Statistica Sinica Preprint No: SS-2016-0340.R3	
Title	Generalized method of moments for nonignorable missing data
Manuscript ID	SS-2016-0340.R3
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0340
Complete List of Authors	Li Zhang Cunjie Lin and Yong Zhou
Corresponding Author	Cunjie Lin
E-mail	lincunjie@ruc.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

Statistica Sinica

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*

Abstract: In this study, we consider the problem of nonignorable missingness in the framework of generalized method of moments. To model the missing propensity, a semiparametric logistic regression model is adopted and we modify this model with nonresponse instrumental variables to overcome the identifiability issue. Under the identifiability conditions, we mitigate the effects of nonignorable missing data through reformulated estimating equations imputed via a kernel regression method, then the idea of generalized method of moments is applied to estimate the parameters of interest and the tilting parameter in propensity simultaneously. Moreover, the consistency and the asymptotic normality of the proposed estimators are established and we find that the price we pay for estimating unknown tilting parameter is an increased variance for the estimator of

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

population parameters, which is quite acceptable in contrast with validation sample especially for practical problems. At last, the proposed method is evaluated through simulation studies and demonstrated on the real data example.

Key words and phrases: Estimating equations, exponential tilting, generalized method of moments, kernel regression, nonignorable missing, nonresponse instrument.

1. Introduction

Missing data is a common occurrence in many applications, including clinical trials, sampling survey and observational studies, among others. It may arise due to subjects' refusal to undergo complete examinations, unavailability of measurements and loss of data. Most statistical models for dealing with the missing data depend on a missing data mechanism which is described by Little and Rubin (1987). They defined missing completely at random (MCAR) to be a process in which the probability of being observed is independent of observed or missing quantities. And missing at random (MAR) refers to the case where the propensity of missing data is conditionally independent of unobserved quantities given the observed quantities. Both MCAR and MAR are said to be ignorable in the sense that the propensity of missing data depends only on the observed data. If the missingness also depends on the unobserved quantities, the missing

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

data mechanism is termed nonignorable. For example, people with high incomes may be less likely to report their incomes, and in clinical trials, people who are getting worse are more likely to drop out than people who are getting better. In contrast to the ignorable mechanism, nonignorable missingness is associated with the unobserved values, and it leads to much more complexity for subsequent statistical inference.

Various methods have been developed to handle missing data, especially when missing mechanism is ignorable. But for nonignorable missing data, statistical inference usually depends on some unverifiable assumptions, and incorrect use of methods under ignorable assumptions may result in biased estimates. In this study, we focus on the identifiability and estimation for parameters of interest with nonignorable missing data. Let y be the response of interest subject to missingness, δ be the response status indicator of y. Suppose that a vector of covariates x are always observed, and given x, the conditional density of y is $f(y|x)$. In literatures, the conditional probability $\pi(x, y) = P(\delta = 1|x, y)$ is called the propensity of missing data. Under some parametric assumptions on both $\pi(x, y)$ and $f(y|x)$, Greenlees, Reece and Zieschang (1982) and Baker and Laird (1988) studied likelihood estimators with nonignorable missing data. However, the fully parametric assumption for joint modeling of the propensity and the population

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

model is very restrictive and the estimates are sensitive to failure of the assumed models. More efforts have been made to develop semiparametric approaches because $\pi(x,y)f(y|x)$ may be nonidentifiable when both $\pi(x,y)$ and $f(y|x)$ are purely nonparametric (Robins and Ritov (1997)). For example, Tang, Little and Raghunathan (2003) proposed a pseudo-likelihood method with a parametric model for $f(y|x)$ but an unspecified $\pi(x,y)$. Zhao and Shao (2015) studied the identifiability and estimation in a generalized linear model with a nonparametric missing mechanism.

On the other hand, Qin, Leung and Shao (2002) and Kott and Chang (2010) studied likelihood-based estimation and calibration weighting approach, respectively, for data with nonignorable nonresponse, assuming a parametric model for $\pi(x,y)$ and a nonparametric model for $f(y|x)$. Wang, Shao and Kim (2014) utilized nonresponse instrument, an auxiliary variable related to y but not related to the nonresponse probability, to overcome the difficulty of identifiability and applied the generalized method of moments to estimate the parameters in parametric propensity and nonparametric population. But it is difficult to verify the model assumption on propensity under nonignorable missingness, and a weaker assumption for $\pi(x,y)$ is more desirable in applications. Recently, Kim and Yu (2011) proposed a semiparametric logistic regression model for $\pi(x,y)$ and studied the semi-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

parametric estimation of mean functional. This model assumption is weaker than the parametric assumption and some refined methods based on this model can be found in recent literatures, see Zhao, Zhao and Tang (2013), Tang, Zhao and Zhu (2014) and Niu et al. (2014). However, to estimate the parameters of population and avoid the identifiability issue, they all assumed that the tilting parameter in the propensity is known or can be estimated using external data, which limits its applications to a great extent. To remove this serious limitation on methodology, Shao and Wang (2016) proposed to estimate the propensity using the generalized method of moments. Then other population parameters can be estimated using the inverse propensity weighting approach.

In this study, we consider the problem of nonignorable missingness in the framework of generalized method of moments with the propensity serving as auxiliary information. The properties of population are characterized by some parameters of interest via estimating equations without specifying distribution for the underlying population. The semiparametric logistic regression model is also adopted to model the propensity. Different from Shao and Wang (2016) that estimates the tilting parameter and population mean in two steps, we propose to estimate the parameters of interest and the tilting parameter of propensity simultaneously by the assistance of generalized

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

method of moments. To estimate the parameters, we impute the estimating equations by transforming the distribution of the unobserved data into that of the observed data based on the exponential tilting model. Then we get unbiased estimating equations consisted of both observed and missing information of data through a kernel regression method. The key advantage of this approach is that the parameters of interest and the tilting parameter can be estimated simultaneously without a validation sample and restrictive assumptions concerning population and propensity. At last, we establish the consistency and asymptotic normality of the proposed estimators for both parameters of interest and tilting parameter of propensity.

The rest of this article is organized as follows. In Section 2, we discuss the identifiability of the model and describe the model formulation. Then we explicate the estimation procedure in Section 3. In Section 4, we discuss the theoretical results for the two cases where the true value of the tilting parameter is known and unknown, respectively. We also propose the method to estimate the asymptotic variance. The results of simulation studies are reported in Section 5 and the real data examples are studied in Section 6. Some concluding remarks are given in Section 7 and the proofs are included in appendix.

2. Basic Setup and Identifiability

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

Let $(X_i, Y_i), 1 \leq i \leq n$, be n independent realizations of random variables (X, Y). Y is a response variable and X are d-dimensional covariates. Suppose that there are q estimating functions $\psi(y, x, \boldsymbol{\theta}) = (\psi_1(y, x, \boldsymbol{\theta}), \cdots, \psi_q(y, x, \boldsymbol{\theta}))^\tau$ satisfying $E\psi(Y, X, \boldsymbol{\theta}_0) = 0$, where $\boldsymbol{\theta}_0$ is the true value of p-dimensional parameter $\boldsymbol{\theta}$ and $q > p$. In this study, we are interested in making statistical inference on $\boldsymbol{\theta}$. If Y is fully observed, we can estimate $\boldsymbol{\theta}_0$ by minimizing

$$\left[\frac{1}{n}\sum_{i=1}^{n}\psi(Y_i, X_i, \boldsymbol{\theta})\right]^\tau W \left[\frac{1}{n}\sum_{i=1}^{n}\psi(Y_i, X_i, \boldsymbol{\theta})\right],$$

where W is a $q \times q$ weight matrix. However, this method can not be used directly with missing data.

