MULTIPLE IMPUTATION OF MISSING INCOME DATA IN THE NATIONAL HEALTH INTERVIEW SURVEY*

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

Joint work with
Pei-Lu Chiu, Alan J. Cohen, Diane M. Makuc, Trivellore E. Raghunathan, and Guangyu Zhang

Presented at the
NCHS Data Users Conference
July 12, 2006

*** Article to appear in the**
Journal of the American Statistical Association
Manuscript available by request

CONTENTS

1. THE NHIS AND MISSING DATA ON INCOME

- **National Health Interview Survey (NHIS)**

  - Principal source of information on the health of the civilian noninstitutionalized population

  - Data collected at both family and person levels

  - Multistage area probability sample of about 40,000 households including about 100,000 persons, with oversampling of Blacks and Hispanics

    - **Allows national and subnational estimation**

  - Contains items on health, demographic, and socioeconomic characteristics

    - **Allows the study of relationships between health and other characteristics**

- **Income items collected in NHIS**

  - Personal earnings (wages, salaries, tips, commissions) in past calendar year for each adult who worked for pay

  - Family income (including personal earnings and other sources) in past calendar year
    - ♦ **"Exact" dollar amount or**
    - ♦ **One of 2 categories (< $20,000 or ≥ $20,000) or**
    - ♦ **One of 44 categories (up to $75,000+)**

- **Family income (or its ratio to Federal poverty threshold) frequently used in analyses of NHIS data because of strong association with health and relevance to health policy**

