Statistics and Probability Letters 215 (2024) 110246

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A kernel-type regression estimator for NMAR response variables with applications to classification

Majid Mojirsheibani *, Arin Khudaverdyan

Department of Mathematics, California State University, Northridge, CA, USA

ARTICLE INFO

MSC:
primary 62G05
secondary 62G08

Keywords:
Regression
Missing data
Kernel
Convergence
Margin

ABSTRACT

This work deals with the problem of nonparametric estimation of a regression function when the response variable may be missing according to a *not-missing-at-random* (NMAR) setup. To assess the theoretical performance of our estimators, we study their strong convergence properties in L_p norms where we also look into their rates of convergence. We also study applications of our results to the problem of statistical classification in semi-supervised learning.

1. Introduction

The problem of statistical estimation, prediction, and inference with nonignorable missing data (i.e., data missing but not at random) has received considerable attention in recent years. This paper considers the problem of nonparametric regression in the presence of missing response variables, Y, for the *not-missing-at-random* (NMAR) setup where the mechanism that causes the absence of Y is allowed to depend on both the predictor $\mathbf{X}$ and the response variable Y.

To present our results, let $(\mathbf{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$ be a random vector and consider the nonparametric estimation of the regression function $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ using the independent and identically distributed (iid) observations $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, drawn from the distribution of $(\mathbf{X}, Y)$. When the data is fully observable, the classical Nadaraya–Watson kernel estimator of $m(\mathbf{x})$ is given by

$$\widehat{m}_n(\mathbf{x}) = \sum_{i=1}^{n} Y_i \, \mathcal{K}((\mathbf{x} - \mathbf{X}_i)/h) \Big/ \sum_{i=1}^{n} \mathcal{K}(\mathbf{x} - \mathbf{X}_i)/h \,, \tag{1}$$

where the function $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}_+$ is the kernel used with bandwidth $h > 0$. A global measure of performance of $\widehat{m}_n(\cdot)$ as an estimator of $m(\cdot)$ is the L_p-type statistic $I_n(p) = \int |\widehat{m}_n(\mathbf{x}) - m(\mathbf{x})|^p \, \mu(d\mathbf{x})$, $1 \le p < \infty$, where μ is the probability measure of $\mathbf{X}$. The quantity $I_n(1)$ plays an important role in statistical classification; see for example Devroye et al. (1996), Sec. 6.2). For the *almost sure* convergence of $I_n(1)$ to zero see, for example, Devroye and Krzyzak (1989).

Now suppose that the response variable Y may be missing according to the NMAR mechanism, then the estimator $\widehat{m}_n(\mathbf{x})$ in (1) is no longer available and, furthermore, the estimator based on the complete cases alone is not the correct estimator in general. To appreciate this, define the indicator variable $\Delta = 0$ if Y is missing, and $\Delta = 1$ otherwise. Similarly, for $i = 1, \ldots, n$, let $\Delta_i = 0$ when Y_i is missing (and $\Delta_i = 1$ otherwise). Then, the complete-case estimator

$$m_n^{\mathrm{cc}}(\mathbf{x}) \; := \sum_{i=1}^{n} \Delta_i Y_i \, \mathcal{K}((\mathbf{x} - \mathbf{X}_i)/h) \Big/ \sum_{i=1}^{n} \Delta_i \mathcal{K}((\mathbf{x} - \mathbf{X}_i)/h). \tag{2}$$

* Corresponding author.
E-mail address: majid.mojirsheibani@csun.edu (M. Mojirsheibani).

https://doi.org/10.1016/j.spl.2024.110246
Received 5 February 2024; Received in revised form 11 August 2024; Accepted 12 August 2024
Available online 21 August 2024

turns out to be the kernel-type estimator of the quantity $E(\Delta Y|\mathbf{X}=\mathbf{x})/E(\Delta|\mathbf{X}=\mathbf{x})$ which, in general, is not equal to the regression function $m(\mathbf{x})= E(Y|\mathbf{X}=\mathbf{x})$ under the NMAR missing response mechanism. Of course, when the MAR assumption holds, i.e., when $E(\Delta|\mathbf{X},Y)=E(\Delta|\mathbf{X})$, the expression in (2) is the correct estimator of $m(\mathbf{x})$ because $E(\Delta Y|\mathbf{X})=E(\Delta|\mathbf{X})m(\mathbf{X})$.

In order to present and construct our proposed estimator, we start by considering a flexible logistic-type NMAR selection probability model that works as follows. For any real-valued functions $\varphi>0$ on $\mathbb{R}$ and g on $\mathbb{R}^d$, define

$$\pi_\varphi(\mathbf{x},y) := \left[1+\exp\{g(\mathbf{x})\}\cdot\varphi(y)\right]^{-1}. \tag{3}$$

Then we consider the following generalization of the popular model of Kim and Yu (2011)

$$P\left\{\Delta=1\,\middle|\,\mathbf{X}=\mathbf{x},Y=y\right\} = \left[1+\exp\{g(\mathbf{x})\}\cdot\varphi^*(y)\right]^{-1} := \pi_{\varphi^*}(\mathbf{x},y), \tag{4}$$

where φ^* represents the true function φ that could depend on unknown parameters and g is a completely unspecified function. The case where $\varphi^*(y)=\exp(\gamma^* y)$ in (4) for some parameter γ^* corresponds to the original model of Kim and Yu (2011) which has been studied and explored extensively in the literature; see, for example, Zhao and Shao (2015), Shao and Wang (2016), Morikawa et al. (2017), Morikawa and Kim (2018), Morikawa and Kano (2018), Fang et al. (2018), O'Brien et al. (2018), Maity et al. (2019), Sadinle and Reiter (2019), Chen et al. (2020), Liu and Yau (2021), and Mojirsheibani (2022).

