Statistica Sinica Preprint No: SS-2019-0025	
Title	Propensity model selection with nonignorable nonresponse and instrument variable
Manuscript ID	SS-2019-0025
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0025
Complete List of Authors	Lei Wang Jun Shao and Fang Fang
Corresponding Author	Fang Fang
E-mail	ffang@sfs.ecnu.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

Propensity Model Selection with Nonignorable Nonresponse and Instrument Variable

Lei Wang[1], Jun Shao[2,3] and Fang Fang[2]

[1] *Nankai University,* [2]*East China Normal University*

and [3]*University of Wisconsin-Madison*

Abstract: Handling data with nonignorable missing responses is difficult because of the identifiability issue caused by nonignorable nonresponse. An effective approach in the literature is to impose a parametric model on the nonresponse propensity (while the conditional distribution of the response given covariates is totally unspecified), and to utilize a nonresponse instrument, a useful covariate vector that can be excluded from the propensity given the response and other covariates. However, how to find a nonresponse instrument from a given set of covariates is not well addressed. Further, one may also want to select a parametric propensity model from a set of candidate models. We propose a simultaneous propensity model and instrument selection criterion. In the presence of nonignorable nonresponse, the proposed method can consistently select the most compact correct parametric propensity model and instrument from a group of candidate models, if one of these candidate models is correct and an instrument exists. Simulation results show that our proposed method works quite well. A real data example is presented for illustration.

Key words and phrases: Generalized method of moments; identifiability; penalized validation criterion; misspecified model; nonignorable propensity; nonresponse instrument.

1. Introduction

Consider the problem in which a univariate outcome or response Y is subject to nonresponse and a vector $\boldsymbol{X}$ of covariates is always observed, and estimation or inference on unknown quantities in F_Y, the distribution of Y, or in $F_{Y|\boldsymbol{X}}$, the conditional distribution of Y given $\boldsymbol{X}$, is of interest. The conditional probability $\Pr(\delta = 1|Y, \boldsymbol{X})$ is called the nonresponse propensity or simply propensity, where δ is the indicator of observing Y. When Y can be excluded from the propensity $\Pr(\delta = 1|Y, \boldsymbol{X})$ so that it is a function of $\boldsymbol{X}$ only, the propensity is ignorable and missing data are at random (Little and Rubin, 2002), in which case unknown quantities in F_Y or $F_{Y|\boldsymbol{X}}$ can be estimated using $F_{Y|\boldsymbol{X},\delta=1}$ and $F_{\boldsymbol{X}}$ because $F_{Y|\boldsymbol{X}} = F_{Y|\boldsymbol{X},\delta=1}$; see, for example, Rubin (1987), Cheng (1994), Robins et al. (1994), Ibrahim et al. (2005), Kim and Shao (2013), and the references therein. When Y cannot be excluded from $\Pr(\delta = 1|Y, \boldsymbol{X})$, the propensity is nonignorable and developing a valid estimation method is notoriously challenging. In this case, the population parameters are generally not identifiable and estimates based on the assumption of ignorable

nonresponse may have large biases (Fitzmaurice et al., 1995; Wang et al., 2014). Thus, methods very different from those for ignorable propensity have to be applied; see, for example, Scharfstein et al. (1999), Qin et al. (2002), Tang et al. (2003), Kim and Yu (2011), Xie et al. (2011), Wang et al. (2014), Tang et al. (2014), Zhao and Shao (2015), Shao and Wang (2016), Guan and Qin (2017), and the references therein.

When the propensity is nonignorable, the distribution of $(\delta, Y, \boldsymbol{X})$ is typically not identifiable (Robins and Ritov, 1997; Wang et al., 2014). Two general sufficient conditions for the identifiability are:

$$\begin{aligned} &\Pr(\delta = 1|Y, \boldsymbol{X}) = \pi(Y, \boldsymbol{U}), \quad \boldsymbol{X} = (\boldsymbol{U}, \boldsymbol{Z}), \\ &F_{Y|\boldsymbol{X}} \text{ depends on } \boldsymbol{Z}, \end{aligned} \tag{1.1}$$

and

$$\text{there is a parametric component in either } F_{Y|\boldsymbol{X}} \text{ or } \pi(Y, \boldsymbol{U}). \tag{1.2}$$

Condition (1.1) means that, when Y cannot be excluded from the propensity, a sub-vector $\boldsymbol{Z}$ of $\boldsymbol{X}$ can be excluded, and $\boldsymbol{Z}$ is still a useful covariate for Y. Such a $\boldsymbol{Z}$ was named as nonresponse instrument by Wang et al. (2014). Excluding Y or $\boldsymbol{Z}$ from the propensity simplifies the form of propensity and enables us to identify it. Although (1.1) and (1.2) are sufficient conditions, without either of them leads to examples of nonidentifiable distribution of

$(\delta, Y, \boldsymbol{X})$; see Wang et al. (2014) for (1.1) and Robins and Ritov (1997) for (1.2).

For condition (1.2), of course one can impose parametric models on both $\pi(Y, \boldsymbol{U})$ and $F_{Y|\boldsymbol{X}}$; see, for example, Molenberghs and Kenward (2007). Efforts have been made in deriving results under semiparametric models. In one direction, Tang et al. (2003) and Zhao and Shao (2015) studied a pseudo-likelihood method under a parametric model on $F_{Y|\boldsymbol{X}}$ but an unspecified propensity $\pi(Y, \boldsymbol{U})$ with a given instrument $\boldsymbol{Z}$. In the other direction, which is the focus of the current paper, Wang et al. (2014) derived estimators under a parametric model on the propensity $\pi(Y, \boldsymbol{U})$ but allowed an unspecified $F_{Y|\boldsymbol{X}}$, i.e., (1.2) is replaced by

$$\pi(Y, \boldsymbol{U}) \text{ follows a parametric model but } F_{Y|\boldsymbol{X}} \text{ is unspecified.} \qquad (1.2\text{A})$$

The main technique in Wang et al. (2014) is to use a given instrument $\boldsymbol{Z}$ to create enough estimating equations for the estimation of the parametric propensity $\pi(Y, \boldsymbol{U})$ in (1.2A); once the propensity is estimated, unknown quantities can be estimated by inverse propensity weighting. However, two important issues for carrying out this approach have not been studied. The first one is how to find an instrument, a sub-vector $\boldsymbol{Z}$ of $\boldsymbol{X}$ satisfying (1.1). The second issue is how to select a parametric model for the propensity $\pi(Y, \boldsymbol{U})$. Although there are many publications on model selection with ignorable missing responses,

to the best of our knowledge, there are only two papers on model selection with nonignorable missing responses: Fang and Shao (2016) and Zhao et al. (2018) considered model/variable selection for $F_{Y|\boldsymbol{X}}$ when $F_{Y|\boldsymbol{X}}$ is parametric and $\pi(Y, \boldsymbol{U})$ is unspecified, which is different from (1.2A), i.e., the situation we focus. Further, they did not consider how to search for an instrument.

The purpose of this paper is to propose a method to simultaneously search for an instrument satisfying (1.1) and select a parametric model for the propensity from a set of available models. We formulate the search for an instrument and propensity model selection into one model selection framework. Our key idea is to construct and compare two estimators of $F_{\boldsymbol{X}}$, the cumulative distribution function of the covariate vector $\boldsymbol{X}$. Since $\boldsymbol{X}$ is always observed, a simple consistent estimator not depending on any model and instrument is the empirical cumulative distribution function $\hat{F}_{\boldsymbol{X}}$ based on $\boldsymbol{X}$ data. On the other hand, for candidate parametric propensity model k on $\pi(Y, \boldsymbol{U})$ with a possible instrument $\boldsymbol{Z}$, we construct an inverse propensity estimator $\hat{F}_k$ of $F_{\boldsymbol{X}}$ using both Y and $\boldsymbol{X}$ data and the model information in the presence of nonignorable missing Y data. Since only a correct candidate model and a correct instrument can produce a consistent estimator $\hat{F}_k$ close to $\hat{F}_{\boldsymbol{X}}$, we select a model from a group of candidate models and an instrument by minimizing a closeness measure between the two estimators $\hat{F}_{\boldsymbol{X}}$ and $\hat{F}_k$. Because some

propensity models may be correct but overfitted, we add a penalty term in our model selection criterion, following the general principle of the well-known BIC model selection.

When there exists an instrument and the group of candidate models contains at least one correct propensity model, our theory shows that, with probability tending to one as the sample size increases to infinity while the dimension of $\boldsymbol{X}$ remains fixed, the proposed method can simultaneously select the most compact correct parametric propensity model and a correct instrument. Consequently, parameter estimators using the inverse propensity weighting approach based on the selected model and instrument are consistent and asymptotically normal. Our proposed method also works well in simulation studies and is illustrated in a data example.

2. Methodology and Theory

Under conditions (1.1) and (1.2A), we would like to select sub-vectors $\boldsymbol{Z}$ and $\boldsymbol{U}$ such that $\boldsymbol{X} = (\boldsymbol{U}, \boldsymbol{Z})$, $\boldsymbol{Z}$ is an instrument, and $\pi(Y, \boldsymbol{U})$ is the propensity. Since choosing different components of $\boldsymbol{X}$ as $\boldsymbol{Z}$ and $\boldsymbol{U}$ can be viewed as selecting different models, instrument and propensity model selection can be combined into a general model selection problem.

To illustrate, consider a 3-dimensional $\boldsymbol{X} = (X_1, X_2, X_3)$ and $\pi(Y, \boldsymbol{U}, \theta)$

being logistic in a linear combination of X_j's and Y. Then we have a total of seven models:

$$\begin{aligned}
&\pi_0(Y, \boldsymbol{U}_0, \theta_0) = 1/\{1 + \exp(\alpha_0 + \gamma_0 Y)\},\\
&\pi_j(Y, \boldsymbol{U}_j, \theta_j) = 1/\{1 + \exp(\alpha_j + \beta_j X_j + \gamma_j Y)\}, \quad j = 1, 2, 3,\\
&\pi_4(Y, \boldsymbol{U}_4, \theta_4) = 1/\{1 + \exp(\alpha_4 + \beta_{41} X_1 + \beta_{42} X_2 + \gamma_4 Y)\},\\
&\pi_5(Y, \boldsymbol{U}_5, \theta_5) = 1/\{1 + \exp(\alpha_5 + \beta_{51} X_1 + \beta_{52} X_3 + \gamma_5 Y)\},\\
&\pi_6(Y, \boldsymbol{U}_6, \theta_6) = 1/\{1 + \exp(\alpha_6 + \beta_{61} X_2 + \beta_{62} X_3 + \gamma_6 Y)\},
\end{aligned} \tag{2.1}$$

where $\boldsymbol{U}_0 = 0$, $\boldsymbol{U}_j = X_j$, $j = 1, 2, 3$, $\boldsymbol{U}_4 = (X_1, X_2)$, $\boldsymbol{U}_5 = (X_1, X_3)$, and $\boldsymbol{U}_6 = (X_2, X_3)$. The model with $\boldsymbol{U} = \boldsymbol{X}$ is excluded because we assume the existence of an instrument. These seven models actually correspond to selection of both propensity and instrument, because if model k is selected, then the selected instrument is $\boldsymbol{Z}_k$ containing components in $\boldsymbol{X}$ but not in $\boldsymbol{U}_k$ included in the propensity. If we would like to select a model between logistic and another model such as probit, then replacing $1/\{1 + \exp(\cdot)\}$ with another function results in another seven models and the total number of models becomes 14. Alternatively, we may want to add a nonlinear term such as Y^2 in the linear combination of the logistic model, which results in a total of $3 \times 7 = 21$ models, because we may have a Y term only, a Y^2 term only, or both Y and Y^2 terms.

