Repeated Measures Data Analysis With Missing Values: An Overview of Methods

By

K. C. Carriere, T.S. Park, Y. Liang

University of Alberta and Seoul National University

OUTLINE

- Introduction
- Missing Data Mechanisms
- Available Data Analysis
- Imputations
- Numerical Examples
- Conclusion

Introduction

Objectives in Missing Data Analysis

- Remove Bias
- Reduce Variance
- Improve Efficiency

Missing Data Mechanism

- MCAR-- Missing cases are a random sample of observed cases. No danger of biased estimation
- MAR-- Cases with incomplete data are different from cases with complete data. LR method leads to consistent estimation.
- NIM—Reason for missing data is explainable but unmeasurable. Special attention needed.

General Approaches

- Explicit variance formulas that allow for nonresponse
- Available case analysis
- Resampling and/or imputation
- Single/Multiple imputation

Imputation Approaches

- Single imputation method
  - Generally underestimate the variability of the missing data
  - Produce simple estimators
- Multiple imputation method (Rubin, 1987)
  - Intensive computation
  - Large memory space for storing multiply-imputed data
  - No work on small-sample repeated measure data

Available Case/Data Analyses

- Maximum likelihood methods (Carriere 1994 and 1999)
- Jackknife and bootstrap methods (Miller 1974; Efron 1994)
- Data augmentation (Tanner and Wong 1987)
- Gibbs sampler (Gelfand and Smith 1990; Gelman and Rubin 1992)

Available Data Analysis

- Available case analysis - Generally lack practical appeal due to imbalance of sample bases
- Available data analysis – Almost ML method with large sample theory
- Approximate solutions (SAS, SPSS, etc)
- Limited in scope

Multiple Imputation Theory

- Draw missing values from posterior distribution $f(\mathbf{y}_{mis} \mid \mathbf{y}_{obs})$
- Posterior of the parameter of interest $\theta_{q\times 1}$

$$\theta = \int g(\theta \mid \mathbf{y}_{obs}, \mathbf{y}_{mis}) f(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}) d\mathbf{y}_{mis}$$

- Imputed data set $\mathbf{y}^{(i)} = (\mathbf{y}_{obs}, \mathbf{y}_{mis}^{(i)}), i = 1, \ldots, M$
- Estimators and associated variances

$\hat{\theta}_{(i)}$ and $\mathbf{U}_{(i)},\ i = 1, \ldots, M$

Multiple Imputation Theory

$$\bar{\theta}_M = \sum_{i=1}^{M} \hat{\theta}_{(i)} / M \qquad (1)$$

$$V(\bar{\theta}_M) = \mathbf{T}_M = \bar{\mathbf{U}}_M + (1 + M^{-1})\mathbf{B}_M \qquad (2)$$

where

$$\bar{\mathbf{U}}_M = \sum_{i=1}^{M} \mathbf{U}_{(i)} / M$$

$$\mathbf{B}_M = \sum_{i=1}^{M} (\hat{\theta}_{(i)} - \bar{\theta}_M)(\hat{\theta}_{(i)} - \bar{\theta}_M)^T / (M-1)$$

Multiple Imputation Theory

- Consider a linear transformation

$$\eta = \boldsymbol{l}^T \theta$$

- Approximate distribution

$$(\eta - \bar{\eta}_M)[\boldsymbol{l}^T \mathbf{T}_M \boldsymbol{l}]^{-1/2} \sim t_v$$

where $\bar{\eta}_M = \boldsymbol{l}^T \bar{\theta}_M$

Multiple Imputation Theory

- Degree of freedom
  - Rubin 1987: $v = (M-1)r_M^{-2}$

$$r_M = (1+M^{-1})tr(\mathbf{B}_M \mathbf{T}_M^{-1})/q$$

  - Rubin 1999: $\tilde{v} = v_0\{[f(v_0)(1-r_M)]^{-1} + \frac{v_0}{v}\}^{-1}$ (3)

$$f(v_0) = (v_0+1)/(v_0+3)$$

v_0 : df based on the complete data

RMD(t, p, s) Model

$$\mathbf{y} = \mu + \varepsilon = \mathrm{X}\beta + \varepsilon$$

$$\boldsymbol{X} = (\mathbf{1}_{N_1}^T \otimes \boldsymbol{X}_1^T, \ldots, \mathbf{1}_{N_s}^T \otimes \boldsymbol{X}_s^T)^T$$

$$\beta = (m_0,\ \pi^T,\ \tau^T,\ \gamma^T,\ \lambda^T)^T$$

Assumption

- Missing at random
- Monotonic missing pattern

Notation: (Carriere 1999)

$$N_k^{(l)} \qquad l = 1, \ldots, L$$

$$N^{(l)} = \sum_k N_k^{(l)}$$

$$\bar{y}_{i.k}^{(l)} = \sum_{j=1}^{N_k^{(l)}} y_{ijk} / N_k^{(l)}$$

Improper Imputation

- Valid if using proper imputation strategy
- Improper imputations can still be confidence-valid
- True even if some important predictors are left out of the strategy given that fraction of missing is not large.
- Simpler improper strategies.
- Proxy data from caregivers.

