Statistics and Its Interface Volume 9 (2016) 123–130

Regression analysis with nonignorably missing covariates using surrogate data

Fang Fang*

The paper considers parameter estimation in regression analysis with missing covariates when the missing data mechanism is nonignorable and unspecified, which is quite common in practice but has rarely been discussed in the literature. Assuming that surrogate data for the missed covariates is available for all the subjects, we propose a novel approach that constructs estimating equations based on the conditional expectation of the outcome given the always observed covariates and the surrogate data. Asymptotic properties and variance estimation of the parameter estimators from the new approach are established. Some simulation results are presented to compare the finite sample performance of various estimators. A real data set from the National Health and Nutrition Examination Survey is analyzed to illustrate the application of the method.

AMS 2000 subject classifications: Primary 62J05; secondary 62G20.
Keywords and phrases: Imputation, Nonignorably missing covariates, Pseudo likelihood, Regression analysis, Surrogate data.

1. INTRODUCTION

Missing covariate data is quite common in many application areas such as sample survey, medical studies and social sciences. Conventional statistical methods can not be directly applied to the data with missing covariate values since the observed data may not be representative of the population and is likely to lead to inefficient or inconsistent estimates. When the missing probability only depends on the observed data, the missing data mechanism is called ignorable or missing at random (MAR), which has been discussed in a rich literature (see, for example, Little, 1992; Robins et al., 1994; Lipsitz et al., 1999; Little and Rubin, 2002; Ibrahim et al., 2005; Qin et al., 2009; Kim and Shao, 2013, among others). However, in many applications the missing data mechanism is believed to be nonignorable, i.e., the missing probability depends on the unobserved data itself even after controlling for the observed data. For example, in health sciences research, missing on self-reported socially unacceptable behaviors such as alcohol consumption or drug use is often considered to be nonignorable (Rotnitzky and Robins, 1997). In labor force surveys, subjects with very high or very low income usually are not willing to report their true income. Parameter estimation with nonignorably missing covariate data is quite challenging since the missing data mechanism involves unobserved data and is hard to identify. Applying methods for MAR to nonignorably missing data may result in serious estimation biases and incorrect inference.

In this paper, we consider regression analysis when some covariates are nonignorably missing. When the missing data mechanism is assumed to have a parametric form, methods based on maximum likelihood estimation (Ibrahim et al., 1999b, 2005; Stubbendick and Ibrahim, 2003, 2006), fully Bayesian (Huang et al., 2005) or inverse probability weighted estimating equations (Rotnitzky and Robins, 1997) have been developed. However, these methods are sensitive to the parametric model assumptions on the missing data mechanism. Limited work has been done when the missing covariate data mechanism is totally unspecified other than the nonignorable assumption. The pseudo likelihood method in Tang et al. (2003) and Zhao and Shao (2014) can be applied in generalized linear models with nonignorably missing covariate data. Unfortunately, as we show in Section 2, for the problem we consider in this paper, this method can not be used to estimate the regression parameters since the parameters are not identifiable from the pseudo likelihood.

In many applications, surrogate data is available for the covariates having missing values. For example, in the National Health and Nutrition Examination Survey of the Unites States, body fat percentage is measured by dual-energy x-ray absorptiometry, which is accurate but expensive and hence only available for part of the sampled subjects. Body mass index is considered as a surrogate measurement of body fat and is available for all the subjects. In the UK Labor Force Survey, hourly payment is directly reported by part of the sampled subjects, while approximated hourly payment can be calculated indirectly from some other variables available for all the subjects.

When surrogate data is available, we propose a novel estimating equation approach with imputation to estimate regression parameters without any parametric model assumption on the missing data mechanism. The surrogate variables are not of direct statistical interest and can not directly replace the missing covariates in the regression model. But

*Partially supported by the Shanghai Scientific Grant 15ZR1410300 and the 111 Project (B14019).

they can be used to identify the conditional distribution of the missing covariates given the observed covariates and the surrogate variables, which is then used for imputation.

The rest of the paper is organized as follows. After introducing details for the model and assumption in Section 2, we propose the new approach in Section 3. Some asymptotic theorems are derived in Section 4. Simulation results and an illustrative real data example are presented in Section 5 and Section 6. Some concluding remarks are given in Section 7. The proofs are sketched in the Appendix.

2. MODEL AND ASSUMPTION

Let Y denote a fully observed outcome variable and the covariate vector $X = (U', Z')'$, where t' denotes the transpose of a column vector t, U is a p-dimensional covariate vector with missing data and Z is a q-dimensional fully observed covariate vector. Our main interest is to estimate the regression parameter $\beta = (\beta_c, \beta_u', \beta_z')'$ defined by $E(Y|X) = \mu(\beta_c + \beta_u' U + \beta_z' Z)$, where $\mu(\cdot)$ is a known monotone and continuously differentiable link function. When there is no missing data, β can be estimated by solving

$$(1) \qquad \frac{1}{n}\sum_{i=1}^{n} D(U_i, Z_i, \beta)\{Y_i - \mu(\beta_c + \beta_u' U_i + \beta_z' Z_i)\} = 0,$$

where $\{Y_i, U_i, Z_i, i = 1, \cdots, n\}$ are independent and identically distributed realization from (Y, U, Z) and $D(U_i, Z_i, \beta)$ is a user-specified function with dimension $p + q + 1$. For example, in a generalized linear model, $D(U_i, Z_i, \beta) = \{\partial\mu(\beta_c + \beta_u' U_i + \beta_z' Z_i)/\partial\beta\}/\mathrm{Var}(Y_i|U_i, Z_i)$.

When there is missing data in covariate U, we denote R as the response indicator of U, i.e., $R = 1$ if U is fully observed and $R = 0$ otherwise. We assume that, other than the outcome and the covariates, a p-dimensional surrogate vector S of U is available for all the subjects. The observed data are $\{R_i, Y_i, R_i U_i, Z_i, S_i, i = 1, \cdots, n\}$. Let $[\cdot]$ or $[\cdot|\cdot]$ be a generic notation for unconditional or conditional probability density function. We assume

$$(2) \qquad [Y|U, Z, S] = [Y|U, Z],$$

i.e., Y and S are conditionally independent given (U, Z). The conditional independence assumption is commonly made when U is a gold standard for some characteristics of the subject and S is a surrogate variable measured with error (Reilly and Pepe, 1995; Bashir and Duffy, 1997; Horton and Laird, 2001). Similarly, we assume

$$(3) \qquad [R|Y, U, Z, S] = [R|Y, U, Z],$$

i.e., R and S are conditionally independent given (Y, U, Z).

