MARGINAL METHODS FOR INCOMPLETE CLUSTERED LONGITUDINAL BINARY DATA

Richard Cook and Grace Yi
Department of Statistics and Actuarial Science
University of Waterloo

Missing Data Workshop
Fields Institute Toronto, Ontario

OUTLINE

1 LONGITUDINAL BINARY DATA

2 CLUSTERED INCOMPLETE LONGITUDINAL DATA

3 IPW ESTIMATING EQUATIONS WITH CLUSTERED DROP-OUT PROCESS

4 SIMULATION STUDY AND APPLICATION

5 GENERAL REMARKS

NOTATION

- Y_{jk} - response for jth subject at kth assessment
- $\boldsymbol{Y}_j = (Y_{j1}, ..., Y_{jK})'$
- $\boldsymbol{x}_{jk}$ - covariate vector for jth subject at kth assessment; $\boldsymbol{x}_j = (\boldsymbol{x}'_{j1}, \ldots, \boldsymbol{x}'_{jK})'$
- $\mu_{jk} = E(Y_{jk}|\boldsymbol{x}_j) = g(\boldsymbol{x}'_{jk}\boldsymbol{\beta})$
- $\boldsymbol{\mu}_j = (\mu_{j1}, ..., \mu_{jK})'$
- $V_j(\boldsymbol{\beta}, \boldsymbol{\delta}) = \mathrm{cov}(\boldsymbol{Y}_j|\boldsymbol{x}_j)$

LIANG AND ZEGER (1986)

$$U_1(\boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{j=1}^{J} \left(\frac{\partial \boldsymbol{\mu}'_j}{\partial \boldsymbol{\beta}}\right) [V_j(\boldsymbol{\beta}, \boldsymbol{\delta})]^{-1} (\boldsymbol{Y}_j - \boldsymbol{\mu}_j)$$

PRENTICE (1988)

- $\boldsymbol{W}_j = (Y_{j1}Y_{j2}, Y_{j1}Y_{j3}, ..., Y_{j,K-1}Y_{jK})'$
- $\boldsymbol{\nu}_j = E(\boldsymbol{W}_j|\boldsymbol{x}_j)$

$$U(\boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{j=1}^{J} \left(\frac{\partial(\boldsymbol{\mu}'_i, \boldsymbol{\nu}'_i)}{\partial(\boldsymbol{\beta}', \boldsymbol{\delta}')'}\right) [\mathrm{cov}(\boldsymbol{Y_j}, \boldsymbol{W_j})]^{-1} \begin{pmatrix} \boldsymbol{Y_j} - \boldsymbol{\mu_j} \\ \boldsymbol{W_j} - \boldsymbol{\nu_j} \end{pmatrix}$$

CAREY ET AL. (1993)

- $\xi_{jk:k'} = E(Y_{jk}|Y_{jk'} = y_{jk'}, \boldsymbol{x}_j)$
- $\boldsymbol{\xi}_j = (\xi_{j1:2}, \xi_{j1:3}, ..., \xi_{j,K-1:K})'$
- $E_{jkk'} = Y_{jk} - \xi_{jk:k'}$, $1 \leq k < k' \leq K$
- $\boldsymbol{E}_j = (Y_{j1} - \xi_{j1:2}, Y_{j1} - \xi_{j1:3}, ..., Y_{j,K-1} - \xi_{j,K-1:K})'$
- $G_j(\boldsymbol{\beta}, \boldsymbol{\delta}) = \text{diag}(\xi_{j1:2}(1 - \xi_{j1:2}), ..., \xi_{j,K-1:K}(1 - \xi_{j,K-1:K}))$

$$
\begin{aligned}
U_1(\boldsymbol{\beta}, \boldsymbol{\delta}) &= \sum_{j=1}^{J} \left(\frac{\partial \boldsymbol{\mu}_j'}{\partial \boldsymbol{\beta}}\right) [V_j(\boldsymbol{\beta}, \boldsymbol{\delta})]^{-1} (\boldsymbol{Y}_j - \boldsymbol{\mu}_j) = \mathbf{0} \\
U_2(\boldsymbol{\beta}, \boldsymbol{\delta}) &= \sum_{j=1}^{J} \left(\frac{\partial \boldsymbol{\xi}_j'}{\partial \boldsymbol{\delta}}\right) [G_j(\boldsymbol{\beta}, \boldsymbol{\delta})]^{-1} \boldsymbol{E}_j = \mathbf{0}
\end{aligned}
$$

