Strength of Parameter Identifiability in NMAR Missing Data

Hua Yun Chen

Division of Epidemiology and Biostatistics
School of Public Health
University of Illinois at Chicago

November, 2015

Parametric Models for data subject to missing

- Let $\mathbf{Y}_i = (Y_{i1}, \cdots, Y_{ip})$ be the full data from subject i, where $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$ are independent identically distributed.
- Assume a parametric model for $\mathbf{Y}_i$, i.e.,

$$(Y_{i1}, \cdots, Y_{ip}) \sim f(y_1, \cdots, y_p \mid \theta),$$

  where $\theta \in \Theta$.
- Assume that $Y_{i1}, \cdots, Y_{ip}$ are subject to missing values. Let $\mathbf{R}_i = (R_{i1}, \cdots, R_{ip})$ be the respective missing data indicators. The missing data mechanism is modeled as

$$P(R_{i1} = r_1, \cdots, R_{ip} = r_p \mid Y_1, \cdots, Y_p) = \pi_{\mathbf{r}}(Y_1, \cdots, Y_p \mid \gamma),$$

  where $\mathbf{r} = (r_1, \cdots, r_p)$ and $\gamma \in \Gamma$.

The observed data likelihood

- Assume that θ is identifiable from the model $f(y_1, \cdots, y_p \mid \theta), \theta \in \Theta$.
- Assume that γ is identifiable from the model $\pi_{\mathbf{r}}(y_1, \cdots, y_p \mid \gamma), \gamma \in \Gamma$.
- Denote the observed data by $\mathbf{R}_i(\mathbf{Y}_i) = (r_{i1}(Y_{i1}), \cdots, r_{ip}(Y_{ip}))$ where $r_{ij}(Y_{ij}) = Y_{ij}$ if $r_{ij} = 1$, and all possible values if $r_{ij} = 0$.
- The observed data likelihood is

$$\prod_{i=1}^{n} \int \pi_{\mathbf{R}_i}(Y_{i1}, \cdots, Y_{ip} \mid \gamma) f(Y_{i1}, \cdots, Y_{ip} \mid \theta) d\bar{\mathbf{R}}_i(\mathbf{Y}_i),$$

where $\bar{\mathbf{R}}_i = \mathbf{1} - \mathbf{R}_i$.

The maximum likelihood approach

- Maximize the likelihood to obtain the parameter estimator for θ and γ. Denote by $\hat{\theta}$ and $\hat{\gamma}$.
- Use the inverse of the observed-data information matrix to approximate the variance-covariate of the maximum likelihood estimator.
- Make inference on the model parameters using a normal approximation to the distribution of the maximum likelihood estimator, or approximating the distribution of the likelihood ratio statistic by a χ^2.

Parametric likelihood approach for MAR and NMAR data

- For MAR data, the parametric likelihood approach is advocated (Little and Rubin, 2002).
- For NMAR data, the parametric likelihood approach is also proposed, e.g., Ibrahim et al (1999).
- In practice, the parametric likelihood approach can be more difficult to apply to NMAR data. It is not uncommon the computational algorithms behave erratically.
- Sensitivity analysis is often recommended to deal with the problem.

Questions with the NMAR data

- Why does the maximum likelihood approach mostly behave well with MAR data, while it often does not with NMAR data?
- When the maximum likelihood approach can still be trusted under NMAR?
- Does a sensitivity analysis tell the story as we have hoped?
- What is a reasonable (right?) way to use the likelihood approach and the sensitivity analysis with NMAR data?

Partial answers to the questions

- The cause of the problems with NMAR is the varying strength of parameter identifiability.
- MAR missing data mostly preserve parameter identifiability of the full data model.
- NMAR missing data often change parameter identifiability of the full data model. Fully identifiable parameters may be changed into unidentifiable, or more frequently weakly identifiable parameters in parametric models.
- A weakly identifiable parameter (Chen, 2011) is defined as identifiable for some values of the other parameters in the model and unidentifiable at other values of the other parameters.

