Analysis of Incomplete Longitudinal Data: Some Issues and Methods

Grace Y. Yi

Department of Statistics and Actuarial Science

University of Waterloo

Tour Guide

- Visit 1:
  Missing Data Mechanisms and Inference Methods

- Visit 2:
  Methods for Incomplete Data

- Visit 3:
  Additional Challenge

$\Downarrow$

Concluding Remarks/Messages

Visit 1:

Missing Data Mechanisms and Inference Methods

Example: Waterloo Smoking Prevention Project (WSPP)

(Brown et al. 2002)

- 100 schools with 6294 students participated
- schools were randomized to receive either regular health education program or one of four anti-smoking programs
- smoking behavior questionnaire was scheduled annually from grades 6 to 12

6	7	8	9	10	11	12
N	N	N	N	N	Y	Y
N	N	Y	Y	N	N	N
N	Y	Y	Y	•	•	•
Y	Y	•	•	N	N	•

Illustration: Ad Hoc Methods (Cook, Zeng & Yi 2004)

- Setup:
  - response: $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{im})^\mathsf{T}$
  - mean: $\boldsymbol{\mu}_{ij} = \mathrm{E}(Y_{ij}|\mathbf{X}_i)$
  - regression model: $\mathrm{logit}\ \mu_{ij} = \mathbf{X}_{ij}^\mathsf{T}\boldsymbol{\beta}$
- Generalized Estimating Equations (GEE):
  - Use available observations
- Last Observation Carry Forward (LOCF):
  - pseudo-response Z_{ij} obtained by LOCF:

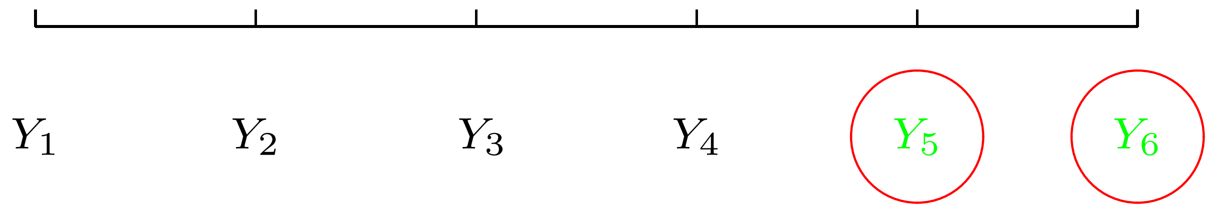
$$Z_{ij} = Y_{ij} \quad \text{if } j \le m_i$$
$$Z_{ij} = Y_{im_i} \quad \text{if } j > m_i$$

Numerical Study/Messages

e^{β_1}	e^{β_2}	e^{α_3}	LOCF BIAS	LOCF ESE	GEE BIAS	GEE ESE	IPWGEE BIAS	IPWGEE ESE
1.0	0.5	1.0	0.003	0.100	0.001	0.093	0.002	0.104
1.0	0.5	2.0	0.159	0.100	-0.016	0.093	-0.003	0.102
1.0	0.5	4.0	0.348	0.102	-0.033	0.097	0.001	0.104
1.0	2.0	1.0	-0.00	0.101	-0.000	0.095	0.001	0.106
1.0	2.0	2.0	0.066	0.102	-0.012	0.097	0.000	0.105
1.0	2.0	4.0	0.144	0.101	-0.031	0.096	0.003	0.105

- Imputation by LOCF and ordinary GEE can lead to considerable bias. LOCF tends to perform worse than unweighted GEE.
- IPWGEE leads to consistent estimators.
- There is a price of increased variability in the estimates arising from the IPWGEE.

