7-2014

An instrument variable approach for identification and estimation with nonignorable nonresponse

Mathematica Policy Research

Jun Shao
East China Normal University

Jae Kwang Kim
Iowa State University, jkim@iastate.edu

Statistica Sinica **24** (2014), 1097-1116
doi:http://dx.doi.org/10.5705/ss.2012.074

AN INSTRUMENTAL VARIABLE APPROACH FOR IDENTIFICATION AND ESTIMATION WITH NONIGNORABLE NONRESPONSE

Sheng Wang[1], Jun Shao[2,3], and Jae Kwang Kim[4]

[1]*Mathematica Policy Research,* [2]*East China Normal University,*
[3]*University of Wisconsin and* [4]*Iowa State University*

Abstract: Estimation based on data with nonignorable nonresponse is considered when the joint distribution of the study variable y and covariate $\boldsymbol{x}$ is nonparametric and the nonresponse probability conditional on y and $\boldsymbol{x}$ has a parametric form. The likelihood based on observed data may not be identifiable even when the joint distribution of y and $\boldsymbol{x}$ is parametric. We show that this difficulty can be overcome by utilizing a nonresponse instrument, an auxiliary variable related to y but not related to the nonresponse probability conditional on y and $\boldsymbol{x}$. Under some conditions we can apply the generalized method of moments (GMM) to obtain estimators of the parameters in the nonresponse probability and the nonparametric joint distribution of y and $\boldsymbol{x}$. Consistency and asymptotic normality of GMM estimators are established. Simulation results and an application to a data set from the Korean Labor and Income Panel Survey are also presented.

Key words and phrases: Consistency and asymptotic normality, generalized method of moments, missing not at random, nonparametric distribution, nonresponse instrument, parametric propensity.

1. Introduction

Nonresponse at an appreciable rate exists in many applications. Let y be the value of a study variable subject to nonresponse, δ be the response indicator of y ($\delta = 1$ if y is observed and $\delta = 0$ otherwise), and $\boldsymbol{x}$ be a vector of covariates that are always observed, where $\boldsymbol{x}$ is either deterministic or random with inference conditional on values of $\boldsymbol{x}$. We assume that an independent sample of size n is obtained with $(y_i, \delta_i, \boldsymbol{x}_i)$ being the realized value of $(y, \delta, \boldsymbol{x})$ for sampled unit $i = 1, \ldots, n$, where y_i is observed if and only if $\delta_i = 1$. Let $p(y|\boldsymbol{x})$ be the conditional density of y given $\boldsymbol{x}$ and $p(y)$ be the marginal density of y. The joint distribution of y and δ given $\boldsymbol{x}$ is determined by $p(y|\boldsymbol{x})$ and the nonresponse mechanism $P(\delta = 1|y, \boldsymbol{x})$. Nonresponse is said to be ignorable if the nonresponse mechanism is a function of the observed data (Little and Rubin (2002)). Since $(y_1, \delta_1, \boldsymbol{x}_1), \ldots, (y_n, \delta_n, \boldsymbol{x}_n)$ are independent, ignorable nonresponse in this case

means that $P(\delta = 1|y, \boldsymbol{x}) = P(\delta = 1|\boldsymbol{x})$. For ignorable nonresponse, there is a rich literature on deriving valid estimators of unknown parameters in $p(y|\boldsymbol{x})$ or $p(y)$. When $P(\delta = 1|y, \boldsymbol{x})$ depends on y that may be missing, which is the focus of this paper, nonresponse is nonignorable and the construction of valid estimators is a challenging problem.

Greenlees, Reece, and Zieschang (1982) and Baker and Laird (1988) proposed likelihood methods under some parametric assumptions on both $P(\delta = 1|y, \boldsymbol{x})$ and $p(y|\boldsymbol{x})$. However, a fully parametric approach is sensitive to the parametric model assumptions. Since the population is not identifiable when both $P(\delta = 1|y, \boldsymbol{x})$ and $p(y|\boldsymbol{x})$ are nonparametric (Robins and Ritov (1997)), efforts have been made in some cases where one of $P(\delta = 1|y, \boldsymbol{x})$ and $p(y|\boldsymbol{x})$ is parametric and the other is nonparametric. Tang, Little, and Raghunathan (2003) considered the situation where $p(y|\boldsymbol{x})$ is parametric but $P(\delta = 1|y, \boldsymbol{x})$ is nonparametric, whereas Qin, Leung, and Shao (2002), Chang and Kott (2008), and Kott and Chang (2010) focused on the case where $P(\delta = 1|y, \boldsymbol{x})$ is parametric but $p(y|\boldsymbol{x})$ is nonparametric. In many applications, such as survey problems, it is difficult to find a suitable parametric model for $p(y|\boldsymbol{x})$, but a parametric model for $P(\delta = 1|y, \boldsymbol{x})$ such as the logistic may be reasonable.

Although Greenlees, Reece, and Zieschang (1982), Qin, Leung, and Shao (2002), Chang and Kott (2008), and Kott and Chang (2010) proposed some estimation methods, their results rely on the assumption that the observed likelihood is identifiable. Identifiability is necessary for the existence of consistent estimators of parameters (Gelfand and Sahu (1999)). It has been studied in the case of parametric $p(y|\boldsymbol{x})$ (Chen (2001)); Tang, Little, and Raghunathan (2003)) and some semiparametric $p(y|\boldsymbol{x})$ (Rotnitzky and Robins (1997)), but it is not well studied in the case of nonparametric $p(y|\boldsymbol{x})$.

In Section 2, we establish a sufficient condition for the identifiability of observed likelihood assuming a parametric model for $P(\delta = 1|y, \boldsymbol{x})$ but without assuming parametric model for $p(y|\boldsymbol{x})$. The key is to utilize a nonresponse instrument, a component of $\boldsymbol{x}$ that is related to y but not related to the nonresponse conditional on y and other components of $\boldsymbol{x}$. Without such an auxiliary variable, the observed likelihood may be nonidentifiable even when both $p(y|\boldsymbol{x})$ and $P(\delta = 1|y, \boldsymbol{x})$ are parametric, as shown in a simple example in Section 2.

When the observed likelihood is identifiable, efforts are still needed to develop an estimation method for unknown quantities in $p(y|\boldsymbol{x})$ or $p(y)$. Qin, Leung, and Shao (2002) applied the empirical likelihood approach, while Kott and Chang (2010) used calibration. In Section 3, we propose the generalized method of moments (GMM) for estimation and establish the consistency and asymptotic normality of the GMM estimators. An advantage of the proposed GMM approach is that the asymptotic covariance matrices of the estimators can be explicitly

derived and their consistent estimators can be easily computed, which is useful for statistical inference such as setting confidence regions.

In Section 4, some simulation results are on the finite sample performance of the GMM estimators and the related confidence intervals for the population mean. An application to a data example is also included. Section 5 contains some concluding remarks. Proofs are given in the Appendix.

2. Identifiability

Since y_i is observed if and only if $\delta_i = 1$, the observed likelihood is

$$\prod_{i:\ \delta_i=1} P(\delta_i = 1|y_i, \boldsymbol{x}_i)p(y_i|\boldsymbol{x}_i) \prod_{i:\ \delta_i=0} \int [1 - P(\delta_i = 1|y, \boldsymbol{x}_i)]p(y|\boldsymbol{x}_i)dy, \tag{2.1}$$

where each of $P(\delta = 1|y, \boldsymbol{x})$ and $p(y|\boldsymbol{x})$ may be parametric or nonparametric. It is identifiable if two different populations do not produce the same observed likelihood. Because the second product in (2.1) involves integrals of the quantities in the first product in (2.1), identifiability comes to whether two different populations give the same $P(\delta = 1|y, \boldsymbol{x})p(y|\boldsymbol{x})$ for all possible values of $(y, \boldsymbol{x})$.

Even if both $P(\delta = 1|y, \boldsymbol{x})$ and $p(y|\boldsymbol{x})$ are parametric, identifiability under nonignorable nonresponse is not trivial, as an example indicates.

Example 1. Suppose there is no covariate and $p(y)$ is normal with unknown mean μ and variance σ^2. Let $P(\delta = 1|y) = [1 + \exp(\alpha + \beta y)]^{-1}$ with unknown real-valued α and β. Nonresponse is ignorable if and only if $\beta = 0$. Here

$$P(\delta = 1|y)p(y) = \frac{\exp[-(y - \mu)^2/2\sigma^2]}{\sqrt{2\pi}\sigma[1 + \exp(\alpha + \beta y)]}.$$

The observed likelihood is not identifiable if $(\alpha, \beta, \mu, \sigma)$ and $(\alpha', \beta', \mu', \sigma')$ produce

$$\frac{\exp[-(y - \mu)^2/2\sigma^2]}{\sigma[1 + \exp(\alpha + \beta y)]} = \frac{\exp[-(y - \mu')^2/2\sigma'^2]}{\sigma'[1 + \exp(\alpha' + \beta' y)]} \qquad \text{for all } y. \tag{2.2}$$

But (2.2) holds if $\sigma = \sigma'$, $\alpha' = -\alpha$, $\beta' = -\beta$, $\alpha = (\mu'^2 - \mu^2)/2\sigma^2$, and $\beta = (\mu' - \mu)/\sigma^2$. Hence, the observed likelihood is not identifiable unless $\beta = \beta' = 0$ (ignorable nonresponse).

