Subsample ignorable likelihood methods for regression with missing values of covariates (throwing data away can actually pay!)

Roderick J. Little

University of Michigan
Outline

• Likelihood methods for repeated-measures with missing data
  – Missing data for outcomes and predictors

• Tools: Ignorable likelihood methods, selective discarding of incomplete cases
  – Positive feature: no model required for missing-data mechanism, even though some models are not MAR

• Also apply ideas to missing covariates in survival analysis

• Little & Zhang (in press)
Key Idea

• Rubin’s (1976) MAR theory does not distinguish between missing outcomes and predictors
  – Here adopt a “divide and conquer” strategy

• An alternative to MNAR modeling the missing data mechanism is to drop cases with missing values of predictors from the analysis
  – Valid when missingness does not depend on outcomes
Missing Data: General Strategies

Complete-Case Analysis

Imputation

Analyze Incomplete

Weights

Imputations
e.g. maximum likelihood

UNC 2011 SSIL
Likelihood approaches

- Maximum Likelihood (ML, REML) for large samples
- Bayes for small samples
- Multiple imputation (MI) of missing values based on predictive distribution for a Bayesian model, with Bayesian MI combining rules (SAS PROC MI, IVEWARE, MICE, etc.)
- “Ignorable likelihood” – no model for missing-data mechanism
- Assumes data are missing at random (MAR): “missingness does not depend on missing data, given observed data”
- When not MAR, ML generally requires a model for the mechanism, which is often weakly identified and vulnerable to misspecification
- Here, discuss methods that avoid modeling the mechanism
Unweighted CC analysis

• Drops incomplete cases
• Hence inefficient if there is substantial information in these cases
• Loss of information depends on pattern and estimand
• E.g. Figure 1: for mean of $Y$ the incomplete cases have substantial information, when $X$’s are predictive
• For regression of $Y$ on $X$, incomplete cases have no information, under MAR
• But there is info under NMAR

Figure 1
Missing data in $X$

Target: regression of $Y$ on $X$, $Z$; missing data on $X$

IL methods include information for the regression in the incomplete cases (particularly intercept and coefficients of $Z$) and are valid assuming MAR:

$$\text{Pr}(X \text{ missing}) = g(Z, Y)$$

**BUT**: if $\text{Pr}(X \text{ missing}) = g(Z, X)$

CC analysis is consistent, but IL methods (or weighted CC) are inconsistent since mechanism is not MAR

Simulations favoring IL often generate data under MAR, hence are biased against CC
(A) Missing data in $X$

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Observation, $i$</th>
<th>$z_i$</th>
<th>$x_i$</th>
<th>$y_i$</th>
<th>$R_{x_i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>$i = 1, \ldots, m$</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>$u_x = (1, \ldots, 1)$</td>
</tr>
<tr>
<td>P2</td>
<td>$i = m + 1, \ldots, n$</td>
<td>√</td>
<td>?</td>
<td>√</td>
<td>$\bar{u}_x$</td>
</tr>
</tbody>
</table>

Key: √ denotes observed, ? denotes observed or missing

Could be vector

P1: complete cases

P2: incomplete $X$
XMAR ⇒ Ignorable Likelihood

Target: Parameters $\phi$ of regression of $Y$ on $X, Z$

Full Model: $p(x_i, y_i | z_i, \theta) \times p(R_{x_i} | z_i, x_i, y_i, \psi); \phi = \phi(\theta)$

If we assume XMAR:

$p(R_{x_i} | z_i, x_i, y_i, \psi) = p(R_{x_i} | z_i, x_{\text{obs},i}, y_i, \psi)$ for all $x_{\text{mis},i}$

Then $L_{\text{full}}(\theta, \psi) = L_{\text{ign}}(\theta) \times L_{\text{md}}(\psi)$, can base inference on