Under the assumptions (2) and (3), we have

$$(4) \qquad \begin{aligned}[S|Y, U, Z, R = 1] &= [S|Y, U, Z] \\ &= \frac{[Y|U, Z, S][U|Z, S][Z|S][S]}{\int [Y|U, Z, s][U|Z, s][Z|s][s]ds} \\ &= \frac{[U|Z, S][Z|S][S]}{\int [U|Z, s][Z|s][s]ds}.\end{aligned}$$

This actually is the key idea of the pseudo likelihood method in Zhao and Shao (2014). However, $[Y|U, Z]$ is not identifiable from the pseudo likelihood since (4) does not involve in Y. So we can not directly use the pseudo likelihood method to estimate the regression parameter β.

Further we assume that $[U|Z, S]$ has a parametric form $p(U|Z, S, \alpha)$ with an unknown parameter vector α. The conditional distribution $[Z|S]$ could have either a parametric form $p(Z|S, \gamma)$ with an unknown parameter vector γ or a nonparametric form $p(Z|S)$.

3. THE PROPOSED METHOD

Since U has some missing values and the nonignorable missing mechanism is difficult to handle, our proposed method focuses the fully observed data (Y, Z, S) and considers the conditional expectation of Y given (Z, S). Based on the fact that

$$\begin{aligned} E(Y|Z, S) &= E[E(Y|U, Z, S)|Z, S] \\ &= E\left[\mu(\beta_c + \beta_u' U + \beta_z' Z)|Z, S\right], \end{aligned}$$

we can get a consistent estimator of β by solving the following estimating equation

$$(5) \qquad \begin{aligned} \psi(\beta, \hat\alpha) &= \frac{1}{n}\sum_{i=1}^{n}\psi_i(\beta, \hat\alpha) \\ &= \frac{1}{n}\sum_{i=1}^{n} D(U_i^{imp}, Z_i, \beta)\left\{Y_i - \mu_i^{imp}\right\} = 0, \end{aligned}$$

where $\hat\alpha$ is a consistent estimator of α that will be given later, function $D(\cdot)$ is defined in (1), and for $i = 1, \cdots, n$, $U_i^{imp} = E(U_i|Z_i, S_i, \hat\alpha)$ and $\mu_i^{imp} = E(\mu(\beta_c + \beta_u' U_i + \beta_z' Z_i)|Z_i, S_i, \hat\alpha)$ are imputed values of U_i and $\mu_i = \mu(\beta_c + \beta_u' U_i + \beta_z' Z_i)$, respectively. When $\mu(\cdot)$ is a nonlinear function, μ_i^{imp} usually does not have an explicit form. In practice, μ_i^{imp} can be replaced by its Monte Carlo approximation $\frac{1}{L}\sum_{l=1}^{L}\mu(\beta_c + \beta_u' U_i^l + \beta_z' Z_i)$, where $\{U_i^l, l = 1, \cdots, L\}$ is a random sample generated from $p(U_i|Z_i, S_i, \hat\alpha)$. Similar techniques have been widely used in the Monte Carlo EM algorithms (for example, Wei and Tanner, 1990; Ibrahim et al., 1999a; Lipsitz et al., 1999).

Now we consider how to obtain a consistent estimator of α. Based on (4), we can try to estimate α using the pseudo likelihood method. First we need to check whether α is identifiable from (4) or not. By the discussions in Zhao and Shao (2014), a necessary and almost sufficient condition for the identifiability of α is that $p(U|Z, S, \alpha)$ should depend on S. Since S is a surrogate of U, this condition is almost always satisfied. So in what follows we assume that α is identifiable from (4).

When $[Z|S]$ has a parametric form $p(Z|S,\gamma)$, γ can be estimated by an estimator $\hat{\gamma}$ based on the fully observed (Z,S) data. If we replace $[Z|S]$ in (4) by $p(Z|S,\hat{\gamma})$ and replace $[S]$ by the empirical distribution of S-data, we can get the pseudo likelihood of α as

$$L(\alpha)=\prod_{i=1}^{n} R_i\frac{p(U_i|Z_i,S_i,\alpha)p(Z_i|S_i,\hat{\gamma})}{\sum_{j=1}^{n} p(U_i|Z_i,S_j,\alpha)p(Z_i|S_j,\hat{\gamma})}. \tag{6}$$

When $[Z|S]$ has a nonparametric form $p(Z|S)$, it can be estimated using the standard nonparametric product kernel estimator (Li and Racine, 2007)

$$\hat{p}(Z|S)=\frac{\frac{1}{nh^{q+p}}\sum_{k=1}^{n}\prod_{d=1}^{q}K\left(\frac{Z_d-Z_{dk}}{h}\right)\prod_{d=1}^{p}K\left(\frac{S_d-S_{dk}}{h}\right)}{\frac{1}{nh^{p}}\sum_{k=1}^{n}\prod_{d=1}^{p}K\left(\frac{S_d-S_{dk}}{h}\right)},$$

where $Z=(Z_1,\cdots,Z_q)'$, $S=(S_1,\cdots,S_p)'$, $K(\cdot)$ is a one-dimensional kernel function, and h is a bandwidth. The pseudo likelihood of α in this case is given by

$$L(\alpha)=\prod_{i=1}^{n} R_i\frac{p(U_i|Z_i,S_i,\alpha)\hat{p}(Z_i|S_i)}{\sum_{j=1}^{n} p(U_i|Z_i,S_j,\alpha)\hat{p}(Z_i|S_j)}. \tag{7}$$

Then $\hat{\alpha}$ is obtained by maximizing the pseudo likelihood (6) or (7). Our proposed estimator $\hat{\beta}$ is the solution to the estimating equation in (5).

