An Approximated Expectation-Maximization Algorithm for Analysis of Data with Missing Values

Gong Tang

Department of Biostatistics, GSPH
University of Pittsburgh

NISS Workshop on Nonignorable Nonresponse
November 12-13, 2015
1. Introduction
   - Background & Current Approaches
   - Expectation-Maximization (EM) Algorithm for Regression Analysis of Data with Nonresponse

2. An approximated EM algorithm
   - The algorithm
   - Application in contingency table analysis
   - Application in regression analyses with a continuous outcome

3. Simulation studies

4. Discussion
Consider bivariate data \( \{x_i, y_i, \; i = 1, 2, \ldots, n\} \) where
- \( x_i \)'s are fully observed,
- \( y_i \)'s are only observed for \( i = 1, \ldots, m \).

Denote \( R_i \) to be the missing-data indicator:
\[ R_i = 1 \text{ if } y_i \text{ is observed and } R_i = 0 \text{ otherwise}. \]

Assume that
\[
[y_i \mid x_i] \sim g(y_i \mid x_i; \theta) \propto \exp\{\theta S(x_i, y_i) + a(\theta)\}
\]
\[
pr[R_i = 1 \mid x_i, y_i] = w(x_i, y_i; \psi)
\]

and the parameter of interest is \( \theta \).

Goal: to avoid modeling \( w(x_i, y_i; \psi) \).
Likelihood-based Inference (I)

The observed data are $\mathcal{D}_{obs} = \{x_i, R_i, R_i y_i; i = 1, \ldots, n\}$.

When $w(x_i, y_i; \psi)$ is parametrically modeled, the likelihood function is

$$L(\theta, \psi; \mathcal{D}_{obs}) = \prod_{i=1}^{n} p(R_i, R_i y_i | x_i; \theta, \psi)$$

$$= \prod_{i=1}^{m} g(y_i | x_i; \theta) w(x_i, y_i; \psi) \prod_{i=m+1}^{n} \int g(y_i | x_i; \theta) \{1 - w(x_i, y_i; \psi)\} \, dy_i$$

When $w(x_i, y_i; \psi) = w(x_i; \psi)$, data are called missing at random (MAR) and

$$L(\theta, \psi; \mathcal{D}_{obs}) \propto L(\theta; \mathcal{D}_{obs}) L(\psi; \mathcal{D}_{obs})$$  (Rubin, 1976).
When data are MAR plus $\theta$ and $\psi$ are distinct, the modeling of $w(x, y; \psi)$ is not necessary under the likelihood-based inference.

When data are not MAR, the missingness has to be modeled and the inference on $\theta$ and $\psi$ are made together.

Misspecification of the missing-data model $w(x, y; \psi)$ often leads to biased estimate of $\theta$. 
A conditional likelihood

Assume that

\[ [y_i \mid x_i] \sim g(y_i \mid x_i; \theta) \quad (A \text{ parametric regression}) \]

\[ pr[R_i = 1 \mid x_i, y_i] = w(x_i, y_i; \psi) = w(y_i; \psi). \]

Then \([X \mid Y, R = 1] = [X \mid Y, R = 0] = [X \mid Y]\).

Consider the following conditional likelihood:

\[
CL(\theta) = \prod_{R_i = 1} p(x_i \mid y_i; \theta, F_X) \quad (F_X \text{ is the CDF of } X)
\]

\[
= \prod_{R_i = 1} \frac{g(y_i \mid x_i; \theta)p(x_i)}{p(y_i; \theta, F_X)} \quad \text{(Bayes formula)}
\]

\[
\propto \prod_{R_i = 1} \frac{g(y_i \mid x_i; \theta)}{\int g(y_i \mid x; \theta) dF_X(x)}
\]

Requires knowing \( F_X(\cdot) \)!
A pseudolikelihood method
(Tang, Little & Raghunathan, 2003)

In alternative, we may either

- model $X \sim f(x; \alpha)$, obtain $\hat{\alpha} = \arg \max_\alpha \prod_{i=1}^n f(x_i; \alpha)$, then consider a pseudolikelihood function

$$PL_1(\theta) = \prod_{R_i=1} \frac{g(y_i|x_i; \theta)}{\int g(y_i|x; \theta) \, dF_X(x; \hat{\alpha})}$$

