GENERALIZED METHOD OF MOMENTS FOR NONIGNORABLE MISSING DATA

Li Zhang[1], Cunjie Lin[2,3] and Yong Zhou[4,5]

[1] *School of Economics and Management, Northwest University, Xian, China*

[2] *Center for Applied Statistics, Renmin University of China, Beijing, China*

[3] *School of Statistics, Renmin University of China, Beijing, China*

[4] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China*

[5] *School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China*

Supplementary Material

In this supplementary material, we first give the main technically derivations for the theoretical results. And we include some numerical studies and discussions on the instrumental variables in the second part. More details about a data example are presented in the last part.

S1 Proof of Theorem 1

The proof is similar to that of Wang, Shao, and Kim (2014). For a given u, define $f_1(y) = f(y|u, z_1)$, $f_2(y) = f(y|u, z_2)$. And we will show that, for

all $y \in \mathcal{R}$, if

$$\begin{aligned} H(g(u)+\alpha y)f_1(y) &= H(g^{'}(u)+\alpha^{'}y)f_1^{'}(y), \\ H(g(u)+\alpha y)f_2(y) &= H(g^{'}(u)+\alpha^{'}y)f_2^{'}(y) \end{aligned} \tag{S1.1}$$

then, $g = g^{'}$, $\alpha = \alpha^{'}$, $f_1 = f_1^{'}$, and $f_2 = f_2^{'}$.

Since $f_i, f_i^{'}, i = 1, 2$ are density functions, (S1.1) implies

$$\int \left[\frac{H(g(u)+\alpha y)}{H(g^{'}(u)+\alpha^{'}y)} - 1\right] f_1(y)dy = \int \left[\frac{H(g(u)+\alpha y)}{H(g^{'}(u)+\alpha^{'}y)} - 1\right] f_2(y)dy = 0. \tag{S1.2}$$

We now show in two steps that $(S1.2)$ implies $\alpha = \alpha^{'}$.

First, we show that, when $\alpha \neq \alpha^{'}$, the function $K(y) = \frac{H(g(u)+\alpha y)}{H(g^{'}(u)+\alpha^{'}y)} - 1$ has a single change of sign. Under condition (C1), $H(\cdot)$ is strictly monotone and we consider a strictly increasing H here. The proof for a strictly decreasing is similar. Note that if one of α and $\alpha^{'}$ is 0 or if α and $\alpha^{'}$ have different signs, then $K(y)$ is a strictly monotone function having a unique root and, hence, it has a single change of sign. For the case where α and $\alpha^{'}$ have the same sign, we assume that $\alpha > \alpha^{'} > 0$. Let $y^* = \frac{g(u)-g^{'}(u)}{\alpha^{'}-\alpha}$, which implies $g(u) + \alpha y^* = g^{'}(u) + \alpha^{'}y^*$ and consequently $K(y^*) = 0$. For any $y > y^*$, it is easy to show that $g(u) + \alpha y < g^{'}(u) + \alpha^{'}y$. Since $H(\cdot)$ is strictly increasing, we get $H(g(u) + \alpha y) < H(g^{'}(u) + \alpha^{'}y)$. Therefore, $K(y) < 0$. Similarly, when $y < y^*$, $K(y) > 0$. This proves that $K(y)$ has a single change of sign.

Next, we prove that if $\alpha \neq \alpha^{'}$ and the first integral in (S1.2) is 0, then the second integral is not 0. Let X be a random variable having f_1 or f_2 as its probability density and E_j denote the expectation when X has density f_j. And we will prove that if $E_1[K(X)] = 0$, then $E_2[K(X)] \neq 0$.

Let t_0 be the change point of $K(\cdot)$, i.e., $K(t) < 0$ if $t < t_0$ and $K(t) > 0$ if $t > t_0$. Define $c = \sup_{t<t_0} f_2(t)/f_1(t)$. Under condition (C2), $f_2(t)/f_1(t)$ is a nondecreasing function of t. Hence, when $f_1(t_0) > 0$, $c = f_2(t_0)/f_1(t_0) < \infty$. When $f_1(t_0) = 0$, there exists t_1 such that $t_1 > t_0$ and $f_1(t_1) > 0$ under the assumption that $E_1[K(X)] = 0$. Hence, $c \leq f_2(t_1)/f_1(t_1) < \infty$. Thus, $c < \infty$ is always true. Let $A = \{t : f_1(t) = 0, f_2(t) > 0\}$ and $B = \{t : f_1(t) > 0, f_2(t) > 0\} \cup \{t : f_1(t) > 0, f_2(t) = 0\}$ and $E_2(K(X))$ can be written as $E_2(K(X)) = \int K(t) f_2(t) dt = \int_A K(t) f_2(t) ft + \int_B K(t) f_2(t) dt$. If $t \in A$, then $f_2(t)/f_1(t) = \infty$, and, therefore, $t > t_0$ and $K(t) > 0$ for $t \in A$, which yields that $\int_A K(t) f_2(t) \geq 0$. Then

