

UCLA

Department of Statistics Papers

Title

R&D, Attrition and Multiple Imputation in The Business Research and Development and Innovation Survey (BRDIS)

Permalink

<https://escholarship.org/uc/item/1bx747j2>

Authors

Sanchez, Juana

Kahmann, Sydney N

Li, Dennis

Publication Date

2017-09-14

R&D, Attrition and Multiple Imputation in The Business Research and Development and Innovation Survey (BRDIS)

Juana Sanchez
Sydney Noelle Kahmann
Dennis Li

UCLA Department of Statistics



Presented by the three authors at the Annual Conference of the Federal Statistical Research Data Centers on "Big Data" on September 14, 2017, Los Angeles, California

The results regarding BRDIS data were obtained while Juana Sanchez 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 authors and do not necessarily reflect the views of the Census Bureau. The research has been screened to insure that no confidential data are revealed.

- Item nonresponse is a source of non-sampling error. Its impact on error may vary considerably by survey (Dixon, 2002).
- Impact on population estimates of R&D based on BRDIS are unknown
- **Goal: improving accuracy of estimates of the effects of firm and economic environment variables on R&D expenses using MI.**



- NSF/Census Bureau BRDIS: annual mandatory survey of about 40,000 US nonprofits. Manufacturing ($\sim 42\%$), services and research business ($\sim 58\%$) included. Linked to LBD administrative data.
- 3 strata: Unknown R&D, $R\&D > 0$, and $R\&D = 0$.
- NSF provides national estimates of total R&D and R&D employment based on BRDIS



- **R&D not imputed by NSF or Census Bureau.**
- False impression of constant annual data quality disappears when studying missing data patterns over time
- This research shows that survey design characteristics lead to attrition at a higher rate for higher R&D performers. After MI, we show that estimates of total R&D vary considerably.

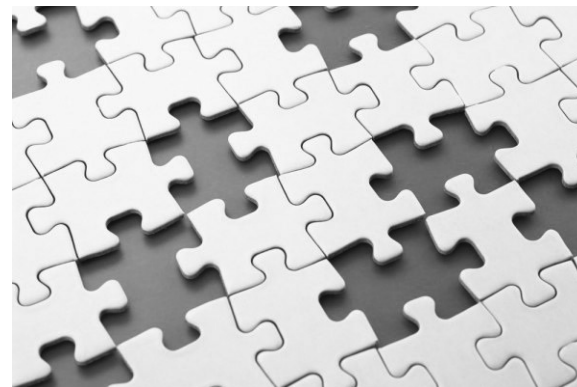
Before Conducting MI, Missing Data Patterns Were Explored!



“Well, this certainly explains much of the company’s missing data. Who else thought the ‘DEL’ key on their computer was for delegating work?”

Mechanisms describe the assumptions about the nature of the missing data and can be categorized as follows:

1. MCAR (Missing Completely at Random)
2. MNAR (Missing Not at Random)
3. MAR (Missing at Random)



Re: Little and Rubin(1987)

- Probability of missing values has nothing to do with the observed or missing values
- **R&D question is compulsory. Not MCAR.**



- Probability of missing values depends on the missing values themselves, and can also depend on observed values too
- **BRDIS is not MNAR based on our study because missingness is due to survey design characteristics**

- Probability of missing values depends only on the observed values of other variables in the dataset (not the missing variable itself)
- **In BRDIS, the unit and item nonresponse in the R&D field is MAR and due to survey design characteristics.**

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
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 NAICS	2	3.6		y	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



- Packages in R that can be used to visualize the missing data through plots include VIM and Amelia. We use **simulated data**.