The proposed method needs to specify a parametric model for $[U|Z,S]$. When U has nonignorably missing data, this is not trivial in general. However, when S is a surrogate for U, we may have some background information for a rough relationship between S and U, which makes the parametric modeling on $[U|Z,S]$ much easier. For example, in some cases, S is just a measurement of U with a completely random error, we may just assume that $[U|Z,S]\sim N(S,\alpha^2)$. Note that even in this simple situation, the method that imputes the missing U by S followed by a regular estimation procedure may have serious bias.

When $[Z|S]$ has a parametric form, we also need to specify a parametric model for $[Z|S]$. There are several strategies to do this. For example, Lipsitz and Ibrahim (1996) and Ibrahim et al. (1999b) write $p(Z|S,\gamma)$ as

$$\begin{aligned} & p(Z_1,\cdots,Z_q|S,\gamma) \\ = \; & p(Z_q|Z_1,\cdots,Z_{q-1},S,\gamma_1) \\ & \times p(Z_{q-1}|Z_1,\cdots,Z_{q-2},S,\gamma_2)\cdots p(Z_1|S,\gamma_q), \end{aligned} \tag{8}$$

where γ_j is a vector of parameters for the jth conditional distribution, the γ_j's are distinct, and $\gamma=(\gamma_1',\cdots,\gamma_q')'$. Sometimes it may be easier to model $[S|Z]$, then we can write $p(Z|S,\gamma)=p(S|Z,\gamma)p(Z,\gamma)/p(S,\gamma)$, where $p(S|Z,\gamma)$ can be modeled in a similar way to (8). When we choose appropriate $p(S|Z,\gamma)$ and $p(Z,\gamma)$, $p(S,\gamma)=\int p(S|z,\gamma)p(z,\gamma)dz$ could have an explicit form. For example, the real data analysis in Section 6 just uses this modeling strategy.

4. ASYMPTOTIC THEORY

In this section, we establish some asymptotic properties of the proposed estimator $\hat{\beta}$ as $n\to\infty$. The asymptotic properties of $\hat{\beta}$ is related to the asymptotic properties of $\hat{\alpha}$, which has been studied in Zhao and Shao (2014). When $\hat{\alpha}$ is obtained by maximizing (6), denote

$$\begin{aligned} H_i(\alpha,\gamma,F) \;\; = \;\; & R_i\Big\{\log p(U_i|Z_i,S_i,\alpha) \\ & - \log\int p(U_i,Z_i,S,\alpha)p(Z_i|S,\gamma)dF(s)\Big\} \end{aligned}$$

and $H(\alpha,\gamma,F)$ as $H_i(\alpha,\gamma,F)$ with (R_i,U_i,Z_i,S_i) replaced by (R,U,Z,S). Then maximizing (6) is the same as maximizing

$$l(\alpha,\hat{\gamma},\hat{F})=\frac{1}{n}\sum_{i=1}^{n}H_i(\alpha,\hat{\gamma},\hat{F}), \tag{9}$$

where $\hat{\gamma}$ is an estimator of the nuisance parameter γ and $\hat{F}$ is the empirical distribution of S-data. The following theorem shows the consistency and asymptotic normality of $\hat{\beta}$ under some regularity conditions.

Theorem 4.1. *Assume that α is identifiable from (4) and*

(i) $\hat{\gamma}$ is consistent and

$$\sqrt{n}(\hat{\gamma}-\gamma_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}T(Z_i,S_i,\gamma_0)+o_p(1), \tag{10}$$

where γ_0 is the true value of γ, $o_p(1)$ denotes a quantity converging to 0 in probability as $n\to\infty$, and $T(Z,S,\gamma_0)$ is the influence function.

(ii) As $n\to\infty$, $E\{H(\alpha,\hat{\gamma},\hat{F})-H(\alpha,\gamma_0,F_0)\}\to 0$, where F_0 is the true distribution of S, and there exist $\epsilon_1>0$ and $\epsilon_2>0$ such that

$$\sup_{\substack{\|\gamma-\gamma_0\|<\epsilon_1\\ \|F-F_0\|<\epsilon_2}}\left|\frac{1}{n}\sum_{i=1}^{n}H_i(\alpha,\gamma,F)-E\{H(\alpha,\gamma,F)\}\right|\to 0.$$

(iii) $H(\alpha,\gamma,F)$ is continuously twice differentiable with respect to α, $E[\nabla^2_{\alpha\alpha}H(\alpha_0,\gamma_0,F_0)]$ is positive definite, where α_0 is the true value of α, and $\|\nabla^2_{\alpha\alpha}H(\alpha,\gamma_0,F_0)\|$ and $\|\nabla^2_{\alpha\gamma}H(\alpha,\gamma_0,F_0)\|$ are bounded by integrable functions in a neighborhood of α_0.

(iv) The estimating equation $\psi(\beta,\alpha)$ in (5) is continuously differentiable with respect to $\theta=(\beta,\alpha)$, $E[\nabla_\beta\psi(\beta_0,\alpha_0)]$ is positive definite, and $\|\nabla_\theta\psi(\beta,\alpha)\|$ is bounded by an integrable function in a neighborhood of $\theta_0=(\beta_0,\alpha_0)$, where β_0 is the true value of β.

Then as $n\to\infty$,

$$\sqrt{n}(\hat{\beta}-\beta_0)\to_d N(0,\Sigma)$$

for a covariance matrix Σ and $\to_d$ denotes convergence in distribution.

For the variance estimation of $\hat{\beta}$, we derive (in the Appendix) that

$$\sqrt{n}(\hat{\beta}-\beta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E_i+o_p(1), \tag{11}$$

where $E_i = f(W_i,\beta_0,\alpha_0,\gamma_0,F_0,A_1,A_2,B_1,B_2)$, $W_i = (R_i,Y_i,U_i,Z_i,S_i)$, f is defined in (15) in the Appendix, $A_1 = E[\nabla^2_{\alpha\alpha}H(\alpha_0,\gamma_0,F_0)]$, $A_2 = E[\nabla^2_{\alpha\gamma}H(\alpha_0,\gamma_0,F_0)]$, $B_1 = E[\nabla_\beta\psi(\beta_0,\alpha_0)]$ and $B_2 = E[\nabla_\alpha\psi(\beta_0,\alpha_0)]$. So $\Sigma = Var(E_i)$ can be estimated by $\hat{\Sigma}$, which is the sample covariance matrix based on $\hat{E}_i = f(W_i,\hat{\beta},\hat{\alpha},\hat{\gamma},\hat{F},\hat{A}_1,\hat{A}_2,\hat{B}_1,\hat{B}_2)$,$i = 1,\cdots,n$, where

