Chapter 8: Nonignorable Missing Data (Part 1)

Introduction

- (X, Y): random variable, y is subject to missingness
- Response indicator function

$$\delta_i = \begin{cases} 1 & \text{if } y_i \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

- Nonignorable nonresponse (or Non-MAR)

$$f(y \mid \mathbf{x}) \neq f(y \mid \mathbf{x}, \delta = 1).$$

- In general,

$$f(y \mid \mathbf{x}, \delta = 1) = \frac{P(\delta = 1 \mid \mathbf{x}, y)}{P(\delta = 1 \mid \mathbf{x})} f(y \mid \mathbf{x}).$$

Thus, $P(\delta = 1 \mid \mathbf{x}, y) \neq P(\delta = 1 \mid \mathbf{x})$ implies nonignorable nonresponse.

Observed likelihood

- Assume that $\delta \mid (x, y) \sim \text{Bernoulli}\{\pi(x, y)\}$.
- If $\pi(X, Y) = \pi(X, Y; \phi)$ for some ϕ, we may use the observed likelihood to find the MLE of ϕ:

$$\begin{aligned} L_{\text{obs}}(\phi) &= \prod_{\delta_i=1} f\left(y_i \mid \mathbf{x}_i\right) \pi\left(\mathbf{x}_i, y_i; \phi\right) \\ &\quad \times \prod_{\delta_i=0} \int f\left(y \mid \mathbf{x}_i\right) \left\{1 - \pi\left(\delta_i \mid \mathbf{x}_i, y; \phi\right)\right\} dy. \end{aligned}$$

- Under what conditions are the parameters identifiable (or estimable)?

Definition

Identifiability

Let $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$ be a statistical model with parameter space in Θ. We say that $\mathcal{P}$ is identifiable if the mapping $\theta \to P_\theta$ is one-to-one:

$$P_{\theta_1} = P_{\theta_2} \quad \text{implies} \quad \theta_1 = \theta_2 \quad \text{for all } \theta_1, \theta_2 \in \Theta.$$

That is, if $F(\mathbf{z}; \theta)$ is the distribution function from P_θ then for any θ_1 and θ_2 in Θ such that $\theta_1 \neq \theta_2$, it implies

$$F(\mathbf{z}; \theta_1) \neq F(\mathbf{z}, \theta_2)$$

for some $\mathbf{z}$.

Remark

Identifiability is a concept closely related to the ability to estimate the parameters of a model from a sample generated by the model.

Lemma 8.1

Define

$$O(X, Y) = \frac{P(\delta = 0 \mid X, Y)}{P(\delta = 1 \mid X, Y)}$$

and

$$\tilde{O}(X) = \frac{P(\delta = 0 \mid X)}{P(\delta = 1 \mid X)}.$$

Then, $\tilde{O}(X)$ satisfies

$$\tilde{O}(X) = E\left\{O(X, Y) \mid X, \delta = 1\right\}.$$

Alternative expression

$$
\begin{aligned}
L_{obs}(\phi) &= \prod_{\delta_i=1} f\left(y_i \mid \mathbf{x}_i\right) \pi\left(\mathbf{x}_i, y_i; \phi\right) \times \prod_{\delta_i=0} \int f\left(y \mid \mathbf{x}_i\right) \left\{1 - \pi\left(\delta_i \mid \mathbf{x}_i, y; \phi\right)\right\} dy \\
&= \prod_{i=1}^{n} \left\{f(y_i \mid x_i, \delta_i = 1)\right\}^{\delta_i} \times \prod_{i=1}^{n} \{\tilde{\pi}(x_i; \phi)\}^{\delta_i} \{1 - \tilde{\pi}(x_i; \phi)\}^{1-\delta_i} \qquad (1)
\end{aligned}
$$

where

$$
\tilde{\pi}(x; \phi) = P(\delta = 1 \mid x; \phi) = \int \pi(x, y; \phi) f(y \mid x) dy.
$$

- To investigate the identifiability, we may consider the second term only.
- The log-likelihood function based on the conditional response probability

$$\tilde{\ell}(\phi) = \sum_{i=1}^{n} [\delta_i \log \tilde{\pi}(X_i; \phi) + (1 - \delta_i) \log\{1 - \tilde{\pi}(X_i, \phi)\}]$$

where

$$\tilde{\pi}(X; \phi) = \frac{1}{1 + \tilde{O}(X; \phi)}$$

and $\tilde{O}(X; \phi) = E\{O(X, Y; \phi) \mid X, \delta = 1\}$.