NOTATION

- Y_{jk} response for jth subject at kth assessment
- $\boldsymbol{Y}_j = (Y_j^{(o)}, Y_j^{(m)})$
- $R_{jk} = I(Y_{jk}$ is observed)
- $\boldsymbol{R}_j = (R_{j1}, R_{j2}, ..., R_{jK})$

TYPES OF MISSING DATA MECHANISMS

MISSING COMPLETELY AT RANDOM

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

MISSING AT RANDOM

$$f(\boldsymbol{R}_j|\boldsymbol{Y}_j, \boldsymbol{X}_j; \boldsymbol{\alpha}) = f(\boldsymbol{R}_j|Y_j^{(o)}, \boldsymbol{X}_j; \boldsymbol{\alpha})$$

MISSING NOT AT RANDOM

$$f(\boldsymbol{R}_j|\boldsymbol{Y}_j, \boldsymbol{X}_j; \boldsymbol{\alpha}) = f(\boldsymbol{R}_j|\boldsymbol{Y}_j^{(o)}, \boldsymbol{Y}_j^{(m)}, \boldsymbol{X}_j; \boldsymbol{\alpha})$$

PROBLEM WITH NAIVE GEE WHEN DATA ARE MAR

We can write the equations as

$$\sum_{j=1}^{J} \left(\frac{\partial \boldsymbol{\mu}_j'}{\partial \boldsymbol{\beta}}\right) [V_j(\boldsymbol{\beta}, \boldsymbol{\delta})]^{-1} \; D_j \; (\boldsymbol{Y}_j - \boldsymbol{\mu}_j) = \mathbf{0}$$

where $D_j = \mathrm{diag}(I(R_{jk} = r_{jk})r_{jk}, k = 1, \ldots, K)$.

Taking the expectation conditional on $\boldsymbol{Y}_i^{(o)}$ and $\boldsymbol{x}_j$ gives

$$\sum_{j=1}^{J} \left(\frac{\partial \boldsymbol{\mu}_j'}{\partial \boldsymbol{\beta}}\right) [V_j(\boldsymbol{\beta}, \boldsymbol{\delta})]^{-1} E(D_j|\boldsymbol{Y}_j^{(o)}, \boldsymbol{x}_j)(\boldsymbol{Y}_j - \boldsymbol{\mu}_j)$$

where $E(D_j|\boldsymbol{Y}_j^{(o)}, \boldsymbol{X}_j) = \mathrm{diag}(P(R_{jk} = 1|\boldsymbol{Y}_j^{(o)}, \boldsymbol{X}_j)r_{jk}, k = 1, \ldots, K)$.

INVERSE PROBABILITY WEIGHTED ESTIMATING EQUATIONS (ROBINS et al., 1995)

$$\sum_{j=1}^{J} \left(\frac{\partial \boldsymbol{\mu}_j'}{\partial \boldsymbol{\beta}}\right) [V_j(\boldsymbol{\beta}, \boldsymbol{\delta})]^{-1} \Delta_j(\boldsymbol{Y}_j - \boldsymbol{\mu}_j) = \mathbf{0}$$

where $\Delta_j = \mathrm{diag}(r_{jk}/P(R_{jk} = 1|\boldsymbol{Y}_j^{(o)}, \boldsymbol{X}_j), k = 1, \ldots, K)$

SETTINGS WITH CLUSTERED LONGITUDINAL DATA

EXAMPLE 1: COMMUNITY INTERVENTION TRIALS

- communities are randomized to interventions
- individuals assessed periodically to collect data on behaviour

