Doubly robust estimates for longitudinal data analysis with missing response and missing covariates

Xiao-Hua Andrew Zhou, Ph.D

Co-Investigator and Senior Biostatistician, NACC
Professor, Department of Biostatistics
University of Washington

October, 2009

A NACC example

- Using the National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS), we are interested in assessing he association between patient's characteristics and the onset of dementia.
- The response is the diagnosis of dementia (Yes/No).
- The covariates that may be related to the status of dementia include sex, congestive heart failure (CVCHF, yes/no), family history of dementia (FHDEM, yes/no), diabetes (yes/no), behavioral assessment (depression or dysphoria, yes/no), hypertension (yes/no), education (years), Mini-Mental State Exam (MMSE) score, and age.

A NACC example, continued

- There are 16223 subjects from 29 Alzheimer's Disease Centers included at the entry of this study.
- Follow-up visits for subjects are scheduled at approximately one-year intervals, with up to three follow-ups at present.

An example, continued

- Due to some reasons, there are some missing data for the response and the behavioral assessment covariate.
- There are 8724 subjects with complete data on scheduled visits.
- About 11.9% subjects miss both the response and behavioral assessment; about 31.2% subjects miss the response but observe behavioral assessment; about 3.2% subjects miss the behavioral assessment but observe the response; and about 53.7% subjects observe both the response and the behavioral assessment covariate.

GEE Approach with Complete Longitudinal Data

- The method of generalized estimating equations (GEE) is a popular method for analyzing longitudinal data.
- It requires only the specification of a model for the marginal mean and variance of each measurement and of a "working" matrix for the correlation between measurements in a cluster.

Notations

- Let Y_{ij} denote the response of individual i at time j $(i = 1, \ldots, N;\ j = 1, \ldots, M_i)$. Let $Y_i = (Y_{i1}, \ldots, Y_{iM_i})^T$.
- Let x_{ij} denote a vector of covariates for individual i at time j, and $x_i = (x_{i1}^T, \ldots, x_{iM_i}^T)^T$. $\mathbf{x}_i = (\mathbf{x}_{i1}^T, \ldots, \mathbf{x}_{iM_i}^T)^T$.
- Let $\mu_{ij} = E(Y_{ij} \mid x_{ij})$, $g(\mu_{ij}) = \beta^T x_{ij}$; let $\mu_i = (\mu_{i1} \ \ldots, \mu_{iM_i})^T$.

GEE for Complete Data Analysis

- The GEE for complete data are

$$\sum_{i=1}^{N} U_i(\beta, \rho; Y_i, x_i) = 0,$$

where

$$U_i(\beta, \rho; Y_i, x_i) = \frac{\partial \mu_i^T}{\partial \beta} V_i(\rho)^{-1}(Y_i - \mu_i),$$

and $V_i(\rho)$ is the working covariance matrix of Y_i.

Asymptotic results

- When x_i contains only time-independent covariates, under some regularity conditions, the GEE yields estimators that are consistent.
- If x_i includes some time-dependent covariates, the GEE still yields consistent estimators under one additional assumption that $E(Y_{ij} \mid x_i) = E(Y_{ij} \mid x_{ij})$. If this is not the case, then for consistency the independent working correlation should be used.

Time-dependent Covariates

- Let $\mathbf{L}_{ij}$ denote all the data that should be collected on individual i at time j.
- Let $\overline{\mathbf{L}}_{ij}$ denote the data available on individual i by time j.
- Let $\underline{\mathbf{L}}_{ij}$ denote the data not yet available by time j.
- Note that $\mathbf{L}_{ij}$ includes both Y_{ij} and $\mathbf{x}_{ij}$.

Drop-out

- Let $R_{ij} = 1$ if measurement j on individual i is observed and $R_{ij} = 0$ otherwise.
- Assume monotone drop-out: $R_{ij} = 0$ implies $R_{ik} = 0$ for all times $k > j$.
- Let $C_{ij} = 1$ if subject is last observed measurement is at time j and 0 otherwise.

We assume that the covariates included in $\mathbf{L}_{ij}$ are chosen so that the data can assumed to be Missing at Random (MAR):

$$P(R_{ij} = 1|\overline{\mathbf{L}}_{iM_i}, R_{i,j-1} = 1) = P(R_{ij} = 1|\overline{\mathbf{L}}_{i,j-1}, R_{i,j-1} = 1).$$

i.e., the probability of missingness only depends on the observed data.