Here we focus on the case where Y_i is subject to missingness and X_i is always observed. Let δ_i be the missing indicator for Y_i, where $\delta_i = 1$ if Y_i is observed and $\delta_i = 0$ otherwise. We assume that δ_i is independent of δ_j for any $i \neq j$ and the response mechanism is $\delta_i|(X_i, Y_i) \sim Bernoulli(\pi_i)$. The nonignorable missingness means π_i depends on X_i as well as Y_i, then write $\pi_i = \pi(X_i, Y_i)$. We consider a semiparametric logistic regression model for the propensity (Kim and Yu (2011)):

$$\pi(X, Y) = P(\delta = 1|X, Y) = \frac{\exp\{\alpha Y + g(X)\}}{1 + \exp\{\alpha Y + g(X)\}}, \tag{2.1}$$

where $g(\cdot)$ is an unspecified function and α is the tilting parameter. Since g and α are not identifiable without any further assumptions, we study

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

the identifiability of the model before estimation. Similar to the discussion of Wang, Shao and Kim (2014), the identifiability can be resolved by the aid of nonresponse instrument, which means that the covariates X has two components, $X = (U, Z)$, and Z acts as the instrumental variable satisfying that Z is independent of δ given Y and U but is associated with Y even in the presence of U. For the general case with semiparametric propensity, we extent the results in Wang, Shao and Kim (2014) and have the following theorem.

Theorem 1. *For missing data (X_i, Y_i, δ_i), the observed likelihood*

$$\prod_{i:\delta_i=1} \pi(X_i, Y_i) f(Y_i|X_i) \prod_{i:\delta_i=0} \int [1 - \pi(X_i, y)] f(y|X_i) dy.$$

is identifiable under the following conditions:

(C1) The covariates X can be decomposed into two components, $X = (U, Z)$, such that $P(\delta = 1|Y, X) = P(\delta = 1|Y, U) = H(g(U) + \alpha Y)$, where α is an unknown parameter and g is a continuously differentiable function not depending on z. $H(\cdot)$ is a known, strictly monotone, and twice differentiable function.

(C2) For any given u, there exist two values of Z, z_1 and z_2, such that $f(y|u, z_1) \neq f(y|u, z_2)$. And $f(y|u, z)$ has monotone likelihood ratio

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

in the sense that $f(y|u,z_1)/f(y|u,z_2)$ *is nondecreasing in* y *for any given* u.

According to the identifiability conditions, we can reformulate the response probability model (2.1) as

$$\pi(X,Y)=\pi(U,Y)=\frac{\exp\{\alpha Y+g(U)\}}{1+\exp\{\alpha Y+g(U)\}}. \tag{2.2}$$

Here, Z doesn't appear in model (2.2) but assists in resolving identifiability issue. Based on model (2.2), we can identify all parameters including $\boldsymbol{\theta}$, α and g. The question then is how to estimate these parameters using the available data.

3. Estimation Procedure

To estimate the unknown parameters $\boldsymbol{\theta}_0$ of interest, we propose to impute the estimating functions $\psi(Y,X,\boldsymbol{\theta})$ using the observed data. Under the ignorable missing mechanism condition, Zhou, Wan and Wang (2008) proposed to estimate parameters based on the following estimating functions

$$\psi^*(Y_i,X_i,\boldsymbol{\theta})=\delta_i\psi(Y_i,X_i,\boldsymbol{\theta})+(1-\delta_i)\hat{m}(X_i,\boldsymbol{\theta}),$$

where $\hat{m}(X_i,\boldsymbol{\theta})$ is a consistent estimator of $m(X_i,\boldsymbol{\theta})=E[\psi(Y,X,\boldsymbol{\theta})|X=X_i]$. Under the nonignorable propensity (2.2), we consider the adjusted

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

functions

$$\widetilde{\psi}(Y_i, X_i, \boldsymbol{\theta}) = \delta_i \psi(Y_i, X_i, \boldsymbol{\theta}) + (1 - \delta_i) m_0(X_i, \boldsymbol{\theta}), \tag{3.1}$$

where $m_0(x, \boldsymbol{\theta}_0) = E\{\psi(Y, X, \boldsymbol{\theta}_0)|X = x, \delta = 0\}$ is the conditional expectation of $\psi(Y, X, \boldsymbol{\theta}_0)$ given $X = x$ and $\delta = 0$ and it can be expressed based on the observed data. Actually, the conditional distribution of the missing data given x can be written as

$$f(y|x, \delta = 0) = f(y|x, \delta = 1) \times \frac{\exp\{\gamma y\}}{E\{\exp(\gamma Y)|x, \delta = 1\}}, \tag{3.2}$$

where $\gamma = -\alpha$ and it describes the deviation from the ignorable assumption. Equation (3.2) also shows that the density for the nonrespondents is an exponential tilting of the density for the respondents, which yields the following expression for $m_0(X, \boldsymbol{\theta})$,

$$\begin{aligned} m_0(X, \boldsymbol{\theta}) &= E\left[\psi(Y, X, \boldsymbol{\theta}) \times \frac{\exp\{\gamma Y\}}{E\{\exp(\gamma Y)|X, \delta = 1\}}\middle| X, \delta = 1\right] \\ &= \frac{E\{\psi(Y, X, \boldsymbol{\theta}) \exp(\gamma Y)|X, \delta = 1\}}{E\{\exp(\gamma Y)|X, \delta = 1\}} = \frac{E\{(1 - \delta)\psi(Y, X, \boldsymbol{\theta})|X\}}{E\{1 - \delta|X\}}. \end{aligned}$$

Then we have

$$\begin{aligned} E\{\tilde{\psi}(Y, X, \boldsymbol{\theta})\} &= E\{\delta\psi(Y, X, \boldsymbol{\theta}) + (1 - \delta)m_0(X, \boldsymbol{\theta})\} \\ &= E\left\{\delta\psi(Y, X, \boldsymbol{\theta}) + (1 - \delta)\frac{E\{(1 - \delta)\psi(Y, X, \boldsymbol{\theta})|X\}}{E\{1 - \delta|X\}}\right\} = 0. \end{aligned}$$

Hence, we can estimate $\boldsymbol{\theta}_0$ based on $\tilde{\psi}(Y, X, \boldsymbol{\theta})$ under the propensity model

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

(2.2). However, $m_0(x, \boldsymbol{\theta})$ is always unknown in presence of missing data and we need to estimate it consistently in advance.

Let $K(\cdot)$ be a d-variate kernel function satisfying $\int K(\mathbf{u})d\mathbf{u} = 1$. Furthermore, we assume that $K(\cdot)$ has a compact support and is a higher-order kernel of order m, i.e., $\int u_1^{\alpha_1} \cdots u_d^{\alpha_d} K(\mathbf{u})d\mathbf{u} = 0$, for $0 < \alpha_1 + \cdots + \alpha_d < m$, where $m > d$. Let $\mathbf{H}$ be a diagonal bandwidth matrix, then $K_h(\mathbf{u}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{u})$. For simplicity, we take the same bandwidth for each component in $\mathbf{H}$. Thus, with a known tilting parameter $\gamma = \gamma_0$, we can estimate $m_0(x, \boldsymbol{\theta})$ through the kernel regression method, i.e.,

$$\hat{m}_0(x, \boldsymbol{\theta}) = \frac{\sum_{i=1}^n \delta_i \psi(Y_i, X_i, \boldsymbol{\theta}) \exp(\gamma_0 Y_i) K_h(X_i, x)}{\sum_{i=1}^n \delta_i \exp(\gamma_0 Y_i) K_h(X_i, x)}, \tag{3.3}$$

where $K_h(u, x) = h^{-d} K(\frac{u-x}{h}) = h^{-d} K\left(\frac{u_1-x_1}{h}, \cdots, \frac{u_d-x_d}{h}\right)$. According to the consistency of nonparametric kernel estimator, $\hat{m}_0(X, \boldsymbol{\theta})$ is a consistent estimator of $m_0(X, \boldsymbol{\theta})$. By substituting $\hat{m}_0(X_i, \boldsymbol{\theta})$ for $m_0(X_i, \boldsymbol{\theta})$ in (3.1), we obtain the estimating functions:

$$\hat{\psi}(Y_i, X_i, \boldsymbol{\theta}) = \delta_i \psi(Y_i, X_i, \boldsymbol{\theta}) + (1 - \delta_i)\hat{m}_0(X_i, \boldsymbol{\theta}).$$