Lemma 2

- Let $f(y \mid x, \delta = 1; \gamma_0)$ be a parametric model known up to γ; $f(y \mid x, \delta = 1; \gamma)$ is identifiable. Also let $E_1(\cdot \mid x) = E(\cdot \mid x, \delta = 1)$.
- Model $\pi(X, Y; \phi)$ is identified if and only if the mapping

$$\phi \mapsto E_1\{O(x, Y; \phi) \mid x\}$$

  is one-to-one, almost everywhere.
- The proof is given in Morikawa and Kim (2021).

Identifiable Model

Example 8.1

- Suppose that
$$y_i \mid (x_i, \delta_i = 1) \sim N(\tau(x_i), \sigma^2).$$
Assume that x_i is always observed but we observe y_i only when $\delta_i = 1$ where $\delta_i \sim Bernoulli\left[\pi_i\left(\phi\right)\right]$ and
$$\pi_i\left(\phi\right) = \frac{\exp\left(\phi_0 + \phi_1 x_i + \phi_2 y_i\right)}{1 + \exp\left(\phi_0 + \phi_1 x_i + \phi_2 y_i\right)}.$$

- Then,
$$\begin{aligned} E_1\{O(x, Y; \phi) \mid x\} &= E_1\{\exp\left(-\phi_0 - \phi_1 x - \phi_2 Y\right) \mid x\} \\ &= \exp\{-\phi_0 - \phi_1 x - \phi_2 \tau(x) - {\phi_2}^2 \sigma^2 / 2\}. \end{aligned}$$

- Therefore, the model is identifiable unless $\tau(x)$ is constant or linear.

Theorem 1 (Wang et al., 2014)

Suppose that we can decompose the covariate vector $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ such that

$$g(\delta|y, \mathbf{x}) = g(\delta|y, \mathbf{x}_1) \tag{2}$$

and, for any given $\mathbf{x}_1$, there exist $\mathbf{x}_2^{(1)}$ and $\mathbf{x}_2^{(2)}$ such that

$$f(y|\mathbf{x}_1, \mathbf{x}_2 = \mathbf{x}_2^{(1)}) \neq f(y|\mathbf{x}_1, \mathbf{x}_2 = \mathbf{x}_2^{(2)}). \tag{3}$$

Under some other minor conditions, all the parameters in f and g are identifiable.

Back to Example 8.1

- Suppose that $\mathbf{x} = (x_1, x_2)$ and

$$y_i \mid (\mathbf{x}_i, \delta_i = 1) \sim N(\tau(\mathbf{x}_i), \sigma^2).$$

Assume $\delta_i \sim Bernoulli\left[\pi_i\left(\phi\right)\right]$ where

$$\pi_i\left(\phi\right) = \frac{\exp\left(\phi_0 + \phi_1 x_{1i} + \phi_2 y_i\right)}{1 + \exp\left(\phi_0 + \phi_1 x_{1i} + \phi_2 y_i\right)}.$$

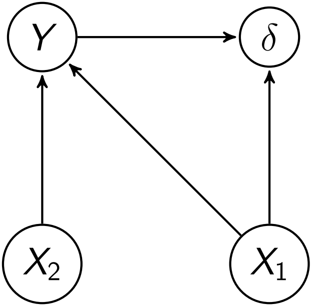
- Then,

$$\begin{aligned} E_1\{O(x_1, Y; \phi) \mid \mathbf{x}\} &= E_1\{\exp\left(-\phi_0 - \phi_1 x_1 - \phi_2 Y\right) \mid \mathbf{x}\} \\ &= \exp\{-\phi_0 - \phi_1 x_1 - \phi_2 \tau(\mathbf{x}) - {\phi_2}^2 \sigma^2 / 2\}. \end{aligned}$$

- Therefore, the model is identifiable as long as $\tau(\mathbf{x})$ is a non-constant function of x_2

Remark

- Condition (2) means
$$\delta \perp \mathbf{x}_2 \mid y, \mathbf{x}_1.$$
- That is, given $(y, \mathbf{x}_1)$, $\mathbf{x}_2$ does not help in explaining δ.



- We may call $\mathbf{x}_2$ the nonresponse instrument variable.