$$
\begin{aligned}
E_2[K(X)] &\geq \int_B K(t) f_2(t) dt \\
&= \int_{B_1} K(t) f_2(t) dt + \int_{B_2} K(t) f_2(t) dt \\
&= \int_{B_1} K(t) \frac{f_2(t)}{f_1(t)} f_1(t) dt + \int_{B_2} K(t) \frac{f_2(t)}{f_1(t)} f_1(t) dt \\
&\geq \int_{B_1} cK(t) f_1(t) dt + \int_{B_2} cK(t) f_1(t) dt \\
&= cE_1[K(T)] = 0,
\end{aligned}
$$

where $B_1 = \{t : t \in B, t < t_0\}$, $B_2 = \{t : t \in B, t > t_0\}$. If A has a positive Lebesgue measure, $\int_A K(t)f_2(t)dt > 0$ and, hence, $E_2[K(X)] > 0$. If A has Lebesgue measure 0, the support sets of f_1 and f_2 are subsets of B. If $E_2[K(X)] = 0$, then $f_2(t) = cf_1(t)$ a.e. on B. But f_1 and f_2 are densities, thus $c = 1$, which is a contradiction to condition (C1). Therefore, we have $E_2[K(X)] > 0$.

Thus, (S1.2) implies that $\alpha = \alpha^{'}$ and (S1.2) reduces to

$$\int \left[\frac{H(g(u) + \alpha y)}{H(g^{'}(u) + \alpha y)} - 1\right] f_1(y)dy = \int \left[\frac{H(g(u) + \alpha y)}{H(g^{'}(u) + \alpha y)} - 1\right] f_2(y) = 0.$$

which implies $g(u) = g^{'}(u)$ since $H(\cdot)$ is strictly monotone function. Combine these results and (S1.1), we get $f_1 = f_1^{'}$ and $f_2 = f_2^{'}$. □

S2 Proof of Lemmas 1 and 2

By the definition of $\hat{\psi}(Y_i, X_i, \boldsymbol{\theta})$, we have the following decomposition:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\hat{\psi}(Y_i, X_i, \boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\left[\delta_i\psi(Y_i, X_i, \boldsymbol{\theta}) + (1-\delta_i)\hat{m}_0(X_i, \boldsymbol{\theta})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i[\psi(Y_i, X_i, \boldsymbol{\theta}) - m_1(X_i, \boldsymbol{\theta})]\right\} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i m_1(X_i, \boldsymbol{\theta}) + (1-\delta_i)m_0(X_i, \boldsymbol{\theta})\right\} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\left\{\hat{m}_0(X_i, \boldsymbol{\theta}) - m_0(X_i, \boldsymbol{\theta})\right\} \\
&:= I_1 + I_2 + I_3,
\end{aligned}$$

where $m_1(X_i, \boldsymbol{\theta}) = E[\psi(Y_i, X_i, \boldsymbol{\theta})|X_i, \delta_i = 1]$.

The first two terms I_1 and I_2 are sums of independent random variables. So we only need to deal with I_3.

Proof of Lemma 1. Note that

$$\begin{aligned} I_3 &= \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\Big\{\hat{m}_0(X_i,\boldsymbol{\theta}) - m_0(X_i,\boldsymbol{\theta})\Big\} \\ &= \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i) \times \frac{\sum_{j=1}^{n}\delta_j K_h(X_j, X_i)\exp(\gamma_0 Y_j)[\psi(Y_j, X_j, \boldsymbol{\theta}) - m_0(X_i, \boldsymbol{\theta})]}{\sum_{j=1}^{n}\delta_j K_h(X_j, X_i)\exp(\gamma_0 Y_j)}. \end{aligned}$$

For the denominator of I_3, we have the following representation for a fixed point x_0,

$$\begin{aligned} p\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\delta_i K_h(X_i, x_0)\exp(\gamma_0 Y_i) &= E\left\{\delta K_h(X, x_0)\exp(\gamma_0 Y)\right\} \\ &= E\left\{K_h(X, x_0)\exp\{g(U)\}(1-\pi(U,Y))\right\} \\ &= E\left\{K_h(X, x_0)\exp\{g(U)\}(1-\varrho(X))\right\}, \end{aligned}$$

where $\varrho(X) = E\{\pi(U,Y)|X\}$. By some calculations, we get that

$$\begin{aligned} E\{K_h(X, x_0)\exp\{g(U)\}(1-\varrho(X))\} &= \frac{1}{h^d}\int K\left(\frac{x-x_0}{h}\right)\exp\{g(u)\}(1-\varrho(x))f(x)dx \\ &= \int K(t)\exp\{g(th+u_0)\}(1-\varrho(th+x_0))f(th+x_0)dt \\ &= f(x_0)(1-\varrho(x_0))\exp\{g(u_0)\} + O_p(h^{2d}), \end{aligned}$$

where $f(x)$ is the marginal density of X. Let $H(x_0) = f(x_0)(1-\varrho(x_0))\exp\{g(u_0)\}$,

