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

Regression Analysis of Data with Nonresponse

Consider bivariate data $\{x_i, y_i,\ i = 1, 2, \ldots, n\}$ where

- x_is are fully observed,
- y_is are only observed for $i = 1, \ldots, m$.

Denote R_i to be the missing-data indicator:

$R_i = 1$ if y_i is observed and $R_i = 0$ otherwise.

Assume that

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

and the parameter of interest is θ.

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

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

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

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

$$L(\theta, \psi; \mathcal{D}_{obs}) \propto L(\theta; \mathcal{D}_{obs}) L(\psi; \mathcal{D}_{obs}) \quad \text{(Rubin, 1976)}.$$

Likelihood-based Inference (II)

When data are MAR plus θ and ψ 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 θ and ψ are made together.

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

A conditional likelihood

Assume that

$$[y_i \mid x_i] \sim g(y_i|x_i;\theta) \quad \text{(A parametric regression)}$$
$$pr[R_i = 1 \mid x_i, y_i] = w(x_i, y_i;\psi) = w(y_i;\psi).$$

Then $[X|Y, R = 1] = [X|Y, R = 0] = [X|Y]$.
Consider the following conditional likelihood:

$$\begin{aligned} CL(\theta) &= \prod_{R_i=1} p(x_i|y_i;\theta, F_X) \qquad (F_X \text{ is the CDF of X}) \\ &= \prod_{R_i=1} \frac{g(y_i|x_i;\theta)p(x_i)}{p(y_i;\theta, F_X)} \qquad \text{(Bayes formula)} \\ &\propto \prod_{R_i=1} \frac{g(y_i|x_i;\theta)}{\int g(y_i|x;\theta)\, dF_X(x)} \end{aligned}$$

Requires knowing $F_X(\cdot)$!

A pseudolikelihood method

(Tang, Little & Raghunathan, 2003)

In alternative, we may either

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

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

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

$$\begin{aligned} PL_2(\theta) &= \prod_{R_i=1} \frac{p(y_i|x_i;\theta)}{\int p(y_i|x;\theta)\, dF_n(x)} \\ &= \prod_{R_i=1} \frac{p(y_i|x_i;\theta)}{\frac{1}{n}\sum_{j=1}^n p(y_i|x_j;\theta)} \end{aligned}$$

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 ψ is either known or estimated from an external dataset with the missing values recovered.
Subsequently we would have:

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

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

$$\widehat{\theta} = \frac{1}{n}\{\sum_{r_i=1} H(x_i, y_i) + \sum_{r_i=0}^{n} \widehat{E}[H(x_i, y_i)|x_i, r_i = 0]\}$$

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 ψ:

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

Then estimate β's via

$$\sum_{i=1}^{n} \frac{r_i}{\pi(\widehat{\psi}; u_i, y_i)}(y_i - \beta_0 - \beta_1 u_i + \beta_2 z_i)(1, u_i, z_i) = 0.$$

The likelihood function

From the following representation:

$$
\begin{aligned}
L(\theta, \psi; \mathcal{D}_{obs}) &= \prod_{i=1}^{n} p(R_i, R_i y_i \mid x_i; \theta, \psi) \\
&= \prod_{i=1}^{n} \frac{p(R_i, R_i y_i, (1-R_i) y_i \mid x_i; \theta, \psi)}{p((1-R_i) y_i \mid R_i, R_i y_i, x_i; \theta, \psi)} \\
&= \prod_{i=1}^{n} \frac{p(R_i, y_i \mid x_i; \theta, \psi)}{p((1-R_i) y_i \mid R_i, R_i y_i, x_i; \theta, \psi)} \\
&= \prod_{i=1}^{n} \frac{p(y_i \mid x_i; \theta) p(R_i \mid x_i, y_i; \psi)}{p((1-R_i) y_i \mid R_i, R_i y_i, x_i; \theta, \psi)},
\end{aligned}
$$

The log-likelihood function

We have

$$\begin{aligned} l(\theta, \psi; y_{obs}) &= \log L(\theta, \psi; y_{obs}) \\ &= \sum_{i=1}^{n} \{\log p(y_i \mid x_i; \theta) + \log p(R_i \mid x_i, y_i; \psi) \\ &\quad - \log p((1 - R_i)y_i \mid R_i, R_i y_i, x_i; \theta, \psi)\}. \end{aligned}$$

Let

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

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

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 ψ, ψ_0, is known:

$$\begin{aligned} l(\theta, \psi_0; y_{obs}) &= \log L(\theta, \psi_0; y_{obs}) \\ &= \sum_{i=1}^{n} \{\log p(y_i \mid x_i; \theta) + \log p(R_i \mid x_i, y_i; \psi_0) \\ &\quad - \log p((1 - R_i) y_i \mid \mathcal{D}_{i,obs}; \theta, \psi_0)\}. \end{aligned}$$

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

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

because the second term does not involve θ.