The observed likelihood in Example 1 is identifiable when there is a covariate $\boldsymbol{z}$ such that the conditional distribution of y given $\boldsymbol{z}$ depends on the value of $\boldsymbol{z}$, and $P(\delta = 1|y, \boldsymbol{z})$ does not depend on $\boldsymbol{z}$.

Theorem 1. *The observed likelihood* (2.1) *is identifiable under the following conditions.*

(C1) *The covariate* $\boldsymbol{x}$ *has two components,* $\boldsymbol{x} = (\boldsymbol{u}, \boldsymbol{z})$*, such that*

$$P(\delta = 1|y, \boldsymbol{x}) = P(\delta = 1|y, \boldsymbol{u}) = \Psi(\alpha_{\boldsymbol{u}} + \beta_{\boldsymbol{u}} y), \tag{2.3}$$

where $\alpha_{\boldsymbol{u}}$ *and* $\beta_{\boldsymbol{u}}$ *are unknown parameters not depending on* $\boldsymbol{z}$ *but may depend on* $\boldsymbol{u}$*,* Ψ *is a known, strictly monotone, and twice differentiable function from* $\mathcal{R}$ *to* $(0, 1]$*, and, for any given* $\boldsymbol{u}$*, there exist two values of* $\boldsymbol{z}$*,* $\boldsymbol{z}_1$ *and* $\boldsymbol{z}_2$ *(which may depend on* $\boldsymbol{u}$*), such that* $p(y|\boldsymbol{u}, \boldsymbol{z}_1) \neq p(y|\boldsymbol{u}, \boldsymbol{z}_2)$*.*

(C2) *For any given* $\boldsymbol{u}$*,* $p(y|\boldsymbol{u}, \boldsymbol{z})$ *has a Lebesgue density* $f(y|\boldsymbol{u}, \boldsymbol{z})$ *with a monotone likelihood ratio.*

When a covariate $\boldsymbol{x}^*$ associated with a study variable y^* is measured with error, valid estimators of regression parameters can be obtained by utilizing an instrument $\boldsymbol{z}$ that is correlated with $\boldsymbol{x}^*$ but independent of y^* conditioned on $\boldsymbol{x}^*$. In (C1), we decompose the covariate vector $\boldsymbol{x}$ into $\boldsymbol{u}$ and $\boldsymbol{z}$, such that $\boldsymbol{z}$ is correlated with $\boldsymbol{x}^* = (y, \boldsymbol{u})$, a "covariate" associated with the "study variable" $y^* = \delta$, and $\boldsymbol{z}$ is independent of $y^* = \delta$ conditioned on $\boldsymbol{x}^* = (y, \boldsymbol{u})$. Unconditionally, $\boldsymbol{z}$ may still be related to δ. Since y is subject to nonresponse, not measurement error, we call $\boldsymbol{z}$ a *nonresponse instrument*, it helps to identify the observed likelihood so that valid estimators of unknown quantities can be obtained (Section 3). The existence of $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ such that $p(y|\boldsymbol{u}, \boldsymbol{z}_1) \neq p(y|\boldsymbol{u}, \boldsymbol{z}_2)$ means that $\boldsymbol{z}$ is associated with y even in the presence of $\boldsymbol{u}$.

The nonresponse mechanism in (2.3) has a parametric model. Popular parametric models are the logistic model with $\Psi(t) = [1 + \exp(t)]^{-1}$, and the probit model with Ψ the distribution function of the standard normal.

Here $f(y|\boldsymbol{u}, \boldsymbol{z})$ in (C2) is nonparametric, since its form is not specified. The monotone likelihood ratio property in (C2) is satisfied for many Lebesgue density families, for example, many one-parameter exponential families, the logistic distribution with location parameter $\boldsymbol{z}_j$, and the uniform distribution on the interval $(\boldsymbol{z}_j, \boldsymbol{z}_j + 1)$. The following result provides another example in which condition (C2) holds.

Corollary 1. *Suppose* (C1) *holds and* $\partial \log(f(y|\boldsymbol{u}, \boldsymbol{z}))/\partial y$ *is a monotone function on the support of* $f(y|\boldsymbol{u}, \boldsymbol{z})$*, where* $f(y|\boldsymbol{u}, \boldsymbol{z})$ *is given in* (C2)*. The observed likelihood* (2.1) *is identifiable if*

(i) $f(y|\boldsymbol{u}, \boldsymbol{z}) = f(y - \varphi)$ *with a parameter* $\varphi \in \mathcal{R}$ *and a Lebesgue density* f *(which may be unknown), or*

(ii) $f(y|\boldsymbol{u}, \boldsymbol{z}) = \varphi_j f(\varphi y)$ *with a parameter* $\varphi > 0$ *and a Lebesgue density* f *(which may be unknown), and either* $f(y) = 0$ *or* $f(y) = f(-y)$ *for* $y \leq 0$*.*

3. Estimation

Using the data with nonresponse and a parametric $p(y|\boldsymbol{x})$, we can estimate parameters by maximizing the observed likelihood (2.1). Here we consider nonparametric $p(y|\boldsymbol{x})$ and the GMM (Hansen (1982); Hall (2005)) for estimation.

The key idea of the GMM is to construct a set of L estimating functions

$$g_l(\vartheta, y, \delta, \boldsymbol{x}), \qquad l = 1, \ldots, L, \quad \vartheta \in \Theta,$$

where Θ is the parameter space containing the true parameter value θ, $L \geq$ the dimension of Θ, the g_l's are non-constant functions with $E[g_l(\theta, y, \delta, \boldsymbol{x})] = 0$ for all l, and are not linearly dependent. Let

$$G(\vartheta) = \Big(\frac{1}{n}\sum_i g_1(\vartheta, y_i, \delta_i, \boldsymbol{x}_i), \ldots, \frac{1}{n}\sum_i g_L(\vartheta, y_i, \delta_i, \boldsymbol{x}_i)\Big)^T, \quad \vartheta \in \Theta, \tag{3.1}$$

where a^T denotes the transpose of the vector a. If L is the same as the dimension of Θ, then we may be able to find a $\hat{\theta}$ such that $G(\hat{\theta}) = 0$. If L is larger than the dimension of Θ, however, a solution to $G(\vartheta) = 0$ may not exist. A GMM estimator of θ can be obtained using a two-step algorithm.

1. Obtain $\hat{\theta}^{(1)}$ by minimizing $G^T(\vartheta)G(\vartheta)$ over $\vartheta \in \Theta$.
2. Let $\hat{W}$ be the inverse matrix of the $L \times L$ matrix whose (l, l') element is $n^{-1}\sum_i g_l(\hat{\theta}^{(1)}, y_i, \delta_i, \boldsymbol{x}_i)g_{l'}(\hat{\theta}^{(1)}, y_i, \delta_i, \boldsymbol{x}_i)$. The GMM estimator $\hat{\theta}$ is obtained by minimizing $G^T(\vartheta)\hat{W}G(\vartheta)$ over $\vartheta \in \Theta$.

We first consider the situation where $\boldsymbol{x} = (\boldsymbol{u}, \boldsymbol{z})$, $\boldsymbol{z} = (\boldsymbol{t}, z)$ has a q-dimensional continuous component $\boldsymbol{t}$ and a discrete component z taking values $1, \ldots, J$, and $\boldsymbol{u}$ is a continuous p-dimensional covariate.

Although the observed likelihood (2.1) is identifiable under (C1) and (C2) $\alpha_{\boldsymbol{u}}$ and $\beta_{\boldsymbol{u}}$ in (2.3) depend on values of $\boldsymbol{u}$ and, hence, there may be uncountably many parameters for the case of continuous $\boldsymbol{u}$. We assume therefore a condition to replace (2.3):

$$P(\delta = 1|y, \boldsymbol{x}) = P(\delta = 1|y, \boldsymbol{u}) = \Psi(\alpha + \beta y + \gamma \boldsymbol{u}^T), \tag{3.2}$$

where Ψ is as in (2.3), and (α, β, γ) is a $(p+2)$-dimensional unknown parameter not depending on values of $\boldsymbol{x}$. A similar assumption to (3.2) was made in Qin, Leung, and Shao (2002) and Kott and Chang (2010).

