UNECE Work Session on Statistical Data Editing
Budapest, Hungary, 14-16 September 2015

Multiple Ratio Imputation by the EMB Algorithm

National Statistics Center (Japan)
Masayoshi Takahashi

Notes: The views and opinions expressed in this presentation are the authors' own, not necessarily those of the institution.

National Statistics Center
NSTAC

Outline

1. Missing Data Problems and Existing Imputation Methods
2. Theory of Multiple Ratio Imputation
3. Monte Carlo Evidence
4. Empirical Example

Missing Data Problems and Existing Imputation Methods

National Statistics Center

NSTAC

Missing Data and Ratio Imputation

- Missing data problems are ubiquitous in many fields.
- In official statistics, one of the common treatments of missing data is ratio imputation.

Single Ratio Imputation Model

- Form of a simple regression model without an intercept
- Slope coefficient calculated by the ratio between the means of two variables

$$\hat{Y}_{i1} = \hat{\omega} Y_{i2} \text{ (Deterministic)}$$

$$\hat{Y}_{i1} = \hat{\omega} Y_{i2} + \hat{u}_i \text{ (Stochastic)}$$

where $\hat{\omega} = \bar{Y}_{1,obs} / \bar{Y}_{2,obs}$

de Waal *et al.* (2011)

Thompson & Washington (2012)

Office for National Statistics (2014)

Multiple Imputation

- Recommended practice from statisticians
- Known to be the gold standard of treating missing data

Rubin (1987)
Little & Rubin (2002)
Baraldi & Enders (2010)
Cheema (2014)

Multiple Imputation

- Multiple imputation in theory
  - Randomly draw several imputed values from the distribution of missing data.
- True distribution of missing data
  - Unobserved by definition
  - Always unknown
- Solution
  - Estimate the posterior distribution of missing data based on observed data, and make a random draw of imputed values.

Existing Software for Multiple Imputation

- R-Packages
  - Amelia II (EMB)
  - MICE (FCS)
  - NORM (MCMC)
- Commercial Software Programs
  - SAS Proc MI (MCMC/FCS)
  - SOLAS (FCS)
  - SPSS Missing Values (FCS)

None of them allows to perform multiple ratio imputation.

In the Literature

	Deterministic Single Imputation	Stochastic Single Imputation	Multiple Imputation
Regression Imputation	Exist	Exist	Exist
Ratio Imputation	Exist	Exist	Not Exist

