Robust Methods for Missing Covariates in Longitudinal Studies with Application to Biomarker Research in Parkinson's Disease Dementia

Leah H. Suttner, Panpan Zhang and Sharon X. Xie

University of Pennsylvania

Abstract

Objective: Develop a nonparametric method for handling different types of missing covariates in longitudinal studies.
Background: There are various reasons causing missing data in longitudinal studies (e.g., study design which only requires a subgroup of Parkinson's disease (PD) patients to take invasive biomarker tests for cerebrospinal fluid (CSF) amyloid β 1-42). More than 50% of the patients from the motivating dataset in this research missed the measurements of CSF $A\beta$ 1-42.
Method: Propose a new statistical method utilizing auxiliary variables that are related to the missing covariates without assuming any distributional assumptions. Different types of missing covariates have been considered. The proposed method has been compared with other conventional missing data methods like available case analysis (ACA) and multiple imputation (MI).
Results: The proposed method is
(1) consistently more efficient than ACA, and recovers a great deal of efficiency loss caused by missing data and almost recovers all efficiency loss when the correlation between missing covariates and auxiliary variables is high;
(2) more flexible than standard MI methods, as it performs well when the relation between missing covariates and auxiliary variables becomes extremely nonlinear.
Conclusion: The method proposed in our study provides a robust and efficient approach to dealing with different types of missing covariates in longitudinal studies, and is particularly attractive when the underlying relationship between missing covariates and auxiliary information is unknown.

Preliminaries

Notations:

- $\boldsymbol{Y}$: continuous outcomes with repeated measures;
- $\boldsymbol{X}$: covariates measured for a subsample;
- $\boldsymbol{Z}$: covariates measured for the entire sample;
- $\boldsymbol{Z}^*$: a subset of $\boldsymbol{Z}$, included in the analysis model;
- $\boldsymbol{S}$: a subset of $\boldsymbol{Z}$, containing information about $\boldsymbol{X}$.

The entire dataset has been divided into a validation set and a non-validation set.

- Validation set: consisting of the subjects whose $\boldsymbol{X}$ must be observed on the time points at which $\boldsymbol{Y}$ are observed;
- Non-validation set: consisting of the subjects whose $\boldsymbol{X}$'s are either completely missing, or for whom at least one $\boldsymbol{Y}$ corresponding to a missing $\boldsymbol{X}$ is observed.

	$\boldsymbol{Y}$			$\boldsymbol{X}$		
Time	1	2	3	1	2	3
Subj.1	O	O	O	O	O	O
Subj.2	O	O	M	O	O	O
Subj.3	O	M	M	O	M	M
Subj.4	O	O	O	M	O	M
Subj.5	O	M	M	M	M	M
Subj.6	O	M	O	O	M	M

Note: In the toy example above, O and M respectively represents "observed" and "missing"; Subjects 1, 2 and 3 belong to the validation set, whereas the rest belong to the non-validation set.

Statistical Method Outline

Assumptions:

- All missing data are missing at random (MAR);
- Missing $\boldsymbol{X}$ does not depend on auxiliary variable $\boldsymbol{S}$, but may depend on $\boldsymbol{Z}^*$ or observed outcome $\boldsymbol{Y}$.
- The analysis model is linear mixed effects (LME) model.

Approach:

1. Divide the full likelihood into a likelihood for the validation set V and a likelihood for the non-validation set $\bar{V}$.
2. For each $j \in \bar{V}$, assign values to missing $\boldsymbol{X}_j$ through the corresponding (observed) $\boldsymbol{X}_i$ for all eligible $i \in V$, where the likelihood of each assignment is evaluated empirically by using auxiliary variable.
   (1) For discrete $\boldsymbol{S}$, the empirical probability estimate uses the proportion of matchings;
   (2) For continuous $\boldsymbol{S}$, the empirical probability estimate uses a kernel density.

Additional Notes:

1. Throughout the analysis, we focus on missing covariates **only**.
2. The method requires **sufficient** candidates in the validation set for "mathcing" or "kernel estimation".

Appealing properties:

1. The maximum likelihood estimator (MLE) of the estimated likelihood is **consistent**.
2. The asymptotic distribution of MLE is **normal**.

Simulations

Simulated data are generated from LME, with missing proportion controlled around 50%. There are **three** correlation levels for missing covariates and auxiliary variables: 0.50 (moderate), 0.75 (high) and 0.99 (extremely high).

Simulation set 1: time-independent, discrete $\boldsymbol{X}$ with time-independent, discrete $\boldsymbol{S}$; the relationship between $\boldsymbol{X}$ and $\boldsymbol{S}$ is nonlinear.
Results:

- Both the proposed method and ACA are unbiased, but the proposed method is obviously more efficient than ACA.
- Standard MI is biased.

Simulation set 2: time-varying, continuous $\boldsymbol{X}$ with time-independent, continuous $\boldsymbol{S}$; the relationship between $\boldsymbol{X}$ and $\boldsymbol{S}$ is linear.
Results:

- The proposed method, ACA and MI are all unbiased.
- The proposed method is obviously more efficient than ACA, and slightly more efficient than standard MI.

Simulation set 3: time-varying, continuous $\boldsymbol{X}$ with time-varying, continuous $\boldsymbol{S}$; the relationship between $\boldsymbol{X}$ and $\boldsymbol{S}$ is linear.
Results:

- The proposed method, ACA and MI are all unbiased.
- The proposed method is obviously more efficient than ACA, and slightly more efficient than standard MI.

UPenn Parkinson's Disease Research Center Cohort

- Contains 408 PD patients taking cognitive assessments on an annual base for the first four years, and biennially thereafter.
- Longitudinal variable of interest: age-adjested dementia rating scale total (DRStotalAge).
- Covariates with missing data: CSF $A\beta$ discretized at cutoff 192ng/L.
- Auxiliary variable: apolipoprotein E (APOE) genotype discretized to APOE4, reflecting the number of APOE $\varepsilon 4$ allels.
- Missing proportion: 58%.
- We did not find empirical evidence against missing completely at random (MCAR) mechanism.

Analysis Results

	ACA		MI		Proposed	
	Estimate	SE	Estimate	SE	Estimate	SE
(Intercept)	2.286	0.461	2.166	0.387	1.539	0.298
ABETA	−0.309	0.304	−0.297	0.294	−0.251	0.268
YEAR	−0.197	0.077	−0.206	0.003	−0.254	0.059
SEX	−0.340	0.228	−0.573	0.260	−0.320	0.164
baseDRS	0.827	0.038	0.828	0.033	0.886	0.025
ABETA:YEAR	−0.448	0.187	−0.135	0.073	−0.445	0.187

Conclusions:

1. The proposed method provides unbiased estimates, but is more efficient than ACA.
2. Standard MI produces some obviously biased estimates (see those in red in the table above).

Discussions

1. We have proposed a nonparametric estimator for longitudinal data with missing covariates by using related auxiliary variables.
2. The proposed estimator is consistent and asymptotic normal.
3. The proposed method is more efficient than standard methods like ACA, and more robust than other popular methods like standard MI.

Limitation: We have not yet considered the setting of multiple missing covariates with multiple auxiliary variables.

Acknowledgements

This work was supported by National Institute of Health grant R01-NS102324. This work was also supported in part by GlaxoSmithKline (GSK).