$L_{\text{ign}}(\theta) = \text{const.} \times \prod_{i=1}^{n} p(x_{\text{obs},i}, y_i | z_i, \theta)$
XMAR ⇒ Ignorable Likelihood

Target: \( \phi = \phi(\theta) = \) parameters of regression of \( Y \) on \( X, Z \)

ML: \( \hat{\phi} = \phi(\hat{\theta}) \)

Bayes: draw \( \phi^{(d)} = \phi(\theta^{(d)}) \)

Multiple imputation: draw \( X_{\text{mis}}^{(d)} \sim P(X_{\text{mis}} \mid \text{data}) \), apply MI combining rules to estimates of \( \phi \)
XCOV $\Rightarrow$ Complete-Case Analysis

Assume XCOV: completeness of $X$ depends on covariates, not outcomes:

$$p\left( R_{x_i} = u_x \mid z_i, x_i, y_i, \psi \right) = p\left( R_{x_i} = u_x \mid z_i, x_i, \psi \right)$$
for all $y_i$ (MNAR)

$$L_{\text{full}}(\theta, \psi) = \prod_{i=1}^{m} p(R_{x_i} = u_x, x_i, y_i \mid z_i, \theta, \psi) \prod_{i=m+1}^{n} p(R_{x_i}, x_{\text{obs},i}, y_i \mid z_i, \theta, \psi)$$

$$= \prod_{i=1}^{m} p(y_i \mid x_i, R_{x_i} = u_x, z_i, \theta, \psi) p(R_{x_i} = u_x, x_i \mid z_i, \theta, \psi)$$

By XCOV

$$\times \prod_{i=m+1}^{n} p(R_{x_i}, x_{\text{obs},i}, y_i \mid z_i, \theta, \psi)$$

$$= \prod_{i=1}^{m} p(y_i \mid x_i, z_i, \phi) \times p(R_{x_i} = u_x, x_i \mid z_i, \theta, \psi) \prod_{i=m+1}^{n} p(R_{x_i}, x_{\text{obs},i}, y_i \mid z_i, \theta, \psi)$$

$$= L_{\text{cc}}(\phi) \times L_{\text{rest}}(\theta, \psi)$$

Maximizing $L_{\text{cc}}(\phi)$ is valid, but info in $L_{\text{rest}}(\theta, \psi)$ except in special cases
(B) Missing data on $X$ and $Y$

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Observation, $i$</th>
<th>$z_i$</th>
<th>$x_i$</th>
<th>$y_i$</th>
<th>$R_{x_i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>$i = 1,\ldots,m$</td>
<td>$\sqrt{;}$</td>
<td>$\sqrt{;}$</td>
<td>?</td>
<td>$u_x = (1,\ldots,1)$</td>
</tr>
<tr>
<td>P2</td>
<td>$i = m + 1,\ldots,n$</td>
<td>$\sqrt{;}$</td>
<td>?</td>
<td>?</td>
<td>$\bar{u}_x$</td>
</tr>
</tbody>
</table>

Key: $\sqrt{\;}$ denotes observed, ? denotes observed or missing

Could be vector

P1: covariates complete

P2: $x$ incomplete
XYMAR ⇒ Ignorable Likelihood

Target: Parameters $\phi$ of regression of $Y$ on $X, Z$

Model: $p(x_i, y_i \mid z_i, \theta)$

Assume XYMAR:

$$p(R_{x_i}, R_{y_i} \mid z_i, x_i, y_i, \psi) = p(R_{x_i}, R_{y_i} \mid z_i, x_{obs,i}, y_{obs,i}, \psi)$$

for all $x_{mis,i}, y_{mis,i}$

Then $L_{full}(\theta, \psi) = L_{ign}(\theta) \times L_{md}(\psi)$, can base inference on

$$L_{ign}(\theta) = \text{const.} \times \prod_{i=1}^{n} p(x_{obs,i}, y_{obs,i} \mid z_i, \theta)$$