- or substitute $F_X(\cdot)$ by the empirical distribution $F_n(\cdot)$:

$$PL_2(\theta) = \prod_{R_i=1} \frac{p(y_i|x_i; \theta)}{\int p(y_i|x; \theta) \, dF_n(x)}$$

$$= \prod_{R_i=1} \frac{p(y_i|x_i; \theta)}{\frac{1}{n} \sum_{j=1}^n p(y_i|x_j; \theta)}$$
Exponential tilting (Kim & Yu, 2011)

Consider the following semiparametric logistic regression model:

\[
pr[R_i = 1 \mid x_i, y_i] = \logit^{-1}\{h(x_i) - \psi y_i\},
\]

where \(h(\cdot)\) is unspecified and \(\psi\) is either known or estimated from an external dataset with the missing values recovered. Subsequently we would have:

\[
p(y\mid x, r = 0) = p(y\mid x, r = 1) \frac{\exp(\psi y)}{E[\exp(\psi y)\mid x, r = 1]}.
\]

With both \(p(y\mid x, r = 1)\) and \(E[\exp(\psi y)\mid x, r = 1]\) empirically estimated, one will obtain a density estimator for \(p(y\mid x, r = 0)\) and estimate \(\theta = E[H(X, Y)]\) via:

\[
\hat{\theta} = \frac{1}{n} \left\{ \sum_{r_i=1} H(x_i, y_i) + \sum_{r_i=0} \hat{E}[H(x_i, y_i)\mid x_i, r_i = 0] \right\}
\]
An instrumental variable approach
(Wang, Shao & Kim, 2014)

Assume that \( X = (Z, U) \) and

\[
E(Y|X) = \beta_0 + \beta_1 u + \beta_2 z
\]

\[
pr(R = 1|Z, U, Y) = pr(R = 1|U, Y) = \pi(\psi; U, Y)
\]

one may use the following estimating equations to estimate \( \psi \):

\[
\sum_{i=1}^{n} \left\{ \frac{r_i}{\pi(\psi; u_i, y_i)} - 1 \right\}(1, u_i, z_i) = 0.
\]

Then estimate \( \beta \)'s via

\[
\sum_{i=1}^{n} \frac{r_i}{\pi(\hat{\psi}; u_i, y_i)}(y_i - \beta_0 - \beta_1 u_i + \beta_2 z_i)(1, u_i, z_i) = 0.
\]
The likelihood function

From the following representation:

\[
L(\theta, \psi; D_{\text{obs}}) = \prod_{i=1}^{n} p(R_i, R_i y_i \mid x_i; \theta, \psi)
\]

\[
= \prod_{i=1}^{n} \frac{p(R_i, R_i y_i, (1 - R_i) y_i \mid x_i; \theta, \psi)}{p((1 - R_i) y_i \mid R_i, R_i y_i, x_i; \theta, \psi)}
\]

\[
= \prod_{i=1}^{n} \frac{p(R_i, y_i \mid x_i; \theta, \psi)}{p((1 - R_i) y_i \mid R_i, R_i y_i, x_i; \theta, \psi)}
\]

\[
= \prod_{i=1}^{n} \frac{p(y_i \mid x_i; \theta)p(R_i \mid x_i, y_i; \psi)}{p((1 - R_i) y_i \mid R_i, R_i y_i, x_i; \theta, \psi)}.
\]
The log-likelihood function

We have

\[ l(\theta, \psi; y_{obs}) = \log L(\theta, \psi; y_{obs}) = \sum_{i=1}^{n} \{ \log p(y_i \mid x_i; \theta) + \log p(R_i \mid x_i, y_i; \psi) - \log p((1 - R_i)y_i \mid R_i, R_iy_i, x_i; \theta, \psi) \}. \]

Let

\[ Q(\theta, \psi; \tilde{\theta}, \tilde{\psi}) = \sum_{i=1}^{n} E[\log p(y_i \mid x_i; \theta) + \log p(R_i \mid x_i, y_i; \psi) \mid R_i, R_iy_i, x_i; \tilde{\theta}, \tilde{\psi}], \]

\[ H(\theta, \psi; \tilde{\theta}, \tilde{\psi}) = \sum_{i=1}^{n} E[\log p((1 - R_i)y_i \mid R_i, R_iy_i, x_i; \theta, \psi) \mid R_i, R_iy_i, x_i; \tilde{\theta}, \tilde{\psi}]. \]

