Title
Sensitivity of Econometric Estimates to Item Non-response Adjustment

Author
Sanchez, Juana

Publication Date
2016-10-06

Sensitivity of Econometric Estimates to Item Non-response Adjustment [1]

Juana Sanchez [2]

Abstract

Non-response in establishment surveys is a very important problem that can bias results of statistical analysis. The bias can be considerable when the survey data is used to do multivariate analysis that involve several variables with different response rates, which can reduce the effective sample size considerably. Fixing the non-response, however, could potentially cause other econometric problems. This paper uses an operational approach to analyze the sensitivity of results of multivariate analysis to multiple imputation procedures applied to the U.S. Census Bureau/NSF's Business Research and Development and Innovation Survey (BRDIS) to address item non-response. Multiple imputation is first applied using data from all survey units and periods for which there is data, presenting scenario 1. A scenario 2 involves separate imputation for units that have participated in the survey only once and those that repeat. Scenario 3 involves no imputation. Sensitivity analysis is done by comparing the model estimates and their standard errors, and measures of the additional uncertainty created by the imputation procedure. In all cases, unit non-response is addressed by using the adjusted weights that accompany BRDIS micro data. The results suggest that substantial benefit may be derived from multiple imputation, not only because it helps provide more accurate measures of the uncertainty due to item non-response but also because it provides alternative estimates of effect sizes and population totals.

Key Words: Item non-response, unit non-response, business establishment survey data, statistical models, multiple imputation, BRDIS, LBD, multivariate analysis, standard errors, non-sampling errors.

1. Introduction

Researchers seeking understanding of business innovation in the United States have the option of using the Business Research and Development and Innovation Survey (BRDIS) microdata (NSF, 2016). If BRDIS suffered only from sampling error, multivariate statistical methods that control for the survey design effect could be used, and business innovation research would be straightforward using procedures that exist now in many software packages such as SAS, Stata, SPSS and R, among others. But BRDIS is a 5-section, large survey. As is the case with many surveys with complex design, it suffers from non-sampling errors due to unit and item non-response. Appraising the extent of these non-sampling errors and their effect on total error is not easy for any survey (Mason et al., 2002), but multiple imputation (MI) used in this paper offers an alternative that helps shed some light on those effects. In this paper, we take an operational approach to the problem. Comparisons are made among different methods of making use of BRDIS' responders data to estimate the same statistical, microeconometric, model of the determinants of R&D, and population total R&D.

[1] Format for referencing this paper: Sanchez, J. (2016). "Sensitivity of Econometric Estimates to Item Non-response," Proceedings of the Fifth International Conference of Establishment Surveys, June 20-23, 2016, Geneva, Switzerland: American Statistical Association. (To appear at http://www.amstat.org/ASA/Meetings/International-Conference-on-Establishment-Surveys.aspx

[2] Juana Sanchez, University of California Los Angeles, Department of Statistics, 8125 Math Sciences Building, Box 951554, Los Angeles, CA 90095, USA, email: jsanchez@stat.ucla.edu The research in this paper was conducted while the author was Special Sworn Status researcher of the U.S. Census Bureau at the Center for Economic Studies. Research results and conclusions expressed are those of the author and do not necessarily reflect the views of the Census Bureau. This paper has been screened to insure that no confidential data are revealed.

The paper compares the estimates of the model parameters and their standard errors under different multiple imputation scenarios (with or without the survey‘s adjusted weights) and the traditional direct analysis of incomplete data practiced by most researchers. Stata' MICE and SVY are the software used (StataCorp, 2015) . The analysis is restricted to variables and years that are expected to report data for all the items used. There are approximately 25% unit responders in the sample that have item non-response in at least one of the 4 variables of interest, resulting in a 76% item response rate. Multiple imputation accounts for the uncertainty due to that item non-response and imputation. That additional uncertainty is reflected in the standard errors obtained with multiple imputation procedures, which indicate the percentage increase in total error due to non-sampling error caused by item non-response.

