Chapter 8: Nonignorable Missing Data (Part 2)

§8.4. Propensity model approach

Response models for nonignorable nonresponse

- Parametric response model

$$P(\delta = 1 \mid X, Y) = \frac{\exp(\phi_0 + \phi_1 X + \phi_2 Y)}{1 + \exp(\phi_0 + \phi_1 X + \phi_2 Y)} \quad (1)$$

- Semiparametric response model

$$P(\delta = 1 \mid X, Y) = \frac{\exp\{g(X) + \phi Y\}}{1 + \exp\{g(X) + \phi Y\}} \quad (2)$$

where $g(\cdot)$ is completely unspecified.

Parameter estimation

- Assume parametric response model (1).
- How to estimate ϕ?
  1. Method-of-moments (MOM) estimation
  2. Maximum likelihood estimation

- First we assume nonresponse instrumental variable X_2 in $X = (X_1, X_2)$ such that

$$P(\delta = 1 \mid X, Y) = \pi(\phi_0 + \phi_1 X_1 + \phi_2 Y)$$

  for some (ϕ_0, ϕ_1, ϕ_2).

- Kott and Chang (2010): Construct a set of estimating equations such as

$$\sum_{i=1}^{n} \left\{ \frac{\delta_i}{\pi(\phi_0 + \phi_1 X_{1i} + \phi_2 Y_i)} - 1 \right\} (1, X_{1i}, X_{2i}) = (0, 0, 0)$$

  that are unbiased to zero.

- Rigorous theory developed by Wang et al. (2014).

Optimal MOM estimation

- Consider the class of estimating equations for ϕ:

$$\hat{U}_b(\phi) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\delta_i}{\pi(x_i, y_i; \phi)} - 1\right\}\mathbf{b}(\mathbf{x}_i; \phi) = 0 \tag{3}$$

  such that the solution exists uniquely.
- Note that the solution $\hat{\phi}_b$ to (3) is asymptotically unbiased regardless of choice of $\mathbf{b}(X; \phi)$.
- What is the optimal choice of b in the sense of minimizing the asymptotic variance of $\hat{\phi}_b$?

Theorem 8.2

- The asymptotic variance is

$$V(\hat{\phi}_b) = \frac{1}{n} \cdot A_b^{-1} B_b (A_b^{-1})' \tag{4}$$

where

$$\begin{aligned} A_b &= E\{\mathbf{b} E(O \cdot S_0' \mid X)\} \\ B_b &= E\{E(O \mid X)\mathbf{b}\mathbf{b}'\}, \end{aligned}$$

$O(x,y) = \{1 - \pi(x,y)\}/\pi(x,y)$, $S_0 = S_0(\phi; x, y)$ with

$$\begin{aligned} S_\delta(\phi; x, y) &= \frac{\partial}{\partial \phi}\{\delta \ln \pi(x,y;\phi) + (1-\delta)\ln(1-\pi(x,y;\phi))\} \\ &= \frac{\{\delta - \pi(x,y;\phi)\}}{\pi(x,y;\phi)\{1-\pi(x,y;\phi)\}} \frac{\partial \pi(x,y;\phi)}{\partial \phi}. \end{aligned} \tag{5}$$

Proof

Corollary 8.1

- The asymptotic variance in (4) is minimized at

$$\begin{aligned}\mathbf{b}^*(X;\phi) &= \frac{E(O\cdot S_0 \mid X)}{E(O \mid X)}, \\ &= \frac{E_1\{\pi^{-1}O\cdot S_0 \mid X\}}{E_1\{\pi^{-1}O \mid X\}} \qquad (6)\end{aligned}$$

  where $E_1(\cdot \mid X)$ is the expectation with respect to $f_1(y \mid x) = f(y \mid x, \delta = 1)$.

- Thus, we can use

$$\hat{\mathbf{b}}^*(X;\phi) = \frac{\hat{E}_1(\pi^{-1}O\cdot S_0 \mid X)}{\hat{E}_1(\pi^{-1}O \mid X)} \qquad (7)$$

  in (3) to obtain the optimal MOM estimator of ϕ, where $\hat{E}_1(\cdot)$ is a consistent estimator of $E_1(\cdot)$.

Proof

- We can use the matrix extension of the Cauchy-Schwarz inequality (Tripathi, 1999): For two random vectors $\mathbf{x}$ and $\mathbf{y}$ of the same dimension, we have

$$E(\mathbf{x}'\mathbf{y})\{E(\mathbf{y}\mathbf{y}')\}^{-1}E(\mathbf{y}\mathbf{x}') \leq E(\mathbf{x}\mathbf{x}')$$

  where $A \geq B$ if and only if $A - B$ is non-negative definite. The equality holds if $\mathbf{x} = c \cdot \mathbf{y}$.

