Strategies for Handling Missing Outcomes in Longitudinal Questionnaire Data

Nazanin Nooraee

Post-Doctoral Fellow in Statistics

4th International Conference on Biometrics & Biostatistics
November 16 -18, 2015
San Antonio, USA

Technische Universiteit
Eindhoven
University of Technology

Content

- **Background**
- **Missing data methods for questionnaire**
- **Simulation**
- **Results**
- **Conclusion**

Background: Importance of Missing Data Methods

- **Missingness is a pervasive issue in research**
  - Mistake in input information: typo,…
  - Drop out subjects before completing longitudinal survey,
  - Unanswered questions in questionnaire
- **Potential source of bias**
- **Reduction of power due to reduced sample size**

Types of Missingness

- **Missing Completely At Random (MCAR) :**

  The probability that an item is missing is related to neither predictors nor other items.

- **Missing At Random (MAR):**

  The probability that an item is missing is related only to the observed data.

- **Missing Not At Random (MNAR):**

  The probability that an item is missing is related to the (unknown) value of the unobserved data and observed data

Methods for Handling Missing Data in Longitudinal Settings

- **Complete case analysis**
- **Imputation**
  - **Single imputation:**
    - Mean imputation,
    - Individual imputation, …
  - **Multiple Imputation (MI):**
    - Joint Modeling, mainly with multivariate normal distribution(MVN)
    - Fully Conditional Specification (FCS)
  - **Ad hoc methods**
    - Predictive mean matching (PMM)
- **Maximum likelihood inference**
- **Advanced statistical methods:**
  - Selection models
  - Pattern mixture models

Advantages and Limitations

Method		Advantage	Disadvantages
Complete case analysis		Simple to apply	Reduce sample size, information loss, influence on precision and power
		Valid under MCAR	Biased results under MAR and MNAR
Imputation	**Single imputation**	Unbiased under MCAR and MAR	Underestimate standard error
	Multiple imputation	Uses all available data	Requires some decisions (each involves uncertainty) apply JM or FCS, which technique of FCS, how many imputed data, how many iterations is sufficient
		Incorporate auxiliary variables	A imputed setting cannot be used for different analysis
	PMM	Retains the distribution of variables, robust to transformation, less sensitive to mis-specification of the model	Lack of (mathematical) theory
Maximum likelihood		For a given data set, always gives the same results	Commonly cannot incorporate auxiliary variables
		Default of mixed models	Cannot handle missing covariates

Applied Statistical Methods in Our Study

- **Analyze sum score of the items using marginal models**
- **Missing data methods**
  - **Multiple imputation**
    - At item level
      - Logistic regression imputation (LR_{item})
      - PMM (PMM_{item})
    - At scale level
      - Multivariate normal imputation using MCMC, with an addition constraint ($MCMC_{scale}$)
      - PMM, with an addition constraint (PMM_{scale})
  - **Maximum likelihood**
    - All items are missing (ML_{10})
    - At least one item was missing (ML_1)

Combine Advantage of ML and MI: A Hybrid Approach

- **Suggestion of Von Hipple** [1] **and White et al.** [2]
  - **Including imputed outcomes adds noise to the parameter estimates in the final analysis**

- **We proposed this approach for questionnaire survey (hybrid approach)**
  - **With all the applied imputation methods**

[1] Von Hippel, P. T. (2007). Regression with missing y's: an improved strategy for analyzing multiply imputed data. *Sociological Methodology*, 37(1):83–117.
[2] White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30:377–399.

Simulation: Full Data Set

- **Simulate 10 items at 4 follow-ups**
  - **Generate covariates (X_{it})**
    - A binary variable representing gender
    - 4-dimensional variables representing age
    - 3 binary and 1 continuous (correlated) covariates
  - **Generate items**
    - 4-dimensional normally distributed random variables as correlated ($Z_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})$)
    - Set parameters for difficulties (a_{tj}) and discriminations (b_{tj})
    - Create binary items with success probability

$$\pi_{it}(j) = \frac{exp(a_{tj} + b_{tj} Z_{it} + X'_{it} c_{tj})}{1 + exp(a_{tj} + b_{tj} Z_{it} + X'_{it} c_{tj})}$$

Simulation: Missingness Framework

- **Full observation at baseline**
- **Create missingness indicator variable**
  - with success probability for $t > 2$ for each item j

$$\tilde{\pi}_{it}(j) = \frac{exp(\tilde{a}_{tj} + \tilde{b}_{tj} Z_{it} + X'_{it} \tilde{c}_{tj})}{1 + exp(\tilde{a}_{tj} + \tilde{b}_{tj} Z_{it} + X'_{it} \tilde{c}_{tj})}$$