- **Amelia**

- missmap

Re: Honaker (2011)

- **VIM**

- aggr
- marginplot
- pbox
- spineMiss
- matrixplot

Re: Templ & Filzmoser (2008)

Table 2: Data as it comes in BRDIS

ID	COUNT	MYOBS	YEAR	R&D	UR
222	3	1	2008	20000	1
222	3	2	2010	15000	1
222	3	3	2012	.	1
541	3	1	2008	}	1
541	3	2	2009		0
541	3	3	2010		689

- COUNT=how many years company is surveyed
- Companies within a count are similar in R&D, payroll, employment, stratum.

Table 3: Reshaping the data

ID	COUNT	MYOBS	year	RD1	RD2	RD3	UR1	UR2	UR3
222	3	1	2008	20000	15000	.	1	1	1
541	3	1	2008	.	.	689	1	0	1



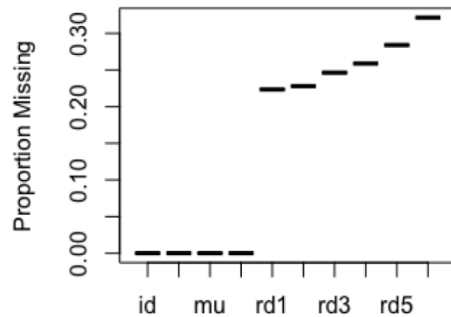
The data shown are artificial, for illustrative purposes.

- Simulated data of companies that had COUNT=6.

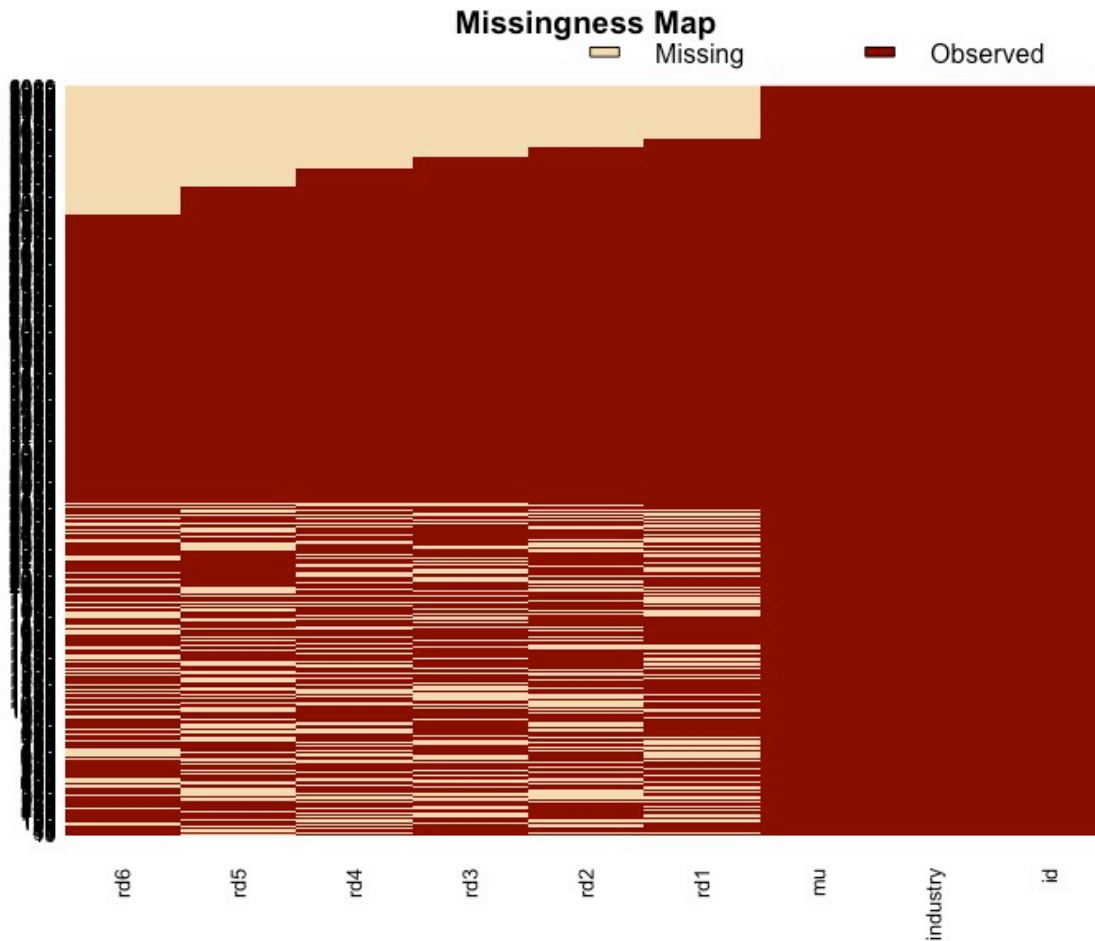
ID	RD1	RD2	RD3	RD4	RD5	RD6	UR1	UR2	UR3	UR4	UR5	UR6	MU	ind
234	.	25	21	11	.	.	1	1	1	1	1	.	1	2
456	4	1	.	.	.	3	1	1	1	.	.	1	0	3

