Chapter 5: Propensity Score Approach

Jae-Kwang Kim

Department of Statistics, Iowa State University

Outline

Basic Setup

- $z_i = (x_i, y_i), i = 1, 2, \cdots, n$: random sample
- Parameter of interest (θ_0): defined by the (unique) solution to $E\{U(\theta; Z)\} = 0$.
- Under complete response of $z_1, \cdots, z_n$, a consistent estimator of θ is obtained by solving

$$\hat{U}_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} U(\theta; z_i) = 0$$

  for θ. We assume that the solution $\hat{\theta}_n$ is unique.
- Under some conditions, $\hat{\theta}_n$ converges in probability to θ_0.
- Note that $\hat{\theta}_n$ is asymptotically unbiased for θ^* if $E\{U(\theta^*; Z)\} = 0$.
- What if some of y_i are missing ?

Two approaches

1. Prediction model approach: use a model for y. solve

$$n^{-1}\sum_{i=1}^{n}[\delta_i U(\theta; x_i, y_i) + (1-\delta_i)E\{U(\theta; x_i, y_i) \mid x_i, \delta_i = 0\}] = 0$$

   Prediction model approach was discussed in Chapter 2-4.

2. Response model approach: use a model for δ_i (response indicator function).

Complete-case (CC) method

- Solve $\sum_{i=1}^{n} \delta_i U(\theta; z_i) = 0$ for θ.
- The CC method lead to biased estimator unless $Cov(\delta_i, U_i) = 0$, where $U_i = U(\theta_0; z_i)$. So, unless the missing mechanism is missing completely at random (MCAR), the CC method leads to biased estimation.
- Furthermore, the CC method does not make use of the observed information of x_i for $\delta_i = 0$. Thus, it is not efficient.

Weighted Complete-case (WCC) method

- Solve

$$\sum_{i=1}^{n} \delta_i \frac{1}{\pi_i} U(\theta; z_i) = 0 \tag{1}$$

  for θ, where $\pi_i = Pr(\delta_i = 1 \mid z_i)$
- The WCC method leads to unbiased estimator of θ_0 if $1/\pi_i$ is used as the weight for unit i.
- In survey sampling, π_i are known and the WCC method is very popular (Horvitz-Thompson estimation) since it does not require the model assumptions about unobserved y.

WCC method (Cont'd)

- In survey sampling, δ_i is the sampling indicator function. The sampling indicator functions are not necessarily independent. Two parameters can be considered, θ_N (finite population quantity) and θ_0 (infinite population quantity). When the finite population is a random sample from an infinite population, called superpopulation, and the parameter θ_0 is the superpopulation parameter.
- We will only consider estimation of θ_0 and independent sampling (Poisson sampling).

Properties of WCC

- Asymptotically unbiased
- Asymptotic variance: Assuming that $Cov(\delta_i, \delta_j) = 0$ for $i \neq j$,

$$V\left(\hat{\theta}_W\right) \cong \tau^{-1} V\left\{\hat{U}_W(\theta_0)\right\} \tau^{-1'}$$

where $\tau = E\left\{\dot{U}(\theta_0; Z)\right\}$ and

$$\begin{aligned} V\left\{\hat{U}_W(\theta_0)\right\} &= V\left\{\hat{U}_n(\theta_0)\right\} + E\left\{n^{-2}\sum_{i=1}^{n}\left(\pi_i^{-1}-1\right)U(\theta_0; z_i)^{\otimes 2}\right\} \\ &= n^{-1}E\left\{n^{-1}\sum_{i=1}^{n}\pi_i^{-1}U(\theta_0; z_i)^{\otimes 2} - \bar{U}_n(\theta_0)^{\otimes 2}\right\} \\ &\cong E\left\{n^{-2}\sum_{i=1}^{n}\pi_i^{-1}U(\theta_0; z_i)^{\otimes 2}\right\}. \end{aligned} \tag{2}$$

Properties of WCC

- A consistent estimator for the variance of $\hat{\theta}_W$ is computed by

$$\hat{V}\left(\hat{\theta}_W\right) = \hat{\tau}^{-1}\hat{V}_u\hat{\tau}^{-1'}$$

where

$$\hat{\tau} = n^{-1}\sum_{i=1}^{n}\delta_i\pi_i^{-1}\dot{U}(\hat{\theta}_W; z_i)$$

and

$$\hat{V}_u = n^{-2}\sum_{i=1}^{n}\delta_i\pi_i^{-2}U(\hat{\theta}_W; z_i)^{\otimes 2}.$$

Example 5.1

- Let the parameter of interest be $\theta = E(Y)$ and we use $U(\theta; z) = (y - \theta)$ to compute θ. The WCC estimator of θ can be written

$$\hat{\theta}_W = \frac{\sum_{i=1}^{n} \delta_i y_i / \pi_i}{\sum_{i=1}^{n} \delta_i / \pi_i}. \tag{3}$$

- The asymptotic variance of $\hat{\theta}_W$ in (3) is equal to, by (2),

$$n^{-2} \sum_{i=1}^{n} \pi_i^{-1} (y_i - \theta)^2 \tag{4}$$

which is consistently estimated by

$$n^{-2} \sum_{i=1}^{n} \delta_i \pi_i^{-2} \left(y_i - \hat{\theta}_W \right)^2.$$

Example 5.1 (Cont'd)

- In survey sampling, the estimator (3) is called the Hajek estimator. The asymptotic variance in (4) represents the asymptotic variance of the Hajek estimator under Poisson sampling when the parameter θ is the superpopulation parameter.
- If the parameter is the finite population parameter, the asymptotic variance of $\hat{\theta}_W$ in (3) is equal to

$$n^{-2}\sum_{i=1}^{n}(\pi_i^{-1}-1)\,(y_i-\theta_N)^2$$

which is consistently estimated by

$$n^{-2}\sum_{i=1}^{n}\delta_i\pi_i^{-1}(\pi_i^{-1}-1)\left(y_i-\hat{\theta}_W\right)^2.$$

Remark

If parameter θ is estimated by solving $\sum_{i=1}^{n} U(\theta; y_i) = 0$ under full sample. Let δ_i be independently generated from $Bernoulli(\pi_i)$ distribution with $\pi_i = \pi(y_i, z_i)$ and $\pi(\cdot)$ is a known function. We observe (y_i, z_i) only when $\delta_i = 1$. In this case, we can consider two types of propensity weights:

1. Obtain $\hat{\theta}_1$ by solving

$$\hat{U}_1(\theta) \equiv \sum_{i=1}^{n} \frac{\delta_i}{\pi(y_i, z_i)} U(\theta; y_i) = 0.$$

2. Obtain $\hat{\theta}_2$ by solving

$$\hat{U}_2(\theta) \equiv \sum_{i=1}^{n} \frac{\delta_i}{\tilde{\pi}(y_i)} U(\theta; y_i) = 0,$$

where $\tilde{\pi}(y) = E\{\pi(y, z) \mid y\}$.

