Imputing the Missing Y's: Implications for Survey Producers and Survey Users

Rebekah Young[1] and David R. Johnson[2]

[1]The Pennsylvania State University, 211 Oswald Tower, University Park, PA 16801
[2]The Pennsylvania State University, 211 Oswald Tower, University Park, PA 16801

Abstract

Multiple imputation is a popular technique used to handle item-level missing data. Recent studies, however, have generated serious concerns about the best practices for statistical analysis with an imputed dependent variable. We use an example from observed data to examine three multiple imputation strategies: (1) excluding the dependent variable from the imputation model, (2) multiple imputation then deletion, and (3) including the dependent variable in the imputation model and retaining the imputed values in the subsequent analysis. Consistent with previous research, our results suggest that the dependent variable should be included in the imputation model. Under conditions where it is most practical to do so, survey users may be able to retain the imputed values in their analysis, provided that a sufficient number of datasets was generated.

Key Words: Multiple imputation, dependent variable, survey data, missing data, planned missing designs

1. Introduction

Survey practitioners have an increasing number of tools for handling item-level missing data in an unbiased manner. Modern methods have increased the utility of planned missing designs, which can increase the efficiency of surveys. Modern imputation techniques have motivated the release of datasets for public archive in which all missing values have been imputed with the goal of facilitating analysis of the data. Data missing on the dependent variable presents a special dilemma for survey producers and users. In this paper we use an example from observed data to examine three different multiple imputation strategies when data are missing on the dependent variable: (1) excluding the dependent variable from the imputation model, (2) multiple imputation then deletion, and (3) including the dependent variable in the imputation model and retaining the imputed values in the subsequent analysis.

2. Background

Multiple imputation (MI) is a popular method for parameter estimation with missing data (Graham 2009; Rubin 1996). Three basic steps characterize a MI procedure (Little and Rubin 2002; Rubin 1987). First, *m* number of replicate datasets are created and the missing values in each dataset are separately filled in with plausible random values drawn from the conditional distribution given the observed data. Second, each analysis model is estimated separately in each dataset. Third, the estimates are pooled using Rubin's (1987) rules to yield coefficients and standard errors that reflect the uncertainty about the missing values. Multiple imputation is widely regarded as both an unbiased and efficient modern method for the treatment of missing data

(Allison 2001; Schafer and Graham 2002). This approach has been implemented in a variety of popular software packages such as SPSS, SAS, R and Stata (Graham 2009).

For survey producers, widespread access to MI software implementations has several practical consequences. First, the availability of MI has increased the utility of planned missing data designs. Such an approach may make data collection less costly, more efficient, and reduce respondent burden. When some questions are intentionally not asked of some respondents (planned missing data), the data are likely to be missing completely at random (MCAR) or at least missing at random (MAR) (Schafer and Graham 2002). Under this assumption, MI can be used to fill in the missing values. Second, MI has motivated the release of datasets for public archive in which all missing values have been imputed. Although providing multiply imputed data sets for large public release data surveys has not yet seen extensive use, the increase in economical data storage and faster computers has made it a practical approach.

For survey users, imputation by survey producers is an attractive option for the treatment of missing values in a dataset. Most dataset users are focused on their substantive scientific analysis and view missing data as a nuisance (Rubin 1996). With pre-imputed data sets the researcher, using any of the widely available packages which allow statistical analysis with multiply imputed datasets, can basically proceed with their substantive analysis without having to fiddle with the imputation process. This would be an "impute it and forget it" model. In addition to simplicity, another advantage of pre-imputed data is that multiple users can conduct analysis with the same exact data, giving consistent results among analyses.

Although there are many advantages of survey producers releasing multiply imputed datasets for public archive, several statistical limitations may stand in the way of fuller utilization of such standard imputed datasets. These include practical difficulties of imputing a large number of variables in a single imputation model, the possibility that the results could be biased if the variables used in the analysis model (such as interactions and polynomial terms) were not included in the imputation model, and the use of data where the dependent variable was imputed may yield biased and inefficient results. We focus here on the last concern.

3. Imputing the Missing Y's

Within the MI framework, there are two common strategies for handling missing data on the dependent variable (DV), with a third method recently proposed. The first technique for handling missing data on the DV is to exclude all cases with missing data on the DV from the imputation model and from the analysis. This method uses complete case analysis for the DV, but MI for all of the independent variables. The popularity of this method likely stems from a widely held belief by many researchers that they are doing something "wrong" by imputing the DV.