Figure 1: Proportion of missing values by variables in the simulated data set

Proportion of Missing Values by Variable

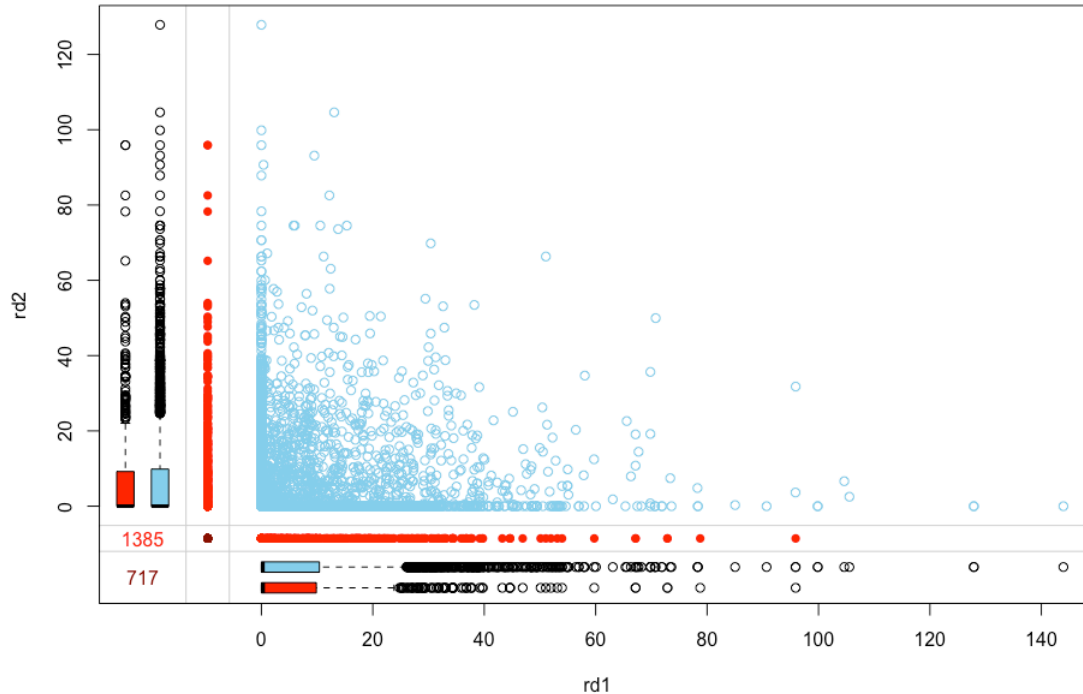


- Year when missing is randomly chosen by the company.
- Item non-response higher in RD6
- In general, item nonresponse due to unit nonresponse.