Unit non-response is addressed in this paper by using the data producers‘ adjusted sampling weights. Univariate estimates of total R&D published by the NSF account for unit non-response that way. Not using adjusted weights would produce very different, much smaller, estimates of total R&D. Model coefficients and standard errors are also different if they are not used. The results presented in this paper indicate that standard errors increase when multiple imputation is done (but less so if the adjusted weights are used), and the estimated effects and their direction is not affected much by multiple imputation (when the adjusted weights are used).

MI methodologies that preserve the underlying relationships among variables and the distributional properties of the data have been proposed for other government surveys when multivariate analysis is anticipated. NASS's ARMS survey imputation using a customized version of MI called Iterative Sequential Regression (ISR) is an example at hand, with good results (Robbins et al., 2013; Miller et al., 2016) But the use of multiple imputation in business survey data obtained via complex designs is not as prevalent as it is in consumer surveys, and has not been studied much. Thus the researcher faces challenges that have already been overcome by consumer surveys or other nonbusiness establishment data. Luckily, some multiple imputation methodologies are easily implemented in major commercial packages such as SAS, Stata, SPSS, R and others (Horton and Lipsits, 2001; White et al., 2011). Numerous areas of research have used them (Schaffer, 1999; Little and Rubin, 1987; Little, 1988).

This paper takes BRDIS‘ survey design characteristics into account both in the multiple imputation and the estimation of the multivariate model. The strata and the survey weights are used as auxiliary variables in the imputation and as information in the estimation. Everything that is known about the survey and the data is used. Results are then compared to those obtained when that information is not used.

The organization of this paper is as follows. Section 2 of this paper introduces the data and the motivation for addressing item non-response. Section 3 then explains how multiple imputation was conducted, and the rest of the methodology used. This is followed by section 4, where the population estimates of total R&D and the statistical model parameter estimates are presented. The last section contains the conclusions and further remarks.

2. BRDIS and LBD

This paper uses BRDIS data linked with the Longitudinal Business Database (LBD) (NSF, 2016; Jarmin and Miranda, 2012). BRDIS is survey data, LBD is administrative data. Linking was done separately for multi and single business units using appropriate identifiers. Not only does LBD contribute auxiliary variables for the multiple imputation but it also plays a very important role in the multivariate study of R&D, as it provides exogenous business metrics not requested in BRDIS. The rest of this section describes in more detail

the data used in this paper.

Table 1: Summary statistics and model inclusion for variables appearing in the regression models or imputation models. Unweighted. Source: BRDIS and LBD 2009-2013.

Var name	Mean	Std	ProbM	ProbRD	InRM	InIM
Domestic R&D employee count	39	457			y	y
R&D performed paid by others	2706	87543			y	y
R&D EXPENSE	11002	168517			y	y
Multi unit	0.32	0.5		y	y	y
Number of states	2.7	6			y	y
Number of legal forms	1.10	0.4				y
Number of NAICS	2	3.6		y	y	y
Total Employment	966	11126			y	y
Annual payroll (in $1000)	60768	544263		y	y	y
R&D establishments	0.14	3			y	y
Age of oldest est	22	12	y	y	y	y
Years in BRDIS	2.5	2				
Industry				y	y	y
Stratum			y	y		y
Sampling weight						y
Survey form			y	y		y
Year						y

2.1 BRDIS

BRDIS is an annual survey of about 40,000 for-profit non agricultural companies with at least 5 employees (NSF, 2016). It is administered annually by the NSF and the U.S. Census Bureau, and it replaced the Survey of Industrial Research and Development in 2008. The data user community is broad, and includes the U.S. Bureau of Economic Analysis, businesses, the National Science Foundation and academic researchers.

The survey is the primary source of information on U.S. business R&D expenditures, R&D work force, R&D management and intellectual property. The innovative capacity of the U.S. is measured almost exclusively by companies‘ R&D outlays (somewhat over 2.8% of GDP in the U.S.A. in 2010) because R&D expenditures activities, when accumulated, are believed to create the stock of knowledge (Cohen, 2010). Hence BRDIS is a very important economic survey that adds to the U.S. National accounts. Response is mandatory and confidential under Title 13 U.S. Code, thus the unit response rate is high. In 2008 the overall unit response rate was 77.4%, and the unit response rate for the top 500 domestic R&D performing companies was 92.6% (Wolfe, 2010). The survey frame is extracted from the U.S. Business Register (BR).