IL Inference about $\phi(\theta)$, as before
XCOV, YSMAR $\Rightarrow$ IL on cases with $X$ observed

Target: Parameters $\phi$ of regression of $Y$ on $X, Z$

Assume:

XCOV: completeness of $X$ depends on covariates, not outcomes:

$$p\left( R_{x_i} = u_x \mid z, x, y_i, \psi \right)$$

$$= p\left( R_{x_i} = u_x \mid z, x, \psi \right) \quad \text{for all } y_i$$

YSMAR: $Y$ is MAR in subsample with $X$ observed:

$$p\left( R_{y_i} \mid R_{x_i} = u_x, z, x, y_i, \psi \right)$$

$$= p\left( R_{y_i} \mid R_{x_i} = u_x, z, x, y_{\text{obs}, i}, \psi \right) \quad \text{for all } y_i$$

SSIL: Apply IL to subsample with $X$ fully observed (P1)
(B) Missing data on $X$ and $Y$

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Observation, $i$</th>
<th>$z_i$</th>
<th>$x_i$</th>
<th>$y_i$</th>
<th>$R_{x_i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>$i = 1, \ldots, m$</td>
<td>$\sqrt{}$</td>
<td>$\sqrt{}$</td>
<td>?</td>
<td>$u_x = (1, \ldots, 1)$</td>
</tr>
<tr>
<td>P2</td>
<td>$i = m + 1, \ldots, n$</td>
<td>$\sqrt{}$</td>
<td>?</td>
<td>?</td>
<td>$\bar{u}_x$</td>
</tr>
</tbody>
</table>

**Key:** $\sqrt{}$ denotes observed, ? denotes observed or missing
SSIL likelihood under XCOV, YSMAR

\[ L_{\text{full}}(\theta, \psi) = \prod_{i=1}^{n} p(R_{x_i}, x_{\text{obs},i}, R_{y_i}, y_{\text{obs},i} \mid z_i, \theta, \psi) \]

\[ = \prod_{i=1}^{m} p(R_{x_i} = u_x, x_i, R_{y_i}, y_{\text{obs},i} \mid z_i, \theta, \psi) \times L_{\text{rest}}(\theta, \psi) \]

\[ = L_{\text{rest}}(\theta, \psi) \times \prod_{i=1}^{m} \left( p(R_{x_i} = u_x, x_i \mid z_i, \theta, \psi) \right) \]

\[ \times \prod_{i=1}^{m} \left( \int p(y_i \mid x_i, R_{x_i} = u_x, z_i, \theta, \psi) p(R_{y_i} \mid y_i, x_i, R_{x_i} = u_x, z_i, \theta, \psi) dy_{\text{mis}} \right) \]

\[ = L_{\text{rest}}^{*}(\theta, \psi) \times \prod_{i=1}^{m} \int p(y_i \mid x_i, z_i, \phi) p(R_{y_i} \mid y_{\text{obs},i}, x_i, R_{x_i} = u_x, z_i, \psi) dy_{\text{mis}} \]

\[ = L_{\text{rest}}^{*}(\theta, \psi) \times \prod_{i=1}^{m} p(y_{\text{obs},i} \mid x_i, z_i, \phi) \prod_{i=1}^{m} p(R_{y_i} \mid y_{\text{obs},i}, x_i, R_{x_i} = u_x, z_i, \psi) \]

SSIL maximizes this
Two covariates $X, W$ with different mechanisms

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Observation, $i$</th>
<th>$z_i$</th>
<th>$x_i$</th>
<th>$w_i$</th>
<th>$y_i$</th>
<th>$R_{x_i}$</th>
<th>$R_{w_i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>$i = 1, \ldots, m$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>?</td>
<td>$u_x$</td>
<td>$u_w$</td>
</tr>
<tr>
<td>P2</td>
<td>$i = m+1, \ldots, m+r$</td>
<td>✓</td>
<td>✓</td>
<td>?</td>
<td>?</td>
<td>$u_x$</td>
<td>$\bar{u}_w$</td>
</tr>
<tr>
<td>P3</td>
<td>$i = m+r+1, \ldots, n$</td>
<td>✓</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>$\bar{u}_x$</td>
<td>$\bar{u}_w$</td>
</tr>
</tbody>
</table>