§8.2 Conditional Likelihood approach

- We are interested in estimating θ in $f(y \mid \mathbf{x}; \theta)$.
- The response probability $\pi(x, y) = P(\delta = 1 \mid \mathbf{x}, y)$ is known.
- By (1), we can express the observed likelihood as

$$L_{obs}(\theta) = \prod_{\delta_i=1} f(y_i \mid \mathbf{x}_i, \delta_i = 1) \times \prod_{i=1}^{n} \{\tilde{\pi}(\mathbf{x}_i)\}^{\delta_i} \{1 - \tilde{\pi}(\mathbf{x}_i)\}^{1-\delta_i},$$

  where $\tilde{\pi}(\mathbf{x}) = E\{\pi(\mathbf{x}, Y) \mid \mathbf{x}; \theta\}$.
- The conditional likelihood is defined to be the first component:

$$L_c(\theta) = \prod_{\delta_i=1} f_1(y_i \mid \mathbf{x}_i, \delta_i = 1) = \prod_{\delta_i=1} \frac{f(y_i \mid \mathbf{x}_i; \theta)\pi(\mathbf{x}_i, y_i)}{\int f(y \mid \mathbf{x}_i; \theta)\pi(\mathbf{x}_i, y)dy}.$$

Maximum Conditional Likelihood estimation

- The score function derived from the conditional likelihood is

$$
\begin{aligned}
S_c(\theta) &= n^{-1} \frac{\partial}{\partial \theta} \ln L_c(\theta) \\
&= n^{-1} \sum_{i=1}^{n} \delta_i \left[S_i(\theta) - E\left\{ S_i(\theta) \mid \mathbf{x}_i, \delta_i = 1; \theta \right\} \right] \qquad (4) \\
&= n^{-1} \sum_{i=1}^{n} \delta_i \left[S_i(\theta) - \frac{E\left\{ S_i(\theta)\pi_i \mid \mathbf{x}_i; \theta \right\}}{E\left(\pi_i \mid \mathbf{x}_i; \theta \right)} \right], \qquad (5)
\end{aligned}
$$

  where $S_i(\theta) = \partial \ln f\left(y_i \mid \mathbf{x}_i; \theta\right) / \partial \theta$.
- The second term $E\left\{ S_i(\theta) \mid \mathbf{x}_i, \delta_i = 1; \theta \right\}$ can be understood as a bias term of the complete-sample score function.

Remark

- If $\pi(x, y) = \pi(x)$, we have

$$
\begin{aligned}
f(y \mid x, \delta = 1; \theta) &= \frac{f(y \mid x; \theta)\pi(x, y)}{\int f(y \mid x; \theta)\pi(x, y)dy} \\
&= \frac{f(y \mid x; \theta)\pi(x)}{\int f(y \mid x; \theta)\pi(x)dy} \\
&= \frac{f(y \mid x; \theta)}{\int f(y \mid x; \theta)dy} = f(y \mid x; \theta).
\end{aligned}
$$

Thus, propensity score function can be safely ignored and the bias-correction term in (4) is equal to zero.

Computation

- Assume that $\pi(\mathbf{x}, y)$ is known.
- We wish to solve $S_c(\theta) = 0$, by applying the Fisher-scoring method.
- Recall (5):

$$S_c(\theta) = n^{-1} \sum_{i=1}^{n} \delta_i \left[S_i(\theta) - \frac{E\left\{S_i(\theta)\pi_i \mid \mathbf{x}_i; \theta\right\}}{E\left(\pi_i \mid \mathbf{x}_i; \theta\right)} \right].$$

Thus, we can obtain

$$\begin{aligned} \frac{\partial}{\partial \theta'} S_c(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \delta_i \dot{S}_i(\theta) \\ &- \frac{1}{n} \sum_{i=1}^{n} \delta_i \frac{E\left\{\dot{S}_i \pi_i \mid \mathbf{x}_i; \theta\right\} + E\left\{S_i S_i' \pi_i \mid \mathbf{x}_i; \theta\right\}}{E\left(\pi_i \mid \mathbf{x}_i; \theta\right)} \\ &+ \frac{1}{n} \sum_{i=1}^{n} \delta_i \frac{\left\{E\left(S_i \pi_i \mid \mathbf{x}_i; \theta\right)\right\}^{\otimes 2}}{\left\{E\left(\pi_i \mid \mathbf{x}_i; \theta\right)\right\}^2}. \end{aligned}$$