There is one sampling unit (SU) per enterprise, covering all the establishments under common ownership or control that operate in the US (60% single unit and 30% multiple unit). The SU is assigned to U.S. NAICS 2007 industry group in which it has the largest proportion of sales relative to R&D Measure of Size (MOS), although allocation to industry group is later recoded based on business codes reported by the company in BRDIS. For

a given company with more than one establishment, the prior year's annual payroll and employment data for its active establishments are summed to the company level. Multiunit employment and payroll imputation in the BR are done before sampling for BRDIS. In multiunit companies, if a establishment was not in the 50 states or the district of Columbia the establishment was treated as if it did not exist. Measure of size does not include global R&D in BRDIS. For example, multinational corporations in the BRDIS sample frame are assigned industry codes based on their local operations only (their global operations or outsourcing are not taken into account).

All of the above considerations result in a sampling frame of BRDIS that consists of three major strata: (a) known positive R&D in the last 5 years (two treatments). These are companies who are known from prior BRDIS surveys or other sources to have known R&D , and with measure of size their most recently reported domestic R&D; (b) Known zero R&D in the last 5 years. These are companies known from previous R&D surveys or other sources to have zero domestic R&D; (c) Unknown R&D or unkwnown group (two treatments), which consists of companies about which nothing is known of their R&D expenditures. Relatively speaking, the largest group is stratum (c), followed by stratum (a) and (b).

After allocating to strata based on MOS, the companies are allocated to about 60 business strata correponding to 60 industry groups.

Although the model estimates presented later in this paper account for the main strata, the 60 business strata are not accounted in the Stata SVY methodology employed in this paper.

Business surveys in the United States are usually not integrated. BRDIS, like most business establishment surveys, collects from companies hard data for which records are available, but it does not collect all the information that would be relevan to a multivariate analysis of the relation between R&D and company characteristics. After all, the survey is not done to help researchers, but to obtain univariate population descriptive summaries. It is because of this that active cases of BRDIS must be linked to active LBD establishments to obtain auxiliary variables not provided by BRDIS.

2.2 LBD

The LBD is a longitudinal census of business establishments and companies in the U.S. with paid employees. LBD is comprised of survey and administrative records. The LBD covers all industries and all U.S. States (Jarmin and Miranda, 2012). The multiunit nature of the company, the legal form of organization, the age of the company, the number of establishments, the states where the company operates, the zip codes where it has establishments,the number of research establishments, payroll and employment, all are variables measured by LBD. The multivariate analysis of this paper seeks to determine the effects of those on BRDIS‘ R&D.

After linkage with BRDIS, establishment data was compressed to obtain one company aggregate record.

2.3 BRDIS and LBD variables used

Table 1 displays the variables included in the imputation and the regression models that will be described in the section 3. The table displays the mean (Mean) and standard deviation (Std) of the economic variables listed, without weighing or imputation. The other columns of Table 1 can be explained as follows. Logistic regression analysis was done to determine which of the economic variables affect the probability of missingness of R&D (column ProbM), and which of them relate to the probability of being an R&D performer ProbRD).

A letter "y" indicates effect. The last two columns of the table indicate whether the variable was included in the R&D statistical model (InRM) and/or the imputation model (InIM).

As can be seen in Table 1, the probability of missingness in R&D (in column ProbM) is significantly related to the age of the oldest establishment in the company, the stratum, and the survey instrument type. The probability of spenditures on R&D (column ProbRD) is associated significantly with the number of NAICS in which the company operates, whether the company is multiunit or single unit, the survey instrument received, the age of the oldest establishment, the stratum, payroll, and industry.

It must be pointed out that before imputation, the highest correlations of R&D expense is with count (0.11), R&D employees (0.39), R&D performed (0.18), Number of states (0.21), number of NAICS (0.23), total payroll (0.38), R&D establishments (0.20), and total employment (0.17).

2.4 Possible reasons for item non-response in BRDIS

Multiple imputation benefits from an understanding of the reasons for non-response. Quantification of those reasons in the form of variables, can be included as auxiliary information in the imputation step. The following elements of the survey design impact and influence the survey response process (Willimack and Snijkers, 2013):

The nature of the respondents. Response burden is shared by several departments of the company and is impacted by the organizational setting. This is particularly the case in BRDIS, where the survey has 5 parts and companies are asked to send each part to the appropriate department.