It is well-understood in the framework of NMAR missing data that in a fully nonparametric setup where $\pi_\varphi(\mathbf{x},y)$ and the distribution of $(\mathbf{X},Y)$ are completely unknown, one faces the issue of non-identifiability when estimating quantities such as φ (Shao and Wang, 2016). On the other hand, imposing parametric models on both $\pi_\varphi(\mathbf{x},y)$ and the distribution of $(\mathbf{X},Y)$ is too strong to be useful in practice (Molenberghs and Kenward, 2007). Some authors have assumed a fully parametric model for $\pi_\varphi(\mathbf{x},y)$ only, but not the underlying distributions (Qin et al., 2002), but this is also considered to be too strong in practice.

Due to identifiability issues, it turns out that the estimation of (4) can be challenging and that a sufficient condition for model identification is to assume (see, for example, Shao and Wang (2016) that there is a part of $\mathbf{X}$, say $\mathbf{V}$, which is conditionally independent of Δ, given Y and $\mathbf{Z}$, where $\mathbf{X}=(\mathbf{Z},\mathbf{V})\in\mathbb{R}^d$ and $\mathbf{Z}\in\mathbb{R}^q$, with $1\le q<d$. Of course, this approach fails for the important case of $\mathbf{X}\in\mathbb{R}^1$. Furthermore, finding consistent estimators of φ^* based on the above assumption on $\mathbf{X}$ does not necessarily yield strong optimality (in L_p norms) of kernel regression estimators in general. To tackle these difficulties, we consider the approach of Kim and Yu (2011) where one has access to a small follow-up subsample of response values selected from the set of non-respondents. This approach, which enables one to perform various estimations (even for the case where $\mathbf{X}$ is in $\mathbb{R}^1$), does not suffer from identifiability issues; see, for example, Shao and Wang (2016) and Kim and Yu (2011). An even more attractive feature of our approach is that the subsample size can be negligibly small. For example, as our numerical work shows, when $n=100$ and the missing rate is at about 50% then on average a follow-up subsample of size around 2 will be sufficient to carry out all estimations! We have pressed this issue here to emphasize that the seemingly undesirable need for a follow-up subsample can in practice be a non-issue.

Regression function estimation with NMAR response variables is generally viewed to be a challenging problem in the literature. In fact, to the best of our knowledge, there are only a few results available in the literature that also address the theoretical validity of their proposed methods. These include the work of Niu et al. (2014) and Guo et al. (2019) for the case of linear regression, those of Bindele and Zhao (2018) to estimate β in the model $E(Y|\mathbf{X}=\mathbf{x})=g(\mathbf{x},\beta)$, where g is completely known, and the results of Li et al. (2018) for parameter estimation in functional linear regression. The current work does not impose linearity or any other known structures on the underlying regression function. In the case of nonparametric regression, Mojirsheibani (2022) constructed a new regression estimator and derived the limiting distribution of its maximal deviation.

The rest of the paper is organized as follows. Section 2 presents the main results where we outline the mechanics of the proposed regression estimators and study their asymptotic properties in general L_p norms. Section 3 discusses the applications of our estimators to the problem of nonparametric classification with partially observed data, where we also look into the rates of convergence of the proposed classifiers under different conditions. Numerical studies are presented in Section 4. All proofs are deferred to Supplemental materials.

2. Main results

2.1. The proposed estimator

To motivate our proposed estimator, we first observe that the regression function $m(\mathbf{x})=E[Y|\mathbf{X}=\mathbf{x}]$ can also be represented as follows (see Lemma 1 in the Supplementary material)

$$m(\mathbf{x}) = m(\mathbf{x};\varphi^*) := \eta_1(\mathbf{x}) + \frac{\psi_1(\mathbf{x};\varphi^*)}{\psi_2(\mathbf{x};\varphi^*)}\cdot\left(1-\eta_2(\mathbf{x})\right), \tag{5}$$

where φ^* is as in (4) and the functions ψ_k and η_k, $k=1,2$, are conditional expectations given by

$$\psi_k(\mathbf{x};\varphi^*) := E\left[\Delta Y^{2-k}\varphi^*(Y)\middle|\mathbf{X}=\mathbf{x}\right] \quad\text{and}\quad \eta_k(\mathbf{x}) := E\left[\Delta Y^{2-k}\middle|\mathbf{X}=\mathbf{x}\right], \quad \text{for } k=1,2. \tag{6}$$

