Simultaneous Mean and Covariance Estimation of Partially Linear Models for Longitudinal Data with Covaraite Measurement Errors and Missing Responses

Zhongyi Zhu

Department of Statistics, Fudan University

joint work with Guoyou Qin and Jiajia Zhang

Motivation

In practical issues, longitudinal data sets with measurement errors or dropouts or both arise more often. Ignoring them usually results in inconsistent estimators.

- Statistical inference of mean regression for data sets with covariate measurement errors and missing response have attracted considerable interests of research, e.g., Liu and Wu (2007), Yi et al. (2011) and Yi et al. (2012).
- However, less attentions have been paid to simultaneous mean and covariance estimation for partially linear models.

Partially linear models for longitudinal data

We consider the marginal partially linear model as

$$Y_{ij} = X_{ij}^T \beta_0 + f_0(T_{ij}) + \epsilon_{ij}, i = 1, \cdots, n, j = 1, \cdots, m, \quad (1)$$

where

- β_0 is a p-dimensional vector of regression parameters,
- $f_0(\cdot)$ is an unknown smoothing function,
- $\epsilon_i = (\epsilon_{i1}, \cdots, \epsilon_{im})^T$ are independent random error vectors with mean 0 and covariance matrix Σ_0.

Objects: We focus on simultaneous estimation of the mean components β_0 and f_0 and the covariance component Σ_0 when Y_{ij} are subject to missing and X_{ij} are measured with errors.

Model for Measurement errors

Let W_{ij} be the observed version of the error-prone covariate vector X_{ij}. We assume that

$$W_{ij} = X_{ij} + \delta_{ij},$$

where δ_{ij} follow some distribution with mean 0, and are independent of X_i and ϵ_i.

In this article, we assume that there are two replicate measurements for each X_{ij}, i.e.,

$$W_{ij,1} = X_{ij} + \delta_{ij,1} \quad \text{and} \quad W_{ij,2} = X_{ij} + \delta_{ij,2},$$

where δ_{i1} and δ_{i2} are independent.

Missing process

Let R_{ij} be 1 if Y_{ij} is observed, and 0 otherwise.
Let $\tilde{R}_{ij} = (R_{i1}, \cdots, R_{ij-1})^T$ be the history of missing data indicator at time j and Y_i^o contain the observed components of Y_i.

- We assume that missingness of Y_{ij} is allowed to depend on Y_i^o, X_i and T_i but not on W_i.
- To reflect the dynamic feature of the observation process over time, we assume $P(R_{ij} = 1|\tilde{R}_i, Y_i, X_i, T_i) = P(R_{ij} = 1|\tilde{R}_i, Y_i^o, X_i, T_i)$

Missing process, continued

- It is important to emphasize that, under the setting of this article, the missingness of Y does not depend on W.
- Since the true X is not observable, Y is therefore not missing at random.
- Since we will make no further assumption, such as about the distribution of X or about the model on missing probabilities, what we are dealing with here is conceptually quite different from most studies of missing data in which missing at random or missing completely at random is assumed.

Proposed Approach

Naively replacing the error-prone covariates X_{ij} with the observed version W_{ij} and simply excluding the missing data usually result in inconsistent estimators for general approaches.

- To deal with missing response, we use the idea of projection proposed in Qu et al. (2010).
- We propose a new approach to handle the bias induced by measurement errors utilizing the independence of the two replicate measurements.

Estimating equation for the mean component

Regression spline is used to approximated the nonparametric function $f_0(\cdot)$, model (1) is linearized as

$$Y_{ij} = X_{ij}^T \beta_0 + \pi_{ij}^T \alpha_0 + \epsilon_{ij} = D_{ij}^T \theta_0 + \epsilon_{ij}, \qquad (2)$$

where

- $D_{ij} = (X_{ij}^T, \pi_{ij}^T)^T$, $\pi_{ij} = \pi(T_{ij})$,
- $\theta_0 = (\beta_0^T, \alpha_0^T)^T$ is the combined regression parameters.
- $\pi(t) = (B_1(t), \cdots, B_{N_k}(t))^T$ is a vector of basis function and $\alpha_0 \in R^{N_k}$ is the vector of spline coefficient.

Estimating equation for the mean component, continued I

- Let $(Y_i - D_i\theta) = \begin{pmatrix} Y_i^o - D_i^o\theta \\ Y_i^m - D_i^m\theta \end{pmatrix}$ be a decomposition of the data vector into observed Y_i^o and missing Y_i^m variables,
- Further denote $cov(Y_i) = \Sigma_i = \begin{pmatrix} \Sigma_i^{11} & \Sigma_i^{12} \\ \Sigma_i^{21} & \Sigma_i^{22} \end{pmatrix}$ where Σ_i^{11} and Σ_i^{22} respectively denote the covariance of the observed and missing responses.