- Hence,

$$
\begin{aligned}
\mathcal{I}_c(\theta) &= -E\left\{\frac{\partial}{\partial\theta'} S_c(\theta)\right\} \\
&= E\left[n^{-1}\sum_{i=1}^{n} E\left\{S_i S_i' \pi_i \mid \mathbf{x}_i; \theta\right\} - \frac{\left\{E\left(S_i\pi_i \mid \mathbf{x}_i; \theta\right)\right\}^{\otimes 2}}{E\left(\pi_i \mid \mathbf{x}_i; \theta\right)}\right] \quad (6)
\end{aligned}
$$

- The Fisher-scoring method for obtaining the MLE from the conditional likelihood is then given by

$$
\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} + \left\{\mathcal{I}_c(\hat{\theta}^{(t)})\right\}^{-1} S_c(\hat{\theta}^{(t)}), \qquad t = 0, 1, 2, \ldots
$$

Statistical Properties 1

Asymptotic normality

- Under some regularity conditions, the solution $\hat{\theta}_c$ to $S_c(\theta) = 0$ satisfies
$$\sqrt{n}(\hat{\theta}_c - \theta_0) \xrightarrow{\mathcal{L}} N\left(0, \mathcal{I}_c^{-1}\right), \tag{7}$$
where $\mathcal{I}_c = \mathcal{I}_c(\theta_0)$ in (6).
- Works only when $\pi(x, y)$ is a known function.

Statistical Properties 2

- Consider a class of estimating equations of the form

$$\sum_{i=1}^{n} \delta_i U(\theta; \mathbf{x}_i, y_i) = 0, \tag{8}$$

  where the function U satisfies $E\{\delta U(\theta; \mathbf{x}, Y)\} = 0$.

- The choice of

$$S_c(\theta; \mathbf{x}, y) = S(\theta; \mathbf{x}, y) - E\{S(\theta; \mathbf{x}, Y) \mid \mathbf{x}, \delta = 1; \theta\}$$

  belong to the class in (8). Also, $U(\theta; \mathbf{x}, y) = S(\theta; \mathbf{x}, y)/\pi(\mathbf{x}, y)$ also satisfies $E\{\delta U(\theta; \mathbf{x}, y)\} = 0$.

- Which one is better?

Theorem 8.1

Theorem

Let $\hat{\theta}_u$ be the estimator obtained through solving (8), and assume that the regularity conditions for the following standard asymptotic expansion holds.

$$\hat{\theta}_u = \theta - n^{-1} M_u^{-1} \sum_{i=1}^{n} \delta_i U(\theta; \mathbf{x}_i, y_i) + o_p(n^{-1/2}), \tag{9}$$

where $M_u = E\{\delta \dot{U}(; \mathbf{x}, y)\}$ and $\dot{U}(\theta; \mathbf{x}, y) = \partial U(\theta; \mathbf{x}, y)/\partial \theta'$. Then, ignoring the smaller order terms,

$$V(\hat{\theta}_u) \geq n^{-1} \mathcal{I}_c^{-1} = V(\hat{\theta}_c), \tag{10}$$

which suggests that the conditional MLE $\hat{\theta}_c$ achieves the lower bound in (10).

Proof

§8.3 Pseudo Likelihood approach

Idea

- Consider bivariate (x_i, y_i) with density $f(y \mid x; \theta)h(x)$ where y_i are subject to missingness.
- We are interested in estimating θ.
- Suppose that $Pr(\delta = 1 \mid x, y)$ depends only on y. (i.e. x is nonresponse instrument)
- Note that $f(x \mid y, \delta) = f(x \mid y)$.
- Thus, we can consider the following conditional likelihood

$$L_c(\theta) = \prod_{\delta_i=1} f(x_i \mid y_i, \delta_i = 1) = \prod_{\delta_i=1} f(x_i \mid y_i).$$

- We can consider maximizing the pseudo likelihood

$$L_p(\theta) = \prod_{\delta_i=1} \frac{f(y_i \mid x_i; \theta)\hat{h}(x_i)}{\int f(y_i \mid x; \theta)\hat{h}(x)dx},$$

where $\hat{h}(x)$ is a consistent estimator of the marginal density of x.