Researchers are sometimes reluctant to impute values on the dependent variables because they believe that doing so would be treating unknown outcomes as though they were known. There is even some evidence that, under special circumstances, excluding cases missing on the DV and imputing the DV lead to equivalent results. If the missing data are MCAR, or if there are no missing data on any of the independent variables and no strongly correlated auxiliary predictors, MI cannot improve upon complete case analysis (Allison 2001). Routinely dropping all cases with missing values on any of the DVs, however, is a problematic strategy. Additionally, some statistical analyses, such as path analysis, may treat some variables as independent in one equation and dependent in another. In this situation, excluding cases with missing values on variables ever treated as outcomes may result in substantial loss in sample size, as well as

possible selection bias. In general, it is not safe for researchers to simply ignore values missing on the dependent variable.

Despite considerable worry over imputing the DV, the truth is that missing values on the DV and missing values on the independent variables do not fundamentally differ. An imputation model does not represent causal relationships among the data. Rather, the model is a device to preserve important features of the observed data (means, covariances) in the imputed values. In the MI process, all variables in the imputation model are treated as multivariate response. If the dependent variable is omitted from the imputation model, then the correlation between the dependent variable and any of the independent variables is assumed to be zero (Graham 2009). This assumption will systematically bias coefficients downward (Little and Rubin 2002; Graham 2009). One of the important standards of MI, therefore, is that every variable to be included in the analysis model should also be included in the imputation model, including the DV (Schafer and Graham 2002).

While it is clear that distinctions between dependent and independent variables should be left to post-imputation analysis, it is less clear what to do with imputed values during analysis. The second common technique for handling missing data on the DV is to impute the DV and retain the imputed values in the analysis. The MI model should contain all of the variables that will be used in a subsequent analysis model, a condition clearly satisfied by this approach. An advantage of this method is that in subsequent analysis, any variable could be treated as an independent variable or as a dependent variable without changing the number of cases in the dataset used.

The third, recently proposed technique for handling missing data on the DV involves imputing all variables, including the DV, but then deleting the cases with imputed values on the DV prior to analysis. This technique, proposed by von Hippel (2007), is a multiple imputation, then deletion (MID) method. In the analysis phase, the imputed values on the DV may simply be adding useless noise and unnecessarily inflating the standard errors (von Hipple 2007). When only a small number of imputed datasets are generated (e.g., less than 5), or when DVs have large amounts of missingness, this issue may be particularly salient. With more commonly observed levels of missingness, such as five to 15 percent, the MID method may not offer a discernable advantage, which we examine in this paper.

If retaining cases in the analysis with imputed DVs proves to be a serious problem then the simple “impute it and forget it” approach fails. In this case, the researcher would need to take the missingness into account, at least to the point of deleting cases with imputed values on the DV. Of course, this would require that the survey producers release imputed datasets with indicators for all imputed values so that survey users are able to remove the imputed values for variables treated as an outcome. Many imputed, publicly released, datasets do not have this feature and it is important to question how necessary this added step is. To date, the MID method has only been tested in a series of simulations that may not reflect the complexity of how the approach fares in the analysis of observed survey data. In this paper we compare three approaches to handling data missing on a DV in an empirical data set that more closely approximates the conditions a survey user will normally encounter.

4. Data and Method

To compare different approaches to MI when values are missing on the DV, we have selected to use as an example a regression model of predictors of marital happiness. The data were taken

from the first wave of the National Survey of Families and Households (NSFH) (Bumpass and Sweet 1987). The dependent variable, marital happiness, was asked of all married respondents with the question: "Taking things all together, how would you describe your marriage?" Responses were recorded on a scale of *1* (very unhappy) to *7 (*very happy). Independent variables were selected that were expected to predict marital happiness, that spanned a broad range of the proportion of the values in the variable that were missing, and that varied by level of measurement. Descriptive information, presented in Table 1, shows the variety of variables that were included.

Table 1. Descriptive Statistics for Example Model Variables

Variables	*Mean*	*Std. Dev.*	*Min.*	*Max.*	*% Missing*
Marital happiness	6.0	1.4	1.0	7.0	11.4
Marital duration	18.5	15.2	0.1	61.6	4.8
Log of household income	4.4	0.7	0.0	5.4	27.0
Years of education, centered	0.3	3.2	-12.0	8.0	1.2
Female = 1; male = 0	0.6	0.5	0.0	1.0	0.0
Religious fundamentalist = 1	2.8	1.2	1.0	5.0	21.0
Does the wife work? Yes = 1	0.7	0.5	0.0	1.0	1.0
Depression scale (Ces-D)	12.9	15.4	0.0	84.0	12.8

Our goal of this analysis was to both represent realistic data circumstances encountered by researchers and also a situation that was extreme enough that the choice of method could reasonably be expected to make a difference. To achieve the latter, we took a random subsample of the larger dataset to obtain a sample size of 500 cases. We also increased the overall observed missingness from around 40 percent to approximately 50 percent in a way that preserved the observed missing data structure. We divided the sample into two parts; the first part included only those cases with no missing data on any of the variables in Table 1, and the second part included cases with one or more missing values on any of these variables. We then used bootstrap sampling with replacement to draw a disproportionate random sample that doubled the number of cases with a least one missing value, and then randomly reduced the number of cases with no missing values to bring the sample size to 500.