Remark (Cont'd)

In this case, we can prove that

$$E(\hat{\theta}_1) = E(\hat{\theta}_2) \tag{5}$$

and

$$V(\hat{\theta}_1) \geq V(\hat{\theta}_2). \tag{6}$$

§5.2 Regression weighting method

Motivation

- $\mathbf{x}_i$: auxiliary variables (observed throughout the sample)
- Assume that $1 = \mathbf{x}_i'\mathbf{a}$ for some $\mathbf{a}$.
- y_i: study variable (observed only when $\delta_i = 1$).
- Regression weighting technique: Use

$$w_i = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\right)' \left(\sum_{i=1}^{n}\delta_i\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\mathbf{x}_i$$

for the weight associated with unit i with $\delta_i = 1$.

Motivation

- Note that the regression estimator $\hat{\theta}_{reg} = \sum_{i=1}^{n} \delta_i w_i y_i$ of $\theta = E(Y)$ can be written as

$$\hat{\theta}_{reg} = \bar{\mathbf{x}}_n' \hat{\boldsymbol{\beta}}_r \tag{7}$$

  where

$$\hat{\beta}_r = \left(\sum_{i=1}^{n} \delta_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{n} \delta_i \mathbf{x}_i y_i .$$

- Under what conditions, the regression weighting method is justified (in that the resulting estimator is asymptotically unbiased under the response model) ?

Main Result (Fuller et al, 1994)

Assume that auxiliary variables $\mathbf{x}_i$ are observed throughout the sample and the response probability satisfies

$$\frac{1}{\pi_i} = \mathbf{x}_i'\boldsymbol{\lambda} \tag{8}$$

for all unit i in the sample, where $\boldsymbol{\lambda}$ is unknown. We assume that an intercept is included in xi. Under these conditions, the regression estimator defined by (23) is asymptotically unbiased for $\theta = E(Y)$.

Justification

- Because an intercept term is included in $\mathbf{x}_i$, we have

$$\hat{\theta}_n \equiv \bar{y}_n = \bar{\mathbf{x}}_n' \hat{\boldsymbol{\beta}}_n$$

where

$$\hat{\beta}_n = \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i.$$

- Note that we can write

$$\hat{\theta}_{reg} - \hat{\theta}_n = \bar{\mathbf{x}}_n' \left(\sum_{i=1}^{n} \delta_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{n} \delta_i \mathbf{x}_i \left(y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_n \right)$$

and so

$$E\left(\hat{\theta}_{reg} - \hat{\theta}_n \mid \mathbf{X}, \mathbf{Y} \right) \cong \bar{\mathbf{x}}_n' \left(\sum_{i=1}^{n} \pi_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{n} \pi_i \mathbf{x}_i \left(y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_n \right)$$

where the expectation is taken with respect to the response mechanism.

Justification (Cont'd)

- Thus, to show that $\hat{\theta}_{reg}$ is asymptotically unbiased, we have only to show that

$$\sum_{i=1}^{n} \pi_i \mathbf{x}_i \left(y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_n \right) = 0 \tag{9}$$

  holds.

- By (8), we have

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_n \right) \\ &= \sum_{i=1}^{n} \pi_i \left(\boldsymbol{\lambda}' \mathbf{x}_i \right) \left(y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_n \right), \end{aligned}$$

  which implies that (9) holds.

Variance estimation of the regression estimator

- To discuss variance estimation of the regression estimator where the covariates $\mathbf{x}_i$ satisfy (8), note that

$$\begin{aligned}
\bar{\mathbf{x}}_n'\hat{\boldsymbol{\beta}}_r &= \bar{\mathbf{x}}_n'\boldsymbol{\beta} + \bar{\mathbf{x}}_n'\left(\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta}\right) \\
&= \bar{\mathbf{x}}_n'\boldsymbol{\beta} + \bar{\mathbf{x}}_n'\left(\sum_{i=1}^{n}\delta_i\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{n}\delta_i\mathbf{x}_i\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right) \\
&\cong \bar{\mathbf{x}}_n'\boldsymbol{\beta} + \bar{\mathbf{x}}_n'\left(\sum_{i=1}^{n}\pi_i\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{n}\delta_i\mathbf{x}_i\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)
\end{aligned}$$

  where $\boldsymbol{\beta}$ is the probability limit of $\hat{\boldsymbol{\beta}}_r$
- By the fact that 1 is included in $\mathbf{x}_i$ and by (8), it can be shown that

$$\bar{\mathbf{x}}_n'\left(\sum_{i=1}^{n}\pi_i\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{n}\delta_i\mathbf{x}_i\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right) \quad (10)$$

  by some matrix algebra.

Variance estimation of the regression estimator

- Approximate variance

$$V\left(\hat{\theta}_{reg}\right) \cong V\left(\frac{1}{n}\sum_{i=1}^{n} d_i\right) \tag{11}$$

where $d_i = \mathbf{x}_i'\boldsymbol{\beta} + \delta_i \pi_i^{-1}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)$.

- Variance estimation can be implemented by using a standard variance estimation formula applied to $\hat{d}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_r + \delta_i n w_i (y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_r)$. That is,

$$\hat{V}\left(\hat{\theta}_{reg}\right) = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{d}_i - \bar{\hat{d}}_n\right)^2$$

where $\bar{\hat{d}}_n = \sum_{i=1}^{n} \hat{d}_i / n$.

§5.3 Propensity score method

Motivation

- $z_i = (x_i, y_i)$, y_i is subject to missingness
- Interested in estimating θ which is defined by $E\{U(\theta; Z)\} = 0$.
- 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, denoted by $\hat{\theta}_{PS}$, is computed by solving

$$\hat{U}_{PS}(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} \delta_i \frac{1}{\hat{\pi}_i} U(\theta; z_i) = 0, \tag{12}$$

  where $\hat{\pi}_i = \pi(z_i; \hat{\phi})$ and $\hat{\phi}$ is the MLE of ϕ_0.
- What is the asymptotic properties of $\hat{\theta}_{PS}$?