$$\hat{A}_1=\frac{1}{n}\sum_{i=1}^{n}\nabla^2_{\alpha\alpha}H_i(\hat{\alpha},\hat{\gamma},\hat{F}),\hat{A}_2=\frac{1}{n}\sum_{i=1}^{n}\nabla^2_{\alpha\gamma}H_i(\hat{\alpha},\hat{\gamma},\hat{F}),$$

$$\hat{B}_1=\frac{1}{n}\sum_{i=1}^{n}\nabla_\beta\psi_i(\hat{\beta},\hat{\alpha}),\ \text{and}\ \hat{B}_2=\frac{1}{n}\sum_{i=1}^{n}\nabla_\alpha\psi_i(\hat{\beta},\hat{\alpha}).$$

When $\hat{\alpha}$ is obtained by maximizing (7), the regularity conditions for the consistency and asymptotic normality of $\hat{\alpha}$ are quite complicated and hence are omitted here. One may refer to Theorem 2 and Theorem 3 of Zhao and Shao (2014) for the details. Once $\hat{\alpha}$ is $\sqrt{n}$-consistent, the consistency and asymptotic normality of $\hat{\beta}$ can be shown exactly the same as in Theorem 4.1. For the variance estimation, a similar result in (11) is hard to derive, so we suggest using the bootstrap in this case.

5. SIMULATION STUDIES

We performed some simulation studies to examine the finite sample behaviors of the proposed estimator and compare the performance of several methods. The simulation was carried out with 1,000 replications and sample size $n = 500$. The surrogate variable S and covariates (U,Z) were generated with $S \sim N(0,1)$, $Z|S \sim N(1+2S,1)$, and $U|Z,S \sim N(1-Z+3S,1)$. Note that in this setting $U|S \sim N(S,2)$. For the outcome Y and the response indicator R, we considered the following six cases.

(A1) $Y|U,Z \sim N(1+U+Z,1)$, $P(R=1|Y,U,Z) = \Phi(-1+Y^2+2U+|Y|Z)$.
(A2) The same as (A1) except that $P(R=1|Y,U,Z) = \Phi(-1+Y^2+0.5U+|Y|Z)$.
(B1) Y is binary with $P(Y=1|U,Z) = \text{expit}(1+U+Z)$, $P(R=1|Y,U,Z) = \Phi(-2+Y+|U|+Z)$.
(B2) The same as (B1) except that $P(R=1|Y,U,Z) = \Phi(-1+|U|+Z)$.
(C1) Y is Poisson with $E(Y|U,Z) = \exp(-1+0.5U-0.5Z)$, $P(R=1|Y,U,Z) = \Phi(-2+Y+|U|+Z)$.
(C2) The same as (C1) except that $P(R=1|Y,U,Z) = \Phi(-1+|U|+Z)$.

Note that the association between R and U in case (A2) was weaker than case (A1). The missing mechanism in cases (B2) and (C2) did not depend on Y and hence the complete case analysis was valid in these two cases. In the six cases, the response rates $P(R=1)$ were about 69%, 74%, 65%, 69%, 62%, and 69%, respectively.

For all the cases, we compared the following 6 methods: full data analysis assuming no missing data (FULL), complete case analysis (CC), surrogate method (SURRO) that simply imputes the missed U by S, simple imputation method (SIMP) that simply imputes the missed U by $E(U|Z,S,\hat{\alpha})$, and the proposed method with a parametric $[Z|S]$ and a correctly specified $[U|Z,S]$ model (PRO) or a misspecified $[U|Z,S]$ model as $[U|Z,S] \sim N(\alpha_1+\alpha_2\sin(Z/4)+\alpha_3S,\alpha_4)$ (PRO_MU). In addition, we compared several other methods in the following cases.

1. In cases (A1) and (A2), we also considered the maximum likelihood methods with full parametric model. First, we assume all the parametric models including the response model $[R|Y,U,Z]$ are correctly specified (MLE). Second, we assume the $[U|Z,S]$ model is misspecified (MLE_MU) as in PRO_MU. Third, we assume the response model is misspecified (MLE_MR) as $P(R=1|Y,U,Z) = \Phi(\eta_1+\eta_2Y+\eta_3U+\eta_4Z)$. Fourth, we assume the $[U|Z,S]$ model and the response model both are misspecified (MLE_M). Fifth, we assume the missing data mechanism is considered to be missing at random (MAR). The observed likelihood involves an integral and generally is not easy to handle. But in cases (A1) and (A2), the integral has an explicit form and hence the maximization is not difficult.
2. In cases (B1) and (B2), we considered the proposed method assuming $[Z|S]$ has a nonparametric form (PRO_N). We used the Epanechnikov kernel function of order 4: $K(t) = \frac{45}{32}(1-\frac{7}{3}t^2)(1-t^2)I\{|t| \le 1\}$ and the bandwidth $h \propto n^{-\frac{1}{6}}$. The kernel function and the bandwidth were selected to satisfy the regularity condition (H) in Theorem 2 of Zhao and Shao (2014). In our notation, the condition is that $h \to 0$, $nh^{p+q} \to \infty$, $n^{\frac{1}{2}}h^{p+q+2d}/\log(n) \to \infty$ and $nh^{2m} \to 0$, as $n \to \infty$, where m is the order of the kernel function $K(\cdot)$ which has bounded derivatives of order d.

In the methods of PRO, PRO_MU and PRO_N, the Monte Carlo sample size $L = 10{,}000$.

Tables 1–3 report the relative bias, standard deviation, standard error, which is the estimate of standard deviation, and the coverage probability of approximate 95% confidence interval for β based on normal approximation. The standard error based on PRO was obtained by $\hat{\Sigma}$ defined in Section 4. The standard error based on PRO_N was obtained by bootstrapping with bootstrap round 200. The simulation results can be summarized as follows.