A Closer Look at Impact of Missingness



- Missing Observations

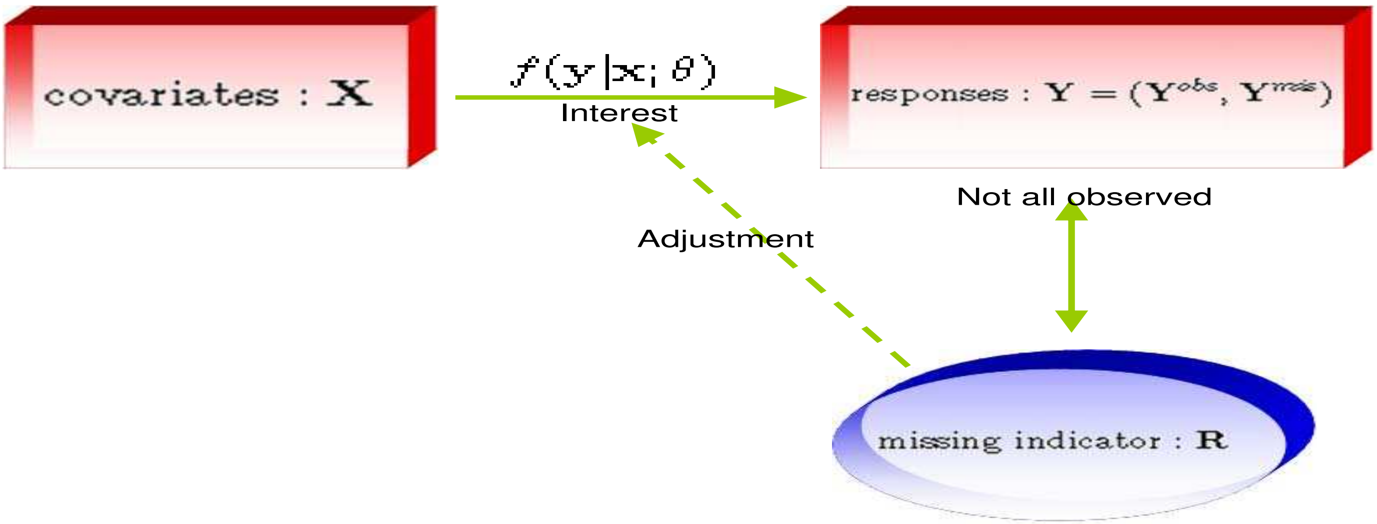
$$R_j = I(Y_j \text{ is observed})$$

- Association Structures

  measurements are correlated within the same cluster/subject

$$\mathbf{Y} = (Y_1, Y_2, ..., Y_m)^\mathsf{T} = (\mathbf{Y}^{obs}, \mathbf{Y}^{mis})^\mathsf{T}$$

Accounting for Response Missingness



Inference Framework:

$$f(\mathbf{Y}, \mathbf{X}, \mathbf{R}) \propto f(\mathbf{Y}, \mathbf{R} | \mathbf{X})$$

The Framework of Likelihood Method

Types of Models:

- Selection Model (e.g., Little and Rubin 1987)

$$f(\mathbf{Y}, \mathbf{R}|\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\alpha}) = f(\mathbf{Y}|\mathbf{X}; \boldsymbol{\beta}) f(\mathbf{R}|\mathbf{Y}, \mathbf{X}; \boldsymbol{\alpha})$$

- Pattern-Mixture Model (e.g., Little 1993)

$$f(\mathbf{Y}, \mathbf{R}|\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\alpha}) = f(\mathbf{Y}|\mathbf{R}, \mathbf{X}; \boldsymbol{\beta}) f(\mathbf{R}|\mathbf{X}; \boldsymbol{\alpha})$$

- Shared-Parameter Model (e.g., Wu and Carroll 1988)

$$f(\mathbf{Y}, \mathbf{R}|\mathbf{X}, \mathbf{u}; \boldsymbol{\beta}, \boldsymbol{\alpha}) = f(\mathbf{Y}|\mathbf{X}, \mathbf{u}; \boldsymbol{\beta}) f(\mathbf{R}|\mathbf{X}, \mathbf{u}; \boldsymbol{\alpha})$$

Implicit Assumption: $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are distinct

Selection Model/Missing Data Mechanism

$$
\begin{aligned}
L &\propto f(\mathbf{Y}^{obs}, \mathbf{R} | \mathbf{X}; \boldsymbol{\theta}, \boldsymbol{\alpha}) \\
&= \int f(\mathbf{R} | \mathbf{Y}^{obs}, \mathbf{Y}^{mis}, \mathbf{X}; \boldsymbol{\alpha}) f(\mathbf{Y}^{obs}, \mathbf{Y}^{mis} | \mathbf{X}; \boldsymbol{\beta}) \, d\mathbf{Y}^{mis}
\end{aligned}
$$

- MCAR: $f(\mathbf{R}|\mathbf{Y}^{obs}, \mathbf{Y}^{mis}, \mathbf{X}; \boldsymbol{\alpha}) = f(\mathbf{R}|\mathbf{X}; \boldsymbol{\alpha})$

  $\Longrightarrow \log L = \log f(\mathbf{R}|\mathbf{X}; \boldsymbol{\alpha}) + \log f(\mathbf{Y}^{obs}|\mathbf{X}; \boldsymbol{\beta})$

- MAR: $f(\mathbf{R}|\mathbf{Y}^{obs}, \mathbf{Y}^{mis}, \mathbf{X}; \boldsymbol{\alpha}) = f(\mathbf{R}|\mathbf{Y}^{obs}, \mathbf{X}; \boldsymbol{\alpha})$

  $\Longrightarrow \log L = \log f(\mathbf{R}|\mathbf{Y}^{obs}, \mathbf{X}; \boldsymbol{\alpha}) + \log f(\mathbf{Y}^{obs}|\mathbf{X}; \boldsymbol{\beta})$

- MNAR: $f(\mathbf{R}|\mathbf{Y}^{obs}, \mathbf{Y}^{mis}, \mathbf{X}; \boldsymbol{\alpha})$ depends on $\mathbf{Y}^{mis}$

  $\Longrightarrow \log \ f(\mathbf{Y}^{obs}|\mathbf{X}; \boldsymbol{\beta})$ is not obviously sorted out from $\log L$

Missing Data Mechanism & Likelihood Method

Remarks:

- such a classification allows us to treat the missing data process differently:
  - under MCAR and MAR, we can leave it unspecified when using likelihood based methods
  - under MNAR, inference based on the observed data is often biased
    - modeling the missing data process is commonly required
    - nonidentifiability could be an issue
- missing data mechanism is generally not verifiable

Key:

- assume distinct parameters for the response and missing data processes
- covariates $\mathbf{X}$ are precisely measured
- conditional inference on covariates is commonly employed

GEE Method

Introduction:

- Generalized Linear Models (GLM):

$$f(y) = \exp\{(y\theta - b(\theta))/a(\psi) + c(y,\psi)\}$$

  - mean: $\mu(\theta) = \mathrm{E}(Y) = b'(\theta)$
  - variance: $V(\theta) = \mathrm{var}(Y) = b''(\theta) \cdot a(\psi)$

- Likelihood Score:

$$S(\theta) = \{y - b'(\theta)\}/a(\psi)$$

GEE: (e.g., Liang & Zeger 1986)

$$\boldsymbol{U}(\boldsymbol{\theta}) = (\partial \boldsymbol{\mu}^{\mathsf{T}}/\partial \boldsymbol{\theta}) \cdot \boldsymbol{V}^{-1} \cdot (\boldsymbol{Y} - \boldsymbol{\mu})$$

- Key:
  - $E[\boldsymbol{U}(\boldsymbol{\theta})] = \mathbf{0}$
  - $\mathbf{V}$ may be replaced with a working matrix (efficiency loss may incur)

Missing Data Mechanism & GEE Method

Impact of Missingness:

- GEE applying to the observed data leads to consistent estimators if MCAR holds.
- GEE is not valid if data are incomplete with MAR or MNAR.

Why? (Robins et al. 1995)

$$
\begin{aligned}
& E_{Y|(X,Z)} E_{R|(Y,X,Z)} \left[\left(\frac{\partial \boldsymbol{\mu}_i^{\mathsf{T}}}{\partial \boldsymbol{\beta}} \right) \mathbf{V}_i^{-1} \text{diag}\,(R_{ij}, j = 1, ..., m)(\mathbf{Y}_i - \boldsymbol{\mu}_i) \right] \\
= \quad & E_{Y|(X,Z)} \left[\left(\frac{\partial \boldsymbol{\mu}_i^{\mathsf{T}}}{\partial \boldsymbol{\beta}} \right) \mathbf{V}_i^{-1} \cdot \text{diag}\,\{P(R_{ij} = 1|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i), j = 1, ..., m\} \cdot (\mathbf{Y}_i - \boldsymbol{\mu}_i) \right] \\
\neq \quad & E_{Y|(X,Z)} \left[\left(\frac{\partial \boldsymbol{\mu}_i^{\mathsf{T}}}{\partial \boldsymbol{\beta}} \right) \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \right] \\
= \quad & \mathbf{0}
\end{aligned}
$$

Take Home Messages

Missingness & Inference Methods:

Classification of the missing data mechanism depends on inference methods. In particular, if MCAR, MAR, and MNAR are three mechanisms to characterize the feature of missing data, then their impact would depend on the form of inference method:

- Likelihood:
  - MCAR and MAR: ignorable
  - MNAR: nonignorable
- GEE:
  - MCAR: ignorable
  - MAR and MNAR: nonignorable

Visit 2: Some Methods

- Likelihood Method

  - Missing Observations in Response Variable

- Marginal Method

  - Missing Observations in Both Response and Covariate Variables

UNIVERSITY OF Waterloo

Part 1: Likelihood Method (Chen, Yi & Cook 2010a)

Psoriatic Arthritis Data: (Gladman et al. 1995)

- Patients assessed annually for 10 years
- **Outcomes:** disease states = # of damaged joints
- **Covariates:**
  - duration of initial psoriasis (DUR)
  - SEX (0–F, 1–M)
  - age at onset of PsA (AGE)
  - family history of psoriasis (FM1, 0–No, 1–Yes)
  - family history of PsA (FM2, 0–No, 1–Yes)
  - erythrocyte sedimentation rate (ESR)

Sample Data of the Example

ASSESSMENT							0	1	2	3	4	5	6	7	8	9	10
ID	DUR	AGE	FM1	FM2	ESR	SEX	STATE										
1	21.5	33	0	0	6	1	1	.	.	.	.	.	1	.	1	.	1
2	38.3	40	1	0	36	0	1	.	.	.	.	.	.	.	.	.	1
3	15.1	25	0	0	4	1	1	.	.	.	.	.	.	.	.	.	4
4	7.1	34	0	0	83	0	1	.	.	1	1	.	1	1	1	.	1
5	7.4	28	1	1	16	1	1	.	.	.	2	.	4	4	4	4	4

Features:

- uni-directional transition (progressive)

$$\boxed{\text{STATE 1: 0 damaged joints}} \xrightarrow{\lambda_1(t)} \boxed{\text{STATE 2: 1–4 damaged joints}} \xrightarrow{\lambda_2(t)} \boxed{\text{STATE 3: 5–9 damaged joints}} \xrightarrow{\lambda_3(t)} \boxed{\text{STATE 4: 10+ damaged joints}}$$