Idea

- The PS estimator $\hat{\theta}_{PS}$ is a function of $\hat{\phi}$.
- Note that $(\hat{\theta}_{PS}, \hat{\phi})$ is the solution to

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

  where $S(\phi)$ is the score function for ϕ.
- Thus, we can apply the sandwitch formula to obtain the asymptotic variance of $(\hat{\theta}_{PS}, \hat{\phi})$.
- When $\hat{\phi}$ is the MLE, then we may use Bartlett identity. (see Lemma 5.1)

Lemma 5.1

Lemma

Let

$$U_1(\theta, \phi) = \sum_{i=1}^{n} u_{i1}(\theta, \phi),$$

where $u_{i1}(\theta, \phi) = u_{i1}(\theta, \phi; z_i, \delta_i)$, *be an estimating equation satisfying*

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

Let $\pi_i = \pi_i(\phi)$ *be the probability of response. Then,*

$$E\{-\partial U_1/\partial\phi\} = Cov(U_1, S) \tag{13}$$

where S *is the score function of* ϕ.

Note: If we set $U_1(\theta, \phi) = S(\phi)$, then (13) reduces to $E\{-\partial S(\phi)/\partial\phi\} = E\{S(\phi)^{\otimes 2}\}$, which is already presented in Chapter 2 (Theorem 2.3).

Proof.

Since $E\{U_1(\theta_0, \phi_0)\} = 0$, we have

$$\begin{aligned}
0 &= \partial E\{U_1(\theta_0, \phi_0)\}/\partial\phi \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial\phi} \int u_{i1}(\theta_0, \phi_0) f(\delta_i \mid z_i, \phi_0) f(z_i) \, d\delta_i dz_i \\
&= \sum_{i=1}^{n} \int \left[\frac{\partial}{\partial\phi} u_{i1}(\theta_0, \phi_0)\right] f(\delta_i \mid z_i, \phi_0) f(z_i) \, d\delta_i dz_i \\
&+ \sum_{i=1}^{n} \int u_{i1}(\theta_0, \phi_0) \frac{\partial}{\partial\phi} [f(\delta_i \mid z_i, \phi_0)] f(z_i) \, d\delta_i dz_i \\
&= E\{\partial U/\partial\phi\} + E\{U(\theta_0, \phi_0) S(\phi_0)\}
\end{aligned}$$

which proves (13). □

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

$$A = \begin{bmatrix} E\left\{-\partial \hat{U}_1/\partial\theta\right\} & E\left\{-\partial U_1/\partial\phi\right\} \\ E\left\{-\partial S/\partial\theta\right\} & E\left\{-\partial S/\partial\phi\right\} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$$

$$B = \begin{bmatrix} V\left(\hat{U}_1\right) & C\left(\hat{U}_1, S\right) \\ C\left(S, \hat{U}_1\right) & V(S) \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}.$$

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 5.1, $B_{22} = A_{22}$ and $B_{12} = A_{12}$. Thus,

$$V(\hat{\theta}_{PS}) \cong A_{11}^{-1}\left[B_{11} - B_{12}B_{22}^{-1}B_{21}\right]A_{11}^{'-1}. \tag{14}$$

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

$$V\left(\hat{\theta}_W\right) \geq V\left(\hat{\theta}_{PS}\right). \tag{15}$$

- The result of (15) means that the PS estimator with estimated π_i is more efficient than the PS estimator with known π_i.

Remark

- 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 (17) as

$$\begin{aligned} \hat{\theta}_{PS} &\cong \hat{\theta}_W - C(\hat{\theta}_W, \hat{\phi}) \left\{ V(\hat{\phi}) \right\}^{-1} \left(\hat{\phi} - \phi_0 \right) \\ &\cong \hat{\theta}_W - C(\hat{\theta}_W, S) \{V(S)\}^{-1} S(\phi_0) \end{aligned}$$

which can be understood as a special case of Taylor linearization

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

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

Remark

- Writing $\hat{\theta}_{PS} = \hat{\theta}_W(\hat{\phi})$, another way of understanding (14) is

$$V\left(\hat{\theta}_{PS}\right) \cong E\left\{V\left(\hat{\theta}_W \mid S^{\perp}\right)\right\}, \tag{16}$$

where

$$V\left(Y \mid X^{\perp}\right) = V(Y) - C(Y,X)\{V(X)\}^{-1} C(X,Y)$$

and $S = S(\phi)$ is the score function of ϕ.

- Thus, the PS estimator $\hat{\theta}_{PS}$ with $\hat{\pi}_i = \pi_i(\hat{\phi})$ with $\hat{\phi}$ from the maximum likelihood method can be viewed as a projection of $\hat{\theta}_W$ to the orthogonal complement of the space generated by $S(\phi)$. That is, we can express

$$\hat{\theta}_{PS} \cong E\{\hat{\theta}_W \mid S^{\perp}\} \equiv \hat{\theta}_W - C(\hat{\theta}_W, S)\{V(S)\}^{-1} S(\phi_0). \tag{17}$$

Variance estimation of the PS estimator

- If we assume that the response mechanism is MAR and follows the following parametric model

$$\pi_i = \pi\left(x_i; \phi_0\right) \tag{18}$$

for some $\phi_0 \in \Omega$, where x_i are completely observed in the sample. In this case, the propensity score can be estimated by the maximum likelihood method that solves

$$S(\phi) \equiv \sum_{i=1}^{n} \left\{\delta_i - \pi\left(x_i; \phi\right)\right\} \frac{1}{\pi\left(x_i; \phi\right)\left\{1 - \pi\left(x_i; \phi\right)\right\}} \dot{\pi}\left(x_i; \phi\right) = 0, \tag{19}$$

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

Variance estimation of the PS estimator (Cont'd)

- Using (14), a plug-in variance estimator of the PS estimator is computed by

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

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

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

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

Improving the efficiency of PS estimator

- We want to improve the efficiency of the PS estimator in (12) by incorporating the auxiliary variable x_i observed throughout the sample.
- One can consider a class of estimating equations of the form

$$\sum_{i=1}^{n} \delta_i \frac{1}{\hat{\pi}_i} \{U(\theta; x_i, y_i) - b(\theta; x_i)\} + \sum_{i=1}^{n} b(\theta; x_i) = 0 \qquad (20)$$

  where $b(\theta; x_i)$ is to be determined.
- We can write the solution to (20) as $\hat{\theta}_b$ as it depends on the particular choice of $b(\theta; x)$ function.
- Note that the solution $\hat{\theta}_b$ is consistent regardless of the choice of $b(\theta; x_i)$.
- We want to find an optimal choice $b^*(\theta; x_i)$ which minimizes the variance of $\hat{\theta}_b$.