Now let $\mathbb{D}_n=\{(\boldsymbol{X}_1,Y_1,\Delta_1),\ldots,(\boldsymbol{X}_n,Y_n,\Delta_n)\}$ represent a sample of size n (iid), where $\Delta_i=0$ if Y_i is missing (and $\Delta_i=1$ otherwise). Then in the hypothetical situation where φ^* is completely known in (4), one can consider the following estimator of $m(\mathbf{x})$

$$\widehat{m}(\mathbf{x};\varphi^*)=\widehat{\eta}_1(\mathbf{x})+\frac{\widehat{\psi}_1(\mathbf{x};\varphi^*)}{\widehat{\psi}_2(\mathbf{x};\varphi^*)}\left(1-\widehat{\eta}_2(\mathbf{x})\right), \tag{7}$$

where $\widehat{\psi}_k(\mathbf{x};\varphi^*)$ and $\widehat{\eta}_k(\mathbf{x})$, $k=1,2$, are the kernel estimators given by

$$\widehat{\psi}_k(\mathbf{x};\varphi^*) = \sum_{i=1}^n \Delta_i Y_i^{2-k}\varphi^*(Y_i)\,\mathcal{K}((\mathbf{x}-\mathbf{X}_i)/h)\Big/\sum_{i=1}^n \mathcal{K}((\mathbf{x}-\mathbf{X}_i)/h), \tag{8}$$

$$\widehat{\eta}_k(\mathbf{x}) = \sum_{i=1}^n \Delta_i Y_i^{2-k}\mathcal{K}((\mathbf{x}-\mathbf{X}_i)/h)\Big/\sum_{i=1}^n \mathcal{K}((\mathbf{x}-\mathbf{X}_i)/h), \tag{9}$$

where, as before, $\mathcal{K}:\mathbb{R}^d\to\mathbb{R}_+$ is the kernel with bandwidth $h>0$. Clearly the estimator in (7) is not available because φ^* is unknown and must be estimated. To that end, we employ a data splitting approach that works as follows. Start by randomly splitting the data into a training sample $\mathbb{D}_m$ of size m and a validation set $\mathbb{D}_\ell$ of size $\ell=n-m$. It is assumed that as $n\to\infty$, both $\ell\to\infty$ and $m\to\infty$; the choices of m and ℓ will be discussed later. Also, define the index sets $\mathcal{I}_m=\big\{i\in\{1,\dots,n\}\,\big|\,(\mathbf{X}_i,Y_i,\Delta_i)\in\mathbb{D}_m\big\}$ and $\mathcal{I}_\ell=\big\{i\in\{1,\dots,n\}\,\big|\,(\mathbf{X}_i,Y_i,\Delta_i)\in\mathbb{D}_\ell\big\}$. Let $\mathcal{F}$ be the class of functions to which the unknown function φ^* in (4) belongs and for each $\varphi\in\mathcal{F}$ consider the following counterpart of (7) constructed based on the training set $\mathbb{D}_m$ only

$$\widehat{m}_m(\mathbf{x};\varphi)=\widehat{\eta}_{m,1}(\mathbf{x})+\frac{\widehat{\psi}_{m,1}(\mathbf{x};\varphi)}{\widehat{\psi}_{m,2}(\mathbf{x};\varphi)}\big(1-\widehat{\eta}_{m,2}(\mathbf{x})\big), \tag{10}$$

where

$$\widehat{\psi}_{m,k}(\mathbf{x};\varphi) = \frac{\sum_{i\in\mathcal{I}_m}\Delta_i Y_i^{2-k}\varphi(Y_i)\,\mathcal{K}((\mathbf{x}-\mathbf{X}_i)/h)}{\sum_{i\in\mathcal{I}_m}\mathcal{K}((\mathbf{x}-\mathbf{X}_i)/h)},\quad k=1,2,\quad \varphi\in\mathcal{F}, \tag{11}$$

$$\widehat{\eta}_{m,k}(\mathbf{x}) = \frac{\sum_{i\in\mathcal{I}_m}\Delta_i Y_i^{2-k}\mathcal{K}((\mathbf{x}-\mathbf{X}_i)/h)}{\sum_{i\in\mathcal{I}_m}\mathcal{K}((\mathbf{x}-\mathbf{X}_i)/h)},\quad k=1,2. \tag{12}$$

Next, to estimate φ^*, we will use the approach based on the approximation theory of totally bounded function spaces. More specifically, let $\mathcal{F}$ be a given class of functions $\varphi\colon\ [-L,L]\to(0,B]$, for some $B<\infty$ and finite $L>0$. Fix $\varepsilon>0$ and suppose that the finite collection of functions $\mathcal{F}_\varepsilon=\{\varphi_1,\dots,\varphi_{\mathrm{N}(\varepsilon)}\}$ $\varphi_i\colon\ [-L,L]\to(0,B]$, is an ε-cover of $\mathcal{F}$, i.e., for each $\varphi\in\mathcal{F}$, there is a $\varphi'\in\mathcal{F}_\varepsilon$ such that $\|\varphi-\varphi'\|_\infty<\varepsilon$; here, $\|\cdot\|_\infty$ is the usual supnorm. The cardinality of the smallest ε-cover of $\mathcal{F}$ is called the *covering number* of $\mathcal{F}$ and will be denoted by $\mathcal{N}_\varepsilon(\mathcal{F})$. If $\mathcal{N}_\varepsilon(\mathcal{F})<\infty$ holds for every $\varepsilon>0$, then the family $\mathcal{F}$ is said to be *totally bounded* (with respect to $\|\cdot\|_\infty$). The quantity $\log(\mathcal{N}_\varepsilon(\mathcal{F}))$ is called Kolmogorov's ϵ-entropy of $\mathcal{F}$. For more on such concepts see, for example, the monograph by van der Vaart and Wellner (1996, p. 83).

