Estimating Nonlinear Models With Multiply Imputed Data

Catherine Phillips Montalto[1] and **Yoonkyung Yuh**[2]

Repeated-imputation inference (RII) techniques for estimating nonlinear models with multiply imputed data are described. RII techniques are used to estimate a logit model using the 1995 Survey of Consumer Finances. RII techniques use all information available in multiply imputed data and incorporate estimates of imputation error. The advantage of RII techniques for analysis of multiply imputed data is that RII techniques produce ***more efficient*** *estimates and provide a basis for* ***more valid inference****. Researchers who do not use RII techniques when estimating nonlinear models on multiply imputed data may incorrectly conclude that some independent variables have statistically significant effects.*

Key Words: *Logit, Probit, Repeated-imputation inference (RII), Survey of Consumer Finances, Tobit*

Multiple imputation is a technique commonly used to deal with missing information on individual items in survey data. Multiple imputation employs multivariate statistical methods to impute missing data resulting in multiple complete data sets.[a] Since 1989, the Survey of Consumer Finances (SCF) data files have contained five complete data sets, referred to as *implicates*.[b] The benefit to researchers of these multiply imputed data files is that the data files contain no missing values; the cost is that researchers must learn how to analyze data appropriately in the presence of five complete data sets.

An appropriate method of analyzing multiply imputed data is to combine the results obtained independently on each of the separate implicates using multiple imputation combining rules. Inferences based on the appropriately combined results are called *repeated-imputation inferences* (RII) (Rubin, 1987, 1996). Montalto and Sung (1996) provide a clear discussion of multiple imputation in the SCF from a user's point of view, and use the multiple imputation combining rules to obtain estimates of descriptive statistics and ordinary least squares regression coefficients. The use of these two specific examples has caused some users of the SCF to question whether RII techniques can be applied more broadly, for example to the estimation of nonlinear models.

The purpose of this research note is to address the appropriateness of RII techniques in a broad range of applications including nonlinear estimation, to briefly explain the intuition behind RII techniques, and to emphasize the advantages of using RII techniques to obtain efficient estimates and make valid inferences from multiply imputed data. An example using RII techniques in nonlinear regression analysis with Survey of Consumer Finances data is presented. A more technical discussion of repeated-imputation inference techniques is presented in the Appendix.

When Is It Appropriate to Use RII Techniques?

RII techniques are appropriate whenever inferences made from the data analysis are based on point estimates and variances. For descriptive statistics, inferences are based on estimates of the mean and the variance of the mean. (The square root of the variance is the standard error of the mean.) For linear regressions, inferences are based on estimates of regression coefficients and the standard errors of these estimates. Similarly, correlations, factor loadings, populations proportions, and nonlinear regressions (including logit, probit and tobit) yield inferences based on estimates and variances.

For descriptive statistics, analysis of each complete data set produces an estimate of a mean and the variance of the mean. When a researcher has an interest in one variable, the estimate of the mean and the variance of the mean are single numbers. When a researcher has an interest in two or more variables, the estimate is represented by a vector, and the variance is represented by a variance-covariance matrix.

For ordinary least squares and nonlinear regression, analysis of each implicate produces a k-dimensional

[1]Catherine Phillips Montalto, Assistant Professor, Consumer and Textile Sciences Department, The Ohio State University, 1787 Neil Avenue, Columbus, OH 43210-1295. Phone: (614) 292-4571. Fax: (614) 292-7536. E-mail: montalto.2@osu.edu

[2]Yoonkyung Yuh received her Ph.D. from The Ohio State University in September, 1998. E-mail: yuh@afcpe.org

vector of coefficients, and a k x k variance-covariance matrix. In the case of nonlinear regression models, the estimate vector and variance-covariance matrix are based on asymptotic calculations, but RII techniques are still appropriate.

The criteria for determining appropriateness of RII techniques are independent of the functional form of the estimation method. RII techniques are appropriate whenever inferences made from the data analysis are based on point estimates and variances of the point estimates.

