Using SRMI for Non-response Adjustments in IRS Taxpayer Compliance Studies

Wei Liu, Karen Masken, Getaneh Yismaw

The views expressed in this paper are those of the authors and do not necessarily represent the views of the Internal Revenue Service.

1. Introduction

The Internal Revenue Service (IRS) has begun conducting annual individual taxpayer reporting compliance studies, the results of which are used in the strategic planning process, to develop workload selection formulas (i.e., which returns to audit), and to provide estimates of the amount of misreported tax and credits, including the Earned Income Tax Credit (EITC). To conduct these studies, taxpayers are randomly selected and their returns are audited to determine if there is any non-compliance. While the selected taxpayers probably do not view an audit of their tax return equivalent to responding to a survey, the non-response in these studies presents the same statistical issues as any other survey non-response. In earlier work (Masken, JSM 2004) the argument for treating non-response in these studies as missing at random was made, but did not come to a conclusion on the exact method that should be implemented. Continuing that effort, we first explore various imputation techniques that we considered and then detail the sequential regression multivariate imputation (SRMI) method that we implemented to estimate the amount of EITC that was overclaimed.

2. Missing Data

Missing data mechanisms are commonly described as falling into one of three categories (Little and Rubin, 1987):

Missing Completely at Random (MCAR)
This assumes that the missing cases are no different than the observed cases, and can be thought of as a random sample of the sample.

Missing at Random (MAR)
This assumes that the missing cases depend on certain observed characteristics, and can be fully represented by the observed cases with those same characteristics. Accounting for those observed characteristics which "cause" the missing data will produce unbiased results in an analysis.

Not Missing at Random (NMAR)
This assumes that the missing cases depend on variables which have not been measured or observed. NMAR is also termed "non-ignorable".

In all IRS compliance studies, there are instances of taxpayers who do not respond to the audit. The reasons for nonresponse can vary and include, but are not limited to, taxpayers the IRS was not able to locate, taxpayers in disaster areas, and taxpayers who do not show up for the audit. In the case of nonresponse, the IRS has data for all the reported items on the original tax return, but the amount allowed after exam is missing (hence the missingness is monotone). **Table 1** shows the distribution of the sample and the missing cases by strata. Since the missing cases are associated with the strata, we chose to treat them as MAR for this study

Table 1. Distribution of Sample and Missing Cases by Strata

	Percent of	
Strata	**Sample**	**Missing**
Schedule C Amount = $0, Married, EITC <= $412	5.3	3.0
Schedule C Amount = $0, Married, EITC <= $2747	13.1	5.0
Schedule C Amount = $0, Married, EITC > $2747	4.2	1.2
Schedule C Amount = $0, Unmarried, EITC <= $412	19.8	29.2
Schedule C Amount = $0, Unmarried, EITC <= $2747	27.8	32.7
Schedule C Amount = $0, Unmarried, EITC > $2747	8.5	10.0
Schedule C Amount < $0	1.9	0.9
Schedule C > $0, Married, EITC <= $412	1.3	0.0
Schedule C > $0, Married, EITC <= $2747	2.4	0.6
Schedule C > $0, Married, EITC > $2747	3.1	0.6
Schedule C > $0, Unmarried, EITC <= $412	4.6	2.4
Schedule C > $0, Unmarried, EITC <= $2747	5.0	7.7
Schedule C > $0, Unmarried, EITC > $2747	3.1	6.8

3. Multiple Imputation

General Overview

When the data are MAR, multiple imputation is an attractive solution to the missing data problem. Missing values for any variable are predicted based on the observed variables correlated with the missing data and causes of missingness. The predicted values, called “imputes”, are substituted for the missing values, producing a full data set. This process is performed multiple times, producing different plausible versions of missing data and multiple imputed data sets. For each imputed data set, a standard statistical analysis can be carried out and then results from each imputed data set can be combined to generate an overall estimate. This incorporates the uncertainty introduced by estimating the missing data.

One theoretical framework in support of imputation methods is the Bayesian theorem. An explicit model for variables with missing values can be specified, conditional on all the observed variables, some unknown parameters, and a prior distribution for the unknown parameters. Then a posterior predictive distribution of the missing values can be derived conditional on the observed values. The imputations are random draws from this posterior predictive distribution.