Now, let $0<\varepsilon_n\downarrow 0$ be a decreasing sequence, as $n\to\infty$, and let $\mathcal{F}_{\varepsilon_n}=\{\varphi_1,\dots,\varphi_{N(\varepsilon_n)}\}\subset\mathcal{F}$ be any ε_n-cover of $\mathcal{F}$; the choice of ε_n will be discussed later in Corollary 1. Also, as explained in the introduction, here we consider the setup in which one has access to response values for a small follow-up subsample selected from the set of non-respondents. More formally, let δ_i, $i=1,\dots,\ell$, be iid Bernoulli random variables, independent of the data $\mathbb{D}_n$, with the probability of success

$$p_n=P\{\delta_i=1\},\quad i=1,\dots,\ell,\quad \text{with } p_n\to 0,\ \text{as } n\to\infty. \tag{13}$$

Then we select a non-respondent in the validation set $\mathbb{D}_\ell$ to be included in the small follow-up subsample only when $(1-\Delta_i)\,\delta_i=1$, $i\in\mathcal{I}_\ell$, where $\Delta_i=0$ if Y_i is missing. Next, for each $\varphi\in\mathcal{F}_{\varepsilon_n}$ define the empirical L_2 error of the estimator $\widehat{m}_m(\mathbf{x};\varphi)$ in (10), based on the validation set $\mathbb{D}_\ell$, by

$$\begin{aligned}\widehat{L}_{m,\ell}(\varphi) &= \frac{1}{\ell}\left[\sum_{i\in\mathcal{I}_\ell}\Delta_i\big|\widehat{m}_m(\mathbf{X}_i;\varphi)-Y_i\big|^2+\sum_{i\in\mathcal{I}_\ell}(1-\Delta_i)(\delta_i/p_n)\big|\widehat{m}_m(\mathbf{X}_i;\varphi)-Y_i\big|^2\right]\\ &= \frac{1}{\ell}\sum_{i\in\mathcal{I}_\ell}\left(\Delta_i+\frac{(1-\Delta_i)\delta_i}{p_n}\right)\big|\widehat{m}_m(\mathbf{X}_i;\varphi)-Y_i\big|^2.\end{aligned} \tag{14}$$

Our estimator of φ^* is then given by

$$\widehat{\varphi}_n = \operatorname*{argmin}_{\varphi\in\mathcal{F}_{\varepsilon_n}}\widehat{L}_{m,\ell}(\varphi). \tag{15}$$

The subscript n at $\widehat{\varphi}_n$ reflects the fact that the entire data of size n has been used here. Finally, our proposed estimator of $m(\mathbf{x})$ is given by

$$\widehat{m}(\mathbf{x};\widehat{\varphi}_n)\ :=\ \widehat{m}_m(\mathbf{x};\varphi)\Big|_{\varphi=\widehat{\varphi}_n}\quad \text{where } \widehat{m}_m(\mathbf{x};\varphi) \text{ is as in (10).} \tag{16}$$

2.2. How good is $\widehat{m}(\mathbf{x};\widehat{\varphi}_n)$ in (16) as an estimator of $m(\mathbf{x})$?

To answer this, we assume that the kernel $\mathcal{K}$ is *regular* in the sense of Devroye and Krzyzak (1989):

Definition. A nonnegative kernel $\mathcal{K}$ is said to be regular if there are real constants $b>0$ and $r>0$ such that $\mathcal{K}(\mathbf{u})\ge b\,I\{\mathbf{u}\in S_{0,r}\}$ and $\int\sup_{\mathbf{y}\in\mathbf{u}+S_{0,r}}\mathcal{K}(\mathbf{y})\,d\mathbf{u}<\infty$, where $S_{0,r}$ is the ball of radius r centered at the origin.

Theorem 1. *Suppose that the regularity conditions described in the supplementary section hold. Then, for every $\varepsilon_n > 0$ satisfying $\varepsilon_n \downarrow 0$, as $n \to \infty$, every $t > 0$, any distribution of $(\mathbf{X}, Y) \in \mathbb{R}^d \times [-L, L]$, $L < \infty$, and n large enough,*

$$P\left\{ \int \left|\widehat{m}(\mathbf{x}; \widehat{\varphi}_n) - m(\mathbf{x})\right|^2 \mu(d\mathbf{x}) > t \right\} \le c_1 \left|\mathcal{F}_{\varepsilon_n}\right| \left(e^{-c_2\, \ell\, p_n^2 t^2} + \ell\, e^{-c_3\, m h^d\, p_n^2 t^2} \right) \tag{17}$$

whenever $\varphi^* \in \mathcal{F}$, where c_1, c_2, and c_3 are positive constants not depending on m, ℓ, n, or t, and where $|\mathcal{F}_{\varepsilon_n}|$ is the cardinality of the set $\mathcal{F}_{\varepsilon_n}$.

Remark 1 ($p \ge 2$)**.** The above theorem is stated in the L_2 sense; it is straightforward to show that the theorem holds for all $p \ge 2$. To see this, observe that if $p > 2$ then one can always write

$$\left|\widehat{m}(\mathbf{x}; \widehat{\varphi}_n) - m(\mathbf{x})\right|^p \le \left(\left|\widehat{m}(\mathbf{x}; \widehat{\varphi}_n)\right| + \left|m(\mathbf{x})\right| \right)^{p-2} \left|\widehat{m}(\mathbf{x}; \widehat{\varphi}_n) - m(\mathbf{x})\right|^2 \le (3L)^{p-2} \left|\widehat{m}(\mathbf{x}; \widehat{\varphi}_n) - m(\mathbf{x})\right|^2.$$

Additionally, we note that if $p \in [1, 2)$, then by Lyapunov's inequality (for expectations) we have $P\{\int |\widehat{m}(\mathbf{x}; \widehat{\varphi}_n) - m(\mathbf{x})|^p \mu(d\mathbf{x}) > t\} \le P\{\int |\widehat{m}(\mathbf{x}; \widehat{\varphi}_n) - m(\mathbf{x})|^2 \mu(d\mathbf{x}) > t^{2/p}\}$.