Imputation Procedures

Step 1: Use the LSE for mean and covariance matrix for

$$y_{p+1,jk} \mid \boldsymbol{y}_{(p)jk} \sim N(\hat{\mu}_{p+1,k}, \hat{\sigma}^2)$$

For the usual conditional mean and conditional variance.

Imputation Procedues

- Step 2: Draw a chi-square random variable g with degrees of freedom $N^{(l)} - s$

  Let $\sigma^* = \hat{\sigma}(N^{(l)} - s)/g$

- Step 3: Draw a random variable Z from a standard normal distribution and let

$$\mu^*_{p_1+1,k} = \hat{\mu}_{p_1+1,k} + \sigma^* z / \sqrt{N_k^{(l)}}$$

Imputation Procedures

- Step 4: Draw a random variable z from a standard normal distribution, and impute for the missing values in the period $p_1 + 1$

  $$y_{p_1+1, jk} = \mu^*_{p_1+1,k} + \sigma^* z$$

  Repeat Step 4 for all missing components in period $p_1 + 1$
- Step 5: Treat the imputed values as if they were actual values and repeat Steps 1-4 for the next periods, with p_1 replaced by $p_1 + 1$

Imputation Procedures

- Step 6: Repeat Steps 1-5, M times to create M multiply-imputed data sets.
  - Substantial empirical work (for example, Rubin 1998) has shown that multiple imputation with $M = 3$ or 5 works well with typical fractions (<30%) of missing data in surveys.

Comparison

- Degrees of freedom
  - Multiple imputation method

$$\tilde{v} = v_0\{[f(v_0)(1-r_M)]^{-1} + \frac{v_0}{v}\}^{-1}$$

  - Carriere (1994 and 1999)
    - Compound symmetric (SYS)

      $(p-1)(N^{(L)}-s)$ for both τ and γ
    - Unspecified (UNS)

      $N^{(L)}-s$ for τ

      $(N+N^{(2)}-2s-p_1p_2)/2$ for γ

	ρ	c_1	Method	Type	$\alpha = 0.01$ size	power	$\alpha = 0.05$ size	power	$\alpha = 0.1$ size	power
γ	.3	1	INC	sys	.011	.025	.051	.121	.104	.229
				uns	.013	.028	.051	.135	.114	.223
			MI	sys	.007	.032	.037	.130	.084	.229
				uns	.013	.021	.044	.124	.109	.204
		4	INC	sys	.014	.009	.054	.082	.114	.147
				uns	.016	.010	.057	.066	.112	.148
			MI	sys	.013	.010	.060	.058	.109	.152
				uns	.013	.007	.051	.075	.104	.146
	.7	1	INC	sys	.011	.017	.047	.106	.094	.199
				uns	.008	.023	.042	.116	.092	.189
			MI	sys	.007	.019	.036	.112	.089	.181
				uns	.007	.026	.038	.116	.090	.188
		4	INC	sys	.011	.016	.051	.067	.112	.125
				uns	.006	.021	.048	.074	.100	.131
			MI	sys	.008	.023	.048	.080	.110	.124
				uns	.004	.027	.036	.083	.086	.144
τ	.3	1	INC	sys	.011	.182	.059	.509	.102	.691
				uns	.010	.189	.056	.497	.099	.701
			MI	sys	.008	.201	.045	.511	.091	.693
				uns	.010	.194	.050	.498	.094	.696
		4	INC	sys	.014	.078	.061	.235	.116	.367
				uns	.011	.075	.060	.232	.107	.351
			MI	sys	.010	.082	.054	.246	.116	.352
				uns	.007	.098	.054	.211	.115	.332
	.7	1	INC	sys	.008	.558	.050	.871	.094	.951
				uns	.010	.541	.051	.865	.092	.946
			MI	sys	.006	.506	.047	.831	.087	.927
				uns	.011	.470	.045	.826	.089	.923
		4	INC	sys	.004	.185	.061	.333	.115	.505
				uns	.004	.170	.058	.348	.111	.490
			MI	sys	.005	.139	.055	.342	.105	.499
				uns	.005	.167	.050	.326	.097	.485

Note: Based on 1000 simulations. Design I includes AB and BA sequences. MI– Multiple imputation approach; INC – incomplete data procedure (18, 19); SYS – compound symmetry covariance structure; UNS – unspecified covariance pattern.