EXAMPLE 2: FAMILY STUDIES

- families form the clusters (J_i subjects in family i)
- each family member is assessed repeatedly for dietary status

EXAMPLE 3: SCHOOL-BASED LONGITUDINAL STUDIES

- schools form the clusters (J_i students for school i)
- each student is followed longitudinally for smoking status
- e.g. Waterloo Smoking Prevention Project (WSPP)

POSSIBLE MISSING DATA STRUCTURES

		ASSESSMENT					
CLUSTER	SUBJECT	1	2	3	4	5	6
1	1	$\surd$	$\surd$	$\surd$	$\bullet$	$\bullet$	$\bullet$
	2	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\bullet$
	$\vdots$						
	J_1	$\surd$	$\surd$	$\bullet$	$\bullet$	$\bullet$	$\bullet$
2	1	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\bullet$
	2	$\surd$	$\bullet$	$\surd$	$\bullet$	$\bullet$	$\surd$
	$\vdots$						
	J_2	$\surd$	$\surd$	$\bullet$	$\surd$	$\surd$	$\bullet$
$\vdots$	$\vdots$						
I	1	$\surd$	$\surd$	$\surd$	$\surd$	$\bullet$	$\surd$
	2	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
	$\vdots$						
	J_I	$\surd$	$\bullet$	$\surd$	$\surd$	$\surd$	$\surd$

- $\boldsymbol{Y}_{ij} = (Y_{ij1}, ..., Y_{ijK})'$ $\qquad \boldsymbol{Y}_i = (\boldsymbol{Y}'_{i1}, \boldsymbol{Y}'_{i2}, ..., \boldsymbol{Y}'_{iJ_i})'$
- $\boldsymbol{\mu}_{ij} = (\mu_{ij1}, ..., \mu_{ijK})'$ $\qquad \boldsymbol{\mu}_i = (\boldsymbol{\mu}'_{i1}, \boldsymbol{\mu}'_{i2}, ..., \boldsymbol{\mu}'_{iJ_i})'$
- $\boldsymbol{R}_{ij} = (R_{ij1}, ..., R_{ijK})'$ $\qquad \boldsymbol{R}_i = (\boldsymbol{R}'_{i1}, \boldsymbol{R}'_{i2}, ..., \boldsymbol{R}'_{iJ_i})'$
- $\boldsymbol{x}_{ijk} = (1, x_{ijk1}, ..., x_{ijk,p-1})'$ $\qquad \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_{p-1})'$
- J_i subjects in cluster i, $i = 1, 2, ..., I$

MODEL FOR MARGINAL MEAN RESPONSE

- $\mu_{ijk} = E(Y_{ijk}|\boldsymbol{x}_i) = P(Y_{ijk} = 1|\boldsymbol{x}_i)$

$$\log(\mu_{ijk}/(1-\mu_{ijk})) = \boldsymbol{x}'_{ijk}\boldsymbol{\beta}$$

where $\boldsymbol{x}_{ijk}$ is a $p \times 1$ vector of covariates.

MODEL FOR ASSOCIATIONS BETWEEN RESPONSES

$$\psi_{ijk:j'k'} = \frac{P(Y_{ijk}=1, Y_{ij'k'}=1|\boldsymbol{x}_i)/P(Y_{ijk}=0, Y_{ij'k'}=1|\boldsymbol{x}_i)}{P(Y_{ijk}=1, Y_{ij'k'}=0|\boldsymbol{x}_i)/P(Y_{ijk}=0, Y_{ij'k'}=0|\boldsymbol{x}_i)}$$

and

$$\log \psi_{ijk:j'k'} = \boldsymbol{z}'_{ijk:j'k'}\boldsymbol{\phi}$$

EXAMPLE:

$$\log \psi_{ijk:j'k'} = \phi_0 + \phi_1 I(j=j') + \phi_2 I(k=k')$$

Longitudinal Association: odds ratio is $\exp(\phi_0 + \phi_1)$

Cross-sectional Association: odds ratio is $\exp(\phi_0 + \phi_2)$

Mixed Association: odds ratio is $\exp(\phi_0)$

$$\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\phi}')'$$

MODEL FOR MISSING DATA PROCESS

- $\boldsymbol{Y}_{ij}^{(o)}$ - observed components of $\boldsymbol{Y}_{ij}$
- $H_{ijk} = \{Y_{ij\ell},\ \ell = 1, 2, ..., k-1\}$
- $H_{ijk}^{(o)}$ = history of observed components
- $H_{ijk}^{R} = \{R_{ij\ell},\ \ell = 1, 2, ..., k-1\}$

MODEL FOR THE MEAN

- $\lambda_{ijk} = P(R_{ijk} = 1 | H_{ijk}^{R}, H_{ijk}^{(o)}, \boldsymbol{x}_{ijk}) = \lambda_{ijk}(\boldsymbol{\vartheta}_k)$

$$\log(\lambda_{ijk}/(1-\lambda_{ijk})) = \boldsymbol{u}'_{ijk}\boldsymbol{\vartheta}_k$$

MODEL FOR THE CROSS-SECTIONAL ASSOCIATION

- $\psi^{*}_{ijk:j'k} = \dfrac{P(R_{ijk}=1, R_{ij'k}=1|\boldsymbol{u}_i)/P(R_{ijk}=0, R_{ij'k}=1|\boldsymbol{u}_i)}{P(R_{ijk}=1, R_{ij'k}=0|\boldsymbol{u}_i)/P(R_{ijk}=0, R_{ij'k}=0|\boldsymbol{u}_i)}$

$$\log \psi^{*}_{ijk:j'k} = \boldsymbol{w}'_{ijk:j'k}\boldsymbol{\gamma}_k$$

where $\boldsymbol{\alpha} = (\boldsymbol{\vartheta}', \boldsymbol{\gamma}')'$, $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}'_1, ..., \boldsymbol{\vartheta}'_K)'$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}'_1, ..., \boldsymbol{\gamma}'_K)'$

EQUATIONS FOR MEAN OF MISSING DATA PROCESS

- $\boldsymbol{R}_{i(k)} = (R_{i1k}, ..., R_{iJ_ik})'$
- $\boldsymbol{\lambda}_{i(k)} = (\lambda_{i1k}, .., \lambda_{iJ_ik})'$
- $\mathrm{var}(\boldsymbol{R}_{i(k)}|H^R_{ijk}, H^{(o)}_{ijk}, \boldsymbol{x}_i, j = 1, 2, \ldots, J_i) = V_{i(k)}(\boldsymbol{\vartheta}_k, \boldsymbol{\gamma}_k)$

$$S_{1k}(\boldsymbol{\vartheta}_k, \boldsymbol{\gamma}_k) = \sum_{i=1}^{I} \left(\frac{\partial \boldsymbol{\lambda}'_{i(k)}}{\partial \boldsymbol{\vartheta}_k}\right) [V_{i(k)}(\boldsymbol{\vartheta}_k, \boldsymbol{\gamma}_k)]^{-1} (\boldsymbol{R}_{i(k)} - \boldsymbol{\lambda}_{i(k)})$$

EQUATIONS FOR ASSOCIATION OF MISSING DATA PROCESS

- $\zeta_{ijk:j'k} = E(R_{ijk} | R_{ij'k} = r_{ij'k})$ where

$$\zeta_{ijk:j'k} = \text{expit}\left(r_{ij'k} \log \psi_{ijk:j'k} + \log\left(\frac{\lambda_{ijk} - \lambda_{ijk:j'k}}{1 - \lambda_{ijk} - \lambda_{ij'k} + \lambda_{ijk:j'k}}\right)\right)$$

  where $\lambda_{ijk:j'k} = P(R_{ijk} = 1, R_{ij'k} = 1 | \boldsymbol{u}_i)$.