National Statistics Center NSTAC

Theory of Multiple Ratio Imputation

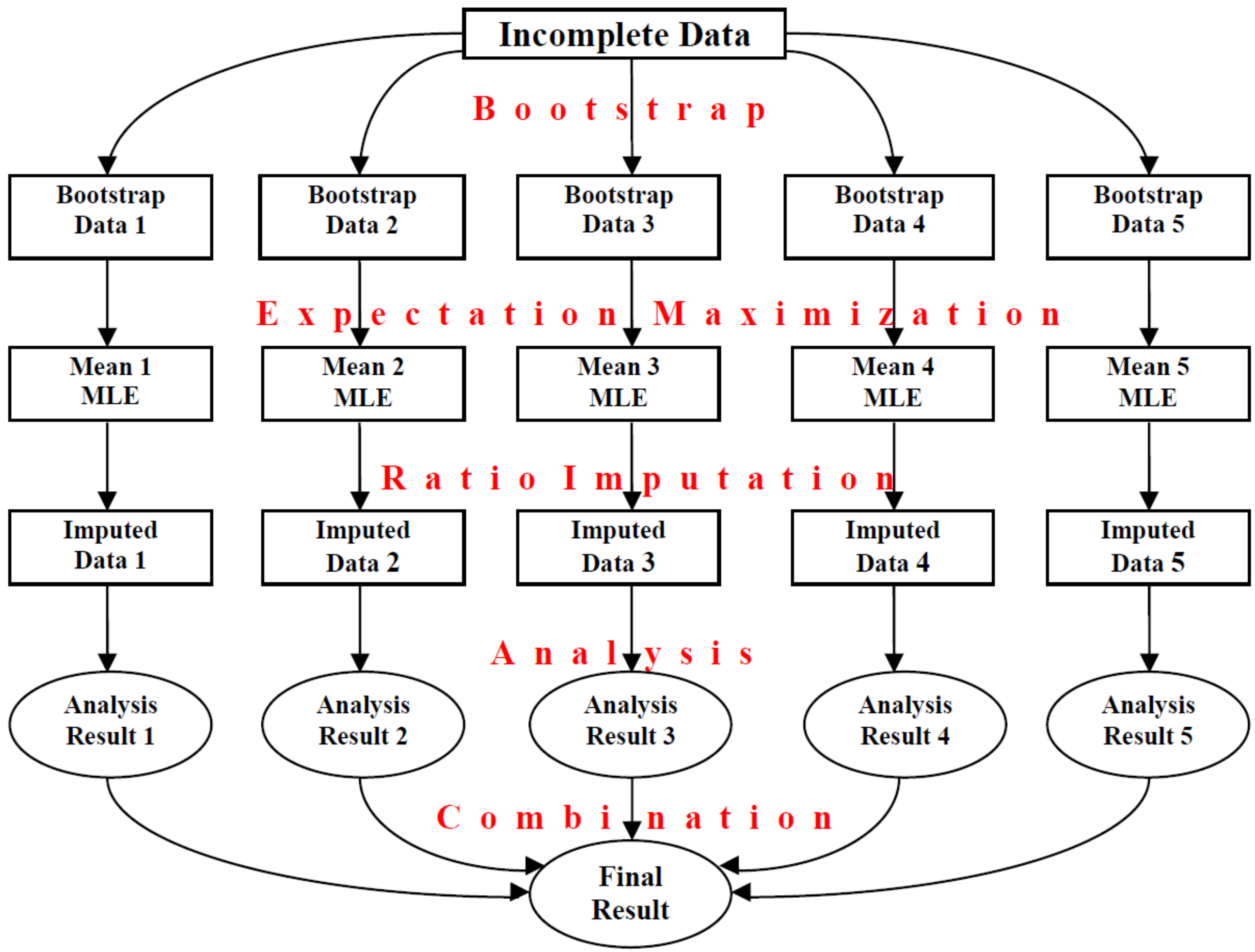
Multiple Ratio Imputation

- Literature
  - Devoid of multiple ratio imputation
- This paper
  - Proposes a novel application of the Expectation-Maximization with Bootstrapping (EMB) algorithm to ratio imputation
  - Proposes multiple ratio imputation

Multiple Ratio Imputation

- Value of ω
  - Estimated by $\widehat{\omega} = \bar{Y}_{1,obs}/\bar{Y}_{2,obs}$
- To create multiple ratio imputation
  - The mean vector is what needs to be randomly drawn from the posterior distribution of missing data given observed data.

Multiple Ratio Imputation by the EMB Algorithm



Monte Carlo Evidence

Monte Carlo Settings 1

- 1,000 iterations
- Random draw from the following multivariate normal distribution:
  - Variables y1 and y2 are normally distributed with the mean vector (6, 10) and the standard deviation vector (1, 1).
  - The correlation between y1 and y2 is set to 0.6.

National Statistics Center
NSTAC

Monte Carlo Settings 2

- Sample Size
  - n = 50, n =100, n =200, n =500, and n =1,000
- Three data generation processes
  - MCAR, MAR, and NI
- Average missing rates
  - 15%, 25%, and 35%

Monte Carlo Settings 3

- RRMSE: Relative Root Mean Square Errors
  - Mean
  - Standard Deviation
  - *t*-statistics in regression
- Comparisons of
  - Deterministic ratio imputation
  - Stochastic ratio imputation
  - Regular multiple imputation (Amelia II)
  - Multiple ratio imputation

Monte Carlo Evidence: Mean

Table 6. RRMSE Comparisons for the Mean (45,000 Datasets)

