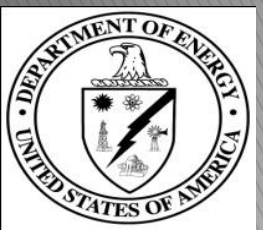
SAND2011-7197C C

Multiple Imputation for Missing Data

Kimberly Proctor
Strategic Studies, Org. 249



Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Introduction

- Datasets, particularly in the social sciences, are notoriously incomplete
- Standard estimation techniques do not account for this missingness
- Our resulting models are likely incorrect = we make the wrong conclusions!

Example: Nuclear Detection Architecture Index

- Ranks countries' social/political/security aspects for detecting and deterring threats to nuclear security
- Example: all countries in the world for 2008

Variable	% Missing
Military expenditure (% GDP)	35
GDP per capita (US$)	10
GDP growth (annual %)	11
Total	37

Assumptions about Missingness

- Missing Completely at Random (MCAR)
  - Only analyze complete cases
  - What we know and what we are missing are completely independent
  - Generally appropriate if less than 5% missing

- Missing at Random (MAR)
  - Missing data and observed data are related through some other information
  - We can explain what we are missing with the information we have

Sandia National Laboratories

Problems with Assuming MCAR

- Most stats software packages omit observations with missing information from the estimation model
- Discards a vast amount of information
  - Almost 40% in our example!
- Introduces bias to the extent that the observed cases differ systematically from missing cases and inefficiency from loss of information

Ad Hoc Methods

- Listwise deletion
  - Our estimates are incorrect
    - Wrong magnitudes, signs, significance, standard error; non-representative sample

- Mean, median, mode substitution
  - Our estimates are incorrect
    - Distorts covariance = biased estimates toward zero
    - Treats “guesses” as if they were real data
    - Does not account for uncertainty

Multiple Imputation (MI)

- MI is the state of the art technique for handling missing data in the social sciences

- MI handles missing data in advance of modeling
- Run M>1 Monte Carlo simulations of complete data
  - Results in M plausible but different versions of complete data
  - Analyze the results
  - Average parameters across the datasets

MI: Technical Details

- Standard errors are averaged,
  - Equation contains elements that reflect the uncertainty due to missing data
- MI estimates are
  - Consistent, normal, efficient and valid
- As M increases, MI estimates become more efficient
  - M >=5 & <=10 is standard practice

MI: Optimizing the Technique

- Auxiliary Variables are KEY!
  - Variables not included in the final estimation model that are used in the imputation model
  - Based on theory and lags and leads
  - Ensures that data is MAR
  - Include more, rather than less (~20)

Sandia National Laboratories

MI: Pros and Cons

- Most (only) ‘sound’ statistical method for handling missing data

- Can create high correlations among variables
- Can lead to nonsensical predictions
  - Negative military expenditure
  - This gets less likely as auxiliary information improves
- No clear guideline for how much missing data is too much missing data

MI in STATA

```
. ice  milex-gro, m(5) clear persist

  #missing
    values |      Freq.     Percent        Cum.
-----------+-----------------------------------
         0 |         141       64.38       64.38
         1 |          55       25.11       89.50
         2 |           1        0.46       89.95
         . |          22       10.05      100.00
-----------+-----------------------------------
     Total |         219      100.00

  Variable | Command | Prediction equation
-----------+---------+--------------------
     milex | regress | gdp gro
       gdp |         | [No missing data in estimation sample]
       gro | regress | milex gdp
-----------------------------------------------

Imputing .....1.....2.....3.....4.....5
[note: imputed dataset now loaded in memory]

Warning: imputed dataset has not (yet) been saved to a file
```

Sandia National Laboratories

var10[25]

Snapshots

	year	cname	milex	gdp	gro	_mi	_mj
9	2008	Albania	1.97151	4076.4	7.7	2	2
10	2008	Albania	1.97151	4076.4	7.7	2	3
11	2008	Albania	1.97151	4076.4	7.7	2	4
12	2008	Albania	1.97151	4076.4	7.7	2	5
13	2008	Algeria	3.02494	4966.57	2.4	3	0
14	2008	Algeria	3.02494	4966.57	2.4	3	1
15	2008	Algeria	3.02494	4966.57	2.4	3	2
16	2008	Algeria	3.02494	4966.57	2.4	3	3
17	2008	Algeria	3.02494	4966.57	2.4	3	4
18	2008	Algeria	3.02494	4966.57	2.4	3	5
19	2008	Andorra	.	44952.4	3.57074	5	0
20	2008	Andorra	.702315	44952.4	3.57074	5	1
21	2008	Andorra	3.61675	44952.4	3.57074	5	2
22	2008	Andorra	.186436	44952.4	3.57074	5	3
23	2008	Andorra	2.31893	44952.4	3.57074	5	4
24	2008	Andorra	.042755	44952.4	3.57074	5	5
25	2008	Angola	2.87967	4666.74	13.8	6	0
26	2008	Angola	2.87967	4666.74	13.8	6	1
27	2008	Angola	2.87967	4666.74	13.8	6	2
28	2008	Angola	2.87967	4666.74	13.8	6	3
29	2008	Angola	2.87967	4666.74	13.8	6	4
30	2008	Angola	2.87967	4666.74	13.8	6	5
31	2008	Antigua and Barbuda	.	13850.1	.172689	7	0
32	2008	Antigua and Barbuda	2.13729	13850.1	.172689	7	1
33	2008	Antigua and Barbuda	.963058	13850.1	.172689	7	2
34	2008	Antigua and Barbuda	3.68883	13850.1	.172689	7	3
35	2008	Antigua and Barbuda	3.49828	13850.1	.172689	7	4
36	2008	Antigua and Barbuda	.0875	13850.1	.172689	7	5

References

- Collins, Linda M., Joseph L. Schafer, and Chi-Ming Kam. 2001. A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods* 64330-351.
- Honaker, James, and Gary King. 2010. What to Do About Missing Values in Time-Series Cross-Section Data. *American Journal of Political Science* 54 (2):561-581.
- King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *The American Political Science Review* 95 (1):49-69.
- Little, Roderick J. A. 1988. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association* 83 (404):1198-1202.
- ———. 1992. Regression with Missing X's: A Review. *Journal of the American Statistical Association* 87 (December):1227-37.
- Little, Roderick J. A, and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. New York.

Sandia National Laboratories

- Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. Edited by V. Barnett, R. A. Bradley, J. S. Hunter, D. G. Kendall, A. F. M. Smith, S. M. Stigler and G. S. Watson, *Wiley Series in Probability and Mathematical Statistics*. New York: John Wiley & Sons.
- ———. 1996. Multiple Imputation after 18+ Years. *Journal of the American Statistical Association* 91 (434):473-489.
- Rubin, Donald B., and Roderick J. A Little. 1987. *Statistical Analysis with Missing Data*. Edited by V. Barnett, R. A. Bradley, J. S. Hunter, D. G. Kendall, R. G. Miller, A. F. M. Smith, S. M. Stigler and G. S. Watson, *Wiley Series in Probability and Mathematical Statistics*. New York: John Wiley & Sons.
- Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall.

MI: Technical Details Appendix

- MI estimates are generated using the Markov Chain Monte Carlo (MCMC) algorithm based on linear regression
- Predicted values based on regression (logit, etc.) and random draws made from a simulated error distribution
  - Errors are added to the predicted value
- Values with no missing data remain the same in each M dataset