MAR preserves parameter identifiability

- MAR defined as

$$P(\mathbf{R} = \mathbf{r} \mid \mathbf{Y}) = P\{\mathbf{R} = \mathbf{r} \mid \mathbf{r}(\mathbf{Y})\},$$

for all $\mathbf{r}$ and $\mathbf{Y}$, is equivalent to

$$p\{\bar{\mathbf{r}}(y) \mid \mathbf{r}(y), \mathbf{R} = \mathbf{r}\} = p\{\bar{\mathbf{r}}(y) \mid \mathbf{r}(y)\},$$

for all $\mathbf{r}$ and y, where $\bar{\mathbf{r}} = \mathbf{1} - \mathbf{r}$.

- Result: For MAR missing data, if $p\{\bar{\mathbf{r}}(y) \mid \mathbf{r}(y)\}$ for all missing data patterns are identifiable from the incompletely observed data, the parameter identifiability under incompletely observed data is unchanged from the fully observed data.

When MAR is "trouble-free"

- Under MAR, if $P(\mathbf{R} = \mathbf{1} \mid \mathbf{Y}) > 0$, parameter identifiability in the full data model $f(y_1, \cdots, y_p \mid \theta)$ is preserved by MAR.
- When $P(\mathbf{R} = \mathbf{1} \mid \mathbf{Y}) > 0$ does not hold, the parameter identifiability may still be preserved as long as any parameter in $p\{\bar{\mathbf{r}}(y) \mid \mathbf{r}(y)\}$ can be identified from $p\{\mathbf{r}(y), \mathbf{R} = \mathbf{r}\}$
- If there is no unidentifiable or weakly identifiable parameter in the full data model, there is no unidentifiable or weakly identifiable parameters in the missing data model when either of the foregoing conditions is satisfied.

Mapping a NMAR model to a MAR model

- Given a missing data model $\pi_{\mathbf{r}}(y \mid \gamma)$ and a full data model $p(y \mid \theta)$. Define a mapping

$$Aq(y) = \sum_{\mathbf{r}} \int \pi_{\mathbf{r}}(y \mid \gamma) p(y \mid \theta) d\bar{\mathbf{r}}(y) q\{\bar{\mathbf{r}}(y) \mid \mathbf{r}(y)\},$$

  where q is a density function on y.

- If there exists a density $q(y)$ such that $Aq(y \mid \theta, \gamma) = q(y \mid \theta, \gamma)$, then a full data model $q(y \mid \theta, \gamma)$ and a MAR missing data probability model

$$P_q(\mathbf{R} = \mathbf{r} \mid \mathbf{Y}) = \int \pi_{\mathbf{r}}(y) p(y) d\bar{\mathbf{r}}(y) / q\{\mathbf{r}(y)\},$$

  is indistinguishable from the NMAR model $\pi_{\mathbf{r}}(y \mid \gamma)$ and $p(y \mid \theta)$ based on the observed data (Molenberghs et al, 2008).

The mapping under monotone NMAR

- Let $0 = \mathbf{r}_0 < \mathbf{r}_1 < \cdots < \mathbf{r}_K$ be all the monotone missing data patterns observed. Then

$$\begin{aligned} q\{\mathbf{r}_1(y)\} &= \frac{P(\mathbf{R} \geq \mathbf{r}_1 \mid \mathbf{r}_1(y)\} p\{\mathbf{r}_1(y)\}}{1 - P(\mathbf{R} = \mathbf{r}_0)}, \\ P_q(R = \mathbf{r}_1 \mid \mathbf{Y}) &= \pi_{\mathbf{r}_1}\{\mathbf{r}_1(y)\} \frac{1 - P(\mathbf{R} = \mathbf{r}_0)}{P(\mathbf{R} \geq \mathbf{r}_1 \mid \mathbf{r}_1(y)\}}, \end{aligned}$$

  where P_q denotes the probability calculated under the equivalent-MAR model.

