An Introduction to Multiple Imputation for Missing Items in Complex Surveys

October 17, 2014

Joe Schafer

Center for Statistical Research and Methodology (CSRM)
United States Census Bureau

Views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

Outline

1. Historical Development
2. “How to” Do MI
3. Complexities for Complex Surveys
4. Looking Ahead

MI for complex surveys: Ready for prime time?

1. Development of MI

Theory

- First proposed by Rubin in 1977 for missing income in the March income supplement to the Current Population Survey (Scheuren 2005)

- Rationale and theory presented by Rubin (1987)
  - ❑ connections to Bayesian inference
  - ❑ general methods for creating MIs
  - ❑ rules for combining (e.g. point estimates and SEs)
  - ❑ definition of "proper"

- Criticism from designed-based perspectives by Fay (1992), with response by Meng (1994) (congeniality)

- Properties of Rubin's "variance estimate" (Wang and Robins, 1998; Robins and Wang, 2000; Kim et al., 2006) with response by Rubin (2003)

Implementation

- Handling of univariate missingness, linear and logistic regression, monotone patterns, Bayesian bootstrap using noniterative methods (Rubin, 1987)

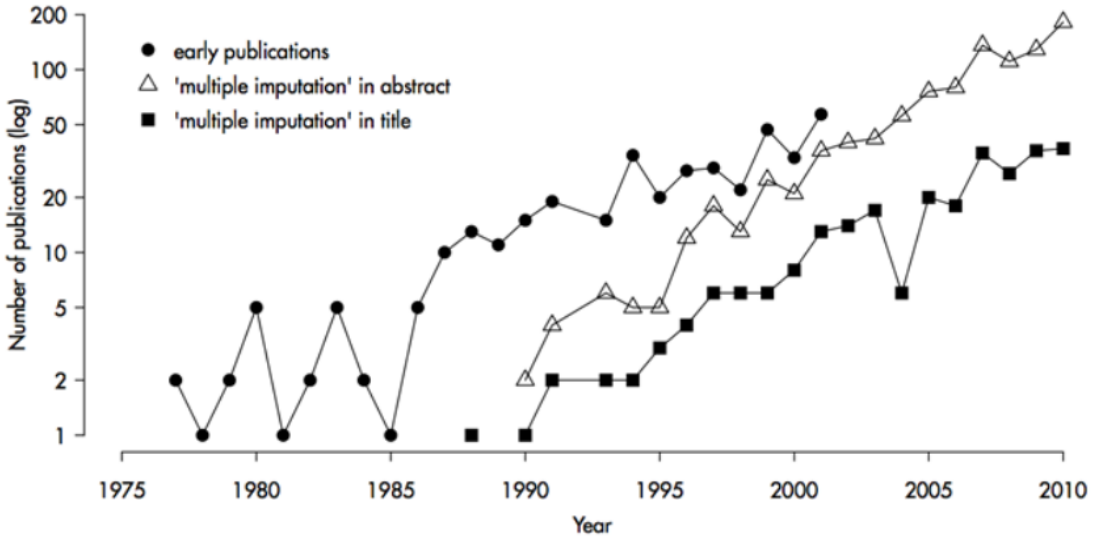
- General multivariate (Swiss cheese) patterns under fully specified joint models for normal, categorical and mixed data via Markov chain Monte Carlo (MCMC) (Schafer, 1997)

- Fully conditional specification / sequential regression / chained equations
  - ❑ Survey of Consumer Finances (Kennickell 1991)
  - ❑ IVEware (Raghunathan, Solenberger and Van Hoewyk, 2002)
  - ❑ mice (Van Buuren and Groothis-Oudshoorn, 2011)

- Many other specialized methods for multivariate data under MAR, and a few under MNAR

Publications

www.multiple-imputation.com



from website of Stef Van Buuren
http://www.stefvanbuuren.nl/mi/

- c. 2005: turning point for acceptance of MI. Now every major statistical package does MI in some fashion (Van Buuren, 2012)

Applications to Surveys, Censuses and Administrative Databases

- Census industry and occupation codes (Clogg et al., 1991)
- Fatality Analysis Reporting System (Heitjan and Rubin, 1991)
- Consumer Expenditure Survey (Raghunathan and Paulin, 1998)
- National Health and Nutrition Examination Survey (Schafer *et al*., 1998)
- Survey of Consumer Finances (Kennickell, 1998)
- National Health Interview Survey (Schenker et al., 2006
- Cancer Care Outcomes Research and Surveillance (He et al., 2009)