To estimate (α, β, γ), the GMM can be applied with the $L = p + q + J$ functions

$$\boldsymbol{g}(\vartheta, y, \delta, \boldsymbol{x}) = \begin{pmatrix} \boldsymbol{d}^T[\delta w(\vartheta) - 1] \\ \boldsymbol{t}^T[\delta w(\vartheta) - 1] \\ \boldsymbol{u}^T[\delta w(\vartheta) - 1] \end{pmatrix}, \tag{3.3}$$

where $\boldsymbol{d}$ is the J-dimensional row vector whose lth component is $I(z = l)$, $I(A)$ the indicator function of A, $w(\vartheta) = [\Psi(\vartheta_1 + \vartheta_2 y + \vartheta_3 \boldsymbol{u}^T)]^{-1}$, and $\vartheta = (\vartheta_1, \vartheta_2, \vartheta_3)$. The function $\boldsymbol{g}$ is motivated by the fact that, when θ is the true parameter value,

$$\begin{aligned} E[\boldsymbol{g}(\theta, y, \delta, \boldsymbol{x})] &= E\{\boldsymbol{\xi}[\delta w(\theta) - 1]\} \\ &= E(E\{\boldsymbol{\xi}[\delta w(\theta) - 1] | y, \boldsymbol{z}, \boldsymbol{u}\}) \\ &= E\left\{\boldsymbol{\xi}\left[\frac{E(\delta|y, \boldsymbol{z}, \boldsymbol{u})}{P(\delta = 1|y, \boldsymbol{z}, \boldsymbol{u})} - 1\right]\right\} \\ &= 0, \end{aligned}$$

where $\boldsymbol{\xi} = (\boldsymbol{d}, \boldsymbol{t}, \boldsymbol{u})^T$. We need $q + J \geq 2$. If $q = 0$, the requirement $J \geq 2$ is satisfied if z is not a constant.

Take G as at (3.1) with g_l the lth function of $\boldsymbol{g}$, $\hat{W}$ given by the two-step algorithm, and $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\gamma})$ the two-step GMM estimator of $\theta = (\alpha, \beta, \gamma)$.

Theorem 2. *As $n \to \infty$, the following conclusions hold under* (C1) *with* (2.3) *replaced by* (3.2), *and* (C3)–(C4).

(i) *There exists $\{\hat{\theta}\}$ such that $P(s(\hat{\theta}) = 0) \to 1$ and $\hat{\theta} \to_p \theta$ as $n \to \infty$, where $s(\vartheta) = -\partial[G^T(\vartheta)\hat{W}G(\vartheta)]/\partial\vartheta$ and $\to_p$ denotes convergence in probability.*

(ii) *For any sequence $\{\tilde{\theta}\}$ satisfying $s(\tilde{\theta}) = 0$ and $\tilde{\theta} \to_p \theta$,*

$$\sqrt{n}(\tilde{\theta} - \theta) \to_d N\left(0, (\Gamma^T \Sigma^{-1} \Gamma)^{-1}\right),$$

where $\to_d$ denotes convergence in distribution, Γ is given in (C4), *and Σ is the positive definite matrix with $E[g_l(\theta, y, \delta, \boldsymbol{x}) g_{l'}(\theta, y, \delta, \boldsymbol{x})]$ as its (l, l')th element, $1 \leq l, l' \leq p + q + J$.*

(iii) *If $\hat{\Gamma}$ is the $(p + q + J) \times (p + 2)$ matrix whose lth row is*

$$\frac{1}{n} \sum_i \left. \frac{\partial g_l(\vartheta, y_i, \delta_i, \boldsymbol{x}_i)}{\partial \vartheta} \right|_{\vartheta = \hat{\theta}}$$

and $\hat{\Sigma}$ is the $L \times L$ matrix whose (l, l')th element is

$$\frac{1}{n} \sum_i g_l(\hat{\theta}, y_i, \delta_i, \boldsymbol{x}_i) g_{l'}(\hat{\theta}, y_i, \delta_i, \boldsymbol{x}_i),$$

$\hat{\Gamma}^T \hat{\Sigma}^{-1} \hat{\Gamma} \to_p \Gamma^T \Sigma^{-1} \Gamma$.

(C3) *The parameter space Θ containing the true value θ is an open subset of $\mathcal{R}^{p+2}$, $E(\|\boldsymbol{u}\|^2 + \|\boldsymbol{t}\|^2) < \infty$, and there is a neighborhood N of θ such that*

$$E\left[\delta \sup_{\vartheta \in N} \{(1 + \|\boldsymbol{t}\|^2 + \|\boldsymbol{u}\|^2) w^2(\vartheta) + (1 + |y| + \|\boldsymbol{u}\|_1)(1 + \|\boldsymbol{t}\|_1 + \|\boldsymbol{u}\|_1) |w'(\vartheta)|\right.$$

$$+(1+y^2+\|\boldsymbol{u}\|^2)(1+\|\boldsymbol{t}\|_1+\|\boldsymbol{u}\|_1)|w''(\vartheta)|\}\Big] < \infty,$$

where $\|\cdot\|$ *is the* L_2*-norm,* $\|\cdot\|_1$ *is the* L_1*-norm,* $w(\vartheta) = [\Psi(\vartheta_1 + \vartheta_2 y + \vartheta_3 \boldsymbol{u}^T)]^{-1}$*, and* w' *and* w'' *are the first and second order derivatives of* $w(\cdot)$*.*

(C4) *The* $(p+q+J)\times(p+2)$ *matrix*

$$\Gamma = \begin{bmatrix} E[\delta \boldsymbol{d}^T w'(\theta)] & E[\delta y \boldsymbol{d}^T w'(\theta)] & E[\delta \boldsymbol{d}^T \boldsymbol{u} w'(\theta)] \\ E[\delta \boldsymbol{t}^T w'(\theta)] & E[\delta y \boldsymbol{t}^T w'(\theta)] & E[\delta \boldsymbol{t}^T \boldsymbol{u} w'(\theta)] \\ E[\delta \boldsymbol{u}^T w'(\theta)] & E[\delta y \boldsymbol{u}^T w'(\theta)] & E[\delta \boldsymbol{u}^T \boldsymbol{u} w'(\theta)] \end{bmatrix}$$

is of full rank.

The asymptotic covariance matrix $(\Gamma^T\Sigma^{-1}\Gamma)^{-1}$ is much simpler than that of the empirical likelihood estimator in Qin, Leung, and Shao (2002), which enables us to obtain an easy-to-compute covariance matrix estimator $(\hat{\Gamma}^T\hat{\Sigma}^{-1}\hat{\Gamma})^{-1}$. Kott and Chang (2010) also derived the asymptotic normality of the calibration estimator and its asymptotic covariance matrix, but they required that y given $\boldsymbol{x}$ follows a linear model.

Consider some special cases in which (C3) or (C4) can be simplified. First, consider that $\boldsymbol{x} = z$ is a discrete nonresponse instrument. Then (C4) can be simplified to the condition that there exist at least j_1 and j_2 in $\{1,\ldots,J\}$ such that

$$\frac{E[yw'(\theta)|\delta=1, z=j_1]}{E[w'(\theta)|\delta=1, z=j_1]} \neq \frac{E[yw'(\theta)|\delta=1, z=j_2]}{E[w'(\theta)|\delta=1, z=j_2]}. \tag{3.4}$$

This condition can be empirically checked using observed y_i's and the covariate d_i's. In this case, if $\Psi(t) = [1+\exp(t)]^{-1}$ is logistic, then (C3) simplifies to $E(\delta y^2) < \infty$ and $E[\delta \exp(\{2\beta \pm \epsilon\}y)] < \infty$ for some $\epsilon > 0$.

A second special case has $\boldsymbol{z} = z$ discrete and $\boldsymbol{u} = u$ a univariate continuous covariate. The full rank assumption on Γ is the key for the results in Theorem 2 and it is implied by any of the following conditions.

(1) $J \geq 3$ and the points (a_j, b_j), $j = 1,\ldots,J$, are not on the same line, where

$$a_j = \frac{E[yw'(\theta)|\delta=1, z=j]}{E[w'(\theta)|\delta=1, z=j]} \quad \text{and} \quad b_j = \frac{E[uw'(\theta)|\delta=1, z=j]}{E[w'(\theta)|\delta=1, z=j]}. \tag{3.5}$$

(2) Condition (3.4) holds, $E[uw'(\theta)|\delta=1] \neq 0$, and the points (a_j, b_j), $j = 1,\ldots, J+1$, are not on the same line, where $a_{J+1} = E[yuw'(\theta)|\delta=1]/E[uw'(\theta)|\delta=1]$, $b_{J+1} = E[u^2w'(\theta)|\delta=1]/E[uw'(\theta)|\delta=1]$, and a_j and b_j, $j = 1,\ldots,J$, are given in (3.5).

(3) Condition (3.4) holds, $E[uw'(\theta)|\delta = 1] = 0$, and either $E[yuw'(\theta)|\delta = 1] = 0$ or

$$\frac{E[u^2w'(\theta)|\delta = 1]}{E[yuw'(\theta)|\delta = 1]} \neq \frac{b_{j_1} - b_{j_2}}{a_{j_1} - a_{j_2}},$$

where j_1 and j_2 are given in (3.4) and a_j and b_j are given in (3.5).