- Thus, as long as the inverses exist, we can obtain

$$\{E(\mathbf{y}\mathbf{x}')\}^{-1}\{E(\mathbf{y}\mathbf{y}')\}\{E(\mathbf{x}'\mathbf{y})\}^{-1} \geq \{E(\mathbf{x}\mathbf{x}')\}^{-1}.$$

- The left side of the above inequality is equal to the asymptotic variance in (4) for some $\mathbf{x}$ and $\mathbf{y}$.

Remark

- The optimal solution in (6) satisfies $A_b = B_b$. That is, it solves

$$E\{\mathbf{b}E(O \cdot S_0' \mid X)\} = E\{E(O \mid X)\mathbf{b}\mathbf{b}'\}.$$

- Note that the above equation is equivalent to

$$-E\left\{\frac{\partial}{\partial \phi'}\hat{U}_b\right\} = V\left(\hat{U}_b\right) \tag{8}$$

  where $\hat{U}_b$ is defined in (3).
- Equation (8) is closely related to the (second) Bartlett identity. It is a key condition for constructing the efficient score function.

Two adaptive methods (Morikawa and Kim, 2021)

How to compute $\hat{\mathbf{b}}^*(X;\phi)$ in (7)?

1. Nonparametric approach: Use a Kernel regression method to estimate the conditional expectation. That is,

$$\hat{E}_1\{g(Y) \mid x\} = \frac{\sum_{i=1}^n \delta_i g(y_i) K_h(x - x_i)}{\sum_{i=1}^n \delta_i K_h(x - x_i)},$$

   where $K_h(\cdot)$ is a Kernel function with bandwidth h.

2. Parametric approach: Use a parametric model for $f_1(y \mid x) = f_1(y \mid x;\gamma)$ and estimate γ from the complete-case analysis.
   1. If the model is correct, then the solution to (3) is optimal.
   2. Even if the model is incorrect, the solution to (3) is still consistent.

Optimal PS estimation (Morikawa and Kim, 2021)

- Consider the following class of estimators of $\theta = E(Y)$:

$$\hat{\theta}_{\mathrm{PS}}(m \mid \phi) = \frac{1}{n}\sum_{i=1}^{n}\left\{ m(\mathbf{x}_i) + \frac{\delta_i}{\pi(\mathbf{x}_i, y_i; \phi)}\left(y_i - m(\mathbf{x}_i)\right)\right\}, \tag{9}$$

  where ϕ satisfies $\hat{U}_b(\phi) = 0$ in (3).

- The optimal estimator among the above class is achieved at

$$m^*(\mathbf{x}) = \frac{E\left(O \cdot Y \mid \mathbf{x}\right)}{E\left(O \mid \mathbf{x}\right)}. \tag{10}$$

- Under MAR, $O(x, y) = O(x)$ and the optimal solution in (10) reduces to $m^*(\mathbf{x}) = E(Y \mid \mathbf{x})$, which is consistent with the result of Robins et al. (1994).

Justification

- Let's first consider the case when ϕ is known.
- Thus, we can express $\hat{\theta}_{\rm PS}(m) = \hat{\theta}_{\rm PS}(m \mid \phi)$ for a fixed ϕ.
- Using m^* in (10), we wish to show that

$$V\left\{\hat{\theta}_{\rm PS}(m)\right\} \geq V\left\{\hat{\theta}_{\rm PS}(m^*)\right\},$$

which is equivalent to

$$Cov\left\{\hat{\theta}_{\rm PS}(m) - \hat{\theta}_{\rm PS}(m^*), \hat{\theta}_{\rm PS}(m^*)\right\} = 0.$$

- Now,

$$
\begin{aligned}
& Cov\left\{\hat{\theta}_{\mathrm{PS}}(m)-\hat{\theta}_{\mathrm{PS}}(m^*), \hat{\theta}_{\mathrm{PS}}(m^*)\right\} \\
= \quad & -E\left[\frac{1}{n^2} \sum_{i=1}^{n} \frac{\pi_i(1-\pi_i)}{\pi_i^2}\left\{m(x_i)-m^*(x_i)\right\}\left\{y_i-m^*(x_i)\right\}\right].
\end{aligned}
$$

- The covariance term is equal to zero if

$$
E\left\{O(x, Y)\{Y-m^*(x)\} \mid x\right\}=0, \tag{11}
$$

which is satisfied at $m^*(x)$ in (10).