then we have

$$\begin{aligned} I_3 &= \frac{1}{n^2}\sum_{i=1}^{n}\frac{1-\delta_i}{H(X_i)}\sum_{j=1}^{n}\delta_j K_h(X_j,X_i)\exp(\gamma_0 Y_j)[\psi(Y_j,X_j,\boldsymbol{\theta})-m_0(X_i,\boldsymbol{\theta})]\{1++O_p(c_n)\} \\ &=: I_3^*\{1+O_p(c_n)\}, \end{aligned}$$

where $c_n = (\log n/(nh^d))^{1/2} + h^m$. Next we give the asymptotic representation of I_3^*. For simplicity, we define a kernel function of the U statistic for all pairs (i,j),

$$\begin{aligned} T(S_i,S_j) &= \frac{1}{2}\frac{1-\delta_i}{H(X_i)}\delta_j K_h(X_j,X_i)e^{(\gamma_0 Y_j)}[\psi(Y_j,X_j,\boldsymbol{\theta}_0)-m_0(X_i,\boldsymbol{\theta}_0)] \\ &\quad +\frac{1}{2}\frac{1-\delta_j}{H(X_j)}\delta_i K_h(X_i,X_j)e^{(\gamma_0 Y_i)}[\psi(Y_i,X_i,\boldsymbol{\theta}_0)-m_0(X_j,\boldsymbol{\theta}_0)] \\ &=: \frac{1}{2}J_1+\frac{1}{2}J_2, \end{aligned}$$

where $S_j = (X_j, Y_j, \delta_j)$. Then I_3^* can be written as

$$I_3^* = \frac{1}{n^2}\sum_{i=1}^{n} T(S_i,S_i) + U_n,$$

where $U_n = \frac{2}{n^2}\sum_{i=1}^{n}\sum_{i<j} T(S_i,S_j)$. For any given $\boldsymbol{\theta}_0 \in \Theta^p$, it is easy to show that $E\{T(S_i,S_j)\} = 0$. By the law of large numbers, the first term of I_3^* approximates $o(n^{-1})$ in probability. Hence, it suffices to consider the U-statistic U_n only. Next we calculate the conditional expectation $T_1(S_j) =$

$E\{T(S_i,S_j)|S_j\}$. For the first part of $T(S_i,S_j)$, we have

$$\begin{aligned}
E\{J_1|S_j\} &= \delta_j e^{\gamma_0 Y_j} E\left\{\frac{1-\delta_i}{H(X_i)}K_h(X_j,X_i)[\psi(Y_j,X_j,\boldsymbol{\theta}_0)-m_0(X_i,\boldsymbol{\theta}_0)]\Big|S_j\right\}\\
&= \delta_j e^{\gamma_0 Y_j} E\left\{\frac{1-\varrho(X_i)}{H(X_i)}K_h(X_j,X_i)[\psi(Y_j,X_j,\boldsymbol{\theta}_0)-m_0(X_i,\boldsymbol{\theta}_0)]\Big|S_j\right\}\\
&= \delta_j e^{\gamma_0 Y_j}\frac{f(X_j)}{H(X_j)}(1-\varrho(X_j))[\psi(Y_j,X_j,\boldsymbol{\theta}_0)-m_0(X_j,\boldsymbol{\theta}_0)]\{1+O(h^{2d})\}\\
&= \delta_j\left\{\frac{1}{\pi(U_j,Y_j)}-1\right\}\{\psi(Y_j,X_j,\boldsymbol{\theta}_0)-m_0(X_j,\boldsymbol{\theta}_0)\}\{1+O(h^{2d})\},
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
E(J_2|S_j) &= \frac{1-\delta_j}{H(X_j)}E\left\{\delta_i K_h(X_i,X_j)e^{\gamma_0 Y_i}[\psi(Y_i,X_i,\boldsymbol{\theta}_0)-m_0(X_j,\boldsymbol{\theta}_0)]\Big|S_j\right\}\\
&= \frac{1-\delta_j}{H(X_j)}E\left\{\delta e^{\gamma_0 Y}[\psi(Y,X,\boldsymbol{\theta}_0)-m_0(X_j,\boldsymbol{\theta}_0)]|X=X_j\right\}f(X_j)\{1+O(h^{2d})\}\\
&= \frac{1-\delta_j}{H(X_j)}f(X_j)\left\{E[\delta e^{\gamma_0 Y}\psi(Y,X,\boldsymbol{\theta}_0)|X=X_j]-E[\delta e^{\gamma_0 Y}|X=X_j]m_0(X_j,\boldsymbol{\theta}_0)\right\}+O(h^{2d})\\
&= O(h^{2d}).
\end{aligned}$$