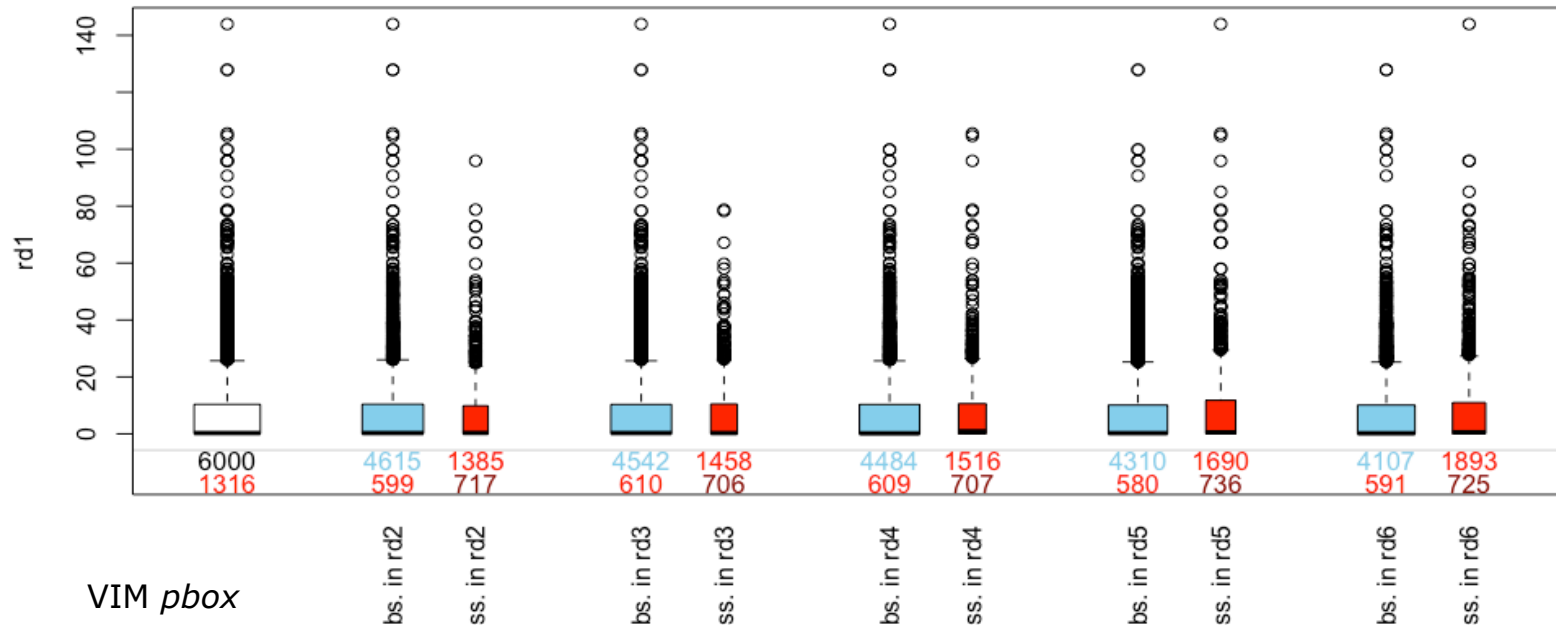
This view of the data is with *Amelia missmap* command showing the data the way we simulated it.

Amelia missmap



VIM *marginplot*

Along the horizontal axis the two parallel box plots both represent the variable rd1, but the red box plot is for those values of rd1, where no values for rd2 are available, and the blue box plot for rd1 values where the information for rd2 is available.



- Distribution of rd1 is white boxplot
- The other boxplots shown also refer to rd1, but they are grouped according to missing-ness (red) or non missing-ness (blue) of each observation in another variable.
- In this plot, there is no dependence between magnitude of rd1 and presence of missing values in the other variables.