Different sampling and follow up strategies according to size and relative weight in the published statistics. Follow up strategies for non-respondents are more intensive for " statistically crucial" companies than for other companies. BRDIS has a special follow-up program for large companies with large R&D, whereas smaller companies are not followed up.

Survey mail out package. BRDIS is offered in two versions: A long form and a short form. The criterion for sending each has changed over the years but it is motivated by the fact that most companies do not have R&D costs. Using an abbreviated form for companies with a low probability of positive R&D activity is expected to reduce both survey costs and respondent burden. However, related to the forms is the threshold for which companies received the long or short form, which also changed, going from over 5 million to 1 million in 2012. Moreover, BRDIS stopped including innovation statistics for companies receiving the short form in 2012. All of those changes increase the business burden, as companies get used to one version and then have to learn a new one.

It is possible that the item non-response is due to attrition. The survey continues for many years, implying that some certainty businesses may be sampled more than once creating missingness due to attrition problems. The current sampling design does not use any procedure to reduce respondent burden by decreasing the probability of selection for companies that were sampled the previous years or for other major business surveys. Other government surveys, such as NASS's ARMS utilizes a Sequential Interval Poisson sampling procedure to reduce respondent burden by decreasing the probability of selection for operations that were sampled for ARMS the previous year or for other major NASS surveys.

The following section explains why this paper uses multiple imputation to address item non-response in BRDIS.

3. Methodology

As mentioned earlier, in this paper item non-response of unit responders is addressed via multiple imputation. Multiple imputation is ideally suited to the setting where disclosure of the identity of respondents is not allowed, which is the public-use survey data setting (Horton and Stuart, 2001): we can use auxiliary and confidential information that is inappropriate to disclose but ok to use to impute the data.

Support for the use of multiple imputation (MI), in conjunction with survey design considerations, comes from many years of its use in consumer surveys, and many other areas of research (Little,1988; Horton and Lipsitz, 2001; Schaffer, 1999; Little and Rubin, 1987). The strong support for this imputation method is based on the fact that MI preserves the correlation structure and distributional properties of the data. Practical implementation is now easy (if not straightforward) because the methods are now implemented in major commercial software packages like SAS, Stata, SPSS, R (Horton and Lipsitz, 2001; White et al., 2011), although discussion of them in the literature when using business survey data is not very prevalent. This paper uses Stata with SVY statements. The motivation for this approach is to improve the accuracy of estimates of the effects of company and economic environment variables on R&D expenses and univariate population estimates of total R&D.

It must be pointed out that multiple imputation in large-scale surveys is never a trivial task. Each survey has its own unique characteristics that can render procedures used on comparable surveys moot (Robbins et al., 2013). The problem is compounded if the survey is a business establishment survey, missing problems of which have received very little attention among researchers.

The results presented in Section 4 are based on 2009-2013 years of data for unit responders. Using all these years of data as information for imputation, the statistical model for the subpopulation of businesses participating in BRDIS in 2013 is estimated. For that year, results are compared under three different multiple imputation scenarios for all the years' data:

- No imputation, complete case, also known as listwise deletion (delete cases with missing data in any variable of interest.
- Multiple imputation for each SU based on information for all SUs, using chained regression.
- Separate multiple imputation for continued participants and one-time participants,using chained regression.

Available case analysis or listwise deletion consists of restricting attention to cases in which the variable of interest is observed. This biases the results of the available case analysis if respondents differ from non-respondents based on the recorded information in the incomplete questionnaire (Little and Rubin, 1987; Little, 1988). It has the undesirable effect that different aspects of a problem are studied with different subsets of data. So different populations are represented in each analysis. But this approach is commonly used by researchers.

MI generates several copies of the data set and fills in (imputes) each copy with different estimates of the missing values. The intention is to reproduce the variability and association among variables that would have prevailed in the full data set. It accounts for the uncertainty associated with the imputed values (Enders, 2010). Each of the simulated complete datasets is analyzed by standard methods, and the results are later combined to produce estimates and confidence intervals that incorporate missing-data uncertainty (Schafer,1999).