Any of (1)–(3) can be empirically checked using observed data.

Once $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\gamma})$ is obtained, we can estimate the marginal distribution of y by the empirical distribution putting mass p_i on each observed y_i, where p_i is proportional to $\delta_i/\Psi(\hat{\alpha} + \hat{\beta}y_i + \hat{\gamma}\boldsymbol{u}_i^T)$.

We consider the estimation of the population mean $\mu = E(y)$. Once we have estimators $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$, μ can be estimated by

$$\tilde{\mu}_1 = \frac{1}{n}\sum_i \frac{\delta_i y_i}{\Psi(\hat{\alpha} + \hat{\beta}y_i + \hat{\gamma}\boldsymbol{u}_i^T)} \tag{3.6}$$

or by

$$\tilde{\mu}_2 = \sum_i \frac{\delta_i y_i}{\Psi(\hat{\alpha} + \hat{\beta}y_i + \hat{\gamma}\boldsymbol{u}_i^T)} \Big/ \sum_i \frac{\delta_i}{\Psi(\hat{\alpha} + \hat{\beta}y_i + \hat{\gamma}\boldsymbol{u}_i^T)}. \tag{3.7}$$

When the number of functions in (3.3) is more than the number of parameters, we can actually obtain a better estimator of μ using the GMM after adding the function $h(\mu, \vartheta, y, \delta, \boldsymbol{x}) = \mu - \delta y w(\vartheta)$, to (3.3). The parameters in this GMM are μ, α, β, and γ, and the number of equations is $L = p + q + J + 1$. The resulting GMM estimators $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$ are the same as those obtained by solving $R = p + q + J$ equations, but $\hat{\mu}$ is different from $\tilde{\mu}_1$ in (3.6) or $\tilde{\mu}_2$ in (3.7). The difference is due to the weight matrix in the second step of the GMM. Let W_R be the optimal weight matrix for the GMM based on the R functions in (3.3). After adding $h(\mu, \vartheta, y, \delta, \boldsymbol{x})$, we can easily show that $(\tilde{\mu}_1, \hat{\alpha}, \hat{\beta}, \hat{\gamma})$ is the GMM estimators based on $R + 1$ equations and the weight matrix

$$\tilde{W}_{R+1} = \begin{bmatrix} W_R & 0 \\ 0 & 1 \end{bmatrix}. \tag{3.8}$$

The weight matrix in (3.8) is not necessarily optimal. If $(\hat{\mu}, \hat{\alpha}, \hat{\beta}, \hat{\gamma})$ is the GMM estimator obtained using the two-step algorithm with the $L = R + 1$ functions, then the following result holds and $\hat{\mu}$ is asymptotically more efficient than $\tilde{\mu}_1$ unless $\tilde{W}_{R+1}$ in (3.8) is optimal.

Corollary 2. *Assume the conditions in Theorem 2,* $E(y^2) < \infty$, *and*

$$E\Big[\delta \sup_{\vartheta \in N} \{y^2w^2(\vartheta) + y^2|w'(\vartheta)| + |y|^3|w''(\vartheta)|\}\Big] < \infty.$$

Let $(\hat{\mu}, \hat{\alpha}, \hat{\beta}, \hat{\gamma})$ *be the GMM estimator obtained using the two-step algorithm after adding* $h(\mu, \vartheta, y, \delta, \boldsymbol{x}) = \mu - \delta y w(\vartheta)$ *to the set of functions in* (3.3). *Then the result in Theorem* 2 *holds with* θ *and* $\hat{\theta}$ *replaced by* $(\mu, \alpha, \beta, \gamma)$ *and* $(\hat{\mu}, \hat{\alpha}, \hat{\beta}, \hat{\gamma})$, *respectively.*

Consider now the general situation where both $\boldsymbol{z}$ and $\boldsymbol{u}$ may have continuous and discrete components. Let $\boldsymbol{u} = (\boldsymbol{v}, s)$ and $\boldsymbol{z} = (\boldsymbol{t}, z)$, where $\boldsymbol{v}$ and $\boldsymbol{t}$ are continuous and s and z are discrete taking values $1, \ldots, K$ and $1, \ldots, J$, respectively. Assume (C1) with (2.3) replaced by

$$P(\delta = 1 | y, \boldsymbol{x}) = \Psi(\alpha_s + \beta_s y + \gamma_s \boldsymbol{v}^T).$$

For each $k = 1, \ldots, K$, we consider the category defined by $s = k$ and apply the GMM for the estimation of $\theta_k = (\alpha_k, \beta_k, \gamma_k)$ using

$$\boldsymbol{g}(\vartheta, y, \delta, \boldsymbol{x}) = \begin{pmatrix} \boldsymbol{d}^T[\delta w(\vartheta) - 1] \\ \boldsymbol{t}^T[\delta w(\vartheta) - 1] \\ \boldsymbol{v}^T[\delta w(\vartheta) - 1] \end{pmatrix},$$

where $\boldsymbol{d}$ is defined by (3.3) , $w(\vartheta) = [\Psi(\vartheta_1 + \vartheta_2 y + \vartheta_3 \boldsymbol{v}^T)]^{-1}$, and $\vartheta = (\vartheta_1, \vartheta_2, \vartheta_3)$.

Let n_k be the number of sample units in the category defined by $s = k$. The unconditional distribution of y can be estimated by the weighted average of these K empirical distributions with weights proportional to n_k. Using these estimated distributions, we can estimate parameters in $p(y|\boldsymbol{x})$ or $p(y)$. Asymptotic results for these estimators similar to those in Theorem 2 and Corollary 2 can be established.

4. Empirical Results

We present some results from a simulation study with normally distributed data. Then we apply the proposed method to a data set from the Korean Labor and Income Panel Survey (KLIPS). Finally, we consider another simulation study using a similar population to the real data set. In the two-step algorithm of GMM for the estimation of α, β, γ, and the overall mean $\mu = E(y)$, we used the MATLAB function `fminsearch` to minimize the objective functions $G(\vartheta)^T G(\vartheta)$ and $G(\vartheta)^T \hat{W} G(\vartheta)$. In all numerical studies, the initial values for α, β, and γ were 0, and the initial value of μ was the naive estimate $\breve{\mu}$, the sample mean of the observed y_i's. For the estimation of μ, we compared the proposed GMM estimator $\hat{\mu}$ (Corollary 2), the naive estimate $\breve{\mu}$, the estimators $\tilde{\mu}_1$ and $\tilde{\mu}_2$ given in (3.6) and (3.7), respectively, and $\hat{\mu}_{\mathrm{EL}}$, the estimator of μ based on the empirical likelihood method in Qin, Leung, and Shao (2002). The MATLAB function `fsolve` was used to solve the empirical likelihood estimation equations.

4.1. Simulation from normal populations

We considered $n = 500$ or 2,000 and three populations, each with two sets of parameter values. In the first population, we took $\boldsymbol{x} = z$ as a discrete nonresponse instrument having $J = 2$ categories with $P(z = 1) = 0.4$ and $P(z = 2) = 0.6$. Conditional on z, $y \sim N(20 + 10z, 4^2)$ with unconditional mean 36. Given the generated data, the nonrespondents were generated according to $P(\delta = 1|y, z) = [1 + \exp(\alpha + \beta y)]^{-1}$, where $(\alpha, \beta) = (1, -0.05)$ or $(-2.6, 0.05)$. These values were chosen so that β had different signs. The unconditional nonresponse probability was approximately between 30% and 40%.

The second population was similar. The discrete nonresponse instrument z had $J = 3$ categories with $P(z = 1) = 0.3$, $P(z = 2) = 0.3$, and $P(z = 3) = 0.4$. Given z, $y \sim N(20 + 10z, 4^2)$, with unconditional mean 41. The nonresponse mechanism is $P(\delta = 1|y, z) = [1 + \exp(\alpha + \beta y)]^{-1}$, where $(\alpha, \beta) = (1.2, -0.05)$ or $(-2.6, 0.05)$.

In the last population, a continuous covariate u was added, $\boldsymbol{x} = (u, z)$, while z was the same as in the second case. Given z, $u \sim N(100z, 40^2)$. Given $z = 1$ and u, $y \sim N(u, 20^2)$; given $z = 2$ and u, $y \sim N(1.5u, 20^2)$; given $z = 3$ and u, $y \sim N(300 + 0.5u, 20^2)$. The unconditional mean of y was 300. The nonresponse mechanism was $P(\delta = 1|y, u, z) = [1 + \exp(\alpha + \beta y + \gamma u)]^{-1}$, where $(\alpha, \beta, \gamma) = (0.4, -0.002, -0.003)$ or $(-2, 0.002, 0.003)$.