Therefore, the conditional expectation can be written as $T_1(S_j)=\frac{1}{2}\delta_j\left[\frac{1}{\pi(U_j,Y_j)}-1\right][\psi(Y_j,X_j,\boldsymbol{\theta}_0)-m_0(X_j,\boldsymbol{\theta}_0)]\{1+O(h^{2d})\}$. Denote the U-statistic projection of U_n as

$$\hat{U}_n=\frac{2}{n}\sum_{i=1}^{n}T_1(S_i).$$

Then we will show that U_n can be approximated by $\hat{U}_n$. By some tedious calculations, we have

$$\begin{aligned}
\chi_1(\boldsymbol{\theta}_0) &= Var(T_1(S_i))=\frac{1}{4}E\left\{\delta_i\left[\frac{1}{\pi(U_i,Y_i)}-1\right]^2[\psi(Y_i,X_i,\boldsymbol{\theta}_0)-m_0(X_i,\boldsymbol{\theta}_0)]^{\otimes 2}\right\}\{1+O(h^{2d})\}\\
&= \frac{1}{4}E\left\{\frac{(1-\pi(U,Y))^2}{\pi(U,Y)}[\psi(Y,X,\boldsymbol{\theta}_0)-m_0(X_i,\boldsymbol{\theta}_0)]^{\otimes 2}\right\}\{1+O(h^{2d})\},
\end{aligned}$$

and

$$\begin{aligned}\chi_2(\boldsymbol{\theta}_0) &= Var(T(S_i,S_j)) = \frac{1}{2}E\left\{\frac{1-\delta_i}{H(X_i)^2}\delta_j K_h^2(X_i,X_j)e^{2\gamma_0 Y_j}[\psi(Y_j,X_j,\boldsymbol{\theta}_0)-m_0(X_i,\boldsymbol{\theta}_0)]^{\otimes 2}\right\}\\ &= \frac{1}{2h^d}E\left\{\frac{\delta_j}{H(X_j)^2}e^{2\gamma_0 Y_j}f(X_j)(1-\varrho(X_j))[\psi(Y_j,X_j,\boldsymbol{\theta}_0)-m_0(X_j,\boldsymbol{\theta}_0)]^{\otimes 2}\right\}\int K^2(x)dx+O(1)\\ &= \frac{1}{2h^d}E\left\{\frac{(1-\pi(U,Y))^2}{\pi(U,Y)f(X)(1-\varrho(X))}[\psi(Y,X,\boldsymbol{\theta}_0)-m_0(X,\boldsymbol{\theta}_0)]^{\otimes 2}\right\}\int K^2(x)dx+O(1).\end{aligned}$$

Furthermore, $E(\hat{U}_n^2) = \frac{4}{n}\chi_1(\boldsymbol{\theta}_0)$ and $EU_n^2 = Var(U_n) = \frac{4(n-2)\chi_1(\boldsymbol{\theta}_0)}{n(n-1)} + \frac{2\chi_2(\boldsymbol{\theta}_0)}{n(n-1)}$. Therefore,

$$E(U_n-\hat{U}_n)^2 = \frac{2\chi_2(\boldsymbol{\theta}_0)}{n(n-1)}+O(n^{-2}),$$

which means that

$$\begin{aligned}U_n &= \hat{U}_n+\left\{\frac{2\chi_2(\boldsymbol{\theta}_0)}{n(n-1)}+O(n^{-2})\right\}^{1/2}\\ &= \frac{1}{n}\sum_{i=1}^n\delta_i\left[\frac{1}{\pi(U_i,Y_i)}-1\right][\psi(Y_i,X_i,\boldsymbol{\theta}_0)-m_0(X_i,\boldsymbol{\theta}_0)]\{1+O(c_n)\}+O((n^2h^d)^{-1/2}).\end{aligned}$$

Thus, we summarize the above conclusion as the following equation:

$$\sqrt{n}(I_3-I_3^{**})=o_p(1),$$

which complete the proof. □

Proof of Lemma 2. By Lemma 1, we have

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^n\hat{\psi}(Y_i,X_i,\boldsymbol{\theta}_0) &= \sqrt{n}(I_1+I_2+I_3^{**})+o_p(1)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\left\{m_0(X_i,\boldsymbol{\theta}_0)+\frac{\delta_i}{\pi(U_i,Y_i)}[\psi(Y_i,X_i,\boldsymbol{\theta}_0)-m_0(X_i,\boldsymbol{\theta}_0)]\right\}+o_p(1)\\ &:= \frac{1}{\sqrt{n}}\sum_{i=1}^n\eta_i+o_p(1),\end{aligned}$$

where $\eta_i = \psi(Y_i, X_i, \boldsymbol{\theta}_0) + (\frac{\delta_i}{\pi(U_i,Y_i)} - 1)[\psi(Y_i, X_i, \boldsymbol{\theta}_0) - m_0(X_i, \boldsymbol{\theta}_0)]$. Apparently, we have $E\eta_i = 0$ and the variance of η_i is