Directly applying the estimating equation under MAR proposed in Qu et al. (2010) for our partially linear model (1), we have

$$\sum_{i=1}^{n} D_i^T \Sigma_i^{-1} E(Y_i - D_i\theta | Y_i^o) = 0. \tag{3}$$

Estimating equation for the mean component, continued II

Under the linear conditional mean assumption (LCM) supposed in Qu et al. (2010) which assume that the conditional expectation is linear in Y_i, we have

$$E(Y_i^m - D_i^m\theta | Y_i^o) = \Sigma_i^{21}(\Sigma_i^{11})^{-1}(Y_i^o - D_i^o\theta).$$

Thus, the estimating equation (3) can be written as

$$\begin{aligned} & \sum_{i=1}^{n} D_i^T \Sigma_i^{-1} A_i (Y_i^o - D_i^o\theta) \\ = & \sum_{i=1}^{n} (D_i^o)^T (\Sigma_i^{11})^{-1} (Y_i^o - D_i^o\theta) = 0. \qquad (4) \end{aligned}$$

Estimating equation for the mean component, continued III

- The estimating equation (3) is unbiased and efficient under MAR in the case where the covariates are exactly measured without errors.
- When the covariates are measured with errors , the estimating equation obtained through naively replacing the covariates in the estimating equation (3) with their observed versions is biased.

Therefore, to achieve unbiased estimating equation under situations of missing response and measurement errors, we develop the following novel estimating equation for the mean

$$\sum_{i=1}^{n}(\tilde{D}_{i(1)}^{o})^{T}(\Sigma_i^{11})^{-1}(Y_i^o - \tilde{D}_{i(2)}^{o}\theta) = 0, \quad (5)$$

where $\tilde{D}_{i(1)} = (W_{i1}, M_i)$ and $\tilde{D}_{i(2)} = (W_{i2}, M_i)$.

Estimating equation for the covarinace component I

We assume the covariance matrix depend on some parameters γ. Using similar idea to the construction of the estimating equation for the mean, to achieve consistent estimate of the covariance component, we develop the following estimating equation

$$\sum_{i=1}^{n} \sum_{a \leq b} \frac{\partial}{\partial \gamma} \sigma_i^{ab}(\gamma)[\sigma_i^{ab}(\gamma) - B_{iab}] = 0, \tag{6}$$

where

Estimating equation for the covarinace component

- B_{iab} equals to

$$\begin{cases} (Y_{ia} - \tilde{D}_{ia(1)}^T\theta)(Y_{ib} - \tilde{D}_{ib(2)}^T\theta) & \text{if both } Y_{ia} \text{ and } Y_{ib} \text{ are observed,} \\ (Y_{ia}^p - \tilde{D}_{ia(1)}^T\theta)(Y_{ib} - \tilde{D}_{ib(2)}^T\theta) & \text{if } Y_{ia} \text{ is missing and } Y_{ib} \text{ is observed,} \\ (Y_{ia} - \tilde{D}_{ia(1)}^T\theta)(Y_{ib}^p - \tilde{D}_{ib(2)}^T\theta) & \text{if } Y_{ia} \text{ is observed and } Y_{ib} \text{ is missing,} \\ [\Sigma_i^{p22}]_{ab} + (Y_{ia}^p - \tilde{D}_{ia(1)}^T\theta)(Y_{ib}^p - \tilde{D}_{ib(2)}^T\theta) & \text{if both } Y_{ia} \text{ and } Y_{ib} \text{ are missing.} \end{cases}$$

- Y_{ia}^p, Y_{ib}^p are predicted response values based on LCM assumption,
- $\Sigma_i^{p22} = \Sigma_i^{22} - \Sigma_i^{21}(\Sigma_i^{11})^{-1}\Sigma_i^{12}$.
- $\tilde{D}_{ia(1)} = (W_{ia(1)}^T, \pi_{ia}^T)^T$,
- $\tilde{D}_{ia(2)} = (W_{ia(2)}^T, \pi_{ia}^T)^T$.
- *Remark:* It is not difficult to show that $E(B_{iab}|Y_i^o, X_i, T_i) = E\{(Y_{ia} - D_{ia}^T\theta)(Y_{ib} - D_{ib}^T\theta)|Y_i^o, X_i, T_i\}$. Thus, the influence introduced by the measurement errors is successfully removed.