BRDIS only surveys active companies

Y = non-missing




- YYYYYN, YYYYNN, YYNNNN, YYNNNN, YNNNNN - attrition due to survey response burden
- YYNYNY, NYYYYY, YNYNY, YNNYYN, etc. – examples of temporary attrition, good candidates for imputation

- Count is a proxy for firm size, age, industry, payroll, employment, survey design variables and R&D
- Companies in the same count are similar → they should be imputed using their count group information
- Makes sense to use count as an important variable in the imputation.
- **So... what type of imputation?**

1. Complete Case (CC) Analysis
2. Inverse Probability Weighting (IPW)
3. Last Observation Carried Forward (LOCF) Imputation
4. Unconditional Mean Imputation
5. Single Imputation
6. Stochastic Imputation
7. Multiple Imputation

- Default method in statistical software packages such as R, Stata, SAS. Most commonly used.
- Delete whole row which contains missing data on any variable
- **Advantages:** easiest, default, unbiased with MCAR
- **Disadvantages:** loss of valuable data, mostly biased (MCAR is rarest)

Subject	Weight	Age	Sex
1	150	60	F
2	.	43	M
3	190	20	M
4	210	38	M
5	.	19	F



Subject	Weight	Age	Sex
1	150	60	F
3	190	20	M
4	210	38	M

Megan M. Marron & Abdus S. Wahed (2016) Teaching Missing Data Methodology to Undergraduates Using a Group-Based Project Within a Six-Week Summer Program, *Journal of Statistics Education*, 24:1, 8-15

- Look for similarities between subjects who are missing the outcome of interest vs. those who are not
- Find pairings where similarities exist, and calculate the probability of missing the outcome of interest based on pairings
- **Advantages:** results are unbiased under MAR and MCAR
- **Disadvantages:** reduced sample size, skewed if small predicted probability of complete data

Table 1. Data used to explain IPW.

Subject	Age	Sex	Year in College
1	.	F	Graduated
2	.	F	Junior
3	20	M	Junior
4	24	F	Graduated
5	21	F	Senior
6	19	F	Junior

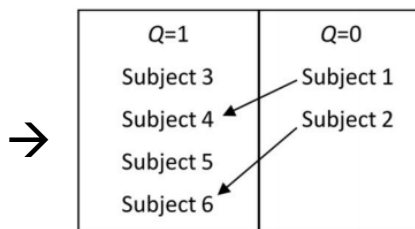



Figure 2. Grouping subjects based on having complete or missing data.

$$\begin{aligned}
 \text{Estimated Mean Age} &= \frac{1}{6} (\text{Subject3's age} + 2 * \text{Subject4's age} \\
 &\quad + \text{Subject5's age} + 2 * \text{Subject6's age}) \\
 &= \frac{1}{6} \left(\frac{Y_{\text{Subject3}}}{1} + \frac{Y_{\text{Subject4}}}{\frac{1}{2}} + \frac{Y_{\text{Subject5}}}{1} + \frac{Y_{\text{Subject6}}}{\frac{1}{2}} \right) \\
 &= \frac{1}{6} \sum_{i=1}^6 \frac{Q_i Y_i}{\hat{P}(Q_i = 1 | X_i)}.
 \end{aligned}$$

Megan M. Marron & Abdus S. Wahed (2016) Teaching Missing Data Methodology to Undergraduates Using a Group-Based Project Within a Six-Week Summer Program, *Journal of Statistics Education*, 24:1, 8-15

- Plug in last available measurement in place of the missing values
- **Advantages:** very simple
- **Disadvantages:** decreased sample variance (replacement with identical values)
- It is the least preferred method because of large bias