1. The proposed methods (PRO and PRO_N) produce nearly unbiased estimators for the regression parameters. The proposed variance estimators also perform well and so is the coverage probability of approximate

Table 1. The Simulation Results when Y is Normal

	Case (A1)			Case (A2)		
Method	$\hat{\beta}_c$	$\hat{\beta}_u$	$\hat{\beta}_z$	$\hat{\beta}_c$	$\hat{\beta}_u$	$\hat{\beta}_z$
			Relative Bias in %			
FULL	0.0	0.1	0.0	-0.3	0.0	0.1
CC	-9.7	2.8	4.5	-6.7	2.7	3.2
SURRO	-7.6	4.1	-6.4	10.8	2.7	-9.5
SIMP	-9.9	0.2	0.1	-2.5	-0.4	-1.1
PRO	0.2	-0.1	-0.1	-0.4	-0.1	0.1
PRO_MU	-0.3	6.7	-1.8	-1.7	5.6	-1.4
MLE	0.1	0.1	-0.1	-0.3	0.0	0.1
MLE_MU	-1.1	0.3	0.5	-2.5	-0.1	1.0
MLE_MR	-2.6	-1.1	1.2	-0.9	-0.3	0.3
MLE_M	-6.2	-0.7	2.4	-5.6	-0.1	1.8
MAR	-13.9	1.4	3.3	-4.8	0.2	1.2
			Standard Deviation			
FULL	0.051	0.026	0.021	0.051	0.026	0.021
CC	0.073	0.032	0.026	0.069	0.029	0.027
SURRO	0.064	0.032	0.027	0.062	0.030	0.029
SIMP	0.070	0.033	0.025	0.066	0.030	0.026
PRO	0.076	0.049	0.028	0.073	0.044	0.028
PRO_MU	0.080	0.063	0.032	0.079	0.058	0.033
MLE	0.059	0.029	0.022	0.059	0.028	0.024
MLE_MU	0.059	0.029	0.023	0.060	0.027	0.024
MLE_MR	0.082	0.034	0.026	0.069	0.028	0.025
MLE_M	0.084	0.034	0.025	0.072	0.028	0.026
MAR	0.061	0.030	0.023	0.059	0.028	0.024
			Standard Error			
FULL	0.049	0.027	0.021	0.049	0.027	0.021
CC	0.070	0.034	0.026	0.067	0.031	0.026
SURRO	0.061	0.038	0.027	0.059	0.035	0.026
SIMP	0.055	0.032	0.023	0.054	0.030	0.023
PRO	0.075	0.047	0.028	0.072	0.044	0.029
PRO_MU	0.081	0.061	0.032	0.078	0.056	0.033
MLE	0.058	0.030	0.023	0.058	0.029	0.023
MLE_MU	0.058	0.029	0.023	0.058	0.028	0.023
MLE_MR	0.079	0.034	0.026	0.069	0.029	0.025
MLE_M	0.090	0.035	0.028	0.072	0.029	0.026
MAR	0.060	0.030	0.023	0.057	0.029	0.023
			Coverage Probability in %			
FULL	93.8	94.4	94.7	95.3	95.1	93.8
CC	71.3	88.1	59.2	81.0	87.9	74.1
SURRO	74.7	85.1	32.2	52.3	91.3	5.6
SIMP	53.0	93.4	93.4	86.5	94.8	89.5
PRO	94.1	93.4	95.6	95.8	94.9	96.0
PRO_MU	94.2	83.0	92.3	95.4	86.0	93.4
MLE	94.7	95.3	95.3	93.6	95.0	93.3
MLE_MU	93.8	94.8	94.6	91.1	94.9	90.4
MLE_MR	93.6	93.8	92.2	95.4	94.8	93.5
MLE_M	90.2	95.7	86.2	87.8	94.5	88.6
MAR	36.1	91.8	70.8	84.8	95.3	89.1

95% confidence interval. When the model $[U|Z, S]$ is misspecified, the proposed method (PRO_MU) will have bias just like most other parametric model based methods, although the biases are not quite serious (less than 10%) in cases (A1), (A2) and (C2).

Table 2. The Simulation Results when Y is Binary

	Case (B1)			Case (B2)		
Method	$\hat{\beta}_c$	$\hat{\beta}_u$	$\hat{\beta}_z$	$\hat{\beta}_c$	$\hat{\beta}_u$	$\hat{\beta}_z$
			Relative Bias in %			
FULL	2.1	1.8	1.9	1.2	1.4	1.4
CC	94.2	17.7	-21.1	0.6	2.7	3.7
SURRO	25.9	-1.8	-22.4	33.0	1.4	-22.8
SIMP	-3.6	-11.8	-5.4	-0.6	-9.9	-5.7
PRO	3.8	4.1	3.1	2.3	2.7	2.3
PRO_MU	-5.4	27.0	26.8	-0.7	25.4	23.4
PRO_N	3.1	3.5	4.3	2.2	1.6	4.1
			Standard Deviation			
FULL	0.178	0.122	0.116	0.160	0.123	0.115
CC	0.446	0.209	0.199	0.301	0.161	0.193
SURRO	0.193	0.127	0.096	0.197	0.138	0.097
SIMP	0.180	0.112	0.109	0.167	0.110	0.107
PRO	0.231	0.204	0.140	0.202	0.188	0.133
PRO_MU	0.278	0.316	0.215	0.241	0.304	0.199
PRO_N	0.235	0.210	0.144	0.204	0.226	0.147
			Standard Error			
FULL	0.166	0.124	0.111	0.165	0.123	0.110
CC	0.430	0.195	0.201	0.299	0.154	0.183
SURRO	0.188	0.137	0.093	0.194	0.137	0.094
SIMP	0.159	0.115	0.102	0.161	0.115	0.102
PRO	0.221	0.202	0.136	0.209	0.197	0.134
PRO_MU	0.263	0.314	0.209	0.252	0.309	0.203
PRO_N	0.231	0.224	0.141	0.198	0.235	0.146
			Coverage Probability in %			
FULL	94.5	95.4	94.9	95.3	95.0	93.9
CC	38.1	91.3	75.7	95.8	94.9	94.6
SURRO	76.2	95.3	32.4	62.6	95.9	33.1
SIMP	90.7	80.6	88.7	93.9	83.3	88.1
PRO	95.8	97.2	96.4	95.7	96.3	95.4
PRO_MU	91.9	99.4	90.4	94.6	99.8	94.2
PRO_N	95.5	97.3	95.9	95.5	96.4	94.2

2. When both of the $[U|Z, S]$ model and the response model are correctly specified, full parametric likelihood method MLE works well and the estimators have smaller standard deviations than PRO. When either parametric model is misspecified (MLE_MU, MLE_MR), although generally the maximum likelihood estimators are biased, they just have some minor biases in the specific cases (A1) and (A2). However, the coverage probabilities may not work as well as MLE which could lead to inaccurate inference. When both models are misspecified, MLE_M works even worse especially for the coverage probabilities.
3. When we apply the methods based on missing at random to the case of nonignorable missingness, the estimators could have large biases and the coverage probabilities perform poorly. When the missing mechanism is close to missing at random in case (A2), the estimation bias of MAR method reduces but the biases in the coverage probabilities are still nonnegligible.