Table 2: Estimates of total R&D and average R&D for 2013 without multiple imputation (N=110000) and with multiple imputation(N=145000). Source: BRDIS and LBD 2009-2013.

Var name	Estimate	Standard error	% increase standard error
	No imputation, only sampling weight		
Total R&D estimate	2.40e+08	2.87e+08	NA
Average R&D estimate	262.42	26.31	NA
	No imputation, adjusted weights		
Total R&D estimate	2.50e+08	2.49e+07	NA
Average R&D estimate	8324.68	828.55	NA
	Multiple imputation, adjusted weights		
Total R&D estimate	3.60e+08	3.31e+07	2.09
Average R&D estimate	9022.83	831.04	2.09
	Multiple Imputation by count, adjusted weights		
Total R&D estimate	3.603+08	3.31e+07	1.45
Average R&D estimate	9019.97	818.28	1.45

In the research presented in this paper, the correlation structure of the imputed data sets is the same as the one of the incomplete data set. The next subsection provides more details.

3.1 Phases of Multiple imputation

We conduct stochastic imputation based on MCMC simulations. This section summarizes the steps (IDRE, 2016) .

1. Imputation phase: Using all years of data, multiple copies of the data (e.g.,m=50) are created, each of which contains different estimates of the missing values. This paper uses Stata's MICE (Multiple Imputation with Chained Equations, which are obtained via MCMC (White et al., 2011)). MICE is used because the missing data pattern is not monotone. The missing value of any of the variables with missing data is imputed with the prediction of that value given the other variables. Thus, R&D expense (R&D), R&D paid by others (R&DFO), R&D employment (R&DEMP) and total employment (TOTEMP) are each imputed according to the following iterative equations, where X1, X2, etc.. represent other variables of interest and auxiliary variables correlated with the missing data patterns but without missing values (see Table 1) :

$$
\begin{aligned}
R\&DFO &= \beta_1 + \beta_2 R\&D + \beta_3 R\&DEMP + \beta_4 TOTEMP + \beta_5 X1 + .. + \beta_k X_k + \epsilon \\
R\&D &= \beta_1 + \beta_2 R\&DFO + \beta_3 TOTEMP + \beta_4 R\&DEMP + .. + \beta_k X_k + \epsilon \\
R\&DEMP &= \beta_1 + \beta_2 R\&D + \beta_3 R\&DFO + \beta_4 TOTEMP + \beta_5 X1 + .. + \beta_k X_k + \epsilon \\
TOTEMP &= \beta_1 + \beta_2 R\&D + \beta_3 R\&DFO + \beta_4 R\&DEMP + \beta_5 X1 + .. + \beta_k X_k + \epsilon
\end{aligned}
$$

Table 3: Imputation Variance

Variable	Within	Between	Total	RVI	FMI	Relative efficiency
			Multiple imputation, adjusted weights			
Total R&D	$1.1e+15$	$4.4e+13$	$1.1e+15$	0.042	0.040	0.999
Average R&D	662688	27385	690621	0 .042	0.040	0.999
	Multiple imputation by count, adjusted weights					
Total R&D	$1.0e+15$	$3.0e+13$	$1.1e+15$	0.03	0.03	0.999
Average R&D	650543	18669	669586	0 .03	0.03	0.999

2. Analysis Phase: Analyze each of the 50 filled in data sets. Yields 50 sets of parameter estimates and standard errors.

3. Pooling Phase: The parameter estimates (e.g. coefficients and standard errors) obtained from each of the 50 data sets are then combined into a single set of results.

3.2 Imputation model

The imputation step of multiple imputation relies on the missingness law assumed. MICE is used because the missing data pattern is not monotone. The data used in this paper has an arbitrary pattern of missing values and is missing at random (MAR). The missing values of R&D, for example, are not due to their size, since rates of missingness are similar for different brackets of R&D. Stata 14's MI sequential regression imputation (also known as chained equations, fully conditional specification or MICE) is used combined with SVY, the latter in order to acknowledge the survey design (StataCorp, 2015).