Theorem 5.1 (Robins et al., 1994)

Theorem

Assume that the probability $Pr(\delta = 1 \mid x, y) = \pi(x)$ *does not depend on the value of* y. *Let* $\hat{\theta}_b$ *be the solution to (20) for given* $b(\theta; x_i)$. *Under some regularity conditions,* $\hat{\theta}_b$ *is consistent and its asymptotic variance satisfies*

$$V\left(\hat{\theta}_b\right) \geq n^{-1}\tau^{-1}\left[V\left\{E(U \mid X)\right\} + E\left\{\frac{1}{\pi(x)}V(U \mid X)\right\}\right](\tau^{-1})', \quad (21)$$

where $\tau = E(\partial U/\partial\theta')$ *and the equality holds when* $b^*(\theta; x_i) = E\left\{U(\theta; x_i, y_i) \mid x_i\right\}$.

Example 5.4

- Consider the sample from a linear regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + e_i \tag{22}$$

where e_i are independent with $E(e_i \mid \mathbf{x}_i) = 0$. Assume that $\mathbf{x}_i$ are available from the full sample and y_i are observed only when $\delta_i = 1$. The response propensity model follows from the logistic regression model with $\text{logit}(\pi_i) = \mathbf{x}_i'\phi$. We are interested in estimating $\theta = E(Y)$ from the partially observed data.

- To construct the optimal estimator that achieves the minimum variance in (21), we can use $U_i(\theta) = y_i - \theta$ and $b_i^*(\theta) = \mathbf{x}_i'\boldsymbol{\beta} - \theta$. Thus, the optimal estimator using $\hat{b}_i^*(\theta) = \mathbf{x}_i'\hat{\boldsymbol{\beta}} - \theta$ in (20) is given by

$$\hat{\theta}_{opt}(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}y_i + \frac{1}{n}\left(\sum_{i=1}^{n}\mathbf{x}_i - \sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}\mathbf{x}_i\right)'\hat{\boldsymbol{\beta}} \tag{23}$$

where $\hat{\boldsymbol{\beta}}$ is any estimator of $\boldsymbol{\beta}$ satisfying $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O_p(1)$, where $X_n = O_p(1)$ denotes that X_n is bounded in probability.

Example 5.4 (Cont'd)

- Note that the choice of $\hat{\boldsymbol{\beta}}$ does not play any leading role in the asymptotic variance of $\hat{\theta}_{opt}(\hat{\boldsymbol{\beta}})$. This is because

$$\hat{\theta}_{opt}(\hat{\boldsymbol{\beta}}) \cong \hat{\theta}_{opt}(\boldsymbol{\beta}_0) + E\left\{\frac{\partial}{\partial \boldsymbol{\beta}}\hat{\theta}_{opt}(\boldsymbol{\beta}_0)\right\}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \tag{24}$$

and, under the correct response model,

$$E\left\{\frac{\partial}{\partial \boldsymbol{\beta}}\hat{\theta}_{opt}(\boldsymbol{\beta}_0)\right\} = E\left\{\frac{1}{n}\left(\sum_{i=1}^{n}\mathbf{x}_i - \sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}\mathbf{x}_i\right)\right\} \cong \mathbf{0}$$

and so the second term of (24) becomes negligible. Furthermore, it can be shown that the choice of $\hat{\phi}$ in $\hat{\pi}_i = \pi_i(\hat{\phi})$ does not matter as long as the regression model holds.

§5.4 Optimal estimation

Motivation

- In Example 5.4, optimal estimator using auxiliary information is considered, under the (outcome) regression model (22).
- We now want to find the optimal estimator (using auxiliary information) without relying on the outcome regression model.
- Note that $\hat{\theta}_{PSA} = n^{-1} \sum_{i=1}^{n} \delta_i y_i / \hat{\pi}_i$ applied to $y_i = x_i$ does not necessarily lead to $\bar{x}_n = n^{-1} \sum_{i=1}^{n} x_i$.
- That is, we have two estimators of $E(X)$, $\hat{\bar{x}}_{PSA}$ and $\bar{x}_n$.
- How to incorporate the extra information without relying on the regression model ?
- More generally, suppose that we have over-identified parameters (i.e. number of estimating equations $>$ number of parameters). How to obtain a best estimator ?

GMM (Generalized method of moment) estimation

- θ: p-dimensional parameter
- $U(\theta; Z) = 0$: a system of estimating equations of size $m > p$.
- No unique solution exists.
- Let $W(\theta)$ be a $m \times m$ symmetric matrix. Define

$$Q_W(\theta) = \{U(\theta; Z)\}' W(\theta) U(\theta; Z).$$

Note that $\hat{\theta}_W = \arg\min Q_W(\theta)$ is now uniquely determined under some regularity conditions. Note that the solution $\hat{\theta}_W$ is obtained by solving

$$U_W(\theta) \equiv \{\dot{U}(\theta; Z)\}' W(\theta) U(\theta; Z) = 0$$

where $\dot{U}(\theta; z) = \partial U(\theta; z)/\partial \theta'$.

GMM estimation (Cont'd)

- Thus, the asymptotic variance of $\hat{\theta}_W$ is

$$\begin{aligned} V\left(\hat{\theta}_W\right) &\cong \left\{E\left(\frac{\partial}{\partial\theta'}U_W(\theta)\right)\right\}^{-1} V\{U_W(\theta)\}\left\{E\left(\frac{\partial}{\partial\theta'}U_W(\theta)\right)'\right\}^{-1} \\ &= \left\{\tau' W\tau\right\}^{-1}\tau' W V(U) W\tau\left\{\tau' W\tau\right\}^{-1}, \end{aligned}$$

  where $W = W(\theta)$ and $\tau = E(\partial U/\partial\theta')$. The asymptotic variance is minimized at

$$W^* = \{V(U)\}^{-1}.$$

- Thus, the GMM estimator of θ is obtained by minimizing

$$Q^*(\theta) = \{U(\theta; Z)\}'\{Var(U(\theta; Z))\}^{-1} U(\theta; Z).$$

  The asymptotic variance of the GMM estimator is $\left[\tau'\{V(U)\}^{-1}\tau\right]^{-1}$.