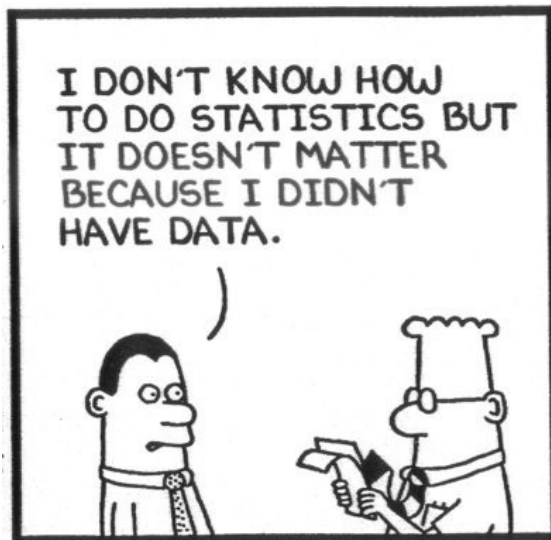
Subject	Age	Sex	Week		
			1	2	3
1	60	F	20.1	20.9	.
2	43	M	13.7	.	15.3
3	20	M	18.0	19.1	20.2
4	38	M	19.3	20.0	.



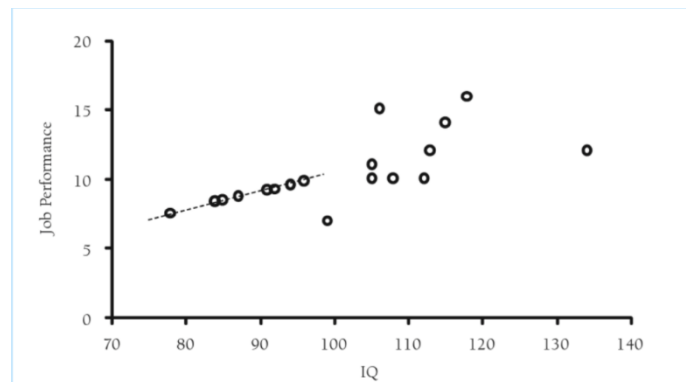
Subject	Age	Sex	Week		
			1	2	3
1	60	F	20.1	20.9	20.9
2	43	M	13.7	13.7	15.3
3	20	M	18.0	19.1	20.2
4	38	M	19.3	20.0	20.0

Megan M. Marron & Abdus S. Wahed (2016) Teaching Missing Data Methodology to Undergraduates Using a Group-Based Project Within a Six-Week Summer Program, *Journal of Statistics Education*, 24:1, 8-15

- Replace missing values with the mean of the available values
- **Advantages:** easy to implement
- **Disadvantages:** leads to a reduction in variability. It also changes the correlation between the imputed variable vs. other variables.

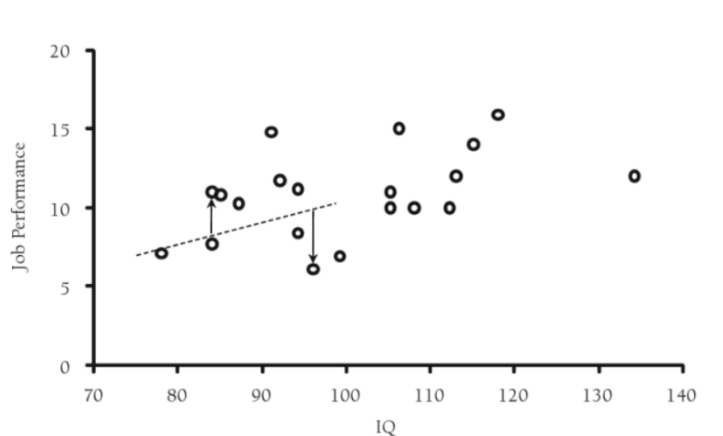


- Also known as deterministic/regression/conditional mean imputation: where missing values are imputed with predicted values from a regression equation
- **Advantages:** usage of complete information to impute
- **Disadvantages:** imputed values are directly from the regression line, decreasing variability. It does not reflect the full uncertainty of the missing data.