GEE for Complete-Data

$$\sum_{i=1}^{N} U_i(\beta, \rho; Y_i, x_i) = 0,$$

where

$$U_i(\beta, \rho; Y_i, x_i) = \frac{\partial \mu_i^T}{\partial \beta} V_i(\rho)^{-1}(Y_i - \mu_i),$$

and $V_i(\rho)$ is the working covariance matrix of Y_i.
These equations yield estimates that are consistent if the data are Missing Completely at Random (MCAR), but not necessarily if they are MAR.

Re-weighting

- With missing data, we can base our estimates on the complete cases, but re-weight them according to the probability of being observed.
- The estimating equations are then

$$\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i(\boldsymbol{\rho})^{-1} \Delta_i(\boldsymbol{\alpha})(\mathbf{Y}_i - \boldsymbol{\mu}_i),$$

  where $\Delta_i(\boldsymbol{\alpha}) = \text{diag}(R_{i1}/\pi_{i1}, \ldots, R_{iM_i}/\pi_{iM_i})$ and
- $\pi_{ij} = \pi_{ij}(\boldsymbol{\alpha})$ is the probability, according to a specified dropout model, that measurement j on subject i is observed.
- Under the drop-out missing data,

$$\pi_{ij}(\boldsymbol{\alpha}) = (1 - \lambda_{i1}(\boldsymbol{\alpha})) \ldots (1 - \lambda_{ij}(\boldsymbol{\alpha})),$$

  where $\lambda_{ij}(\boldsymbol{\alpha}) = P(R_{ij} = 0 \mid \bar{\mathbf{L}}_{ij}, R_{ij} = 1)$.
- The resulting estimates are consistent if the data are MAR, as long as the probability model for the missingness is correctly

Imputation

- Alternatively, we can impute, or "guess", what the missing values are based on some probability model.
- Then the estimates are based on both the observed data and the imputed data.
- The complete case estimating equations are used, but after imputing missing responses with their expected values:

$$E(Y_{ij}|\bar{\mathbf{L}}_{ik}, R_{ik} = 1), \text{ for } j > k.$$

- The imputations are based on specified regression models.
- The resulting estimates are consistent if the data are MAR, as long as the probability model for the imputations is correct.

Doubly-robust Estimating Equations

- The inverse probability weighting estimates make no use of the available data on subjects with missing measurements.
- Let $\mathbf{d}(\overline{\mathbf{L}}_M, \beta) = U(\beta, \rho; \mathbf{Y}, \mathbf{x})$ be the contribution of a fully observed subject to the estimating equations.
- For drop-out missing data, the IPW estimating equations can be augmented by a term $F(C, \overline{\mathbf{L}}_C, \beta)$ satisfying $E_C\{F(C, \overline{\mathbf{L}}_C, \beta)|\overline{\mathbf{L}}_M\} = 0$.
- The resulting augmented estimating equations are

$$\sum_{i=1}^{N} \left\{ \frac{R_{iM_i}}{\pi_{iMi}} \mathbf{d}(\overline{\mathbf{L}}_{Mi}, \beta) + F(C, \overline{\mathbf{L}}_C, \beta) \right\} = 0.$$

Doubly-robust Estimating Equations (2)

- The optimal choice of augmentation term is

$$F_{\text{opt}}(C, \bar{\mathbf{L}}_C, \boldsymbol{\beta}) = \sum_{j=1}^{M-1} \left(\frac{C_j - \lambda_{j+1} R_j}{\pi_{j+1}} \right) \mathbf{H}_j(\boldsymbol{\beta}),$$

  where $\mathbf{H}_j(\boldsymbol{\beta}) = E_{\underline{\mathbf{L}}_j}\{\mathbf{d}(\bar{\mathbf{L}}_M, \boldsymbol{\beta}) | \bar{\mathbf{L}}_j, R_j = 1\}$.
- We specify models for $\mathbf{H}_j(\boldsymbol{\beta}), j = 1, \ldots, M-1$ which involve parameters $\boldsymbol{\gamma}$.
- Let $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ denote consistent estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$.
- Then, in the estimating equations, replace λ_j, π_j, and $\mathbf{H}_j$ with $\lambda_j(\boldsymbol{\alpha})$, $\pi_j(\boldsymbol{\alpha})$, and $\mathbf{H}_j(\boldsymbol{\beta}, \hat{\boldsymbol{\gamma}})$.