Then we have \( l(\theta, \psi; y_{obs}) = Q(\theta, \psi; \tilde{\theta}, \tilde{\psi}) - H(\theta, \psi; \tilde{\theta}, \tilde{\psi}), \) for any \((\tilde{\theta}, \tilde{\psi})\).
The Expectation-Maximization (EM) Algorithm

(Dempster, Laird and Rubin, 1977)

From the Jansen’s inequality, we have

\[ H(\theta, \psi; \tilde{\theta}, \tilde{\psi}) \leq H(\tilde{\theta}, \tilde{\psi}; \tilde{\theta}, \tilde{\psi}). \]

The EM algorithm is an iterative algorithm to find a sequence

\[ \{(\theta(t), \psi(t)), \ t = 0, 1, 2, \ldots\} \]

such that

\[ (\theta(t+1), \psi(t+1)) = \arg \max_{(\theta, \psi)} Q(\theta, \psi; \theta(t), \psi(t)). \]

This assures that

\[ l(\theta(t), \psi(t); y_{obs}) \leq l(\theta(t+1), \psi(t+1); y_{obs}). \]

Typically logistic or probit regression models are used for

\[ w(x, y; \psi). \]

Numerical integrations or the Monte Carlo method are often necessary in the implementation.
When $\psi = \psi_0$ is known

Now consider when the true value of $\psi$, $\psi_0$, is known:

$$l(\theta, \psi_0; y_{obs}) = \log L(\theta, \psi_0; y_{obs})$$

$$= \sum_{i=1}^{n} \{ \log p(y_i \mid x_i; \theta) + \log p(R_i \mid x_i, y_i; \psi_0)$$

$$- \log p((1 - R_i)y_i \mid D_{i,obs}; \theta, \psi_0) \}.$$ 

With current estimate $\theta^{(t)}$, the $Q$-function becomes

$$Q(\theta; \theta^{(t)}) = \sum_{i=1}^{n} E[\log p(y_i \mid x_i; \theta) + \log p(R_i \mid x_i, y_i; \psi_0) \mid R_i, R_i y_i, x_i; \theta^{(t)}, \psi_0]$$

$$\propto \sum_{i=1}^{n} E[\log p(y_i \mid x_i; \theta) \mid R_i, R_i y_i, x_i; \theta^{(t)}, \psi_0],$$

because the second term does not involve $\theta$. 
Another view of the likelihood

Without loss of generality, assume that the regression model follows a canonical exponential family. Then

\[
\Omega(\theta; \theta^{(t)}) = \sum_{i=1}^{m} \log g(y_i | x_i; \theta) + \sum_{i=m+1}^{n} E[\log g(y_i | x_i; \theta) | x_i, R_i = 0; \theta^{(t)}, \psi_0]
\]

\[
\propto \sum_{i=1}^{m} \log g(y_i | x_i; \theta) + \sum_{i=m+1}^{n} \{\theta E[S(x_i, y_i) | x_i, R_i = 0; \theta^{(t)}, \psi_0] + a(\theta)\},
\]

we need to obtain \(E[S(x_i, y_i) | x_i, R_i = 0; \theta^{(t)}, \psi_0]\) in order to carry out the EM algorithm. With \(w(x_i, y_i; \psi_0)\) known,

\[
E[S(x_i, y_i) | x_i, R_i = 0; \theta^{(t)}, \psi_0] = \frac{\int s(x_i, y_i)g(y_i | x_i; \theta^{(t)})\{1 - w(x_i, y_i; \psi_0)\} \, dy_i}{\int g(y_i | x_i; \theta^{(t)})\{1 - w(x_i, y_i; \psi_0)\} \, dy_i}
\]
An alternative look of the E-step

On the other hand,

\[
E[S(x_i, y_i) \mid x_i, \theta^{(t)}] = E[E[S(x_i, y_i) \mid R_i, x_i] \mid x_i]
\]

\[
= E[S(x_i, y_i) \mid x_i, R_i = 1; \theta^{(t)}] pr[R_i = 1 \mid x_i; \theta^{(t)}, \psi_0]
\]

\[
+ E[S(x_i, y_i) \mid x_i, R_i = 0; \theta^{(t)}, \psi_0] pr[R_i = 0 \mid x_i; \theta^{(t)}, \psi_0],
\]