- $\boldsymbol{\zeta}_{ik} = (\zeta_{i1k:2k}, \ldots, \zeta_{i,J_i-1,k:J_i,k})'$

- $\boldsymbol{Q}_{ik} = (Q_{i1k:2k}, Q_{i1k:3k}, ..., Q_{i,J_i-1,k:J_ik})'$ where $Q_{ijk:j'k} = R_{ijk} - \zeta_{ijk:j'k}$

- $G_{ik}(\boldsymbol{\vartheta}_k, \boldsymbol{\gamma}_k) = \text{diag}(\zeta_{i1k:2k}(1 - \zeta_{i1k:2,k}), \ldots, \zeta_{i,J_i-1,k:J_i,k}(1 - \zeta_{i,J_i-1,k:J_i,k}))$

$$S_{2k}(\boldsymbol{\nu}_k, \gamma_k) \ = \ \sum_{i=1}^{I} \left(\frac{\partial \boldsymbol{\zeta}_{ik}{}'}{\partial \boldsymbol{\gamma}_k}\right) [G_{ik}(\boldsymbol{\vartheta}_k, \boldsymbol{\gamma}_k)]^{-1} \boldsymbol{Q}_{ik}$$

IPW EQUATIONS FOR MEAN RESPONSE

$$
\begin{aligned}
U_1(\boldsymbol{\theta}, \boldsymbol{\alpha}_o) &= \sum_{i=1}^{I} U_{1i}(\boldsymbol{\theta}, \boldsymbol{\alpha}_o) \\
&= \sum_{i=1}^{I} \left(\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \right) [V_i(\boldsymbol{\theta})]^{-1} \Delta_i(\boldsymbol{\alpha}_o)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)
\end{aligned}
$$

where

- $\mathrm{var}(\boldsymbol{Y}_i) = V_i(\boldsymbol{\theta})$
- $\Delta_i(\boldsymbol{\alpha}_o) = \mathrm{diag}(r_{ijk}/P(R_{ijk} = 1 | \boldsymbol{Y}_i^{(o)}, \boldsymbol{x}_i; \boldsymbol{\alpha}_o),\ j = 1, \ldots, k = 1, \ldots)$

ESTIMATING EQUATIONS FOR ASSOCIATION BETWEEN RESPONSES

- $v_{ijk:j'k'} = P(Y_{ijk} = 1, Y_{ij'k'} = 1|\boldsymbol{x}_i)$
- $\epsilon_{ijk:j'k'} = E(Y_{ijk}|Y_{ij'k'} = y_{ij'k'})$, for $(j,k) < (j',k')$
- $E_{ijk:j'k'} = Y_{ijk} - \epsilon_{ijk:j'k'}$
- $\boldsymbol{E}_i = (E_{i11:12}, \ldots, E_{iJ_i,K-1:J_iK})'$

$$U_2(\boldsymbol{\theta}, \boldsymbol{\alpha}_o) = \sum_{i=1}^{I} U_{2i}(\boldsymbol{\theta}, \boldsymbol{\alpha}_o) = \sum_{i=1}^{I} \left(\frac{\partial \boldsymbol{\epsilon}_i'}{\partial \boldsymbol{\phi}}\right) [W_i(\boldsymbol{\theta})]^{-1} \Delta_i^*(\boldsymbol{\alpha}_o) \boldsymbol{E}_i$$

where

$$\begin{aligned} W_i &= \text{diag}(\epsilon_{ijk:j'k'}(1 - \epsilon_{ijk:j'k'}), \ \ (j,k) < (j',k')) \\ \Delta_i^* &= \text{diag}\left(\frac{I(R_{ijk} = 1, R_{ij'k'} = 1)}{P(R_{ijk} = 1, R_{ij'k'} = 1|\boldsymbol{Y}_{ij}^{(o)}, \boldsymbol{Y}_{ij'}^{(o)})}, \ \ (j,k) < (j',k')\right) \end{aligned}$$

OBJECTIVE: To examine the performance of IPWGEE with clustered weights, and the relative efficiency compared to IPWGEE ignoring clustering in drop-out rates.