Properties of DR Estimating Equations

If:

- The data are MAR,
- the marginal model is correct, $g(\mu_{ij}) = \boldsymbol{\beta}^T \mathbf{x}_{ij}$, and
- either the dropout model π_j, or the model for $\mathbf{H}_j$ (or both) is correctly specified,

then the solution to the estimating equations $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$.

- Furthermore, if both the dropout model and the model for $\mathbf{H}_j$ are correct, then this solution $\hat{\boldsymbol{\beta}}$ is optimal in the sense that it has the smallest asymptotic variance among estimates from augmented estimating equations. A consistent estimate of this variance exists in closed form.

Methods for Handling Missing Covariates

Lipsitz et al. (1999) considered the doubly robust estimate in the cross-sectional study with a missing covariate

- Notations:
  - y_i: response, x_i: covariate vector that is always observed
  - z_i: covariate that is subject to missing
  - r_i: missing indicator for z_i
- Joint density of $(r_i, y_i, z_i|x_i)$

$$\begin{aligned} p(r_i, y_i, z_i|x_i) &= p(r_i|y_i, z_i, x_i, \omega)p(y_i|z_i, x_i, \beta)p(z_i|x_i, \alpha) \\ &= p(r_i|y_i, x_i, \omega)p(y_i|z_i, x_i, \beta)p(z_i|x_i, \alpha) \quad (MAR) \end{aligned}$$

Score Equation for Complete Data

The likelihood-based score question:

$$\sum_{i=1}^{n} \begin{bmatrix} u_{1i}(\beta) \\ u_{2i}(\alpha) \\ u_{3i}(\omega) \end{bmatrix} = 0,$$

where

- $u_{1i}(\beta; y_i, x_i, z_i) = \frac{\partial \log p(y_i|x_i,z_i,\beta)}{\partial \beta}$
- $u_{2i}(\beta; x_i, z_i) = \frac{\partial \log p(z_i|x_i,\alpha)}{\partial \alpha}$
- $u_{3i}(\beta; r_i, x_i, y_i) = \frac{\partial \log p(r_i|x_i,y_i,z_i,\omega)}{\partial \omega}$

Methods for Handling Missing Covariates

With missing data, the maximum likelihood estimating equations for $\hat{\gamma} = (\hat{\beta}', \hat{\alpha}', \hat{\omega}')'$ solves

$$u^*(\hat{\gamma}) = \sum_{i=1}^{n} u_i^*(\hat{\gamma}) = \sum_{i=1}^{n} E \left[\begin{array}{c} u_{1i}(\hat{\beta}) \\ u_{2i}(\hat{\alpha}) \\ u_{3i}(\hat{\omega}) \end{array} \middle| \text{ observed data } \right] = 0$$

Methods for Handling Missing Covariates

We can further show that

$$u^*(\gamma) = \sum_{i=1}^{n} \begin{bmatrix} r_i u_{1i}(\beta; y_i, x_i, z_i) + (1 - r_i) E_{z_i|y_i,x_i}[u_{1i}(\beta; y_i, x_i, z_i)] \\ r_i u_{2i}(\alpha; z_i, x_i) + (1 - r_i) E_{z_i|y_i,x_i}[u_{2i}(\alpha; z_i, x_i)] \\ u_{3i}(\omega; y_i, x_i, r_i) \end{bmatrix}$$

- Solving $u^*(\hat{\gamma}) = 0$ we get the MLE
- The asymptotic properties of $(\hat{\beta}, \hat{\alpha})'$ don't depend on the missing data model
- If $p(y_i|x_i, z_i)$ and $p(z_i|x_i)$ are correctly specified, we can get consistent estimate of $(\hat{\beta}, \hat{\alpha})'$ by solving $u^*(\hat{\gamma}) = 0$
- If $p(y_i|x_i, z_i)$ or/and $p(z_i|x_i)$ are misspecified, then $\hat{\beta}$ will not be consistent