Let K be the total number of candidate models under all combinations of $\boldsymbol{U}$ and $\boldsymbol{Z}$ decompositions, and let

$$\mathcal{M} = \{\pi_k(Y, \boldsymbol{U}_k, \boldsymbol{\theta}_k),\ k = 1, \ldots, K\}$$

be the collection of all K parametric models, where $\boldsymbol{U}_k$ is the vector $\boldsymbol{U}$ under model k, π_k is a known function of $(Y, \boldsymbol{U}_k, \boldsymbol{\theta}_k)$, and $\boldsymbol{\theta}_k$ is an unknown parameter vector with dimension d_k under model k. If model k is selected, then $\boldsymbol{Z}_k$ with $\boldsymbol{X} = (\boldsymbol{U}_k, \boldsymbol{Z}_k)$ is selected as an instrument and model $\pi_k(Y, \boldsymbol{U}_k, \boldsymbol{\theta}_k)$ is the selected propensity model. We say that model k is correct if and only if $\boldsymbol{Z}_k$ is an instrument satisfying (1.1) and $\pi_k(Y, \boldsymbol{U}_k, \boldsymbol{\theta}_k)$ is a correct propensity. Under this framework, finding an instrument and selecting a propensity model is the same as selecting a model from $\mathcal{M}$.

For simplicity, we now consider a fixed model k and omit the subscript k in $\boldsymbol{U}$ and $\boldsymbol{Z}$ in the following discussion. Let $\boldsymbol{Z} = (\boldsymbol{Z}_c, \boldsymbol{Z}_d)$, where $\boldsymbol{Z}_c$ is a continuous covariate vector and $\boldsymbol{Z}_d$ is the J_k-dimensional vector whose jth component is the indicator of a discrete covariate taking value $j = 1, ..., J_k$. Following the idea in Wang et al. (2014), to estimate the parameter $\boldsymbol{\theta}_k$ we define the vector-valued function

$$\boldsymbol{g}_k(Y, \boldsymbol{X}, \delta, \boldsymbol{\theta}_k) = \boldsymbol{h}_k(\boldsymbol{X})\Big\{\frac{\delta}{\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)} - 1\Big\}, \tag{2.2}$$

where $\boldsymbol{h}_k(\boldsymbol{X})$ is a known vector-valued function of $\boldsymbol{X}$ with dimension $L_k \geq d_k$,

the dimension of $\boldsymbol{\theta}_k$. For example, we can use $\boldsymbol{h}_k(\boldsymbol{X}) = (\boldsymbol{U}, \boldsymbol{Z}_c, \boldsymbol{Z}_d)$ when the dimension of $(\boldsymbol{U}, \boldsymbol{Z}_c)$ plus J_k is larger or equal to d_k. If the dimension of $(\boldsymbol{U}, \boldsymbol{Z}_c)$ plus J_k is smaller than d_k, we may add $\tilde{\boldsymbol{Z}}$ to $(\boldsymbol{U}, \boldsymbol{Z}_c, \boldsymbol{Z}_d)$, where components of $\tilde{\boldsymbol{Z}}$ are higher moments of $\boldsymbol{Z}_c$, such that the dimension of $(\boldsymbol{U}, \boldsymbol{Z}_c, \boldsymbol{Z}_d, \tilde{\boldsymbol{Z}})$ is not smaller than d_k. The efficiency of estimation based on (2.2) depends on the choice of $\boldsymbol{h}_k(\boldsymbol{X})$. Several approaches of choosing $\boldsymbol{h}_k(\boldsymbol{X})$ were proposed by Morikawa et al. (2017) and Ai et al. (2018). Since our focus is on model and instrument selection, in what follows we assume a fixed function $\boldsymbol{h}_k(\boldsymbol{X})$ is used in (2.2).

If model k is correct and $\boldsymbol{\theta}_k^0$ is the unique true parameter value of $\boldsymbol{\theta}_k$, then it can be verified that, under $\Pr(\delta = 1|Y, \boldsymbol{X}) = \pi(Y, \boldsymbol{U})$ and the first part of condition (1.1),

$$E\{\boldsymbol{g}_k(Y, \boldsymbol{X}, \delta, \boldsymbol{\theta}_k^0)\} = 0. \tag{2.3}$$

Thus, the function $\boldsymbol{g}_k$ in (2.2) provides an estimating equation for $\boldsymbol{\theta}_k$. The second part of condition (1.1) ensures that the estimation equations in (2.3) are not linearly dependent so we have enough equations for estimating $\boldsymbol{\theta}_k$.

Throughout, model selection is carried out based on a random sample of size n, $(\boldsymbol{X}_i, Y_i, \delta_i)$, $i = 1, ..., n$, taken from the distribution of $(\boldsymbol{X}, Y, \delta)$, where $\boldsymbol{X}_i$ is always observed and Y_i is observed if and only if $\delta_i = 1$. Since L_k may be larger than d_k, we apply the generalized method of moments to estimate $\boldsymbol{\theta}_k$,

based on (2.2)-(2.3). Specifically, let $\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) = n^{-1}\sum_{i=1}^n \boldsymbol{g}_k(Y_i, \boldsymbol{X}_i, \delta_i, \boldsymbol{\theta}_k)$, $\tilde{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}_k} \bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k)^\top \bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k)$, where $a^\top$ is the transpose of a, and let $\hat{\boldsymbol{W}}_{kn} = n^{-1}\sum_{i=1}^n \boldsymbol{g}_k(Y_i, \boldsymbol{X}_i, \delta_i, \tilde{\boldsymbol{\theta}}_k)\boldsymbol{g}_k(Y_i, \boldsymbol{X}_i, \delta_i, \tilde{\boldsymbol{\theta}}_k)^\top$. The generalized method of moment estimator of $\boldsymbol{\theta}_k$ is

$$\hat{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}_k} \bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k)^\top \hat{\boldsymbol{W}}_{kn}^{-1} \bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k). \tag{2.4}$$

For simplicity, we denote the cumulative distribution function of $\boldsymbol{X}$ by $F = F_{\boldsymbol{X}}$. Once we have $\hat{\boldsymbol{\theta}}_k$ in (2.4), an inverse propensity weighting estimator of $F(\boldsymbol{x})$ is

$$\hat{F}_k(\boldsymbol{x}) = \sum_{i=1}^n \frac{\delta_i \mathrm{I}(\boldsymbol{X}_i \le \boldsymbol{x})}{\pi_k(Y_i, \boldsymbol{U}_i, \hat{\boldsymbol{\theta}}_k)},$$

where $\mathrm{I}(\boldsymbol{X}_i \le \boldsymbol{x})$ is the indicator function of $\boldsymbol{X}_i \le \boldsymbol{x}$ and, for vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, $\boldsymbol{a} \le \boldsymbol{b}$ means that all components of $\boldsymbol{b} - \boldsymbol{a}$ are nonnegative.

If model k is correct, then it can be shown that, as the sample size $n \to \infty$, $\hat{\boldsymbol{\theta}}_k$ is consistent for $\boldsymbol{\theta}_k^0$ and $\hat{F}_k(\boldsymbol{x})$ is consistent for $F(\boldsymbol{x})$. On the other hand, if either $\boldsymbol{Z}$ is not an instrument or $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)$ is wrong, then $\hat{F}_k(\boldsymbol{x})$ is inconsistent.

Without using any model, a consistent estimator of $F(\boldsymbol{x})$ is the empirical cumulative distribution function $\hat{F}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^n \mathrm{I}(\boldsymbol{X}_i \le \boldsymbol{x})$. We then use the closeness between $\hat{F}$ and $\hat{F}_k$ to validate model k. Define the following

model validation criterion:

$$\mathrm{VC}(k) = \frac{1}{n}\sum_{i=1}^{n} |\hat{F}_k(\boldsymbol{X}_i) - \hat{F}(\boldsymbol{X}_i)|. \tag{2.5}$$

As we show later, if model $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)$ with the corresponding instrument is correct, then $\mathrm{VC}(k) \to 0$ in probability as $n \to \infty$. Otherwise, $\mathrm{VC}(k)$ does not converge to 0. Thus, correct and incorrect models can be detected by $\mathrm{VC}(k)$.

A correct model may be an overfitted model including some redundant parameters. For example, suppose that $\boldsymbol{X} = (\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{T})$, $(\boldsymbol{R}, \boldsymbol{T})$ is an instrument, and $\pi(\alpha + \gamma Y + \boldsymbol{\beta}^\top \boldsymbol{S})$ is a correct propensity, where α, γ, and $\boldsymbol{\beta}$ are unknown. Then, $\boldsymbol{T}$ is also an instrument and $\pi(\alpha + \gamma Y + \boldsymbol{\beta}^\top \boldsymbol{S} + \mathbf{0}^\top \boldsymbol{R})$ is a correct propensity containing a redundant $\boldsymbol{R}$, where $\mathbf{0}$ is the vector of zeros. A more compact propensity model may result in more efficient propensity and other parameter estimators (see, e.g., the simulation results in Section 3). Thus, we define the best model to be the most compact correct propensity model in $\mathcal{M}$, and penalize model dimension following the idea of the well-known BIC, i.e., we choose a model by minimizing the penalized validation criterion (PVC):

$$\begin{aligned} \mathrm{PVC}_\lambda(k) &= \mathrm{VC}(k) + \lambda \log(d_k), \\ \hat{k} &= \mathrm{argmin}_{1 \le k \le K} \mathrm{PVC}_\lambda(k), \end{aligned} \tag{2.6}$$

where d_k is the dimension of $\boldsymbol{\theta}_k$ and $\lambda \ge 0$ is a penalization factor that may depend on n and the sample data. The selected instrument is $\boldsymbol{Z}_{\hat{k}}$ with

$\boldsymbol{X} = (\boldsymbol{U}_{\hat{k}}, \boldsymbol{Z}_{\hat{k}})$ and the selected model is $\pi_{\hat{k}}(Y, \boldsymbol{U}_{\hat{k}}, \boldsymbol{\theta}_{\hat{k}})$. Quantities of interest can be estimated using the inverse propensity weighting with the estimated propensity $\pi_{\hat{k}}(Y, \boldsymbol{U}_{\hat{k}}, \hat{\boldsymbol{\theta}}_{\hat{k}})$. For example, the population mean $\mu = E(Y)$ can be estimated by

$$\hat{\mu}_{pvc} = \sum_{i=1}^{n} \frac{\delta_i Y_i}{\pi_{\hat{k}}(Y_i, \boldsymbol{U}_{\hat{k}i}, \hat{\boldsymbol{\theta}}_{\hat{k}})}. \tag{2.7}$$

We now show some asymptotic properties of the proposed method of instrument and model selection. If $\boldsymbol{Z}$ is an instrument and $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)$ is correct, as shown in Wang et al. (2014), $\hat{\boldsymbol{\theta}}_k$ obtained by (2.4) is consistent for $\boldsymbol{\theta}_k^0$ and asymptotically normal under some regularity conditions. When either $\boldsymbol{Z}$ or $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)$ is incorrect, the following lemma shows the property of $\hat{\boldsymbol{\theta}}_k$ under a misspecified model.