The percentage of missingness on each independent variable ranged from none to 27 percent. Some of the independent variables had minimal missing observations (e.g., years of education, gender, number of children) in this dataset. The percent missing was highest for total household income (27%) and for a variable indicating the degree to which the respondent considers himself or herself to be a religious fundamentalist (21%). Relatively high nonresponse rates to income and religion questions are common in survey data, and likely result from the respondent's perceived sensitivity of the questions (Converse 1976; Faulkenberry and Mason 1978; Riphahn and Serfling 2002). The depression scale (CESD) is a summated scale of 10 items from the CESD that were included in the NSFH. For our purposes here, because we wanted to maximize the observed missingness, we constructed the scale so that it was set as missing if any single item was missing. Alternative methods are available for handling missing values in scale item, several of which lead to a lower percent missing (Schafer and Graham 2002). The missingness on the dependent variable was 11.4%. This is lower than the conditions of 20% and 50% missing that

were tested by von Hipple (2007). At the same time, 11% may be a much higher percent missing than many researchers typically encounter.

5. Results

Three techniques for handling missing data on the DV are compared in an OLS regression model predicting marital happiness. As shown in Table 2, the differences in the overall substantive conclusions drawn from the model results would not differ substantially, regardless of the DV imputation strategy. The majority of the coefficients were not significant ($p<.05$) predictors of marital happiness, regardless of imputation strategy. The two coefficients significant at the .05 level, whether or not the wife works outside the home and level of depression, had similar magnitudes regardless of strategy. The coefficient that experienced the most difference in magnitude by technique was the coefficient for total household income (ln), which ranged from -.04 to -.17. Consistent with previous literature, this suggests that excluding the DV from the model appears to bias the coefficients towards zero and that strategies for handling missing data are more consequential as larger amounts of missing data are observed (Allison 2001; Schafer and Graham 2002).

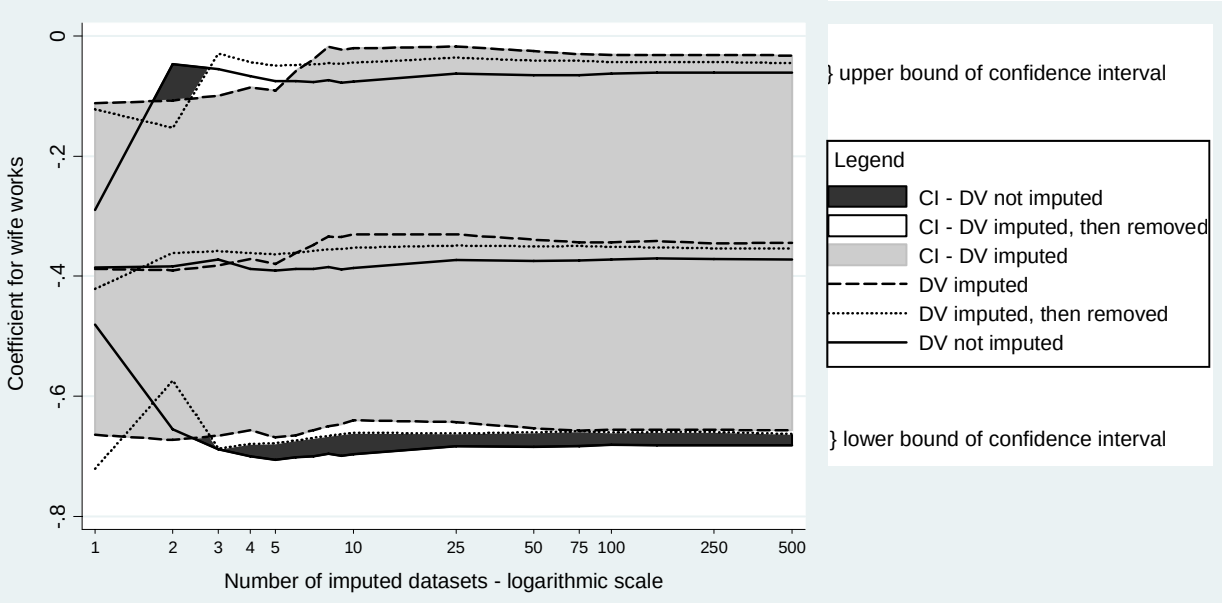
Table 2. Comparison of Three Techniques for Treatment of Data Missing on the Dependent Variable in an OLS Model Predicting Marital Happiness