- In general,

$$\begin{aligned} q\{\mathbf{r}_j(y)\} &= \frac{P(\mathbf{R} \geq \mathbf{r}_j \mid \mathbf{r}_j(y)\} p\{\mathbf{r}_j(y)\}}{1 - \sum_{k=0}^{j-1} P_q\{\mathbf{R} = \mathbf{r}_k \mid \mathbf{r}_k(y)\}}, \\ P_q(R = \mathbf{r}_j \mid \mathbf{Y}) &= \pi_{\mathbf{r}_j}\{\mathbf{r}_j(y)\} \frac{1 - \sum_{k=0}^{j-1} P_q\{\mathbf{R} = \mathbf{r}_k \mid \mathbf{r}_k(y)\}}{P\{\mathbf{R} \geq \mathbf{r}_j \mid \mathbf{r}_j(y)\})}. \end{aligned}$$

Possible benefit of the MAR-equivalent transformation

- The MAR-equivalent mapping may be used for data analysis as if the observed data were MAR.
- The likelihood analysis ignoring the missing data mechanism model, if can be carried out, may be consistent and asymptotically normal, but is not fully efficient.
- Doubly robust estimation approach may be constructed based on the MAR-equivalent model for NMAR.
- All these are contingent on the parameter identifiability of the MAR-equivalent full data model $q(y \mid \theta, \gamma)$.

Why NMAR is "troublesome"?

- The MAR-equivalent full data model $q(y \mid \gamma, \theta)$ often has changed parameter identifiability from the original full data model: weakly identifiable or unidentifiable parameters may emerge.
- The weakly identifiable or unidentifiable parameters are preserved by MAR if the MAR missing data mechanism model is ignored.
- If the MAR missing data mechanism model is taken into consideration, the identifiability may be improved because the MAR missing data model may share parameters with the full data model. However, the improvement may not be enough to restore the parameter identifiability of the original full data model.

A simple example: I

- Consider a simple example of NMAR with (Y_1, Y_2), where Y_1 is binary and Y_2 is continuous random variables.

$$P(Y_1 = 1 \mid Y_2) = \Phi(\alpha_0 + \alpha_1 Y_2),$$

  where Φ is a standard normal distribution function.

- The missing data forms two patterns: one is $\mathbf{R} = (1, 1)$ and the other is $\mathbf{R} = (0, 1)$.

$$P(\mathbf{R} = \mathbf{1} \mid y_1, y_2) = \pi(y_1, y_2).$$

- The MAR-equivalent model has

$$\begin{aligned} q(y_1, y_2) &= \frac{\pi(y_1, y_2)p(y_1, y_2)}{\int \pi(y_1, y_2)p(y_1 \mid y_2)dy_1}, \\ P_q(\mathbf{R} = \mathbf{1} \mid y_1, y_2) &= \int \pi(y_1, y_2)p(y_1 \mid y_2)dy_1. \end{aligned}$$

A simple example: II

- When $\pi(y_1, y_2) = \pi(y_1)$, the MAR-equivalent model has

$$q(y_1 = 1 \mid y_2) = \frac{\exp\left\{\gamma_0 + \log \frac{\Phi(\alpha_0+\alpha_1 Y_2)}{1-\Phi(\alpha_0+\alpha_1 Y_2)}\right\}}{1 + \exp\left\{\gamma_0 + \log \frac{\Phi(\alpha_0+\alpha_1 Y_2)}{1-\Phi(\alpha_0+\alpha_1 Y_2)}\right\}},$$

  where $\pi_k = \pi(k)$, $k = 0, 1$, and $\gamma_0 = log(\pi_1/\pi_0)$.
- γ_0 and α_0 are not identifiable from $q(y_1 \mid y_2)$ when $\alpha_1 = 0$.
- When the missing data mechanism model is taken into consideration,

$$P_q(\mathbf{R} = \mathbf{1} \mid y_1, y_2) = \pi_0 \left[e^{\gamma_0}\Phi(\alpha_0 + \alpha_1 Y_2) + \{1 - \Phi(\alpha_0 + \alpha_1 Y_2)\}\right].$$

- π_0 and thus γ_0 are not identifiable when $\alpha_1 = 0$.