Table 1 reports the following, based on 2,000 simulations: the bias of the GMM estimates, $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$ (for the last case only), $\hat{\mu}$, the naive estimate $\check{\mu}$, $\tilde{\mu}_1$ in (3.6), $\tilde{\mu}_2$ in (3.7), and the empirical likelihood estimate $\hat{\mu}_{\mathrm{EL}}$ (Qin, Leung, and Shao (2002)); the standard deviation (SD) of GMM estimates, $\check{\mu}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$ and $\hat{\mu}_{\mathrm{EL}}$; the standard error (SE) for GMM estimates, the estimated SD using the squared root of the diagonal elements in the matrix $n^{-1}(\hat{\Gamma}^T\hat{\Sigma}^{-1}\hat{\Gamma})^{-1}$ given in Theorem 2(iii), and for $\check{\mu}$ using the sample standard deviation; the coverage probability (CP) of the approximate 95% confidence intervals $[\hat{\mu} - 1.96\mathrm{SE}, \hat{\mu} + 1.96\mathrm{SE}]$ and $[\check{\mu} - 1.96\mathrm{SE}, \check{\mu} + 1.96\mathrm{SE}]$. The values of parameters and J are also included in Table 1.

The simulation results in Table 1 support the asymptotic results for the GMM estimators as well as the consistency of the variance estimators. When there is no covariate u, the GMM estimators work well for $J = 2$ and $J = 3$, although performance is generally better when $J = 3$. The coverage probabilities of the confidence intervals based on $\hat{\mu}$ are all close to the nominal level 95%. The naive estimator $\check{\mu}$ has a positive bias when $\beta < 0$ (larger y has smaller nonresponse probability) and has a negative bias when $\beta > 0$ (larger y has larger nonresponse probability). Although the bias of $\check{\mu}$ may be small compared with the value of μ, it is not small compared with the SD so that it leads to a poor performance of the confidence interval based on $\check{\mu}$. The performance of $\tilde{\mu}_1$, $\tilde{\mu}_2$

Table 1. Simulation results for normal populations.

$n = 2,000$													
	Parameter					Estimate							
J	μ	α	β	γ		$\hat{\mu}$	$\check{\mu}$	$\tilde{\mu}_1$	$\tilde{\mu}_2$	$\hat{\mu}_{\text{EL}}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$
2	36	1	-0.05	0	Bias	0.0054	0.6217				-0.0148	$3.7/10^4$	
					SD	0.1607	0.1646				0.3671	0.0104	
					SE	0.1616	0.1675				0.3523	0.0100	
					CP	94.3%	4.8%						
2	36	-2.6	0.05	0	Bias	0.0018	-0.6115				-0.0222	$4.7/10^4$	
					SD	0.1657	0.1743				0.3908	0.0105	
					SE	0.1622	0.1708				0.3847	0.0103	
					CP	94.3%	4.8%						
3	41	1.2	-0.05	0	Bias	-0.0026	1.2686	-0.0210	-0.0055	-0.0052	-0.0063	$1.2/10^4$	
					SD	0.2207	0.2371	0.2243	0.2214	0.2208	0.2395	0.0061	
					SE	0.2169	0.2417				0.2426	0.0061	
					CP	94.8%	0.0%						
3	41	-2.6	0.05	0	Bias	0.0025	-1.5214	-0.0125	0.0066	0.0025	-0.0321	$6.4/10^4$	
					SD	0.2227	0.2539	0.2225	0.2225	0.2210	0.2717	0.0062	
					SE	0.2207	0.2581				0.2564	0.0059	
					CP	94.1%	0.0%						
3	300	0.4	-0.002	-0.003	Bias	-0.0591	26.9383	-0.2354	-0.0729	-0.0596	0.0086	$5.8/10^5$	$-1.5/10^4$
					SD	3.4081	3.8300	3.4295	3.4091	3.4032	0.1447	0.0018	0.0030
					SE	3.4389	3.8950				0.1429	0.0018	0.0029
					CP	94.6%	0.0%						
3	300	-2	0.002	0.003	Bias	-0.2169	-26.5431	-0.3235	-0.2149	-0.2213	-0.0057	$3.3/10^5$	$5.8/10^5$
					SD	3.3770	4.1205	3.2911	3.4084	3.2961	0.1626	0.0016	0.0024
					SE	3.4413	4.1872				0.1442	0.0015	0.0024
					CP	95.2%	0.0%						
$n = 500$													
2	36	1	-0.05	0	Bias	0.0228	0.6249				-0.0452	0.0011	
					SD	0.3189	0.3429				0.7003	0.0198	
					SE	0.3224	0.3345				0.7110	0.0201	
					CP	94.2%	55.4%						
2	36	-2.6	0.05	0	Bias	0.0025	-0.6166				-0.0517	0.0012	
					SD	0.3354	0.3420				0.7858	0.0209	
					SE	0.3247	0.3418				0.7744	0.0207	
					CP	94.3%	4.8%						
3	41	1.2	-0.05	0	Bias	0.0338	1.2961	-0.0400	0.0195	0.0218	0.0198	$-6.7/10^4$	
					SD	0.4259	0.4775	0.4368	0.4258	0.4240	0.4914	0.0123	
					SE	0.4324	0.4818				0.4916	0.0123	
					CP	95.5%	23.5%						
3	41	-2.6	0.05	0	Bias	-0.0039	-1.5356	-0.0642	0.0015	-0.0087	-0.0701	0.0014	
					SD	0.4382	0.5138	0.4485	0.4448	0.4380	0.5683	0.0130	
					SE	0.4410	0.5165				0.5374	0.0120	
					CP	94.7%	15.1%						
3	300	0.4	-0.002	-0.003	Bias	-0.1934	26.704	-0.8013	-0.2047	-0.1449	0.0192	$2.9/10^4$	$-7.6/10^4$
					SD	6.7338	7.4839	6.7379	6.7295	6.7148	0.2988	0.0035	0.0059
					SE	6.8748	7.8109				0.2900	0.0036	0.0060
					CP	95.5%	7.4%						
3	300	-2	0.002	0.003	Bias	-0.0200	-25.710	-0.3105	0.2197	0.0541	-0.0065	$-5.8/10^5$	$1.4/10^4$
					SD	6.8694	8.1686	6.7988	7.0382	6.7577	0.3376	0.0032	0.0049
					SE	6.8721	8.3815				0.2959	0.0031	0.0049
					CP	94.3%	12.5%						

and $\hat{\mu}_{\mathrm{EL}}$ are similar to that of $\hat{\mu}$ in terms of both bias and standard deviation, indicating that the weight matrix in (3.8) is nearly optimal. When the number of equations is equal to the number of parameters ($J = 2$ case), they are all identical.

4.2. Estimates for the KLIPS data

We applied the proposed method to a data set from the KLIPS. A brief description of this survey can be found at

`http://www.kli.re.kr/klips/en/about/introduce.jsp`.

The data set consists of $n = 2,506$ regular wage earners. The variable of interest, y, is the monthly income in 2006. Covariates associated with y are gender, age group, level of education, and the monthly income in 2005. The variable y has about 35% missing values while all covariate values are observed.

To apply the proposed method, we first used the income in 2005 as a continuous covariate u and the age, gender, and education levels as a discrete nonresponse instrument z. Thus we assumed that these covariates are related to y and u but they are not related to the nonresponse once y and u are given. Unconditionally, these covariates may still be related to the nonresponse. We took age$<$35, 35$\leq$age$<$51, and age$\geq$51, gender as male and female, and education up to high school or beyond. Therefore, z had $3\times2\times2=12$ categories. We assumed model (3.2) with $\Psi(t)=[1+\exp(t)]^{-1}$. The naive estimate $\breve{\mu}$, the GMM estimates, their SE's, $\tilde{\mu}_1$ in (3.6), $\tilde{\mu}_2$ in (3.7), and the empirical likelihood estimator $\hat{\mu}_{\mathrm{EL}}$ were as follows.

	$\breve{\mu}$	$\hat{\mu}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\tilde{\mu}_1$	$\tilde{\mu}_2$	$\hat{\mu}_{\mathrm{EL}}$
Estimate	205.71	184.55	0.6932	−0.0072	−0.0004	183.85	184.77	184.59
SE	2.7407	2.8468	0.1685	0.0024	0.0016			

Here $\hat{\mu}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, and $\hat{\mu}_{\mathrm{EL}}$ are close together but they are significantly different from $\breve{\mu}$. Since $\hat{\gamma}$ is not significantly different from 0, we set $\gamma = 0$ in (3.2) and (3.3) and computed the GMM estimates again. The results were as follows.