It can be shown that $\hat{\psi}(Y_i, X_i, \boldsymbol{\theta})$ is asymptotically unbiased and we can estimate $\boldsymbol{\theta}_0$ by minimizing

$$A_1(\boldsymbol{\theta}) = \left[\frac{1}{n}\sum_{i=1}^n \hat{\psi}(Y_i, X_i, \boldsymbol{\theta})\right]^\tau W_1 \left[\frac{1}{n}\sum_{i=1}^n \hat{\psi}(Y_i, X_i, \boldsymbol{\theta})\right],$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

where W_1 is a positive-definite matrix. We denote the minimizer by $\hat{\boldsymbol{\theta}}_{g1}$, which is called a GMM estimator. Under some mild regularity conditions, it can be shown that $\hat{\boldsymbol{\theta}}_{g1}$ is a consistent estimator of $\boldsymbol{\theta}_0$.

Note that $\hat{m}_0(x, \boldsymbol{\theta})$ depends on γ_0, which is unknown in practice, and it means that $\hat{\boldsymbol{\theta}}_{g1}$ also depends on the unknown quantity. To estimate γ_0, one approach is based on an independent survey or a validation sample which can be a subsample of the nonrespondents (Kim and Yu (2011)). But it is very costly and even infeasible in many cases, because the nonrespondents may still be reluctant to answer questions. Another approach is based on the method proposed by Shao and Wang (2016), that applied the generalized method of moments by profiling the nonparametric component with a kernel-type estimator. And then, the population parameters can be estimated using the inverse probability weighting (IPW) approach. Here, we provide an alternative way to estimate both $\boldsymbol{\theta}_0$ and γ_0. Note that, $A_1(\boldsymbol{\theta})$ can be regarded as a function of $\boldsymbol{\theta}_0$ and γ_0 without involving the nonparametric component $g(\cdot)$, which makes it possible to estimate $\boldsymbol{\theta}_0$ and γ_0 simultaneously.

Denote $\boldsymbol{\beta} = (\boldsymbol{\theta}^\tau, \gamma)^\tau$ and use $m_0(x, \boldsymbol{\beta})$ to stress parameters in $m_0(x, \boldsymbol{\theta})$. The estimating functions for $\boldsymbol{\theta}_0$ and γ_0 can be expressed as:

$$\hat{\psi}(Y_i, X_i, \boldsymbol{\beta}) = \delta_i \psi(Y_i, X_i, \boldsymbol{\theta}) + (1 - \delta_i)\hat{m}_0(X_i, \boldsymbol{\beta}),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

where $\hat{m}_0(X, \boldsymbol{\beta})$ is exactly the same estimate as (3.3), except that the tilting parameter γ_0 is treated as unknown parameter just like the population parameter $\boldsymbol{\theta}$. Since $q \geq (p+1)$ and $p+1$ is just the dimension of $\boldsymbol{\beta}$, we can still use the idea of generalized method of moments to estimate $\boldsymbol{\beta}_0 = (\boldsymbol{\theta}_0^\tau, \gamma_0)^\tau$. The valid objective function can be organized as

$$A_2(\boldsymbol{\beta}) = \left[\frac{1}{n}\sum_{i=1}^{n} \hat{\psi}(Y_i, X_i, \boldsymbol{\beta})\right]^\tau W_2 \left[\frac{1}{n}\sum_{i=1}^{n} \hat{\psi}(Y_i, X_i, \boldsymbol{\beta})\right], \tag{3.4}$$

where W_2 is a positive-definite symmetric weight matrix. We denote the minimizer by $\hat{\boldsymbol{\beta}}_{g2} = (\hat{\boldsymbol{\theta}}_{g2}^\tau, \hat{\gamma}_{g2})^\tau$.

4. Theoretical Results and Asymptotic Variance Estimation

In this section, we study the theoretical properties of estimators $\hat{\boldsymbol{\theta}}_{g1}$ and $\hat{\boldsymbol{\beta}}_{g2}$, corresponding to the cases with known and unknown tilting parameter, respectively, and give the choice of optimal matrices.

Theorem 2. *Suppose that γ_0 is known and there is a unique value $\boldsymbol{\theta}_0$ such that $E[\psi(Y, X, \boldsymbol{\theta}_0)] = 0$. Then under the conditions in Theorem 1 and the conditions (A1)-(A7) stated in the Appendix, as $n \to \infty$, $\hat{\boldsymbol{\theta}}_{g1} \to \boldsymbol{\theta}_0$ in probability. Moreover, $\sqrt{n}(\hat{\boldsymbol{\theta}}_{g1} - \boldsymbol{\theta}_0) \xrightarrow{D} N(0, \Sigma_{g_1})$, where $\Sigma_{g_1} = (\Gamma_\theta^\tau W_1 \Gamma_\theta)^{-1}\Gamma_\theta^\tau W_1 D W_1 \Gamma_\theta (\Gamma_\theta^\tau W_1 \Gamma_\theta)^{-1}$. Here $\Gamma_\theta = \Gamma(\boldsymbol{\theta}_0) = E[\frac{\partial \widetilde{\psi}(Y,X,\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}]$ and $D = D(\boldsymbol{\theta}_0) = E\{\psi(Y, X, \boldsymbol{\theta}_0)^{\otimes 2}\} + E\left\{\left(\frac{1}{\pi(U,Y)} - 1\right)[\psi(Y, X, \boldsymbol{\theta}_0) - m_0(X, \boldsymbol{\theta}_0)]^{\otimes 2}\right\}$, where for a vector $\mathbf{a}$, $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^\tau$.*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

For the asymptotic covariance matrix Σ_{g_1}*, the optimal weight matrix is* $W_1 = D^{-1}$*. With this choice of* W_1*, the asymptotic covariance matrix* Σ_{g_1} *reduces to* $(\Gamma_\theta^\tau D^{-1}\Gamma_\theta)^{-1}$ *and it holds that* $\Sigma_{g_1} - (\Gamma_\theta^\tau D^{-1}\Gamma_\theta)^{-1}$ *is a nonnegative definite matrix.*

Theorem 3. *Assume that the conditions in Theorem 2 are satisfied. Let* γ_0 *be the underlying value of the tilting parameter* γ*. Then, as* $n \to \infty$*, we have that the GMM estimators in (3.4) are consistent, i.e.,* $\hat{\boldsymbol{\theta}}_{g2} \to \boldsymbol{\theta}_0$ *and* $\hat{\gamma}_{g2} \to \gamma_0$ *in probability. Moreover, the estimators are asymptotically normal with* $\sqrt{n}(\hat{\boldsymbol{\beta}}_{g2} - \boldsymbol{\beta}_0) \xrightarrow{D} N(0, \Omega_{g_2})$*, where* $\Omega_{g_2} = (\Gamma_\beta^\tau W_2 \Gamma_\beta)^{-1}\Gamma_\beta^\tau W_2 D W_2 \Gamma_\beta (\Gamma_\beta^\tau W_2 \Gamma_\beta)^{-1}$*. Here* $\Gamma_\beta = \Gamma(\boldsymbol{\beta}_0) = E[\frac{\partial \widetilde{\psi}(Y,X,\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}]$ *and we use* $D = D(\boldsymbol{\beta}_0)$ *to stress the parameter, which is essentially identical with* $D(\boldsymbol{\theta}_0)$ *in Theorem 2.*