Sample Size	Average Missing Rate	Missing Mechanism	Listwise Deletion	Deterministic Ratio Imputation	Multiple Ratio Imputation
50	15%	MCAR	0.009	0.008	0.008
		MAR	0.017	0.008	0.008
		NI	0.026	0.017	0.018
	25%	MCAR	0.014	0.011	0.011
		MAR	0.030	0.010	0.011
		NI	0.048	0.032	0.033
	35%	MCAR	0.017	0.014	0.014
		MAR	0.045	0.012	0.014
		NI	0.075	0.050	0.052
100	15%	MCAR	0.007	0.006	0.006
		MAR	0.016	0.005	0.005
		NI	0.024	0.016	0.016
	25%	MCAR	0.010	0.008	0.008
		MAR	0.028	0.007	0.008
		NI	0.046	0.030	0.030
	35%	MCAR	0.012	0.010	0.010
		MAR	0.044	0.008	0.010
		NI	0.073	0.048	0.050
200	15%	MCAR	0.005	0.004	0.004
		MAR	0.015	0.004	0.004
		NI	0.024	0.016	0.016
	25%	MCAR	0.007	0.005	0.005
		MAR	0.028	0.005	0.005
		NI	0.045	0.029	0.030
	35%	MCAR	0.009	0.007	0.007
		MAR	0.043	0.006	0.007
		NI	0.072	0.048	0.049
500	15%	MCAR	0.003	0.003	0.003
		MAR	0.014	0.002	0.002
		NI	0.024	0.015	0.015
	25%	MCAR	0.004	0.003	0.003
		MAR	0.027	0.003	0.003
		NI	0.045	0.029	0.029
	35%	MCAR	0.006	0.004	0.004
		MAR	0.043	0.004	0.005
		NI	0.072	0.047	0.048
1000	15%	MCAR	0.002	0.002	0.002
		MAR	0.014	0.002	0.002
		NI	0.024	0.015	0.015
	25%	MCAR	0.003	0.003	0.003
		MAR	0.027	0.002	0.002
		NI	0.044	0.029	0.029
	35%	MCAR	0.004	0.003	0.003
		MAR	0.043	0.002	0.003
		NI	0.072	0.047	0.048

Note. Average over the 1,000 simulations for each data type. $M = 100$ for multiple ratio imputation

National Statistics Center
NSTAC

Monte Carlo Evidence: Mean

- In all of the 45 patterns, deterministic ratio imputation and multiple imputation both outperform listwise deletion.
- Between the ratio imputation methods, deterministic ratio imputation slightly performs better than multiple ratio imputation in 32 out of the 45 patterns with 13 ties.
- However, 43 out of the 45 patterns are within a 0.01-point difference in terms of the RRMSE.

National Statistics Center

NSTAC

Monte Carlo Evidence: Mean

Table 7. Mean of y1 (MAR-35%)

	Complete Data	Listwise Deletion	Deterministic Ratio Imputation	Multiple Ratio Imputation
Mean	6.000	5.741	6.000	5.999
BISD	NA	NA	NA	0.029
CI (95%)	NA	NA	NA	5.941, 6.057
n	500	325	500	500

Note. NA means Not-Applicable. Average over the 1,000 simulations. $M = 100$ for multiple ratio imputation

Monte Carlo Evidence: Standard Deviation

Table 8. RRMSE Comparisons for the Standard Deviation (45,000 Datasets)

Sample Size	Average Missing Rate	Missing Mechanism	Listwise Deletion	Stochastic Ratio Imputation	Multiple Ratio Imputation
50	15%	MCAR	0.042	0.048	0.037
		MAR	0.045	0.047	0.038
		NI	0.048	0.052	0.043
	25%	MCAR	0.059	0.062	0.049
		MAR	0.066	0.062	0.054
		NI	0.079	0.074	0.067
	35%	MCAR	0.075	0.075	0.058
		MAR	0.088	0.071	0.067
		NI	0.146	0.117	0.118
100	15%	MCAR	0.029	0.035	0.026
		MAR	0.031	0.034	0.026
		NI	0.035	0.037	0.031
	25%	MCAR	0.040	0.044	0.033
		MAR	0.046	0.044	0.037
		NI	0.064	0.058	0.054
	35%	MCAR	0.052	0.052	0.040
		MAR	0.067	0.054	0.047
		NI	0.121	0.097	0.098
200	15%	MCAR	0.021	0.025	0.018
		MAR	0.022	0.025	0.019
		NI	0.025	0.027	0.023
	25%	MCAR	0.028	0.030	0.023
		MAR	0.036	0.032	0.027
		NI	0.049	0.044	0.042
	35%	MCAR	0.037	0.037	0.028
		MAR	0.053	0.038	0.034
		NI	0.109	0.086	0.088
500	15%	MCAR	0.014	0.016	0.012
		MAR	0.014	0.016	0.012
		NI	0.018	0.019	0.016
	25%	MCAR	0.018	0.020	0.015
		MAR	0.024	0.020	0.017
		NI	0.042	0.038	0.036
	35%	MCAR	0.022	0.023	0.018
		MAR	0.043	0.024	0.021
		NI	0.106	0.083	0.084
1000	15%	MCAR	0.010	0.012	0.008
		MAR	0.010	0.011	0.008
		NI	0.014	0.015	0.013
	25%	MCAR	0.013	0.014	0.011
		MAR	0.019	0.014	0.011
		NI	0.040	0.037	0.033
	35%	MCAR	0.017	0.017	0.013
		MAR	0.038	0.016	0.014
		NI	0.100	0.080	0.079