	$\hat{\mu}$	$\hat{\alpha}$	$\hat{\beta}$	$\tilde{\mu}_1$	$\tilde{\mu}_2$	$\hat{\mu}_{\mathrm{EL}}$
Estimate	184.00	0.7244	−0.0073	183.42	184.35	184.64
SE	2.2735	0.1333	0.0008			

Next, we carried out a sensitivity analysis to see whether results were sensitive to different choices of nonresponse instrument $\boldsymbol{z}$. The following table reports the mean estimates and SE's under different cases, with $\boldsymbol{u} = (u, s)$, u and s continuous and discrete covariates that are related to nonresponse even if y is given, and z is a discrete nonresponse instrument.

u	s	z	$\hat{\mu}$	SE
2005 income		Age, Education, Gender	184.55	2.8468
2005 income	Gender	Age, Education	185.54	3.4032
2005 income	Age	Education, Gender	183.58	3.0727
2005 income	Education	Age, Gender	183.89	3.3423
2005 income	Education, Gender	Age	196.56	6.5353
2005 income	Age, Education	Gender	186.07	4.0512
2005 income	Age, Gender	Education	188.36	4.7384

The results are about the same except for those when the age group is the only covariate used as nonresponse instrument. We think that the age group is not a useful predictor of the 2006 income given 2005 income, education, and gender. It results in a too large $\hat{\mu}$ as well as a large SE.

4.3. Simulation for the KLIPS population

To examine whether estimates for the KLIPS data are adequate, we carried out a simulation study using a population similar to the KLIPS data set with 2005 income treated as a continuous u and the categorical variable formed by age, gender and education treated as z. First, we took an independent probability proportional to $1+\exp(\hat{\alpha}+\hat{\beta}y_i+\hat{\gamma}u_i)$ sample $\mathcal{M} = \{(y_i^*, u_i^*, z_i^*), i = 1, \ldots, 2,506\}$ with replacement from the set of subjects in the KLIPS data set with observed y_i's, where $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$ are the GMM estimates obtained in Section 4.2. Then, we generated independently $n = 2500$ vectors, $(\tilde{y}_i, \tilde{u}_i, \tilde{z}_i)$, by first taking a simple random sample $\mathcal{S} = \{(y_i^{**}, u_i^{**}, z_i^{**}), i = 1, \ldots, n\}$ with replacement from $\mathcal{M}$ and then setting $\tilde{y}_i = y_i^{**} + \epsilon_{yi}$, $\tilde{u}_i = u_i^{**} + \epsilon_{ui}$, and $\tilde{z}_i = z_i^{**}$, $i = 1, \ldots, n$, where ϵ_{yi}, ϵ_{ui}, $i = 1, \ldots, n$, are independent and normally distributed with mean 0 and standard deviation 10.6 (about 1/10 of the standard deviation of y_i^*'s in $\mathcal{M}$). The population mean of $\tilde{y}_i$ is 185.85.

The nonrespondents were generated according to (3.2) with $\Psi(t) = [1 + \exp(t)]^{-1}$. We first considered two sets of parameter values: $\alpha = 0.6932$ and $\beta = -0.0072$ as the GMM estimates in Section 4.2; $\gamma = 0$ in the first set; and $\gamma = -0.0004$ is the GMM estimate in Section 4.2 in the second set. For each set of parameters, we computed

I. GMM estimates using (3.2) and (3.3).

II. GMM estimates using (3.2) and (3.3) but setting $\gamma = 0$.

Table 2 reports the quantities in Table 1 based on 2,000 simulations. All GMM estimators performed well when a correct model on the nonresponse mechanism was used. When method II was used and $\gamma = 0$, the GMM estimators were more efficient. In the case where $\gamma = -0.0004$ but method II was used, the

Table 2. Simulation results for the KLIPS population ($\mu = 185.85$)

	Parameter				Estimate							
	α	β	γ		$\hat{\mu}$	$\check{\mu}$	$\tilde{\mu}_1$	$\tilde{\mu}_2$	$\hat{\mu}_{\mathrm{EL}}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$
I	0.6932	-0.0072	0	Bias	0.7299	21.293	-0.1416	0.7638	0.5056	-0.0352	0.0010	-0.0010
				SD	3.2211	2.7858	3.2246	3.2193	3.2774	0.1639	0.0034	0.0030
				SE	3.1050	2.7871				0.1558	0.0031	0.0026
				CP	93.8%	0.0%						
II	0.6932	-0.0072	0	Bias	0.084		-0.8404	0.126	0.183	-0.019	$3.2/10^5$	
				SD	2.2191		2.2525	2.2242	2.7363	0.1307	0.0008	
				SE	2.3099					0.1287	0.0008	
				CP	94.5%							
I	0.6932	-0.0072	-0.0004	Bias	0.934	21.359	0.038	0.986	0.659	-0.030	0.0011	-0.0011
				SD	3.2087	2.8322	3.2044	3.1994	3.2651	0.1623	0.0033	0.0029
				SE	3.0652	2.7514				0.1589	0.0031	0.0026
				CP	94.6%	0.0%						
II	0.6932	-0.0072	-0.0004	Bias	-0.199		-1.1151	-0.0911	0.2340	0.012	$-5.7/10^4$	
				SD	2.4080		2.4304	2.4040	2.8119	0.1362	0.0008	
				SE	2.2982					0.1325	0.0008	
				CP	92.2%							
I	1.3	-0.003	-0.01	Bias	0.5756	24.669	-0.4070	0.6464	0.2181	-0.0222	$-9.0/10^5$	$2.3/10^5$
				SD	3.2077	2.7148	3.1817	3.2277	3.3850	0.1607	0.0044	0.0043
				SE	3.1528	2.6868				0.1606	0.0045	0.0043
				CP	94.6%	0.0%						
II	1.3	-0.003	-0.01	Bias	-5.4613		-6.5258	-4.4156	-1.9204	0.2123	-0.0113	
				SD	2.2861		2.2974	2.3253	2.6550	0.1587	0.0012	
				SE	2.3142					0.1714	0.0013	
				CP	33.4%							

I: The GMM using (3.2) and (3.3)
II: The GMM using (3.2) and (3.3) but setting $\gamma = 0$

GMM estimators were biased in theory, but still performed well because γ was very small.

To see the effect of incorrectly setting $\gamma = 0$, we carried out the simulation with another set of parameters, $(\alpha, \beta, \gamma) = (1.3, -0.003, -0.01)$. The results are also included in Table 2. Although the GMM estimator based on method I performed well, the GMM estimator based on method II had some biases that resulted in a poor coverage probability of the confidence interval. Similar to the simulation study in Section 4.1, estimators $\tilde{\mu}_1$ in (3.6) and $\tilde{\mu}_2$ in (3.7) have similar performance compared with $\hat{\mu}$. The empirical likelihood estimator $\hat{\mu}_{\mathrm{EL}}$ is slightly worse than the other estimators under this simulation setting. We believe that this is caused by the fact that the numerical solution in empirical likelihood may not be stable when the number of equations (constraints in empirical likelihood) is not small ($J = 8$ in this study compared with $J = 2$ or 3 in Section 4.1).

To summarize, the GMM estimators and their standard deviation estimators have good performance when the nonresponse mechanism model is correct. The naive estimator based on observed y_i's can be seriously biased. The GMM estimators are in general sensitive to the misspecification of the nonresponse

mechanism, although, for the KLIPS data, treating γ in (3.2) and (3.3) as 0 does not create significant biases.

5. Concluding Remarks

We consider parameter estimation under nonignorable nonresponse assuming a parametric model for the probability $P(\delta = 1|y, \boldsymbol{x})$ but without assuming any parametric model for $p(y|\boldsymbol{x})$. The crucial part is to estimate the parameter θ in the nonresponse mechanism. For ignorable nonresponse, the parameters in $P(\delta = 1|y, \boldsymbol{x})$ can be consistently estimated using observed data; see, for example, Nevo (2003) and Chang and Kott (2008). For nonignorable nonresponse, even if both $P(\delta = 1|y, \boldsymbol{x})$ and $p(y|\boldsymbol{x})$ are parametric, θ may not be identifiable without additional auxiliary information. In this paper, we consider the situation where some auxiliary information is provided by a nonresponse instrument $\boldsymbol{z}$ that is useful in predicting the study variable y but is conditionally independent of the response indicator δ given y and values of some other covariates. We show how to use this nonresponse instrument to construct estimating equations for the GMM estimators of θ and other parameters. Consistency and asymptotic normality of the GMM estimators are established. The proposed nonresponse instrument approach sheds light on how to use an auxiliary variable to avoid the notorious nonidentifiability problem associated with the nonignorable nonresponse.

The use of a nonresponse instrument proposed in this paper is different from the approach of using a surrogate variable considered in Chen, Leung, and Qin (2008), which requires that $P(\delta = 1|y, \boldsymbol{x}, \boldsymbol{s}) = P(\delta = 1|\boldsymbol{x}, \boldsymbol{s})$ for an observed surrogate variable $\boldsymbol{s}$. This requirement means that conditional on the surrogate variable $\boldsymbol{s}$, the nonresponse mechanism becomes ignorable, since $(\boldsymbol{x}, \boldsymbol{s})$ is always observed. However, it may not be easy to find a suitable surrogate variable to satisfy the requirement on the nonresponse mechanism.