Methods for Handling Missing Covariates

Weighted GEE

$$S(\gamma) = \sum_{i=1}^{n} \begin{bmatrix} \frac{r_i}{\pi_i} u_{1i}(\beta; y_i, x_i, z_i) + \left(1 - \frac{r_i}{\pi_i}\right) E_{z_i|y_i,x_i}[u_{1i}(\beta; y_i, x_i, z_i)] \\ \frac{r_i}{\pi_i} u_{2i}(\alpha; z_i, x_i) + \left(1 - \frac{r_i}{\pi_i}\right) E_{z_i|y_i,x_i}[u_{2i}(\alpha; z_i, x_i)] \\ u_{3i}(\omega; y_i, x_i, r_i) \end{bmatrix}$$

where $\pi_i = P(r_i = 1|y_i, x_i)$

- Doubly robust estimate, i.e., solving $S(\hat{\gamma}) = 0$ can get asymptotic unbiased estimate for β when either π_i or $p(z_i|x_i)$ is correctly specified
- EM algorithm for the estimate
- Asymptotic variance

$$Var(\hat{\gamma}) = \Big\{ \sum_{i=1}^{n} E\Big[\frac{\partial S_i(\gamma)}{\partial \gamma'}\Big] \Big\}^{-1} \sum_{i=1}^{n} E[S_i(\gamma) S_i'(\gamma)] \Big\{ \sum_{i=1}^{n} E\Big[\frac{\partial S_i(\gamma)}{\partial \gamma}\Big] \Big\}^{-1}$$

Model Formulation

- Notations
  - Response: $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ_i})'$
  - Covariate: $X_i = (X_{i1}, X_{i2}, \ldots, X_{iJ_i})'$
  - $R_{ij} = \begin{cases} 0 & Y_{ij} \text{ and } X_{ij} \text{ are missing} \\ 1 & Y_{ij} \text{ is missing and } X_{ij} \text{ is observed} \\ 2 & Y_{ij} \text{ is observed and } X_{ij} \text{ is missing} \\ 3 & Y_{ij} \text{ and } X_{ij} \text{ are observed} \end{cases}$
  - Covariate: Z_i [always observed]
- Response model: $\mu_{ij} = E(Y_{ij}|X_i, Z_i)$

$$var(Y_{ij}|X_i, Z_i) = \kappa f(\mu_{ij})$$

$$g(\mu_{ij}) = X_{ij}\beta_x + Z'_{ij}\beta_z$$

Model Formulation (Continued)

- Missing data models: $\lambda_{ijk} = P(R_{ij} = k|\bar{R}_{ij}, Y_i, X_i, Z_i),\ k = 0, 1, 2, 3$

$$\log\left(\frac{\lambda_{ijk}}{\lambda_{ij0}}\right) = u_{ijk}{}'\alpha_k \quad k = 1, 2, 3$$

$\bar{R}_{ij}$: missing response indicator history

- Covariate model: $\omega_{ij} = E(X_{ij}|\bar{X}_{ij}, Z_i)$

$$h(\omega_{ij}) = v_{ij}'\gamma$$

$\bar{X}_{ij}$: covariate history

- $\theta = (\beta', \alpha', \gamma')'$, where β is of interest

Model Formulation (Continued)

- MAR assumption:

$$P(R_{ij} = k|\bar{R}_{ij}, Y_i, X_i, Z_i)$$

$$= \quad P(R_{ij} = k|\bar{R}_{ij}, Y_i^{(o)}, X_i^{(o)}, Z_i)$$

  - $Y_i = (Y_i^{(o)}, Y_i^{(m)})$
  - $X_i = (X_i^{(o)}, X_i^{(m)})$

Model Formulation (Continued)

Weighted GEE (WGEE) for β:

$$S_1(\theta) = \sum_{i=1}^{n} \left[D_i M_i (Y_i - \mu_i) + E_{Y_i^{(m)}, X_i^{(m)} | Y_i^{(o)}, X_i^{(o)}, Z_i} [D_i N_i (Y_i - \mu_i)] \right] = 0$$

- $M_i = \kappa^{-1} F_i^{-1/2} [C_i^{-1} \bullet \Delta_i] F_i^{-1/2}$
- $N_i = \kappa^{-1} F_i^{-1/2} [C_i^{-1} \bullet (11' - \Delta_i)] F_i^{-1/2}$
- $F_i = \mathrm{diag}(var(Y_{ij} | X_{ij}, Z_{ij}), \quad j = 1, \ldots, J_i)$
- C_i: working correlation matrix
- $\Delta_i = [\delta_{ijk}]$ with

$$\delta_{ijk} = [I(R_{ij} = 1, R_{ik} = 3) + I(R_{ij} = 3, R_{ik} = 3)] / \pi_{ijk} \text{ for } j \neq k$$

and

$$\delta_{ijj} = I(R_{ij} = 3) / \pi_{ij}$$

Model Formulation (Continued)