Megan M. Marron & Abdus S. Wahed (2016) Teaching Missing Data Methodology to Undergraduates Using a Group-Based Project Within a Six-Week Summer Program, *Journal of Statistics Education*, 24:1, 8-15

- Done by adding randomly drawn residuals from regression imputation, based on residual variance from regression model
- **Advantages:** “adds back” lost variability from regression imputation and produces unbiased correlation estimates under MAR



Megan M. Marron & Abdus S. Wahed (2016) Teaching Missing Data Methodology to Undergraduates Using a Group-Based Project Within a Six-Week Summer Program, *Journal of Statistics Education*, 24:1, 8-15

- Used STATA 14's MICE. Specialized to survey data, allows imputation by count and subpopulation analysis at the estimation stage.

Re: Schafer(1999), Enders(2010), IDRE(2016), Rubin(1987), Little(1988), White et al., (2011).

- Imputation phase: Using all years data, create multiple copies of the data (e.g., $m=50$, each of which contains different estimates of the missing values). R&D, R&DFO, R&DEMP and TOTEMP are imputed. The imputation model is:

$$R\&D = 1 + 2R\&DFO + 3 + 4TOTEMP + 5X1 + \dots + kXk +$$

$$R\&DFO = 1 + 2R\&D + 3R\&DEMP + 4TOTEMP + 5X1 + \dots + kXk +$$

$$R\&DEMP = 1 + 2R\&D + 3R\&DFO + 4TOTEMP + 5X1 + \dots + kXk +$$

$$TOTEMP = 1 + 2R\&D + 3R\&DFO + 4R\&DEMP + 5X1 + \dots + kXk +$$

- Analysis Phase: Analyze each of the 50 filled in data sets. Yields 50 sets of parameter estimates and standard errors.
- Pooling Phase: The parameter estimates (e.g. coefficients and standard errors) obtained from each of the 50 data sets are pooled.

Table 10: Univariate estimates of Total and Average R&D without and with imputation plus Imputation Variance. Year 2013. BRDIS 2009-2013.

Statistic	Estimate	s.e.	No imputation			
			95% CI LB	95% CI UB		
Total R&D	2.61e + 08	2.49e + 07	2.12e + 08	3.10e + 08		
Average R&D	11513	1098.121	9360.691	13665.31		
Multiple Imputation						
Statistic	Estimate	s.e.	95% CI LB	95% CI UB		
Total R&D	3.81e + 08	3.18e + 07	3.18e + 08	4.43e + 08		
Average R&D	12640.4	1054.216	10574.15	14706.65		
Multiple Imputation Diagnostics						
	Within	Between	Total	RVI	FMI	Relative efficiency
Multiple imputation by count, adjusted weights						
Total R&D	1.0e + 15	7.1e + 11	1.0e + 15	0.0008	0.0008	0.9998
Average R&D	1.1e + 06	778.537	1.1e + 06	0.0008	0.0008	0.9998

Table 7: Regression model of R&D against independent variables for 2013, without multiple imputation (CCA). Subpopulation study for 2013 using ($N_{2009-2013} = 110000$; subpopulation $N=23000$). Three industry categories are used as control: research, manufacturing (not research) and service. The last two were used as independent variables and only the manufacturing (non-research) was statistically significant with a large effect. ($p < 0.01$). The service category has a negative effect that is not significant.

Variable	Coef.	Std. Err.	t	$P > t $	95% Conf. Int
l.mu	3246.022	4805.34	0.68	0.499	(-6172.403, 12664.45)
paytotal	0.1426	0.0567	2.51	0.012	(0.0313, 0.2538)
agemax	-372.2174	89.0009	-4.18	0.000	(-546.6584, -197.7764)
nnaics	2684.272	3477.208	0.77	0.440	(-4131.025, 9499.57)
nstate	-617.9197	1403.648	-0.44	0.660	(-3369.058, 2133.219)
rdesttotal	3251.362	3792.992	0.86	0.391	(-4182.87, 10685.59)
constant	871.5226	5845.309	0.15	0.881	(-10585.23, 12328.28)

Table 8: Regression Stata MI with PMM ($N_{2009-2013} = 145000, N_{2013} = 30000$). Number of burn in iterations=10, datasets=5, nearest neighbors=5. Multiple regression results with MI and subpopulation analysis for 2013. Uses all the data for the estimation of standard errors, but only 2013 for the regression coefficient estimates.

