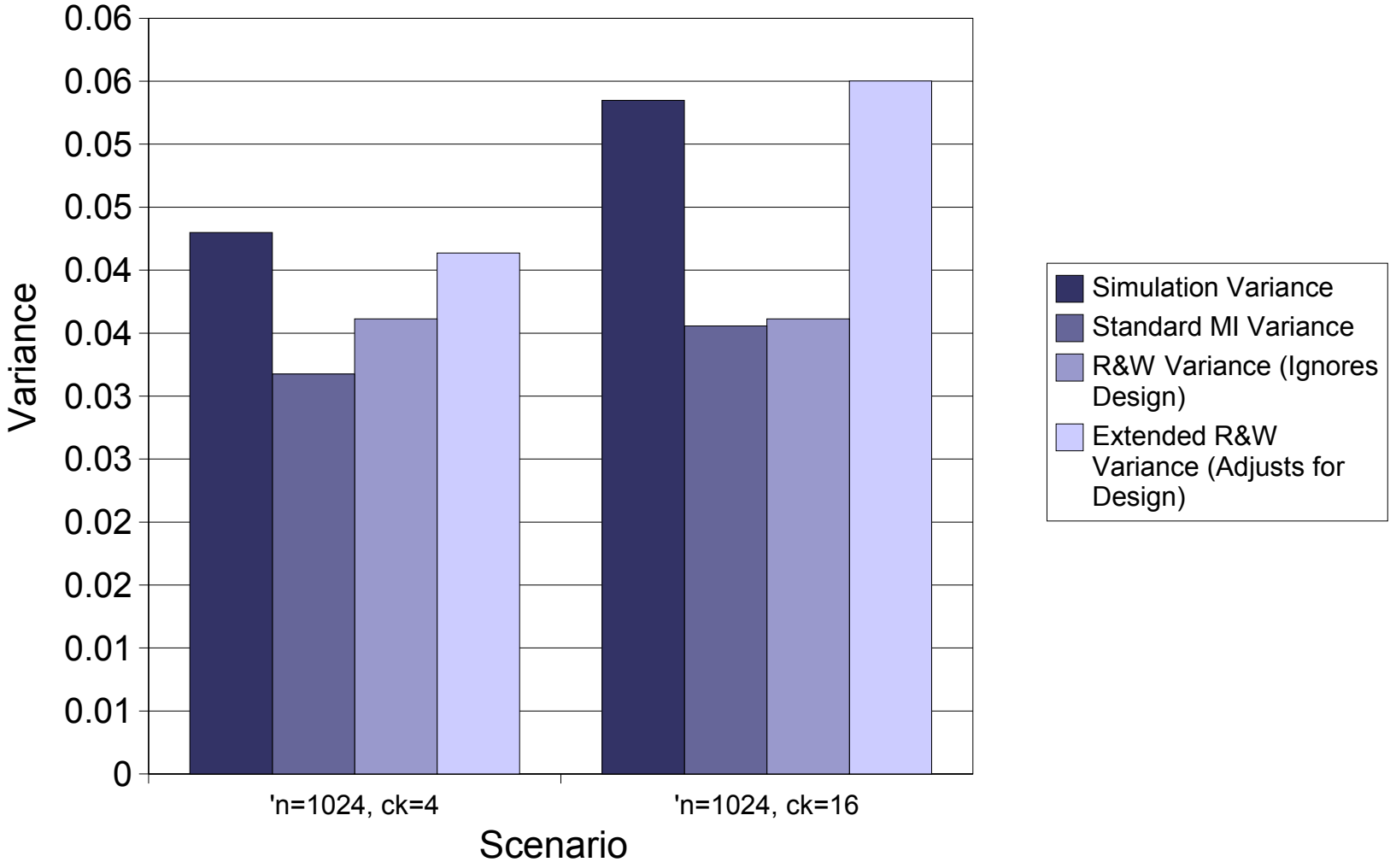
- Imputer and analyst both using age and gender
  - Table shows results for gender parameter

Method	Simulation Variance	Average Variance Estimate	Relative RMSE
Naïve single imputation (complete data variance)	0.0537	0.0211	61%
Extended Robins & Wang	0.0457	0.0426	11%
Standard MI (Rubin)	0.0459	0.0326	34%

- Other simulations show extension usually has an absolute relative bias of less than 10%
  - However it can underestimate variances by approx. 20% when there are heavy weights or few clusters

Effect of Sample Design

Simulation Results for Clustered Designs with Weighting



Application 1

- 2000 National Readership Survey (first month)
  - 43% of values missing for wine drinking in last week
- Imputer: Logistic regression of wine consumption against gender and sex
- Analyst: Logistic regression of wine consumption against gender only

Method	$\hat{\beta}_1$	$\mathrm{Var}(\hat{\beta}_1)$
Extended Robins & Wang	0.2531	0.0407
Standard MI (Rubin)	0.2533	0.0256

Application 2

- 2001 National Crime Victims Survey
  - Measures incidence of victimisation
  - 63% of values missing for offence eligibility
- Multiple imputation with these models:
  - Imputer: Logistic regression of offence eligibility against offence type, gender, age, and living situation
  - Analyst: incidence of each type of (eligible) offence by gender, and also separately by age, ethnicity, NZSEI, employment status and living situation
- Will compare methods for this data

Conclusions

- Standard multiple imputation methods give biased variance estimates for common analyses of complex surveys
  - Underestimates of variance can occur naturally
- Extended estimating equation approach described here gives asympt'ly unbiased variance estimates
  - But more complex to implement than standard MI
  - Extension can underestimate variances when there are few clusters or heavy weights
- More research needed to understand bias & MSE

References

- Barnard, J. and Meng, X.-L. (1999). Application of multiple imputation in medical studies: from AIDS to NHANES, Stat. Methods Med. Res. 8, 17-36.
- Fay, R.E. (1991). A design-based perspective on missing data variance, in Proc. Survey Methods Section, ASA, 266-271.
- Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data, JASA 91, 490-498.
- Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data, 2nd Ed*. Wiley.
- Nielsen, S.F. (2003). Proper and improper multiple imputation, Int. Stat. Review, 593-607.
- Reilly, J.L. (2003). An estimating equations technique for valid inference from imputed survey data, Bull. ISI 55.
- Robins, J.M. and Wang, N. (2000). Inference for imputation estimators, Biometrika 87, 113-124.
- Rubin, D.B. (1987). *Multiple Imputation for Non-response in Surveys*. Wiley, N.Y.
- Rubin, D.B. (1996). Multiple imputation after 18+ years, JASA 91, 473-489.
- SAS Institute (2004). *SAS/STAT 9.1 User's Guide*. SAS Publishing, Cary, N.C.
- Schafer, J.L. (1999). Multiple imputation: a primer, Stat. Methods Med. Res. 8, 3-15.