The performance of multiple imputation has been studied in many missing data applications. It has been shown to perform favorably and balance very well between quality of results and ease of use (Graham and Shafer, 1999; Schafer and Graham, 2002). First, multiple imputation can produce unbiased parameter estimates which reflect the uncertainty associated with estimating missing data. Multiple imputation can also provide adequate results even when the nonresponse rates are high and is shown to be robust to departures from the normality assumption. Sensitivity analysis has been shown to demonstrate that the effects of NMAR, an inaccessible mechanism, are often surprisingly minimal in the implementation of multiple imputation (Graham et al. 1997). Finally, compared with other statistical methods for missing data, for example, EM algorithm based on maximum likelihood estimation, multiple imputation is computationally simpler and less costly.

Auxiliary Variables

In selecting the variables used in the imputation models, we looked for variables that predicted both missingness and noncompliance. The purpose of including these variables is to make the missing data mechanism ignorable (Meng, 1994; Little and Raghunathan, 1997; Collins et al., 2001). For example, **Figure 1** demonstrates that taxpayers over 50 years old are more likely to be compliant with lower nonresponse rates and **Figure 2** shows head of household filers are more likely to be noncompliant with higher nonresponse rates.

Figure 1

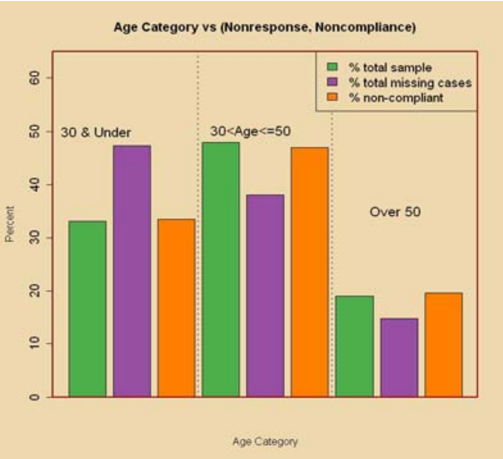
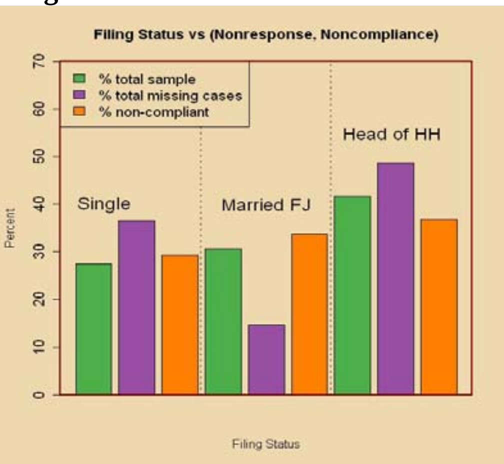


Figure 2



The variables we selected include demographic characteristics (age, gender), family structure (filing status, number of children), income (adjusted gross income, business income), geographic features (region, urbanicity), and tax filing characteristics (preparation method, use of bank product, number of years filing). These variables are of many types and therefore have a variety of distributional forms. For instance, income is continuous, age and filing status are categorical, and number of children is a count. Also, it is often the case that the EITC is disallowed entirely, thus the response variable has a semi-continuous distribution, making it very difficult to postulate a full Bayesian model. To address these issues, and to account for the complex sample design that the data come from (stratified random sample with unequal probabilities of selection), we found sequential regression multivariate imputation (SRMI) to be a good solution.

Sequential Regression Multivariate Imputation

The SRMI approach (Raghunathan et al., 2001) can handle a variety of complex data structures involving many types of variables and bounds, where an explicit full Bayesian model cannot be formulated. It considers imputation on a variable by variable basis and creates imputations through a sequence of multiple regressions. The type of regression model depends on the type of variable imputed. The sequence of imputing missing values is continued in a cyclical manner, each time overwriting previously imputed values. It can build interdependence among imputed values and exploit the correlation among covariates. The SRMI procedure can be easily implemented by the module IMPUTE in a SAS-callable software package IVEware (http://www/isr.umich.edu/src/smp/ive) by Raghunathan et al as well as an R package, MICE (Multivariate Imputation by Chained Equations) by Buuren and Groothuis –Oudshoorn.

