7.3 Propensity score method

Jae-Kwang Kim

ISU

Motivation

- $z_i = (x_i, y_i)$, y_i is subject to missingness.
- Interested in estimating θ_0 which is defined by $E\{U(\theta_0; Z)\} = 0$.
- The true response probability follows from a parametric model

$$\pi_i = \pi(z_i; \phi_0)$$

  for some $\phi_0 \in \Omega$.
- The propensity score (PS) estimator of θ_0, denoted by $\hat{\theta}_{\rm PS}$, is computed by solving

$$\hat{U}_{\rm PS}(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} \delta_i \frac{1}{\hat{\pi}_i} U(\theta; z_i) = 0, \tag{1}$$

  where $\hat{\pi}_i = \pi(z_i; \hat{\phi})$ and $\hat{\phi}$ is a $\sqrt{n}$-consistent esstimator of ϕ_0.

Research Questions

- How to estimate ϕ?
  1. Maximum Likelihood approach
  2. Estimating equation approach
- What is the asymptotic properties of $\hat{\theta}_{\rm PS}$?
  - Consistency
  - Asymptotic normality
  - Variance estimation
- How to improve the efficiency of $\hat{\theta}_{\rm PS}$?

Maximum Likelihood estimation approach - Idea

- Note that we observe $(x_i, \delta_i, \delta_i y_i)$, for $i = 1, \ldots, n$.
- The joint density of $(\delta_i, \delta_i y_i, x_i)$ is

$$f(x_i)\left\{f(y_i \mid x_i)\pi(x_i, y_i; \phi)\right\}^{\delta_i}\left[\int f(y \mid x_i)\{1 - \pi(x_i, y; \phi)\}dy\right]^{1-\delta_i}$$

- Thus, assuming that $f(y \mid x)$ is known for now, the MLE $\hat{\phi}$ is obtained by maximizing

$$\ell_{\mathrm{obs}}(\phi) = \sum_{i=1}^{n}\left[\delta_i \log \pi(x_i, y_i; \phi) + (1 - \delta_i)\log\{1 - \tilde{\pi}(x_i; \phi)\}\right] \quad (2)$$

with respect to ϕ, where

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

- Under MAR, we have $\pi(x, y; \phi) = \tilde{\pi}(x; \phi)$. The observed likelihood in (2) reduces to

$$\ell_{\rm obs}(\phi) = \sum_{i=1}^{n} \left[\delta_i \log \pi(x_i; \phi) + (1 - \delta_i) \log\{1 - \pi(x_i; \phi)\}\right].$$

  Under MAR, the subscript "obs" can be safely removed.

- Note that $(\hat{\theta}_{PS}, \hat{\phi})$ is the solution to

$$\begin{aligned} \hat{U}_{\rm PS}(\theta, \phi) &= \mathbf{0} \\ \hat{S}_{\rm obs}(\phi) &= 0 \end{aligned}$$

  where $\hat{S}_{obs}(\phi) = \partial \ell_{\rm obs}(\phi) / \partial \phi$ is the observed score function for ϕ.

- Thus, we can apply the sandwich formula to obtain the asymptotic variance of $(\hat{\theta}_{PS}, \hat{\phi})$.

Lemma 7.3

Lemma

Let $U_1(\theta, \phi) = U_1(\theta, \phi; Z, \delta)$ *be an estimating function satisfying*

$$E\{U_1(\theta_0, \phi_0)\} = 0.$$

Then,

$$E\{-\partial U_1/\partial \phi\} = Cov(U_1, S) \tag{3}$$

where S *is the score function of* ϕ*.*

Note: If we set $U_1(\theta, \phi) = S(\phi)$, then (3) reduces to $E\{-\partial S(\phi)/\partial \phi\} = E\left\{S(\phi)^{\otimes 2}\right\}$, which is already presented in Chapter 2 (Theorem 2.3).