The following result follows from Theorem 1 in conjunction with the Borel–Cantelli lemma.

Corollary 1. *Consider the estimator in (16) and let p_n be as in (13). If, as $n \to \infty$,*

$$\varepsilon_n \downarrow 0, \quad (\ell\, p_n^2)^{-1} \log(n \vee |\mathcal{F}_{\varepsilon_n}|) \longrightarrow 0, \quad \text{and} \quad (m h^d p_n^2)^{-1} \log(n \vee |\mathcal{F}_{\varepsilon_n}|) \longrightarrow 0, \tag{18}$$

then, under the conditions of Theorem 1, $E\left[\left|\widehat{m}(\mathbf{X}; \widehat{\varphi}_n) - m(\mathbf{X})\right|^p \middle| \mathbb{D}_n\right] \to^{a.s.} 0$, for all $p \in [2, \infty)$.

We also note that by Lebesgue dominated convergence theorem, under the conditions of Corollary 1 without further ado we have $E\left|\widehat{m}(\mathbf{X}; \widehat{\varphi}_n) - m(\mathbf{X})\right|^p \to 0$, for all $p \in [2, \infty)$. However, to study the rates of convergence here, we first state the following theorem.

Theorem 2. *Suppose that the conditions of Theorem 1 hold. Then, for n large enough,*

$$E\left|\widehat{m}(\mathbf{X}; \widehat{\varphi}_n) - m(\mathbf{X})\right|^p \le \sqrt{\frac{c_5 + \log \ell + \log |\mathcal{F}_{\varepsilon_n}|}{c_6 \cdot (\ell \wedge m h^d)\, p_n^2}} + \sqrt{\frac{1}{c_7 \cdot (\ell \wedge m h^d)\, p_n^2 \left[c_5 + \log \ell + \log |\mathcal{F}_{\varepsilon_n}|\right]}}$$

for all $p \in [2, \infty)$, where c_5, c_6, and c_7 are positive constants not depending on m, ℓ, or n.

The following result looks at the rate of convergence of the proposed regression estimator.

Corollary 2. *Suppose that (18) holds. Then, under the conditions of Theorem 1, for all $p \ge 2$,*

$$E\left|\widehat{m}(\mathbf{X}; \widehat{\varphi}_n) - m(\mathbf{X})\right|^p = \mathcal{O}\left(\sqrt{\frac{\log(\ell \vee |\mathcal{F}_{\varepsilon_n}|)}{(\ell \wedge m h^d) \cdot p_n^2}} \right).$$

In the special case where $m = \alpha \cdot n$ and $\ell = (1 - \alpha) \cdot n$, where $\alpha \in (0, 1)$, one finds for all $p \ge 2$,

$$E\left|\widehat{m}(\mathbf{X}; \widehat{\varphi}_n) - m(\mathbf{X})\right|^p = \mathcal{O}\left(\sqrt{\frac{\log(n \vee |\mathcal{F}_{\varepsilon_n}|)}{n h^d \cdot p_n^2}} \right).$$

A close inspection of Corollary 2 shows that choosing ℓ and m to satisfy $\ell/n \to 0$ or $m/n \to 0$ results in estimators with rates of convergence worse than the case $m = \lfloor \alpha n \rfloor$ for any $\alpha \in (0, 1)$.

Remark 2 (*Rates of Convergence*)**.** The rates of convergence in Corollary 2 are generally not optimal as compared to those of kernel estimators based on no missing data. A better rate would be of order $\mathcal{O}\big(\sqrt{\log n / n h^d}\big)$ which is achievable if the following conditions hold: (i) $p_n = c \in (0, 1]$ for some fixed probability c instead of $p_n = o(1)$, (ii) the cardinality of the ε_n-cover satisfies $\log |\mathcal{F}_{\varepsilon_n}| = \mathcal{O}(n)$, and (iii) m is chosen as $m = \lfloor \alpha n \rfloor$ for some $\alpha \in (0, 1)$. When there are no missing data, it is well known in the framework of kernel regression that with additional assumptions such as Lipschitz continuity of the regression function $m(\mathbf{x})$, one can establish rates as fast as $\mathcal{O}\big((n h^d)^{-1} + h^2\big)$ for the usual kernel estimator in (1) based on the *naive* kernel; see, for example, Györfi et al. (2002; Sec. 5.3). Unfortunately, such rates do not seem to be available for our estimators with NMAR missing data where the estimation process involves many steps and many components. In fact, to the best of our knowledge, such fast rates are not available even for the simpler case of MAR missing data. The dependence of the rate of convergence on p_n in Corollary 2 shows that if obtaining a follow-up subsample is not too inconvenient, then one can have good rates by taking p_n to be a fixed percentage, such as 15%, of the entire data (as in Kim and Yu, 2011). Otherwise, by choosing $p_n = o(1)$ appropriately, one requires a much smaller subsample size while still retaining the convergence in Corollary 2, but at rates slower than $\mathcal{O}\big(\sqrt{\log n / n h^d}\big)$.