A simple example: III

- When $\alpha_1 \neq 0$, γ_0 and α_0 are identifiable from $q(y_1 \mid y_2)$ if the observed y_2 values have at least 3 distinctive values (Chen, 2011).
- When the missing data mechanism model is taken into consideration, π_0 is also identifiable when $\gamma_0, \alpha_0, \alpha_1$ are identifiable.
- Estimation of $\pi_0, \pi_1, \alpha_0, \alpha_1$ can be computationally difficult when α_1 is close to 0 but is not 0. The likelihood inference on these parameters can also be problematic because it is hard to tell if the true $\alpha_1 = 0$ or not from the data.

A general example: I

- Tang et al (2003) considered a general parametric model

$$P(Y_1 \mid Y_2) = f(Y_1 \mid Y_2, \theta),$$

and Y_1 is subject to missing values with

$$P(R = 1 \mid Y_1, Y_2) = \pi(Y_1).$$

- Zhao and Shao (2014) considered an extension with

$$P(Y_1 \mid Y_2, Y_3) = f(Y_1 \mid Y_2, Y_3, \theta),$$

and Y_1 is subject to missing values with

$$P(R = 1 \mid Y_1, Y_2, Y_3) = \pi(Y_1, Y_2).$$

A general example: II

- The MAR equivalent has

$$\begin{aligned} q(y_1 \mid y_2, y_3) &= \frac{\pi(y_1, y_2) f(y_1 \mid y_2, y_3)}{\int \pi(y_1, y_2) f(y_1 \mid y_2, y_3) dy_1}, \\ P_q(R = 1 \mid Y_1, Y_2, Y_3) &= \int \pi(y_1, y_2) f(y_1 \mid y_2, y_3) dy_1. \end{aligned}$$

- Let the odds ratio representation of $f(y_1 \mid y_2, y_3)$ be

$$f(y_1 \mid y_2, y_3) = \frac{\eta(y_1; y_3 \mid y_2) f(y_1 \mid y_2, y_{30})}{\int \eta(y_1; y_3 \mid y_2) f(y_1 \mid y_2, y_{30}) dy_1},$$

where $\eta(y_1, y_3 \mid y_2)$ is the conditional odds ratio of y_1 versus y_3 conditional on y_2.

A general example: III

- The odds ratio representation of $q(y_1 \mid y_2, y_3)$ is

$$q(y_1 \mid y_2, y_3) = \frac{\eta(y_1; y_3 \mid y_2) f^*(y_1 \mid y_2, y_{30})}{\int \eta(y_1; y_3 \mid y_2) f^*(y_1 \mid y_2, y_{30}) dy_1},$$

  where

$$f^*(y_1 \mid y_2, y_{30}) = \frac{\pi(y_1, y_2) f(y_1 \mid y_2, y_{30})}{\int \pi(y_1, y_2) f(y_1 \mid y_2, y_{30}) dy_1}.$$

- $\eta((y_1; y_3 \mid y_2)$ is identified from the likelihood ignoring the MAR-equivalent missing data mechanism model.
- If $\eta(y_1; y_3 \mid y_2) \equiv 1$, parameters in $f(y_1 \mid y_2)$ are unidentifiable. In general, they are weakly identifiable.

Weakly identifiability in general NMAR problems

- For a general NMAR problem, weakly identifiability is almost always an issue.
- Weakly identifiable model may or may not affect a specific data analysis depending on where the actual parameter values are.
- For the probit model, if $\alpha_1 \gg 0$, there is no problem with the MLE. For $\alpha_1 \approx 0$, the MLE does not behave well.
- For a specific NMAR data, there is uncertainty about wherea the true parameter values are. So the weakly identifiability is relevant to statistical inference even if the true parameter value is close to but not equal to the unidentifiable point.