- missing data

Inference Strategy

Two Processes:

- Response process $\{Y(t), t > 0\}$
- $t_{i1} < t_{i2} < \cdots < t_{iJ_i}$: variable assessment times
- History:

$$H_{ij}^y = \{Y_i(t_{ik}), k = 1, \ldots, j-1\}$$
$$H_{ij} = \{(t_{ik}, Y_i(t_{ik})), k = 1, \ldots, j-1\}$$

Likelihood:

- $L_i = \prod_{j=2}^{J_i} P(t_{ij}, Y_i(t_{ij})|H_{ij}) \propto \prod_{j=2}^{J_i} P(Y_i(t_{ij})|H_{ij}^y) P(t_{ij}|Y_i(t_{ij}), H_{ij})$
    - if the time of the assessment does not depend on the state process, then we can treat $\prod_{j=2}^{J_i} P(Y_i(t_{ij})|H_{ij}^y)$ the same as if it were the probability of the observed states
    - If $P(t_{ij}|Y_i(t_{ij}), H_{ij})$ does depend on $Y_i(t_{ij})$, then we must consider the full likelihood.

Handling Incompleteness

Notation:

- fix / pre-specify assessment times for every subject:

$$a_1, a_2, \ldots, a_J$$

- $R_{ij} = I\big($response is observed at time a_j for subject $i\big)$
- $\lambda^*_{ij} = P(R_{ij} = 1 | H^r_{ij}, \mathbf{Y}_i, \mathbf{X}_i)$: conditional probability

  H^r_{ij}: the history of the missing indicators until the $(j-1)$st time point

Logistic regression:

$$\mathrm{logit}(\lambda^*_{ij}) = \mathbf{u}^{\mathsf{T}}_{ij}\boldsymbol{\alpha}$$

$\mathbf{u}_{ij}$ features various missing mechanisms:

MCAR; MAR; MNAR

Inference Methods

Continuous Time Models:

- Transition Intensity:

$$\lambda_k(t|\mathbf{X}_k) = \lambda_{0k}(t)\exp(\mathbf{X}_k^{\mathsf{T}}\boldsymbol{\beta}_k), \quad k = 1, \ldots, K-1$$

Method 1: Observed likelihood (e.g. 1st order)

$$\begin{aligned} P(\mathbf{Y}_i^{obs}, \mathbf{R}_i|\mathbf{X}_i) &= \int P(\mathbf{R}_i|\mathbf{Y}_i, \mathbf{X}_i) \cdot P(\mathbf{Y}_i|\mathbf{X}_i) d\mathbf{Y}_i^{mis} \\ &\propto \int \prod_{j=2}^{J} P(R_{ij}|R_{i,j-1},, \mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\alpha}) \cdot \prod_{j=2}^{J} P(Y_{ij}|Y_{i,j-1}, \mathbf{X}_i, \boldsymbol{\beta}) dY_i^{mis} \end{aligned}$$

Method 2: EM algorithm

$\Rightarrow$ the parameter estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}^{\mathsf{T}}, \hat{\boldsymbol{\beta}}^{\mathsf{T}})^{\mathsf{T}}$

Numerical Assessment

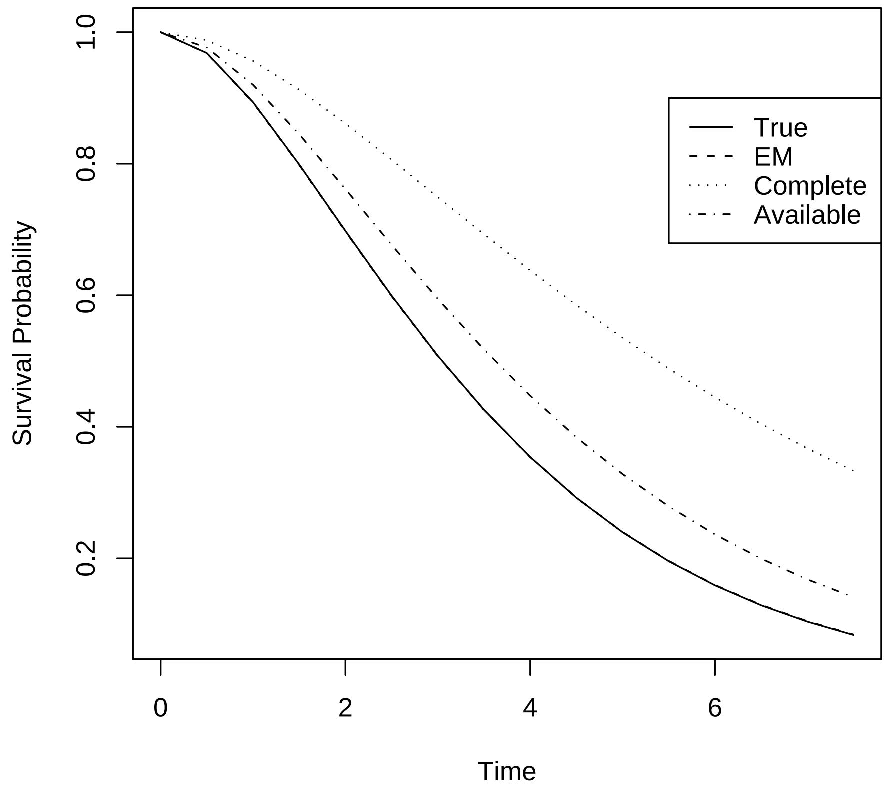
$$\boxed{1} \xrightarrow{\lambda_1} \boxed{2} \xrightarrow{\lambda_2} \boxed{3}$$