Intuition Behind RII Techniques

The multiple imputation combining rules are straightforward, and require only the calculation of means and variances of the results obtained independently from the separate implicates (Rubin, 1987). Point estimates from the separate implicates are averaged to create a single parameter estimate. The average variance within each implicate and the variance between the implicates are summed to create an estimate of the total variance.

Advantages of Using RII Techniques

Two advantages of RII techniques will be emphasized:
(1) RII techniques produce more efficient estimates, and
(2) RII techniques provide a basis for more valid inference.

With respect to efficiency, since the RII estimates use data from all implicates they are more efficient than estimates that use data from a single implicate. The multiple imputation combining rules average over the variability between the individual implicates to produce the best estimate of what the results would have been if the missing data had been observed.

An extremely important advantage of RII techniques is that they provide a basis for more valid inference since the variability due to missing values (i.e. imputation error) is incorporated into the variance estimates. In general, this will increase the estimate of variance compared to estimates that ignore this variability, resulting in more stringent tests for statistical significance. When imputation error is ignored, the variance estimate will be biased and will underestimate the true variance. Inferences based on the biased variance estimate may incorrectly indicate that some relationships are statistically significant.

RII versus Alternative Analytical Approaches

Analytical approaches other than RII have been used to analyze multiply imputed data. Two approaches that have often been used include analysis of data from only a single implicate, and analysis of data "averaged" across the implicates. Montalto and Sung (1996) cite numerous studies that have analyzed single implicates of the SCF. Kennickell (1997, lines 152-156) describes analysis of *averaged* data. There are limitations to both approaches.

Analysis of data from only one implicate of a multiply imputed data set implicitly treats the imputed values as if they are known with certainty. Since the variability due to missing values is ignored, the estimates of variance will be too small, and the statistical significance of relationships will be overestimated. Additionally, parameter estimates from only one implicate will be less efficient than parameter estimates that use data from all implicates.

Another approach commonly used to analyze multiply imputed data is to average the variables across the multiple complete data sets, and then analyze the *averaged* data. Point estimates derived in this method are equivalent to point estimates derived by RII techniques, and therefore are efficient. However, the variance estimates ignore the variability due to missing values (i.e. imputation error), and as a result the statistical significance of relationships will be overestimated.

The risk of not using RII techniques to analyze multiply imputed data is highest when the extent of variability between the imputed values is high. The extent of variability between the separate implicates depends on the proportion of information that has been imputed as well as the variation within the stochastic imputation process.

Empirical Example: Retirement Wealth Adequacy

The analysis of retirement wealth adequacy by Yuh, Montalto and Hanna (1998) is used to illustrate the risk of ignoring imputation error.[c] The dependent variable is an indicator variable for retirement wealth adequacy. Due to the dichotomous nature of the dependent variable, logistic regression is used for the analysis. Independent variables include demographic characteristics of the householder, financial characteristics, saving/investment decision variables, and attitude/expectation variables. There are a total of

28 independent variables in the model (Table 1).

Table 1
Logistic analysis of retirement wealth adequacy

Variables	Sig.
Demographic Characteristics	
Age (reference category: 55 and over)	
35-44	V
45-54	V
Education (reference category: less than high school grad.)	
high school grad.	N
some college	N
college and more	N
Marital Status (reference category: couple)	
unmarried male	N
unmarried female	N
Race/ Ethnicity (reference category: White non-Hispanic)	
Black non-Hispanic	V
Hispanic	V
Other (including Asian American)	N
Financial Characteristics	
Log of normal income	+
DB ownership	+
DC ownership	+
Housing Tenure (reference category: own without mortgage)	
rent	-
own with mortgage	-
Saving/ Investment Decision Variables	
Retirement Age (reference category: retire 61 or earlier)	
retire 62-65	+
retire 66 or later	+
Stock Shares(of assets excluding housing asset). (reference:0%)	
0%<stock <13.5%	+
13.5%≤stock<36.5%	+
stock≥36.5%	+
Retirement as a saving goal	N
Spending≥income	-
Attitude/ Expectation Variables	
Subjective Life Expectancy (reference: expect to live > 42)	
expect to live ≤ 24 years	V
24 < expect to live ≤32	V
32 < expect to live ≤42	N
High risk taking	N
Expect enough pension	N
Expect income growth	V