MODEL FOR MEAN RESPONSE

$$\log(\mu_{ijk}/(1-\mu_{ijk})) = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k}$$

- $x_{1k} = 1$ treated; 0 control $\qquad x_{2k} = I(k=2) \qquad x_{3k} = I(k=3)$

MODEL FOR RESPONSE ASSOCIATIONS

$$\log \psi_{ijk:j'k'} = \delta_0 + \delta_1 I(j=j') + \delta_2 (k=k')$$

$$\psi_{ijk:j'k'} = \begin{cases} \psi_R & j=j',\ k \neq k' \\ \psi_C & j \neq j',\ k = k' \\ \psi_{CR} & j \neq j',\ k \neq k' \end{cases}$$

MODEL FOR MISSING DATA PROCESS (MONOTONE ONLY)

$$\begin{aligned}
\log(\lambda_{ijk}/(1-\lambda_{ijk})) &= \alpha_{0k} + \alpha_{1k} y_{ij,k-1} + \alpha_{2k} x_{ik} + \alpha_{3k} y_{ij,k-1} x_{ik}, \quad k = 2, 3 \\
\log \psi^*_{ijk:j'k} &= \gamma_0 + \gamma_1 I(k=3), \ k = 2, 3 \\
\psi^*_2 &= \exp(\gamma_0), \quad \psi^*_3 = \exp(\gamma_0 + \gamma_1)
\end{aligned}$$

ANALYSES

- Analysis assuming conditionally independent missing data (INDEP)
- Analysis accommodating “missingness” dependence within clusters (CLUST)

- $\text{expit}(\beta_0) = .4$, $\exp(\beta_1) = .50$, $\text{expit}(\beta_0 + \beta_2) = .5$, $\text{expit}(\beta_0 + \beta_3) = .6$
- $\psi_{CR} = 1.2$, $\psi_C = 1.5$, $\psi_R = 2.0$
- $\alpha_{\ell k} = \alpha_\ell$, $\ell = 0, 1, 2, 3$, $k = 2, 3$, $\text{expit}(\alpha_0) = 0.70$, $\alpha_3 = 0$
- $I = 100$, $J_i = 3$, $K = 3$, 500 simulations

PARAMETERS FOR MISSING DATA PROCESS				RELATIVE EFFICIENCIES $\left(\frac{\text{var}_{\text{CLUST}}(\hat{\boldsymbol{\theta}})}{\text{var}_{\text{INDEP}}(\hat{\boldsymbol{\theta}})}\right)$			
$\exp(\alpha_1)$	$\exp(\alpha_2)$	ψ_2^*	ψ_3^*	β_0	β_1	β_2	β_3
0.5	0.5	4	4	100.0	99.8	99.9	100.2
0.5	1	4	4	100.1	100.1	99.8	100.0
1	2	4	4	100.0	100.2	100.2	99.9
1	1	4	4	100.0	100.0	100.3	99.5
0.5	0.5	4	2	100.0	100.2	100.3	100.2
0.5	1	4	2	100.0	100.3	100.0	100.1
1	2	4	2	100.0	100.0	100.2	100.1
1	1	4	2	100.0	99.9	100.0	100.0
0.5	0.5	2	2	100.0	100	100.1	99.8
0.5	1	2	2	100.0	99.9	99.9	99.8
1	2	2	2	100.0	100.0	100.0	100.0
1	1	2	2	100.0	100.0	100.0	100.0
0.5	0.5	1	1	100.0	100.0	100.0	99.9
0.5	1	1	1	100.0	100.0	100.0	99.9
1	2	1	1	100.0	100.0	100.0	99.9
1	1	1	1	100.0	100.0	99.9	100.0