Estimation and inference with weakly identifiable parameters

- Estimation and inference by the likelihood approach under singular information matrix had been studied in Rotnitzky et al (2000) when the parameters are identifiable.
- For model with weakly identifiable parameters, likelihood type of inference was studied by Andrews and Cheng (2012).
- Likelihood ratio test under loss of identifiability were carefully studied by Liu and Shao (2003).
- The results show that the likelihood ratio test at the point of loss-of-identifiability follows a mixture of χ^2 distributions which can be hard to apply in practice. But the behavior of the likelihood estimator is even more involved.

Measuring the strength of parameter identifiability by Fisher information

- If there is an unidentifiable parameter in the model, Fisher information matrix is singular everywhere.
- If there is a weakly identifiable parameter in the model, Fisher information matrix is singular at some parameter values, but is positive definite at other parameter values.
- Conditional Fisher information for a given parameter may serve as a measure of strength of identifiability in weakly identified model.
- The conditional Fisher information may be difficult to compute because the maximum likelihood estimator itself is computationally difficult to find around the loss-of-identifability point.

Measuring the strength of parameter identifiability by K-L information

- Given independent priors for each parameter in the model.
- Find the marginal posterior for each parameter.
- Compare the posterior and prior distributions for each parameter by the Kullback-Leibler information.

$$KL(\alpha) = -\int p(\alpha) \log \frac{p\{\alpha \mid \mathbf{R}, \mathbf{R}(\mathbf{Y})\}}{p(\alpha)} d\alpha,$$

where $p(\alpha)$ is the prior for α.

Possible problems with identifiability strength measurements

- For conditional Fisher information,
  - finding the maximum likelihood estimator can be difficult.
  - finding the conditional Fisher information can be difficult because of the near singularity of the matrix.
- For the K-L information,
  - implementing the approach can be computationally difficult,
  - specifying different priors may lead to different strength measurements.

Sensitivity analysis

- Global sensitivity analysis varies the model parameters in the missing data mechanism model and examines θ as a function of γ.
- Local sensitivity analysis examines the derivative of $\theta(\gamma)$ with respect to γ.
- Insensitive to γ change means nonignorable missing data do not affect the parameter estimator much.
- Sensitive to γ change means nonignorable missing data affect the parameter estimator.

Interpretation of the sensitivity results

- Almost all the models for sensitivity analysis are wrong even if the original model is correctly specified.
- Fixing parameters in a model, even at the correct values, reduces the variability of the estimator.
- Use Bayesian approach and treat fixed γ as a sample from posterior distribution, then

$$Var(\theta_k \mid data) = E\{Var(\theta_k \mid \gamma, data)\} + Var\{E(\theta_k \mid \gamma, data)\}.$$

Inference based on the sensitivity analysis

- If $p(\gamma \mid data) \approx p(\gamma)$, Fixing parameter values is justifiable.
- The estimated variance of the θ_k estimator for approximate inference may be approximated by

$$\frac{1}{m}\sum_{j=1}^{m} v_{kj} + \frac{1}{m-1}\sum_{j=1}^{m}(\hat{\theta}_{kj} - \bar{\hat{\theta}}_k)^2,$$

where

$$\bar{\hat{\theta}}_k = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}_{kj}.$$

- Approximate inference on the parameter θ_k may now be carried out.

Inference based on the sensitivity analysis

- If $p(\gamma \mid data) \neq p(\gamma)$, Arbitrarily fixing parameter values is questionable.
- Fixed γ should be sampled from the posterior of $p(\gamma \mid data)$ and then

$$\frac{1}{m}\sum_{j=1}^{m} v_{kj} + \frac{1}{m-1}\sum_{j=1}^{m}(\hat{\theta}_{kj} - \bar{\hat{\theta}}_k)^2,$$

where

$$\bar{\hat{\theta}}_k = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}_{kj}.$$

- Approximate inference on the parameter θ_k may now be carried out.