- Now, we consider the second case of an <u>unknown</u> ϕ.
- If ϕ is unknown and estimated by solving (3) with the optimal $\mathbf{b}^*(\mathbf{x})$ in (6), we can approximate $\hat{\theta}_{\mathrm{PS}}(m \mid \hat{\phi}^*)$ in (9) by

$$\begin{aligned} \hat{\theta}_{\mathrm{PS},\ell}(m) &= E\left\{\hat{\theta}_{\mathrm{PS}}(m) \mid \hat{U}_{b^*}^{\perp}\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{m(\mathbf{x}_i) + \mathbf{b}_i^{*\prime}\gamma^* + \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)}\left(y_i - m(\mathbf{x}_i) - \mathbf{b}_i^{*\prime}\gamma^*\right)\right\}, \end{aligned}$$

where $\mathbf{b}_i^* = \mathbf{b}^*(\mathbf{x}_i)$ and $\gamma = \gamma(m)$ satisfies

$$E\left[O(x, Y)\left\{Y - m(x) - \mathbf{b}^*(x)'\gamma\right\}\mathbf{b}^*(x)\right] = 0. \tag{12}$$

- For the choice of $m^*(\mathbf{x})$ in (10), we have $\gamma(m^*) = \mathbf{0}$.

- We have only to check

$$Cov\left\{\hat{\theta}_{\mathrm{PS},\ell}(m)-\hat{\theta}_{\mathrm{PS},\ell}(m^*),\hat{\theta}_{\mathrm{PS},\ell}(m^*)\right\}=0.$$

- Note that

$$\hat{\theta}_{\mathrm{PS},\ell}(m)-\hat{\theta}_{\mathrm{PS},\ell}(m^*)=-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\delta_i}{\pi_i}-1\right)\{m(\mathbf{x}_i)-m^*(\mathbf{x}_i)\}$$

and

$$\begin{aligned}
&Cov\left\{\hat{\theta}_{\mathrm{PS},\ell}(m)-\hat{\theta}_{\mathrm{PS},\ell}(m^*),\hat{\theta}_{\mathrm{PS},\ell}(m^*)\right\}\\
=\quad &-E\left[\frac{1}{n^2}\sum_{i=1}^{n}\frac{\pi_i(1-\pi_i)}{\pi_i^2}\{m(x_i)-m^*(x_i)\}\{y_i-m^*(x_i)\}\right]
\end{aligned}$$

- The covariance term is equal to zero as (11) holds.

Maximum likelihood estimation

- Note that, assuming for now that $f(y \mid \mathbf{x})$ is known, the observed likelihood function is

$$L_{\text{obs}}(\phi) = \prod_{\delta_i=1} f(y_i \mid \mathbf{x}_i) P(\delta_i = 1 \mid \mathbf{x}_i, y_i; \phi) \times \prod_{\delta_i=0} \int f(y \mid \mathbf{x}_i) P(\delta_i = 0 \mid \mathbf{x}_i, y; \phi)\, dy.$$

- Note that

$$\begin{aligned} S_{\text{obs}}(\phi) &\equiv \frac{\partial}{\partial \phi} \ln L_{\text{obs}}(\phi) \\ &= \sum_{i=1}^{n} [\delta_i S_1(\phi; \mathbf{x}_i, y_i) + (1-\delta_i) E\{S_0(\phi; \mathbf{x}_i, Y) \mid \mathbf{x}_i, \delta_i = 0\}] \end{aligned}$$

where $S_\delta(\phi; \mathbf{x}, y)$ is defined in (5).

Maximum Likelihood Estimation

How to compute the conditional expectation?

- Classical approach (Baker and Laird (1988); Ibrahim et al. (1999)): Assume a parametric model on $f(y \mid \mathbf{x}) = f(y \mid \mathbf{x}; \theta)$ and use the EM to solve the mean score equation of the parameters in the full joint distribution.

$$E\{S_0(\phi; \mathbf{x}_i, Y) \mid \mathbf{x}_i, \delta_i = 0\} = \frac{\int S_0(\phi; \mathbf{x}_i, y) f(y \mid \mathbf{x}_i; \theta)\{1 - \pi(\mathbf{x}_i, y; \phi)\} dy}{\int f(y \mid \mathbf{x}_i; \theta)\{1 - \pi(\mathbf{x}_i, y; \phi)\} dy}.$$

- Requires correct specification of $f(y \mid \mathbf{x}; \theta)$. Known to be sensitive to the choice of $f(y \mid \mathbf{x}; \theta)$.