For the asymptotic covariance matrix Ω_{g_2}*, the optimal weight matrix is* $W_2 = D^{-1}$*. With this choice of* W_2*,* Ω_{g_2} *reduces to* $(\Gamma_\beta^\tau D^{-1}\Gamma_\beta)^{-1}$ *and it holds that* $\Omega_{g_2} - (\Gamma_\beta^\tau D^{-1}\Gamma_\beta)^{-1}$ *is a nonnegative definite matrix.*

From Theorem 2 and Theorem 3, we can see that the GMM estimators $\hat{\boldsymbol{\theta}}_{g1}$ and $\hat{\boldsymbol{\beta}}_{g2}$ share the same optimal weight matrix in theory. In practice, we usually use the identity matrix in the first step to obtain a GMM estimator, and based on the first-step GMM estimator, we can obtain an estimated optimal matrix, which is the matrix we utilize to get the final

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

GMM estimator. If we denote $\Gamma_\gamma = \Gamma(\gamma_0) = E[\frac{\partial\widetilde{\psi}(Y,X,\boldsymbol{\beta}_0)}{\partial\gamma}]$, we have

$$\Gamma_\beta^\tau D^{-1}\Gamma_\beta = \begin{pmatrix} \Gamma_\theta^\tau D^{-1}\Gamma_\theta^\tau & \Gamma_\theta^\tau D^{-1}\Gamma_\gamma \\ \Gamma_\gamma^\tau D^{-1}\Gamma_\theta & \Gamma_\gamma^\tau D^{-1}\Gamma_\gamma \end{pmatrix}.$$

Thus with the optimal weight matrix, the asymptotic normality for $\hat{\boldsymbol{\theta}}_{g2}$ and $\hat{\gamma}_{g2}$ can be expressed separately as

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{g2} - \boldsymbol{\theta}_0) \xrightarrow{D} N(0, \Sigma_{g_2}), \qquad \sqrt{n}(\hat{\gamma}_{g2} - \gamma_0) \xrightarrow{D} N(0, \sigma_{g_2}),$$

where $\Sigma_{g2} = [\Gamma_\theta^\tau D^{-1}\Gamma_\theta - \Gamma_\theta^\tau D^{-1}\Gamma_\gamma(\Gamma_\gamma^\tau D^{-1}\Gamma_\gamma)^{-1}\Gamma_\gamma^\tau D^{-1}\Gamma_\theta]^{-1}, \sigma_{g2} = [\Gamma_\gamma^\tau D^{-1}\Gamma_\gamma - \Gamma_\gamma^\tau D^{-1}\Gamma_\theta(\Gamma_\theta^\tau D^{-1}\Gamma_\theta)^{-1}\Gamma_\theta^\tau D^{-1}\Gamma_\gamma]^{-1}$. An appealing feature of this result is that our method does not require a validation sample for estimating γ, but only at the cost of bigger variance of estimator for $\boldsymbol{\theta}$. And we can treat the bigger variance Σ_{g2} as the price we pay for estimating unknown tilting parameter, which is quite acceptable in contrast with validation sample especially for practical problems.

Take the estimation of mean function for example, the interested parameter is $\theta_0 = E(Y)$. With a known γ_0, the observed likelihood is identifiable under propensity (2.1). We can estimate θ_0 using the estimating function $\psi_1(y,\theta) = y - \theta$ and it can be shown from Theorem 2 that $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, \sigma_1^2)$, where $\sigma_1^2 = E\{(Y-\theta_0)^2\} + E\left\{\left(\frac{1}{\pi(X,Y)} - 1\right)(Y - m_0(X))^2\right\}$, and $m_0(x) = E(Y|X = x, \delta = 0)$. That is the result of Theorem 1 in Kim and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

Yu (2011). If $\pi(X,Y)$ does not depend on Y, σ_1^2 reduces to the asymptotic variance of Cheng (1994). If γ_0 is unknown, the estimation function $\psi_1(y,\theta)$ is not enough to estimate θ and γ_0 simultaneously. Under this case, we suppose that the distribution of Y is symmetric and construct another estimating function: $\psi_2(y,\theta) = (y-\theta)^3$. In principle, other higher odd moments can also be used. Then we can use the proposed method to estimate $\boldsymbol{\beta}_0 = (\theta_0, \gamma_0)^\tau$ by minimizing $A_2(\boldsymbol{\beta})$ in (3.4). By Theorem 3, we have that both $\hat{\theta}_{g2}$ and $\hat{\gamma}_{g2}$ are asymptotically normal.

The results for nonignorable missing data are also applied for ignorable case where $\gamma_0 = 0$. In this case, the observed likelihood is identifiable and the propensity may depends on the whole X, i.e., $\pi(X,Y) = \pi(X)$, which can be regarded as a nonparametric model because $\gamma_0 = 0$. Then our results are consistent with those of Zhou, Wan and Wang (2008).

The asymptotic normality results provide a basis for estimating the variances of the proposed estimators. Based on these results, it suffices to estimate D, Γ_θ and Γ_β . First, we can consistently estimate Γ_θ and Γ_β by

$$\hat{\Gamma}_\theta = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{\psi}(Y_i, X_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{g_1}}, \qquad \hat{\Gamma}_\beta = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{\psi}(Y_i, X_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{g_2}},$$

respectively. And the consistent estimators for $D(\boldsymbol{\theta}_0)$ and $D(\boldsymbol{\beta}_0)$ are $\hat{D}(\hat{\boldsymbol{\theta}}_{g1}) =$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

$\frac{1}{n}\sum_{i=1}^n \hat{\eta}_i\hat{\eta}_i^\tau$ and $\hat{D}(\hat{\boldsymbol{\beta}}_{g2}) = \frac{1}{n}\sum_{i=1}^n \widetilde{\eta}_i\widetilde{\eta}_i^\tau$, respectively, where

$$\begin{aligned}\hat{\eta}_i &= \hat{m}_0(X_i,\hat{\boldsymbol{\theta}}_{g1}) + \frac{\delta_i}{\hat{\pi}(U_i,Y_i)}[\psi(Y_i,X_i,\hat{\boldsymbol{\theta}}_{g1}) - \hat{m}_0(X_i,\hat{\boldsymbol{\theta}}_{g1})],\\ \widetilde{\eta}_i &= \hat{m}_0(X_i,\hat{\boldsymbol{\beta}}_{g2}) + \frac{\delta_i}{\widetilde{\pi}(U_i,Y_i)}[\psi(Y_i,X_i,\hat{\boldsymbol{\theta}}_{g2}) - \hat{m}_0(X_i,\hat{\boldsymbol{\beta}}_{g2})].\end{aligned}$$

Hence, we need to estimate the propensity $\pi(U,Y)$, which involves estimating $g(U)$. For any given γ, denote $\zeta(U,\gamma) = \exp\{-g(U)\}$, which can be estimated by its kernel regression estimator:

$$\hat{\zeta}(U,\gamma) = \frac{\sum_{j=1}^n (1-\delta_j)K_h(U,U_j)}{\sum_{j=1}^n \delta_j \exp(\gamma Y_j)K_h(U,U_j)}.$$