- $\text{expit}(\beta_0) = .4$, $\exp(\beta_1) = .50$, $\text{expit}(\beta_0 + \beta_2) = .5$, $\text{expit}(\beta_0 + \beta_3) = .6$
- $\psi_{CR} = 1.2$, $\psi_C = 1.5$, $\psi_R = 2.0$
- $\alpha_{\ell k} = \alpha_\ell$, $\ell = 0, 1, 2, 3$, $k = 2, 3$, $\text{expit}(\alpha_0) = 0.70$, $\alpha_3 = 0$
- $I = 100$, $J_i = 3$, $K = 3$, 500 simulations

PARAMETERS FOR MISSING DATA PROCESS				RELATIVE EFFICIENCIES		
$\exp(\alpha_1)$	$\exp(\alpha_2)$	ψ_2^*	ψ_3^*	$\log(\psi_{CR})$	$\log(\psi_C)$	$\log(\psi_R)$
0.5	0.5	4	4	89.8	91.0	100.2
0.5	1	4	4	94.4	92.6	100.0
1	2	4	4	97.0	97.2	99.9
1	1	4	4	97.7	97.0	99.8
0.5	0.5	4	2	87.3	91.3	100.2
0.5	1	4	2	93.7	93.5	100.0
1	2	4	2	96.7	97.2	99.9
1	1	4	2	96.2	97.0	100.1
0.5	0.5	2	2	94.9	95.0	100.1
0.5	1	2	2	96.4	96.6	100.0
1	2	2	2	98.5	98.4	99.9
1	1	2	2	98.6	98.9	100.0
0.5	0.5	1	1	101.3	101.1	100.1
0.5	1	1	1	100.7	100.6	100.0
1	2	1	1	99.8	99.9	100.0
1	1	1	1	100.2	100.8	100.0

- $\text{expit}(\beta_0) = .4$, $\exp(\beta_1) = .50$, $\text{expit}(\beta_0 + \beta_2) = .5$, $\text{expit}(\beta_0 + \beta_3) = .6$
- $\psi_{CR} = 1.2$, $\psi_C = 1.5$, $\psi_R = 2.0$
- $\alpha_{\ell k} = \alpha_\ell$, $\ell = 0, 1, 2, 3$, $k = 2, 3$, $\text{expit}(\alpha_0) = 0.70$, $\alpha_3 = 0$
- $I = 100$, $J_i = 3$, $K = 3$, 500 simulations

PARAMETERS FOR MISSING DATA PROCESS				RELATIVE EFFICIENCIES		
$\exp(\alpha_1)$	$\exp(\alpha_2)$	ψ_2^*	ψ_3^*	δ_0	δ_2	δ_1
0.5	0.5	4	4	89.8	89.0	95.2
0.5	1	4	4	94.4	92.8	97.3
1	2	4	4	97.0	96.6	98.8
1	1	4	4	97.7	96.1	98.5
0.5	0.5	4	2	87.3	88.6	93.6
0.5	1	4	2	93.7	92.5	98.0
1	2	4	2	96.7	96.6	98.5
1	1	4	2	96.2	95.3	98.3
0.5	0.5	2	2	94.9	93.7	97.4
0.5	1	2	2	96.4	96.0	98.7
1	2	2	2	98.5	98.3	99.1
1	1	2	2	98.6	98.8	99.0
0.5	0.5	1	1	101.3	101.5	100.4
0.5	1	1	1	100.7	100.5	100.4
1	2	1	1	99.8	99.8	100.0
1	1	1	1	100.2	100.7	100.2

WATERLOO SMOKING PREVENTION PROJECT

OBJECTIVE

To assess effect of smoking prevention programs on smoking behaviour of children in grades 6, 7 and 8 at 45 schools from three school boards.

DESIGN

- 100 schools randomized to an anti-smoking program or control
- each child's smoking status assessed annually
- 12% of children have one or more missing observations

RESPONSE MODEL:

Covariates include treatment, grade and sex.