We would have

\[
E[S(x_i, y_i) \mid x_i, R_i = 0; \theta^{(t)}, \psi_0] = E[S(x_i, y_i) \mid x_i, \theta^{(t)}] - \frac{E[S(x_i, y_i) \mid x_i, R_i = 1; \theta^{(t)}, \psi_0] pr[R_i = 1 \mid x_i; \theta^{(t)}, \psi_0]}{pr[R_i = 0 \mid x_i; \theta^{(t)}, \psi_0]}
\]
Empirical replacements

It is noted that if we use some empirical estimates of $E[S(x_i, y_i) \mid x_i, R_i = 1; \theta^{(t)}, \psi_0]$ and $pr[R_i = 0 \mid x_i; \theta^{(t)}, \psi_0]$ to replace them, we would be able to carry out an iterative algorithm as the following:

At the E-step

$$\hat{E}[S(x_i, y_i) \mid x_i, R_i = 0; \theta^{(t)}, \psi_0] = \frac{E[S(x_i, y_i) \mid x_i, \theta^{(t)}] - \hat{E}[S(x_i, y_i) \mid x_i, R_i = 1] \hat{pr}[R_i = 1 \mid x_i]}{1 - \hat{pr}[R_i = 1 \mid x_i]}$$

At the M-step, find $\theta = \theta^{(t+1)}$ that solves:

$$\theta^{(t+1)} = \arg \max_\theta Q^*(\theta; \theta^{(t)})$$

$$:= \arg \max_\theta \left\{ \sum_{i=1}^{m} S(x_i, y_i) + \sum_{i=m+1}^{n} \hat{E}[S(x_i, Y) \mid x_i, R_i = 0; \theta^{(t)}] \right\} + na(\theta)$$
An important observation

- In usual an EM algorithm, $\theta^{(t)}$ may be far away from the truth $\theta_0$.

- However, in order for the empirical estimates $\hat{pr}[R_i = 1 \mid x_i]$ and $\hat{E}[S(x_i, y_i) \mid x_i, R_i = 1]$ to be self-consistent with the corresponding terms under $(\theta^{(t)}, \psi_0)$, it is required that $\theta^{(t)}$ is a consistent estimate of $\theta$.

- Therefore we should start with a consistent initial estimate of $\theta$ and carry out this iterative algorithm to obtain another consistent estimate of $\theta$ at convergence.
A short summary of the modified EM

1. Choose initial value of $\theta^{(0)}$ from
   - either a complete dataset from subjects with similar characteristics
   - or a completed subset recovered from recalls.

2. Compute the Nadaraya-Watson estimates or local polynomial estimates for $pr[R = 1 | x_i]$ and $E[S(x_i, Y) | x_i, R = 1]$, for each $i = m + 1, m + 2, \ldots, n$.

3. At the E-step of the $t$th iteration, calculate
   $\hat{E}[S(x_i, y_i) | x_i, R_i = 0; \theta^{(t)}, \psi_0]$.

4. At the M-step of the $t$th iteration, find $\theta = \theta^{(t+1)}$ that solves:

$$\sum_{i=1}^{n} E[S(x_i, y_i) | x_i; \theta] = \sum_{i=1}^{m} S(x_i, y_i) + \sum_{i=m+1}^{n} \hat{E}[S(x_i, Y) | x_i, R_i = 0; \theta^{(t)}]$$

With a complete external dataset $\{x_i, y_i; i = n + 1, \ldots, n + n_E\}$, we would solve $\theta = \theta^{(t+1)}$ from:

$$\sum_{i=1}^{n} E[S(x_i, y_i) | x_i; \theta] = \sum_{i=n+1}^{n} E[S(x_i, y_i) | x_i; \theta] + \sum_{i=m+1}^{n+n_E} \hat{E}[S(x_i, Y) | x_i, R_i = 0; \theta^{(t)}] + \sum_{i=m+1}^{n+n_E} S(x_i, y_i)$$
Here we design an iterative algorithm with a sequence
\[ \{\theta^{(t)}, \ t = 0, 1, 2, \ldots\} \] such that
\[
\theta^{(t+1)} = \arg \max_\theta Q^*(\theta; \theta^{(t)}) \\
= \arg \max_\theta \{Q(\theta; \theta^{(t)}) + o(1)\}
\]