$$\begin{aligned} D_1(\boldsymbol{\theta}_0) &= E\Bigg\{\psi(Y_i, X_i, \boldsymbol{\theta}_0)^{\otimes 2} + 2\left(\frac{\delta_i}{\pi(U_i, Y_i)} - 1\right)\psi(Y_i, X_i, \boldsymbol{\theta}_0)[\psi(Y_i, X_i, \boldsymbol{\theta}_0) - m_0(X_i, \boldsymbol{\theta}_0)]^{\tau} \\ &\quad + \left(\frac{\delta_i}{\pi(U_i, Y_i)} - 1\right)^2 [\psi(Y_i, X_i, \boldsymbol{\theta}_0) - m_0(X_i, \boldsymbol{\theta}_0)]^{\otimes 2}\Bigg\} \\ &= E\{\psi(Y_i, X_i, \boldsymbol{\theta}_0)^{\otimes 2}\} + E\Bigg\{[\psi(Y_i, X_i, \boldsymbol{\theta}_0) - m_0(X_i, \boldsymbol{\theta}_0)]^{\otimes 2} E\left[\left(\frac{\delta_i}{\pi(U_i, Y_i)} - 1\right)^2 \Bigg| X_i, Y_i\right]\Bigg\} \\ &= E\{\psi(Y_i, X_i, \boldsymbol{\theta}_0)^{\otimes 2}\} + E\Bigg\{\left(\frac{1}{\pi(U_i, Y_i)} - 1\right)[\psi(Y_i, X_i, \boldsymbol{\theta}_0) - m_0(X_i, \boldsymbol{\theta}_0)]^{\otimes 2}\Bigg\} \quad , \end{aligned}$$

therefore, by the central limit theorem, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i \xrightarrow{D} N(0, D_1(\boldsymbol{\theta}_0)),$$

which complete the proof.

When the tilting parameter γ is unknown, the proof is similar, we omit the details. □

S3 The Study of Instrumental Variable

In this section, we study the choice of instrumental variable and the impact on the estimators through some numerical examples.

Test 1. Violation of conditions for instrument

We consider a two dimensional covariates vector $X = (Z, U)$, where Z is generated from a discrete distribution with $P(Z = 1) = 0.25, P(Z = 2) =$

$0.25, P(Z = 3) = 0.3, P(Z = 4) = 0.1, P(Z = 5) = 0.1$, and $U \sim N(0, 1)$. The response variable Y is generated from the linear regression model $Y = \theta_0 + \theta_1 Z + \theta_2 U + \epsilon$, where $\epsilon \sim N(0, 1)$. Z is treated as the true instrument and the propensity model is given by

$$\pi(Y, X) = \frac{\exp(\phi_0 + \phi_1 Z + \phi_2 U + \phi_3 Y)}{1 + \exp(\phi_0 + \phi_1 Z + \phi_2 U + \phi_3 Y)}.$$

To estimate $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)$ and $\gamma = -\phi_3$, we construct the following estimating equations

$$\boldsymbol{\psi}(Y, U, Z, \boldsymbol{\theta}) = \begin{pmatrix} Y - \theta_0 - \theta_1 U - \theta_2 Z \\ U(Y - \theta_0 - \theta_1 U - \theta_2 Z) \\ I(Z = 1)(Y - \theta_0 - \theta_1 U - \theta_2 Z) \\ I(Z = 2)(Y - \theta_0 - \theta_1 U - \theta_2 Z) \\ I(Z = 3)(Y - \theta_0 - \theta_1 U - \theta_2 Z) \end{pmatrix},$$

where $I(\cdot)$ is an indicator function.

We examine the performance of the proposed method under different combinations of the identifiability conditions (C1) and (C2) as follows:

V1. (C1) is satisfied while (C2) is violated with that Y is not associated with instrument Z. We choose $(\theta_0, \theta_1, \theta_2) = (1, 0, 1)$, $(\phi_0, \phi_1, \phi_2, \phi_3) = (0.8, 0, 0.5, 0.1)$, the corresponding missing rate is about 30%.

V2. (C2) is satisfied while (C1) is violated with Z, U and Y in the propen-

sity model. We choose $(\theta_0, \theta_1, \theta_2) = (1, 1, 1)$, $(\phi_0, \phi_1, \phi_2, \phi_3) = (0.8, 0.5, 0.5, 0.1)$. Missing rate is around 10%.

V3. Both (C1) and (C2) are violated. We choose $(\theta_0, \theta_1, \theta_2) = (1, 0, 1)$, $(\phi_0, \phi_1, \phi_2, \phi_3) = (0.8, 0.5, 0.5, 0.1)$. Missing rate is about 12%.

V4. Both (C1) and (C2) are satisfied. We choose $(\theta_0, \theta_1, \theta_2) = (1, 1, 1)$, $(\phi_0, \phi_1, \phi_2, \phi_3) = (0.8, 0, 0.5, 0.1)$. Missing rate is about 25%.

For each case, we conduct 1000 replications with sample size 500. Table 1 summarizes the simulated results. From this table, we can see that the bias and MSE (mean squared error) of the estimates in V4 are the smallest since the identifiability conditions are satisfied. The estimation has a poor performance in V3 with biased estimate for γ due to the violation of the identifiability conditions. And the estimates for cases V1 and V2, which violate one of the identifiability conditions respectively, have a comparable performances between that of V3 and V4. Above all, the ideal situation is that we can choose an instrumental variable meeting the two identifiability conditions. Or at least, one of the conditions should be satisfied.