Another view of the likelihood

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

$$\begin{aligned}\Omega(\theta;\theta^{(t)}) &= \sum_{i=1}^{m} logg(y_i \mid x_i;\theta) + \sum_{i=m+1}^{n} E[\log g(y_i \mid x_i;\theta) \mid x_i, R_i = 0;\theta^{(t)},\psi_0] \\ &\propto \sum_{i=1}^{m} logg(y_i \mid x_i;\theta) + \sum_{i=m+1}^{n} \{\theta E[S(x_i,y_i) \mid x_i, R_i = 0;\theta^{(t)},\psi_0] + a(\theta)\},\end{aligned}$$

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

$$E[S(x_i,y_i) \mid x_i, R_i = 0;\theta^{(t)},\psi_0] = \frac{\int s(x_i,y_i)g(y_i \mid x_i;\theta^{(t)})\{1 - w(x_i,y_i;\psi_0)\}\,dy_i}{\int g(y_i \mid x_i;\theta^{(t)})\{1 - w(x_i,y_i;\psi_0)\}\,dy_i}$$

An alternative look of the E-step

On the other hand,

$$\begin{aligned} E[S(x_i, y_i) \mid x_i, \theta^{(t)}] &= E[E[S(x_i, y_i)|R_i, x_i]|x_i] \\ &= E[S(x_i, y_i) \mid x_i, R_i = 1; \theta^{(t)}]pr[R_i = 1 \mid x_i; \theta^{(t)}, \psi_0] \\ &\quad + E[S(x_i, y_i) \mid x_i, R_i = 0; \theta^{(t)}, \psi_0]pr[R_i = 0 \mid x_i; \theta^{(t)}, \psi_0], \end{aligned}$$

We would have

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

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

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

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

$$\begin{aligned}\theta^{(t+1)} &= \arg\max_\theta Q^*(\theta; \theta^{(t)}) \\ &:= \arg\max_\theta [\theta\{\sum_{i=1}^{m} S(x_i, y_i) + \sum_{i=m+1}^{n} \widehat{E}[S(x_i, Y) \mid x_i, R_i = 0; \theta^{(t)}]\} + na(\theta)]\end{aligned}$$

An important observation

- In usual an EM algorithm, $\theta^{(t)}$ may be far away from the truth θ_0.

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

- Therefore we should start with a consistent initial estimate of θ and carry out this iterative algorithm to obtain another consistent estimate of θ 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 $pr[R = 1 \mid x_i]$ and $E[S(x_i, Y) \mid x_i, R = 1]$, for each $i = m + 1, m + 2, \ldots, n$.
3. At the E-step of the tth iteration, calculate $\widehat{E}[S(x_i, y_i) \mid x_i, R_i = 0; \theta^{(t)}, \psi_0]$.
4. At the M-step of the tth iteration, find $\theta = \theta^{(t+1)}$ that solves:

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

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

$$\sum_{i=1}^{n} E[S(x_i, y_i) \mid x_i; \theta] + \sum_{i=n+1}^{n+n_E} E[S(x_i, y_i) \mid x_i; \theta]$$

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

Here we design an iterative algorithm with a sequence $\{\theta^{(t)},\ t = 0, 1, 2, \ldots\}$ such that

$$\begin{aligned} \theta^{(t+1)} &= \arg\max_\theta Q^*(\theta; \theta^{(t)}) \\ &= \arg\max_\theta \{Q(\theta; \theta^{(t)}) + o(1)\} \end{aligned}$$

This algorithm will yield
$l(\theta^{(t+1)}, \psi_0; y_{obs}) \geq l(\theta^{(t)}, \psi_0; y_{obs}) + o(1)$.

An example in contingency table analysis

Consider discrete data $\{x_i, y_i, z_i, i = 1, \ldots, n\}$:

- x_is and y_is 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 $\mathcal{D}_E = \{x_i, y_i, z_i, i = n+1, \ldots, n+n^E\}$ are fully observed.
- A log-linear model is assumed with parameter θ, 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 $\{\pi^{(0)}_{jkl}\}$ are derived from the external complete data $\mathcal{D}_E$.

