Discussion of Three Papers on Treatment of Missing Data

Nathaniel Schenker
Senior Scientist for Research and Methodology
National Center for Health Statistics
nschenker@cdc.gov

Presented at the FCSM Research Conference
November 14, 2001

Introduction

- I enjoyed reading the three papers and listening to the presentations of them.

- First two papers (Fetter; Piela and Laaksonen): regression-based methods for imputing continuous and/or categorical missing data

- Third paper (Greene, Smith, Levenson, Hiser, and Mah): raking-based methods for handling missing data when the variables are categorical and form a contingency table with several dimensions and many cells

- I will discuss the first two papers first and discuss the third paper afterwards.

Explicit models vs. implicit models

- **Fetter's models:**
  - **MCMI procedure based on explicit model**
  - **RER procedure has both explicit (regression) and implicit (empirical residual) components**

- **Piela & Laaksonen's models:**
  - **CART procedures based on implicit models**

- **Implicit models often have a nonparametric flavor; attempt to be more robust**

- **Schenker and Taylor (1996) studied "partially parametric" techniques**

- Results from Schenker and Taylor (1996, Table 4) on estimating the distribution function at the median, when the regression model underlying the multiple-imputation method is misspecified regarding the transformation of the outcome variable:

	Imputation Method			
	Fully Parametric	Predictive Mean Matching	Local Residual Draw	No Missing Data
MSE	2.37	1.43	1.31	1.00
Coverage of Nominal 95% Interval	86.6	93.2	94.1	94.9

Multiple imputation

- ***M*** **independent draws from**

$$p(Y_{mis} \mid Y_{obs}) = \int p(Y_{mis} \mid Y_{obs}, \theta)\, p(\theta \mid Y_{obs})\, d\theta$$

- **For many models, can use two-step procedure to produce each draw of** Y_{mis}**:**

1. **Draw a value** θ^* **from** $p(\theta \mid Y_{obs})$
2. **Draw a value** Y^*_{mis} **from** $p(Y_{mis} \mid Y_{obs}, \theta^*)$

- **Can follow two-step paradigm for partially-parametric and/or nonparametric models**

  • e.g., for RER, for each of the *M* sets of imputations, draw regression parameters from approximate posterior distribution prior to calculating predicted values and residuals (see Schenker and Taylor 1996)

  • e.g., for each of the *M* imputations of Y_{mis}, run CART on a bootstrap sample to determine the tree

Additional comments on Fetter

- **Designed missing data to reduce respondent burden is an attractive idea**
  - **Reminiscent of one-sixth sampling for census "long form"**

- **Consider one multivariate procedure for all of the logistic regressions?**
  - **e.g., sequential regression imputation (Raghunathan *et al.* 2001)**
  - **Might help to preserve relationships among the variables**

- **Don't forget to reflect uncertainty in estimating logistic regression parameters**

- **Unclear of the need to set some zero values to "missing" before running MCMI**
  - **Could cause bias due to nonignorable missingness?**
  - **Reason for lower precision of MCMI relative to RER?**
  - **Seems preferable to condition on zero values**

- **Drawing from "local" empirical residuals rather than "global" empirical residuals might improve robustness to model misspecification (see Schenker and Taylor 1996)**

Additional comments on Piela and Laaksonen

- **Potential for achieving robust imputations**

- **Can the method be used when there are missing values in the covariates?**

- **Difficult to judge performance based on one data set. Could just be "unlucky".**
  - **Useful to examine performance under repeated sampling**
  - **Useful to consider properties of inferences (multiple imputation?)**

- **Is it possible to build an assumption of nonignorable missing data into CART-based imputation?**

- **Problems with mode or mean imputation**
  - **Distorts distribution of variables**
  - **Biases when estimator is nonlinear in data**

- **Choosing the number of explanatory variables and the number of terminal nodes**
  - **Bias/variance trade-off**
  - **Analogous to choosing the number of donor cells in a hot-deck scheme**
  - **Schenker and Taylor (1996) used an adaptive method for choosing the number of prospective donors for each missing value**

Comments on Greene *et al.*

- **Greene *et al.* method has desirable properties relative to "national estimates method"**
  - **All marginals are preserved**
  - **Independent of ordering of variables**

- **Might be interesting to compare Greene *et al.* method with the "national estimates" method with respect to models underlying:**
  - **contingency table**
  - **missing-data mechanism**

- **Consider prior distributions to handle sparse data?**
  - **Rubin and Schenker (1987) and Clogg *et al.* (1991) discussed simple Bayesian methods for logistic regression**

- **Raking generally is useful when the marginal distributions for a table are known but the distributions inside the table are not known. In the application to fire data:**
  - **How precisely are the marginals known?**
  - **Could other methods for handling missing data in contingency tables be useful?**

- **Consider Table 1 of Greene *et al.* (this is Table 1 of the draft that was sent to me)**

	Female	Male	Unknown	Total
Old	**65**	**30**	**5**	**100**
Young	**25**	**50**	**25**	**100**
Unknown	**10**	**2000**	**70**	**2080**
Total	**100**	**2080**	**100**	**2280**

- **Marginal distribution of age not known very precisely, since 2080 values of age are missing**

- **Is it reasonable to distribute the 2080 missing values on age 50/50 into young and old, and then treat the resulting marginals as the known "population" values for raking, as is done in Greene *et al.*?**
    - **Note that 2000 of the missing values on age are for males**

- **Results of a few iterations of Greene *et al.* procedure:**

	Female	Male	Total	"Population"
Old	**84.3**	**1055.8**	**1140.1**	**1140.0**
Young	**20.6**	**1119.3**	**1139.9**	**1140.0**
Total	**104.9**	**2175.1**	**2280.0**	**2280.0**
"Population"	**104.6**	**2175.4**	**2280.0**	

- **"Population" marginals preserved**
- **Odds ratio from original table preserved**
- **Distributions of age by gender from original table not preserved**
- **Some young females from original table "removed"; i.e., cell count for young females smaller than that in original table**

- **Results of a few iterations of EM algorithm (done by hand, with three significant digits of precision) for maximum likelihood under a saturated multinomial model, assuming ignorable missing data (see Little and Rubin 1987, Section 9.3):**

	Female	Male	Total
Old	74.9	798	873
Young	29.3	1378	1407
Total	104	2176	2280

- **Gender marginals close to those for raking, but age marginals much different**
- **Odds ratio from original table nearly preserved**
- **Distributions of age by gender from original table nearly preserved**
- **Cell counts all greater than those in original table**

References

Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B., and Weidman, L. (1991), "Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression," *Journal of the American Statistical Association*, 86, 68-78.

Little, R.J.A., and Rubin, D.B. (1987), "Statistical Analysis with Missing Data," Wiley: New York.

Raghunathan, T.E., Lepkowski, J.M., Van Howewyk, J., and Solenberger, P. (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," *Survey Methodology*, 27, 85-95.

Rubin, D.B., and Schenker, N. (1987), "Logit-Based Interval Estimation for Binomial Data Using the Jeffreys Prior," *Sociological Methodology*, 17, 131-144.

Schenker, N., and Taylor, J.M.G. (1996), "Partially Parametric Techniques for Multiple Imputation," *Computational Statistics & Data Analysis*, 22, 425-446.