Test 2. The choice of instrument

As discussed in Shao and Wang (2016), we need to choose an instrument Z to meet the conditions (C1) and (C2) of Theorem 1. Suppose that Z is

related to Y, otherwise, Z can be excluded from the study. Next, we need to check that Z can be excluded from the propensity $\pi(X, Y)$, i.e., (C2) holds. If we can specify the right formulation for the propensity $\pi(U, Y) = \pi(U, Y; \alpha, g(U))$ with true instrument Z, the tilting parameter α can be estimated by maximizing the profile likelihood function

$$L(\alpha) = \prod_{i=1}^{n} \pi(U_i, Y_i; \alpha, g(U_i))^{\delta_i} (1 - \pi(U_i, Y_i; \alpha, g(U_i)))^{1-\delta_i},$$

where $g(U_i)$ is replaced by its kernel estimate. And it is equivalent to solve the score equation $S(\alpha) = \partial \log L(\alpha)/\partial\alpha = 0$, which can be expressed as $S(\alpha) = \sum_{i=1}^{n} s(\delta_i, U_i, Y_i; \alpha) = \sum_{i=1}^{n} \{\delta_i - \pi(U_i, Y_i; \alpha)\} h(U_i, Y_i; \alpha) = 0$, and $h(U_i, Y_i; \alpha) = \pi(U_i, Y_i; \alpha)^{-1}\{1 - \pi(U_i, Y_i; \alpha)\}^{-1} \partial\pi(U_i, Y_i; \alpha)/\partial\alpha$. Considering that Y_i is missing when $\delta_i = 0$, we propose to estimate α by solving the imputed mean score equation

$$\tilde{S}(\alpha) = \sum_{i=1}^{n} \delta_i s(\delta_i, U_i, Y_i; \alpha) + (1 - \delta_i) E[s(\delta, U, Y; \alpha) | X = X_i, \delta_i = 0] = 0,$$

where $E[s(\delta, U, Y; \alpha) | X = X_i, \delta_i = 0]$ can be estimated using the kernel regression method. Denote the resulting estimator by $\tilde{\pi}(U, Y)$, which does not converge to the true propensity $\pi(U, Y)$ if Z can not be excluded from the propensity. Consequently,

$$D = \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i X_i}{\tilde{\pi}(U_i, Y_i)} - \frac{1}{n} \sum_{i=1}^{n} X_i \right\|$$

does not converge to zero in probability. In other word, D converges to zero

if and only if Z is an instrument and $\pi(U, Y)$ is a correct model. Hence, we can select an instrument by minimizing D over a group of candidate variables.

To test the reliability of the criterion D, we conduct a numerical study. Let $X = (U, Z)$, where $U \sim N(1, 1)$ and $Z \sim Bernoulli(0.5)$ are independent. Given X, $Y \sim \lambda N(U, 1)^2 + (1 - \lambda)N(Z, 1)^2$, $0 \leq \lambda \leq 1$. We consider two cases for the propensity:

$$\pi(U, Y) = \frac{\exp(\phi_0 + \phi_1 U + \alpha Y)}{1 + \exp(\phi_0 + \phi_1 U + \alpha Y)}, \tag{S3.1}$$

and

$$\pi(U, Y) = \frac{\exp(\phi_2 + \phi_3 U + \sin(Y))}{1 + \exp(\phi_2 + \phi_3 U + \sin(Y))}, \tag{S3.2}$$

which correspond to the right and misspecified model, respectively. For both cases, Z plays the role of instrument. Several settings are considered for $(\lambda, \phi_0, \phi_1, \alpha, \phi_2, \phi_3)$ to assess the performance of D.

(a) $(\lambda, \phi_0, \phi_1, \alpha, \phi_2, \phi_3) = (0.5, 0.6, 0.4, 0.2, 0.6, 0.5)$.

(b) $(\lambda, \phi_0, \phi_1, \alpha, \phi_2, \phi_3) = (0.5, 0.6, 0.4, -0.3, -0.5, 0.5)$.

(c) $(\lambda, \phi_0, \phi_1, \alpha, \phi_2, \phi_3) = (0, 0.2, 0.8, 0.3, 1, 0.2)$.

(d) $(\lambda, \phi_0, \phi_1, \alpha, \phi_2, \phi_3) = (0, 0.6, 0.4, -0.5, 0.2, -0.2)$.

(e) $(\lambda, \phi_0, \phi_1, \alpha, \phi_2, \phi_3) = (0.9, 0.2, 0.7, 0.4, 0.6, 0.5)$.

(f) $(\lambda, \phi_0, \phi_1, \alpha, \phi_2, \phi_3) = (0.9, 0.2, 0.7, -0.2, -0.5, 0.6)$.