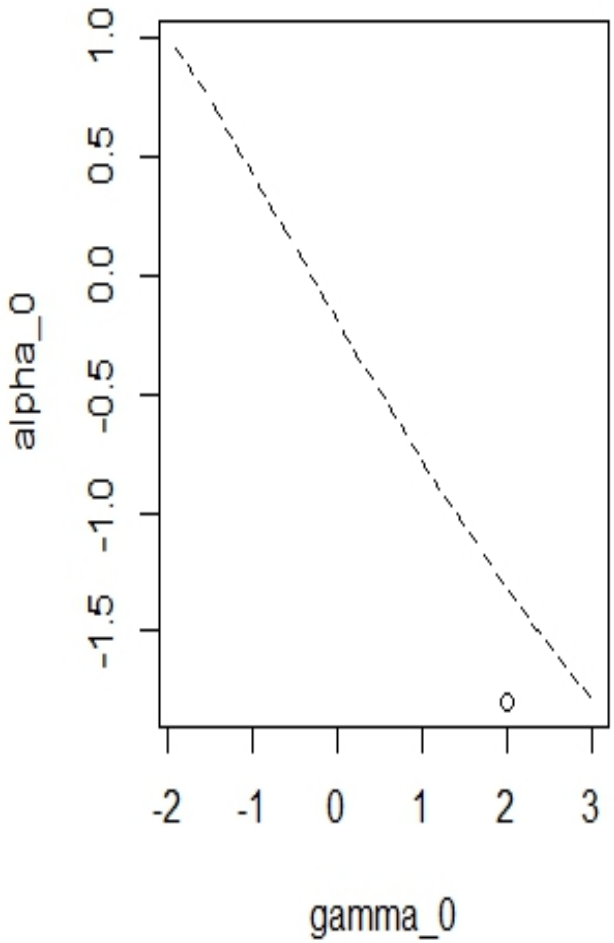
A simulation

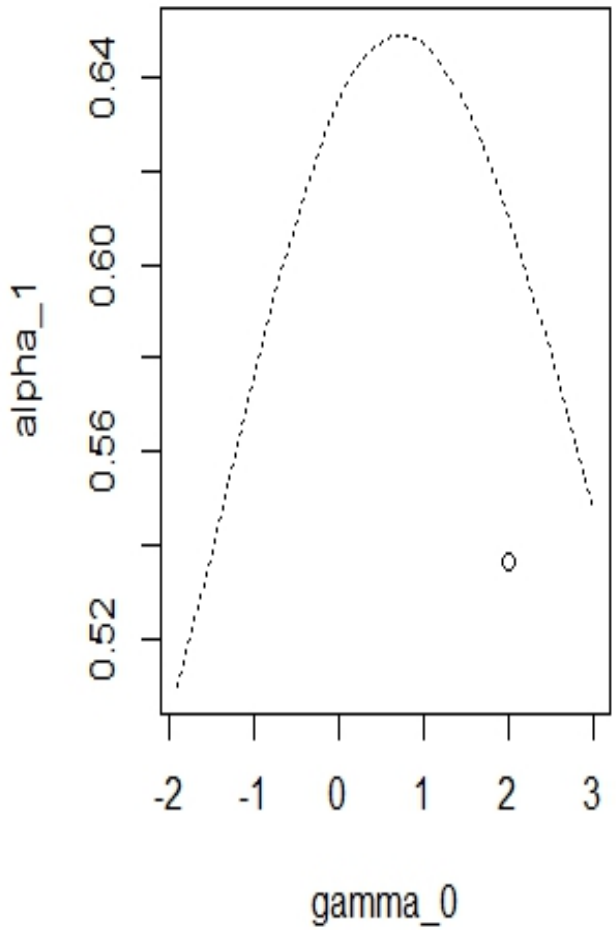
- The model:

$$q(y_1 = 1 \mid y_2) = \frac{\exp\left\{\gamma_0 + \log \frac{\Phi(\alpha_0+\alpha_1 Y_2)}{1-\Phi(\alpha_0+\alpha_1 Y_2)}\right\}}{1 + \exp\left\{\gamma_0 + \log \frac{\Phi(\alpha_0+\alpha_1 Y_2)}{1-\Phi(\alpha_0+\alpha_1 Y_2)}\right\}}.$$

- $\gamma_0 = 0.693$, $\alpha_0 = -2$, and $\alpha_1 = 0.3$.