GEE2 with associations expressed as correlations ρ_{CR}, ρ_C, and ρ_R

MISSING DATA MODELS:

Covariates include past responses and school risk indicators

Weights derived from independence and clustered models

Table 1: Estimates and Standard Errors from the Analysis of the Waterloo Smoking Prevention Project Data

	UNWEIGHTED			INDEPENDENCE WEIGHTS			CLUSTERED WEIGHTS		
	EST.	S.E.	95% CI	EST.	S.E.	95% CI	EST.	S.E.	95% CI
RESPONSE MODELS:									
Intercept (β_0)	-2.795	0.198	(-3.182, -2.408)	-2.589	0.136	(-2.856, -2.321)	-2.579	-2.579	(-2.950,-2.207)
Treatment (β_1)	-0.133	0.167	(-0.460, 0.193)	-0.325	0.158	(-0.635, -0.016)	-0.308	0.186	(-0.672, 0.056)
Sex (β_2)	0.002	0.086	(-0.166, 0.169)	0.067	0.116	(-0.160, 0.294)	0.051	0.124	(-0.192, 0.294)
Grade 7 (β_3)	0.770	0.091	(0.592, 0.948)	0.683	0.207	(0.278, 1.088)	0.664	0.232	(0.209, 1.119)
Grade 8 (β_4)	1.567	0.092	(1.387, 1.748)	1.743	0.238	(1.277, 2.209)	1.723	0.269	(1.195, 2.251)
ϕ_0	0.135	0.052	(0.034, 0.236)	0.184	0.080	(0.027, 0.342)	0.186	0.094	(0.001, 0.370)
ϕ_1	2.404	0.114	(2.181, 2.627)	2.265	0.144	(1.982, 2.548)	2.267	0.133	(2.007, 2.527)
ϕ_2	0.067	0.028	(0.012, 0.122)	0.091	0.067	(-0.040, 0.222)	0.100	0.073	(-0.043, 0.242)
$\psi_{i;jk;j'k'}$	1.144		(1.034, 1.266)	1.203		(1.027, 1.408)	1.204		(1.001, 1.448)
$\psi_{i;jk;jk'}$	12.660		(10.072, 15.913)	11.579		(8.669, 15.466)	11.617		(9.016,14.968)
$\psi_{i;jk;j'k}$	1.224		(1.118, 1.339)	1.317		(1.066, 1.627)	1.330		(1.059, 1.671)
MISSING DATA MODELS:									
GRADE 7 :									
Intercept (α_{20})				2.731	0.131	(2.474, 2.988)	2.741	0.125	(2.496, 2.985)
Prev. Res. (α_{21})				-1.250	0.175	(-1.593, -0.908)	-1.188	0.202	(-1.585, -0.792)
Treatment (α_{22})				-0.163	0.144	(-0.444, 0.118)	-0.103	0.158	(-0.413, 0.207)
ϕ_2^*							0.211	0.089	(0.037, 0.385)
$\psi_{i;j2;j'2}$							1.235		(1.038, 1.470)
GRADE 8 :									
Intercept (α_{30})				-0.840	0.157	(-1.147, -0.532)	-0.726	0.188	(-1.094, -0.358)
Prev. Res. (α_{31})				-0.714	0.167	(-1.041, -0.387)	-0.712	0.139	(-0.984, -0.439)
Prev. Mis. (α_{32})				3.424	0.134	(3.161, 3.688)	3.374	0.152	(3.077, 3.671)
Treatment (α_{33})				0.197	0.131	(-0.059, 0.453)	0.195	0.184	(-0.164, 0.555)
ϕ_3^*							0.119	0.050	(0.021, 0.218)
$\psi_{i;j3;j'3}$							1.127		(1.021, 1.243)

SUMMARY

- Whenever interest lies primarily in association parameters (i.e. genetic studies) efficiency gains could be important when
  - modest degree of missing data
  - modest within cluster association for missing indicators
- Little loss in efficiency if missing mechanism is not clustered

FUTURE WORK

- Adaptations to deal with incomplete multivariate longitudinal data are under study