Weighted GEE (WGEE) for γ:

$$S_2(\theta) = \sum_{i=1}^{n} \left[v_i \Delta_i^* (X_i - \omega_i) + E_{X_i^{(m)}|X_i^{(o)},Z_i} [v_i (I - \Delta_i^*)(X_i - \omega_i)] \right] = 0$$

- $\Delta_i^* = \text{diag}(I(R_{ij} = 1 \text{ or } 3)/\pi_{ij}^x, \quad j = 1, \ldots, J_i)$
- $\pi_{ij}^x = P(R_{ij} = 1 \text{ or } 3 | Y_i, Z_i, X_i)$

Model Formulation (Continued)

Estimation function for missing data parameter α:

$$S_3(\alpha) = \sum_{i=1}^{n} \sum_{j=1}^{J_i} \sum_{k=0}^{3} \frac{I(R_{ij} = k)}{\lambda_{ijk}} \frac{\partial \lambda_{ijk}}{\partial \alpha} = 0$$

Estimation and Inference

Solve estimating equations

$$S(\hat{\theta}) = \begin{bmatrix} S_1(\hat{\theta}) \\ S_2(\hat{\theta}) \\ S_3(\hat{\alpha}) \end{bmatrix} = \sum_{i=1}^{n} S_i(\theta) = 0$$

- EM algorithm for the estimation
- Variance estimate

$$Var(\hat{\theta}) = \Big\{ \sum_{i=1}^{n} E\Big[\frac{\partial S_i(\theta)}{\partial \theta}\Big] \Big\}^{-1} \sum_{i=1}^{n} E[S_i(\theta) S_i'(\theta)] \Big\{ \sum_{i=1}^{n} E\Big[\frac{\partial S_i(\theta)}{\partial \theta}\Big]' \Big\}^{-1}.$$

Estimation and Inference (Continued)

Doubly robust estimate

- If missing data model is correctly specified, we get asymptotic unbiased estimate for β no matter the model for the covariate is correctly specified or not
- If covariate model is correctly specified, we get asymptotic unbiased estimate for β no matter the model for the missing data is correctly specified or not

Simulations

- Response model is $\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{ij} + \beta_2 Z_{ij}$, $j = 1, 2, 3$, with exchangeable correlation ρ.
- Covariate model

$$\text{logit}\omega_{ij} = \gamma_0 + \gamma_1 X_{i,j-1} + \gamma_2 Z_{ij}$$

- Missing data model

$$\begin{aligned} \log\left(\frac{\lambda_{ijk}}{\lambda_{ij0}}\right) &= \alpha_{0k} + \alpha_{1k1} I(R_{i,j-1} = 1) + \alpha_{1k2} I(R_{i,j-1} = 2) \\ &\quad + \alpha_{1k3} I(R_{i,j-1} = 3) + \alpha_{2k} y^{(o)}_{i,j-1} + \alpha_{3k} x^{(o)}_{i,j-1} \end{aligned}$$

Simulations (Continued)

Methods considered

1. EM($x+$): EM with correct covariate model
2. WGEE($x+, r+$): WGEE with correct covariate and missing data models
3. WGEE($x-, r+$): WGEE with incorrect covariate and correct missing data models
4. WGEE($x+, r-$): WGEE with correct covariate and incorrect missing data models
5. WGEE($x-, r-$): WGEE with incorrect covariate and incorrect missing data models
6. cc: complete case MLE

Simulations (Continued)

Table: Empirical bias, standard deviation and coverage probabilities for six approaches to estimation and inference with incomplete covariate and response data ($\rho = 0.6$, $\alpha_2 = \gamma_2 = -2$)

	β_0			β_1			β_2		
Method	Bias%	SD	CP%	Bias%	SD	CP%	Bias%	SD	CP%
EM($x+$)	-0.3	0.102	94.9	-1.1	0.077	94.3	0.5	0.091	94.8
$(x+, r+)$	0.7	0.104	95.1	0.8	0.080	94.5	-0.9	0.093	94.9
$(x+, r-)$	-1.0	0.110	95.2	-1.6	0.088	94.9	1.6	0.102	95.0
$(x-, r+)$	0.4	0.105	94.4	1.0	0.084	94.8	-0.3	0.096	94.5
$(x-, r-)$	-20.1	0.094	91.4	12.0	0.081	92.9	3.0	0.096	93.9
cc	-302.0	0.876	53.8	49.9	1.077	96.8	0.4	1.218	94.6