Sensitivity analysis and MLE





A simulation study

- Covariate Z follows a normal distribution with mean zero and unit variance.
- The observed data follow a nonlinear logistic regression model,

$$\text{logit}P(Y=1|Z) = \alpha_0 + \log \eta(U; Z|\beta_0, \beta_1, p_{10}, p_{01}),$$

where $\eta(Y=1; Z|\beta_0, \beta_1, p_{10}, p_{01})\}$ is

$$\frac{\{p_{10}+(1-p_{01})e^{\beta_0+\beta_1 Z}\}\{1-p_{10}+p_{01}e^{\beta_0+\beta_1 Z_0}\}}{\{p_{10}+(1-p_{01})e^{\beta_0+\beta_1 Z_0}\}\{1-p_{10}+p_{01}e^{\beta_0+\beta_1 Z}\}}.$$

- The parameters were set to $\beta_0 = -1$, $\beta_1 = 1$, $p_0 = 0.1$, and $p_1 = 0.2$.
- The design has $1:1$ ratio of cases versus controls.
- We assume p_{10} and p_{01} are known and estimate $(\alpha_0, \beta_0, \beta_1)$ from the observed data.

Simulation results

	Slope($\beta_1 = 1.0$)			Intercept($\beta_0 = -1.0$)		
Methods	Bias	Mvar	Evar	Bias	Mvar	Evar
			Sample Size=5000			
Clogist.	-0.37	0.0014	0.0017			
MLE	0.01	0.0052	0.0067	0.01	0.00041	0.088
RMLE(T)	0.00	0.0051	0.0054			
			Sample Size=1000			
Clogist.	-0.36	0.0073	0.0082			
MLE	0.08	0.029	0.037	0.00	0.0020	0.43
RMLE(T)	0.02	0.027	0.027			
			Sample Size=200			
Clogist.	-0.34	0.039	0.044			
MLE	0.52	0.24	0.97	-0.04	0.010	3.89
RMLE(T)	0.07	0.16	0.16			

Table legends

Bias=estimated-truth, Mvar:average of the estimated variance, Evar:empirical estimate of the variance based on the parameter estimates; Logist: logistic regression applied directly to the case-control data, Clogist: Corrected logistic regression parameter estimator, MLE: maximum likelihood estimator, PMLE: penalized maximum likelihood estimator with $\lambda = n^{1/4}$ and m guessed at $\beta_0 + 1$, RMLE(T): maximum likelihood estimator with β_0 fixed at the true value, RMLE(G): maximum likelihood estimator with β_0 fixed at the guessed value.

What does the simulation result tell us?

- The MLE have small biases when the sample size is very large. It can have a significantly large bias when the sample size is not very large.
- The variance estimate of the MLE poorly estimate the true variance, especially when the sample size is not very large.
- The variance of the MLE for β_0 is particular difficult to estimate which may reflect the limited information on β_0 in the case-control sample.
- The corrected logistic regression has very large biases which are not reduced as the sample size increases.

Summary

- NMAR may be tranformed into MAR for data analysis.
- The transformation changes the parameter identifiability, and as a result the transformed missing data mechanism model is MAR but not ignorable.
- Weakly identifiable parameters in a model can have an effect on statistical inference even when the true parameter values are not at the loss-of-identifiability point.
- A measure of strength of parameter identifiability combined with sensitivity analysis may serve as an approximation for inference in practice.