The amount of EITC a taxpayer is allowed is determined by their correct filing status, correct income, and the number of children they have that qualify. Rather than impute the missing values for the correct amount of EITC directly, we initially planned to impute the missing values for the input variables and then calculate the amount of EITC allowed. However, we were unable to find good models for either the filing status or the number of qualifying children. Since earlier studies have shown that these are both highly correlated with noncompliance, we abandoned this plan.[1]

For this study, we were interested only in the portion of EITC that was overclaimed. Therefore, we created a new variable called 'overclaim' for the completed cases and set it equal to 0 if the initial EITC claim by the taxpayer was less than or equal to the EITC amount allowed after exam. Otherwise, it was the difference of the allowed amount from the claimed amount of EITC. This new overclaim variable is the variable we imputed.

While including the design variables and the interaction terms between design variables and other predictors is one way to incorporate the sample design into the imputation model, we found that we could not fit all the interaction terms into one model. Also the IVEware module IMPUTE can be used only to impute missing values in a simple random sample. Therefore, we decided to develop separate imputation models for each stratum (some strata were collapsed due to small sample sizes). The models were developed using the SAS procedure SURVEYREG and incorporated as much information as possible given the varying sample sizes within each stratum. Once the models were developed, they were input into the IVEware IMPUTE module.

Each imputation process for the overclaim amount in a stratum was done in rounds. In the first round, the imputed value is based on only the completed cases. In the next round, the imputed value is based on both the completed cases and the imputed amounts for the missing cases from the previous round's imputation, with each round overwriting the previous round's imputation. In all, we did ten rounds for each imputation.

Since the overclaim amount is a semi-continuous variable, a two-stage model was fit. In the first stage, logistic regression was used to impute zero/non-zero status for the overclaim amount. If the first stage resulted in non-zero status, then a normal linear regression model was used to impute a positive amount for the overclaim.

The procedure described above was applied sixty times to generate sixty data sets with the final imputed values of the overclaim for each stratum. Then the strata were combined to form sixty full-sample data sets.

[1] *Compliance Estimates for Earned Income Tax Credit Claimed on 1999 Returns,* Department of the Treasury Internal Revenue Service, February 28, 2002.

4. Analyzing Multiply Imputed Data

Once the sixty imputed data sets were obtained, a standard statistical analysis on EITC overclaims was carried out for each data set. The mean and variance of the amount of EITC overclaimed was calculated for each imputed data set and then combined to produce an overall estimate of the mean EITC overclaim and its variance. The formulas below were used to combine the estimates (Rubin 1987a, Chapter 3).

Let $\hat{Q}_l$ be the point estimate of the mean of Y and U_l be its estimated variance from the l^{th} imputed data set, where $l = 1,2,...60$. The combined overall point estimate of the mean of Y is

(1) $$\overline{Q}_m = \sum_{l=1}^{m} \hat{Q}_l / m ,$$

Where $m = 60$. It is the average of the 60 point estimates of the mean of Y .

The overall estimated variance of $\overline{Q}_m$ is

$$T_m = \overline{U}_m + (1 + m^{-1}) B_m$$

(2) where

$$\overline{U}_m = \sum_{i=1}^{m} U_l / m, and$$

$$B_m = \sum_{i=1}^{m} (\hat{Q}_l - \overline{Q}_m)'(\hat{Q}_l - \overline{Q}_m)/(m-1)$$

$\overline{U}_m$ is the average of the 60 variance estimates from the imputed data sets. It measures the original variability in overclaims, hence, it is called “within-imputation” variance. This component can be regarded as the variance estimate we would have generated if we had taken the imputes as the observed values and there had not been any missing data.

B_m is called “between-imputation” variance. It measures the uncertainty caused by estimating missing data, i.e., imputation. This uncertainty is low when the point estimates of the mean of Y are quite similar across different imputed data sets. Finally, $(1 + m^{-1})$ is the correction factor when m is small.