TABLE: Empirical bias, standard deviation and coverage probabilities for six approaches to estimation and inference with incomplete covariate and response data ($\rho = 0.3$, $\alpha_2 = \gamma_2 = -2$)

	β_0			β_1			β_2		
Method	Bias%	SD	CP%	Bias%	SD	CP%	Bias%	SD	CP%
EM($x+$)	-1.6	0.058	94.4	-0.2	0.067	95.3	1.1	0.084	94.4
($x+, r+$)	0.1	0.060	95.4	0.1	0.072	95.1	0.3	0.086	94.6
($x+, r-$)	0.0	0.066	94.3	0.8	0.071	94.9	0.2	0.091	94.7
($x-, r+$)	1.2	0.062	94.7	0.6	0.079	94.8	-0.9	0.087	94.5
($x-, r-$)	-12.4	0.076	93.4	8.4	0.077	94.1	2.0	0.087	94.2
cc	-219.6	0.784	78.6	-27.0	1.065	97.2	0.0	0.930	94.9

Simulations (Continued)

Summary of the Simulations:

- EM algorithm gives consistent and most efficient estimate when the covariate model is correctly specified
- The proposed method yields negligible biases when either the covariate model or the missing data model is correctly specified
- If both the covariate and missing data model are misspecified, the proposed method yield biased result
- The complete case analysis gives biased estimate