Lemma 1. *Assume the following regularity conditions:*

C1. (a) The dimension of $\boldsymbol{X}$, p, and the number of candidate models, K, remain fixed when the sample size $n \to \infty$; (b) The parameter space $\mathcal{A}$ for $\boldsymbol{\theta}_k$ is a compact set of $\mathcal{R}^{d_k}$ and $\boldsymbol{\theta}_k^*$ is the unique minimizer of $\|\boldsymbol{G}_k(\boldsymbol{\theta}_k)\|$ over $\boldsymbol{\theta}_k$, where $\boldsymbol{G}_k(\boldsymbol{\theta}_k) = E\{\boldsymbol{g}_k(Y, \boldsymbol{X}, \delta, \boldsymbol{\theta}_k)\}$ and $\|\cdot\|$ is l_2 norm; (c) $\sup_{\boldsymbol{\theta}_k} \|\boldsymbol{g}_k(Y, \boldsymbol{X}, \delta, \boldsymbol{\theta}_k)\| < \infty$; (d) The matrix $\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k^*) = E\{\boldsymbol{h}_k(\boldsymbol{X})^\top \delta[\partial \pi_k^{-1}(Y, \boldsymbol{U}, \boldsymbol{\theta}_k^*)/\partial \boldsymbol{\theta}_k]\}$ is of full rank and the matrix $\boldsymbol{W}_k(\boldsymbol{\theta}_k^*) = E\{\boldsymbol{g}_k(Y, \boldsymbol{X}, \delta, \boldsymbol{\theta}_k^*)\boldsymbol{g}_k(Y, \boldsymbol{X}, \delta, \boldsymbol{\theta}_k^*)^\top\}$ is positive definite;

C2. (a) $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)$ is twice differentiable with respect to $\boldsymbol{\theta}_k$; (b) $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k^*)$

$\geq C > 0$ *for* $k = 1, ..., K$*; (c)* $\partial\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)/\partial\boldsymbol{\theta}_k$ *is uniformly bounded.*

Then, as $n \to \infty$*,*

$$n^{1/2}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*) \to N(0, \{\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k^*)^\top \boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*)\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k^*)\}^{-1}) \qquad \textit{in distribution.}$$

In the presence of misspecification, the proposed $\hat{\boldsymbol{\theta}}_k$ consistently estimates $\boldsymbol{\theta}_k^*$ minimizing the population version of the empirical generalized moment discrepancy. If model $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)$ is correct, then $\hat{\boldsymbol{\theta}}_k$ is consistent for the true parameter vector $\boldsymbol{\theta}_k^0$.

Define

$$F_k(\boldsymbol{x}) = E\Big\{\frac{\delta \mathrm{I}(\boldsymbol{X} \leq \boldsymbol{x})}{\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k^*)}\Big\} = E\Big[E\Big\{\frac{\pi(Y, \boldsymbol{U})}{\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k^*)}\Big\}\mathrm{I}(\boldsymbol{X} \leq \boldsymbol{x})\Big].$$

Since $\hat{\boldsymbol{\theta}}_k \to \boldsymbol{\theta}_k^*$ in probability, it can be verified that $\hat{F}_k(\boldsymbol{x}) \to F_k(\boldsymbol{x})$ in probability. Define

$$\Delta_k = E|F_k(\boldsymbol{X}) - F(\boldsymbol{X})|.$$

Then, VC(k) defined in (2.5) converges in probability to Δ_k. If $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)$ is a correct model and $\boldsymbol{Z}$ is an instrument, then $F_k(\boldsymbol{x}) = F(\boldsymbol{x})$ and $\Delta_k = 0$. If $\Delta_k > 0$ for any wrong model k, then we can distinguish a wrong model from a correct model.

The order of λ tending to 0 as $n \to \infty$ determines the asymptotic behavior of the proposed model selection procedure. Without loss of generality, we

assume that the most compact correct model is $\pi_1(Y, \boldsymbol{U}, \boldsymbol{\theta}_1)$. The model selection procedure is consistent as $n \to \infty$ if and only if

$$\Pr\{\mathrm{PVC}_\lambda(k) > \mathrm{PVC}_\lambda(1)\} \to 1 \tag{2.8}$$

holds for any model k with $k > 1$. Suppose that $\Delta_k > 0$ when $\pi_k(Y, \boldsymbol{U}_k, \boldsymbol{\theta}_k)$ is a wrong model. To achieve (2.8), we need λ satisfying $\lambda(\log d_1 - \log d_k) < \mathrm{VC}(k) - \mathrm{VC}(1)$; since $\mathrm{VC}(k) - \mathrm{VC}(1) \to \Delta_k > 0$ and d_k may be smaller than d_1, we need $\lambda \to 0$ as $n \to \infty$. Next, let $\pi_k(Y, \boldsymbol{U}_k, \boldsymbol{\theta}_k)$ be a correct model that is overfitted such that $d_k > d_1$. In this case, we need to find a λ such that $\lambda > \{\mathrm{VC}(1) - \mathrm{VC}(k)\}/(\log d_k - \log d_1)$ with probability tending to 1. Since both $\mathrm{VC}(1)$ and $\mathrm{VC}(k)$ converge to 0 under correct models, we need to choose λ converging to 0 at a rate slower than that of $\mathrm{VC}(1) - \mathrm{VC}(k)$. The following lemma proved in the Appendix provides the convergence rate of $\mathrm{VC}(1) - \mathrm{VC}(k)$.

Lemma 2. *Under Conditions in Lemma 1, if $\pi_1(Y, \boldsymbol{U}, \boldsymbol{\theta}_1)$ is the most compact correct model and $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)$ is an overfitted correct model, then $VC(1) - VC(k) = O_p(n^{-1/2})$.*

This result and the previous discussion actually establishes the following result about the consistency of the proposed selection for propensity and instrument.

Theorem 1. *Assume that $\mathcal{M}$ contains a correctly specified propensity model for $\pi(Y, \boldsymbol{U})$ with an instrument $\boldsymbol{Z}$ satisfying (1.1). Under the regularity conditions in Lemmas 1 and 2, if λ in (2.6) is chosen such that $\lambda \to 0$ and $n^{1/2}\lambda \to \infty$, then (2.8) holds, i.e., $\Pr(\hat{k} = 1) \to 1$ as $n \to \infty$, where model 1 is assumed to be the most compact correct model.*

In practice, we propose to use $\lambda = Cn^{-1/2}(\log\log n)^{1/2}$ with a constant C chosen by Cross-Validation (CV). Specifically, we randomly split the set $\{1, ..., n\}$ into J nonoverlapping subsets $\{S_1, \ldots, S_J\}$ of roughly equal sizes $n_1, ..., n_J$. For each $j = 1, ..., J$ and a given C, using all data from $i \notin S_j$, we compute

$$\mathrm{PVC}_{-j}(k) = (n - n_j)^{-1} \sum_{i \notin S_j} |\hat{F}_k(\boldsymbol{X}_i) - \hat{F}(\boldsymbol{X}_i)| + \lambda \log(d_k),$$

$$\hat{k}_{-j} = \mathrm{argmin}_{1 \le k \le K} \mathrm{PVC}_{-j}(k).$$

For a fixed C, we compute the error on the validation set S_j as

$$e_j(C) = \frac{1}{n_j} \sum_{i \in S_j} |\hat{F}_{\hat{k}_{-j}}(\boldsymbol{X}_i) - \hat{F}(\boldsymbol{X}_i)|,$$

and then choose the value C to be $\hat{C}$ that minimizes the average error over all subsets, i.e.,

$$\hat{C} = \arg\min_C \frac{1}{J} \sum_{j=1}^{J} e_j(C). \tag{2.9}$$

The collection $\mathcal{M}$ may be all possible decompositions of $\boldsymbol{X} = (\boldsymbol{U}, \boldsymbol{Z})$, which means at least $K = 2^p - 1$ models when the dimension of $\boldsymbol{X}$ is p. The grid search over all models may be computationally not feasible for a moderate p, e.g., $p \geq 7$. For the purpose of searching for a correct propensity and an instrument, however, it is not necessary to do grid search. Here, we propose a forward instrument selection procedure which can handle a moderate p. Consider a p-dimensional covariates $\boldsymbol{X} = (X_1, \ldots, X_p)$ and a fixed parametric function $\pi(Y, \boldsymbol{U}, \theta)$ (e.g., a logistic function). We follow the following steps.

(i) Start with p models with $\boldsymbol{Z} = X_j$ and $\boldsymbol{U} = (X_t : t \neq j)$, $j = 1, ..., p$. Search through all p models and pick the one with the lowest PVC. Denote the resulting $\boldsymbol{Z}_1^* = X_{1^*}$ and $\boldsymbol{U}_1^* = (X_t : t \neq 1^*)$.

(ii) Consider next $p-1$ models with $\boldsymbol{Z} = (X_{1^*}, X_j)$ and $\boldsymbol{U} = (X_t : t \neq j, t \neq 1^*)$, $j = 1, ..., p$, $j \neq 1^*$. Search through all $p - 1$ models and pick the one with the lowest PVC. If this PVC value is higher than that in step 1, then stop and the model selected is $\pi(Y, \boldsymbol{U}_1^*, \theta)$. Otherwise, denote the resulting $\boldsymbol{Z}_2^* = (X_{1^*}, X_{2^*})$ and $\boldsymbol{U}_2^* = (X_t : t \neq 1^*, t \neq 2^*)$, and continue to the next step.