3. Applications to classification with possibly missing labels

In this section we consider the following two-group classification problem. Let $(\mathbf{X}, Y) \in \mathbb{R}^d \times \{0,1\}$ be a random pair where the class label Y has to be predicted based on the covariate $\mathbf{X}$. More specifically, the goal is to find a function $g : \mathbb{R}^d \to \{0,1\}$ for which the misclassification error, i.e.,

$$L(g) := P\{g(\mathbf{X}) \neq Y\}, \tag{19}$$

is as small as possible. The optimal classifier, also referred to as the Bayes classifier, is given by

$$g_{\mathrm{B}}(\mathbf{x}) = 1 \text{ if } m(\mathbf{x}) > 1/2, \quad (g_{\mathrm{B}}(\mathbf{x}) = 0, \text{ otherwise}) \tag{20}$$

where $m(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}]$; see, for example, Chapter 2 of Devroye et al. (1996). In practice, the distribution of $(\mathbf{X}, Y)$ is almost always unknown and therefore finding g_{B} is impossible. Suppose that we have access to n iid observations (the data), $\mathbb{D}_n := \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$, where $(\mathbf{X}_i, Y_i) \overset{\text{iid}}{=} (\mathbf{X}, Y)$, $i = 1, \ldots, n$, and let $\widehat{g}_n$ be any classifier constructed based on the data $\mathbb{D}_n$. Also, let

$$L_n(\widehat{g}_n) = P\left\{\widehat{g}_n(\mathbf{X}) \neq Y \,\middle|\, \mathbb{D}_n\right\} \tag{21}$$

be the conditional misclassification error of $\widehat{g}_n$. Now, let $\widehat{m}(\mathbf{x})$ be any estimator of the regression function $m(\mathbf{x})$ and consider the plug-in type classifier

$$\widehat{g}_n(\mathbf{x}) = 1 \text{ if } \widehat{m}(\mathbf{x}) > 1/2, \quad (\widehat{g}_n(\mathbf{x}) = 0, \text{ otherwise.}) \tag{22}$$

Then the following bound follows from Devroye et al. (1996); Lemma 6.1)

$$L_n(\widehat{g}_n) - L(g_{\mathrm{B}}) \;\le\; 2E\left[\left|\widehat{m}(\mathbf{X}) - m(\mathbf{X})\right| \,\middle|\, \mathbb{D}_n\right], \tag{23}$$

and thus $E\left[L_n(\widehat{g}_n)\right] - L(g_{\mathrm{B}}) \;\le\; 2E\left|\widehat{m}(\mathbf{X}) - m(\mathbf{X})\right|$. Now, suppose that some of the Y_i's may be missing not at random (NMAR) and consider the plug-in classifier corresponding to (16)

$$\widehat{g}_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \widehat{m}(\mathbf{x}; \widehat{\varphi}_n) > \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{24}$$

where $\widehat{m}(\mathbf{x}; \widehat{\varphi}_n)$ is as in (16). To study the asymptotic performance of the classifier in (24), we also state the following so-called margin condition which can be found in Audibert and Tsybakov (2007).

Assumption (G) *[Margin condition.]* There exist constants $c > 0$ and $\alpha > 0$ such that

$$P\left\{0 < \left|m(\mathbf{X}) - 0.5\right| \le t\right\} \;\le\; c\, t^{\alpha}, \quad \text{for all } t > 0. \tag{25}$$

Several authors have studied the margin condition (25); these include Mammen and Tsybakov (1999), Audibert and Tsybakov (2007), Kohler and Krzyzak (2007), and Döring et al. (2016).

Theorem 3. *Suppose that (18) holds. Then, under the conditions of Theorem 1,*

(i) $P\left\{\widehat{g}_n(\mathbf{X}) \neq Y \,\middle|\, \mathbb{D}_n\right\} \xrightarrow{a.s.} P\{g_B(\mathbf{X}) \neq Y\}$.

(ii) $P\left\{\widehat{g}_n(\mathbf{X}) \neq Y\right\} - P\{g_B(\mathbf{X}) \neq Y\} = \mathcal{O}\left(\left(\frac{\log(\ell^{\vee}|\mathcal{F}_{\epsilon_n}|)}{(\ell^{\wedge} m h^d)\cdot p_n^2}\right)^{1/4}\right)$.

(iii) *If (25) holds then* $P\left\{\widehat{g}_n(\mathbf{X}) \neq Y\right\} - P\{g_B(\mathbf{X}) \neq Y\} = \mathcal{O}\left(\left(\frac{\log(\ell^{\vee}|\mathcal{F}_{\epsilon_n}|)}{(\ell^{\wedge} m h^d)\cdot p_n^2}\right)^{\frac{1+\alpha}{2(2+\alpha)}}\right)$.

Part (iii) shows that the rate in Part (ii) can come closer to the better rate in Corollary 2 whenever the regression function $m(\mathbf{x})$ satisfies condition (25); in fact, $\frac{1+\alpha}{2(2+\alpha)} \to \frac{1}{2}$ as α diverges.