Pseudo Likelihood approach

Idea

- We may use the empirical density in $\hat{h}(x)$. That is, $\hat{h}(x) = 1/n$ if $x = x_i$. In this case,

$$L_c(\theta) = \prod_{\delta_i=1} \frac{f(y_i \mid x_i; \theta)}{\sum_{k=1}^{n} f(y_i \mid x_k; \theta)}.$$

- We can extend the idea to the case of $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ where $\mathbf{x}_2$ is a nonresponse instrument. In this case, the conditional likelihood becomes

$$\prod_{i:\delta_i=1} p(\mathbf{x}_{2i} \mid y_i, \mathbf{x}_{1i}; \theta) = \prod_{i:\delta_i=1} \frac{f(y_i \mid \mathbf{x}_i; \theta) p(\mathbf{x}_{2i} | \mathbf{x}_{1i})}{\int f(y_i \mid \mathbf{x}_i; \theta) p(\mathbf{x}_{2i} | \mathbf{x}_{1i}) d\mathbf{x}_{2i}}. \quad (11)$$

Pseudo Likelihood approach

- Let $\hat{p}(\mathbf{x}_2|\mathbf{x}_1)$ be an estimated conditional probability density of $\mathbf{x}_2$ given $\mathbf{x}_1$. Substituting this estimate into the likelihood in (11), we obtain the following pseudo likelihood:

$$\prod_{i:\delta_i=1} \frac{f(y_i \mid \mathbf{x}_i; \theta)\hat{p}(\mathbf{x}_{2i}|\mathbf{x}_{1i})}{\int f(y_i \mid \mathbf{x}_i; \theta)\hat{p}(\mathbf{x}_{2i}|\mathbf{x}_{1i})d\mathbf{x}_{2i}}. \tag{12}$$

- The pseudo maximum likelihood estimator (PMLE) of θ, denoted by $\hat{\theta}_p$, can be obtained by solving

$$S_p(\theta; \hat{\alpha}) \equiv \frac{1}{n}\sum_{\delta_i=1}[S(\theta; \mathbf{x}_i, y_i) - E\{S(\theta; \mathbf{x}_i, y_i) \mid y_i, \mathbf{x}_{1i}; \theta, \hat{\alpha}\}] = 0$$

for θ, where $S(\theta; \mathbf{x}, y) = \partial \log f(y \mid \mathbf{x}; \theta)/\partial\theta$ and

$$E\{S(\theta; \mathbf{x}_i, y_i) \mid y_i, \mathbf{x}_{1i}; \theta, \hat{\alpha}\} = \frac{\int S(\theta; \mathbf{x}_i, y_i)f(y_i \mid \mathbf{x}_i; \theta)p(\mathbf{x}_{2i} \mid \mathbf{x}_{1i}; \hat{\alpha})d\mathbf{x}_{2i}}{\int f(y_i \mid \mathbf{x}_i; \theta)p(\mathbf{x}_{2i} \mid \mathbf{x}_{1i}; \hat{\alpha})d\mathbf{x}_{2i}}.$$

Pseudo Likelihood approach

- The Fisher-scoring method for obtaining the PMLE is given by

$$\hat{\theta}_p^{(t+1)} = \hat{\theta}_p^{(t)} + \left\{ \mathcal{I}_p \left(\hat{\theta}^{(t)}, \hat{\alpha} \right) \right\}^{-1} S_p(\hat{\theta}^{(t)}, \hat{\alpha})$$

  where

$$\begin{aligned} & \mathcal{I}_p \left(\theta, \hat{\alpha} \right) \\ = \; & \frac{1}{n} \sum_{\delta_i = 1} \left[E\{S(\theta; \mathbf{x}_i, y_i)^{\otimes 2} \mid y_i, \mathbf{x}_i; \theta, \hat{\alpha}\} - E\{S(\theta; \mathbf{x}_i, y_i) \mid y_i, \mathbf{x}_{1i}; \theta, \hat{\alpha}\}^{\otimes 2} \right]. \end{aligned}$$

- First considered by Tang et al. (2003) and further developed by Zhao and Shao (2015).

REFERENCES

Morikawa, K. and J. K. Kim (2021), 'Semiparametric optimal estimation with nonignorable nonresponse data', *Annals of Statistics* **49**, 2991–3014.

Tang, G., R. J. A. Little and T. E. Raghunathan (2003), 'Analysis of multivariate missing data with nonignorable nonresponse', *Biometrika* **90**, 747–764.

Wang, S., J. Shao and J. K. Kim (2014), 'Identifiability and estimation in problems with nonignorable nonresponse', *Statistica Sinica* **24**, 1097 – 1116.

Zhao, J. and J. Shao (2015), 'Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data', *Journal of the American Statistical Association* **110**, 1577–1590.