New approach

Idea

Instead of specifying a parametric model for $f(y \mid \mathbf{x})$, consider specifying a parametric model for $f(y \mid \mathbf{x}, \delta = 1)$, denoted by $f_1(y \mid \mathbf{x})$. In this case,

$$E\{S_0(\phi; \mathbf{x}_i, Y) \mid \mathbf{x}_i, \delta_i = 0\} = \frac{\int S_0(\phi; \mathbf{x}_i, y) f_1(y \mid \mathbf{x}_i) O(\mathbf{x}_i, y; \phi) dy}{\int f_1(y \mid \mathbf{x}_i) O(\mathbf{x}_i, y; \phi) dy}$$

where

$$O(\mathbf{x}_1, y; \phi) = \frac{1 - \pi(\phi; \mathbf{x}, y)}{\pi(\phi; \mathbf{x}, y)}.$$

Remark

- Based on the following identity

$$f(y \mid \mathbf{x}, \delta = 0) = f(y \mid \mathbf{x}, \delta = 1) \frac{O(\mathbf{x}, y; \phi)}{E\{O(\mathbf{x}, y; \phi) \mid \mathbf{x}, \delta = 1\}}. \qquad (13)$$

- Kim and Yu (2011) considered a Kernel-based nonparametric regression method of estimating $f(y \mid \mathbf{x}, \delta = 1)$ to obtain $E(Y \mid \mathbf{x}, \delta = 0)$.

Maximum likelihood estimation

- If $f_1(y \mid x)$ is correctly specified, we can obtain the maximum likelihood estimator of ϕ by solving

$$\sum_{i=1}^{n}\left[\delta_i S_1(\phi; x_i, y_i) + (1-\delta_i)\frac{E_1\{O(x_i, Y; \phi)S_0(\phi; x_i, Y) \mid x_i\}}{E_1\{O(x_i, Y; \phi) \mid x_i\}}\right] = 0. \tag{14}$$

- EM algorithm can be used to solve (14): Update $\hat{\phi}$ by solving

$$\sum_{i=1}^{n}\left[\delta_i S_1(\phi; x_i, y_i) + (1-\delta_i)\frac{E_1\{O(x_i, Y; \hat{\phi}^{(t)})S_0(\phi; x_i, Y) \mid x_i\}}{E_1\{O(x_i, Y; \hat{\phi}^{(t)}) \mid x_i\}}\right] = 0. \tag{15}$$

- Considered by Riddles et al. (2015) for parametric $f_1(y \mid x)$ and by Morikawa et al. (2017) for non-parametric $f_1(y \mid x)$.

Efficiency comparison

- Question: Is the MLE more efficient than the optimal MOM estimator using $\hat{\mathbf{b}}_i^*$ in (7)?

- Answer: It depends...
  1. If we use a parametric model for $f_1(y \mid x)$ and the model is correctly specified, then the MLE is more efficient than the optimal MOM estimator because it uses more model assumption (the parametric model assumption on f_1).
  2. If we use a non-parametric model for $f_1(y \mid x)$, then the MLE is asymptotically equivalent to MOM estimator using

  $$b(X;\phi) = \frac{E_1(O \cdot S_0 \mid X)}{E_1(O \mid X)}.$$

  So, it is less efficient than the optimal MOM estimator using (6).

§5. Semi-parametric response model

Semiparametric response probability model

- The response probability follows from a logistic regression model

$$\pi(\mathbf{x}_i, y_i) \equiv Pr(\delta_i = 1 \mid \mathbf{x}_i, y_i) = \frac{\exp\{g(\mathbf{x}_i) + \phi y_i\}}{1 + \exp\{g(\mathbf{x}_i) + \phi y_i\}}, \qquad (16)$$

  where $g(\mathbf{x})$ is completely unspecified.
- The expression (13) can be simplified to

$$f_0(y_i \mid \mathbf{x}_i) = f_1(y_i \mid \mathbf{x}_i) \times \frac{\exp(\gamma y_i)}{E\{\exp(\gamma Y) \mid \mathbf{x}_i, \delta_i = 1\}}, \qquad (17)$$

  where $\gamma = -\phi$ and $f_1(y \mid \mathbf{x})$ is the conditional density of y given $\mathbf{x}$ and $\delta = 1$.
- Model (17) states that the density for the nonrespondents is an exponential tilting of the density for the respondents. The parameter γ is the tilting parameter that determines the amount of departure from the ignorability of the response mechanism. If $\gamma = 0$, the the response mechanism is ignorable and $f_0(y|\mathbf{x}) = f_1(y|\mathbf{x})$.