Implementation under the discrete setting

- Initial estimations:

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

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

- At the tth step, for (j, k, l), we update

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

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

$$= c_{jkl} + m_{jk+}\frac{\frac{\pi^{(t)}_{jkl}}{\pi^{(t)}_{jk+}} - \frac{c_{jkl}}{n_{jk+}}}{\frac{m_{jk+}}{n_{jk+}}} = n_{jk+}\frac{\pi^{(t)}_{jkl}}{\pi^{(t)}_{jk+}}$$

Impression on the discrete setting

- In the E-step, we ended up as if imputing all Z_is based on (x_i, y_i), including those observed ones.
- If we include the external data $\mathcal{D}_E$ in the algorithm with updating the sufficient statistics through

$$S^*_{jkl} = n^E_{jkl} + S_{jkl} = n^E_{jkl} + n_{jk+} \frac{\pi^{(t)}_{jkl}}{\pi^{(t)}_{jk+}},$$

the modified EM algorithm is equivalent to running a regular EM algorithm on the fully classified table $\{n^E_{jkl}\}$ and the partial classified table $\{n_{jk+}\}$, or, removing the observed z_is 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:

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

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

- If X is continuous, we use either the Nadaraya-Watson estimate or local polynomial estimate as $\widehat{E}[S(x_i, Y)|x_i = k, r_i = 1]$; a kernel estimator for $\widehat{pr}[R_i = 1|x_i = k]$.

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 θ.
- Simulate missing y_is from the rest $n = 1000$ subjects with: $pr[R_i = 1 \mid x_i, y_i] = \Phi(\psi_0 + \psi_1 x_i + \psi_2 y_i)$, where $\Phi()$ is the CDF of the Gaussian distribution.
- Compare the complete-case estimate ($\widehat{\theta}_{cc}$) based on the 1000 observations with missing data, MLE from the external data ($\widehat{\theta}_E$) and the proposed approximated EM estimates ($\widehat{\theta}_{AEM}$).

Simulation results with $X \sim N(0, 1)$

Methods		β_0	β_1	σ^2
Complete-case analysis	Empirical Bias*	2009	-1388	-486
	Empirical SD*	380	422	468
	Coverage of 95% CI	0%	8.8%	79.5%
MLEs from external subset	Empirical Bias*	-26	9	-103
	Empirical SD*	705	707	970
	Coverage of 95% CI	96.1%	93.8%	92.6%
Approximated EM**	Empirical Bias*	-97	-11	-161
	Empirical SD*	551	472	624
	Coverage of 95% CI	95.0%	95.2%	93.4%

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 $X \sim Bin(5, 0.3)$

Methods		β_0	β_1	σ^2
Complete-case analysis	Empirical Bias*	156	-1412	-354
	Empirical SD*	574	446	477
	Coverage of 95% CI	94.6%	11.6%	86.5%
MLEs from external subset	Empirical Bias*	-12	25	-111
	Empirical SD*	1245	688	976
	Coverage of 95% CI	94.7%	94.2%	92.5%
Approximated EM**	Empirical Bias*	-16	27	-50
	Empirical SD*	662	492	731
	Coverage of 95% CI	94.7%	93.9%	93.6%

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 $\widehat{\theta}_{AEM}$.

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 θ.
- Simulate missing y_is from the rest $n = 1000$ subjects with: $pr[R_i = 1 \mid x_i, y_i] = \Phi(\psi_0 + \psi_1 x_i)$, where $\Phi()$ is the CDF of the Gaussian distribution.
- Compare the complete-case estimate ($\widehat{\theta}_{cc}$) based on the 1000 observations with missing data, MLE from the external data ($\widehat{\theta}_E$) and the proposed approximated EM estimates ($\widehat{\theta}_{AEM}$).

Simulation results with $X \sim N(0, 1)$

Methods		β_0	β_1	σ^2
Complete-case analysis	Empirical Bias*	-12	1	-24
	Empirical SD*	386	411	488
	Coverage of 95% CI	94%	94.5%	93.4%
MLEs from external subset	Empirical Bias*	-26	9	-103
	Empirical SD*	705	707	970
	Coverage of 95% CI	96.1%	93.8%	92.6%
Approximated EM**	Empirical Bias*	11	11	-73
	Empirical SD*	556	472	641
	Coverage of 95% CI	93.9%	94.5%	93.7%

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 $\widehat{\theta}_{AEM}$

Without an external complete dataset ...

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

$$\widehat{\theta}^{(t+1)} = arg\ max_{\theta}\ \{(\theta - \theta^{(0)})^T \widehat{V}^{-1}(\theta - \theta^{(0)}) + \frac{1}{n} Q^*(\theta; \theta^{(t)})\}$$

Without external complete dataset ...

Consider missing-data mechanisms such as

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

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 pseudolikehood function as a component in each M-step:

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

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 $\widehat{\theta}_{AEM}$ 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.

Reference

1. Dempster, A. P., N. M. Laird and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B* 39, 1-37.
2. Tang, G., Little, R.J.A. and Raghunathan, T. (2003). Analysis of Multivariate Missing Data with Nonignorable Nonresponse. *Biometrika*, 90, 747-764.
3. Kim, J. K. and C. L. Yu (2011). A semi-parametric estimation of mean functionals with non-ignorable missing data, *Journal of the American Statistical Association* 106, 157-165.
4. Wang, S., J. Shao and J. K. Kim (2014). Identifiability and estimation in problems with nonignorable nonresponse, *Statistica Sinica* 24, 1097-1116.