	DV not imputed			DV imputed, then removed (MID)			DV imputed and included in analysis		
Variables	*B*	*SE (B)*	*p > t*	*B*	*SE (B)*	*p > t*	*B*	*SE (B)*	*p > t*
Marital duration	0.01	0.00	0.84	0.01	0.00	0.90	0.01	0.00	0.73
Log of household income	-0.04	0.14	0.77	-0.17	0.13	0.17	-0.15	0.14	0.30
Years of education, centered	-0.01	0.02	0.53	-0.01	0.02	0.58	-0.01	0.02	0.64
Female = 1; male = 0	0.19	0.13	0.15	0.20	0.13	0.11	0.20	0.13	0.12
Religious fundamentalist = 1	0.09	0.06	0.11	0.09	0.06	0.12	0.09	0.06	0.12
Does the wife work? Yes = 1	-0.39	0.16	0.02	-0.35	0.16	0.03	-0.33	0.16	0.04
Depression scale (Ces-D)	-0.02	0.00	0.00	-0.03	0.00	0.00	-0.03	0.01	0.00
Constant	6.42	0.71	0.00	7.03	0.63	0.00	6.85	0.72	0.00
Sample size (n)	*443*			*443*			*500*		
Imputed datasets (m)	*10*			*10*			*10*		

Perhaps the most important question for researchers is whether or not one approach to handling missing data on the dependent variable offers more efficiency than another. The simulations performed by von Hipple (2007) suggest that as the number of imputed datasets increases, the difference between imputing the dependent variable and retaining it for analysis versus the MID method will become trivial. We explored this with observed data by imputing one to 500 datasets and comparing the results. We examined the behavior of the confidence interval around the coefficient for whether or not the wife was employed in the formal labor market, shown in Figure 1. The y-axis in Figure 1 shows the size of the coefficient for wife working. The x-axis shows the log of the number of datasets generated. By imputation method, the top three lines on the graph

show the upper bound confidence interval, the middle three lines show the actual coefficients, and the bottom three lines show the lower bound of the confidence interval. All coefficients were significant at $p < .05$ in all models.

Table 3. Coefficient Behavior and 95 Percent Confidence Intervals Among Three Techniques for Imputing the Dependent Variable



As shown in Figure 1, the confidence intervals for the estimates are nearly identical once more than five datasets are generated. In the recent past, working with a small number of datasets may have been necessary due to limitations in computing power. Rubin's (1987) early work on multiple imputation, for example, involved the use of only three datasets and he envisioned the use of perhaps as many as five datasets in the future. Today, generating a larger number of datasets is a relatively trivial issue regarding computing time and as many as 20 to 100 datasets are recommended for use in practical applications (Schafer and Graham 2002). Based on our results, we agree.

6. Discussion

An advantage of our method is that we used actually observed data to explore some of the findings from the simulation literature regarding three MI strategies for handling data missing on the DV. We recognize, however, that using observed data is also a limitation because it is only one example and not an empirical test of a method. We do not suggest that our paper provides evidence that the MID method is unnecessary. We merely suggest that in at least in some real world situations, it may be acceptable for values multiply imputed on the DV to be retained in the analysis. One implication of the MID method is that when releasing an imputed dataset for public use, survey producers should include indicators for all imputed values which allow survey users to remove the imputed values for the DVs of interest. Our results suggest that this distinction may not be a crucial step, provided that a sufficient number of datasets has been generated.

References

Allison, Paul D. 2001. *Missing Data*. Thousand Oaks: Sage Publications.

Bumpass, Larry L., and James A. Sweet. 1987. *National Survey of Families and Households, 1987-1988*. Madison, WI: University of Wisconsin, Center for Demography and Ecology.

Converse, Jean M. 1976. "Predicting no opinion in the pools." *Public Opinion Quarterly* 40:515-530.

Faulkenberry, G. David, and Robert Mason. 1978. "Characteristics of nonopinion and no opinion response groups." *Public Opinion Quarterly* 42 (4):533-543.

Graham, John W. 2009. "Missing Data Analysis: Making it Work in the Real World." *Annual Review of Psychology* 60:549-576.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data, Second Edition*. New Jersey: John Wiley and Sons.

Riphahn, Regina T, and Oliver Serfling. 2002. Title. Switzerland: University of Basel.

Rubin, Donald B. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association* 91 (434):473-489.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Schafer, Joseph L, and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2):147-177.

von Hippel, Paul T. 2007. "Regression With Missing Y's: An Improved Strategy for Analyzing Multiply Imputed Data." *Sociological Methodology* 37:83-117.