2. "How to" Do MI

First step: Generate the MIs

Complete data: $Y = (Y_{obs}, Y_{mis})$
Response indicators: R

Simulate m independent draws of Y_{mis} from

$$P(Y_{mis} \mid Y_{obs}) = \int P(Y_{mis} \mid Y_{obs}, \theta)\, P(\theta \mid Y_{obs})\, d\theta \qquad \text{MAR}$$

$$P(Y_{mis} \mid Y_{obs}, R) = \int P(Y_{mis} \mid Y_{obs}, R, \varphi)\, P(\varphi \mid Y_{obs}, R)\, d\varphi \qquad \text{MNAR}$$

- Under MAR, requires a parametric joint model for the incomplete variables (monotone patterns can be handled noniteratively; Swiss cheese patterns require iteration)
- Under MNAR, requires a model for joint distribution of the incomplete variables and response indicators

Second step: Analyze the completed datasets

In the simplest case, save point and variance estimates from each version of imputed data $j = 1, \ldots, m$

$$\hat{Q}_j = \hat{Q}\left(Y_{obs}, Y_{mis}^{(j)}\right)$$
$$U_j = U\left(Y_{obs}, Y_{mis}^{(j)}\right)$$

- What does the estimand represent, and how is it related (if at all) to the parameters of the imputation model?
  - In surveys, estimand may be a total, mean, ratio, etc. in the finite population
  - okay if imputation method is "proper" in the sense defined by Rubin (1987); see Van Buuren (2012)
  - Meng's (1994) discussions of congeniality
  - Theory still not widely understood by practitioners, but intuitive heuristics go a long way

Third step: Consolidate the results

- Best known, and most widely used, technique is still Rubin's (1987) rules for scalar estimands

  Total variance = Within-imputation + Between-imputation

  analogous to one-way ANOVA, but motivated by Bayesian arguments, and assumes $(\hat{Q} - Q) | Y \sim N(0, U)$

- At least ten other methods (Reiter and Raghunathan, 2007), including
  - ❑ small-sample df (Barnard and Rubin, 1999)
  - ❑ vector Q (Li, Raghunathan and Rubin, 1991)
  - ❑ LR test statistics (Meng and Rubin, 1992)
  - ❑ p-values (Li, Raghunathan, Meng and Rubin, 1991)
  - ❑ rules for partially or fully synthetic data (Reiter, 2003; Raghunathan, Reiter and Rubin, 2003)
  - ❑ rules for "nested MI" (Shen, 2000; Harel, 2003)

How many imputations are needed?

- Early references (e.g., Rubin, 1987) suggest that only a few (say, $m = 3$ or $m = 5$) imputations are needed when the rate of missing information λ is moderate

$$\text{relative efficiency of point estimate} = \frac{1}{(1 + \lambda/m)}$$

- With modern computational speed and storage capacity, other considerations suggest that we should take more (say, 25-50) (Graham, Olchowski and Gilreath, 2007)

- If we need an accurate estimate of the rate of missing information, we will need $m > 100$. Harel (2003) shows that

$$\sqrt{m}(\hat{\lambda} - \lambda) \rightarrow N(0, 2\lambda^2(1-\lambda)^2)$$
$$\sqrt{m}\left(\text{logit}(\hat{\lambda}) - \text{logit}(\lambda)\right) \rightarrow N(0, 2)$$

3. Complexities for Complex Samples

Issue: compatibility of imputation model and analysis procedures

- Rubin's definition of proper is difficult to verify in practice (Van Buuren, 2012)

- Meng's (1994) discussion of congeniality and superefficiency

- Imputers often have access to extra information and may make extra assumptions

- Mismatch between models may be harmful or helpful, and it depends on whether the extra assumptions are true; see heuristic discussion by Schafer (2003)

Issue: Popular MI software (joint modeling) assumes multivariate normality, but survey variables tend to be categorical or mixed types

- Loglinear and general location models (Schafer, 1997) are okay when number of variables is small (say, <20)

- Impute as normal, then categorize the imputed values by rounding or coin flipping (Allison, 2005, 2006; Bernaards, Belin and Schafer, 2007; Yucel and Zaslavsky, 2008; Demirtas, 2009, 2010)

- Models for mixed variables based on latent normal structure (Boscardin, Zhang and Belin, 2008; He, 2012); this is a special case of multivariate copula models (Pitt, Chan and Kohn, 2006; Smith and Khaled, 2011)

Issue: potentially large number of variables to be imputed

With a joint normal model, we can reduce the dimensionality of covariance parameters