(iii) At the kth step, consider $p-k+1$ models with $\boldsymbol{Z} = (X_{1^*}, ..., X_{(k-1)^*}, X_j)$ and $\boldsymbol{U} = (X_t : t \neq j, t \neq 1^*, ..., t \neq (k-1)^*)$, $j = 1, ..., p$, $j \neq 1^*, ..., j \neq$

$(k-1)^*$. Search through all $p-k+1$ models and pick the one with the lowest PVC. If this PVC value is higher than that in step $k-1$, then stop and the model selected is $\pi(Y, \boldsymbol{U}_{(k-1)^*}, \theta)$. Otherwise, continue until $k = p$.

The number of models considered in this procedure is at most $p+(p-1)+\cdots+2+1 = p(p+1)/2$. Further, if we want select $\pi(Y, \boldsymbol{U}, \theta)$ between logistic and probit models or adding a nonlinear term Y^2 in the linear combination of X_j's and Y, we can apply the previous idea and establish a similar multi-step procedure. An asymptotic result similar to that in Theorem 1 can also be established.

3. Simulation Studies

Under assumptions (1.1) and (1.2A), we study in this section by simulation the finite sample performance of the proposed method in terms of the rate of selecting the most compact correct model, as well as the bias and root mean squared error (RMSE) of the resulting inverse propensity weighting estimator $\hat{\mu}_{pvc}$ defined in (2.7). All results are based on 1,000 simulation replications.

In simulation 1, selecting a model from seven models in (2.1) is considered, where $\boldsymbol{X} = (X_1, X_2, X_3)$ is generated from a 3-dimensional normal distribution with mean 1, and covariance $\mathrm{Cov}(X_j, X_{j'}) = 0.5$ for $1 \le j < j' \le 3$ and

$\mathrm{Var}(X_j) = 1$, $Y = X_1^2 + X_2^2 + X_3^2 + \varepsilon$ with ε from $N(0, 2)$, and ε is independent of $\boldsymbol{X}$. For convenience, we denote the seven models in (2.1) by M_0, $M_1(X_1)$, $M_1(X_2)$, $M_1(X_3)$, $M_2(X_1, X_2)$, $M_2(X_1, X_3)$ and $M_2(X_2, X_3)$, respectively, where the subscript s in $M_s(\boldsymbol{U})$ is the dimension of $\boldsymbol{U}$; for example, M_0 is the model with $\boldsymbol{U} = 0$ and $\boldsymbol{Z} = (X_1, X_2, X_3)$; $M_2(X_1, X_3)$ is the model with a 2-dimensional $\boldsymbol{U} = (X_1, X_3)$ and $\boldsymbol{Z} = X_2$.

Given $(Y, \boldsymbol{X})$, we generate δ from the Bernoulli distribution with the logistic function in (2.1) as the probability and parameter vector $\boldsymbol{\theta}^0 = (-0.4, -0.3)$ for M_0, $\boldsymbol{\theta}^0 = (-0.8, 1.2, -0.3)$ for $M_1(X_j)$ with $j = 1, 2, 3$, and $\boldsymbol{\theta}^0 = (-0.8, 1.2, 1.2, -0.3)$ for $M_2(X_j, X_{j'})$, $1 \le j < j' \le 3$. The coefficients in the propensity models are chosen so that the unconditional rates of missing data are between 20% and 40%. As in Wang et al. (2014), we use $\boldsymbol{h}_k(\boldsymbol{X}) = (1, \boldsymbol{U}, \boldsymbol{Z})$ in (2.2) and (2.4) to obtain the GMM estimator $\hat{\boldsymbol{\theta}}_k$.

If the true propensity $\pi(Y, \boldsymbol{U}) = M_1(X_1)$, then $\boldsymbol{Z} = (X_2, X_3)$ is an instrument; models M_0, $M_1(X_2)$, $M_1(X_3)$, and $M_2(X_2, X_3)$ are wrong; $M_2(X_1, X_2)$ and $M_2(X_1, X_3)$ are also correct propensity models with $\boldsymbol{Z} = X_3$ and $\boldsymbol{Z} = X_2$ as instruments, respectively. Since both $M_2(X_1, X_2)$ and $M_2(X_1, X_3)$ are overfitted, the penalty term in (2.6) forces us to choose $M_1(X_1)$ more frequently. The discussion is similar if $M_1(X_2)$ or $M_1(X_3)$ is correct. If the true propensity $\pi(Y, \boldsymbol{U}) = M_2(X_1, X_2)$, then $M_2(X_1, X_2)$ is the only correct propensity

model and $\boldsymbol{Z} = X_3$ is the only correct instrument. Finally, if $\pi(Y, \boldsymbol{U}) = M_0$, then all models are correct and M_0 is the most compact model with $\boldsymbol{Z} = \boldsymbol{X}$.

For n= 300, 500, and 1,000, we implement the penalized validation criterion in (2.6) with a 10-folds CV method to determine the tuning parameter λ as discussed at the end of Section 2, where the range for the minimization in (2.9) is $(0.1, 20)$. Table 1 reports the rates in 1,000 Monte Carlo replications of selecting each model by the proposed penalized validation criterion under different best models, i.e., the most compact correct models. It can be seen that the proposed method selects the best model most of the time, i.e., the simulation rates of selecting the best model are very high when the sample size $n = 300$ and are close to 1 when $n = 500$ or 1,000.

Put Tables 1-2 about here.

Besides model selection, we also consider the estimation of $\mu = E(Y)$ after model and instrument selection, using the proposed estimator $\hat{\mu}_{pvc}$ defined in (2.7) based on the selected and estimated propensity. To compare, we also study the following three estimators, $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$, the sample mean when there is no missing data, which is used as a benchmark; $\hat{\mu}_{cc} = \sum_{i=1}^{n} \delta_i Y_i / \sum_{i=1}^{n} \delta_i$, the sample mean of observed Y data, which is a biased estimator; and the inverse propensity weighting estimators $\hat{\mu}_k = \sum_{i=1}^{n} \delta_i Y_i / \pi_k(Y_i, \boldsymbol{U}_i, \hat{\boldsymbol{\theta}}_k)$, $k = 0, \ldots, 6$, which is different from $\hat{\mu}_{pvc}$ in (2.7) be-

cause $\hat{\mu}_k$ uses a fixed propensity without model selection and may be biased when the propensity model is wrong. The mean μ is 6 in all cases.

Due to symmetry, the simulation results when $M_1(X_2)$ or $M_1(X_3)$ is the best are about the same as the results when $M_1(X_1)$ is the best, and the results when $M_2(X_1, X_3)$ or $M_2(X_2, X_3)$ is the best are about the same as those when $M_2(X_1, X_2)$ is the best. Hence, we only present the results when M_0, $M_1(X_1)$ and $M_2(X_1, X_2)$ are the best. Table 2 shows the biases and RMSE of point estimators based on different methods. In terms of bias and RMSE, when M_0 is the best, it can be seen that the proposed PVC estimator and the estimator based on the seven propensity model are comparable. When $M_1(X_1)$ is the best, it can be seen that the proposed PVC estimator and the estimator based on $M_1(X_1)$ are comparable, which have negligible biases and slightly larger RMSEs than that of $\bar{Y}$ in all cases; as expected, the estimators based on $M_1(X_1, X_2)$ and $M_1(X_1, X_3)$ are also unbiased but less efficient; the estimators based on observed Y values and other propensity models have larger biases and RMSE, which agrees with our theory. Similar results are obtained when $M_2(X_1, X_2)$ is the best. In this case, since only $M_2(X_1, X_2)$ is correct, it further can be seen that $\hat{\mu}_k$ based on incorrect models even have much larger biases and RMSE than those of the method based on observed Y values, which is consistent with findings in Shao and Wang (2016) and is our

motivation in studying model and instrument selection.

Our simulation 2 is used to check the performance of proposed method in selecting Y or Y^2 in the logistic propensity model, i.e., in addition to the seven candidate models in (2.1), we further include the following seven candidate models,

$$\tilde{M}_0 = 1/\{1+\exp(\alpha+\gamma Y^2)\}, \quad \boldsymbol{\theta} = (\alpha,\gamma)$$
$$\tilde{M}_1(X_j) = 1/\{1+\exp(\alpha+\beta X_j+\gamma Y^2)\}, \quad \boldsymbol{\theta} = (\alpha,\beta,\gamma)$$
$$\tilde{M}_2(X_j,X_{j'}) = 1/\{1+\exp(\alpha+\beta_1 X_j+\beta_2 X_{j'}+\gamma Y^2)\}, \quad \boldsymbol{\theta} = (\alpha,\beta_1,\beta_2,\gamma)$$

with $\boldsymbol{\theta}^0 = (0.8,-0.1)$ for $\tilde{M}_0$, $\boldsymbol{\theta}^0 = (-0.8,2,-0.3)$ for $\tilde{M}_1(X_j)$ and $\boldsymbol{\theta}^0 = (-0.8,1.5,1.5,-0.1)$ for $\tilde{M}_2(X_j,X_{j'})$ in the simulation, $1 \le j \le j' \le 3$.

If the true propensity $\pi(Y,\boldsymbol{U}) = M_0$, $\boldsymbol{X}$ is an instrument and $\boldsymbol{Z}$ = sub-vectors of $\boldsymbol{X}$ under all other models are correct instruments, although M_0 is the most compact model. Also, models $M_1(X_j)$ and $M_2(X_j,X_{j'})$ are correct, but $\tilde{M}_0$, $\tilde{M}_1(X_j)$ and $\tilde{M}_2(X_j,X_{j'})$ are incorrect because of using Y^2 instead of Y; the discussion is similar if $\tilde{M}_0$ is correct. If $\pi(Y,\boldsymbol{U}) = M_1(X_1)$, only models $M_1(X_1)$, $M_2(X_1,X_2)$, $M_2(X_1,X_3)$, $\tilde{M}_1(X_1)$, $\tilde{M}_2(X_1,X_2)$ and $\tilde{M}_2(X_1,X_3)$, give correct instruments $\boldsymbol{Z} = X_2$, X_3 or (X_2,X_3); the rest models do not give correct instruments. However, $\tilde{M}_1(X_1)$, $\tilde{M}_2(X_1,X_2)$ and $\tilde{M}_2(X_1,X_3)$ are wrong models; $M_2(X_1,X_2)$ and $M_2(X_1,X_3)$ are correct but overfitted. The discus-

sion is similar if $M_1(X_j)$ or $\tilde{M}_1(X_j)$ is correct. If $\pi(Y, \boldsymbol{U}) = M_2(X_1, X_2)$, only $M_2(X_1, X_2)$ and $\tilde{M}_2(X_1, X_2)$ give correct instrument $\boldsymbol{Z} = X_3$, but $\tilde{M}_2(X_1, X_2)$ is incorrect. The discussion is similar if $M_2(X_j, X_{j'})$ or $\tilde{M}_2(X_j, X_{j'})$ is correct.

The model selection probabilities and estimation results of $\mu = E(Y)$ are shown in Figure 1 and Table 3, respectively. Figure 1 shows that our proposed method performs well on selecting the best propensity model and instrument simultaneously. Compared with the results in Table 1, the selection rates for the best model decrease a little bit when M_0 or $M_1(X_j)$ is the best, but they are close to 1 when $n = 1,000$.