V = variable was statistically significant in some but not all of the implicates and the RII results.
N = variable was not statistically significant in any of the implicates and thus was not statistically significant in the RII results.
+ = variable was positive and statistically significant across the five implicates and in the RII results.
- = variable was negative and statistically significant across the five implicates and in the RII results.

Eleven of the independent variables are statistically significant and consistent in terms of sign across the five implicates and in the RII results.[d] These variables include the variables measuring financial characteristics and saving/investment decisions (with the exception of the variable indicating if retirement is a saving goal).

Ten of the independent variables are NOT statistically significant in any of the implicates and thus are not statistically significant in the RII results. Seven of the independent variables are statistically significant in some but not all of the five implicates and the RII results. These variables include the variables measuring the demographic characteristics of age and race/ethnicity of the respondent (with the exception of the indicator variable for other race); two of the three attitude/expectation variables measuring subjective life expectancy, and if the household expects income growth in the future. These variables illustrate clearly how empirical results from individual implicates can differ from one another, as well as from the RII results.

Selected variables are used to illustrate the risks of not using RII techniques to analyze multiply imputed data. Results for Implicate 1 and Implicate 2 indicate that households with a Black non-Hispanic householder are less likely to have adequate retirement wealth than otherwise similar households with a White non-Hispanic householder. However, the effect of having a Black non-Hispanic householder is not statistically significant in the results for Implicates 3, 4, or 5, or in the RII results. Thus, a researcher basing inferences on analysis of only Implicate 1 or 2 would incorrectly conclude that households with a Black non-Hispanic householder are less likely to have adequate retirement wealth.

Since the logit coefficients cannot be directly compared to evaluate the magnitude of differences in the effects of estimated coefficients across the five implicates, odds ratios are calculated for selected variables.[e] There is some evidence of more variability in the magnitude of effects for variables that have inconsistent results (in terms of statistical significance) across the implicates, compared to variables that are statistically significant across all implicates.

The variability in the magnitude of effects for variables

that have inconsistent results (in terms of statistical significance) across the five implicates is illustrated with the indicator variable for households with a Black, non-Hispanic householder. This odds ratio ranges from .549 to .819 across the five implicates, indicating that these households are only 55% to 82% as likely as otherwise similar households with a White non-Hispanic householder to be adequately prepared for retirement. The odds ratio based on the RII coefficient is .630. Thus, the odds ratio based on results of single implicates ranges from 13% below to 30% above the odds ratio based on the RII results -- a range of 43 percentage points.[f]

The variability in the magnitude of effects for variables that are statistically significant across all implicates is illustrated with the variable indicating whether the household spent at least as much as income. This odds ratio ranges from .102 to .125 across the five implicates, compared to an odds ratio of .114 based on the RII coefficient. The odds ratios based on results of single implicates ranges from 10% below to 10% above the odds ratio based on the RII results — a range of only 20 percentage points.

In most (but not all cases) the range of the odds ratios based on results of single implicates relative to the odds ratio based on the RII coefficient is larger for variables with inconsistent results across the five implicates compared to variables that are statistically significant across all implicates, and therefore, in the RII results. To some extent this finding is expected, but nonetheless, it provides another way of assessing the practical implications of alternative approaches to analyzing multiply imputed data.

