Stratified Multiple Imputation Estimation for Complex Surveys

Juana Sanchez

UCLA Department of Statistics

The results regarding BRDIS data presented here were obtained while the author was Special Sworn Status researcher at the U.S. Census Bureau (Center for Economic Studies), using the California Census Research Data Center at UCLA. Research results and conclusions expressed are those of the author and do not reflect the views of the Census Bureau. The research has been screened to ensure that no confidential data are revealed.

Overview

- This research presents an approach to conducting multiple imputation of the R&D expenditures of businesses. It could be titled: don't forget the past when imputing the present.

- The approach uses all past years of survey data to learn about the attrition patterns of businesses. This gives a criterion for imputing the data. Based on that criterion, all years of data are imputed at once, each year's imputation gaining strength from the other year's information.

- The approach is purely empirical/

- STATA's MICE is an ideal software to conduct the multiple imputation because although we do pooled multiple imputation it gives us at the same time the estimates of R&D for the subpopulation of the current year.

1. The Data used: BRDIS and LBD

- The data comes from NSF/Census Bureau BRDIS 2009-2013. This is an annual mandatory survey of about 40,000 US businesses with 5 or more employees to collect data on full range of R&D activities and innovation. The survey has changed name several times since then (BRDS since 2017 with no innovation, BERD since 2020).

- Manufacturing business(~42%), services and research business (~58%) included.

- The BRDIS data was linked to LBD administrative data while at CCRDC.

- NSF is in charge of the annual national estimates of R&D intensity (~3% of GDP) and R&D employment (https://ncses.nsf.gov/pubs/nsb20201)

2. Survey design

- Two survey design variables:
- Complex stratified design: 3 strata by size: Unknown R&D, R&D>0, and R&D=0.

- A second layer of stratification based on industry.

- Different sampling methods: with certainty, PPS (Probability Proportional to Size), SRS (simple Random Sampling).

3. Nature of the missing Data in BRDIS

- **R&D not imputed by NSF or Census Bureau. Weights account for missing values.**

- Simulating data of missing data patterns like those observed in BRDIS (Sanchez and Kahmann, 2017), revealed that the probability of missing values in R&D depends only on the observed values of other variables in the dataset (not the variable itself) and in particular the survey design variables.

  - **R&D missing data is MAR.**

  - **In BRDIS, the unit and item nonresponse in the R&D could be attributed to survey design.**

For example, patterns of nonresponse vary by strata

BRDIS only surveys active companies

Y =response that year; N=no response that year

- YYYYYN, YYYYNN, YYYNNN, YYNNNN, YNNNNN - attrition due to survey response burden
- YYNYNY, NYYYNY, YNYNYY, YNNYYN, etc. – examples of temporary attrition, good candidates for imputation. This is the pattern of high R&D performers.

4. Creating an index by which to impute.

- **New variable “Coun**t” (how many years the company was included in the data frame) is a proxy for firm size, age, industry, payroll, employment, survey design variables (strata, industry, sampling method) and R&D.

- Companies in the same count are similar in missing data patterns → they should be imputed using their count group information → **MI done by count in this research**

- Makes sense to use count as an important variable in the imputation.

5. BRDIS Imputed Using Stata's MICE

- Used STATA 14's MICE. Specialized to survey data `(svyset id, strata(stratm))`, 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).

Code with names modified

```
mi impute chained (pmm,knn(5)) weightedRD      = i.mu    ///
       i.industry age pay stratumvar samplingweightvars///
         geographicdispersionUS
industrialsectors,   ///
 add(50) burnin(1000) by(count)
```

Phases of MI for BRDIS

- Imputation phase: Using four years survey data (2009-2013 surveys), create multiple copies of the data (e.g., m=50, each of which contains different estimates of the **missing values obtained by count**). Weighted R&D, R&DFO, R&DEMP and TOTEMP are imputed. The imputation model is:

R&D = 1 + 2R&DFO + 3 + 4TOTEMP + 5X1 + :::::: + kXk +
R&DFO = 1 + 2R&D + 3R&DEMP + 4TOTEMP + 5X1 + :::::: + kXk +
R&DEMP = 1 + 2R&D + 3R&DFO + 4TOTEMP + 5X1 + :::::: + kXk +
TOTEMP = 1 + 2R&D + 3R&DFO + 4R&DEMP + 5X1 + :::::: + 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.

National indicators of R&D with and without MI

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

			No imputation			
Statistic	Estimate	s.e.	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			
	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

To summarize

- Multiple imputation pooling all years of data and done **by count** (the variable that is related to survey variables) to gain strength with more information and exploit the fact that missing patterns vary by count.

- Estimate of the total and average R&D for the subpopulation of 2013 using the data as imputed.

- Using the strength of all the years of data to impute, the estimate of a subpopulation of 2013's R&D is higher than the one obtained without imputation.

- If a survey is repeated every year, why forget what is learned from the past when it could help us provide more accurate estimates?

Predictors of R&D without MI

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
1.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	$(-10585.23, 12328.28)$

Predictors of R&D with MI

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

Conclusions

- Study of missing data patterns in BRDIS linked to LBD suggests that attrition due to survey response burden (proxied by count) is the main reason for item nonresponse, more so in higher R&D companies.
- MI of the data that uses that information (MI by count) 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 and average 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.

Appendix

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.**

References

- 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. https://www2.census.gov./ces/wp/2017/CES-WP-17-13.pdf
- 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.