If we use $\hat{\zeta}(U,\gamma_0)$ and $\hat{\zeta}(U,\hat{\gamma}_{g2})$ to distinguish between the two cases where γ_0 is known and unknown, we can estimate the propensity $\pi(U,Y)$ with

$$\hat{\pi}(U_i,Y_i) = \frac{1}{1+\hat{\zeta}(U_i,\gamma_0)\exp\{\gamma_0 Y_i\}}, \qquad \widetilde{\pi}(U_i,Y_i) = \frac{1}{1+\hat{\zeta}(U_i,\hat{\gamma}_{g2})\exp\{\hat{\gamma}_{g2} Y_i\}},$$

respectively. Finally, the asymptotic variances of the GMM estimators can be estimated consistently by $\hat{\Sigma}_{g_1} = (\hat{\Gamma}_\theta^\tau W_1\hat{\Gamma}_\theta)^{-1}\hat{\Gamma}_\theta^\tau W_1\hat{D}(\hat{\boldsymbol{\theta}}_{g1})W_1\hat{\Gamma}_\theta(\hat{\Gamma}_\theta^\tau W_1\hat{\Gamma}_\theta)^{-1}$, and $\hat{\Omega}_{g_2} = (\hat{\Gamma}_\beta^\tau W_2\hat{\Gamma}_\beta)^{-1}\hat{\Gamma}_\beta^\tau W_2\hat{D}(\hat{\boldsymbol{\beta}}_{g2})W_2\hat{\Gamma}_\beta(\hat{\Gamma}_\beta^\tau W_2\hat{\Gamma}_\beta)^{-1}$.

5. Simulation Studies

In this section, we conduct simulation studies to evaluate the finite sample performance of the proposed estimators.

Experiment 1.

In this experiment, we consider a simple case where the only covariate is the instrumental variable, i.e $X = Z$, and the propensity model is given

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

by $\pi(Y_i) = \exp(\alpha_0 Y_i)/\{1+\exp(\alpha_0 Y_i)\}$, and $\gamma_0 = -\alpha_0$ is used to control the missing rate. We generate data from the model

$$Y = \theta Z + \theta(Z-1)^2 + \varepsilon,$$

where the true value of θ is $\theta_0 = 1$ and Z are generated from $N(1,1)$ and $\varepsilon \sim N(0,1)$. Similar to Zhou, Wan and Wang (2008), the estimating functions are given by

$$\psi(Y,Z,\theta) = \begin{pmatrix} \psi_1(Y,Z,\theta) \\ \psi_2(Y,Z,\theta) \end{pmatrix} = \begin{pmatrix} Y^2 - 2\theta^2 - 2\theta^2 Z(Z-1) - \theta^2(Z-1)^4 - 1 \\ Y - \theta Z - \theta \end{pmatrix}.$$

We carry out 1000 replications with sample size $n = 1000$ and use the proposed methods to estimate θ and γ. In estimation, the Gaussian kernel $K(u) = \exp(-u^2/2)/\sqrt{2\pi}$ is adopted. The selected bandwidth for estimating $\hat{m}_0(Z,\theta)$ is $h = c\hat{\sigma}_Z n^{-1/3}$, where $\hat{\sigma}_Z$ is the standard deviation of Z_i in the sample and c is a constant and we use the optimal Gaussian kernel bandwidth $h = 1.06\hat{\sigma}_Z n^{-1/5}$ to estimate $\hat{\pi}(Y_i)$. The results are summarized in Table 1.

In Table 1, Bias and SE are the bias, estimated standard error based on the asymptotic normality results, averaged over 1000 replications. SD is the standard deviation calculated using the estimated values from 1000 replications. CP is the actual coverage probability of the nominal 95%

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

confidence interval. The estimator $\hat{\theta}_{g1}$ is based on the kernel-assisted estimating equation imputation scheme when γ_0 is known. The estimators $\hat{\theta}_{g2}$ and $\hat{\gamma}_{g2}$ are obtained based on the proposed method when γ_0 is unknown. From Table 1, we see that the bias, SE and SD of $\hat{\theta}_{g1}$ are smaller than that of $\hat{\theta}_{g2}$ under two settings with different missing rates. When γ_0 is unknown, the estimate $\hat{\gamma}_{g2}$ is also unbiased. Comparing across the results, we see that the proposed estimates are unbiased and the estimated variances are close to the true sampling variation. Overall, this provides empirical evidence for the asymptotic properties of the proposed estimators.

Experiment 2.

In the second experiment, we add another covariate U, i.e $X = (Z, U)$, and assess the performance of the proposed estimators under several missingness mechanisms. First, we generate Z from binomial distribution with success probability 0.5. Given Z, $U \sim N(Z, 1)$. We standardized U and Z, and generate Y from the model $Y = \theta_1 U + \theta_2 Z + \epsilon$, where $\epsilon \sim N(0, 1)$, the true value of $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is $\boldsymbol{\theta}_0 = (-1, 1)$. The estimating functions are given by

$$\psi(Y, X, \theta) = \begin{pmatrix} \psi_1(Y, X, \boldsymbol{\theta}) \\ \psi_2(Y, X, \boldsymbol{\theta}) \\ \psi_3(Y, X, \boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} Y - \theta_1 U - \theta_2 Z \\ UY - \theta_1 U^2 - \theta_2 UZ \\ ZY - \theta_1 UZ - \theta_2 Z^2 \end{pmatrix}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

The missing indicator δ is generated from the Bernoulli distribution with probability $\pi(U, Y)$. We consider the following response probability models similar to Kim and Yu (2011):

M1. (Linear Ignorable): $\pi(U_i, Y_i) = \frac{\exp(\phi_0+\phi_1 U_i)}{1+\exp(\phi_0+\phi_1 U_i)}$, where $(\phi_0, \phi_1) = (1.2, 0.1)$ for missing rate about 23%, $(\phi_0, \phi_1) = (0.4, 0.3)$ for missing rate about 40%.

M2. (Nonlinear Nonignorable): $\pi(U_i, Y_i) = \frac{\exp(\phi_0+\phi_1 U_i+\phi_2 U_i^2+\phi_3 Y_i)}{1+\exp(\phi_0+\phi_1 U_i+\phi_2 U_i+\phi_3 Y_i)}$, where $(\phi_0, \phi_1, \phi_2, \phi_3) = (1, 0.5, 0.2, 0.1)$ for missing rate about 24%, $(\phi_0, \phi_1, \phi_2, \phi_3) = (0.3, 0.5, 0.2, 0.1)$ for missing rate about 40%.

For each missing case, we carry out 1000 replications with sample size $n = 1000$ and use the proposed methods to estimate $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and γ_0. The Gaussian kernel is also adopted in all cases, and we use the bandwidth selection method described in experiment 1 to choose the bandwidth. The results for missing mechanisms M1 and M2 are presented in Tables 2 and 3, respectively. From these tables, we can see that the estimates derived when γ_0 is unknown are comparable with the results when γ_0 is known. Besides, under high missing rate, the proposed methods still give reliable results. The bias are all negligible, SEs and SDs are close, and CP are all around 95%, thus the asymptotic approximations work well for the proposed

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

approaches.