Endnotes

a. *A complete data set is a data set free of missing data.*
b. *Kennickell, Starr-McCluer & Suden (1997) describe the survey, the survey procedures and the statistical measures.*
c. *Refer to Yuh, Montalto & Hanna (1998) for specific information on the method of analysis and variable measurement, as well as discussion of the results.*
d. *A table summarizing the logistic results for each of the separate implicates, as well as the RII results is available at: www.afcpe.org/nonr.htm*
Additionally, this table illustrates the calculation of the test statistic for the overall significance of the RII logit equation that is described in the following Appendix.
e. *The odds ratio for a dichotomous independent variable is calculated as e^{β}. These odds ratios are presented in the table cited in endnote d.*
f. *Calculation: (.549-.630)/.630=-.13; (.819-.630)/.630=.3*

Appendix

An Example Using RII Techniques for Nonlinear Regression Analysis

(Notation closely follows Montalto and Sung (1996) which closely follows Rubin (1987)). The nonlinear regression analysis (for example, logit, probit, or tobit) is conducted on each of the five implicates separately. The results obtained independently from the five separate implicates are combined to obtain the RII estimates.

The best estimate of the nonlinear regression coefficients is the average of the results from the five implicates (m=5) where Q_i is a 1xk vector.

$$\overline{Q}_m = \frac{\sum_{i=1}^{m} Q_i}{m} \qquad \textbf{(1)}$$

The within imputation variance is the average of the variance-covariance matrices from the five implicates where U_i is a kxk matrix.

$$\overline{U}_m = \frac{\sum_{i=1}^{m} U_i}{m} \qquad \textbf{(2)}$$

The between imputation variance is the sample variance in the estimates of Q_i from the five implicates and is estimated by

$$B_m = \frac{\sum_{i=1}^{m} \left(Q_i - \overline{Q}_m\right)^t \left(Q_i - \overline{Q}_m\right)}{m - 1} \qquad \textbf{(3)}$$

The transpose of the vector is indicated by *t*.

The total variance-covariance matrix is given by

$$T_m = \overline{U}_m + \left(1 + m^{-1}\right) B_m \qquad \textbf{(4)}$$

A Wald chi-squared statistic is used to test whether each estimated coefficient is significantly different from zero (Maddala, 1992, pp. 120-124). The Wald chi-squared statistic can be computed by dividing the squared parameter estimate by its variance estimate. The Wald chi-squared statistic for testing an individual coefficient is distributed chi-square with one degree of freedom.

$$\chi^2 = \frac{\left(\overline{Q_m}\right)^2}{T_m} \qquad \textbf{(5)}$$

The test statistic for the overall significance of the nonlinear regression can be computed from the χ^2 statistics (-2 Log Likelihood for the contribution of the explanatory variables only) from the analysis conducted on each of the five implicates separately. The test statistic has an *F* distribution with k and (k + 1)*v*/2 degrees of freedom where k is equal to the number of independent variables excluding the intercept in the regression.

$$\hat{D}_m = \frac{\frac{\overline{d_m}}{k} - \frac{m-1}{m+1} r_m}{1 + r_m} \qquad \textbf{(6)}$$

$$where\ \overline{d_m} = \frac{\sum_{l=1}^{m} d_{*l}}{m} = average\ of\ the\ five\ \chi^2\ statistic$$

$$r_m = \frac{\left(1 + m^{-1}\right) Tr\left(B_m \overline{U}_m^{-1}\right)}{k} \qquad \textbf{(7)}$$

where Tr (A) is the sum of the diagonal elements in the kxk matrix A.

$$v = (m-1)\left(1 + r_m^{-1}\right)^2 \qquad \textbf{(8)}$$

SAS code for estimation of selected nonlinear models will be available at www.afcpe.org/nonr.htm

References

Kennickell, A. B. (1997). Codebook for 1995 Survey of Consumer Finances. Washington, D.C.: Board of Governors of the Federal Reserve System.

Kennickell, A. B., Starr-McCluer, M. & Sundén, A. E. (1997). Family finances in the U.S.: Recent evidence from the Survey of Consumer Finances. *Federal Reserve Bulletin, 83*(1), 1-24.