- exploratory factor models (Song and Belin, 2004)
- confirmatory factor models for multi-themed questionnaires (Liu, 2010)
- hierarchical Bayesian smoothing toward a structured covariance matrix (Boscardin and Zhang, 2004; He, 2012)

Sequential regression / chained equations can handle large numbers of variables

- conditionals may not be compatible with a true joint distribution (Gelman and Speed, 1993), but in practice this doesn't seem to matter (Van Buuren, 2012)
- Rubin (2003) uses incompatible chained equations only to complete the monotone pattern

Issue: preserving complicated interactions

- Normal models have no interactions; we can preserve some by data splitting

- Interactions can be included in sequential regression models

- Sequential regression with random forests (Doove, Van Buuren and Dusseldorp, 2014)

- Jerry Reiter's presentation today on Bayesian mixtures

Issue: Important features of sample design ought to be reflected in the imputation model

- fixed effects for stratifying variables or stratum indicators
- multilevel multivariate models with random effects for clusters (Schafer et al., 1998)
- cross-wave correlations in longitudinal surveys (Schafer and Yucel, 2002)
- spline bases for functions of sample weights (Zhang and Little, 2009)
- mixed-effects models in sequential regression (Yucel, Schenker and Raghunathan, 2006; Van Buuren, 2012)

Issue: Hierarchical or multilevel data structures with missing values at multiple levels

- Earlier procedures for multilevel imputation assumed missing values at only one level (Schafer and Yucel, 2002)

- Yucel (2008) chained multivariate models for different levels; Mistler (2013) SAS macro for PROC MI

- Carpenter, Kenward and Vansteelandt (2006) REALCOM-IMPUTE software

- Book by Carpenter and Kenward (2013) with applications in MLwin; also see book chapter by Van Buuren (2011)

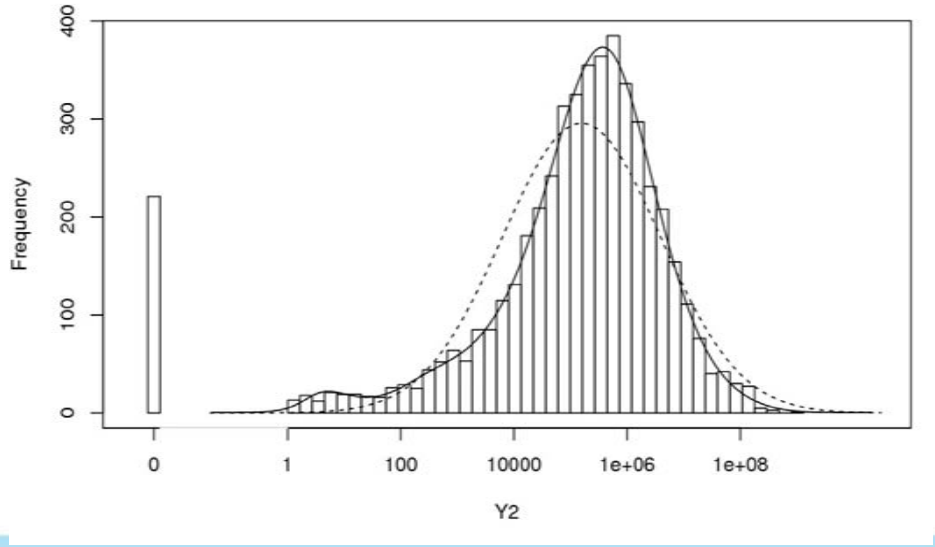
- What about hierarchical categorical data with complicated cross-level relationships and constraints? (e.g., Census short form)

Issue: imputed values need to satisfy logical constraints

- questionnaire skip patterns (He et al., 2009)
- sum constraints (Kim et al., 2014)
- logical zeroes induced by edit rules (Reiter et al.)
- Even if observed values pass edit tests, they might not be error free; consider multiple imputation to account for response errors (Ghosh-Dastidar and Schafer, 2003)

Issue: semicontinuous variables and unusual marginal distributions

- two-part models in sequential regression
- joint models for semicontinuous variables (Schafer and Olsen, 1998; Javaras and Van Dyk, 2003)
- log and power transformations might still not work for continuous part; may need Bayesian nonparametric modeling



4. Looking Ahead

Explosion of new models, techniques, algorithms over last 15 years. But are they ready for prime time?

Many nonstatistical issues remain…

- availability, reliability, sustainability of software
- perceived and actual difficulty of implementation for production
- perceived and actual difficulty of explaining to policymakers and public
- organizational culture and priorities

References

(in a separate file)