Table 3. The Simulation Results when Y is Poisson

	Case (C1)			Case (C2)		
Method	$\hat{\beta}_c$	$\hat{\beta}_u$	$\hat{\beta}_z$	$\hat{\beta}_c$	$\hat{\beta}_u$	$\hat{\beta}_z$
Relative Bias in %						
FULL	-0.9	0.5	0.1	-0.7	-0.1	0.0
CC	42.3	-21.8	-7.1	-1.6	0.3	-0.2
SURRO	13.1	-2.8	-10.4	20.8	-6.0	-23.2
SIMP	0.8	-2.7	0.7	4.7	-5.6	-2.2
PRO	-1.2	0.4	0.1	-1.0	-0.3	0.3
PRO_MU	-9.1	11.2	2.2	-8.6	7.4	2.7
Standard Deviation						
FULL	0.084	0.040	0.032	0.085	0.040	0.032
CC	0.108	0.048	0.040	0.147	0.056	0.065
SURRO	0.076	0.038	0.035	0.081	0.045	0.054
SIMP	0.090	0.043	0.035	0.092	0.046	0.044
PRO	0.101	0.062	0.044	0.104	0.061	0.045
PRO_MU	0.115	0.083	0.049	0.118	0.078	0.051
Standard Error						
FULL	0.086	0.040	0.031	0.086	0.040	0.031
CC	0.115	0.048	0.037	0.147	0.056	0.064
SURRO	0.079	0.036	0.033	0.077	0.041	0.039
SIMP	0.085	0.040	0.031	0.084	0.042	0.032
PRO	0.102	0.059	0.040	0.104	0.058	0.041
PRO_MU	0.116	0.075	0.045	0.116	0.072	0.046
Coverage Probability in %						
FULL	96.1	95.8	95.5	95.8	95.1	95.0
CC	6.0	37.5	85.7	95.6	95.1	94.8
SURRO	61.4	91.9	64.1	25.6	85.0	22.7
SIMP	93.4	92.5	92.2	87.2	86.7	84.5
PRO	95.0	93.3	94.1	95.4	93.9	94.4
PRO_MU	91.5	88.7	91.2	91.7	92.0	90.4

4. When the missing mechanism depends on Y in cases (A1), (A2), (B1) and (C1), CC method is biased as expected. When the missing mechanism does not depend on Y in cases (B2) and (C2), CC method is nearly unbiased. The standard deviations of $\hat{\beta}_c$ and $\hat{\beta}_z$ for PRO are smaller than CC, while the standard deviation of $\hat{\beta}_u$ for PRO is larger than CC. This is also expected since our proposed method discards the observed U in the estimating equation (5).
5. The estimators from the trivial methods of SURRO and SIMP have large biases and the coverage probabilities usually are much less than 95%.

6. A REAL DATA ANALYSIS

To illustrate the application of our proposed method, we analyze a data set from the National Health and Nutrition Examination Survey (NHANES 2005, the United States Centers for Disease Control and Prevention). The data is available at http://www.cdc.gov/nchs/nhanes.htm. NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States. In our analysis, we investigate how adults' hypertension is related to age, gender and body fat. Dual-energy x-ray absorptiometry (DXA) has been accepted as the gold standard direct measurement of body fat. However, some of the DXA data are missing and the missing pattern seems to be systematic and non-random. Use of only the measured variables could lead to biased results. Fortunately, body mass index (BMI) is available for almost all the subjects and can be considered as a surrogate variable of DXA although it is less accurate than DXA.

In our analysis, the binary outcome $Y = 1$ if the subject has hypertension, i.e., the systolic blood pressure (average of BPXSY1, BPXSY2, BPXSY3, and BPXSY4) is greater than 140 or the diastolic blood pressure (average of BPXDI1, BPXDI2, BPXDI3, and BPXDI4) is greater than 90, and $Y = 0$ otherwise. The covariate U is the body fat percentage measured by DXA (variable name DXDTOPF), which has a surrogate variable S (BMI, defined as $log(BMXBMI)$). The other two covariates contained in the regression model are Z_1 (gender, 1 for male and 0 for female, variable name RIAGENDR) and Z_2 (age, defined as $log(RIDAGEYR)$). Logarithmic transformations are made to BMXBMI and RIDAGEYR to adjust the skewness. There are $n = 3{,}707$ subjects and 1,111 (29.9%) of them have missing U. We assume that $P(Y = 1) = \text{expit}(\beta_c + \beta_u U + \beta_{z1} Z_1 + \beta_{z2} Z_2)$ and the main interest is to estimate the regression parameter $\beta = (\beta_c, \beta_u, \beta_{z1}, \beta_{z2})'$.

To apply the proposed method, we need to specify a parametric model on $[U|Z_1, Z_2, S]$. We assume that $U|Z_1, Z_2, S \sim N(\alpha_1 + \alpha_2 S, \alpha_3)$. We also need to specify a parametric model on $[Z_1, Z_2|S]$. Since it is much reasonable to model on $[S|Z_1, Z_2]$ in this example, we consider the following parametric modeling methods. We assume that $Z_1 \sim Bin(1, \rho)$, $Z_2 \sim N(\mu, \sigma^2)$ and the marginal distributions of Z_1 and Z_2 are independent. The conditional distribution of S given (Z_1, Z_2) is assumed as $S|Z_1 = 1, Z_2 \sim N(\gamma_1 + \gamma_2 Z_2, \gamma_3)$ and $S|Z_1 = 0, Z_2 \sim N(\gamma_4 + \gamma_5 Z_2, \gamma_6)$. The parameter $\gamma = (\rho, \mu, \sigma^2, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6)'$. Under these assumptions, the marginal distribution of S is a mixed normal and the joint distribution of $[Z_1, Z_2, S]$ is also easy to calculate. Thus the conditional distribution $[Z_1, Z_2|S] = [Z_1, Z_2, S]/[S]$ is easy to obtain.