Lemma 5.2

Lemma

Assume that $\hat{X}_1$ and $\hat{X}_2$ are two unbiased estimators of μ_x and $\hat{Y}$ is an unbiased estimator of μ_y. Let

$$Q = \begin{pmatrix} \hat{X}_1 - \mu_x \\ \hat{X}_2 - \mu_x \\ \hat{Y} - \mu_y \end{pmatrix}' \begin{pmatrix} V(\hat{X}_1) & C(\hat{X}_1, \hat{X}_2) & C(\hat{X}_1, \hat{Y}) \\ C(\hat{X}_1, \hat{X}_2) & V(\hat{X}_2) & C(\hat{X}_2, \hat{Y}) \\ C(\hat{X}_1, \hat{Y}) & C(\hat{X}_2, \hat{Y}) & V(\hat{Y}) \end{pmatrix}^{-1} \begin{pmatrix} \hat{X}_1 - \mu_x \\ \hat{X}_2 - \mu_x \\ \hat{Y} - \mu_y \end{pmatrix} \tag{25}$$

The optimal estimator of (μ_x, μ_y) that minimizes Q in (25) is

$$\hat{\mu}_x^* = \alpha^* \hat{X}_1 + (1 - \alpha^*) \hat{X}_2 \tag{26}$$

and

$$\hat{\mu}_y^* = \hat{Y} + B_1 \left(\hat{\mu}_x^* - \hat{X}_1 \right) + B_2 \left(\hat{\mu}_x^* - \hat{X}_2 \right) \tag{27}$$

Lemma (Cont'd)

where

$$\alpha^* = \frac{V(\hat{X}_2) - C(\hat{X}_1, \hat{X}_2)}{V(\hat{X}_1) + V(\hat{X}_2) - 2C(\hat{X}_1, \hat{X}_2)}$$

and

$$\begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \begin{pmatrix} V(\hat{X}_1) & C(\hat{X}_1, \hat{X}_2) \\ C(\hat{X}_1, \hat{X}_2) & V(\hat{X}_2) \end{pmatrix}^{-1} \begin{pmatrix} C(\hat{X}_1, \hat{Y}) \\ C(\hat{X}_2, \hat{Y}) \end{pmatrix}.$$

Proof.

Using the inverse of the partitioned matrix, we can write

$$Q = Q_1 + Q_2$$

where

$$Q_1 = \begin{pmatrix} \hat{X}_1 - \mu_x \\ \hat{X}_2 - \mu_x \end{pmatrix}' \begin{pmatrix} V(\hat{X}_1) & C(\hat{X}_1, \hat{X}_2) \\ C(\hat{X}_1, \hat{X}_2) & V(\hat{X}_2) \end{pmatrix}^{-1} \begin{pmatrix} \hat{X}_1 - \mu_x \\ \hat{X}_2 - \mu_x \end{pmatrix},$$

$$Q_2 = \left\{ \hat{Y} - E(\hat{Y} \mid \hat{X}_1, \hat{X}_2) \right\}' V_{ee}^{-1} \left\{ \hat{Y} - E(\hat{Y} \mid \hat{X}_1, \hat{X}_2) \right\},$$

$$E(\hat{Y} \mid \hat{X}_1, \hat{X}_2) = \mu_y + B_1(\hat{X}_1 - \mu_x) + B_2(\hat{X}_2 - \mu_x),$$

and $V_{ee} = V(\hat{Y}) - (B_1, B_2)\{V(\hat{X}_1, \hat{X}_2)\}^{-1}(B_1, B_2)'$.
Minimizing Q_1 with respect to μ_x gives $\hat{\mu}_x^*$ in (26) and minimizing Q_2 with respect to μ_y for given $\hat{\mu}_x^*$ gives $\hat{\mu}_y^*$ in (27). □

Remark

- The optimal estimator of μ_y takes the form of the regression estimator with $\hat{\mu}_x^*$ as the control.
- Using (26), we can also express

$$\hat{\mu}_y^* = \hat{Y} - C\left(\hat{Y}, \hat{X}_2 - \hat{X}_1\right)\left\{V\left(\hat{X}_2 - \hat{X}_1\right)\right\}^{-1}(\hat{X}_2 - \hat{X}_1).$$

Remark

- Under the missing data setup where $\mathbf{x}_i$ is always observed and y_i is subject to missingness, if we know π_i, then we can use $\hat{X}_1 = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i = \hat{X}_n$, $\hat{X}_2 = n^{-1}\sum_{i=1}^{n}\delta_i\mathbf{x}_i/\pi_i = \hat{X}_W$, and $\hat{Y} = n^{-1}\sum_{i=1}^{n}\delta_i y_i/\pi_i = \hat{Y}_W$.
- In this case, we can obtain $\hat{\mu}_x^* = \bar{X}_1$ and the optimal estimator of μ_y reduces to

$$\begin{aligned}\hat{\mu}_y^* &= \hat{Y} + C\left(\hat{Y}, \hat{X}_2 - \hat{X}_1\right)\left\{V\left(\hat{X}_2 - \hat{X}_1\right)\right\}^{-1}(\hat{X}_1 - \hat{X}_2) \\ &= \hat{Y}_W + (\hat{X}_n - \hat{X}_W)'B^*\end{aligned}$$

where

$$B^* = E\left(\sum_{i=1}^{n}\frac{1-\pi_i}{\pi_i}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}E\left(\sum_{i=1}^{n}\frac{1-\pi_i}{\pi_i}\mathbf{x}_i y_i\right).$$

Optimal PS estimation

GMM approach

- Let $\theta = (\mu_x, \mu_y)$. We have three estimators for two parameters.
- Find θ that minimizes

$$Q_{PS}(\theta) = \begin{pmatrix} \bar{x}_n - \mu_x \\ \hat{\theta}_{x,PS} - \mu_x \\ \hat{\theta}_{y,PS} - \mu_y \end{pmatrix}' \left\{ \hat{V} \begin{pmatrix} \bar{x}_n \\ \hat{\theta}_{x,PS} \\ \hat{\theta}_{y,PS} \end{pmatrix} \right\}^{-1} \begin{pmatrix} \bar{x}_n - \mu_x \\ \hat{\theta}_{x,PS} - \mu_x \\ \hat{\theta}_{y,PS} - \mu_y \end{pmatrix} \tag{28}$$

  where $\hat{\theta}_{PS} = \hat{\theta}_{PS}(\hat{\phi})$ is the propensity score estimator using $\hat{\pi}_i$.
- Computation for $\hat{V}$ is somewhat cumbersome.

Optimal PS estimation (Cont'd)

Alternative GLS (or GMM) approach

- Find (θ, ϕ) that minimizes

$$\begin{pmatrix} \bar{x}_n - \mu_x \\ \hat{\theta}_{x,PS}(\phi) - \mu_x \\ \hat{\theta}_{y,PS}(\phi) - \mu_y \\ S(\phi) \end{pmatrix}' \left\{ \hat{V} \begin{pmatrix} \bar{x}_n \\ \hat{\theta}_{x,PS}(\phi) \\ \hat{\theta}_{y,PS}(\phi) \\ S(\phi) \end{pmatrix} \right\}^{-1} \begin{pmatrix} \bar{x}_n - \mu_x \\ \hat{\theta}_{x,PS}(\phi) - \mu_x \\ \hat{\theta}_{y,PS}(\phi) - \mu_y \\ S(\phi) \end{pmatrix}.$$

- Computation for $\hat{V}$ is easier since we can treat ϕ as if known.
- Let $Q^*(\theta, \phi)$ be the above objective function. It can be shown that $Q^*(\theta, \hat{\phi}) = Q_{PS}(\theta)$ in (28) and so minimizing $Q^*(\theta, \hat{\phi})$ is equivalent to minimizing $Q_{PS}(\theta)$.