P1: covariates complete
P2: $x$ obs, $w$, $y$ may be mis
P3: $x$ mis, $w$, $y$ may be mis

 UNC 2011 SSIL
XCOV, WYSMAR $\Rightarrow$ IL on cases with $X$ observed

- Target: regression of $Y$ on $Z$, $X$, and $W$
- Assume:
  (XCOV) Completeness of $X$ can depend on covariates but not $Y$:
  \[
p\left( R_{x_i} = u_x \mid z_i, x_i, w_i, y_i, \psi_x \right) = p\left( R_{x_i} = u_x \mid z_i, x_i, w_i, \psi_x \right) \text{ for all } y_i
  \]
  (WYMAR) Missingness of $(W, Y)$ is MAR within subsample of cases with $X$ observed:
  \[
p(R_{(w_i, y_i)} \mid z_i, x_i, w_i, y_i, R_{x_i} = u_x; \psi_{wy.x}) = \]
  \[
p(R_{(w_i, y_i)} \mid z_i, x_i, w_{obs,i}, y_{obs,i}, R_{x_i} = u_x; \psi_{wy.x}) \text{ for all } w_{mis,i}, y_{mis,i}
  \]
- SSIL: apply IL method (e.g. ML) to the subsample of cases for which $X$ is observed
- Proof of consistency: similar to previous case, treating $W$ and $Y$ as block
Two covariates $X, W$ with different mechanisms

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Observation, $i$</th>
<th>$z_i$</th>
<th>$x_i$</th>
<th>$w_i$</th>
<th>$y_i$</th>
<th>$R_{x_i}$</th>
<th>$R_{w_i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>$i = 1, \ldots, m$</td>
<td>√</td>
<td>√</td>
<td>?</td>
<td>?</td>
<td>$u_x$</td>
<td>$u_w$</td>
</tr>
<tr>
<td>P2</td>
<td>$i = m+1, \ldots, m+r$</td>
<td>√</td>
<td>√</td>
<td>?</td>
<td>?</td>
<td>$u_x$</td>
<td>$\bar{u}_w$</td>
</tr>
<tr>
<td>P3</td>
<td>$i = m+r+1, \ldots, n$</td>
<td>√</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>$\bar{u}_x$</td>
<td>$\bar{u}_w$</td>
</tr>
</tbody>
</table>

SSIL: analyze cases in patterns 1 and 2

P1: covariates complete
P2: $x$ obs, $w$, $y$ may be mis
P3: $x$ mis, $w$, $y$ may be mis

UNC 2011 SSIL
Simulation Study

• For each of 1000 replications, 5000 observations $Z$, $W$, $X$ and $Y$ generated as:

$$(y_i | z_i, w_i, x_i) \sim_{\text{ind}} N(1 + z_i + w_i + x_i, 1)$$

$$(z_i, w_i, x_i) \sim_{\text{ind}} N(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho & \rho \\
\rho & 1 & \rho \\
\rho & \rho & 1 \end{pmatrix}$$

• 20-35% of missing values of $W$ and $X$ generated by four mechanisms
Simulation: missing data mechanisms