In settings (a) and (b), both Z and U are related to Y. And, only Z is related to Y in settings (c) and (d) due to that $\lambda = 0$. Settings (e) and (f) correspond to the case where the instrument Z has a relatively little impact on Y. The missing rates in settings (a), (c) and (e) are about 20% and the others are about 42%. For each setting, the sample size n is 500 or 1000. The simulated results based on 1000 replications for model (S3.1) and model (S3.2) are summarized in Tables 2 and 3, respectively. From Table 2, we can see that the D values based on the true instrument Z are always smaller than that using U as instrument. Moreover, both D values and the standard deviation of D are reduced with increasing sample size. We also report the selected probabilities in 1000 replications that the correct instrument is selected by the criterion D. For different cases, the proposed criterion selects the correct instrument in almost all replications, which means that the proposed criterion D is a reliable method to select instrument. From Table 3, we can see that the probability of correctly selecting the instrument is lower than that in Table 2 as expected, which means that an accurate propensity model is also important in selecting the instrument using the criterion D.

Remark 1. In this study, we use the nonresponse instrument, which is

related to the response but can be excluded from the propensity, to avoid the identifiability issue. This nonresponse instrument is different from the usual instrument in standard IV (instrumental variable) estimation where IV is independent of the error term. The validity of standard instrument can be tested by examining the orthogonality conditions in an overidentified model. In the context of GMM, the overidentifying restrictions may be tested via Anderson and Rubin (1949) test statistic or the commonly employed J-statistic of Hansen (1982). In the IV context, this statistic is known as the Sargan (1958) statistic. The Hansen-Sargan tests for overidentification evaluate the entire set of overidentifying restrictions. When the researcher has prior suspicions about the validity of a subset of instruments, a "difference-in-Sargan" statistic can be employed to test a subset of orthogonality conditions.

In our paper, all covariates are exogenous and thus the standard instruments are themselves, however, we still need the nonresponse instrument to identify the response model. Our main interest is to estimate θ, the titling parameter γ is a nuisance parameter, when γ is unknown the target is to estimate θ and γ simultaneously. We assume that the estimating equations are overidentified for θ, thus for $\boldsymbol{\beta} = (\theta, \gamma)^T$, the equations are at least exactly identified. To test whether the model fits the data, tests for

overidentification presented above can be employed for our equations.

To verify the validity of a nonresponse instrumental variable, we should check the two conditions in Theorem 1. Condition (C1) can be assessed by an examination of the significance of the excluded instruments in estimation. For Condition (C2), we should demonstrate the conditional independence of Z and δ, however this independence is not embodied in the estimating equations as in the standard IV estimator context. In our paper, we use the criterion D to select the nonresponse instrument and verify the effect of the conditions through simulation studies. Theoretical proofs are our future research.

S4 Data example

In the data example, we use the proposed criterion D to identify the instrumental variable and find that years in the major leagues (X_1) is the best candidate instrument with the smallest D value. To investigate the effect of invalid instrumental variable, we consider the estimates with all possible instrument subsets. We use S1-S6 to represent the scenarios with different instrumental variables respectively, and the corresponding estimates are reported in Tables 4 and 5.

The D values based on years in the major leagues or players' division

are smaller than that based on hits. When hits is added to the instrument, the D value increases a lot, which means that hits should be excluded from the instrument. Although the D value using years in the major leagues is smaller than that of division, the estimates of parameters for S1 and S2 are not significantly different. And, the estimates of γ indicate that the nonignorable missing assumption holds for the response variable. In addition, the estimates of S3-S6 are very similar, except for the estimate of γ in S5 with $Z = (X_1, X_3)^\tau$, that has the largest D value. Overall, the estimates of $(\theta_0, \theta_1, \theta_2, \theta_3)$ are not much different among these six scenarios.

References

Anderson, T. W. and Rubin, H. (1949), Estimation of the parameters of single euqation in a complete system of stochastic equation. *Annals of Math. Stat.* **20**, 46-63.

Hansen, L. (1982), Large sample properties of generalized method of moments estimators. *Econometrica* **50**: 1029-1054.

Sargan, J. D. (1958), The estimation of economic relationships using instrumental variables. *Econometrica* **26**, 393-415.

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

Wang, S., Shao, J. and Kim, J. (2014), An instrumental variable approach for identification and

estimation with nonignorable nonresponse. *Statist. Sinica* **24**, 1097-1116.

Table 1: Simulation results for V1–V4

	True	Bias	MSE	True	Bias	MSE
	$V1$			$V2$		
$\hat{\theta}_0$	1.0000	0.0423	0.0143	1.0000	-0.0027	0.0688
$\hat{\theta}_1$	0	-0.0083	0.0029	1.0000	0.0046	0.0041
$\hat{\theta}_2$	1.0000	0.0041	0.0114	1.0000	0.0044	0.0052
$\hat{\gamma}$	-0.1000	0.0969	0.2954	-0.1000	-0.0162	0.3839
	$V3$			$V4$		
$\hat{\theta}_0$	1.0000	0.0542	0.0883	1.0000	0.0028	0.0097
$\hat{\theta}_1$	0	-0.0096	0.0051	1.0000	0.0045	0.0033
$\hat{\theta}_2$	1.0000	0.0045	0.0159	1.0000	0.0038	0.0093
$\hat{\gamma}$	-0.1000	0.2321	0.4228	-0.1000	0.0025	0.0017