$$\lambda_k = \lambda_{0k} e^{\beta_k x}, \quad k = 1, \ldots, K-1,$$

$$\text{logit}(\lambda_{ij}^*) = \alpha_0 + \alpha_1 r_{i,j-1} + \alpha_2 y_{i,j-1} + \alpha_3 y_{ij} + \alpha_4 x_i \quad \Longrightarrow \text{MNAR}$$

	EM				Complete Case			
Para.	Bias	SEL	ESE	CP%	Bias	SEL	ESE	CP%
λ_{01}	0.002	0.037	0.037	95.6	-0.162	0.078	0.088	39.7
λ_{02}	0.005	0.057	0.057	95.2	-0.086	0.124	0.154	60.9
β_1	0.003	0.116	0.116	95.4	0.529	0.478	0.460	86.1
β_2	-0.003	0.126	0.127	94.7	0.215	0.534	0.533	93.4

Graphical Comparison



Comparisons of the estimated survival functions $(S(t) = 1 - P_{13}(t))$ obtained from the three analyses with the true curve for the case without covariates ($K = 3$ and $J = 5$).

Part 2: Marginal Method (Chen, Yi & Cook 2010b)

Revisit WSPP Data: (Brown et al. 2002)

- 100 schools participated; questionnaire was scheduled to be administered annually from grades 6 to 8

Objectives: include evaluating

1. whether the intensive anti-smoking education program is more effective than standard school education program
2. whether students' smoking behavior changes over time
3. what factors have influence on the children's smoking behavior

Response:

- smoking status

Covariates:

- treatment, social models risk score (SMR), sex, grade

Summary of WSPP Data

- Data: 4400 students from grades 6 to 8

response			SMR		
6	7	8	6	7	8
N	N	Y	$\surd$	•	$\surd$
N	Y	•	$\surd$	$\surd$	$\surd$
N	•	Y	$\surd$	•	•

variable	proportion
missing Y	13.7%
missing X (SMR)	15.2%
missing both X and Y	5.1%

Features and Objective

Features:

- individuals are followed over time
- response Y with covariates $(X, \mathbf{Z})$ is scheduled to be recorded at each assessment
- missing observations arise in both response Y and covariate X
- Interest:

$$P(Y = 1|X, \mathbf{Z}) = E[Y|X, \mathbf{Z}]$$ - mean structure

UNIVERSITY OF Waterloo

Model Setup

Response Model:

- Mean and Variance:
  - $\mu_{ij} = \mathrm{E}(Y_{ij} | \mathbf{X}_i, \mathbf{Z}_i)$
  - $v_{ij} = \mathrm{var}(Y_{ij} | \mathbf{X}_i, \mathbf{Z}_i)$
- Regression Model:

$$\mu_{ij} = g^{-1}(X_{ij}\beta_x + \mathbf{Z}_{ij}^{\mathsf{T}}\boldsymbol{\beta}_z)$$
$$v_{ij} = \phi h^{-1}(g^{-1}(X_{ij}\beta_x + \mathbf{Z}_{ij}^{\mathsf{T}}\boldsymbol{\beta}_z))$$

- Interest: $\boldsymbol{\beta} = (\beta_x, \boldsymbol{\beta}_z^{\mathsf{T}})^{\mathsf{T}}$

Two Missing Data Processes:

$$R_{ij}^{y} = I(Y_{ij} \text{ is observed})$$
$$R_{ij}^{x} = I(X_{ij} \text{ is observed})$$

Inference Strategy

Usual GEE:

$$\sum_{i=1}^{n} D_i[A_i^{-1/2} C_i^{-1} A_i^{-1/2}](\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

- $A_i = \mathrm{diag}\{v_{ij}\}, \quad D_i = \partial \boldsymbol{\mu}_i^T / \partial \boldsymbol{\beta}$
- C_i: working correlation matrix

IPWGEE:

$$\mathbf{U}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = D_i(A_i^{-1/2}[C_i^{-1} \bullet \Delta_i(\boldsymbol{\alpha})]A_i^{-1/2})(\mathbf{Y}_i - \boldsymbol{\mu}_i)$$

- $\Delta_i(\boldsymbol{\alpha}) = [I(R_{ij}^x = 1, R_{ik}^x = 1, R_{ik}^y = 1)/\pi_{ijk}^{xy}]$
- $\pi_{ijk}^{xy} = P(R_{ij}^x = 1, R_{ik}^x = 1, R_{ik}^y = 1 | \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$