Asymptotic properties

Under some regularity conditions, we have

- For the mean component:
  Theorem 1.

$$\{cov(\hat{\beta}|X, T)\}^{-1/2}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} N(0, I_{p\times p}),$$

and for any $t \in (a_i, a_{i+1}], i = 0, \cdots, k$ and
$f^*(t) = f_0(t) + b(t) + o_p(h^r)$,

$$var(\hat{f}(t)|X, T)^{-1/2}\{\hat{f}(t) - f^*(t)\} \xrightarrow{\mathcal{L}} N(0, 1),$$

where $\xrightarrow{\mathcal{L}}$ *denotes convergence in distribution.*

- For the covariance component:
  Theorem 2. $n^{1/2}(\hat{\gamma} - \gamma_0)$ *converges to normal distribution.*

Simulation models

We consider a partial linear model with covariate measurement errors and dropouts as

$$Y_{ij} = X_{ij}\beta_0 + \sin(2\pi T_{ij}) + \epsilon_{ij}, i = 1, \ldots, 400, m = 1, \ldots, 4,$$

where

- $\beta_0 = 1$, X_{ij} and T_{ij} are independently drawn from normal distribution with mean one and standard deviation one and uniform distributions on $(0, 1)$ respectively,
- $\epsilon_i = (\epsilon_{i1}, \cdots, \epsilon_{im})^T$ are generated from multivariate normal distribution with mean zero and covariance matrix Σ_0 which is taken to be exchangeable (EX), one-order autoregressive (AR1) and unstructured (UN) structures

Simulation models, continued I

- Specification of the measurement error model: the surrogate value w_{ij} were generated from the following model

$$W_{ij,1} = X_{ij} + \delta_{ij,1} \quad \text{and} \quad W_{ij,2} = X_{ij} + \delta_{ij,2},$$

where δ_{i1} and δ_{i2} are independently generated from normal distribution with mean zero and standard deviation σ_m. In the simulation, we take $\sigma_m = 0.4$ and 0.6 respectively.

- Specification of the dropouts model, the missing data indicators were generated from the model

$$ln \frac{\lambda_{ij}}{1 - \lambda_{ij}} = \varphi_0 + \varphi_1 Y_{ij-1} + \varphi_2 X_{ij},$$

where $(\varphi_0, \varphi_1, \varphi_2)^T$ is taken to be $(1, 1, -0.5)^T$ which yields about 33% missingness.

Simulation models, continued II

- We calculated the bias, standard error (SE), and mean squared error (MSE) of $\hat{\beta}$, as well as the the integrated mean squared error (IMSE) of $\hat{f}(\cdot)$ for the estimators of the mean.
- To investigate the performance of the proposed method in estimating the covariance, we calculate the entropy loss $\Delta_E(\Sigma, \hat{\Sigma}) = trace(\Sigma^{-1}\hat{\Sigma}) - log|\Sigma^{-1}\hat{\Sigma}| - m$ and quadratic loss $\Delta_Q(\Sigma, \hat{\Sigma}) = trace(\Sigma^{-1}\hat{\Sigma} - I)^2$ which means accuracy in estimating the covariance matrix where Σ is the true covariance matrix and $\hat{\Sigma}$ is its estimator.
- For each case, we performed 500 simulations.

Methods in comparison

We compares the proposed estimator (P) with other two estimators

- One is the naive method (N) which ignores both missing response and covariate measurement errors, performed by classical generalized estimating equation using the average values of two replicate measurements as the observation values for the error-prone covariate X and adopting the AR correlation matrix as the working correlation matrix.
- The other method (Q), performed by Qu et al. (2010)'s estimating equation also using the average values of two replicate measurements as the observation values for X, which accounts for the missingness but ignores the measurement errors.

Results in simulation

Table 1 Simulation results for the mean model in Study 1

		$\hat{\beta}$ BIAS	MSE	CP	$\hat{f}$ IMSE
			$\sigma_m = 0.3$		
EX	P	0.0023	0.0007	0.949	0.0073
	Q	-0.0430	0.0024	0.531	0.0084
	N	-0.0235	0.0013	0.843	0.0259
AR	P	0.0022	0.0008	0.952	0.0074
	Q	-0.0434	0.0025	0.571	0.0084
	N	-0.0215	0.0012	0.869	0.0184
UN	P	0.0025	0.0007	0.939	0.0065
	Q	-0.0433	0.0024	0.461	0.0074
	N	-0.0183	0.0009	0.875	0.0217
			$\sigma_m = 0.5$		
EX	P	0.0066	0.0012	0.944	0.0095
	Q	-0.1119	0.0131	0.003	0.0185
	N	-0.0945	0.0097	0.066	0.0496
AR	P	0.0063	0.0013	0.942	0.0097
	Q	-0.1128	0.0134	0.007	0.0188
	N	-0.0928	0.0093	0.067	0.0384
UN	P	0.0076	0.0012	0.927	0.0092
	Q	-0.1127	0.0132	0.000	0.0178
	N	-0.0899	0.0087	0.060	0.0446

Notes: MSE: mean squared error; CP: coverage probability; IMSE: integrated MSE; EX: Data generated from exchangeable correlation structure; AR: autoregressive correlation; UN: unstructured correlation ; P: proposed method; Q: Qu et al's method; N: naive method.