The literature recommends to include in the imputation model: (a) variables included in the statistical model; (b) factors that correlate with missingness; (c) factors that explain a lot of variance in the target variable; (d) design and weight variables (Enders, 2010). This paper incorporates everything correlated to response. The sampling weight and the main strata and survey instrument information are used both as auxiliary variables in the imputation model, and as independent variables in the statistical model according to indications in Table 1. The items included try to capture the following concepts widely believed to affect non-response:

- business complexity (motive, authority and capacity) ia measured by multi/singleunit, total number of LBD establishments, legal forms of organization, LBD NAICS, number of states where the company operates, age of the oldest establishment (not included in the econometric model), amount of R&D research paid by others and the total R&D employment. The latter is included because, according to Hough et al., reporting it may increase the burden and complexity, as more documents have to be looked at.

- economic environment (total employment, total payroll, industry)

- survey collection and design (type of survey instrument (not included in econometric model), strata, sampling weight, number of years participating in BRDIS).

Table 4: Regression model of R&D against independent exogenous variables without imputation (N=110,000), with multiple imputation (N=145000) using all information and with multiple imputation based on number of times appearing in the survey. Subpopulation Year 2013 (N=23000 and 30000) Source: BRDIS and LBD 2009-2013. Weighted with adjustment weights for unit non-response. Controlling for industry.

R&D expense	MI full	No MI	MI by count
R&D performed	-0.228^{**}	-0.235	-0.223^{**}
R&D employee	357.17^{**}	362.51^{**}	356.86^{**}
Total employment	-1.25^{*}	-1.881^{*}	-1.25^{*}
Multi Unit	2294.52	3417.4^{*}	2244.56
Number of states	-368.6	-470.16	-354.57
Number of naics	-2196.66^{*}	-3467.8^{**}	-2183.64^{*}
Annual payroll	0.038^{*}	0.068^{*}	0.0384^{*}
Age	-22.96	-8.242	-21.61
R&D establishments	-656.26	-993.7	-677.27
Constant term	-1144.5	-916.28	1063.88

Respondent's personal characteristics are another concept but they are confidential, not available to RDC researchers, so they could not be included in the model.

The above procedures preserve the correlation structure in our sample, although they shift slightly the upper quartiles of the distribution of R&D.

Table 5: Imputation Variance. Variance information for statistical model coefficient estimates. Adjusted weight and industry control case

Variable	Within	Between	Total	RVI	FMI	Relative efficiency
R&D performed	.074	.000	.0074	0.0025	0.0025	.999
R&D employee	733.7	.16	733.86	.0002	.0002	.999
Total employment	.3627	.0001	.362	.0003	.0003	.999
Multi unit	1.7e+06	204896	1.9e+06	.1211	.1085	.997
Number of states	89541	2282.99	91869.9	.026	.025	.999
Number of naics	812461	1875	814374	.0023	.0023	.999
Total payroll	.0003	.0000	.0003	.0006	.0006	.999
Age	986	225	1215	.232	.189	.996
R&D establishments	625092	319	625418	.0005	.0005	.999
Constant	2.2e+06	232231	2.5e+06	.1065	.0966	.998

4. Results

4.1 Estimates of total and average R&D with and without multiple imputation

Historically, the objective of BRDIS data analysis has been to obtain population estimates of total R&D and related variables (NSF, 2016). As we can see in Table 2, the population estimate of total R&D without imputation for the year 2013 is 240,000 million (or

$240,000,000 thousand). With imputation, the estimate is $360,300 million with a standard error of 33,100 million (a relative standard error, se/estimate, of 9%). When doing the imputation separately for multi-year and single-year participants, the standard error is smaller, with smaller percentage increase in standard error due to imputation than when doing the imputation without discriminating according to that criterion.

Table 3 contains variance information for MI univariate estimates of R&D presented in Table 2. These are the imputation diagnostics. The *within* variance describes the sampling variance expected if there had not been missing data. It is calculated by taking the average of the 50 means obtained from the 50 imputed data sets. In some way, it is the expected variance due to only sampling. The *between* variance is the additional variance because of imputation (i.e., due to missing data). It is calculated as the variance of the 50 estimates around their mean. RVI stands for the relative increase in variance, representing the proportional increase in total sampling variance due to missing information. The FMI is the fraction of missing observations. It is the proportion of total sampling variance that is due to missing data. So it is directly related to RVI. Relative (variance) efficiency tells us how well the true population parameters are estimated. It is related to the amount of missing data and the number of imputations used. It tells us how well the variance is estimated with just 50 imputations instead of an infinite number of them.