Experiment 3. In the third experiment, we conduct simulations to compare the proposed methods with the following two estimators: (1) the benchmark estimator that uses the complete data; (2) the naive method that uses the observed data and ignore the missing part. First, we generate data based on the logistic regression model

$$P(Y=1|Z,U)=\frac{\exp(\theta_1 Z+\theta_2 U)}{1+\exp(\theta_1 Z+\theta_2 U)},$$

where $Z\sim U[0,2]$, $U\sim N(0,1)$. The true values of θ_1 and θ_2 are $\theta_1=1$ and $\theta_2=-1$. The estimating functions are

$$\psi(Y,Z,U,\theta_1,\theta_2)=(1,Z,U)^T\left\{Y-\frac{\exp(\theta_1 Z+\theta_2 U)}{1+\exp(\theta_1 Z+\theta_2 U)}\right\}.$$

To generate the missing indicator, we consider the following model:

M3. (Linear Nonignorable): $\pi(U_i,Y_i)=\frac{\exp(\phi_0+\phi_1 U_i+\phi_2 Y_i)}{1+\exp(\phi_0+\phi_1 U_i+\phi_2 Y_i)}$, where $(\phi_0,\phi_1,\phi_2,\phi_3)=(0.7,0.45,0.5,0.2)$ for the missing rate about 23% and $(\phi_0,\phi_1,\phi_2,\phi_3)=(0.45,0.1,-0.15,-0.2)$ for the missing rate 40%.

We conduct 1000 replications with $n=1000$, and still adopt the Gaussian kernel and the same method to select bandwidth as above. The results are summarized in Table 4. For the benchmark and the naive estimator, we denote as $\hat{\boldsymbol{\theta}}_b$ and $\hat{\boldsymbol{\theta}}_n$, respectively. Table 4 shows that the naive estimator

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

performs worst without surprise. The other three estimators are comparable in terms of bias, but the SE and SD increase in order of $\hat{\boldsymbol{\theta}}_b$, $\hat{\boldsymbol{\theta}}_{g1}$ and $\hat{\boldsymbol{\theta}}_{g2}$. The coverage probabilities of the three estimators are all close to 95%. Over all, the results indicate that the proposed method can give close estimators to the no missing data estimators and the proposed methods are reliable and effective.

6. Real Data Example

We apply the proposed method to the Baseball data, which is described in Michael (1991). A total of 322 baseball players' information were collected in this dataset, including the annual salary on opening day (in USD 1000) in 1987, experience as measured by years in the major leagues and players' division, as well as some performance metrics such as times at Bat, hits, the number of runs scored by a player (Runs), Runs Batted In (RBI) and so on. Some studies indicate that the baseball players are paid based on their on-the-field performance(Hoaglin and Velleman (1995), Magel and Hoffman (2015)). In this study, we are interested in estimating the players' annual salaries using the players' performance statistics. Thus, the response variable Y is the log of annual salary and its missing rate is about 18.3%. As indicated by Stone and Pantuosco (2008), years in the major leagues and players' division are significant predictors for the baseball players' salaries.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

And our initial analysis also shows that these two variables are very useful covariates in estimating the salary. In addition to the players' experiences, the performance in baseball field is the primary variable. However, among all performance metrics, hits is highly correlated with other variables. Thus, hits is the only incorporated measure of players' ability in our model. We consider the linear regression model

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \epsilon,$$

where X_1, X_2, X_3 stand for years in the major leagues, players' division and hits, respectively. We assume that $E(\epsilon|X_1, X_2, X_3) = 0$ and $E(\epsilon^2|X_1, X_2, X_3) = \sigma^2$. To estimate the parameters, we use the following estimating functions:

$$\psi(Y, Z, \theta) = \begin{pmatrix} \psi_1(Y, X, \theta) \\ \psi_2(Y, Z, \theta) \\ \psi_3(Y, X, \theta) \\ \psi_4(Y, X, \theta) \\ \psi_5(Y, X, \theta) \end{pmatrix} = \begin{pmatrix} Y - \theta_0 - \theta_1 X_1 - \theta_2 X_2 - \theta_3 X_3 \\ X_1(Y - \theta_0 - \theta_1 X_1 - \theta_2 X_2 - \theta_3 X_3) \\ X_2(Y - \theta_0 - \theta_1 X_1 - \theta_2 X_2 - \theta_3 X_3) \\ X_3(Y - \theta_0 - \theta_1 X_1 - \theta_2 X_2 - \theta_3 X_3) \\ X_1 X_2(Y - \theta_0 - \theta_1 X_1 - \theta_2 X_2 - \theta_3 X_3) \end{pmatrix}.$$

The nonignorable missing assumption appears reasonable here, because the player with high income tends not to report salary. To apply the proposed method, we need to determine which covariate can be used as the instrumental variable Z. In this study, we consider the estimates with all possible instrument subsets to investigate the effect of invalid instrumental

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

variable, and find that the estimates of the regression coefficients are not sensitive to the choices of instrumental variable. Here, we only include the result with years in the major leagues (X_1) serving as the instrumental variable. For other scenarios with different instrumental variables, the results are reported in the supplementary material. From the result in Table 5, we can see that the players with longer time in the major leagues tend to have higher salaries. And the players' division is also an important factor of salary. Moreover, a high hits, as a measure of player's on-field ability, can increase the salary in a certain extent. Besides, the estimate of γ indicates that the nonignorable missing assumption holds for the response variable.

7. Discussion

This study provides an alternative method to handle nonignorable missing data in the framework of GMM. To apply the method, we should have more unbiased estimating equations than the population parameters to account for the tilting parameter. And we use the nonresponse instrument, which is related to the response but can be excluded from the propensity, to avoid the identifiability issue. Similar to Shao and Wang (2016), we can select an instrument using the criterion D

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

which converges to zero if and only if Z is an instrument and $\pi(U, Y)$ is

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

a correct model and consistently estimated by $\tilde{\pi}(U_i, Y_i)$. Hence, we can select an instrument by minimizing D over a group of candidate variables. Further discussions and simulation studies about the instrumental variable and the performance of D are included in the supplementary material.

In this study, we focused on the situation where only the response is subject to missing, for the case with missing observations in both response and covariates, the identifiability needs a more thorough discussion. Besides, the idea of the proposed method can be applied to other types of data with a more complex structure, including longitudinal data and censored survival data. However, with these types of data, the model and missing mechanism can be more complicated. The identifiability of model as well as theoretical analysis and computational implementation will also be more difficult. These are interesting and important problems that require a considerable amount of further work.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

Table 1: Simulation results for Experiment 1.

	$\gamma_0 = 0.7$, $MR = 26.06\%$				$\gamma = 0.5$, $MR = 30.62\%$			
	Bias	SE	SD	CP(%)	Bias	SE	SD	CP(%)
$\hat{\theta}_{g1}$	0.0009	0.0202	0.0199	94.70	0.0003	0.0202	0.0203	94.80
	γ_0 unknown, $MR = 26.06\%$				γ_0 unknown, $MR = 30.62\%$			
	Bias	SE	SD	CP(%)	Bias	SE	SD	CP(%)
$\hat{\theta}_{g2}$	0.0016	0.0280	0.0272	94.60	0.0048	0.0302	0.0300	94.70
$\hat{\gamma}_{g2}$	-0.0021	0.3456	0.3377	96.70	0.0528	0.2937	0.2916	96.00

Table 2: Simulation results for M1.