Remark:

- Key:

$$E_{(R_i^y, R_i^x)|(Y_i, X_i, Z_i)}[C_i^{-1} \bullet \Delta_i(\boldsymbol{\alpha})] = C_i^{-1}$$

Improving Efficiency

- Remark:
  $\mathrm{U}_i(\beta, \alpha)$ includes merely the measurements with the patterns:
  $$(Y_{ij}, X_{ij}) = (\surd, \surd), (\bullet, \surd), \text{ but not } (\surd, \bullet)$$

- Augmented IPWGEE:
  $$\mathbf{U}_i^{\dagger}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{U}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) - \eta \mathbf{A}_i$$

- Key: make $\mathbf{A}_i$
  - have zero mean
  - be expressed in terms of the observed data

Asymptotic Properties

Estimators:

- $\tilde{\beta}^{\dagger}$: estimator obtained from the augmented IPWGEE
- $\hat{\beta}$: estimator obtained from the IPWGEE

Theorem: Under regularity conditions,

1. $\sqrt{n}(\tilde{\boldsymbol{\beta}}^{\dagger} - \boldsymbol{\beta}) \rightarrow_d N(\mathbf{0}, \Gamma^{-1}\Sigma^{\dagger}[\Gamma^{-1}]^{\mathsf{T}})$, as $n \rightarrow \infty$

   where $\Sigma^{\dagger} = \mathrm{var}\{\mathrm{Res}(\mathbf{U}_i(\beta,\alpha), \mathbf{H}_i^*)\}$

   $$\mathbf{H}_i^* = (\mathbf{A}_i^T(\alpha), \mathbf{S}_i^T(\alpha))^T$$

2. If $\eta \neq 0$, then $\tilde{\boldsymbol{\beta}}^{\dagger}$ is more efficient than $\hat{\boldsymbol{\beta}}$ asymptotically.

UNIVERSITY OF Waterloo

Empirical Studies

			$\alpha_2 = 0.1$						$\alpha_2 = 2.0$					
			β_0			β_1			β_0			β_1		
ψ_2	ψ_3	Method*	Bias†	ESE‡	CP%	Bias	ESE	CP%	Bias	ESE	CP%	Bias	ESE	CP%
2	4	GEE	-20.3	0.11	85	2.8	0.14	94	10.7	0.11	80	-1.9	0.11	81
		IPWGEE	-0.1	0.35	95	-1.4	0.37	95	-0.9	0.13	94	-0.9	0.12	95
		AIPWGEE	0.7	0.33	94	1.2	0.34	94	-1.1	0.12	94	-1.0	0.12	95
2	2	GEE	-21.4	0.12	85	2.8	0.17	93	10.8	0.12	78	1.2	0.12	81
		IPWGEE	0.2	0.35	95	-0.8	0.38	95	-0.1	0.14	95	-0.6	0.12	95
		AIPWGEE	-0.6	0.32	95	-0.6	0.38	95	-0.9	0.13	95	-0.8	0.12	94
1	2	GEE	-19.0	0.11	87	2.1	0.16	93	9.0	0.12	90	-0.6	0.12	84
		IPWGEE	0.9	0.35	94	0.7	0.39	95	-1.3	0.14	95	-0.8	0.13	95
		AIPWGEE	0.3	0.34	95	-0.8	0.38	95	-1.4	0.13	94	-1.1	0.13	94
1	1	GEE	-19.1	0.12	87	3.9	0.16	93	8.5	0.12	84	-2.0	0.12	84
		IPWGEE	0.7	0.37	95	-0.3	0.45	94	-0.4	0.15	95	0.3	0.13	95
		AIPWGEE	-0.9	0.33	95	-0.7	0.44	94	-0.5	0.13	95	0.1	0.13	95

* true values: $\beta_0 = \log(1.5)$ and $\beta_1 = \log(0.5)$. Correlation coefficient for responses: 0.6

† Relative bias defined by $(\hat{\beta} - \beta_{true})/\beta_{true} \times 100$.

‡ ESE: empirical standard error for the 2000 times simulation

Validity of the Method

Remarks:

- Correct Mean Structure: $\mu_{ij} = E[Y_{ij}|\mathbf{X}_i, \mathbf{Z}_i]$
- Correct Weight: consistent estimate of π_{ij}
  - correct modeling the missing data process
  - MAR ensures the possibility that the α parameter may be consistently estimated and avoids nonidentifiability of model parameters
- Questions:

  MAR or MNAR is not testable simply based on data; how can we gain confidence in the model we use?

  - sensitivity analyses
  - alternative: assess a particular model by comparing its fit to expanded models including additional terms. This provides focused tests of the adequacy of a particular model which are easily interpreted.