Variable	Coef.	Std. Err.	t	$P > t $	DF	% increase s.e.
1.mu	3613.862	3768.045	0.96	0.338	110956.6	0.00
paytotal	0.1378	0.0411	3.36	0.001	110810.4	0.01
agemax	-355.5897	69.3308	-5.13	0.000	106687.9	0.06
nnaics	2037.27	2760.437	0.74	0.461	110231.4	0.02
nstate	23.9638	1121.139	0.02	0.983	107453.1	0.05
rdesttotal	4114.519	4379.975	0.94	0.348	110987.3	0.00
constant	-195.6143	6125.513	-0.03	0.975	110637.7	0.02

- Our study of missing data patterns in BRDIS linked to LBD suggests that attrition due to survey response burden is the main reason for item nonresponse, more so in higher R&D companies.
- MI of the data that uses that information provides us with more observation for regression analysis to study economic theories that matter (without changing the correlation structure of the data).
- We found that estimates of total R&D are higher than estimates obtained with complete case analysis.
- Recommendations: Moving to Poisson sequential sampling might be a good idea to adopt by NSF/Census Bureau.
- More information on this research can be found in CES working paper 17-13. This presentation will appear in UC e-scholarship.

Does Item Non-Response Help Predict Unit Non-Response?

Table 5: Recoding the simulated data

ID	COUNT	MYOBS	year	RD1	RD2	RD3	UR1	UR2	UR3
222	3	1	2008	0	0	1	1	1	1
541	3	1	2008	1	1	0	1	0	1

- A common way to do this in the response literature is binary logistic regression.
- By modeling the probability of unit nonresponse in the last year, j , as a function of unit nonresponse and item non response in period $j-1$, we can test the hypothesis that item nonresponse helps predict future unit nonresponse.
- **In BRDIS, we found item nonresponse in recent years to be significant predictor of unit nonresponse in the next year.**

- Enders, C.K. *Applied Missing Data analysis*. The Guilford Press.
- Honaker, J., King, G. and Blackwell. Amelia II. A program for Missing Data. *Journal of Statistical Software*, 45(7):1-47,2011.
- IDRE. https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/
- Little, R.J.A. Missing Data Adjustment in Large Surveys. *Journal of Business and Economic Statistics*, 6(3):287-296,1988.
- Little, R. J.A. and Rubin, D. B. *Statistical Analysis with Missing Data*, 2nd edition. Wiley and Sons, 1987
- Marron, M.M. and Wahed, A.S. Teaching Missing Data Methodology to Undergraduates Using a Group-Based Project Within a Six-Week Summer Program, *Journal of Statistics Education*, 24:1, 8-15, 2016.
- NCSES/Census Bureau, BRDIS (<http://www.nsf.gov/statistics/srvyindustry/about/brdis/>)
- Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 1987.
- Sanchez, J. and Kahmann, S.N. R&D, Attrition and Multiple Imputation in BRDIS. *CES working paper* 17-13. U.S. Census Bureau.
- Schafer, J. Multiple Imputation: a primer. *Statistical Methods in Medical Research*, 8:3-15,1999.
- Templ, M. Alfons, A. and Filzmoser, P. Exploring Incomplete Data using Visualization Techniques. *Adv. Data Anal Classif*, 6:29-47,2012
- White, T., Reiter, J. and Petrin, A. Plant-level Productivity and Imputation of Missing Data in the Census of Manufactures. *CES Working Paper* 11-02, 2011. U.S. Census Bureau.