4.2 Pooled estimates of the statistical model parameters with and without MI

Table 4 shows the estimated coefficients of the statistical model, adjusting for industry (industry coefficients not shown). The columns with MI estimates are averages of coefficients across all 50 imputations. The results are sensitive to the weighing procedure used. The table presents the results with adjusted weights. A double asterisk superscript means highly significant (p-value less than 0.01); a single asterisk means significant (p-value less than 0.05), and nothing means that the variables is not significantly related to R&D.

Something relevant to point out is that if we impute by R&D status or some other criteria such as by industry (these criteria not shown here), instead of by participant status, the relative results are consistent with those given in Table 4 if the same adjusted weight is used. The differences are mostly in the standard errors, which will always be larger to account for the uncertainty due to missing data and multiple imputation, but the percentage increase will be smaller if, say, imputing by industry.

Another observation derived from Table 4 is that the direction of the effect of the independent variables (negative or positive effect) is the same in all models. However, he effects based on non-imputed data are higher than those based on imputed data, except in the case of the effect of age. And some estimates that are significant when based on imputed data are not significant when based on nonimputed data. R&D performed paid by others is not significant without MI, but it is with MI. The multiunit nature of the company is significant without MI, but not with MI. Thus not only the magnitude but also the significance of some effects are different between multiply imputed and unimputed data.

Table 5 shows the multiple imputation diagnostics for the full imputation case, including industry in the model. We can see in the RVI column that the estimated sampling variance of the multi unit variable is 12.1% larger than its sampling variance would have been had the data been complete. Variables with lots of missing data or uncorrelated with the others tend to have high RVI. To look at another example, the estimated sampling variance for the age variable is 23.2% larger than its sampling variance would have been had the data been complete.

For all other variables in Table 5 it is the case that variance is higher due to item nonresponse adjustment, as expected.

5. Conclusions

This paper took an operational approach to shed light on the problem posed by item non-response of unit responders participating in BRDIS, a U.S. business establishment survey. Comparisons were made among different methods of making use of the data to estimate the same statistical, microeconometric, model of the determinants of R&D for the subpopulation of 2013 survey participants. One method multiply imputed the data with an imputation model that uses all 5 years of data to impute each item. Another method used all years of data but imputing companies that participated only once separately from companies that participated more than once. Those multiple imputation methods allow researchers to quantify the uncertainty due to the non-sampling error caused by item non-response, producing more accurate estimates. Their results were compared to those obtained without multiple imputation. In all the the methods used, unit responders‘ data were used, and adjustment for unit non response was done using the NSF/Census Bureau's adjustment weight. Not using this adjusted weight results in dramatically different results in all cases, but the results without the adjusted weights were not reported in this paper.

The paper showed that non-sampling errors indeed have an effect on estimates and their statistical significance. Although the direction of the effect of the independent variables (negative or positive effect) on R&D investment is the same under all methods, effects based on non-imputed data are higher than those based on imputed data and differ in statistical significance in some crucial variables. The conclusion derived from those results is that without taking into account the uncertainty due to item non-response researchers may be exagerating the effects of exogenous variables on R&D expenditures, and even judging some as significant when they are not.

Population estimates of total R&D were also compared under the same three methods, with the highest estimates obtained when using multiple imputation. The paper showed that the latter method produces higher standard errors, which incorporate the nonsampling error due to item non-response, thus are more accurate.

The main conclusion obtained in this paper is that multiple imputation makes a difference not only in the magnitudes of the standard errors of estimates, which account for the non-sampling uncertainty that unit non-response adds to estimates, but also in the size of the effects themselves.

More precision and more fined tuned estimates can be obtained by imputing groups that affect the magnitude of R&D, such as industrial groups, using only data for those groups. Thus, although the imputation model used is very complete in the sense that it includes all possible variables that affect item non-reponse, imputing within groups gives more precision. Regardless of what method is used, the standard errors of the estimates based on multiply imputed data are higher and more appropriate, because they reflect the uncertainty due to non-response, and hence give us some indication of the magnitude of the non-sampling error due to item non-response.