Put Figure 1 and Table 3 about here.

For the seven propensity models using Y, due to symmetry, we only present the results when M_0, $M_1(X_1)$, $M_2(X_1, X_2)$ are the best. Table 3 shows the biases and RMSE of point estimators based on different methods. In terms of bias and RMSE, when M_0 is the best, it can be seen that the proposed PVC estimator and the estimator based on the seven propensity model using Y are comparable. When $M_1(X_1)$ is the best, the proposed PVC estimator and the estimators based on $M_1(X_1)$ and $\tilde{M}_1(X_1)$ are comparable, which have negligible biases and slightly larger RMSEs than that of $\bar{Y}$ in all cases; the estimators based on $M_2(X_1, X_2)$, and $M_2(X_1, X_3)$ are also unbiased, but $\tilde{M}_2(X_1, X_3)$ and $\tilde{M}_2(X_1, X_2)$ have much larger RMSEs using Y^2; the esti-

mators based on observed Y values and other propensity models have larger biases and RMSE, which agrees with our theory. When $M_2(X_1, X_2)$ is the best, the proposed PVC estimator and the estimator based on $M_2(X_1, X_2)$ are comparable, which has negligible biases and slightly larger RMSEs than that of $\bar{Y}$ in all cases; it further can be seen that $\hat{\mu}_k$ based on incorrect models even have much larger biases and RMSE than those of the method based on observed Y values. For the seven propensity models using Y^2, the conclusions are similar and hence the results are omitted.

Our last simulation studies the forward instrument selection procedure discussed in the end of Section 3 when the dimension of $\boldsymbol{X}$ is 10. As in simulation 1, $\boldsymbol{X} = (X_1, X_2, \ldots, X_{10})$ is generated from the 10-dimensional normal distribution with mean 1 and covariance $\mathrm{Cov}(X_j, X_{j'}) = 0.5$ for $1 \leq j < j' \leq 10$ and $\mathrm{Var}(X_j) = 1$, Y is generated from

$$Y = X_1^2 + X_2^2 + \ldots + X_{10}^2 + \varepsilon,$$

with ε from $N(0, 2)$ and independent of $\boldsymbol{X}$. We denote all possible ten models in (2.1) by M_0 and $M_j(\boldsymbol{U}_j) = 1/\{1 + \exp(\alpha + \beta^T \boldsymbol{U}_j + \gamma Y)\}$ with $\boldsymbol{U}_j = (X_1, \ldots, X_j)$ for $j = 1, \ldots, 9$, respectively, and consider that $\boldsymbol{\theta}^0 = (0.2, -0.2)$ for M_0, $\boldsymbol{\theta}^0 = (-0.8, 0.8, -0.2)$ for $M_1(U_1)$, $\boldsymbol{\theta}^0 = (-0.8, 0.8, 0.8, -0.2)$ for $M_2(\boldsymbol{U}_2)$, $\boldsymbol{\theta}^0 = (-0.8, 0.8, 0.8, 0.8, -0.4)$ for $M_3(\boldsymbol{U}_3)$, $\boldsymbol{\theta}^0 = (-0.8, 1, 1, 1, 1, -0.6)$ for $M_4(\boldsymbol{U}_4)$, $\boldsymbol{\theta}^0 = (-2, 0.8, \ldots, 0.8, -0.4)$ for $M_5(\boldsymbol{U}_5)$, $\boldsymbol{\theta}^0 = (-2, 0.8, \ldots, 0.8, -0.4)$

for $M_6(\boldsymbol{U}_6)$, $\boldsymbol{\theta}^0 = (-0.8, 0.8, \ldots, 0.8, -0.8)$ for $M_7(\boldsymbol{U}_7)$, $\boldsymbol{\theta}^0 = (-1, 0.8, \ldots, 0.8, -0.6)$ for $M_8(\boldsymbol{U}_8)$, $\boldsymbol{\theta}^0 = (-2, 0.8, \ldots, 0.8, -0.6)$ for $M_9(\boldsymbol{U}_9)$. For $n = 1,000$ or 1,500, we compute the simulation rates of the forward instrument selection procedure in selecting the correct, best (the most compact correct), and wrong models, respectively. The results shown in Table 4 indicate that the forward instrument selection procedure performs well, especially when the dimension of $\boldsymbol{U}$ is small.

Put Table 4 about here.

4. Real data example

We consider a data set from the National Health and Nutrition Examination Survey (NHNES) conducted in 2005 by the United States Centers for Disease Control and Prevention, which was designed to assess the health and nutritional status of adults and children in the United States. The data are available at www.cdc.gov/nchs/nhanes.htm.

As in Fang and Shao (2016), we consider body fat percentage measured by dual-energy X-ray absorptiometry (dxa) as the response variable Y, body mass index (bmi), gender, and age as covariates, i.e., $\boldsymbol{X}$ = (bmi, gender, age), and middle-aged and old people (age ≥ 45) with a total of $n = 1591$ subjects, 393 (24.7%) of which have missing Y data.

As in the first simulation study in Section 3, we consider the seven candidate propensity models in (2.1) based on the assumption that the underlying true propensity model is a logistic model linear in Y and $\boldsymbol{X}$, and implement the proposed method to select an instrument $\boldsymbol{Z}$. Thus, we have seven choices of instrument: $\boldsymbol{Z}$ = bmi, $\boldsymbol{Z}$ = gender, $\boldsymbol{Z}$ = age, $\boldsymbol{Z}$ = (bmi, gender), $\boldsymbol{Z}$ = (bmi, age), $\boldsymbol{Z}$ = (gender, age), and $\boldsymbol{Z}$ = (bmi, gender, age). For each choice of $\boldsymbol{Z}$, the covariates not included in $\boldsymbol{Z}$ are treated as $\boldsymbol{U}$. We use the proposed PVC in (2.6) with a tuning parameter $\hat{\lambda}$ obtained from 10-folds cross-validation. For each candidate propensity model, values of the proposed PVC, estimate of population mean of dxa and its standard error based on the bootstrap with 200 replications are given in Table 4. The proposed PVC method selects M_1(bmi), i.e., $\boldsymbol{Z}$ =(gender, age), which is consistent with the results in Fang and Shao (2016) obtained under a different setting where $F_{Y|\boldsymbol{X}}$ is parametric and the propensity $\pi(Y, \boldsymbol{U})$ is unspecified. Among the seven choices of instruments, the mean estimates based on $\boldsymbol{Z}$ = bmi, $\boldsymbol{Z}$ = (bmi, age), $\boldsymbol{Z}$ = (bmi, gender) and $\boldsymbol{Z}$ = (bmi, age, gender) are different from the proposed mean estimate based on $\boldsymbol{Z}$ = (gender, age), which indicates that these are wrong choices of instruments. On the other hand, as we mentioned in Section 2, $\boldsymbol{Z}$ = age and $\boldsymbol{Z}$ = gender are correct choices of instruments if $\boldsymbol{Z}$ = (gender, age) is an instrument; they provide similar mean estimates, but the estimates based

on $\boldsymbol{Z}$ = age and $\boldsymbol{Z}$ = gender have much larger SEs. This result indicates that, for the efficiency of mean estimator, we should select an instrument with the largest possible dimension.

Put Table 5 about here.

As in the second simulation study in Section 3, we further include the seven candidate propensity models that are logistically linear in $(Y^2, \boldsymbol{X})$. The results are also presented in Table 5. The proposed PVC method still selects $M_1(\text{bmi})$ from all the fourteen candidate models, indicating that the propensity models with a term Y^2 but not Y are wrong.

The United States Centers for Disease Control and Prevention indicated that, in this problem, missing responses may not be ingorable through examination of missing items in the data files. To see the effect of addressing nonignorable nonresponse, we computed estimates of $E(Y)$ based on the assumption of ignorable nonresponse by excluding the Y term in the logistic propensity previously discussed. The resulting models are denoted by $M_s^{-Y}(\boldsymbol{U})$ etc. For example, $M_1^{-Y}(\text{bmi}) = 1/\{1+\exp(\alpha+\beta\times\text{bmi})\}$. The results are included in Table 5. Note that the estimate under M_0^{-Y} is equal to the sample mean of observed Y values, which is 34.44. Regardless of which ignorable propensity model is used, all estimates of $E(Y)$ are between 34.17 and 34.68, which are close to the sample mean of observed Y data. Thus, in this example we do see

some effect of addressing nonignoraable nonresponse, although no one knows the truth in a real dataset.

5. Discussion

Handling nonignorable nonresponse is a challenging problem, mainly due to the issue of identifiability of the nonresponse propensity. A nonresponse instrument plays a crucial role in identifiability, but it is often assumed to be given in the existing literature. Further, to obtain consistent estimators, imposed parametric propensity model is also needed to be verified. In this paper, we propose a simultaneous propensity model and instrument selection criterion in the presence of nonignorable nonresponse. We showed that the proposed method can consistently select the most compact correct parametric propensity model and instrument from a group of candidate models, if one of these candidate models is correct and an instrument exists. The simulation studies and data analysis show that the proposed method has good performances.

It can be seen that the proposed method based on (1.1), (1.2A), (2.2)-(2.6) can be extended to the situiation where Y is multivariate (with δ changed to a vector of indicators) or the situation where both Y and $\boldsymbol{X}$ have missing data. For illustration, we consider the situation where $\boldsymbol{Z}$ is always observed, and $\boldsymbol{U}$ and Y have missing values. Let δ_Y and δ_U be the indicators of observing Y

and $\boldsymbol{U}$, respectively. Instead of (1.1), we assume that

$$\Pr(\delta_Y = t, \delta_U = s|Y, \boldsymbol{U}, \boldsymbol{Z}) = \Pr(\delta_Y = t, \delta_U = s|Y, \boldsymbol{U}),$$

where $t = 0, 1$, and s is a vector of 0-1 values, and consider the collection of all K parametric models $\mathcal{M} = \{\pi_k(Y, \boldsymbol{U}_k, \boldsymbol{\theta}_k),\ k = 1, \ldots, K\}$ for $\Pr(\delta_Y = t, \delta_U = s|Y, \boldsymbol{U})$. The proposed penalized validation criterion can be adopted.

There are several limitations of our proposed method. First, our proposed method is for small or moderate p. For high dimensional covariates, model selection and instrument search with nonignorable nonresponse is challenging. One possible solution is to apply a proper variable/feature screening method first to reduce the dimensionality of covariates and then use the proposed method based on the reduced small or moderate number of covariates. Second, stepwise selection procedure itself has some limitations. For example, the order of covariate entry and the number of covariates may affect the selected model. Third, the proposed method relies on the assumption that one of the candidate models is correct and an instrument exists. In practice, we may consider a number of potential candidate models and try to make sure that one of the candidate models is correct. When all candidate models are incorrect or no instrument exists, we may only derive some procedures that are approximately valid. But research on the situation where no correct candidate model or instrument exists is still interesting although challenging, since we

all know that no model is perfect in practical applications. All these issues will be explored in our further research.