Other than the proposed method, we also consider the other two methods: the complete case analysis (CC) and the maximum likelihood method assuming missing at random (MAR, $[R = 1|Y, U, Z_1, Z_2, S] = [R|Y, Z_1, Z_2, S]$).

The analysis results of the three methods are reported in Table 4. All the three methods agree that intercept, gender and age have significant effects on hypertension. The complete case analysis fails in detecting the significant effect of body fat (p-value is 0.0766) while the proposed method and MAR method both suggest that body fat has a significantly positive effect on hypertension. The proposed method and MAR method have similar results in the intercept, body fat and age. But the proposed method suggests smaller gender effect (estimated gender coefficient is 0.394) than the

Table 4. The analysis results for NHANES data

	Estimate	Standard Error	Z-Value	P-Value
		Complete Case Analysis		
$\hat{\beta}_c$	-14.499	0.9582	-15.131	0.0000
$\hat{\beta}_u$	0.019	0.0108	1.771	0.0766
$\hat{\beta}_{z1}$	0.492	0.1793	2.742	0.0061
$\hat{\beta}_{z2}$	3.080	0.2400	12.834	0.0000
		The Proposed Method		
$\hat{\beta}_c$	-14.657	0.7795	-18.803	0.0000
$\hat{\beta}_u$	0.037	0.0062	5.900	0.0000
$\hat{\beta}_{z1}$	0.394	0.1061	3.715	0.0002
$\hat{\beta}_{z2}$	2.959	0.1904	15.540	0.0000
		Maximum Likelihood Estimation Assuming MAR		
$\hat{\beta}_c$	-14.271	0.6698	-21.307	0.0000
$\hat{\beta}_u$	0.038	0.0081	4.673	0.0000
$\hat{\beta}_{z1}$	0.705	0.1196	5.897	0.0000
$\hat{\beta}_{z2}$	2.836	0.1694	16.745	0.0000

MAR method (estimated gender coefficient is 0.705) although both of them are significant.

7. CONCLUDING REMARKS

Handling nonignorable missing covariates is a difficult problem without a parametric model on the missing data mechanism. When surrogate data is available, we propose a novel approach that constructs unbiased estimating equations based on the conditional expectation of the outcome given the always observed covariates and the surrogate data. The proposed estimator works well both theoretically and empirically. To gain the flexibility of not specifying any parametric model for the missing data process, we have to pay some prices such as discarding the observed U in (5) and extra parametric model assumptions on $[U|Z,S]$ and $[Z|S]$. However, such prices are either worthy or not quite expensive as one may think.

First, the fact that we have to discard the observed U in (5) unravels one major difficulty in handling nonignorable missing data, that is, the missing data process and the data generating process usually are hard to be handled separately in the estimation procedure unless we are willing to give up some information (in our case, part of the observed covariates). Meanwhile, it should be noted that, the information carried in the observed U is actually retrieved partially when we estimate α by maximizing (6) or (7).

Second, although the proposed method relies on a parametric assumption on the distribution of $[U|Z,S]$ (just like most other parametric model based methods) that may not be testable in the presence of nonignorable missing data, it is still useful, for example, in exploring data or in a sensitivity analysis. Meanwhile, since S is a surrogate for U, the parametric modeling for $[U|Z,S]$ in our case is much easier than usual as we have discussed in Section 3. On the other hand, the parametric model $[Z|S]$ is testable since Z and S are fully observed. Moreover, $[Z|S]$ can also be estimated nonparametrically.

ACKNOWLEDGEMENTS

The author's research is partially supported by the Shanghai Scientific Grant 15ZR1410300 and the 111 Project (B14019). The author would like to thank the editor, the associate editor and two anonymous referees for their helpful suggestions.

APPENDIX A. PROOF OF THEOREM 4.1

Proof. Following Zhao and Shao (2014), under the regularity conditions (i) and (ii), $\hat{\alpha}$ is a consistent estimator of α_0. Then under the regularity condition (iii) and by Taylor's expansion, we have

$$\begin{aligned} (12) \quad 0 &= \nabla_\alpha l(\hat{\alpha}, \hat{\gamma}, \hat{F}) \\ &= \nabla_\alpha l(\alpha_0, \gamma_0, \hat{F}) + E[\nabla^2_{\alpha\alpha} H(\alpha_0, \gamma_0, F_0)](\hat{\alpha} - \alpha_0) \\ &\quad + E[\nabla^2_{\alpha\gamma} H(\alpha_0, \gamma_0, F_0)](\hat{\gamma} - \gamma_0) + o_p(n^{-\frac{1}{2}}), \end{aligned}$$

where l is defined in (9). By the theory of V-statistics,

$$\begin{aligned} (13) \quad & \nabla_\alpha l(\alpha_0, \gamma_0, \hat{F}) = \nabla_\alpha l(\alpha_0, \gamma_0, F_0) \\ & \qquad + \frac{1}{n} \sum_{i=1}^{n} V(S_i, \alpha_0, \gamma_0, F_0) + o_p(n^{-\frac{1}{2}}), \end{aligned}$$

where

$$\begin{aligned} & V(S_i, \alpha_0, \gamma_0, F_0) \\ = \; & \mathrm{E}\left\{ \frac{R \int \nabla_\alpha p(U|Z,s,\alpha_0) p(Z|s,\gamma_0) dF_0(s)}{[\int p(U|Z,s,\alpha_0) p(Z|s,\gamma_0) dF_0(s)]^2} \right. \\ & \times p(U|Z,S_i,\alpha_0) p(Z|S_i,\gamma_0) \\ & \left. - \left. \frac{R \nabla_\alpha p(U|Z,S_i,\alpha_0) p(Z|S_i,\gamma_0)}{\int p(U|Z,s,\alpha_0) p(Z|s,\gamma_0) dF_0(s)} \right| S_i \right\}. \end{aligned}$$