Once a consistent estimator $\hat{\theta}$ is obtained, consistent estimators of unknown quantities in $p(y|\boldsymbol{x})$, such as the mean and the distribution function of y, can be obtained by a weighted mean where the weights are proportional to the inverse of the estimated response probabilities. We also show that efficient estimators may be obtained by applying the GMM with some additional estimating functions related to the quantities to be estimated.

To apply the proposed method, one needs to carefully choose a nonresponse instrument $\boldsymbol{z}$ among a set of covariates to meet the conditions (i) $\boldsymbol{z}$ is related to the study variable y and (ii) $\boldsymbol{z}$ can be excluded from the nonresponse mechanism when y and some other covariates $\boldsymbol{u}$ are included. In the KLIPS, for example, $\boldsymbol{z} =$ (age group, gender, level of education) is related to the 2006 monthly income, but when the 2006 monthly income as well as the 2005 monthly income are included in the nonresponse mechanism for the 2006 monthly income, it is likely that $\boldsymbol{z}$

or part of $\boldsymbol{z}$ is not needed for the nonresponse model and, hence, $\boldsymbol{z}$ can be used as a nonresponse instrument.

The assumption on the nonresponse mechanism is crucial to the proposed GMM estimators. Unfortunately, we are not able to check this assumption due to the presence of missing values. This issue also exists for methods developed under the ignorable nonresponse assumption, since we are not able to check the ignorable nonresponse assumption using observed data only. It is then important to develop methods under various assumptions. The results can be compared in applications and are useful for a sensitivity analysis.

While the proposed method provides a useful tool for handling the nonignorable nonresponse, there is no guarantee that the proposed GMM estimators are optimal. Since estimating functions for the GMM are not uniquely defined. The function in (3.3), for example, is related to the first order moments of covariates. Other moments or characteristics of $\boldsymbol{u}$ and $\boldsymbol{z}$ may provide more information and, hence, result in more efficient GMM estimators. Finding more efficient GMM estimators or other types of estimators under the setup of this paper is a topic of our future research.

Acknowledgements

We would like to thank the anonymous referees and an associate editor for helpful comments and suggestions. The research of the second author was partially supported by the NSF Grant DMS-1007454. The research of the third author was partially supported by USDA NRCS CESU agreement 68-7482-11-534.

Appendix

Proof of Theorem 1. Since we consider the conditional distribution for given $\boldsymbol{u}$, we can treat $\boldsymbol{u}$ as fixed and omit $\boldsymbol{u}$ in the notation. Let $f_1(y)$ and $f_2(y)$ be the Lebesgue density functions in (C2). To show identifiability, it suffices to show that, if

$$\begin{aligned} \Psi(\alpha+\beta y)f_1(y) &= \Psi(\alpha'+\beta' y)f_1'(y) \\ \Psi(\alpha+\beta y)f_2(y) &= \Psi(\alpha'+\beta' y)f_2'(y) \end{aligned} \quad \text{for all } y\in\mathcal{R}, \tag{A.1}$$

then $\alpha=\alpha'$, $\beta=\beta'$, $f_1=f_1'$, and $f_2=f_2'$. Since f_1, f_2, f_1', and f_2' are density functions, (A.1) implies

$$\int\Big[\frac{\Psi(\alpha+\beta y)}{\Psi(\alpha'+\beta' y)}-1\Big]f_1(y)dy=\int\Big[\frac{\Psi(\alpha+\beta y)}{\Psi(\alpha'+\beta' y)}-1\Big]f_2(y)dy=0. \tag{A.2}$$

We now show in two steps that (A.2) implies $\beta=\beta'$.

Step I. We prove that, when $\beta \neq \beta'$, the function $K(y) = [\Psi(\alpha + \beta y)/\Psi(\alpha' + \beta' y)] - 1$ has a single change of sign. Under (C1), Ψ is strictly monotone. We consider a strictly decreasing Ψ, proof for a strictly increasing Ψ is similar. Now if one of β and β' is 0 or if β and β' have different signs, then $K(y)$ is a strictly monotone function having a unique root and, hence, it has a single change of sign.

It remains to consider the case where β and β' have the same sign, say $\beta' > \beta > 0$. Let $y^* = (\alpha - \alpha')/(\beta' - \beta)$. Since $\alpha + \beta y^* = \alpha' + \beta' y^*$, we have $K(y^*) = 0$. For any $y > y^*$,

$$\alpha + \beta y = \alpha + \beta y^* + \beta(y - y^*) < \alpha + \beta y^* + \beta'(y - y^*) = \alpha' + \beta' y. \tag{A.3}$$

Since Ψ is strictly decreasing, it follows from (A.3) that $\Psi(\alpha + \beta y) > \Psi(\alpha' + \beta' y)$ and, therefore, $K(y) = [\Psi(\alpha + \beta y)/\Psi(\alpha' + \beta' y)] - 1 > 0$. Similarly, when $y < y^*$, $K(y) < 0$. This proves that $K(y)$ has a single change of sign.

Step II. We prove that, if $\beta \neq \beta'$ and if the first integral in (A.2) is 0, then the second integral is not 0. Let X be a random variable having f_1 or f_2 as its probability density and let E_j denote the expectation when X has density f_j. We show that if $E_1[K(X)] = 0$, then $E_2[K(X)] \neq 0$, where K is the function defined in Step I with $\beta \neq \beta'$.

Let $K(x) < 0$ if $x < x_0$ and $K(x) > 0$ if $x > x_0$, and put

$$c = \sup_{x < x_0} \frac{f_2(x)}{f_1(x)}.$$

Under (C2), $f_2(y)/f_1(y)$ is a nondecreasing function of y. Hence, when $f_1(x_0) > 0$, $c = f_2(x_0)/f_1(x_0) < \infty$. When $f_1(x_0) = 0$, because $E_1[K(X)] = 0$, there exists x_1 such that $x_1 > x_0$ and $f_1(x_1) > 0$, which implies that $c \leq f_2(x_1)/f_1(x_1) < \infty$. Thus, $c < \infty$. Write

$$E_2(K(X)) = \int K(x) f_2(x) dx = \int_A K(x) f_2(x) dx + \int_B K(x) f_2(x) dx,$$

where $A = \{x : f_1(x) = 0, f_2(x) > 0\}$ and $B = \{x : f_1(x) > 0, f_2(x) > 0\} \cup \{x : f_1(x) > 0, f_2(x) = 0\}$. If $x \in A$, then $f_2(x)/f_1(x) = \infty$ and, therefore, $x > x_0$. This shows that $K(x) > 0$ for $x \in A$ and $\int_A K(x) f_2(x) \geq 0$. Then

$$\begin{aligned}
E_2[K(X)] &\geq \int_B K(x) f_2(x) dx \\
&= \int_{B_1} K(x) f_2(x) dx + \int_{B_2} K(x) f_2(x) dx \\
&= \int_{B_1} K(x) \frac{f_2(x)}{f_1(x)} f_1(x) dx + \int_{B_2} K(x) \frac{f_2(x)}{f_1(x)} f_1(x) dx \\
&\geq \int_{B_1} cK(x) f_1(x) dx + \int_{B_2} cK(x) f_1(x) dx
\end{aligned}$$

$$= cE_1[K(X)]$$
$$= 0,$$

where $B_1 = \{x : x \in B, x < x_0\}$, $B_2 = \{x : x \in B, x > x_0\}$, and the last inequality follows from the definition of c and the fact that $K(x) < 0$ for $x \in B_1$ and $K(x) > 0$ for $x \in B_2$.

If A has a positive Lebesgue measure, then $\int_A K(x)f_2(x)dx > 0$ and, hence, $E_2[K(X)] > 0$. If A has Lebesgue measure 0, then the support sets of f_1 and f_2 are subsets of B. If $E_2[K(X)] = 0$, then $f_2(x) = cf_1(x)$ a.e. on B. Then, $c = 1$ because f_1 and f_2 are densities. This contradicts (C1). Therefore, we have $E_2[K(X)] > 0$.

Thus, (A.2) implies $\beta = \beta'$ and reduces to

$$\int \left[\frac{\Psi(\alpha + \beta y)}{\Psi(\alpha' + \beta y)} - 1\right] f_1(y)dy = \int \left[\frac{\Psi(\alpha + \beta y)}{\Psi(\alpha' + \beta y)} - 1\right] f_2(y)dy = 0,$$

which implies $\alpha = \alpha'$ since $\Psi(x)$ is a strictly monotone function. These results and (A.1) imply that $f_1 = f_1'$ and $f_2 = f_2'$, which shows identifiability.

Proof of Theorem 2.

(i) Suppose that $\tilde{W} \to_p W$, where W is a positive definite matrix. First, we prove that there exists $\bar{\theta}$ such that, as $n \to \infty$,

$$P(\tilde{s}(\bar{\theta}) = 0) \to 1 \qquad \text{and} \qquad \bar{\theta} \to_p \theta, \tag{A.4}$$

where $\tilde{s}(\vartheta) = -\partial[G^T(\vartheta)\tilde{W}G(\vartheta)]/\partial\vartheta$. Since Γ is of full rank and W is positive definite, $\Gamma^T W \Gamma$ is positive definite. Therefore, there exists a matrix A such that $A^2 = 2\Gamma^T W \Gamma$.