	$\gamma_0 = 0$, $MR = 23.20\%$				$\gamma_0 = 0$, $MR = 40.35\%$			
	Bias	SE	SD	CP(%)	Bias	SE	SD	CP(%)
$\hat{\boldsymbol{\theta}}_{g1}$	-0.0010	0.0394	0.0408	96.20	-0.0008	0.0458	0.0468	94.80
	0.0017	0.0405	0.0403	94.40	-0.0012	0.0460	0.0456	94.50
	γ_0 unknown, $MR = 23.20\%$				γ_0 unknown, $MR = 40.29\%$			
	Bias	SE	SD	CP(%)	Bias	SE	SD	CP(%)
$\hat{\boldsymbol{\theta}}_{g2}$	-0.0011	0.0418	0.0409	93.90	0.0002	0.0481	0.0468	94.70
	0.0003	0.0399	0.0403	94.70	0.0015	0.0457	0.0456	94.80
$\hat{\gamma}_{g2}$	-0.0002	0.1622	0.1582	94.30	-0.0006	0.1055	0.1033	94.40

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

Table 3: Simulation results for M2.

	$\gamma_0 = -0.1$, $MR = 24.43\%$				$\gamma_0 = -0.1$, $MR = 40.00\%$			
	Bias	SE	SD	CP(%)	Bias	SE	SD	CP(%)
$\hat{\boldsymbol{\theta}}_{g1}$	-0.0003	0.0399	0.0407	94.50	-0.0005	0.0430	0.0451	95.90
	-0.0014	0.0413	0.0409	94.20	-0.0019	0.0469	0.0457	94.10
	γ_0 unknown, $MR = 24.49\%$				γ_0 unknown, $MR = 40.00\%$			
	Bias	SE	SD	CP(%)	Bias	SE	SD	CP(%)
$\hat{\boldsymbol{\theta}}_{g2}$	-0.0009	0.0407	0.0410	95.70	-0.0015	0.0438	0.0453	95.80
	0.0003	0.0398	0.0411	96.70	-0.0000	0.0462	0.0458	95.00
$\hat{\gamma}_{g2}$	-0.0059	0.1624	0.1520	95.20	0.0004	0.1120	0.1072	93.80

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

Table 4: Simulation results for M3.

	MR=25%				MR=45%			
	Bias	SE	SD	CP(%)	Bias	SE	SD	CP(%)
$\hat{\boldsymbol{\theta}}_b$	0.0071	0.0777	0.0764	94.70	0.0073	0.0755	0.0764	95.80
	-0.0062	0.0944	0.0902	93.70	-0.0079	0.0941	0.0902	94.20
$\hat{\boldsymbol{\theta}}_n$	0.0712	0.0921	0.0916	89.80	-0.0609	0.1014	0.1020	89.80
	-0.0429	0.1110	0.1066	92.60	-0.0139	0.1237	0.1214	95.20
$\hat{\boldsymbol{\theta}}_{g1}$	0.0042	0.0945	0.0888	93.60	0.0086	0.1031	0.0999	94.40
	-0.0088	0.1108	0.1046	93.30	-0.0115	0.1190	0.1188	95.40
$\hat{\boldsymbol{\theta}}_{g2}$	-0.0047	0.1670	0.1632	93.30	0.0015	0.1753	0.1801	95.10
	0.0004	0.1294	0.1277	95.40	-0.0033	0.1257	0.1192	94.10
$\hat{\gamma}_{g2}$	0.0106	0.7553	0.7596	95.80	0.0069	0.4357	0.4438	96.10

Table 5: Result for Baseball data

	Estimates	SE	Confidence interval		Estimates	SE	Confidence interval
θ_0	4.0252	0.1303	[3.7698, 4.2807]	θ_2	0.2084	0.0685	[0.0741, 0.3427]
θ_1	0.0963	0.0069	[0.0829, 0.1098]	θ_3	0.0095	0.0010	[0.0076, 0.0114]
γ	-3.1300	0.0094	[-3.1484, -3.1117]				

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

Supplementary Materials

Supplementary material contains some proofs and further numerical studies.

Acknowledgements

We thank the associate editor and two referees for their helpful comments. Zhang's work was supported by National Natural Science Foundation of China (NSFC)(11601424), Youth Foundation of the Ministry of Education of China (15YJC910009), Science Foundation of Northwest University (14NW31), China Postdoctoral Science Foundation Funded Project (2015M580867; 2016T90940). Lin's work was supported by National Natural Science Foundation of China (NSFC)(11701561) and the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities (16JJD910002). Zhou's work was supported by the State Key Program of National Natural Science Foundation of China (71331006), the State Key Program in the Major Research Plan of National Natural Science Foundation of China (91546202), National Center for Mathematics and Interdisciplinary Sciences (NCMIS), Key Laboratory of RCSDS, AMSS, CAS (2008DP173182) and Innovative Research Team of Shanghai University of Finance and Economics (IRTSHUFE13122402).

Appendix

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

To prove the results of Theorems 2 and 3, we first introduce some notation. Denote the Euclidean norm of a matrix B by $\|B\|$. Let $|\mathbf{a}| = \max_{1\le i\le q}|a_i|$ for any vector $\mathbf{a} = (a_1, ..., a_q)^\tau$. Write $\mathbf{a} = O(b_n)$ if all elements a_i's satisfying $a_i = O(b_n)$. Define $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^\tau$. In proving our results, we need the following assumptions and regularity conditions, as in Newey and McFadden (1994) and Khan and Powell (2001).

(A1) The kernel function $K(\cdot)$ is a probability density function such that

(i) it is bounded and has compact support;

(ii) it is symmetric with $\mu_l = \int x^l K(x)dx$, and $\mu_2 < \infty$;

(iii) $K(x) \ge c$ for some $c > 0$ in some closed interval centered at zero.

(A2) The bandwidth h satisfies: $h \to 0$, $nh^d \to \infty$, $nh^{2m} \to 0$, and $n^{1/2}h^d/\log n \to \infty$ as $n \to \infty$.

(A3) The probability density function of X is $f(\cdot)$, which is bounded away from ∞ in the support of X, and the second derivatives of $f(x)$ is continuous and bounded.

(A4) (i) $E[\exp(2\gamma_0 y)]$ is finite;

(ii) $\pi(x, y) > c_2 > 0$ and $p(x) = E[\pi(x, y)|x] \ne 1$ almost surely.

(A5) $\psi(\cdot, \boldsymbol{\theta})$ is twice continuously differentiable in the neighborhood of $\boldsymbol{\theta}_0$,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

and $m_0(x, \boldsymbol{\beta})$ is twice continuously differentiable in the neighborhood of $\boldsymbol{\beta}_0$.

(A6) (i) $0 < E|\psi(Y, X, \boldsymbol{\theta}_0)|^2 < \infty$;

(ii) $0 < E|a^\tau \psi'(Y, X, \boldsymbol{\theta}_0)|^2 < \infty$ for any constant vector a.

(A7) $\psi'(\cdot, \boldsymbol{\theta})$ and $\psi^{(3)}(\cdot, \boldsymbol{\theta})$ are bounded by some integrable function $M(x)$ in the neighborhood of $\boldsymbol{\theta}_0$.

These are reasonable assumptions that are commonly used in the literature on nonparametric kernel estimation and estimating equations. We sketch the proof of Theorems 2 and 3 and leave the details in supplementary material. By the definition of $\hat{\psi}(Y_i, X_i, \boldsymbol{\theta})$, we have the decomposition:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\psi}(Y_i, X_i, \boldsymbol{\theta}) = I_1 + I_2 + I_3,$$

where $I_1 = \frac{1}{n}\sum_{i=1}^{n} \{\delta_i[\psi(Y_i, X_i, \boldsymbol{\theta}) - m_1(X_i, \boldsymbol{\theta})]\}$, $I_2 = \frac{1}{n}\sum_{i=1}^{n} \{\delta_i m_1(X_i, \boldsymbol{\theta}) + (1-\delta_i)m_0(X_i, \boldsymbol{\theta})\}$ and $I_3 = \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\{\hat{m}_0(X_i, \boldsymbol{\theta}) - m_0(X_i, \boldsymbol{\theta})\}$, here $m_1(X_i, \boldsymbol{\theta}) = E[\psi(Y_i, X_i, \boldsymbol{\theta})|X_i, \delta_i = 1]$. The first two terms I_1 and I_2 are sums of independent random variables. For I_3, We have the following lemma.