Many software packages are available to implement the above combining rules to analyze the multiply imputed data sets. After we obtained the multiple imputed datasets through the module IMPUTE in IVEware, we used another module DESCRIBE in IVEware to get the combined estimates on the mean value of EITC overclaims. The module DESCRIBE in IVEware supports complex survey design by incorporating design features such as strata and weights and using Taylor series linearization to estimate variances in the analysis of each completed data set. Unfortunately, at the time of this writing, we are not yet able to release the results publicly.

Evaluation of Multiple Imputation Method

Since there was one stratum with no missing data, we decided to use it to evaluate the performance of our multiple imputation method. The logistic regression model of missingness from a similar stratum was imposed to set fifteen percent of the complete stratum cases to missing. We then applied the SRMI procedure described above and compared the results of the original complete data to the imputed data for this stratum. Our results showed no statistically significant difference between the complete and imputed data in the mean EITC overclaimed at the 95% confidence level.

5. Conclusions

The analysis of missing cases in our study supports the MAR mechanism, therefore multiple imputation is an attractive solution to our missing data problem. Since there are many types of predictors in our study and the response variable is semi-continuous, SRMI was an appealing solution. We found the imputation models in our study to be robust to the specific models used and also to the type of software used. We plan to conduct future research to continue improving our imputation models.

References

Allison, P.D. (2001). *Missing Data*. Thousand Oaks: Sage.

Barnard, J., Rubin, D.B. and Schenker, N. (1998). Multiple imputation methods. In: P. Armitage and T. Colton (Eds.), *Encyclopedia of Biostatistics* (pp. 2772-2780). New York: Wiley.

Brick, M. Review of "Multiple imputation for nonresponse in surveys" (Auth D.B. Rubin). *Metrika*, *36*, 249-250.

Ezzati-Rice, T.M., Khare, M., Rubin, D.B., Little, R.J.A., Schafer, J.L. (1993). A Comparison of Imputation Techniques in the Third National Health and Nutrition Examination Survey. *Proceedings of the Survey Research Methods Section of the American Statistical Association 1993*, 303-308.

Graham, J.W., Hofer, S.M., Donaldson, S.I., Mackinnon, D.P. and Schafer, J.L. (1997). Analysis with missing data in prevention research. In K. Bryant, M. Windle and S. West (Eds.), The science of prevention: Methodological advances from alcohol and substance abuse research, 325-366. Washington, D.C.: American Psychological Association.

Little, R.J.A. (1988). Missing Data in Large Surveys. *Journal of Business and Economic Statistics*, *6*, 2, 287-301 (with discussion).

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

Meng, X.L. (1990). *Towards Complete Results for Some Incomplete-Data Problems*. Ph.D. Thesis, Department of Statistics, Harvard University, Cambridge MA.

Raghunathan, T.E., Solenberger, P., van Hoewyk, J. (2000). IVEware: Imputation and Variance Estimation Software: Installation Instructions and User Guide. Survey Research Center, Institute of Social Research, University of Michigan.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 85-95.

Raghunathan, T.E. and Paulin, G.D. (1998). Multiple imputation in the Consumer Expenditure Survey: Evaluation of statistical inference. *Proceedings of the Business and Economics Section of the American Statistical Association*, 1-10.

Rubin, D.B. (1978). Multiple Imputations in Sample Surveys -- A Phenomenological Bayesian Approach to Nonresponse. *The Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20-34.

Rubin, D.B. (1986). Basic ideas of multiple imputation for nonresponse. *Survey Methodology*, *12*, 37-47.

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

Rubin, D.B. (1988). An Overview of Multiple Imputation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 79-84.

Schafer, J.L., Ezzatti-Rice, T.M., Johnson, W. Khare, M., Little, R.J.A. and Rubin, D.B. (1996). The NHANES III multiple imputation project. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 28-27.

Schafer J.L., Olsen, M.K. (1999). Modeling and imputation of semicontinuous survey variables. Federal Committee on Statistical Methodology Research Conference: Complete Proceedings, 2000.

Schafer JL, Graham JW (2002). Missing data: our view of the state of the art. *Psychological Methods*;7(2):147-77.