This algorithm will yield
\[
l(\theta^{(t+1)}, \psi_0; y_{obs}) \geq l(\theta^{(t)}, \psi_0; y_{obs}) + o(1).
\]
Consider discrete data \( \{x_i, y_i, z_i, i = 1, \ldots, n\} \):

- \( x_i \)'s and \( y_i \)'s are fully observed.
- \( z_i \) is observed for \( i = 1, \ldots, c \); missing for \( i = c + 1, \ldots, n \).
  Let \( m = n - c \). Essentially the observed include the fully classified table \( \{c_{jkl}\} \) and the partially classified table \( \{m_{jk+}\} \).

- Complete external data
  \( D_E = \{x_i, y_i, z_i, i = n + 1, \ldots, n + n^E\} \) are fully observed.
- A log-linear model is assumed with parameter \( \theta \), which leads to \( \pi_{jkl} = pr[X = i, Y = j, Z = l] \),
  \( j = 1, \ldots, J; \, k = 1, \ldots, K; \, l = 1, \ldots, L \).
- Initial estimates \( \{\pi_{jkl}^{(0)}\} \) are derived from the external complete data \( D_E \).
Implementation under the discrete setting

- Initial estimations:

\[
\hat{p}r[z = l|x = j, y = k, R = 1] = \frac{\sum_{i=1}^{c} I\{x_i = j, y_i = k, z_i = l\}}{\sum_{i=1}^{c} I\{x_i = j, y_i = k\}} = \frac{c_{jkl}}{c_{jk+}}
\]

\[
\hat{p}r[R = 1|x = j, y = k] = \frac{\sum_{i=1}^{c} I\{x_i = j, y_i = k\}}{\sum_{i=1}^{n} I\{x_i = j, y_i = k\}} = \frac{c_{jk+}}{n_{jk+}}.
\]

- At the \(t\)th step, for \((j, k, l)\), we update

\[
S_{jkl} = c_{jkl} + \sum_{i=c+1}^{n} I\{x_i = j, y_i = k\}\hat{p}r[z = l|x = j, y = k, R = 0]
\]

\[
= c_{jkl} + m_{jk+} \frac{pr[z = l|x = j, y = k; \theta(t)] - \frac{c_{jkl}}{c_{jk+}} \frac{c_{jk+}}{n_{jk+}}}{1 - \frac{c_{jk+}}{n_{jk+}}}
\]

\[
= c_{jkl} + m_{jk+} \frac{\frac{\pi_{jkl}}{\pi_{jk+}} - \frac{c_{jkl}}{n_{jk+}}}{m_{jk+}} = n_{jk+} \frac{\frac{\pi_{jkl}}{\pi_{jk+}}}{n_{jk+}}
\]
Impression on the discrete setting

- In the E-step, we ended up as if imputing all $Z_i$s based on $(x_i, y_i)$, including those observed ones.
- If we include the external data $D_E$ in the algorithm with updating the sufficient statistics through

$$S_{jkl}^* = n_{jkl}^E + S_{jkl} = n_{jkl}^E + n_{jk} + \frac{\pi_j(t)}{\pi_{jk+}},$$

the modified EM algorithm is equivalent to running a regular EM algorithm on the fully classified table $\{n_{jkl}^E\}$ and the partial classified table $\{n_{jk+}\}$, or, removing the observed $z_i$s from the complete cases (not part of the external complete data).
Implementation under the continuous setting

Consider the regression analysis of \( [Y|X] \) where \( Y \) is continuous and subject to nonresponse

- If \( X \) is discrete, the empirical approximations are:

  \[
  \hat{E}[S(x_i, Y)|x_i = k, r_i = 1] = \frac{\sum_{r_j=1, x_j=x_i=k} S(x_i, y_j)}{\#\{j : r_j = 1, x_j = x_i = k\}}
  \]

  \[
  \hat{pr}[R_i = 1|x_i = k] = \frac{\#\{j : r_j = 1, x_j = x_i = k\}}{\#\{j : x_j = x_i = k\}}
  \]