- **Missing data on income in NHIS**

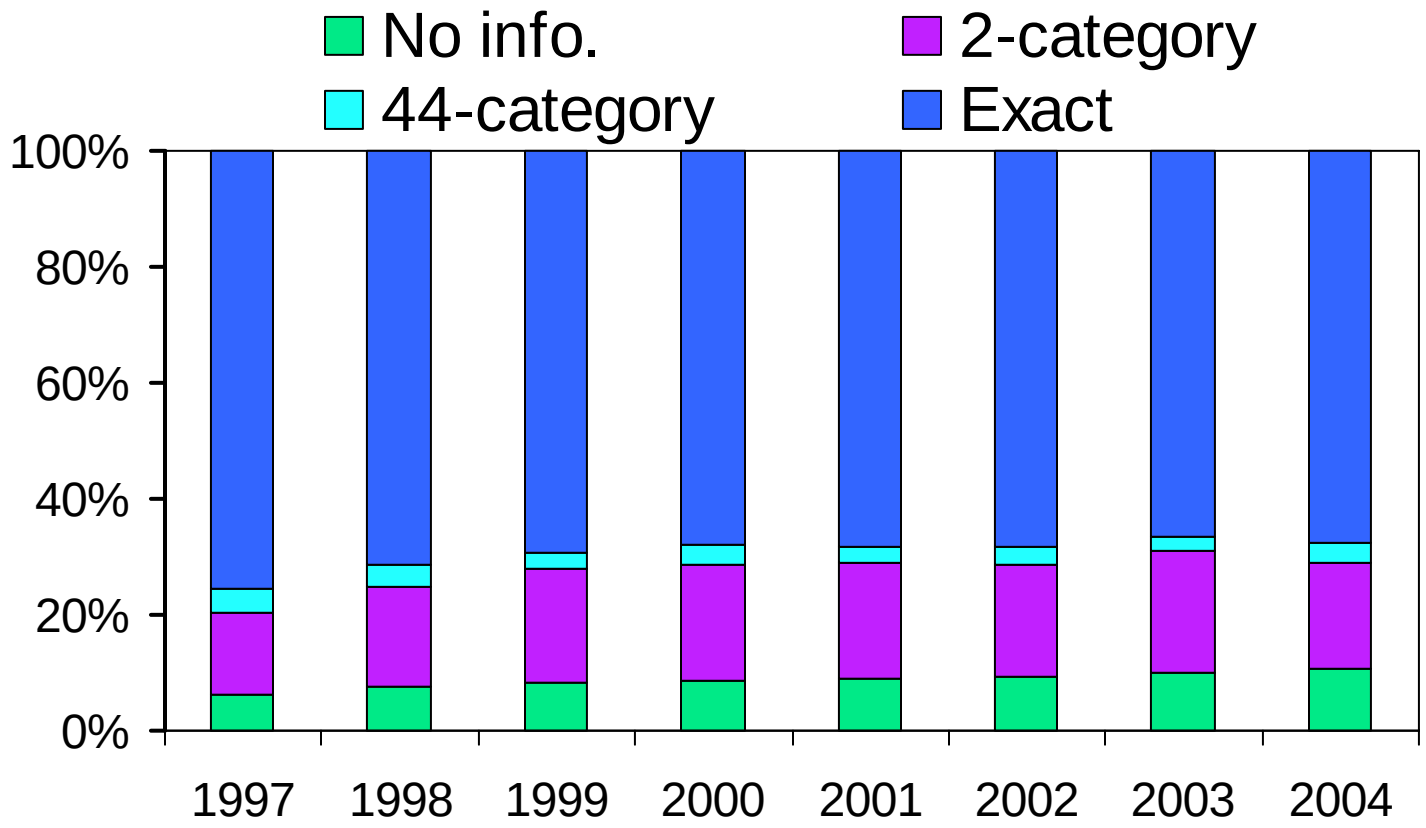
  - High item nonresponse rates

  - Missingness appears to be related to several other characteristics, such as health, health insurance, age, race, country of birth, and region of residence

  $\Rightarrow$ Possible bias and higher variance in analyses that delete observations with missing data

- **Nonresponse for most other variables is low (< 2%)**

Percent distribution of types of family income responses by year for the NHIS in 1997 – 2004



- **Nonresponse rates for personal earnings are similar to those for exact family income**

Results from a logistic regression, with an indicator variable for nonresponse on both the exact and 44-category values of family income as the outcome, and selected variables as predictors, for persons of ages less than 65: 2001 NHIS

Variable	Odds Ratio	95% Confidence Interval
Has health insurance?		
No	**1.54**	**(1.43, 1.66)**
Yes (reference)	**---**	**---**
Has limitations of activities?		
Yes	**0.77**	**(0.72, 0.83)**
No (reference)	**---**	**---**
Age		
< 18	**0.64**	**(0.60, 0.69)**
18 – 24	**0.69**	**(0.63, 0.76)**
25 – 34	**0.58**	**(0.53, 0.62)**
35 – 44	**0.70**	**(0.64, 0.76)**
45 – 54	**0.82**	**(0.76, 0.88)**
55 – 64 (reference)	**---**	**---**
Gender		
Male	**0.98**	**(0.96, 1.01)**
Female (reference)	**---**	**---**
Race/Ethnicity		
Hispanic	**1.01**	**(0.91, 1.12)**
Non-Hispanic Black	**1.21**	**(1.09, 1.34)**
Non-Hispanic Other	**0.92**	**(0.78, 1.08)**
Non-Hispanic White (reference)	**---**	**---**
Born in the US?		
No	**1.14**	**(1.05, 1.25)**
Yes (reference)	**---**	**---**
Region of residence		
Northeast	**1.05**	**(0.90, 1.23)**
South	**0.79**	**(0.71, 0.88)**
West	**0.94**	**(0.84, 1.06)**
Midwest (reference)	**---**	**---**
Resides in metropolitan area?		
No	**0.91**	**(0.79, 1.04)**
Yes (reference)	**---**	**---**

2. MULTIPLE IMPUTATION FOR THE NHIS

- **Imputing just once and treating imputed values as true values $\Rightarrow$ underestimates of uncertainty**

  - **Standard errors too small**

  - **P-values too small (i.e., tests too significant)**

- **Multiple imputation (Rubin 1987)**

  - Impute for missing values several (*M*) times using random draws from the predictive distribution of the missing data given the observed data

  - Analyze each of the *M* completed data sets using methods designed for complete data; then combine point estimates and estimated variances

  - **Combined point estimate is average of point estimates from *M* data sets**
  - **Total estimated variance is:**

  (1) average of variances from *M* data sets, plus

  (2) variation among point estimates from *M* data sets

  - **component (2) reflects extra uncertainty due to missing data**

• Project to multiply impute income items in the NHIS, beginning with 1997

- *M* = 5 sets of imputations of:

- **♦ employment status for adults (< 4% missing)**
- **♦ personal earnings for adults who worked for pay**
- **♦ family income (and ratio of family income to Federal poverty threshold)**

- Imputed income files for 1997 – 2004, with documentation, available at NHIS Web site:

www.cdc.gov/nchs/nhis.htm

- Used Sequential Regression Multivariate Imputation (Raghunathan *et al.* 2001), as implemented in IVEware (Institute for Social Research, University of Michigan)

- **Complicating issues handled during imputation**

  - **Hierarchical structure of data**
    - ♦ **Families and persons**

  - **Structural dependencies between variables**
    - ♦ **e.g., employment status and personal earnings for adults**

  - **Imputation within bounds**
    - ♦ **e.g., families for which income not reported exactly, but rather within coarser categories**

  - **Several variables used as predictors**
    - ♦ **Different types (continuous, categorical, count)**
    - ♦ **Small amounts of missingness (mostly $<$ 2%)**

- **Used about 60 covariates for person-level imputations and for family-level imputations, including:**

  - **Demographic variables**
  - **Family structure**
  - **Geographic variables**
  - **Education**
  - **Employment status**
  - **Hours worked per week**
  - **Sources of income**
  - **Limitations of activities**
  - **Health conditions that caused limitations**
  - **Overall health**
  - **Health care use**
  - **Health insurance**
  - **Indicators for stratum-by-PSU combinations**
  - **Survey weights**
  - **SSU-level summaries of family income**

- **Question: OK to use health items as covariates in the model for imputing income, given that filled-in data will be used to analyze health by levels of income?**

- **Answer: Yes.**

  - Theory of multiple imputation implies that all observed data should be conditioned upon in drawing imputed values for missing data (Rubin 1987)

  - If health items were not included as covariates in imputation, then the relationship between health and income in the filled-in data would be attenuated

  - See Little (1992) and Little and Raghunathan (1997) for further discussion

3. RESULTS FOR FAMILY INCOME IN THE 2001 NHIS

• **Estimated percentage of persons of ages 45-64 in fair or poor health, by ratio of family income to Federal poverty threshold: 2001 NHIS**

Ratio to Poverty Threshold	No Imp. (NI)		Single Imp. (SI)		Mult. Imp. (MI)		Ratio of SEs	
	Est.	SE	Est.	SE	Est.	SE	NI ÷ MI	SI ÷ MI
< 1.00	45.6	1.68	39.4	1.34	39.9	1.54	1.09	0.87
1.00 – 1.99	32.7	1.32	29.8	1.03	29.3	1.11	1.19	0.93
2.00 – 3.99	16.1	0.63	16.0	0.51	15.9	0.55	1.15	0.94
4.00+	5.9	0.34	6.1	0.27	6.2	0.30	1.11	0.90

- **Estimated percentage of persons of ages 45-64 in fair or poor health, by 2-category family income, for reporters and non-reporters of exact family income: 2001 NHIS**

2-Category Family Income	Exact Family Income Reported	Exact Family Income Not Reported
< 20k	41.6	33.5
≥ 20k	10.6	11.0

4. FUTURE WORK

- Additional research is needed regarding possible inconsistencies between total family income and total of personal earnings within family
(total family income < total personal earnings within family)

  - Attempts to enforce consistency through imputation appeared to increase bias

- Imputations for future years as data become available

REFERENCES

Little, R.J.A. (1992), "Regression With Missing X's: A Review," *Journal of the American Statistical Association*, 87, 1227-1237.

Little, R.J.A., and Raghunathan, T.E. (1997), "Should Imputation of Missing Data Condition on All Observed Variables?" *American Statistical Association Proceedings of the Section on Survey Research Methods*, 617-622.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, 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. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

Schenker, N., Raghunathan, T.E., Chiu, P.-L., Makuc, D.M., Zhang, G., and Cohen, A.J. (forthcoming), "Multiple Imputation of Missing Income Data in the National Health Interview Survey," to appear in the *Journal of the American Statistical Association*.