Semiparametric imputation approach

- Kim and Yu (2011): If γ is known, we can estimate $E(Y \mid \mathbf{x}, \delta = 0)$ by

$$\hat{E}_0(Y \mid \mathbf{x}; \gamma) = \frac{\sum_{i=1}^n \delta_i \exp(\gamma y_i) K_h(x - x_i) y_i}{\sum_{i=1}^n \delta_i \exp(\gamma y_i) K_h(x - x_i)},$$

  where $K_h(x)$ is a Kernel function with bandwidth h.
- Semiparametric imputation estimator for $\theta = E(Y)$:

$$\hat{\theta}_I = \frac{1}{n} \sum_{i=1}^n \left\{ \delta_i y_i + (1 - \delta_i) \hat{E}_0(Y \mid \mathbf{x}_i; \gamma) \right\}.$$

Semiparametric inverse propensity weighting method

- Based on the semiparametric response model (16).
- Under this model, we can obtain

$$E\left\{\frac{\delta}{\pi(\mathbf{x}, y)} - 1 \mid \mathbf{x}\right\} = 0,$$

which implies

$$exp\{g(\mathbf{x})\} = \frac{E\{\delta \exp(\gamma y) \mid \mathbf{x}\}}{E\{1-\delta \mid \mathbf{x}\}}.$$

- For known γ case, we can use Kernel regression estimator

$$\exp\{\hat{g}_\gamma(x)\} = \frac{\sum_{i=1}^n \delta_i \exp(\gamma y_i) K_h(x-x_i)}{\sum_{i=1}^n (1-\delta_i) K_h(x-x_i)}$$

to obtain

$$\hat{\pi}(x_i, y_i; \gamma) = \frac{\exp\{\hat{g}_\gamma(x_i) - \gamma y_i\}}{1+\exp\{\hat{g}_\gamma(x_i) - \gamma y_i\}}.$$

Semiparametric inverse propensity weighting method

Estimation of γ:

- Shao and Wang (2016) idea: Use GMM method based on some moments conditions
- Profile ML method: EM algorithm using profile likelihood
  1. E-step: Compute

$$Q_p(\gamma \mid \hat{\gamma}^{(t)}) = E\{\ell_p(\gamma) \mid \text{obs}, \hat{\gamma}^{(t)}\}$$

  where

$$\ell_p(\gamma) = \sum_{i=1}^{n} \{\delta_i \log \hat{\pi}(x_i, y_i; \gamma) + (1 - \delta_i) \log (1 - \hat{\pi}(x_i, y_i; \gamma))\} .$$

  2. M-step: Maximize $Q_p(\gamma \mid \hat{\gamma}^{(t)})$ wrt γ to obtain $\hat{\gamma}^{(t+1)}$.

REFERENCES

Baker, S. G. and N. M. Laird (1988), 'Regression analysis for categorical variables with outcome subject to nonignorable nonresponse', *Journal of the American Statistical Association* **83**, 62–69.

Ibrahim, J. G., S. R. Lipsitz and M. H. Chen (1999), 'Missing covariates in generalized linear models when the missing data mechanism is non-ignorable', *Journal of the Royal Statistical Society, Series B* **61**, 173–190.

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.

Kott, P. S. and T. Chang (2010), 'Using calibration weighting to adjust for nonignorable unit nonresponse', *Journal of the American Statistical Association* **105**, 1265–1275.

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

Morikawa, K., J. K. Kim and Y. Kano (2017), ‘Semiparametric maximum likelihood estimation under nonignorable nonresponse’, *Canadian Journal of Statistics* **45**, 393–409.

Riddles, M. K., J. K. Kim and J. Im (2015), ‘Propensity score adjustment for nonignorable nonresponse’, *Journal of Survey Statistics and Methodology* **4**, 215–245.

Robins, James M, Andrea Rotnitzky and Lue Ping Zhao (1994), ‘Estimation of regression coefficients when some regressors are not always observed’, *Journal of the American statistical Association* **89**(427), 846–866.

Shao, J. and L. Wang (2016), ‘Semiparametric inverse propensity weighting for nonignorable missing data’, *Biometrika* **103**, 175–187.

Tripathi, G. (1999), ‘A matrix extension of the cauchy-schwarz inequality’, *Economic Letters* **63**, 1–3.

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