- If \( X \) is continuous, we use either the Nadaraya-Watson estimate or local polynomial estimate as

  \[
  \hat{E}[S(x_i, Y)|x_i = k, r_i = 1]; \text{ a kernel estimator for}
  \]

  \[
  \hat{pr}[R_i = 1|x_i = k].
  \]
Simulation settings when data are NMAR

- We simulated bivariate data \( \{x_i, y_i, i = 1, 2, \ldots, N \} \) following:
  (i) \( x_i \sim N(0, 1) \) or \( x_i \sim Bin(5, 0.3) \).
  (ii) \( [y_i | x_i] \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \), where \( \theta = (\beta_0, \beta_1, \sigma^2) = (1, 1, 1) \).
- Keep \( n_E = 200 \) of the above observations as the external datasets for obtaining initial values for \( \theta \).
- Simulate missing \( y_i \)'s from the rest \( n = 1000 \) subjects with:
  \( pr[R_i = 1 | x_i, y_i] = \Phi(\psi_0 + \psi_1 x_i + \psi_2 y_i) \), where \( \Phi() \) is the CDF of the Gaussian distribution.
- Compare the complete-case estimate (\( \hat{\theta}_{cc} \)) based on the 1000 observations with missing data, MLE from the external data (\( \hat{\theta}_E \)) and the proposed approximated EM estimates (\( \hat{\theta}_{AEM} \)).
Simulation results with $X \sim N(0, 1)$

<table>
<thead>
<tr>
<th>Methods</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete-case analysis</td>
<td>2009</td>
<td>-1388</td>
<td>-486</td>
</tr>
<tr>
<td>Empirical Bias*</td>
<td>380</td>
<td>422</td>
<td>468</td>
</tr>
<tr>
<td>Empirical SD*</td>
<td>0%</td>
<td>8.8%</td>
<td>79.5%</td>
</tr>
<tr>
<td>Coverage of 95% CI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLEs from external subset</td>
<td>-26</td>
<td>9</td>
<td>-103</td>
</tr>
<tr>
<td>Empirical Bias*</td>
<td>705</td>
<td>707</td>
<td>970</td>
</tr>
<tr>
<td>Empirical SD*</td>
<td>96.1%</td>
<td>93.8%</td>
<td>92.6%</td>
</tr>
<tr>
<td>Coverage of 95% CI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximated EM**</td>
<td>-97</td>
<td>-11</td>
<td>-161</td>
</tr>
<tr>
<td>Empirical Bias*</td>
<td>551</td>
<td>472</td>
<td>624</td>
</tr>
<tr>
<td>Empirical SD*</td>
<td>95.0%</td>
<td>95.2%</td>
<td>93.4%</td>
</tr>
<tr>
<td>Coverage of 95% CI</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The proportion of nonresponse was about 40%.

* Empirical biases and SDs were the actual numbers times 1000.

** The Epanechnikov kernel was used in the approximated EM. Bootstrap was used to derive the standard errors of the $\widehat{\theta}_{AEM}$.
Simulation results with $X \sim Bin(5, 0.3)$

<table>
<thead>
<tr>
<th>Methods</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete-case analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical Bias*</td>
<td>156</td>
<td>-1412</td>
<td>-354</td>
</tr>
<tr>
<td>Empirical SD*</td>
<td>574</td>
<td>446</td>
<td>477</td>
</tr>
<tr>
<td>Coverage of 95% CI</td>
<td>94.6%</td>
<td>11.6%</td>
<td>86.5%</td>
</tr>
<tr>
<td>MLEs from external subset</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical Bias*</td>
<td>-12</td>
<td>25</td>
<td>-111</td>
</tr>
<tr>
<td>Empirical SD*</td>
<td>1245</td>
<td>688</td>
<td>976</td>
</tr>
<tr>
<td>Coverage of 95% CI</td>
<td>94.7%</td>
<td>94.2%</td>
<td>92.5%</td>
</tr>
<tr>
<td>Approximated EM**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical Bias*</td>
<td>-16</td>
<td>27</td>
<td>-50</td>
</tr>
<tr>
<td>Empirical SD*</td>
<td>662</td>
<td>492</td>
<td>731</td>
</tr>
<tr>
<td>Coverage of 95% CI</td>
<td>94.7%</td>
<td>93.9%</td>
<td>93.6%</td>
</tr>
</tbody>
</table>

The proportion of nonresponse was about 40%

* Empirical biases and SDs were the actual numbers times 1000.
** The empirical averages were used in the approximated EM. Bootstrap was used to derive the standard errors of the $\hat{\theta}_{AEM}$. 
Simulation settings when data MAR

- We simulated bivariate data \( \{x_i, y_i, i = 1, 2, \ldots, N\} \) following:
  
  (i) \( x_i \sim N(0, 1) \).
  