Table 2: Simulation results for D based on 1000 replications

		$n = 500$			$n = 1000$		
Scenario	instrument	D	SD(D)	Selects(%)	D	SD(D)	Selects(%)
(a)	Z	0.0002	0.0003	93.0	0.0001	0.0002	98.4
	U	0.0022	0.0020		0.0017	0.0013	
(b)	Z	0.0009	0.0014	92.7	0.0004	0.0007	99.0
	U	0.0841	0.0585		0.0883	0.0434	
(c)	Z	0.0006	0.0009	99.8	0.0003	0.0006	99.9
	U	0.0297	0.0127		0.0289	0.0102	
(d)	Z	0.0012	0.0021	98.6	0.0006	0.0015	99.8
	U	0.0345	0.0256		0.0314	0.0169	
(e)	Z	0.0002	0.0004	99.7	0.0001	0.0002	100
	U	0.0065	0.0040		0.0058	0.0028	
(f)	Z	0.0007	0.0009	99.4	0.0003	0.0004	100
	U	0.2688	0.1263		0.2822	0.0911	

Table 3: Simulation results for D based on 1000 replications

		$n = 500$			$n = 1000$		
Scenario	instrument	D	SD(D)	Selects(%)	D	SD(D)	Selects(%)
(a)	Z	0.0007	0.0018	73.8	0.0005	0.0010	80.8
	U	0.0020	0.0039		0.0019	0.0035	
(b)	Z	0.0020	0.0041	78.9	0.0015	0.0040	86.3
	U	0.0060	0.0078		0.0058	0.0100	
(c)	Z	0.0013	0.0017	63.2	0.0012	0.0015	61.2
	U	0.0021	0.0024		0.0017	0.0019	
(d)	Z	0.0068	0.0096	61.2	0.0069	0.0087	62.5
	U	0.0085	0.0105		0.0085	0.0087	
(e)	Z	0.0006	0.0015	74.5	0.0003	0.0009	84.1
	U	0.0029	0.0053		0.0032	0.0066	
(f)	Z	0.0019	0.0047	72.7	0.0011	0.0035	76.5
	U	0.0061	0.0101		0.0076	0.0154	

Table 4: Results for Baseball data with one instrument variable

Scenario	Instrument	D	parameters	Estimates	SE	Confidence interval
S1	years	0.0327	θ_0	4.0252	0.1303	[3.7698, 4.2807]
			θ_1	0.0963	0.0069	[0.0829, 0.1098]
			θ_2	0.2084	0.0685	[0.0741, 0.3427]
			θ_3	0.0095	0.0010	[0.0076, 0.0114]
			γ	-3.1300	0.0094	[-3.1484, -3.1117]
S2	division	2.1027	θ_0	4.0255	0.1289	[3.7729, 4.2782]
			θ_1	0.0963	0.0068	[0.0830, 0.1097]
			θ_2	0.2080	0.0690	[0.0727, 0.3432]
			θ_3	0.0095	0.0010	[0.0076, 0.0114]
			γ	-3.1301	0.0093	[-3.1482, -3.1119]
S3	hits	33.5695	θ_0	3.8162	0.1278	[3.5657, 4.0666]
			θ_1	0.1021	0.0075	[0.0874 , 0.1167]
			θ_2	0.2568	0.0763	[0.1073 , 0.4063]
			θ_3	0.0100	0.0010	[0.0080, 0.0120]
			γ	-34.8898	0.0001	[-34.8900, -34.8895]

Table 5: Results for Baseball data with two instrument variables

Scenario	Instrument	D	parameters	Estimates	SE	Confidence interval
S4	years, division	0.1284	θ_0	3.8158	0.1278	[3.5653 , 4.0664]
			θ_1	0.1021	0.0075	[0.0875 , 0.1167]
			θ_2	0.2571	0.0763	[0.1076 , 0.4066]
			θ_3	0.0100	0.0010	[0.0080 , 0.0120]
			γ	-34.8891	0.0001	[-34.8893 , -34.8888]
S5	years, hits	39.3146	θ_0	3.7738	0.2017	[3.3784 , 4.1693]
			θ_1	0.1058	0.0217	[0.0632 , 0.1484]
			θ_2	0.2508	0.0760	[0.1018 , 0.3998]
			θ_3	0.0102	0.0011	[0.0081, 0.0122]
			γ	-48.1348	57.2586	[-160.3617 , 64.0920]
S6	division, hits	34.7616	θ_0	3.8162	0.1278	[3.5658 , 4.0666]
			θ_1	0.1021	0.0074	[0.0875, 0.1167]
			θ_2	0.2564	0.0762	[0.1070 , 0.4058]
			θ_3	0.0100	0.0010	[0.0080 , 0.0120]
			γ	-34.8906	0.0001	[-34.8908 ,-34.8903]