4. Numerical studies

For the numerical work, we generated $n = 50, 100$ observations from the following two models:

Model A. $\mathbf{X} \sim N_5(\mathbf{1}, \boldsymbol{\Sigma})$ and $Y = \mu_y - X_1 + X_3 X_4 - X_2^2 + \exp(-X_5) + N(0, \sigma_y^2)$

Model B. $\mathbf{X} \sim N_4(\mathbf{0}, \boldsymbol{\Sigma})$ and $Y = X_1 + (2X_2 - 1)^2 + \frac{\sin(2\pi X_3)}{2 - \sin(2\pi X_3)} + \sin(2\pi X_4) + 2\cos(2\pi X_4)$
$+ 3\sin^2(2\pi X_4) + 4\cos^2(2\pi X_4) + N(0, \sigma_y^2)$,

where $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the d-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j\ge 1}$ with $\sigma_{ij} = 2^{-|i-j|+1}$ in Model A and $\sigma_{ij} = 2^{-|i-j|}$ in Model B. As for σ_y, two values are considered, 0.5 and 4 (high variance model); in Model A we used two values of μ_y: 1 and 2.6. Here, Model B is as in Meier et al. (2009). Next, we also considered two choices for the function φ^* in (4), $\varphi^*(y) = \exp(\gamma^* y)$ as in Kim and Yu (2011) and $\varphi^*(y) = \left[0.1 + (\gamma^* y)^2\right]^{-1}$.

The following choice of coefficients result in approximately 50% missing rate in Model A:

(A1) $\pi(\mathbf{x}, y) = \left(1 + \exp\left\{\beta_0 + \sum_{j=1}^5 \beta_j x_j\right\} \cdot \exp\{\gamma y\}\right)^{-1}$
with $(\beta_0, \ldots, \beta_5) = (0.6,\ 0.8,\ 0.25,\ -0.35,\ -0.3,\ 0.75)$, $\gamma = -0.98$, and $\mu_y = 2.6$.

Table 1
Empirical L_1 and L_2 errors corresponding selection probabilities (A1) and (A2). Here, the proposed estimator $\widehat{m}(\mathbf{x}; \widehat{\pi}_{\widehat{\varphi}_n})$ is as in (16), the complete-case estimator $m_n^{cc}(\mathbf{x})$ is given by (2), and the estimator $\widehat{m}_n(\mathbf{x})$ based on no missing data is given by (1). The numbers in parentheses are standard errors and those in square brackets are average follow-up subsample sizes drawn from the non-respondents in $\mathbb{D}_\ell$.

Selection Prob Model	Sample Size (n)	Noise in Model A	Error Type	$\widehat{m}(\mathbf{x}; \widehat{\pi}_{\widehat{\varphi}_n})$	$m_n^{cc}(\mathbf{x})$	$\widehat{m}_n(\mathbf{x})$
(A1)	50	$\sigma_y = 0.5$	L_2	24.14 (0.5904), **[1.46]**	37.46 (1.1949)	16.91 (0.6497)
			L_1	3.10 (0.0252)	4.23 (0.0741)	2.50 (0.0427)
		$\sigma_y = 4$	L_2	47.22 (1.0872), **[1.47]**	67.56 (0.8659)	40.07 (0.6031)
			L_1	5.11 (0.0485)	6.35 (0.0440)	4.73 (0.0283)
	100	$\sigma_y = 0.5$	L_2	20.93 (0.5256), **[2.10]**	32.47 (0.8113)	11.91 (0.5364)
			L_1	2.82 (0.0219)	3.82 (0.0587)	2.00 (0.0264)
		$\sigma_y = 4$	L_2	40.99 (0.7606), **[2.17]**	65.29 (0.7808)	34.95 (0.7369)
			L_1	4.75 (0.0321)	6.22 (0.0309)	4.39 (0.0320)
(A2)	50	$\sigma_y = 0.5$	L_2	26.66 (1.0315), **[1.44]**	36.36 (1.0980)	17.55 (0.9804)
			L_1	3.23 (0.0263)	4.10 (0.0451)	2.51 (0.0420)
		$\sigma_y = 4$	L_2	47.63 (0.9706), **[1.50]**	65.58 (0.7706)	40.57 (0.5967)
			L_1	5.14 (0.0398)	6.26 (0.0412)	4.76 (0.0273)
	100	$\sigma_y = 0.5$	L_2	21.78 (0.3435), **[2.20]**	30.50 (0.5564)	11.53 (0.3971)
			L_1	2.94 (0.0210)	3.68 (0.0466)	2.02 (0.0299)
		$\sigma_y = 4$	L_2	42.88 (0.7521), **[2.07]**	64.06 (0.6096)	35.41 (0.6778)
			L_1	4.86 (0.0323)	6.17 (0.0303)	4.44 (0.0332)

(A2) $\pi(\mathbf{x}, y) = \left(1 + \exp\{\beta_0 + \sum_{j=1}^4 \beta_j x_j + \beta_5 \gamma\, x_5\} \cdot \exp\{\gamma y\}\right)^{-1}$
with $(\beta_0, \beta_1, \ldots, \beta_5) = (0.50,\ 0.75,\ -0.25,\ 0.25,\ -0.25,\ 0.75)$, $\gamma = -0.98$, and $\mu_y = 1$.
The following choices of coefficients result in approximately 50% missing rate in Model B:
(B1) $\pi(\mathbf{x}, y) = \left(1 + \exp\{\beta_0 \gamma + \sum_{j=1}^4 \beta_j x_j\} \cdot \left(0.1 + \gamma^2 y^2\right)^{-1}\right)^{-1}$
with $(\beta_0, \ldots, \beta_4) = (0.85,\ 0.6,\ 0.35,\ -0.45,\ 0.55)$ and $\gamma = 0.16$.
(B2) $\pi(\mathbf{x}, y) = \left(1 + \exp\{\beta_0 + \sum_{j=1}^3 \beta_j x_j + \beta_4 \gamma\, x_4\} \cdot \exp\{\gamma y\}\right)^{-1}$
with $(\beta_0, \beta_1, \ldots, \beta_4) = (2.6,\ 0.6,\ 0.35,\ -0.45,\ 0.4)$ and $\gamma = -0.36$.