Visit 3: Additional Challenge

An Example

A Data Set from Framingham Heart Study:

- 1672 patients were scheduled for 5 visits
- 24% patients drop out of the study
- response: obesity status
- covariates: age
  systolic blood pressure (SBP)

Additional Challenge

Features:

$$\begin{array}{cccccc} Y_{i1} & Y_{i2} & Y_{i3} & Y_{i4} & Y_{i5} & Y_{i6} \\ X_{i1}/X^*_{i1} & X_{i2} & X_{i3} & X_{i4} & X_{i5} & X_{i6} \end{array}$$

- missing observations
- measurement error in covatiates

$(X_{ij}, \mathbf{Z}_{ij}^{\mathsf{T}})^{\mathsf{T}}$: covariate vector

X_{ij}: error-prone (observed version: X^*_{ij})

$\mathbf{Z}_{ij}$: error-free

An Illustration of Error Effects

Response Model:

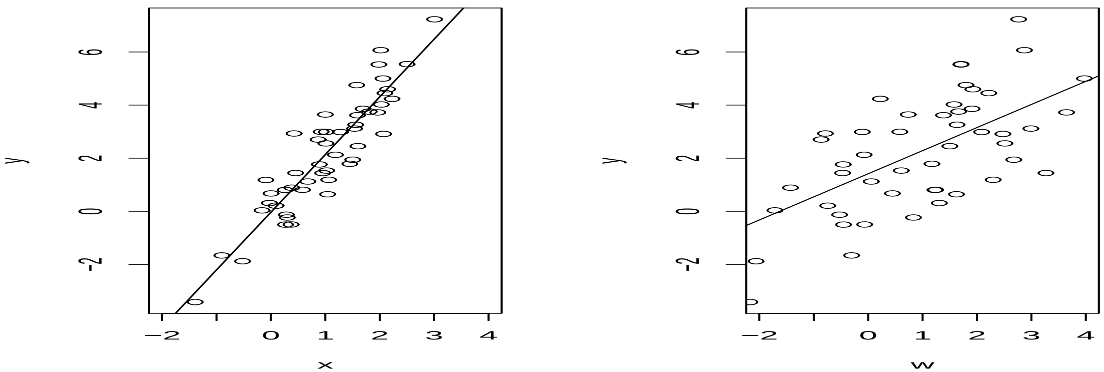
$$Y = \beta_0 + \beta X + \epsilon, \quad X \sim (\mu_x, \sigma_x^2), \; \epsilon \sim (0, \sigma_\epsilon^2), \text{ indep.}$$

Error Model:

$$X^* = X + e, \qquad e \sim (0, \sigma_e^2)$$

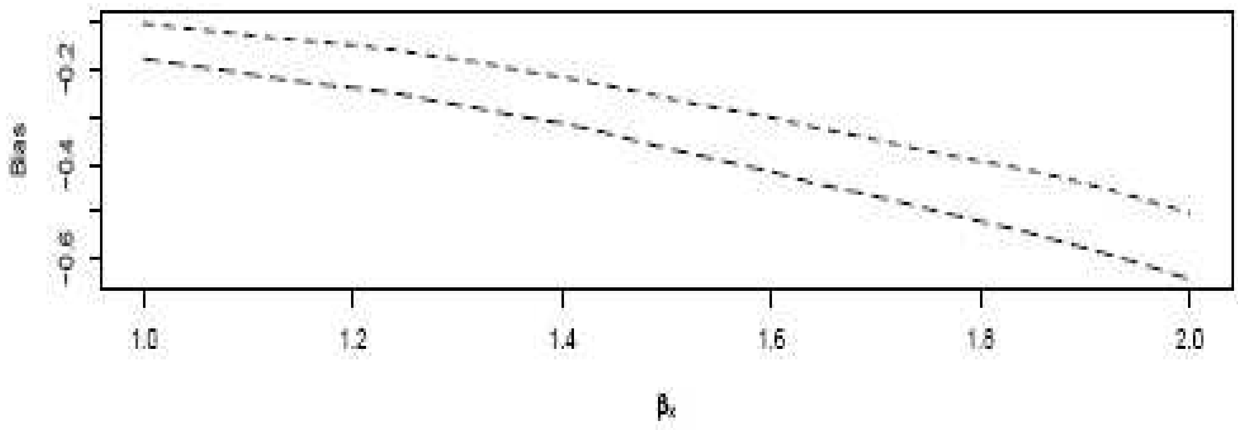
If naively replacing X with X^*, then

- $\beta^* = \left(\frac{\sigma_x^2}{\sigma_x^2+\sigma_e^2}\right)\beta$
- $\mathrm{var}(Y|X^*) = \mathrm{var}(Y|X) + \frac{\beta^2\sigma_e^2\sigma_x^2}{\sigma_x^2+\sigma_e^2}$



Why? - Joint Impact of Missingness and Error

Bias Analysis: (Yi, Liu & Wu 2010)



Messages:

- Estimates of response parameters are usually biased if missingness and measurement error are not properly accounted for.
- Biases induced by ignoring missingness and measurement error are usually complex.