Impact of model misspecification

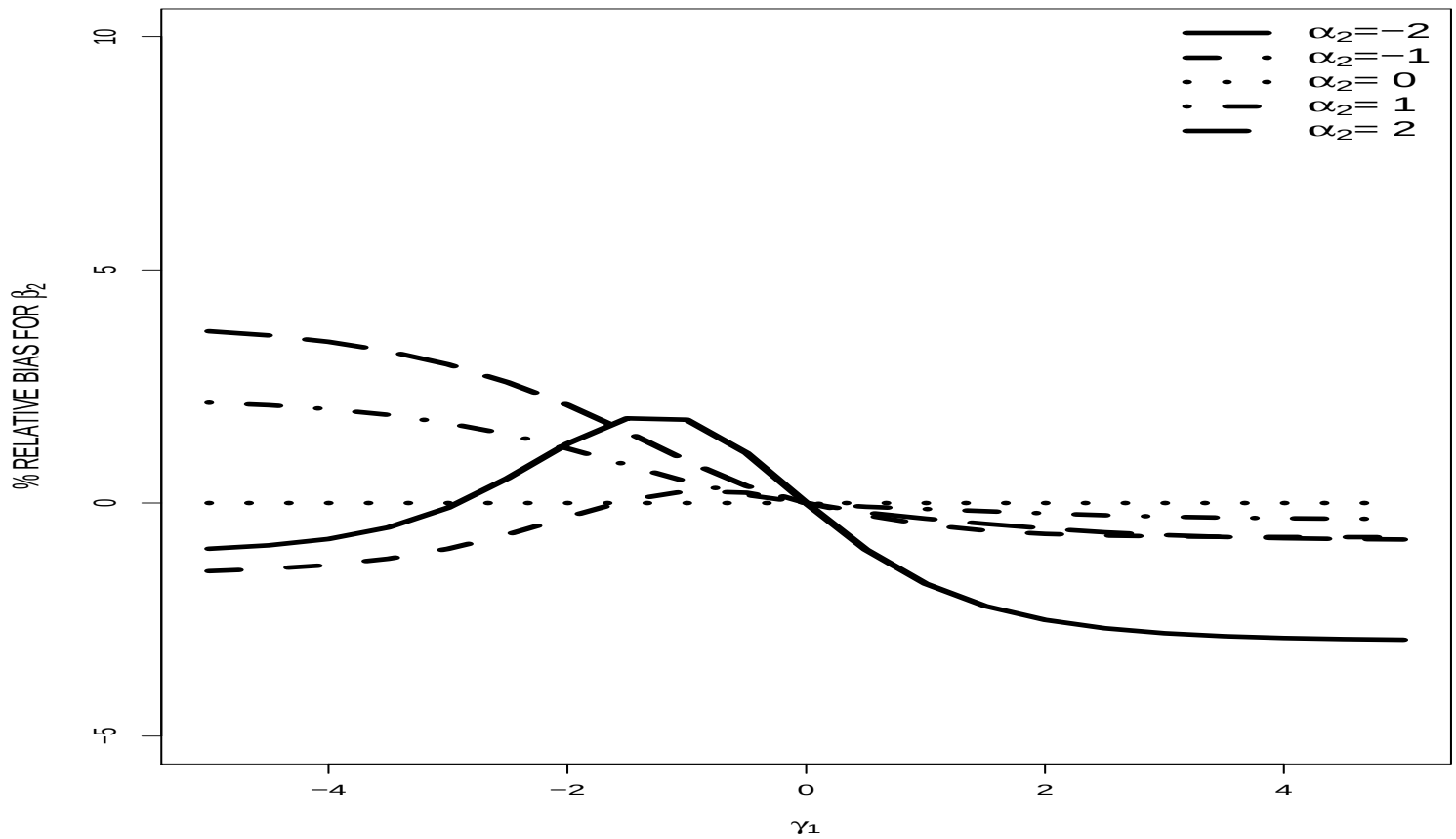

Figure: Asymptotic percent relative bias of β_2 with misspecified covariate model and missing data model

Application to the NACCUDS

Table: Frequency table for the responses and covariate for the missingness (X, Y)

Time	(m, m)	(o, m)	(m, o)	(o, o)
1	6.0	28.8	8.9	56.3
2	10.3	31.7	3.9	54.1
3	12.8	31.1	2.7	53.4
4	14.1	31.3	1.6	52.9

Application to the NACCUDS

Table: Parameter estimate for the NACCUDS: proposed method, $n = 16223$

Parameter	Est.	SE	p
(Intercept)	-0.136	0.106	0.198
SEX(F)	-0.203	0.025	<0.001
CVCHF	-0.031	0.063	0.618
DEPRESSION	0.679	0.029	<0.001
MMSE	-0.002	0.001	<0.001
FHDEM	0.181	0.028	<0.001
DIABETE	-0.124	0.038	0.001
HYPERT	-0.195	0.026	<0.001
EDUC	-0.002	0.001	0.040
AGE	0.006	0.001	<0.001

Application to the NACCUDS

Table: Parameter estimate for the NACCUDS: missing response only, $n = 15416$

Parameter	Est.	SE	p
(Intercept)	-0.272	0.110	0.013
SEX(F)	-0.113	0.026	<0.001
CVCHF	0.123	0.066	0.063
DEPRESSION	0.505	0.030	<0.001
MMSE	-0.007	0.001	<0.001
FHDEM	-0.004	0.029	0.897
DIABETE	-0.176	0.038	<0.001
HYPERT	-0.220	0.027	<0.001
EDUC	0.000	0.001	0.670
AGE	0.013	0.001	<0.001

Application to the NACCUDS

Table: Parameter estimate for the NACCUDS: missing covariate only, $n = 10755$

Parameter	Est.	SE	p
(Intercept)	0.198	0.142	0.163
SEX(F)	-0.040	0.032	0.215
CVCHF	0.044	0.080	0.579
DEPRESSION	0.451	0.034	<0.001
MMSE	-0.019	0.001	<0.001
FHDEM	-0.048	0.036	0.177
DIABETE	-0.177	0.047	<0.001
HYPERT	-0.212	0.034	<0.001
EDUC	-0.000	0.002	0.904
AGE	0.011	0.002	<0.001

Application to the NACCUDS

Table: Parameter estimate for the NACCUDS: complete case analysis, $n = 8724$

Parameter	Est.	SE	p
(Intercept)	0.283	0.162	0.081
SEX(F)	-0.022	0.037	0.551
CVCHF	-0.019	0.092	0.834
DEPRESSION	0.416	0.039	<0.001
MMSE	-0.021	0.001	<0.001
FHDEM	-0.067	0.040	0.099
DIABETE	-0.168	0.054	0.002
HYPERT	-0.212	0.039	<0.001
EDUC	0.002	0.002	0.252
AGE	0.013	0.002	<0.001

Summary

- Likelihood-based method is robust to the misspecification of the missing data process model
- Weighted GEE method is robust to the misspecification of the covariate model
- Our proposed method is robust to the misspecification of either the missing data process model or the covariate model
- Our proposed method can deal with intermittent missingness pattern for longitudinal data with both missing response and missing covariate

Questions?

Thank You!!!