Optimal PS estimation (Cont'd)

Justification for the equivalence

- May write

$$\begin{aligned} Q^*(\theta,\phi) &= \begin{pmatrix} \hat{U}_{PS}(\theta,\phi) \\ S(\phi) \end{pmatrix}' \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}^{-1} \begin{pmatrix} \hat{U}_{PS}(\theta,\phi) \\ S(\phi) \end{pmatrix} \\ &= Q_1(\theta \mid \phi) + Q_2(\phi) \end{aligned}$$

  where

$$\begin{aligned} Q_1(\theta \mid \phi) &= \left(\hat{U}_{PS} - V_{12}V_{22}^{-1}S\right)' \left\{V\left(U_{PS} \mid S^{\perp}\right)\right\}^{-1} \left(\hat{U}_{PS} - V_{12}V_{22}^{-1}S\right) \\ Q_2(\phi) &= S(\phi)' \left\{\hat{V}(S)\right\}^{-1} S(\phi) \end{aligned}$$

- For the MLE $\hat{\phi}$, we have $Q_2(\hat{\phi}) = 0$ and $Q_1(\theta \mid \hat{\phi}) = Q_{PS}(\theta)$.

Example 5.5

- Response model

$$\pi_i(\phi^*) = \frac{\exp(\phi_0^* + \phi_1^* x_i)}{1 + \exp(\phi_0^* + \phi_1^* x_i)}$$

- Three direct PS estimators of $(1, \mu_x, \mu_y)$:

$$(\hat{\theta}_{1,PS}, \hat{\theta}_{x,PS}, \hat{\theta}_{y,PS}) = n^{-1} \sum_{i=1}^{n} \delta_i \hat{\pi}_i^{-1} (1, x_i, y_i) .$$

- $\bar{x}_n = n^{-1} \sum_{i=1}^{n} x_i$ available.
- What is the optimal estimator of μ_y ?

Example 5.5 (Cont'd)

- Minimize

$$\begin{pmatrix} \bar{x}_n - \mu_x \\ \hat{\theta}_{1,PS}(\phi) - 1 \\ \hat{\theta}_{x,PS}(\phi) - \mu_x \\ \hat{\theta}_{y,PS}(\phi) - \mu_y \\ S(\phi) \end{pmatrix}' \left\{ \hat{V} \begin{pmatrix} \bar{x}_n \\ \hat{\theta}_{1,PS}(\phi) \\ \hat{\theta}_{x,PS}(\phi) \\ \hat{\theta}_{y,PS}(\phi) \\ S(\phi) \end{pmatrix} \right\}^{-1} \begin{pmatrix} \bar{x}_n - \mu_x \\ \hat{\theta}_{1,PS}(\phi) - 1 \\ \hat{\theta}_{x,PS}(\phi) - \mu_x \\ \hat{\theta}_{y,PS}(\phi) - \mu_y \\ S(\phi) \end{pmatrix}$$

with respect to (μ_x, μ_y, ϕ), where

$$S(\phi) = \sum_{i=1}^{n} \left(\frac{\delta_i}{\pi_i(\phi)} - 1 \right) \mathbf{h}_i(\phi) = 0$$

with $\mathbf{h}_i(\phi) = \pi_i(\phi)(1, x_i)'$.

Example 5.5 (Cont'd)

- Equivalently, minimize

$$\begin{pmatrix} \hat{\theta}_{y,PS}(\phi) - \mu_y \\ \hat{\theta}_{1,PS}(\phi) - 1 \\ \hat{\theta}_{x,PS}(\phi) - \bar{x}_n \\ S(\phi) \end{pmatrix}' \left\{ \hat{V} \begin{pmatrix} \hat{\theta}_{y,PS}(\phi) \\ \hat{\theta}_{1,PS}(\phi) \\ \hat{\theta}_{x,PS}(\phi) - \bar{x}_n \\ S(\phi) \end{pmatrix} \right\}^{-1} \begin{pmatrix} \hat{\theta}_{y,PS}(\phi) - \mu_y \\ \hat{\theta}_{1,PS}(\phi) - 1 \\ \hat{\theta}_{x,PS}(\phi) - \bar{x}_n \\ S(\phi) \end{pmatrix}$$

with respect to (μ_y, ϕ), since the optimal estimator of θ_x is $\bar{x}_n$.

Example 5.5 (Cont'd)

- The solution can be written as

$$\hat{\mu}_{y,opt} = \hat{\theta}_{y,PS} + \left(1 - \hat{\theta}_{1,PS}\right)\hat{B}_0 + \left(\bar{x}_n - \hat{\theta}_{1,PS}\right)\hat{B}_1 + \left\{0 - S(\hat{\phi})\right\}\hat{C}$$

  where

$$\begin{pmatrix} \hat{B}_0 \\ \hat{B}_1 \\ \hat{C} \end{pmatrix} = \left\{\sum_{i=1}^{n} \delta_i b_i \begin{pmatrix} 1 \\ x_i \\ \mathbf{h}_i \end{pmatrix}\begin{pmatrix} 1 \\ x_i \\ \mathbf{h}_i \end{pmatrix}'\right\}^{-1} \sum_{i=1}^{n} \delta_i b_i \begin{pmatrix} 1 \\ x_i \\ \mathbf{h}_i \end{pmatrix} y_i$$

  and $b_i = \hat{\pi}_i^{-2}(1 - \hat{\pi}_i)$.

- Note that the last term $\{0 - S(\hat{\phi})\}\hat{C}$, which is equal to zero, does not contribute to the point estimation. But, it is used for variance estimation.

Example 5.5 (Cont'd)

- That is, for variance estimation, we simply express

$$\hat{\mu}_{y,opt} = n^{-1} \sum_{i=1}^{n} \hat{\eta}_i$$

where

$$\hat{\eta}_i = \hat{B}_0 + x_i \hat{B}_1 + \mathbf{h}_i' \hat{C} + \frac{\delta_i}{\hat{\pi}_i} \left(y_i - \hat{B}_0 - x_i \hat{B}_1 - \mathbf{h}_i' \hat{C} \right)$$

and apply the standard variance formula to $\hat{\eta}_i$.