Acknowledgement

We are grateful to the editor, the associate editor, and two anonymous referees for their insightful comments and suggestions, which have led to significant improvements. Our research was supported by the National Natural Science Foundation of China (11831008, 11501208, 11871287, 11601156), the Natural Science Foundation of Tianjin (18JCYBJC41100), the Fundamental Research Funds for the Central Universities, the Key Laboratory for Medical Data Analysis and Statistical Research of Tianjin, the Chinese 111 Project (B14019), and the U.S. National Science Foundation (DMS-1612873).

Appendix

Proof of Lemma 1: Recall

$$\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) = n^{-1}\sum_{i=1}^{n}\boldsymbol{g}_k(Y_i, \boldsymbol{X}_i, \delta_i, \boldsymbol{\theta}_k) \text{ and } \boldsymbol{G}_k(\boldsymbol{\theta}_k) = E\{\boldsymbol{g}_k(Y, \boldsymbol{U}, \delta, \boldsymbol{\theta}_k)\}.$$

By the law of large number (LLN), it can be shown that $\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) - \boldsymbol{G}_k(\boldsymbol{\theta}_k) = o_p(1)$ for all $\boldsymbol{\theta}_k \in \mathcal{A}$. Since both $\boldsymbol{g}_k(Y, \boldsymbol{U}, \delta, \boldsymbol{\theta}_k)$ and $\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k)$ are continuous at

each $\boldsymbol{\theta}_k \in \mathcal{A}$,

$$\sup_{\boldsymbol{\theta}_k \in \mathcal{A}} \|\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) - \boldsymbol{G}_k(\boldsymbol{\theta}_k)\| = o_p(1).$$

This, coupled with GMM identification (i.e., Lemma 2.3 of Newey and McFadden, 1994), shows that the first-step estimator

$$\tilde{\boldsymbol{\theta}}_k = \boldsymbol{\theta}_k^* + o_p(1).$$

By the LLN, it can be shown that $\hat{\boldsymbol{W}}_{kn}^{-1} = \boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*) + o_p(1)$. Let

$$Q_k(\boldsymbol{\theta}_k) = \boldsymbol{G}_k(\boldsymbol{\theta}_k)^\top \boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*)\boldsymbol{G}_k(\boldsymbol{\theta}_k) \text{ and } \bar{Q}_k(\boldsymbol{\theta}_k) = \bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k)^\top \hat{\boldsymbol{W}}_{kn}^{-1}\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k).$$

Based on Lemma 2.3 and Theorem 2.1 of Newey and McFadden (1994), to prove $\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^* = o_p(1)$, it is enough to show that

$$\sup_{\boldsymbol{\theta}_k \in \mathcal{A}} |\bar{Q}_{kn}(\boldsymbol{\theta}_k) - Q_k(\boldsymbol{\theta}_k)| = o_p(1).$$

Using the triangle and Cauchy-Schwartz inequalities, we have

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}_k \in \mathcal{A}} |\bar{Q}_{kn}(\boldsymbol{\theta}_k) - Q_k(\boldsymbol{\theta}_k)| \\
&\quad \le \sup_{\boldsymbol{\theta}_k \in \mathcal{A}} \left| \{\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) - \boldsymbol{G}_k(\boldsymbol{\theta}_k)\}^\top \hat{\boldsymbol{W}}_{kn}^{-1} \{\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) - \boldsymbol{G}_k(\boldsymbol{\theta}_k)\} \right| \\
&\qquad + \sup_{\boldsymbol{\theta}_k \in \mathcal{A}} \left| \boldsymbol{G}_k(\boldsymbol{\theta}_k)^\top (\hat{\boldsymbol{W}}_{kn}^{-1} + (\hat{\boldsymbol{W}}_{kn}^{-1})^\top) \{\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) - \boldsymbol{G}_k(\boldsymbol{\theta}_k)\} \right| \\
&\qquad + \sup_{\boldsymbol{\theta}_k \in \mathcal{A}} \left| \boldsymbol{G}_k(\boldsymbol{\theta}_k)^\top (\hat{\boldsymbol{W}}_{kn}^{-1} - \boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*)) \boldsymbol{G}_k(\boldsymbol{\theta}_k) \right| \\
&\quad \le \sup_{\boldsymbol{\theta}_k \in \mathcal{A}} \left\| \bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) - \boldsymbol{G}_k(\boldsymbol{\theta}_k) \right\|^2 \|\hat{\boldsymbol{W}}_{kn}^{-1}\| \\
&\qquad + 2 \sup_{\boldsymbol{\theta}_k \in \mathcal{A}} \left\| \boldsymbol{G}_k(\boldsymbol{\theta}_k) \right\| \left\| \bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) - \boldsymbol{G}_k(\boldsymbol{\theta}_k) \right\| \|\boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*)\| \\
&\qquad + \sup_{\boldsymbol{\theta}_k \in \mathcal{A}} \left\| \boldsymbol{G}_k(\boldsymbol{\theta}_k) \right\|^2 \|\hat{\boldsymbol{W}}_{kn}^{-1} - \boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*)\| = o_p(1).
\end{aligned}$$

Thus, we prove that $\hat{\boldsymbol{\theta}}_k = \boldsymbol{\theta}_k^* + o_p(1)$.

Next, we derive the asymptotic normality of $\hat{\boldsymbol{\theta}}_k$. With probability approaching one, we have the first-order condition

$$2\boldsymbol{\Gamma}_k(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{W}}_{kn}^{-1}\bar{\boldsymbol{G}}_{kn}(\hat{\boldsymbol{\theta}}_k) = 0,$$

where $\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k) = \partial\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k)/\partial\boldsymbol{\theta}_k$. Expanding $\bar{\boldsymbol{G}}_{kn}(\hat{\boldsymbol{\theta}}_k)$ around $\boldsymbol{\theta}_{k^*}$, we have

$$n^{1/2}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*) = -[\boldsymbol{\Gamma}_k^\top(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{W}}_{kn}^{-1}\boldsymbol{\Gamma}_k(\check{\boldsymbol{\theta}}_k)]^{-1}\boldsymbol{\Gamma}_k^\top(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{W}}_{kn}^{-1}n^{1/2}\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k^*),$$

where $\check{\boldsymbol{\theta}}_k$ is between $\hat{\boldsymbol{\theta}}_k$ and $\boldsymbol{\theta}_k^0$. By simple calculation and the LLN, for all $\boldsymbol{\theta}_k \in \mathcal{A}$,

$$\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k) = E\Big\{h_k(\boldsymbol{X})^\top\delta\frac{\partial\pi_k(Y,\boldsymbol{U},\boldsymbol{\theta}_k)^{-1}}{\partial\boldsymbol{\theta}_k}\Big\} + o_p(1).$$

This, together with $\hat{\boldsymbol{W}}_{kn}^{-1} = \boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*) + o_p(1)$ and $\hat{\boldsymbol{\theta}}_k = \boldsymbol{\theta}_k^* + o_p(1)$, implies that

$$\begin{aligned}&[\boldsymbol{\Gamma}_k^\top(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{W}}_{kn}^{-1}\boldsymbol{\Gamma}_k(\check{\boldsymbol{\theta}}_k)]^{-1}\boldsymbol{\Gamma}_k^\top(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{W}}_{kn}^{-1}\\ =&[\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k^*)^\top\boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*)\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k^*)]^{-1}\Gamma_k(\boldsymbol{\theta}_k^*)^\top\boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*) + o_p(1).\end{aligned}$$

By the Slutzky theorem, we can show

$$n^{1/2}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*) \xrightarrow{\mathcal{L}} N(0, (\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k^*)^\top\boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*)\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k^*))^{-1}).$$

Particularly, when the intermittent propensity model is correctly specified, $\pi_k(Y,\boldsymbol{U},\boldsymbol{\theta}_k^*) = \pi_k(Y,\boldsymbol{U},\boldsymbol{\theta}_k^0)$, $\hat{\boldsymbol{W}}_{kn} = \boldsymbol{W}_k(\boldsymbol{\theta}_k^0) + o_p(1)$ and $\hat{\boldsymbol{\theta}}_k = \boldsymbol{\theta}_k^0 + o_p(1)$.

Proof of Lemma 2: Notice

$$|\mathrm{VC}(1) - \mathrm{VC}(k)| \le \frac{1}{n}\sum_{i=1}^{n} |\hat{F}_1(\boldsymbol{X}_i) - \hat{F}_k(\boldsymbol{X}_i)|.$$

We just need to show $n^{-1/2}\sum_{i=1}^{n} |\hat{F}_1(\boldsymbol{X}_i) - \hat{F}_k(\boldsymbol{X}_i)| = O_p(1)$. Note that

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n} |\hat{F}_k(\boldsymbol{X}_i) - \hat{F}_1(\boldsymbol{X}_i)| \\
=&n^{-1/2}\sum_{i=1}^{n} \Big| \frac{1}{n}\sum_{j=1}^{n} \delta_j \mathrm{I}(\boldsymbol{X}_j \le \boldsymbol{X}_i) \Big\{ \frac{1}{\pi_k(Y_j, \boldsymbol{U}_j, \hat{\boldsymbol{\theta}}_k)} - \frac{1}{\pi_1(Y_j, \boldsymbol{U}_j, \hat{\boldsymbol{\theta}}_1)} \Big\} \Big|
\end{aligned}$$

Let

$$\begin{aligned}
A_{ni}^{(1)} &= \frac{1}{n}\sum_{j=1}^{n} \delta_j \mathrm{I}(\boldsymbol{X}_j \le \boldsymbol{X}_i) \Big\{ \frac{1}{\pi_k(Y_j, \boldsymbol{U}_j, \hat{\boldsymbol{\theta}}_k)} - \frac{1}{\pi_k(Y_j, \boldsymbol{U}_j, \boldsymbol{\theta}_k^*)} \Big\}, \\
A_{ni}^{(2)} &= -\frac{1}{n}\sum_{j=1}^{n} \delta_j \mathrm{I}(\boldsymbol{X}_j \le \boldsymbol{X}_i) \Big\{ \frac{1}{\pi_1(Y_j, \boldsymbol{U}_j, \hat{\boldsymbol{\theta}}_1)} - \frac{1}{\pi_1(Y_j, \boldsymbol{U}_j, \boldsymbol{\theta}_1^*)} \Big\}, \\
A_{ni}^{(3)} &= \frac{1}{n}\sum_{j=1}^{n} \delta_j \mathrm{I}(\boldsymbol{X}_j \le \boldsymbol{X}_i) \Big\{ \frac{1}{\pi_k(Y_j, \boldsymbol{U}_j, \boldsymbol{\theta}_k^*)} - \frac{1}{\pi_1(Y_j, \boldsymbol{U}_j, \boldsymbol{\theta}_1^*)} \Big\}.
\end{aligned}$$