<table>
<thead>
<tr>
<th>Mechanisms</th>
<th>$\alpha_0^{(w)}$</th>
<th>$\alpha_z^{(w)}$</th>
<th>$\alpha_w^{(w)}$</th>
<th>$\alpha_x^{(w)}$</th>
<th>$\alpha_y^{(w)}$</th>
<th>$\alpha_0^{(x)}$</th>
<th>$\alpha_z^{(x)}$</th>
<th>$\alpha_w^{(x)}$</th>
<th>$\alpha_x^{(x)}$</th>
<th>$\alpha_y^{(x)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: All valid</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>II: CC valid</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>III: IML valid</td>
<td>-2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>IV: SSIML valid</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$\logit \left( P(R_{w_i} = 0 \mid z_i, w_i, x_i, y_i) \right) = \alpha_0^{(w)} + \alpha_z^{(w)} z_i + \alpha_w^{(w)} w_i + \alpha_x^{(w)} x_i + \alpha_y^{(w)} y_i$

$\logit \left( P(R_{x_i} = 0 \mid R_{w_i} = 1, z_i, w_i, x_i, y_i) \right) = \alpha_0^{(x)} + \alpha_z^{(x)} z_i + \alpha_w^{(x)} w_i + \alpha_x^{(x)} x_i + \alpha_y^{(x)} y_i$
RMSEs*1000 of Estimated Regression Coefficients for Before Deletion (BD), Complete Cases (CC), Ignorable Maximum Likelihood (IML) and Subsample Ignorable Maximum Likelihood (SSIML), under Four Missing Data Mechanisms.

<table>
<thead>
<tr>
<th></th>
<th>ρ = 0</th>
<th>ρ = 0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I* II III IV</td>
<td>I  II III IV</td>
</tr>
<tr>
<td>BD</td>
<td>27 28 28 27</td>
<td>50 46 50 46</td>
</tr>
<tr>
<td>CC</td>
<td>45 44 553 322</td>
<td>86 71 426 246</td>
</tr>
<tr>
<td>IML</td>
<td>37 231 36 116</td>
<td>58 96 53 90</td>
</tr>
<tr>
<td>SSIML</td>
<td>42 133 360 49</td>
<td>70 80 319 69</td>
</tr>
</tbody>
</table>

Valid: ALL CC IML SSIML ALL CC IML SSIML

UNC 2011 SSIL
Missing Covariates in Survival Analysis

\{t_1, \ldots, t_k \} \text{ distinct survival times, } j = \text{ unit that fails at time } t_j \text{ (no ties)};

\( R_j = \text{ risk set at time } t_j, z_j, x_j, w_j = \text{ covariates, as before.} \)

Complete data: contribution of data at time \( t_j \) to partial likelihood is

\[
L_j = \frac{\lambda(y = t_j | z_j, x_j, w_j, \beta)}{\sum_{k \in R_j} \lambda(y = t_k | z_k, x_k, w_k, \beta)}, \quad \lambda(y = t_j | z_j, x_j, w_j, \beta) = \text{ hazard}
\]

With \( z_j, w_j \) fully observed, \( x_j \) covariate-dependent complete, i.e.:

\[
\Pr(R_{x_j} = u_x | y_j, z_j, x_j, w_j) = \Pr(R_{x_j} = u_x | z_j, x_j, w_j)
\]

then \( \lambda(y = t_j | R_{x_j} = u_x, z_j, x_j, w_j, \beta) = \lambda(y = t_j | z_j, x_j, w_j, \beta) \)

That is, conditioning on \( R_{x_j} = u_x \) for each risk set

gives a valid partial likelihood

also OK for time-varying \( x_j \)
SSIL for Survival Analysis

SSIL for partial likelihood: Assume

XCOV: $x_j$ is covariate-dependent missing:

$$\Pr(R_{x_j} = u_x \mid y_j, z_j, x_j, w_j) = \Pr(R_{x_j} = u_x \mid z_j, x_j, w_j)$$

WSMAR: missing values of $w_j$ are MAR

in subsample with $x_j$ observed:

$$\Pr(R_{w_j} \mid R_{x_j} = u_x, y_j, z_j, x_j, w_j) = \Pr(R_{w_j} \mid R_{x_j} = u_x, y_{obs,j}, z_j, x_j, w_{obs,j})$$

Then can apply SSIL methods to partial likelihood in subsample with $w_j$ observed.
How to choose $X$, $W$