Proof

Theorem 7.2: Asymptotic properties of PS estimator

- Under some regularity conditions, the solution $(\hat{\theta}_{PS}, \hat{\phi})$ to

$$\begin{aligned} \hat{U}_1(\theta, \phi) &= \mathbf{0} \\ S(\phi) &= 0 \end{aligned}$$

is asymptotically normal with mean $(\theta_0, \phi_0)'$ and variance $A^{-1}BA'^{-1}$, where

$$\begin{aligned} A &= \begin{bmatrix} E\left\{-\partial \hat{U}_1/\partial\theta\right\} & E\left\{-\partial U_1/\partial\phi\right\} \\ E\left\{-\partial S/\partial\theta\right\} & E\left\{-\partial S/\partial\phi\right\} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \\ B &= \begin{bmatrix} V\left(\hat{U}_1\right) & C\left(\hat{U}_1, S\right) \\ C\left(S, \hat{U}_1\right) & V(S) \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}. \end{aligned}$$

Asymptotic properties of PS estimator

- Using

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22}^{-1} \\ 0 & A_{22}^{-1} \end{bmatrix},$$

we have

$$Var(\hat{\theta}_{PS}) \cong A_{11}^{-1}\left[B_{11} - A_{12}A_{22}^{-1}B_{21} - B_{12}A_{22}^{-1}A_{12}' + A_{12}A_{22}^{-1}B_{22}A_{22}^{-1}A_{12}'\right]A_{11}^{'-1}.$$

- By Lemma 7.3, $B_{22} = A_{22}$ and $B_{12} = A_{12}$. Thus,

$$V(\hat{\theta}_{PS}) \cong A_{11}^{-1}\left[B_{11} - B_{12}B_{22}^{-1}B_{21}\right]A_{11}^{'-1}. \tag{4}$$

Asymptotic properties of PS estimator

- Note that $\hat{\theta}_W = \hat{\theta}_W(\phi_0)$ with known π_i satisfies

$$V\left(\hat{\theta}_W\right) \cong A_{11}^{-1} B_{11} A_{11}^{-1'}.$$

- Therefore, ignoring the smaller order terms, we have

$$V\left(\hat{\theta}_W\right) \geq V\left(\hat{\theta}_{PS}\right). \tag{5}$$

- The result of (5) means that the PS estimator with estimated π_i is more efficient than the PS estimator with known π_i.

Remark

- Another way of understanding (4) is the projection in the Hilbert space. Given two random variables, say X and Y, with finite second moments with zero means, the projection of Y on the linear space generated by X, denoted by $\Pi(Y \mid X)$ satisfies the following two conditions:
  1. $\Pi(Y \mid X) \in \mathcal{L}(X)$, where $\mathcal{L}(X) = \{a'X; a \in \mathbb{R}^p\}$ is the linear subspace generated by X and $p = \dim(X)$.
  2. $Y - \Pi(Y \mid X)$ is orthogonal to all elements in $\mathcal{L}(X)$.
- In the linear space of random variables, the norm of X and Y is the covariance and

$$\Pi(Y \mid X) = Cov(Y, X)\{Var(X)\}^{-1} X.$$

- Also, we can define the projection of Y to the orthogonal complement of $\mathcal{L}(X)$ as

$$\Pi(Y \mid X^{\perp}) = Y - \Pi(Y \mid X).$$

- Note that, as $E(X) = 0$,

$$E\{\Pi(Y \mid X^{\perp})\} = E(Y)$$

and

$$V\left\{\Pi(Y \mid X^{\perp})\right\} = V(Y) - Cov(Y, X)\{Var(X)\}^{-1} Cov(X, Y) \leq V(Y). \tag{6}$$

- Thus, the projection of Y onto $\mathcal{L}(X^{\perp})$ with $E(X) = 0$ will always improve the efficiency.

Remark

- Standard Taylor linearization of $\hat{\theta}_{PS} \equiv \hat{\theta}_W(\hat{\phi})$ leads to

$$\hat{\theta}_{PS} \cong \hat{\theta}_W(\phi_0) - E\left\{\frac{\partial}{\partial \phi'}\hat{\theta}_W(\phi_0)\right\}\left[E\left(\frac{\partial}{\partial \phi'}S(\phi_0)\right)\right]^{-1} S(\phi_0), \quad (7)$$

when $\hat{\phi}$ is obtained by the MLE.

Remark (Cont'd)

- Using

$$\begin{aligned} \hat{\phi} - \phi_0 &= \{\mathcal{I}(\phi_0)\}^{-1} S(\phi_0) \\ V(S) &= \mathcal{I}(\phi_0) = \{V(\hat{\phi})\}^{-1} \end{aligned}$$

  we can write (7) as

$$\hat{\theta}_{PS} \cong \hat{\theta}_W - C(\hat{\theta}_W, S)\{V(S)\}^{-1} S(\phi_0)$$

  which can be understood as a projection of $\hat{\theta}_W$ to $\mathcal{L}(S^{\perp})$.