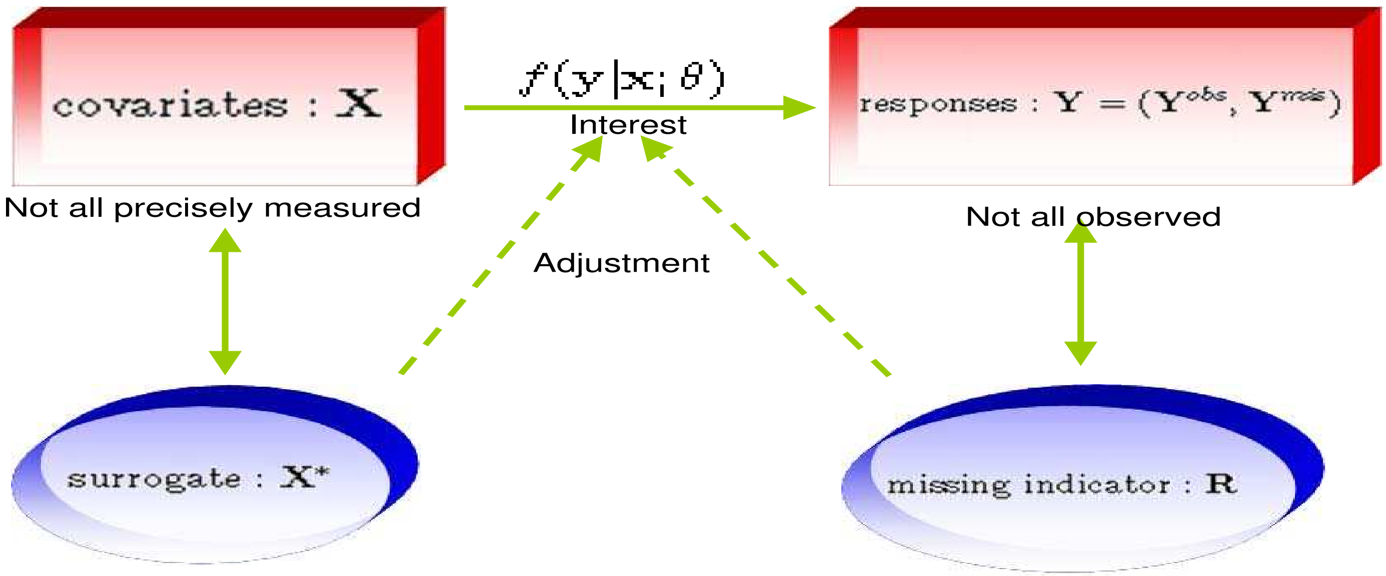
How? - Framework for Valid Inference

$$\text{covariates}: \mathbf{X} \xrightarrow[\text{Interest}]{f(y|\mathbf{x};\,\theta)} \text{responses}: \mathbf{Y} = (\mathbf{Y}^{obs}, \mathbf{Y}^{mis})$$

Not all precisely measured

Not all observed

How? - Framework for Valid Inference



Inference Framework:

$$f(\mathbf{Y}, \mathbf{X}, \mathbf{X}^*, \mathbf{R})$$

UNIVERSITY OF Waterloo

Marginal Analysis (Yi 2005, 2008)

Response Model:

marginal mean and variance structures

Inference Strategy:

- Step 1:
  GEE: constructed under the true model
- Step 2:
  IPWGEE: adjust for bias induced by missingness
- Step 3:
  correct for error effects

Sensitivity Analyses of the Motivating Example

		β_x (SBP)			β_z (AGE)		
σ	Analysis	Est.	SE	p-value	Est.	SE	p-value
0	Naive	3.0819	0.2900	< 0.0001	-0.0058	0.0056	0.3009
	Prop.	3.0653	0.2896	< 0.0001	-0.0055	0.0057	0.3347
0.50	Naive	0.4321	0.0811	< 0.0001	0.0147	0.0053	0.0056
	Prop.	0.7648	0.1215	< 0.0001	0.0120	0.0054	0.0253
0.75	Naive	0.2029	0.0542	0.0002	0.0165	0.0053	0.0019
	Prop.	0.3717	0.0821	< 0.0001	0.0153	0.0054	0.0046
1.00	Naive	0.1142	0.0406	0.0049	0.0172	0.0053	0.0012
	Prop.	0.2136	0.0619	0.0006	0.0166	0.0054	0.0021

- If error is absent, then both analyses yield very compatible results.
- As measurement error becomes more substantial, SBP tends to become less significant while stronger evidence of AGE effects is observed.

Concluding Remarks/Take Home Messages

- Statistical inference methods are commonly challenged by the “imperfectness" of data.
  - Missingness and measurement error exist in many settings.
  - Ignoring these features may lead to seriously biased results.
  - Properly addressing these features is needed:

    modeling additional processes is often required

- In particular, in handling missingness:
  - In the absence of measurement error, whether or not missingness can be ignored depends on the form of inference methods.
    - MCAR and MAR can be ignored if using likelihood based methods.
    - MCAR can be ignored if using the GEE method.
  - In the presence of measurement error, missingness is not ignorable in generable.