Example 5.5 (Cont'd)

- The optimal estimator is linear in y. That is, we can write

$$\hat{\mu}_{y,opt} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}g_i y_i = \sum_{\delta_i=1} w_i y_i$$

  where g_i satisfies

$$\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}g_i(1, x_i, \mathbf{h}_i') = \sum_{i=1}^{n}(1, x_i, \mathbf{h}_i').$$

- Thus, it is doubly robust under the outcome model $E(y \mid x) = \beta_0 + \beta_1 x$ in the sense that $\hat{\mu}_{y,opt}$ is unbiased when either the response model or the outcome model holds.

§5.5 Doubly robust method

5. Doubly robust method

- Two models
  - Response Probability (RP) model: model about δ

$$Pr(\delta = 1 \mid \mathbf{x}, y) = \pi(\mathbf{x}; \phi)$$

  - Outcome Regression (OR) model: model about y

$$E(y \mid \mathbf{x}) = m(\mathbf{x}_i; \beta)$$

- Doubly robust (DR) estimation aims to achieve (asymptotic) unbiasedness under either RP model or OR model.
- For estimation of $\theta = E(Y)$, a doubly robust estimator is

$$\hat{\theta}_{DR} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{y}_i + \frac{\delta_i}{\hat{\pi}_i} (y_i - \hat{y}_i) \right\}$$

where $\hat{y}_i = m(\mathbf{x}_i; \hat{\beta})$ and $\hat{\pi}_i = \pi(\mathbf{x}_i; \hat{\phi})$.

5. Doubly robust method

- Note that

$$\hat{\theta}_{DR} - \hat{\theta}_n = n^{-1} \sum_{i=1}^{n} \left(\frac{\delta_i}{\hat{\pi}_i} - 1 \right) (y_i - \hat{y}_i) \,. \tag{29}$$

  Taking an expectation of the above, we note that the first term has approximate zero expectation if the RP model is true. The second term has approximate zero expectation if the OR model is true. Thus, $\hat{\theta}_{DR}$ is approximately unbiased when either RP model or OR model is true.

- When both models are true, then the choice of $\hat{\beta}$ and $\hat{\phi}$ does not make any difference in the asymptotic sense. Robins et al (1994) called the property local efficiency of the DR estimator.

5. Doubly robust method

- Kim and Riddles (2012) considered an augmented propensity model of the form

$$\hat{\pi}_i^* = \pi_i^*(\hat{\phi}, \hat{\lambda}) = \frac{\pi_i(\hat{\phi})}{\pi_i(\hat{\phi}) + \{1 - \pi_i(\hat{\phi})\} \exp(\hat{\lambda}_0 + \hat{\lambda}_1 \hat{m}_i)}, \tag{30}$$

where $\pi_i(\hat{\phi})$ is the estimated response probability under the response probability model and $(\hat{\lambda}_0, \hat{\lambda}_1)$ satisfies

$$\sum_{i=1}^{n} \frac{\delta_i}{\pi_i^*(\hat{\phi}, \hat{\lambda})} (1, \hat{m}_i) = \sum_{i=1}^{n} (1, \hat{m}_i) \tag{31}$$

with $\hat{m}_i = m(x_i; \hat{\beta})$.

5. Doubly robust method

- The augmented PS estimator, defined by $\hat{\theta}^*_{PS} = n^{-1}\sum_{i=1}^{n}\delta_i y_i/\hat{\pi}^*_i$, based on the augmented propensity in (30) satisfies, under the assumed response probability model,

$$\hat{\theta}^*_{PS} \cong \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{b}_0 + \hat{b}_1\hat{m}_i + \frac{\delta_i}{\hat{\pi}_i}\left(y_i - \hat{b}_0 - \hat{b}_1\hat{m}_i\right)\right\}, \qquad (32)$$

where

$$\begin{pmatrix}\hat{b}_0\\ \hat{b}_1\end{pmatrix} = \left\{\sum_{i=1}^{n}\delta_i\left(\frac{1}{\hat{\pi}_i}-1\right)\begin{pmatrix}1\\ \hat{m}_i\end{pmatrix}\begin{pmatrix}1\\ \hat{m}_i\end{pmatrix}'\right\}^{-1}\sum_{i=1}^{n}\delta_i\left(\frac{1}{\hat{\pi}_i}-1\right)\begin{pmatrix}1\\ \end{pmatrix}$$

5. Doubly robust method

- The augmented PS estimator using

$$\hat{\pi}_i^* = \pi_i^*(\hat{\phi}, \hat{\lambda}) = \frac{\hat{\pi}_i}{\hat{\pi}_i + \{1 - \hat{\pi}_i\} \exp(\hat{\lambda}_0/\hat{\pi}_i + \hat{\lambda}_1 x_i/\hat{\pi}_i)},$$

with $(\hat{\lambda}_0, \hat{\lambda}_1)$ satisfying

$$\sum_{i=1}^{n} \frac{\delta_i}{\pi_i^*(\hat{\phi}, \hat{\lambda})} (1, x_i) = \sum_{i=1}^{n} (1, x_i)$$

is asymptotically equivalent to the optimal regression PS estimator discussed in Example 5.5.

6. Nonparametric method

Motivation

- So far, we have assumed a parametric model for $\pi(x) = Pr(\delta = 1 \mid x)$.
- Using the nonparametric regression technique, we can use a nonparametric estimator of $\pi(x)$ given by a nonparametric regression estimator of $\pi(x) = E(\delta \mid x)$ can be obtained by

$$\hat{\pi}_h(x) = \frac{\sum_{i=1}^n \delta_i K_h(x_i, x)}{\sum_{i=1}^n K_h(x_i, x)}, \tag{33}$$

  where K_h is the kernel function which satisfies certain regularity conditions and h is the bandwidth.
- Once a nonparametric estimator of $\pi(x)$ is obtained, the nonparametric PS estimator $\hat{\theta}_{NPS}$ of $\theta_0 = E(Y)$ is given by

$$\hat{\theta}_{NPS} = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_h(x_i)} y_i. \tag{34}$$

Theorem 5.2

Under some regularity conditions, we have

$$\hat{\theta}_{NPS} = \frac{1}{n}\sum_{i=1}^{n}\left[m(x_i) + \frac{\delta_i}{\pi(x_i)}\{y_i - m(x_i)\}\right] + o_p(n^{-1/2}), \tag{35}$$

where $m(x) = E(Y \mid x)$ and $\pi(x) = P(\delta = 1 \mid x)$. Furthermore, we have

$$\sqrt{n}\left(\hat{\theta}_{NPS} - \theta\right) \to N\left(0, \sigma_1^2\right),$$

where $\sigma_1^2 = V\{m(X)\} + E\left[\{\pi(X)\}^{-1} V(Y \mid X)\right]$.