ρ	c_1 c_2	Method	Type	$\alpha = 0.01$ size	power	$\alpha = 0.05$ size	power	$\alpha = 0.1$ size	power
.3	1	INC	sys	.011	.106	.056	.257	.106	.408
	1		uns	.006	.091	.050	.254	.116	.331
		MI	sys	.007	.100	.037	.255	.111	.383
			uns	.004	.100	.051	.232	.100	.372
	1	INC	sys	.007	.059	.043	.178	.088	.274
	4		uns	.009	.059	.051	.173	.104	.277
		MI	sys	.007	.056	.045	.169	.091	.269
			uns	.009	.054	.049	.191	.106	.281
	4	INC	sys	**.003**	.067	**.029**	.183	**.063**	.297
	1		uns	.010	.051	.051	.152	.106	.263
		MI	sys	**.002**	.065	**.026**	.156	**.064**	.275
			uns	.009	.065	.051	.153	.110	.263
.7	1	INC	sys	.004	.323	.039	.585	.090	.717
	1		uns	.008	.215	.045	.499	.088	.692
		MI	sys	.009	.279	.036	.554	.086	.694
			uns	.005	.235	.042	.516	.083	.663
	1	INC	sys	.008	.105	.036	.280	**.072**	.431
	4		uns	.008	.076	.045	.293	.094	.415
		MI	sys	.007	.099	.043	.257	.082	.406
			uns	.007	.088	.053	.253	.111	.371
	4	INC	sys	**.003**	.110	**.026**	.326	**.055**	.458
	1		uns	.005	.104	.045	.286	.082	.455
		MI	sys	**.002**	.120	**.022**	.304	**.053**	.432
			uns	.007	.106	.055	.259	.104	.403

Note: See notes for Table 1. Design IV includes sequences ABB and BAA.

Available Data or Imputation?

- multiple imputation simple and easy to implement and no special software is required
- As long as all available data are used, all approaches are satisfactory
- generally the multiple imputation methods not superior to the alternative non-imputation ML methods in terms of power of testing hypotheses of parameters of interest

Numerical Example

- Analysis of Bronchial Asthma Data
- Traditional two-period two-sequence two-treatment crossover design
- AB group with 8 subjects
- BA group with 9 subjects
- Goal: estimate the contrast of treatment effect (A-B)

Numerical Example

- Induced missing data in second period from BA group (MAR)
- Proxy estimation:
  - I. smaller value in period 2 than in period 1
  - II. Slight overestimation of actual values
  - Both with no bias and similar in variability
- Multiple imputation

Numerical Example

Table 1: Analysis Results of the Bronchial Asthma Data

Method	$\hat{\tau}$	se($\hat{\tau}$)	df	P-value	$\hat{\gamma}$	se($\hat{\gamma}$)	df	P-value
full original data	-0.384	0.169	15	0.038	-0.512	0.315	15	0.125
complete subset data	-0.404	0.178	11	0.044	-0.503	0.323	11	0.148
incomplete method[1]	-0.384	0.152	11	0.028	-0.471	0.285	11	0.127
incomplete method[2]	-0.384	0.163	10	0.040	-0.470	0.309	10	0.159
proxyI[3]	-0.384	0.177	13	0.049	-0.468	0.340	13	0.193
proxyII[3]	-0.384	0.180	13	0.053	-0.468	0.348	13	0.206
proxyI[4]	-0.384	0.166	13	0.038	-0.468	0.319	13	0.166
proxyII[4]	-0.384	0.169	13	0.041	-0.468	0.326	13	0.174
MI[5]	-0.384	0.164	9.429	0.043	-0.554	0.333	6.554	0.143

Note: 1. Method by Carriere[1]; 2. Method by PROC MIXED of SAS; 3. Method by Huang et al.[23]; 4. Method by PROC MIXED of SAS with df adjustment of Huang et al.[23]; 5. Multiple Imputation method by Huang and Carriere.[27]

Numerical Example

- Original data—
  - Residual effect marginally significant
  - Treatment effect is significant.
- Complete subset data
  - Similar to the original data results
  - Higher standard errors for estimators
- Incomplete Data
  - Similar to the original data results.
  - Power is improved

Numerical Example

- Proxy Information
  - Residual effects not significant
  - Less sensitive test of treatment effect
- Multiple Imputation
  - High penalty
  - Qualitatively similar to the original data results.

Acknowledgements

- Natural Sciences and Engineering Research Council of Canada
- Alberta Heritage Foundation for Medical Research
- Korean Federation of Science and Technology Societies (Brain Pool Program).

Thank you !