Note. Average over the 1,000 simulations for each data type. $M = 100$ for multiple ratio imputation

National Statistics Center
NSTAC

Monte Carlo Evidence: Standard Deviation

- In all of the 45 patterns, multiple ratio imputation always outperforms listwise deletion.
- Between the ratio imputation methods, multiple ratio imputation often performs better than stochastic ratio imputation, 43 out of the 45 patterns.
- Therefore, this study contends that multiple ratio imputation is the preferred method for the estimation of the standard deviation.

Monte Carlo Evidence: *t*-statistics in Regression

Table 9. RRMSE Comparisons for *t*-statistics (45,000 Datasets)

Sample Size	Average Missing Rate	Missing Mechanism	Listwise Deletion	Multiple Imputation Amelia II	Multiple Ratio Imputation
50	15%	MCAR	0.126	0.103	0.087
		MAR	0.137	0.107	0.093
		NI	0.141	0.114	0.099
	25%	MCAR	0.185	0.144	0.113
		MAR	0.220	0.173	0.135
		NI	0.222	0.175	0.138
	35%	MCAR	0.242	0.189	0.134
		MAR	0.317	0.247	0.171
		NI	0.328	0.269	0.179
100	15%	MCAR	0.104	0.075	0.066
		MAR	0.113	0.080	0.071
		NI	0.111	0.081	0.072
	25%	MCAR	0.159	0.109	0.087
		MAR	0.192	0.127	0.101
		NI	0.194	0.136	0.108
	35%	MCAR	0.218	0.153	0.107
		MAR	0.294	0.191	0.131
		NI	0.297	0.224	0.147
200	15%	MCAR	0.091	0.059	0.052
		MAR	0.101	0.064	0.056
		NI	0.101	0.066	0.060
	25%	MCAR	0.145	0.092	0.075
		MAR	0.181	0.106	0.085
		NI	0.177	0.117	0.095
	35%	MCAR	0.208	0.136	0.097
		MAR	0.282	0.159	0.113
		NI	0.282	0.199	0.133
500	15%	MCAR	0.084	0.050	0.044
		MAR	0.094	0.053	0.047
		NI	0.093	0.058	0.051
	25%	MCAR	0.141	0.086	0.066
		MAR	0.171	0.092	0.069
		NI	0.170	0.107	0.083
	35%	MCAR	0.202	0.127	0.086
		MAR	0.279	0.144	0.097
		NI	0.282	0.193	0.121
1000	15%	MCAR	0.080	0.046	0.041
		MAR	0.089	0.046	0.043
		NI	0.091	0.048	0.049
	25%	MCAR	0.137	0.053	0.063
		MAR	0.167	0.084	0.067
		NI	0.168	0.105	0.083
	35%	MCAR	0.198	0.122	0.084
		MAR	0.275	0.132	0.092
		NI	0.275	0.186	0.120

Note. Average over the 1,000 simulations for each data type. $M = 100$ for multiple imputation

National Statistics Center
NSTAC

Monte Carlo Evidence: *t*-statistics in Regression

- In all of the 45 patterns, regular multiple imputation and multiple ratio imputation both outperform listwise deletion.
- Multiple ratio imputation always outperforms regular multiple imputation under the condition where the true population model satisfies the assumption of ratio estimation.

Summary of the Findings

	Mean	Std. Dev.	t-Stats
Listwise Deletion	Poor	Poor	Poor
Existing Method	Excellent	Fair	Fair
Multiple Ratio Imputation	Excellent	Excellent	Excellent

National Statistics Center NSTAC

Empirical Example

Application to Japanese Economic Census

- Data: 2012 Economic Census
  - Division I (Wholesale and Retail Trade)
  - Tokyo
- Target variable for imputation: Turnover
- Quantity of interest: Mean of turnover
- Auxiliary variable: Cost
- Our data
  - Focus on the establishments and enterprises with the number of employees equal to 1.

Results

	Listwise Deletion	Deterministic Ratio Imputation	Multiple Ratio Imputation
Mean	3569.12	3526.73	3526.69
BISD	NA	NA	4.74
CI (95%)	NA	NA	3517.21, 3536.16

Note: BISD = Between-Imputation Standard Deviation

Thank you

National Statistics Center

NSTAC