Results in simulation

Table 2 Simulation results for the covariance component

		QL		EL	
		P	Q	P	Q
				Study 1	
$\sigma_m = 0.3$	EX	0.0411	0.1309	0.0616	0.0649
	AR	0.0402	0.1173	0.0617	0.0648
	UN	0.0444	0.2231	0.0654	0.0823
$\sigma_m = 0.5$	EX	0.0575	0.7403	0.0837	0.1411
	AR	0.0551	0.6484	0.0829	0.1381
	UN	0.0712	1.3788	0.1029	0.2436
				Study 2	
$\sigma_m = 0.3$	EX	0.0475	0.1336	0.0813	0.0791
	AR	0.0481	0.1215	0.0815	0.0792
	UN	0.0533	0.2260	0.0876	0.0958
$\sigma_m = 0.5$	EX	0.0667	0.7378	0.1166	0.1547
	AR	0.0669	0.6480	0.1150	0.1512
	UN	0.0893	1.3766	0.1487	0.2555

Notes: EL: Entropy loss function; QL: Quadratic loss function; Other notations see Tables 1.

Real data analysis

There are total of 197 subjects with 3 observations for each subject. The missing rate is about 21%.
Response is the log of BMI value, covariates include SBP (regarded as covaraite measured with error), DBP(regarded as covaraite measured with error), group, female, race, college and Age.
The partially linear model considered to fit the data is described as

$$\begin{aligned} Y &= SBP\beta_1 + DBP\beta_2 + Gender\beta_3 + Race\beta_4 + college\beta_5 \quad (7) \\ &+group\beta_6 + t_1\beta_7 + t_2\beta_8 + group \times t_1\beta_9 + group \times t_2\beta_{10} \\ &+f(sAge) + \epsilon. \end{aligned}$$

Estimate results for the regression coefficients are summarized in Tables 3 and 4

Real data analysis

Table 3 Regression coefficient estimates in the analysis of the real data

	P	MQ	N
SBP	0.0009 (0.0004)*	0.0007 (0.0003)*	0.0002 (0.0003)
DBP	-0.0001 (0.0005)	0.0003 (0.0004)	0.0001 (0.0004)
Female	-0.0291 (0.0249)	-0.0292 (0.0248)	-0.0284 (0.0249)
Race	0.0246 (0.0237)	0.0241 (0.0237)	0.0348 (0.0241)
College	-0.0860 (0.0234)*	-0.0856 (0.0233)*	-0.0917 (0.0236)*
Group	-0.0065 (0.0251)	-0.0068 (0.0251)	-0.0067 (0.0254)
t1	-0.0139 (0.0075)	-0.0139 (0.0075)	-0.0147 (0.0071)*
t2	-0.0143 (0.0101)	-0.0140 (0.0101)	-0.0149 (0.0100)
Group $\times$ *t*1	-0.0086 (0.0089)	-0.0086 (0.0089)	-0.0089 (0.0087)
Group $\times$ *t*2	-0.0257 (0.0129)*	-0.0260 (0.0128)*	-0.0251 (0.0129)

Notes: The figures in the parenthesis are standard errors.
"*" indicate the effect is significant at the level of $\alpha = 0.05$.

Table 4 Correlation matrix estimates
in the analysis of the real data

	P	
1.0000	0.9415	0.8791
0.9415	1.0000	0.9424
0.8791	0.9424	1.0000
	MQ	
1.0000	0.9415	0.8808
0.9415	1.0000	0.9442
0.8808	0.9442	1.0000

Real data analysis

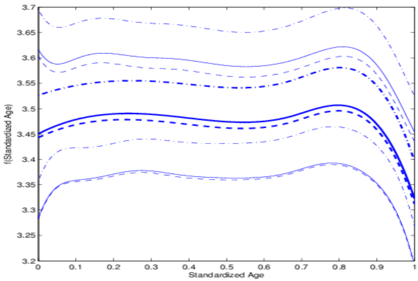

Figure: The estimated function on the standardized age. The heavy solid, dashed, dot-dashed lines represent the curves estimated by the P, Q and N methods respectively. The solid, dashed, dot-dashed lines represent the corresponding confidence bands.

Reference

- Qu. A., Lindsay, B. G. and Lu, L.(2010). Highly efficient aggregate unbiased estimating functions approach for correlated data with missing at random. *Journal of the American Statistical Association*, **105**, 194-204.
- Yi, G, Liu, W. and Wu, L. (2011). Simultaneous inference and bias analysis for longitudinal data with covariate measurement error and missing response. *Biometrics*, **67**, 67-75.
- Yi, G, Ma, Y. and Carroll, R. J. (2012). A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. *Biometrika*, **99**, 151-165.

Thank you!