By (12), (13) and (10), we have

$$\begin{aligned} (14) \quad & \sqrt{n}(\hat{\alpha} - \alpha_0) \\ = \; & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} -A_1^{-1} \left[\nabla_\alpha H_i + V_i + A_2 T_i \right] + o_p(1), \end{aligned}$$

where $H_i = H_i(\alpha_0, \gamma_0, F_0)$, $V_i = V(S_i, \alpha_0, \gamma_0, F_0)$, $T_i = T(Z_i, S_i, \gamma_0)$ defined in (10), $A_1 = E[\nabla^2_{\alpha\alpha} H(\alpha_0, \gamma_0, F_0)]$, and $A_2 = E[\nabla^2_{\alpha\gamma} H(\alpha_0, \gamma_0, F_0)]$. So $\hat{\alpha}$ is $\sqrt{n}$-consistent and $\sqrt{n}(\hat{\alpha} - \alpha_0)$ converges to $N(0, \mathrm{Var}(D_i))$ in distribution, where $D_i = -A_1^{-1} \left[\nabla_\alpha H_i + V_i + A_2 T_i \right]$.

Now we consider $\hat{\beta}$. Since $E[\psi(\beta_0, \alpha_0)] = 0$, where ψ is defined in (5), and $\hat{\alpha}$ is $\sqrt{n}$-consistent, we can show that there exists $\hat{\beta}$ such that $P(\psi(\hat{\beta}, \hat{\alpha}) = 0) \to 1$ and $\hat{\beta}$ is a consistent estimator of β_0 using a standard asymptotic analysis for estimating equations. Then under regularity condition (iv) and by Taylor's expansion, we have

$$\begin{aligned}0 &= \psi(\hat\beta,\hat\alpha)\\ &= \psi(\beta_0,\alpha_0)+B_1(\hat\beta-\beta_0)+B_2(\hat\alpha-\alpha_0)+o_p(n^{-\frac{1}{2}}),\end{aligned}$$

where $B_1 = E[\nabla_\beta \psi(\beta_0,\alpha_0)]$ and $B_2 = E[\nabla_\alpha \psi(\beta_0,\alpha_0)]$. Then by (14) we have

$$\begin{aligned}&\sqrt{n}(\hat\beta-\beta_0)\\ =\ &\frac{1}{\sqrt{n}}\sum_{i=1}^{n} -B_1^{-1}[\psi_i(\beta_0,\alpha_0)+B_2D_i]+o_p(1).\end{aligned}$$

If we denote

$$\begin{aligned}E_i &= f(W_i,\beta_0,\alpha_0,\gamma_0,F_0,A_1,A_2,B_1,B_2)\\ &= -B_1^{-1}[\psi_i(\beta_0,\alpha_0)+B_2D_i],\end{aligned} \tag{15}$$

where $W_i = (R_i, Y_i, U_i, Z_i, S_i)$, then

$$\sqrt{n}(\hat\beta-\beta_0) \to_d N(0,\Sigma) \text{ with } \Sigma = \mathrm{Var}(E_i). \qquad \square$$

Received 22 October 2014

REFERENCES

Bashir, S. A. and Duffy, S. W. (1997). The correction of risk estimates for measurement error. *Annals of Epidemiology* **7** 154–164.

Horton, N. J. and Laird, N. M. (2001). Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. *Biometrics* **57** 34–42. MR1833289

Huang, L., Chen, M.-H. and Ibrahim, J. G. (2005). Bayesian analysis for generalized linear models with nonignorably missing covariates. *Biometrics* **61** 767–780. MR2196165

Ibrahim, J. G., Chen, M.-H. and Lipsitz, S. R. (1999a). Monte carlo EM for missing covariates in parametric regression models. *Biometrics* **55** 591–596.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R. and Herring, A. H. (2005). Missing-Data methods for generalized linear models: A comparative review. *J. Am. Statist. Assoc.* **100** 332–346. MR2166072

Ibrahim, J. G., Lipsitz, S. R. and Chen, M.-H. (1999b). Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *J. R. Statist. Soc. B* **61** 173–190. MR1664045

Kim, J. K. and Shao, J. (2013). *Statistical Methods for Incomplete Data Analysis*. New York: Chapman & Hall/CRC.

Li, Q. and Racine, J. S. (2007). *Nonparametric Ecnometrics: Theory and Practice*. Princeton University Press. MR2283034

Lipsitz, S. R. and Ibrahim, J. G. (1996). A Conditional model for incomplete covariates in parametric regression models. *Biometrika* **83** 916–922.

Lipsitz, S. R., Ibrahim, J. G. & Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *J. Am. Statist. Assoc.* **94** 1147–1160. MR1731479

Little, R. J. A. (1992). Regression with missing X's: A review. *J. Am. Statist. Assoc.* **87** 1227–1237.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*. New York: Wiley, 2nd ed. MR1925014

Qin, J., Zhang, B. and Leung, D. H. (2009). Empirical likelihood in missing data problems. *J. Am. Statist. Assoc.* **104** 1492–1503. MR2750574

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82** 299–314. MR1354230

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimating of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89** 846–866. MR1294730

Rotnitzky, A. and Robins, J. M. (1997). Analysis of semiparmatric regression models with nonignorable nonresponse. *Statist. in Med.* **16** 81–102.

Stubbendick, A. L. and Ibrahim, J. G. (2003). Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics* **59** 1140–1150. MR2025699

Stubbendick, A. L. and Ibrahim, J. G. (2006). Likelihood based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. *Statist. Sinca.* **16** 1143–1167. MR2327484

Tang, G., Little, R. J. and Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika* **90** 747–764. MR2024755

Wei, G. C. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Am. Statist. Assoc.* **85** 699–704.

Zhao, J. W. and Shao, J. (2014). Semiparametric pseudo likelihoods in generalized linear models with nonignorable missing data. *J. Am. Statist. Assoc.* DOI:10.1080/01621459.2014.983234.

Fang Fang
School of Finance and Statistics
East China Normal University
500 Dongchuan Road
Shanghai, 200241
China
E-mail address: ffang@sfs.ecnu.edu.cn