Lemma 1. *Under (A1)-(A7), we have* $\sqrt{n}(I_3 - I_3^{**}) = o_p(1)$, *where* $I_3^{**} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \left[\frac{1}{\pi(U_i, Y_i)} - 1\right][\psi(Y_i, X_i, \boldsymbol{\theta}_0) - m_0(X_i, \boldsymbol{\theta}_0)]$.

Then we can prove the following lemma.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

Lemma 2. *Under (A1)-(A7), we have* $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\psi}(Y_i, X_i, \boldsymbol{\theta}_0) \xrightarrow{D} N(0, D_1(\boldsymbol{\theta}_0))$, *where* $D_1(\boldsymbol{\theta}) = E\{\psi(Y_i, X_i, \boldsymbol{\theta})^{\otimes 2}\} + E\left\{ \left(\frac{1}{\pi(U_i, Y_i)} - 1\right)[\psi(Y_i, X_i, \boldsymbol{\theta}) - m_0(X_i, \boldsymbol{\theta})]^{\otimes 2} \right\}$. *And* $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\psi}(Y_i, X_i, \boldsymbol{\beta}_0) \xrightarrow{D} N(0, D_1(\boldsymbol{\beta}_0))$, *with* $\boldsymbol{\beta}_0 = (\boldsymbol{\theta}_0, \gamma_0)$.

Proof of Theorem 2. Let $\psi_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}(Y_i, X_i, \boldsymbol{\theta})$, $\Gamma_n(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\psi_n(\boldsymbol{\theta})$, we have $\Gamma_n^{\tau}(\hat{\boldsymbol{\theta}}_g)W_1\psi_n(\hat{\boldsymbol{\theta}}_g) = 0$. Applying Taylor's expansion to $\psi_n(\hat{\boldsymbol{\theta}}_g)$ at $\boldsymbol{\theta}_0$, we have $\psi_n(\hat{\boldsymbol{\theta}}_g) = \psi_n(\boldsymbol{\theta}_0) + \Gamma_n(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_0) + o_p(\|\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_0\|)$, where $\boldsymbol{\theta}^*$ lies between $\hat{\boldsymbol{\theta}}_g$ and $\boldsymbol{\theta}_0$, then

$$0 = \Gamma_n^{\tau}(\hat{\boldsymbol{\theta}}_g)W_1\psi_n(\hat{\boldsymbol{\theta}}_g) = \Gamma_n^{\tau}(\hat{\boldsymbol{\theta}}_g)W_1\psi_n(\boldsymbol{\theta}_0) + \Gamma_n^{\tau}(\hat{\boldsymbol{\theta}}_g)W_1\Gamma_n(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_0) + o_p(\|\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_0\|) \ ,$$

and $\|\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_0\| = O_p(n^{-\frac{1}{2}})$, thus

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_0) = -[\Gamma_n^{\tau}(\hat{\boldsymbol{\theta}}_g)W_1\Gamma_n(\boldsymbol{\theta}^*)]^{-1}\Gamma_n^{\tau}(\hat{\boldsymbol{\theta}}_g)W_1\sqrt{n}\psi_n(\boldsymbol{\theta}_0) + o_p(1) \ .$$

Since $-[\Gamma_n^{\tau}(\hat{\boldsymbol{\theta}}_g)W_1\Gamma_n(\boldsymbol{\theta}^*)]^{-1}\Gamma_n^{\tau}(\hat{\boldsymbol{\theta}}_g)W_1 \to^{\mathcal{P}} -[\Gamma^{\tau}W_1\Gamma]^{-1}\Gamma^{\tau}W_1$, where $\Gamma = \Gamma(\boldsymbol{\theta}_0) = E[\frac{\partial\widetilde{\psi}(Y,X,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}]$, by Lemma 1 and Slutsky theorem, we complete the proof. □

Proof of Theorem 3. The proof is similar to that of Theorem 2, we omit the details. □

References

Baker, S. G., and Laird, N. M. (1988), Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *J. Amer. Statist. Assoc.*, **83**, 62-69.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

REFERENCES

Cheng, P. E. (1994), Nonparametric estimation of mean functionals with data missing at random. *J. Amer. Statist. Assoc.*, **89**, 81-87.

Greenlees, W. S., Reece, J. S., and Zieschang, K. D. (1982), Imputation of missing values when the probability of response depends on the variable being imputed. *J. Amer. Statist. Assoc.*, **77**, 251-261.

Hoaglin, D. C. and Velleman, P. F. (1995). A critical look at some analyses of major league baseball salaries. *Amer. Statist.*, **49**, 277-285.

Khan, S. and Powell, J. L. (2001), Two-step estimation of semiparametric censored regression models. *J. Econometrics*, **103**, 73-110.

Kim, J. K. and Yu, C. L. (2011), A semiparametric estimation of mean functionals with nonignorable missing data, *J. Amer. Statist. Assoc.*, **106**, 157-165.

Kott, P. S., and Chang, T. (2010), Using calibration weighting to adjust for nonignorable unit nonresponse. *J. Amer. Statist. Assoc.*, **105**, 1265-1275.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data (2nd ed.)*, New York: Wiley.

Magel, R. and Hoffman, M. (2015), Predicting salaries of major league baseball players, *International Journal of Sports Scinece*, **5**, 51-58.

Michael, F. (1991), *SAS System for Statistical Graphics, First Edition (SAS Series in Statistical Applications),* SAS Institute.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

REFERENCES

Niu, C., Guo, X., Xu, W. and Zhu, L. (2014), Empirical likelihood inference in linear regression with nonignorable missing response. *Comput. Statist. Data Anal.*, **79**, 91-112.

Newey, W. K. and McFadden, D. (1994), Large sample estimation and hypothesis testing. In Handbook of Econometrics, vol. 4, R. F. Engle and D. L. McFadden, eds. Amsterdam: Elsevier, pp. 2111-2245.

Qin, J., Leung, D., and Shao, J. (2002), Estimation with survey data under nonignorable nonresponse or informative sampling, *J. Amer. Statist. Assoc.*, **97**, 193-200.

Robins, J. M. and Ritov, Y. (1997), Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Stat. Med.*, **16**, 285-319.

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

Stone, G. and Pantuosco, L. J. (2008), Estimating baseball salary equations from 1961-2005: A look at changes in major league compensation. *International Journal of Sport Finance*, **3**, 228-238.

Tang, G., Little, R. and Raghunathan, T. (2003), Analysis of multivariate missing data with nonignorable nonresponse. *Biometrka*, **90**, 747-764.

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0340

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

Zhao, J. W. and Shao, J. (2015), Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *J. Amer. Statist. Assoc.*, **110**, 1577-1590.

Zhao, H., Zhao, P. and Tang, N. (2013), Empirical likelihood inference for mean functionals with nonignorably missing response data. *Comput. Statist. Data Anal.*, **66**, 101-116.

Zhou, Y., Wan, A. and Wang, X. (2008), Estimating equations inference with missing data, *J. Amer. Statist. Assoc.*, **103**, 1187-1198.

School of Economics and Management, Northwest University, Xian, China

E-mail: lizhang05@163.com

Center for Applied Statistics of Renmin University of China, Beijing, China

School of Statistics, Renmin University of China, Beijing, China

E-mail: lincunjie@ruc.edu.cn

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai

E-mail: yzhou@amss.ac.cn