- That is, we have

$$\hat{\theta}_{PS} \cong \Pi(\hat{\theta}_W \mid S^{\perp}) \tag{8}$$

  and, by (6), result (5) holds.

Variance estimation

- Under MAR, the score equation for ϕ_0 can be written as

$$S(\phi) \equiv \sum_{i=1}^{n} \{\delta_i - \pi(x_i;\phi)\} \frac{1}{\pi(x_i;\phi)\{1-\pi(x_i;\phi)\}} \dot{\pi}(x_i;\phi) = 0, \quad (9)$$

  where $\dot{\pi}(x_i;\phi) = \partial \pi(x_i;\phi)/\partial\phi$.

- We can express (9) as

$$S(\phi) = \sum_{i=1}^{n} \left\{ \frac{\delta_i}{\pi(x_i;\phi)} - 1 \right\} \mathbf{h}(\mathbf{x}_i;\phi),$$

  where $\mathbf{h}(x_i;\phi) = \dot{\pi}(x_i;\phi)/\{1-\pi(x_i;\phi)\}$.

Variance estimation (Cont'd)

- Using (4), a plug-in variance estimator of the PS estimator is computed by

$$\hat{V}\left(\hat{\theta}_{PS}\right) = \hat{A}_{11}^{-1}\left[\hat{B}_{11} - \hat{B}_{12}\hat{B}_{22}^{-1}\hat{B}_{21}\right]\hat{A}_{11}'^{-1}$$

where $\hat{A}_{11} = n^{-1}\sum_{i=1}^{n}\delta_i\pi_i^{-1}\dot{U}(\hat{\theta}; z_i)$ and

$$\begin{aligned}
\hat{B}_{11} &= n^{-2}\sum_{i=1}^{n}\delta_i\hat{\pi}_i^{-2}U(\hat{\theta}; z_i)^{\otimes 2} \\
\hat{B}_{12} &= n^{-2}\sum_{i=1}^{n}\delta_i\hat{\pi}_i^{-1}(\hat{\pi}_i^{-1} - 1)U(\hat{\theta}; z_i)\mathbf{h}_i \\
\hat{B}_{22} &= n^{-2}\sum_{i=1}^{n}\delta_i\hat{\pi}_i^{-1}(\hat{\pi}_i^{-1} - 1)\mathbf{h}_i\mathbf{h}_i'
\end{aligned}$$

where $\hat{\theta} = \hat{\theta}_{PS}$ and $\mathbf{h}_i = \dot{\pi}_i/(1 - \pi_i)$.

Another approach

- The PS estimator of $\theta_0 = E(Y)$ can be written as

$$\hat{\theta}_{\mathrm{PS}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} y_i$$

  if we impose $\sum_{i=1}^{n} \delta_i / \hat{\pi}_i = n$ as a constraint.

- Estimating equation for ϕ_0:

$$\frac{1}{n}\sum_{i=1}^{n} \left(\frac{\delta_i}{\pi(x_i; \phi)} - 1 \right) \mathbf{b}_i = \mathbf{0}, \tag{10}$$

  where $\mathbf{b}_i = \mathbf{b}(x_i)$ contains an intercept term. As long as the PS model is correctly specified and the solution to (10) exists uniquely, the solution leads to a consistent estimator of ϕ_0.

- Linearization results

$$\hat{\theta}_{\mathrm{PS}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbf{b}_i'\gamma + \frac{\delta_i}{\pi_i}\left(y_i - \mathbf{b}_i'\gamma\right)\right\} + o_p(n^{-1/2}),$$

where $\gamma = p\lim \hat{\gamma}$, $\pi_i = \pi(\mathbf{x}_i;\phi_0)$, and

$$\hat{\gamma} = \left(\sum_{i=1}^{n}\delta_i\hat{\pi}_i^{-2}(1-\hat{\pi}_i)\mathbf{b}_i\mathbf{h}_i'\right)^{-1}\left(\sum_{i=1}^{n}\delta_i\hat{\pi}_i^{-2}(1-\hat{\pi}_i)\mathbf{h}_i'y_i\right)$$

and $\mathbf{h}_i = \dot{\pi}_i/(1-\pi_i)$.

- The linearization formula can be used to construct variance estimation easily.

Justification