Maddala, G. S. (1992) *Introduction to Econometrics, Second Edition*. New York, NY: Macmillan Publishing Company.

Montalto, C. P. & Sung, J. (1996). Multiple imputation in the 1992 Survey of Consumer Finances. *Financial Counseling and Planning*, *7*, 133-46. [also available as a WWW document http://www.hec.ohio-state.edu/hanna/imput.htm]

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*(434), 473-89.

Yuh, Y., Montalto, C. P. & Hanna, S. (1998). Are Americans prepared for retirement? *Financial Counseling and Planning, 9*(1), 1-12.

Association for Financial Counseling and Planning Education
Annual Conference Program. Theme: *Securing Your Financial Future*
Fort Lauderdale Marriott North, Fort Lauderdale, FL November 18-21, 1998

Tentative Schedule of Sessions:

Wednesday, November 18

1:00-7:30 Registration
4:30-6:00 Opening General Session
Speaker: Dallas L. Salisbury, President & CEO, Employee Benefit Research Institute

Retirement Confidence, Investment Behavior, and Savings Education

Thursday, November 19

7:00-8:15 Continental Breakfast and Registration
8:15-9:45 General Session
Speaker: Gordon Sherman, Regional Commissioner of Social Security Administration at Atlanta

Social Security: Today, Tomorrow and Year 2032

9:30-5:00 Exhibits
10:00-11:30 Concurrent Sessions (4 choices)
- Session 1 *New Programs and Opportunities*
- Session 2 *Retirement Planning*
- Session 3 *Financial Planning Instruction*
- Session 4 *Bankruptcy and Counseling*

11:30-1:45 Luncheon and Business meeting
2:00-3:30 Concurrent Sessions (4 choices)
- Session 5 *EFT 99: Update*
- Session 6 *Credit and Mortgage Behavior*
- Session 7 *Housing and Investment Planning*
- Session 8 *Changes of Credit Counseling Industry*

3:30-3:45 Refreshment break
3:45-5:15 Concurrent Sessions (3 choices)
- Session 9 *Workplace Education and Media Coverage*
- Session 10 *Tax Planning and Financial Counseling*
- Session 11 *Building Partnerships to Increase Retirement Planning*

Friday, November 20

7:00-8:30 Continental Breakfast and Registration
7:30-3:00 Exhibits
8:30-10:00 General Session
Speaker: Richard Hinz, CFA, Director, Office of Policy and Analysis,
Pension and Welfare Benefits Administration, U. S. Department of Labor

The Private Pension System: Where Do We Go from Here?

10:30-12:00 Concurrent Sessions (4 choices)
- Session 12 *Alternative Counseling Approaches*
- Session 13 *Helping Low Income Families*
- Session 14 *Saving for Retirement*
- Session 15 *Investment Planning*

12:00-1:30 Awards Luncheon
1:45-3:15 Concurrent Sessions (4 choices)
- Session 16 *Helping Employees and Children*
- Session 17 *Serving Women and Self-employed*
- Session 18 *Serving Special Populations*
- Session 19 *Financial Management Strategies of Family Owned Businesses*

3:30-5:00 Refereed Posters

Saturday, November 21

8:30-10:00 Concurrent Sessions (2 choices)
- Session 20 *Workplace Financial Education: Issues and Answers*
- Session 21 *Collaboration showcase*

10:30-12:00 General Session
Speaker: Harold R. Evensky, CFP, Chair of CFP Board of Standard Board and Author of Wealth Management

Retirement Planning Issues in the Real World

For a detailed program and registration forms, including web links to some presenters, see
WWW.AFCPE.ORG
To request registration material, contact:
Sharon Burns, 6099 Riverside Drive # 100
Dublin, OH 43012-2004

614-791-6560
e-mail: request@afcpe.org
FAX: 614-798-6560