Originally proved by Hirano et al. (2003).

6. Nonparametric method

Remark

- Unlike the usual asymptotic for nonparametric regression, $\sqrt{n}$-consistency is established.
- The nonparametric PS estimator achieves the lower bound of the variance that was discussed in Theorem 5.1.
- Instead of nonparametric PS method, we can use the same Kernel regression technique to obtain a nonparametric imputation estimator given by

$$\hat{\theta}_{NPI} = \frac{1}{n}\sum_{i=1}^{n}\{\delta_i y_i + (1-\delta_i)\hat{m}_h(x_i)\} \tag{36}$$

where

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n}\delta_i K_h(x_i, x) y_i}{\sum_{i=1}^{n}\delta_i K_h(x_i, x)}.$$

Cheng (1994) proves that $\hat{\theta}_{NPI}$ has the same asymptotic variance in Theorem 5.2.

7. Application to longitudinal missing

Basic Setup

- X_i is always observed and remains unchanged for $t = 0, 1, \ldots, T$.
- Y_{it} is the response for subject i at time t.
- δ_{it}: The response indicator for subject i at time t.
- Assuming no missing in the baseline year, Y_0 can be absorbed into X.
- Monotone missing pattern

$$\delta_{it} = 0 \Rightarrow \delta_{i,t+1} = 0, \forall t = 1, \ldots, T-1.$$

- $L_{i,t} = (X_i', Y_{i1}, \ldots, Y_{i,t})'$: Measurement up to t.
- Parameter of interest θ is estimated by solving

$$\sum_{i=1}^{n} U(\theta; L_{i,T}) = 0$$

for θ, under complete response.

Application to longitudinal missing

Missing mechanism (under monotone missing pattern)

- **Missing completely at random (MCAR)** :

$$P(\delta_{it=1}|\delta_{i,t-1}=1, L_{i,T}) = P(\delta_{it=1}|\delta_{i,t-1}=1).$$

- **Covariate-dependent missing (CDM)** :

$$P(\delta_{it}=1|\delta_{i,t-1}=1, L_{i,T}) = P(\delta_{it}=1|\delta_{i,t-1}=1, X_i).$$

- **Missing at random (MAR)** :

$$P(\delta_{it}=1|\delta_{i,t-1}=1, L_{i,T}) = P(\delta_{it}=1|\delta_{i,t-1}=1, L_{i,t-1}).$$

- **Missing not at random (MNAR)** : Missing at random does not hold.

Application to longitudinal missing

Motivation

- Panel attrition is frequently encountered in panel surveys, while classical methods often assume covariate-dependent missing, which can be unrealistic. We want to develop a PS method under MAR.
- Want to make full use of available information.

Application to longitudinal missing

Idea

- Under **MAR**, in the longitudinal data case, we would consider the conditional probabilities:

$$p_{it} := P(\delta_{it} = 1 | \delta_{i,t-1} = 1, L_{i,t-1}),\ t = 1, \ldots, T.$$

Then

$$\pi_{it} = \prod_{j=1}^{t} p_{ij}.$$

π_t then can be modeled through modeling p_t with $p_t(L_{t-1}; \phi_t)$.

- Once we obtain $\hat{\pi}_{iT} = \prod_{t=1}^{T} \hat{p}_{it}$ is obtained, we can use

$$\sum_{i=1}^{n} \frac{\delta_{iT}}{\hat{\pi}_{iT}} U(\theta; L_{i,T}) = 0$$

to obtain a consistent estimator of θ.

Application to longitudinal missing

Score Function for Longitudinal Data Under parametric models for p_t's, the partial likelihood for $\phi_1, \ldots, \phi_T$ is

$$L(\phi_1, \ldots, \phi_T) = \prod_{i=1}^{n} \prod_{t=1}^{T} \left[p_{it}^{\delta_{i,t}} (1 - p_{it})^{1-\delta_{i,t}} \right]^{\delta_{i,t-1}},$$

and the corresponding score function is $(S_1(\phi_1), \ldots, S_T(\phi_T))$, where

$$S_t(\phi_t) = \sum_{i=1}^{n} \delta_{i,t-1} \left\{ \delta_{it} - p_{it}(\phi_t) \right\} \mathbf{q}_{it}(\phi_t) = 0$$

where $\mathbf{q}_{it}(\phi_t) = \partial \text{logit}\{p_{it}(\phi_t)\}/\partial \phi_t$. Under logistic regression model such that $p_t = 1/\{1 + \exp(-\phi_t' L_{t-1})\}$, we have $\mathbf{q}_{it}(\phi_t) = L_{t-1}$.

Application to longitudinal missing

Remark

- Zhou and Kim (2012) proposed an optimal estimator of $\mu_t = E(Y_t)$ incorporating all available information.
- The idea can be extended to non-monotone missing data by re-defining

$$\pi_{it} = P\left(\delta_{i1} = \cdots = \delta_{it} = 1 \mid L_{it}\right) = \prod_{j=1}^{t} p_{ij}$$

where

$$p_{it} := P(\delta_{it} = 1 | \delta_{i1} = \cdots = \delta_{i,t-1} = 1, L_{i,t-1}).$$

- The score equation for ϕ_t in $p_{it} = p(L_{i,t-1}; \phi_t)$ is then

$$S_t(\phi_t) = \sum_{i=1}^{n} \delta^*_{i,t-1} \left\{\delta_{it} - p_{it}(\phi_t)\right\} \mathbf{q}_{it}(\phi_t) = 0$$

where $\delta^*_{i,t-1} = \prod_{j=1}^{t-1} \delta_{ij}$ and $\mathbf{q}_{it}(\phi_t) = \partial \text{logit}\{p_{it}(\phi_t)\}/\partial\phi_t$.

REFERENCES

Cheng, P. E. (1994), 'Nonparametric estimation of mean functionals with data missing at random', *Journal of the American Statistical Association* **89**, 81–87.

Hirano, K., G. Imbens and G. Ridder (2003), 'Efficient estimation of average treatment effects using the estimated propensity score', *Econometrica* **71**, 1161–1189.

Kim, J. K. and M. K. Riddles (2012), 'Some theory for propensity-score-adjustment estimators in survey sampling', *Survey Methodology* **38**, 157–165.

Robins, J. M., A. Rotnitzky and L. P. Zhao (1994), 'Estimation of regression coefficients when some regressors are not always observed', *Journal of the American Statistical Association* **89**, 846–866.

Zhou, M. and J. K. Kim (2012), 'An efficient method of estimation for longitudinal surveys with monotone missing data', *Biometrika* **99**, 631–648.