Statistical Analysis of Incomplete Longitudinal Data Under Different Missing Scenarios

Panpan Zhang and Sharon X. Xie
University of Pennsylvania

Motivation of the Study

Missing data mechanisms:

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

Simplest methods dealing with missing data:

1. complete case analysis (CCA)
2. available case analysis (ACA)

For simple linear regression models, CCA/ACA sometimes may provide unbiased estimates:

- No missing data in X, and the missing data of Y are MAR (Little, 1992).
- Both missing data of X and Y are MAR, but do not depend on observed responses (Little and Rubin, 2002).

Motivated by White and Carlin (2010), we would like to assess the performance of ACA versus that of one of the most practical methods—**multiple imputation** (MI)—in longitudinal setting under a variety of missing data generation scenarios.

Multiple Imputation

Multiple imputation (MI, Rubin, 1987) has been one of the most welcoming methods for dealing with missing data problems in both academia and industry. The fundamental idea of MI is to draw more than one imputed values from the predictive distribution of the missing data reflecting uncertainty.
Popular MI approaches include:

1. Joint modeling (JM, Schafer, 1997)
2. Fully conditional specification (FCS, Van Buuren et al., 1999)
3. Nonparametric imputation, e.g., classification and regression tree (CART, Burgette and Reiter, 2010)
4. Multilevel imputation, e.g., PAN (Schafer and Yucel, 2002)

There have been quite a few well developed `R` packages allowing the users to implement different kinds of imputation methods, such as `mice`, `miceadds`, `jomo`, `pan`, etc.

Linear Mixed Effects Model

The *linear mixed-effects model* (LMM, Laird and Ware, 1982) is given by:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

- $\boldsymbol{Y}_i$: an $m \times 1$ vector of observations;
- $\boldsymbol{X}_i$: an $m \times p$ matrix of fixed-effects covariates;
- $\boldsymbol{\beta}_i$: a p-dimensional vector of regression coefficients;
- $\boldsymbol{Z}_i$: a known $m \times q$ design matrix;
- $\boldsymbol{b}_i$: a q-dimensional vector of random effects;
- $\boldsymbol{\epsilon}_i$: an m-dimensional vector of error terms.

Simulation Results

We consider missing data in

- longitudinal outcome Y_{ij}'s;
- time-invariant fixed covariate X_i's;
- both of Y_{ij}'s and X_i's.

Simulation setup:

- Within each simulation: $n = 400$ subjects and $m = 5$ time points;
- Number of simulation runs: $R = 1,000$.
- For each simulation, we apply Complete data analysis (CDA), ACA, FCS, CART and PAN.
- For each simulation run, we report point estimates (EST), percentage of bias (PB), standard error (SE), relative efficiency (RE) and coverage of probability (CP).

Scenario I Missing data scenario: $\boldsymbol{X}_i$ and/or $\boldsymbol{Y}_i$ under MCAR.

- Missing in $\boldsymbol{Y}_i$ ONLY. Unbiased: ACA, FCS and PAN; RE: ACA $\approx$ PAN $>$ FCS
- Missing in $\boldsymbol{X}_i$ or $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$. Unbiased: ACA and FCS; RE: FCS $>$ ACA

Recommendation:

- Missing in $\boldsymbol{Y}_i$ ONLY: ACA
- Missing in $\boldsymbol{X}_i$ or $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$: FCS

Scenario II Missing data scenario: $\boldsymbol{Y}_i$ under MAR; the missingness may depend on observed responses, fully observed covariates or both.

- Unbiased: ACA, FCS and PAN.
- RE: ACA $\approx$ PAN $>$ FCS (under all three settings)

Recommendation: ACA

Scenario III: $\boldsymbol{X}_i$ under MAR; the missingness may depend on observed responses, fully observed covariates or both.

- Missingness depends on covariates ONLY. Unbiased: ACA, FCS and PAN; RE: ACA $\approx$ FCS $>$ PAN
- Missingness depends on responses ONLY. Unbiased: ACA and FCS; RE: FCS $>$ ACA
- Missingness depends on both. Unbiased: ACA and FCS; RE: FCS $>$ ACA

Recommendation:

- Missingness depends on covariates ONLY: ACA
- Missingness depends on responses ONLY: FCS
- Missingness depends on both: FCS

Scenario IV

- $\boldsymbol{X}_i$ under MAR; the missingness only depends on other fully observes covariates.
- $\boldsymbol{Y}_i$ under MAR; the missingness may depend on observed responses or both observed covariates and responses.
- Unbiased: ACA and FCS
- RE: ACA $>$ FCS (under both combos)

Recommendation: ACA

Simulation Results (Cont'd)

Scenario V

- $\boldsymbol{X}_i$ under MAR; the missingness depends on both observed covariates and responses.
- $\boldsymbol{Y}_i$ under MAR; the missingness may depend on observed responses or both observed covariates and responses.
- Unbiased: FCS
- RE: FCS is the only method providing unbiased estimates.

Recommendation: FCS

PPMI data analysis

- Longitudinal response: Montreal Cognitive Assessment (moca, MAR)
- Temporal covariate: Yearly follow-up
- Time-invariant covariates: age (at baseline) and gender
- Two covariates of primary interest:
  - MRI volume in Frontal ROI (at baseline, MCAR)
  - MRI volume in Parietal ROI (at baseline, MCAR)

		Missing data methods			
Parameter	Name	ACA	FCS	CART	PAN
intercept	EST	30.917	30.549	30.534	30.951
	SE	1.138	0.863	0.826	0.824
time	EST	-0.454	-0.424	-0.409	-0.429
	SE	0.103	0.066	0.065	0.065
age	EST	-0.069	-0.068	-0.069	-0.078
	SE	0.017	0.014	0.012	0.023
gender	EST	0.563	0.489	0.545	0.653
	SE	0.329	0.247	0.233	0.238
Frontal ROI	EST	0.363	0.441	0.440	0.132
	SE	0.170	0.202	0.132	0.151

References

References

[1] Burgette, L.F. and Reiter, J.P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, **172**, 1070–1076.

[2] Little, R.J.A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, **87**, 1227–1237.

[3] Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

[4] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Hoboken, NJ, U.S.A.

[5] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Hoboken, NJ, U.S.A.

[6] Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London, U.K.

[7] Schafer, J.L. and Yucel, R.M. (2002). Computational strategies for multivariate mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, **11**, 437–457.

[8] Van Buuren, S., Boshuizen, H.C. and Knook, D.L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, **18**, 681–694 .

[9] White, I. and Carlin J. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, **29**, 2920–2931.