We have

$$\begin{aligned}
n^{-1/2}\sum_{i=1}^{n} |\hat{F}_k(\boldsymbol{X}_i) - \hat{F}_1(\boldsymbol{X}_i)| &= n^{-1/2}\sum_{i=1}^{n} |A_{ni}^{(1)} + A_{ni}^{(2)} + A_{ni}^{(3)}| \\
&\le n^{-1/2}\Big(\sum_{i=1}^{n} |A_{ni}^{(1)}| + \sum_{i=1}^{n} |A_{ni}^{(2)}| + \sum_{i=1}^{n} |A_{ni}^{(3)}| \Big).
\end{aligned}$$

For $n^{-1/2}\sum_{i=1}^{n}|A_{ni}^{(1)}|$, we have

$$
\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}|A_{ni}^{(1)}| \\
& \le n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big|\delta_j \mathrm{I}(\boldsymbol{X}_j \le \boldsymbol{X}_i)\Big\{\frac{1}{\pi_k(Y_j,\boldsymbol{U}_j,\hat{\boldsymbol{\theta}}_k)}-\frac{1}{\pi_k(Y_j,\boldsymbol{U}_j,\boldsymbol{\theta}_k^*)}\Big\}\Big| \\
& \le n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big|\frac{1}{\pi_k(Y_j,\boldsymbol{U}_j,\hat{\boldsymbol{\theta}}_k)}-\frac{1}{\pi_k(Y_j,\boldsymbol{U}_j,\boldsymbol{\theta}_k^*)}\Big| \\
& = n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big|\frac{\partial\pi_k^{-1}(Y_j,\boldsymbol{U}_j,\boldsymbol{\theta}_k^*)}{\partial\boldsymbol{\theta}_k}(\hat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k^*)+o_p(\hat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k^*)\Big| \\
& \le |\sqrt{n}(\hat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k^*)|\times\frac{1}{n}\sum_{j=1}^{n}\Big|\frac{\partial\pi_k^{-1}(Y_j,\boldsymbol{U}_j,\boldsymbol{\theta}_k^*)}{\partial\boldsymbol{\theta}_k}\Big|+o_p(1) \\
& = |\sqrt{n}(\hat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k^*)|\times E\Big|\frac{\partial\pi_k^{-1}(Y,\boldsymbol{U},\boldsymbol{\theta}_k^*)}{\partial\boldsymbol{\theta}_k}\Big|+o_p(1) \\
& = O_p(1).
\end{aligned}
$$

Similarly, we can show that $n^{-1/2}\sum_{i=1}^{n}|A_{ni}^{(2)}| = O_p(1)$ and $n^{-1/2}\sum_{i=1}^{n}|A_{ni}^{(3)}| = O_p(1)$.

References

Ai, C., Linton, O. and Zhang, Z. (2018). A simple and efficient estimation method for models with nonignorable missing data. *Statistica Sinica*, to appear.

Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of American Statistical Association*,

89, 81–87.

Fang, F. & Shao, J. (2016). Model selection with nonignorable nonresponse. *Biometrika*, **103**, 861–874.

Fitzmaurice, G.M., Molenberghs, G. & Lipsitz, S.R. (1995). Regression models for longitudinal binary responses with informative dropouts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **57**, 691–704.

Guan, Z. & Qin, J. (2017). Empirical likelihood method for non-ignorable missing data problems. *Lifetime Data Analysis*, **23**, 113–135.

Ibrahim, J.G., Chen, M.H., Lipsitz, S.R. & Herring, A.H. (2005). Missing data methods for generalized linear models: a comparative review. *Journal of American Statistical Association*, **100**, 332–346.

Kim, J.K. & Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*, London: Chapman & Hall/CRC.

Kim, J.K. & Yu, C.L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of American Statistical Association*, **106**, 157–165.

Little, R.J.A. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data. (Second Edition)*, New York: Wiley.

Molenberghs, G. & Kenward, M.G. (2007). *Missing Data in Clinical Studies*. New York: Wiley.

Morikawa, K., Kim, J.K. & Kano, Y. (2017). Semiparametric maximum likelihood estimation with data missing not at random. *Canadian Journal of Statistics*, **45**, 393–409.

Newey, W.K. & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, **4**, 2111–2245.

Qin, J., Leung, D. & Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of American Statistical Association*, **97**, 193–200.

Robins, J.M. (1987) Inference and missing data. *Biometrika*, **63**, 581–592.

Robins, J.M. & Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, **16**, 285–319.

Robins, J.M., Rotnitzky, A. & Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal*

of American Statistical Association, **89**, 846–866.

Scharfstein, D.O., Rotnitzky, A. & Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of American Statistical Association*, **94**, 1096–1120.

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

Tang, G., Little, R.J.A. & Raghunathan, T.E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, **90**, 747–764.

Tang, N., Zhao, P. & Zhu, H. (2014). Empirical likelihood for estimating equations with nonignorably missing data. *Statistica Sinica*, **24**, 723–747.

Wang, S., Shao, J. & Kim, J.K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, **24**, 1097–1116.

Xie, H., Qian, Y. & Qu, L. (2011). Semiparametric approach for analyzing nonignorable missing data. *Statistica Sinica*, **21**, 1881–1899.

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

Zhao, J., Yang, Y. & Ning, Y. (2018). Penalized pairwise pseudo likelihood for variable selection with nonignorable missing data. *Statistica Sinica*. **28**, 2125–2148.

Table 1: Simulated probability (÷1000) of selecting each model in simulation 1

n	Best model	Selected model						
		M_0	$M_1(X_1)$	$M_1(X_2)$	$M_1(X_3)$	$M_2(X_1, X_2)$	$M_2(X_1, X_3)$	$M_2(X_2, X_3)$
300	M_0	996	2	1	1	0	0	0
	$M_1(X_1)$	33	945	1	0	11	10	0
	$M_1(X_2)$	29	0	956	0	9	0	6
	$M_1(X_3)$	37	1	0	942	0	11	9
	$M_2(X_1, X_2)$	0	22	20	0	958	0	0
	$M_2(X_1, X_3)$	0	23	0	19	0	959	1
	$M_2(X_2, X_3)$	0	0	18	22	0	0	960
500	M_0	996	1	2	1	0	0	0
	$M_1(X_1)$	1	975	0	0	11	13	0
	$M_1(X_2)$	0	0	980	0	15	0	5
	$M_1(X_3)$	0	0	0	976	0	9	15
	$M_2(X_1, X_2)$	0	1	2	0	997	0	0
	$M_2(X_1, X_3)$	0	3	0	2	0	995	0
	$M_2(X_2, X_3)$	0	0	4	3	0	0	993
1000	M_0	1000	0	0	0	0	0	0
	$M_1(X_1)$	0	988	0	0	4	8	0
	$M_1(X_2)$	0	0	987	0	7	0	6
	$M_1(X_3)$	0	0	0	989	0	8	3
	$M_2(X_1, X_2)$	0	0	0	0	1000	0	0
	$M_2(X_1, X_3)$	0	0	0	0	0	1000	0
	$M_2(X_2, X_3)$	0	0	0	1	0	0	999

Table 2: Simulated bias and RMSE in estimating $E(Y)$ in simulation 1

Best model	Method	$n = 300$		$n = 500$		$n = 1000$	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
M_0	PVC	-0.022	0.357	-0.035	0.273	0.002	0.199
	FULL	-0.001	0.337	-0.022	0.261	0.006	0.188
	CC	0.815	0.899	0.785	0.837	0.814	0.841
	M_0	-0.021	0.357	-0.034	0.272	0.002	0.199
	$M_1(X_1)$	-0.017	0.376	-0.032	0.283	0.006	0.204
	$M_1(X_2)$	-0.008	0.380	-0.032	0.285	0.005	0.204
	$M_1(X_3)$	-0.020	0.376	-0.029	0.280	0.002	0.205
	$M_2(X_1, X_2)$	0.005	0.429	-0.032	0.319	0.010	0.224
	$M_2(X_1, X_3)$	-0.013	0.423	-0.027	0.322	0.004	0.226
	$M_2(X_2, X_3)$	-0.002	0.438	-0.024	0.322	0.002	0.226
$M_1(X_1)$	PVC	-0.004	0.368	-0.028	0.279	-0.006	0.206
	FULL	-0.001	0.337	-0.025	0.261	0.001	0.192
	CC	0.789	0.893	0.766	0.831	0.795	0.829
	M_0	0.096	0.382	0.090	0.296	0.083	0.231
	$M_1(X_1)$	-0.007	0.368	-0.027	0.279	-0.006	0.207
	$M_1(X_2)$	0.502	0.882	0.450	0.710	0.447	0.665
	$M_1(X_3)$	0.503	0.868	0.439	0.681	0.476	0.700
	$M_2(X_1, X_2)$	0.005	0.517	-0.014	0.443	-0.010	0.247
	$M_2(X_1, X_3)$	0.008	0.497	-0.035	0.360	-0.011	0.243
	$M_2(X_2, X_3)$	1.816	2.056	1.924	2.115	2.127	2.334
$M_2(X_1, X_2)$	PVC	0.018	0.453	-0.030	0.356	-0.008	0.251
	FULL	-0.001	0.337	-0.025	0.261	0.001	0.192
	CC	0.309	0.618	0.253	0.469	0.291	0.414
	M_0	0.611	0.923	0.626	0.840	0.692	0.804
	$M_1(X_1)$	0.458	0.783	0.477	0.668	0.507	0.585
	$M_1(X_2)$	0.478	0.884	0.465	0.648	0.501	0.592
	$M_1(X_3)$	2.633	2.826	2.756	2.912	2.922	3.072
	$M_2(X_1, X_2)$	-0.012	0.476	-0.033	0.364	-0.008	0.251
	$M_2(X_1, X_3)$	2.807	3.088	2.941	3.205	3.139	3.371
	$M_2(X_2, X_3)$	2.922	3.225	3.026	3.297	3.039	3.244