Define $Q(\vartheta) = G^T(\vartheta)\tilde{W}G(\vartheta)$. To prove (A.4), it suffices to prove that, for any $\epsilon > 0$, there exists $c > 0$ such that, for sufficiently large n,

$$P\{Q(\theta) - Q(\vartheta) < 0 \quad \text{for all } \vartheta \in B_n(c)\} \geq 1 - \epsilon, \tag{A.5}$$

where $B_n(c) = \{\vartheta : \|A(\vartheta - \theta)\| = c/\sqrt{n}\}$ and $\|A\| = \sqrt{\text{trace}(A^T A)}$ for a vector or matrix A. When n is large enough, $B_n(c)$ is inside the parameter space Θ and $B_n(c)$ shrinks to θ as $n \to \infty$. By Taylor's expansion, there exists θ^* between θ and ϑ such that

$$Q(\theta) - Q(\vartheta) = (\vartheta - \theta)^T \tilde{s}(\theta) + \frac{1}{2}(\vartheta - \theta)^T \nabla\tilde{s}(\theta^*)(\vartheta - \theta)$$
$$= \frac{c}{\sqrt{n}}\lambda^T A^{-1}\tilde{s}(\theta) + \frac{c^2}{2n}\lambda^T A^{-1}\nabla\tilde{s}(\theta^*)A^{-1}\lambda,$$

where $\nabla\tilde{s}(\vartheta) = \partial\tilde{s}(\vartheta)/\partial\vartheta$, $\lambda = \sqrt{n}A(\vartheta - \theta)/c$, and $\|\lambda\| = 1$ for $\vartheta \in B_n(c)$. Using (C3), $\tilde{W} \to_p W$, the proof of Theorem 2.6 in Newey and Mcfadden (1994) and

the fact that every component in $G(\vartheta)$, $\partial G(\vartheta)/\partial\vartheta$, and $\partial^2 G(\vartheta)/\partial\vartheta\partial\vartheta^T$ is an average over independent and identically distributed samples, we obtain that

$$\sup_{\vartheta\in N}\|\nabla\tilde{s}(\vartheta)-\psi(\vartheta)\|\to_p 0,$$

where $\psi(\vartheta)=-\partial^2\{E[G^T(\vartheta)]WE[G(\vartheta)]\}/\partial\vartheta\partial\vartheta^T$. Since $\psi(\theta)=-2\Gamma^TW\Gamma$,

$$\begin{aligned}\|\nabla\tilde{s}(\theta^*)-(-2\Gamma^TW\Gamma)\| &\le \|\nabla\tilde{s}(\theta^*)-\psi(\theta^*)\|+\|\psi(\theta^*)-(-2\Gamma^TW\Gamma)\|\\ &\le \sup_{\vartheta\in N}\|\nabla\tilde{s}(\vartheta)-\psi(\vartheta)\|+\|\psi(\theta^*)-\psi(\theta)\|\\ &\to_p 0\end{aligned}$$

by the continuity of ψ at θ. Hence $A^{-1}\nabla\tilde{s}(\theta^*)A^{-1}\to_p -I_{2\times 2}$. Then,

$$\begin{aligned}Q(\theta)-Q(\vartheta)&=\frac{1}{n}\left\{c\lambda^TA^{-1}\sqrt{n}\tilde{s}(\theta)-\frac{c^2}{2}[1+o_p(1)]\right\}\\ &\le\frac{1}{n}\left\{c\max_{\lambda}[\lambda^TA^{-1}\sqrt{n}\tilde{s}(\theta)]-\frac{c^2}{2}[1+o_p(1)]\right\}\\ &=\frac{1}{n}\left\{c\|A^{-1}\sqrt{n}\tilde{s}(\theta)\|-\frac{c^2}{2}[1+o_p(1)]\right\}.\end{aligned}\tag{A.6}$$

Let $\nabla G(\vartheta)=\partial G(\vartheta)/\partial\vartheta$. Then $A^{-1}\sqrt{n}\tilde{s}(\theta)=-2A^{-1}\nabla G(\theta)\tilde{W}\sqrt{n}G(\theta)$. Under (C3), $\nabla G(\theta)\to_p\Gamma$ by the Law of Large Numbers and $\sqrt{n}G(\theta)\to_d N(0,\Sigma)$ by the Central Limit Theorem. By the fact that $\tilde{W}\to_p W$,

$$A^{-1}\sqrt{n}\tilde{s}(\theta)\to_d N(0,4A^{-1}\Gamma^TW\Sigma W\Gamma A^{-1}).$$

Therefore, there exists a c such that $P(\|A^{-1}\sqrt{n}\tilde{s}(\theta)\|<c/4)\ge 1-\epsilon$. Now $\|A^{-1}\sqrt{n}\tilde{s}(\theta)\|<c/4$ and (A.6) imply $Q(\theta)-Q(\vartheta)<0$ for all $\vartheta\in B_n(c)$. Hence, result (A.5) follows and the proof of (A.4) is complete.

By (A.4) with $\tilde{W}=I_{L\times L}$, we obtain that $\hat{\theta}^{(1)}\to_p\theta$, which, combined with (C3), implies that $\hat{W}\to_p\Sigma^{-1}$. Then the result in (i) follows from (A.4) with $\tilde{W}=\hat{W}$ and $W=\Sigma^{-1}$, where Σ^{-1} is a positive definite matrix.

(ii) By Taylor's expansion, there exists a θ^* between θ and $\tilde{\theta}$ such that $G(\tilde{\theta})=G(\theta)+\nabla G(\theta^*)(\tilde{\theta}-\theta)$, which implies that

$$[\nabla G(\tilde{\theta})]^T\hat{W}G(\tilde{\theta})=[\nabla G(\tilde{\theta})]^T\hat{W}G(\theta)+[\nabla G(\tilde{\theta})]^T\hat{W}\nabla G(\theta^*)(\tilde{\theta}-\theta).$$

Since $-2[\nabla G(\tilde{\theta})]^T\hat{W}G(\tilde{\theta})=s(\tilde{\theta})=0$, we obtain that

$$\sqrt{n}(\tilde{\theta}-\theta)=-\{[\nabla G(\tilde{\theta})]^T\hat{W}\nabla G(\theta^*)\}^{-1}[\nabla G(\tilde{\theta})]^T\hat{W}\sqrt{n}G(\theta).\tag{A.7}$$

Since $\tilde{\theta}\to_p\theta$, we have $\theta^*\to_p\theta$ which, together with $\hat{W}\to_p\Sigma^{-1}$ and $\nabla G(\tilde{\theta})\to_p\Gamma$, imply that

$$\{[\nabla G(\tilde{\theta})]^T\hat{W}\nabla G(\theta^*)\}^{-1}[\nabla G(\tilde{\theta})]^T\hat{W}\to_p(\Gamma^T\Sigma^{-1}\Gamma)^{-1}\Gamma^T\Sigma^{-1}.\tag{A.8}$$

By the Central Limit Theorem, $\sqrt{n}G(\theta)\to_d N(0,\Sigma)$ which, combined with (A.7) and (A.8), proves part (ii) of the theorem.

References

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

Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika* **95**, 557-571.

Chen, S. X., Leung, D. H. and Qin, J. (2008). Improving semiparametric estimation by using surrogate data. *J. Roy. Statist. Soc. Ser. B* **70**, 803-823.

Chen, K. (2001). Parametric models for response-biased sampling. *J. Roy. Statist. Soc. Ser. B* **63**, 775-789.

Gelfand, A. E. and Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *J. Amer. Statist. Assoc.* **94**, 247-253.

Greenlees, J. S., Reece, W. 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.

Hall, A. R. (2005). *Generalized Method of Moments.* Oxford University Press, New York.

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

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. (2002). *Statistical Analysis with Missing Data.* Second edition. Wiley, New York.

Nevo, A. (2003). Using weights to adjust for sample selection when auxiliary information is available. *J. Business and Economic Statistics* **21**, 43-52.

Newey, W. and Mcfadden, D. (1994). *Large Sample Estimation and Hypothesis Testing.* Springer, New York.

Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under non-ignorable 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. *Statistics in Medicine* **16**, 285-319.

Rotnitzky, A. and Robins, J. M. (1997). Analysis of semi-parametric regression models with non-ignorable nonresponse. *Statist. Medicine* **16**, 81-102.

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

Mathematica Policy Research, Princeton, NJ 08540, U.S.A.

E-mail: kingsun2002@gmail.com

School of Finance and Statistics, East China Normal University, Shanghai 200241, China.

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, U.S.A.

E-mail: shao@stat.wisc.edu

Department of Statistics, Iowa State University, 1208 Snedecor Hall, Iowa State University, Ames, IA 50011, USA.

E-mail: jkim@iastate.edu

(Received March 2012; accepted July 2013)