  - If the indicator is 0, then item j at time t is removed from the full data set
- **Different values for $\tilde{a}$'s, $\tilde{b}$'s, and $\tilde{c}$'s lead to different proportion of intermittent missing items**
- **$\tilde{\pi}_{it}$ can be simulated dependent or independent (Subject missing and item missing)**

Simulation

- **Set up**
  - Each simulate data set contains 1000 individuals
  - Generate small, medium, and large proportions of missing items and subjects (if applicable)
  - Each incomplete data set was imputed 10 times
  - Repeat each simulation 500 times
- **Analysis conducted in SAS**
  - Each data set was analyzed with marginal (population-average) models via Proc MIXED
  - Proc MI was applied for multiple imputation
  - Proc MIANALYZE for pooling the analysis results from imputed data sets
- **Comparison criteria: bias and mean square error**

Proportion of Missingness

Missingness indicator	Proportion of missingness	Unit missing			Item missing		
		Visit 2	Visit 3	Visit 4	Visit 2	Visit 3	Visit 4
Independent	Small	0.004	0.72	1.46	4.19	13.95	10.15
	Medium	0.000	1.75	8.80	7.20	26.23	24.04
Dependent	Medium	4.95	23.43	23.56	7.26	26.33	24.06
	Large	8.93	45.73	37.06	12.75	53.12	44.03

Results: Bias in Fixed Effects

	ML_{10}	ML_1	PMM_{item}	LR_{item}	PMM_{scale}	$MCMC_{scale}$	$H\text{-}PMM_{item}$	$H\text{-}LR_{item}$	$H\text{-}PMM_{scale}$	$H\text{-}MCMC_{scale}$
β_0	-1.38	-1.20	-0.47	-2.03	-1.08	-0.90	-0.93	-1.22	-1.15	-1.25
β_1	0.13	0.12	0.04	0.19	0.09	0.08	0.09	0.12	0.11	0.12
β_2	0.002	0.00	0.00	0.001	0.00	0.00	0.00	0.00	0.00	0.001
β_3	-0.06	-0.12	0.08	-0.05	0.07	0.004	-0.06	-0.09	-0.07	-0.05
β_4	0.08	0.14	-0.23	-0.21	-0.09	-0.008	0.03	0.04	0.07	0.06
β_5	-0.17	-0.32	0.04	-0.43	-0.06	-0.11	-0.18	-0.27	-0.24	-0.27
β_6	0.12	0.01	0.13	0.27	0.23	0.14	0.02	0.02	0.07	0.22

Results: Bias in Variance Components

	ML_{10}	ML_1	PMM_{item}	LR_{item}	PMM_{scale}	$MCMC_{scale}$	H-PMM_{item}	H-LR_{item}	H-PMM_{scale}	H-$MCMC_{scale}$
σ_1^2	0.36	0.48	-0.40	-0.63	-0.74	-0.84	0.22	0.22	0.01	0.06
σ_2^2	-22.67	-13.63	5.85	-2.08	2.35	4.30	-4.49	-7.02	-3.18	-6.39
σ_3^2	-108.05	-49.39	19.64	-44.6	-3.40	-16.49	-23.32	-32.08	-26.41	-31.77
σ_4^2	-69.91	-30.69	13.22	-61.5	-4.07	-5.33	-14.22	-30.73	-18.07	-20.43

Results: Comparison Biases

- **Largest bias: Imputation at scales, imputation at item level using logistic regression**
- **"Wilcoxon signed rank" test showed significate bias for most of the parameter estimates**

- **Among four other methods: no clear pattern to choose the best method**
  - Maximum likelihood provide somewhat smaller biases, for the follow-up times parameters
  - For correlation coefficients: ML_{10} performs best in presence of large proportion of missingness while H-PMM_{item} does best for all other settings.

Results: Comparison MSEs

- **The hybrid method at item level outperform their original imputation on almost all parameters, though differences are never very large.**

- **H-PMM$_{item}$ performs generally best on almost all fixed effects parameters and on the correlation parameters**

- **When it is outperformed by another method for a specific parameter, the hybrid method is still close to the other method**

Conclusion

- **Results showed that MI at item level outperforms imputation at scale level, (consistent with findings in cross-sectional studies)**

- **Hybrid approach with PMM at item level revealed smaller MSE, however the differences were not substantial**

TU/e Technische Universiteit Eindhoven University of Technology

Thank you for your attention

Where innovation starts