Analysis with different transformations of the data and narrower subpopulations (not shown in this paper) support the conclusions given above. However, the imputation method could be improved by using models that are more appropriate for such a skewed data. That is the case because, although the correlation structure of the data is maintained after imputation and so is the median, the imputed data sets have larger percentiles above the median than the complete data. Work in progress is using alternative models that take into account skewness and multiple zeros within narrower subgroups.

The conclusions presented in this paper may not apply to other type of establishment surveys. As pointed out by Dixon, item non-response is a source of nonsampling error. Its impact vary considerably by survey (Dixon, 2012).

Acknowledgements

The participation in ICES-V and this research were partially supported by Professional Development Awards and travel support received from the University of California Los Angeles.

REFERENCES

Cohen, W. M.(2010), "Fifty years of Empirical Studies of Innovative Activity and Performance," in *Economics of Innovation*, eds. B. H. Hall and N. Rosenberg, North Holland, pp=129–213.

Dixon, J. (2002), The effect of Item and Unit non-response on estimates of Labor Force Participation . JASA proceedings, 2002.?

Enders, C. K. (2010), *Applied Missing Data Analysis*, The Guilford Press.

Horton, N. J. and Lipsitz, S. R. (2001) Multiple Imputation in Practice, The American Statistician, 55:3, 244-254, DOI: 10.1198/000313001317098266

Hough,R.S and Curcio, Kayla and Keller, S.F. and Wilkinson, S. *Characteristics of Large Company Respondents to an Economic Survey*. Internal Census Bureau document.

IDRE, Statistical Consulting Group, (2016), "Missing Data Techniquest with Stata," `http://www.ats.ucla.edu/stat/Stata/seminars/missing_data/Multiple_imputation/Missing%20Data%20Techniques_UCLA_Stata.pdf`

Jarmin, R.S. and Miranda, J. (2012), "The Longitudinal Business Database," *Center for Economic Studies*, Working paper 12-17, U.S. Census Bureau.

Little, R. J. A (1988), "Missing-Data Adjustments in Large Surveys," *Journal of Business & Economic Statistics*, 6:3, pp. 287–296.

Little, R. J. A and Rubin, D. B. (1987), *Statistical Analysis with Missing Data, 2nd edition*, Wiley.

Mason, R., Lesser, V., and Traugget, M. (2002), " Effect of Item non-response on non-response error and Inference." In *Survey non-response*, eds. R. Groves, D. Dillman, J. Eltinge and R. Little, Wiley, N.Y., pp. 149–162.

Miller, D., Lisic, J., and Dau, A. (2016), "Imputations' Reaction to Data: Exploring the Boundaries and Utility of IVEware and Iterative Sequential Regression (ISR)," paper presented at ICES-V, Geneva.

NSF (2016). BRDIS Web Page at `https://www.nsf.gov/statistics/srvyindustry/`

Robbins, M. W., Ghosh, S. K., and Habiger, J.D. (2013), "Imputation in High-Dimensional Economic Data as Applied to the Agricultural Resource Management Survey," *Journal of the American Statistical Association*, 108-501, 81–95.

Schafer, J.L. (1999). *Multiple Imputation: a primer* . Statistical Methods in Medical Research, 8. Pp 3-15.

StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP.

Willimack, D.K and Snijkers, G. (2013). *The Business Context and its Implications for the Survey Response Process*. In *Designing and Conducting Business Surveys*, eds. G. Snijkers, G. Haraldsen, J. Jones, and D. K. Willimack, Wiley, N.Y. pp. 39–82.

White I. R., Royston, P., and Wood, A.M. (2011), "Multiple imputation using chained equations: Issues and guidance for practice," *Statistics in Medicine*, 30 (4): pp. 377–99.

Wolfe, R. M. (2010), "NSF Announces New U.S. Business R&D and Innovation Survey," *NCSES Infobrief*, NSF 09-304, Arlington,VA:National Science Foundation, Division of Science Resources Statistics.