- Choice requires understanding of the mechanism:
- Variables that are missing based on their underlying values belong in $X$
- Variables that are SMAR belong in $W$
- Collecting data about why variables are missing is obviously useful to get the model right
- But this applies to all missing data adjustments…
Other questions and points

• How much is lost from SSIL relative to full likelihood model of data and missing data mechanism?
  – In some special cases, SSIL is efficient for a pattern-mixture model
  – In other cases, trade-off between additional specification of mechanism and loss of efficiency from conditional likelihood

• MAR analysis applied to the subset does not have to be likelihood-based
  – E.g. weighted GEE, AIPWEE

• Assuming XCOV, MNAR analysis for $Y$ can also be applied to subset with $X$ observed
A longitudinal application of SMAR

• Zhou, Little and Kalbfleisch (2011) consider block-sequential missing-data models factored as

\[ f (V_i, R_i \mid \theta, \psi) \]

\[ = f (V_{i(1)}, R_{i(1)} \mid \theta^{(1)}, \psi^{(1)}) \]

\[ \times f (V_{i(2)}, R_{i(2)} \mid H_{i(1)}, \theta^{(2)}, \psi^{(2)}) \]

\[ \times f (V_{i(3)}, R_{i(3)} \mid H_{i(2)}, \theta^{(3)}, \psi^{(3)}) \]

\[ \times \cdots \times f (V_{i(B)}, R_{i(B)} \mid H_{i(B-1)}, \theta^{(B)}, \psi^{(B)}) \]

\[ H_{i(j)} = \text{history up to } j, \text{ including missing-data indicators} \]

\[ f (V_{i(j)}, R_{i(j)} \mid H_{i(j-1)}, \theta^{(j)}, \psi^{(j)}) \]

could have selection or pattern-mixture factorization
Special case: Block-conditional MAR models

\[
f(V_{i(j)}, R_{i(j)} \mid H_{i(j-1)}, \theta^{(j)}, \psi^{(j)})
\]

\[
= f(V_{i(j)} \mid H_{i(j-1)}, \theta^{(j)}) \times f(R_{i(j)} \mid H_{i(j-1)}, V_{i(j)}, \psi^{(j)})
\]

\[
= f(V_{i(j)} \mid V_{i(1)}, \ldots, V_{i(j-1)}, \theta^{(j)}) \times f(R_{i(j)} \mid H_{i(j-1)}, V_{\text{obs},i(j)}, \psi^{(j)})
\]

\[
L_{\text{full}}(\theta, \psi \mid Y_{\text{obs}}, M) = L_{\text{bm}}(\theta) \times L_{\text{rest}}(\theta, \psi)
\]

\[
L_{\text{bm}}(\theta) = \prod_{j=1}^{B} \prod_{i \in Q_j} f(V_{\text{obs},i(j)} \mid V_{i(1)}, \ldots, V_{i(j-1)}, \theta^{(j)})
\]

\[
Q_j = \{\text{Set of cases with } V_{i(1)}, \ldots, V_{i(j-1)} \text{ observed}\}
\]

\[
L_{\text{bm}}(\theta) = \text{Block-monotone reduced likelihood}
\]

Inference based on \(L_{\text{bm}}(\theta)\) is simpler, since does not involve \(\psi\)

Extends SMAR

UNC 2011 SSIL 28
Conclusions

• Sometimes discarding data is useful!
• SSIL: selectively discards data based on assumed missing-data mechanism
• More efficient than CC
• Valid for mechanisms where IL, CC are inconsistent
• rlittle@umich.edu for copy of papers
and thanks to my recent students…
Hyonggin An, Qi Long, Ying Yuan, Guangyu Zhang, Xiaoxi Zhang, Di An, Yan Zhou, Rebecca Andridge, Qixuan Chen, Ying Guo, Chia-Ning Wang, Nanhua Zhang