Table 3: Simulated bias and RMSE in estimating $E(Y)$ in simulation 2

Best model	Method	$n = 300$		$n = 500$		$n = 1000$	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
M_0	PVC	-0.029	0.349	0.000	0.285	0.003	0.195
	FULL	-0.020	0.336	0.003	0.274	0.005	0.188
	CC	0.791	0.878	0.814	0.869	0.819	0.846
	M_0	-0.037	0.349	-0.005	0.285	0.001	0.196
	$M_1(X_1)$	-0.021	0.359	-0.002	0.297	0.002	0.203
	$M_1(X_2)$	-0.036	0.359	-0.003	0.292	0.001	0.202
	$M_1(X_3)$	-0.034	0.364	-0.005	0.299	0.002	0.202
	$M_2(X_1, X_2)$	-0.006	0.412	0.002	0.342	0.001	0.223
	$M_2(X_1, X_3)$	-0.006	0.411	0.003	0.335	0.002	0.227
	$M_2(X_2, X_3)$	-0.013	0.415	-0.001	0.328	0.001	0.226
	$\tilde{M}_0$	-0.017	0.345	0.015	0.282	0.020	0.195
	$\tilde{M}_1(X_1)$	-0.014	0.356	0.018	0.291	0.023	0.200
	$\tilde{M}_1(X_2)$	-0.018	0.354	0.018	0.286	0.021	0.199
	$\tilde{M}_1(X_3)$	-0.013	0.359	0.016	0.294	0.022	0.199
	$\tilde{M}_2(X_1, X_2)$	0.011	0.399	0.020	0.312	0.015	0.223
	$\tilde{M}_2(X_1, X_3)$	0.014	0.397	0.011	0.330	0.017	0.232
	$\tilde{M}_2(X_2, X_3)$	-0.019	0.399	0.014	0.321	0.017	0.222
$M_1(X_1)$	PVC	-0.005	0.354	-0.007	0.287	0.000	0.205
	FULL	0.003	0.329	0.004	0.269	0.006	0.190
	CC	0.790	0.894	0.794	0.861	0.797	0.832
	M_0	0.097	0.382	0.101	0.308	0.114	0.240
	$M_1(X_1)$	-0.004	0.357	-0.004	0.288	0.002	0.206
	$M_1(X_2)$	0.516	0.891	0.492	0.815	0.464	0.684
	$M_1(X_3)$	0.510	0.866	0.455	0.679	0.468	0.648
	$M_2(X_1, X_2)$	0.003	0.490	-0.011	0.391	0.011	0.296
	$M_2(X_1, X_3)$	-0.009	0.444	0.000	0.351	-0.006	0.245
	$M_2(X_2, X_3)$	1.836	2.071	1.957	2.175	2.103	2.284
	$\tilde{M}_0$	0.096	0.397	0.092	0.323	0.109	0.265
	$\tilde{M}_1(X_1)$	-0.026	0.376	-0.029	0.304	-0.022	0.220
	$\tilde{M}_1(X_2)$	0.575	1.038	0.549	1.017	0.520	0.982
	$\tilde{M}_1(X_3)$	0.589	1.063	0.595	1.053	0.505	0.993
	$\tilde{M}_2(X_1, X_2)$	-0.026	0.527	-0.026	0.451	0.016	0.520
	$\tilde{M}_2(X_1, X_3)$	-0.011	0.560	-0.009	0.537	0.024	0.577
	$\tilde{M}_2(X_2, X_3)$	1.973	2.247	2.218	2.473	2.465	2.694

Table 3: Continued.

Best model	Method	$n = 300$		$n = 500$		$n = 1000$	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
$M_2(X_1, X_2)$	PVC	-0.010	0.448	-0.023	0.353	-0.001	0.247
	FULL	0.005	0.338	-0.005	0.267	0.002	0.189
	CC	0.282	0.595	0.280	0.497	0.290	0.412
	M_0	0.599	0.894	0.651	0.826	0.698	0.818
	$M_1(X_1)$	0.473	0.838	0.479	0.689	0.534	0.811
	$M_1(X_2)$	0.487	0.859	0.475	0.655	0.513	0.594
	$M_1(X_3)$	2.662	2.864	2.853	3.076	2.918	3.055
	$M_2(X_1, X_2)$	-0.034	0.482	-0.033	0.378	0.000	0.248
	$M_2(X_1, X_3)$	2.863	3.122	3.018	3.301	3.074	3.336
	$M_2(X_2, X_3)$	2.914	3.231	3.109	3.423	3.099	3.321
	$\tilde{M}_0$	0.632	1.174	0.591	1.193	0.627	1.287
	$\tilde{M}_1(X_1)$	0.959	1.363	0.949	1.288	1.018	1.351
	$\tilde{M}_1(X_2)$	0.960	1.356	0.963	1.286	1.018	1.313
	$\tilde{M}_1(X_3)$	1.800	2.102	1.913	2.318	2.062	2.477
	$\tilde{M}_2(X_1, X_2)$	0.847	1.318	0.837	1.245	0.984	1.306
	$\tilde{M}_2(X_1, X_3)$	2.025	2.440	2.169	2.601	2.268	2.727
	$\tilde{M}_2(X_2, X_3)$	1.992	2.407	2.201	2.671	2.308	2.763

Table 4: Instrument selection rates when the dimension of $\boldsymbol{X}$ is ten

	$n = 1000$			$n = 1500$		
Best model	Correct	Best	Wrong	Correct	Best	Wrong
M_0	1.000	0.960	0.000	1.000	0.969	0.000
$M_1(X_1)$	0.995	0.955	0.005	1.000	0.988	0.000
$M_2(X_1, X_2)$	0.980	0.795	0.020	0.986	0.916	0.014
$M_3(X_1, X_2, X_3)$	0.975	0.815	0.025	0.978	0.856	0.022
$M_4(X_1, \ldots, X_4)$	0.963	0.563	0.038	0.992	0.711	0.008
$M_5(X_1, \ldots, X_5)$	0.918	0.664	0.092	0.950	0.745	0.050
$M_6(X_1, \ldots, X_6)$	0.900	0.342	0.100	0.915	0.375	0.085
$M_7(X_1, \ldots, X_7)$	0.693	0.288	0.317	0.909	0.424	0.091
$M_8(X_1, \ldots, X_8)$	0.739	0.320	0.261	0.865	0.351	0.135
$M_9(X_1, \ldots, X_9)$	0.530	0.530	0.470	0.733	0.733	0.267

Table 5: Values of PVC, $\hat{\mu}$, standard error (SE) based on NHNES data

Model ($\boldsymbol{U}$)	$\boldsymbol{Z}$	PVC	$\hat{\mu}$	SE
M_0	bmi, age, gender	0.24	32.03	0.94
M_1(bmi)	age, gender	0.19	35.28	0.67
M_1(age)	bmi, gender	0.22	33.19	0.54
M_1(gender)	bmi, age	0.24	31.94	2.09
M_2(bmi, age)	gender	0.21	35.40	1.17
M_2(bmi, gender)	age	0.25	35.96	1.28
M_2(age, gender)	bmi	0.26	36.06	1.24
$\tilde{M}_0$	bmi, age, gender	0.25	31.92	0.74
$\tilde{M}_1$(bmi)	age, gender	1.38	31.52	0.58
$\tilde{M}_1$(age)	bmi, gender	1.37	31.49	0.79
$\tilde{M}_1$(gender)	bmi, age	1.39	31.36	0.49
$\tilde{M}_2$(bmi, age)	gender	0.22	35.40	1.09
$\tilde{M}_2$(bmi, gender)	age	1.40	31.50	0.67
$\tilde{M}_2$(age, gender)	bmi	0.22	36.02	1.30
M_0^{-Y}			34.44	0.95
M_1^{-Y}(bmi)			34.68	1.08
M_1^{-Y}(age)			34.46	0.95
M_1^{-Y}(gender)			34.16	1.13
M_2^{-Y}(bmi, age)			34.68	1.09
M_2^{-Y}(bmi, gender)			34.33	1.08
M_2^{-Y}(age, gender)			34.17	1.15
M_3^{-Y}(bmi,age, gender)			34.33	1.10

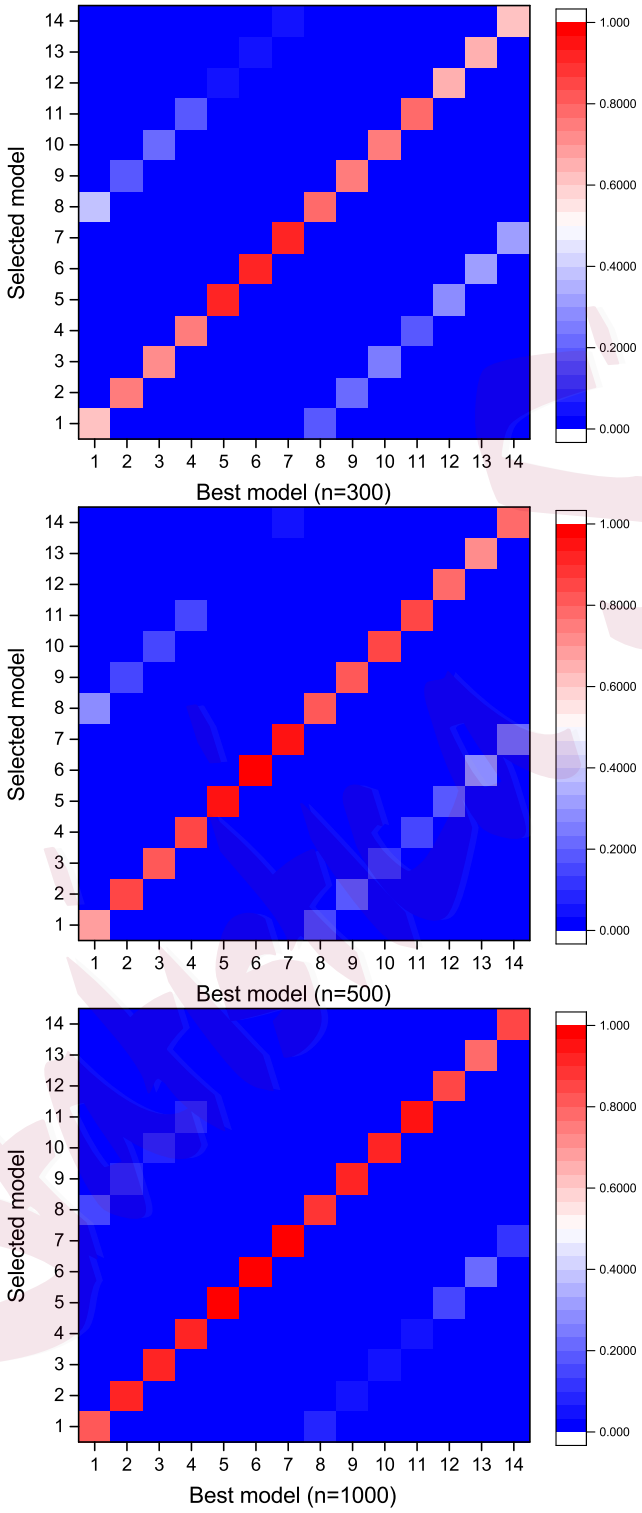

Figure 1: Hotmaps of true selection rates in simulation 2. The model numbers $\{1, 2, \ldots, 13, 14\}$ denote for models $\{M_0,\ M_1(X_1),\ \ldots,\ M_1(X_2, X_3),\ \tilde{M}_0,\ \tilde{M}_1(X_1), \ldots, \tilde{M}_1(X_2, X_3)\}$ in the second column in Table 3, respectively.