  (ii) \( [y_i | x_i] \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \), where \( \theta = (\beta_0, \beta_1, \sigma^2) = (1, 1, 1) \).

- Keep \( n_E = 200 \) of the above observations as the external datasets for obtaining initial values for \( \theta \).

- Simulate missing \( y_i \)'s from the rest \( n = 1000 \) subjects with:
  
  \( pr[R_i = 1 | x_i, y_i] = \Phi(\psi_0 + \psi_1 x_i) \), where \( \Phi() \) is the CDF of the Gaussian distribution.

- Compare the complete-case estimate \( (\hat{\theta}_{cc}) \) based on the 1000 observations with missing data, MLE from the external data \( (\hat{\theta}_E) \) and the proposed approximated EM estimates \( (\hat{\theta}_{AEM}) \).
Simulation results with $X \sim N(0, 1)$

<table>
<thead>
<tr>
<th>Methods</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete-case analysis</td>
<td>Empirical Bias*</td>
<td>-12</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Empirical SD*</td>
<td>386</td>
<td>411</td>
</tr>
<tr>
<td></td>
<td>Coverage of 95% CI</td>
<td>94%</td>
<td>94.5%</td>
</tr>
<tr>
<td>MLEs from external subset</td>
<td>Empirical Bias*</td>
<td>-26</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Empirical SD*</td>
<td>705</td>
<td>707</td>
</tr>
<tr>
<td></td>
<td>Coverage of 95% CI</td>
<td>96.1%</td>
<td>93.8%</td>
</tr>
<tr>
<td>Approximated EM**</td>
<td>Empirical Bias*</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Empirical SD*</td>
<td>556</td>
<td>472</td>
</tr>
<tr>
<td></td>
<td>Coverage of 95% CI</td>
<td>93.9%</td>
<td>94.5%</td>
</tr>
</tbody>
</table>

The proportion of nonresponse was about 40%.

* Empirical biases and SDs were the actual numbers times 1000.

** The Epanechnikov kernel was used in the modified EM. Bootstrap was used to derive the standard errors of the $\hat{\theta}_{AEM}$.
Without an external complete dataset ...

If one could obtain a *reliable* initial estimate $\theta^{(0)}$ and the associated variance-covariance matrix estimate $\hat{V}$, then we can use it as the initial and perform the M-steps via

$$\hat{\theta}^{(t+1)} = \arg \max_\theta \left\{ (\theta - \theta^{(0)})^T \hat{V}^{-1} (\theta - \theta^{(0)}) + \frac{1}{n} Q^*(\theta; \theta^{(t)}) \right\}$$
Without external complete dataset ...

Consider missing-data mechanisms such as

\[ \text{pr}[R_i = 1 \mid x_i, y_i] = w(x_i, y_i; \psi) = w(y_i; \psi), \]

There are two approaches for implementing the approximated EM algorithm for data with outcome-dependent nonresponses:

1. Use the pseudolikelihood estimate as the initial estimate and run the approximated EM.

2. Use the pseudolikelihood estimate as the initial estimate and incorporate the pseudolikelihood function as a component in each M-step:

\[ \hat{\theta}^{(t+1)} = \arg \max_{\theta} \{ l_{pl}(\theta) + Q^*(\theta; \theta^{(t)}) \} \]

The second approach is more computationally intensive.
Future works

- Need to monitor
  \[ \{ l(\theta(t); \psi_0), Q^*(\theta(t+1); \theta(t)) - Q(\theta(t+1); \theta(t)); t = 0, 1, 2 \ldots \} . \]

- Search for a target function \( l(\theta; P_n^{(X,Y,R)}, \theta(0)) \) so that \( \hat{\theta}_{AEM} \) is a stationary point of \( l(\theta; P_n^{(X,Y,R)}, \theta(0)) \).

- Variance estimation.

- Link to integrative data analysis.
Acknowledgment

Megan Hunt Olson, University of Wisconsin, Green Bay. Yang Zhang, Amgen Inc.