To estimate γ^* in $\varphi^*(y) = \exp(\gamma^* y)$, we employed the data-splitting approach of Section 2 with $m = 0.7n$ and $\ell = 0.3n$, where the estimator of γ^* is the minimizer of (14) with respect to γ over a grid of equally-spaced values of γ in $[-M, M]$; here, we took $M = 15$. Next, a small follow-up subsample was selected from the set of non-respondents in $\mathbb{D}_\ell$ where we took $p_n = \left((\log n)^{0.25}/(n\lambda^d)^{1-\alpha}\right)^{1/2}$ in (13) with $\lambda = 0.95$ and $\alpha = 0.01$. This choice of p_n assures a very small subsample size (see the results in Tables 1 and 2). We employed the Gaussian kernel in our estimators where the bandwidths were selected using the cross-validation method of Racine and Li (2004) available from the R package "np"; see Racine and Hayfield (2008). To assess the performance of the proposed estimators we computed their empirical L_2 errors (mean squared prediction errors) committed on a validation set of 1000 additional observations. We also computed the empirical L_1 errors of our estimators. The entire above process was repeated 200 times, each time using a sample of size n (50 and 100) and a validation set of size 1000, and the average errors were computed.

Table 1 gives the results for models (A1) and (A2). Table 1 also gives the corresponding results for two other estimators: the complete-case estimator $m_n^{cc}(\mathbf{x})$ in (2) and the usual estimator with no missing data; this latter estimator allows one to see how different the results could have been (and how close our results are to them) if there had not been any missing values. We also performed the same computations for Model B with the two selections probabilities (B1) and (B2); the corresponding results appear in Table 2. As the tables show, the proposed estimator's error rates are significantly lower than those of the complete case estimators. But more importantly, the tables also show that the average follow-up subsample sizes needed for the proposed estimators are around 2.1 when $n = 100$ and about 1.5 when $n = 50$ (see the boldfaced values in square brackets). In other words, the undesirable need for a follow-up subsample here is virtually a non-issue in practice. As mentioned in Section 1, the complete-case estimator will be

Table 2
Empirical L_1 and L_2 errors corresponding selection probabilities (B1) and (B2). Here, the proposed estimator $\widehat{m}(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_n})$ is as in (16), the complete-case estimator $m_n^{cc}(\mathbf{x})$ is given by (2), and the estimator $\widehat{m}_n(\mathbf{x})$ based on no missing data is given by (1). The numbers in parentheses are standard errors and those in square brackets are the average follow-up subsample sizes drawn from the set of non-respondents in $\mathbb{D}_\ell$.

Selection Prob Model	Sample Size (n)	Noise in Model B	Error Type	$\widehat{m}(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_n})$	$m_n^{cc}(\mathbf{x})$	$\widehat{m}_n(\mathbf{x})$
(B1)	50	$\sigma_y = 0.5$	L_2	51.85 (0.8269), **[1.55]**	53.35 (1.1698)	44.57 (0.7264)
			L_1	4.90 (0.0505)	5.71 (0.0924)	4.55 (0.0509)
		$\sigma_y = 4$	L_2	67.30 (0.7762), **[1.32]**	70.75 (0.9546)	61.62 (0.5441)
			L_1	5.99 (0.0398)	6.63 (0.0616)	5.71 (0.0255)
	100	$\sigma_y = 0.5$	L_2	49.13 (0.8893), **[2.04]**	52.52 (1.0010)	42.40 (0.8758)
			L_1	4.73 (0.0560)	5.74 (0.0821)	4.43 (0.0593)
		$\sigma_y = 4$	L_2	66.21 (0.7095), **[2.06]**	72.10 (0.9379)	60.9001 (0.6425)
			L_1	5.90 (0.0360)	6.75 (0.0576)	5.67 (0.0307)
(B2)	50	$\sigma_y = 0.5$	L_2	53.32 (0.7165), **[1.40]**	57.52 (1.2368)	44.87 (0.7478)
			L_1	4.87 (0.0527)	6.07 (0.0954)	4.58 (0.0542)
		$\sigma_y = 4$	L_2	71.03 (1.2348), **[1.51]**	81.53 (1.4935)	61.87 (0.6288)
			L_1	6.17 (0.0656)	7.33 (0.0862)	5.74 (0.0315)
	100	$\sigma_y = 0.5$	L_2	50.95 (0.7834), **[1.96]**	54.85 (1.1026)	44.13 (0.7440)
			L_1	4.68 (0.0492)	5.97 (0.0873)	4.55 (0.0509)
		$\sigma_y = 4$	L_2	67.44 (0.9639), **[2.04]**	79.26 (1.1646)	61.48 (0.6427)
			L_1	5.94 (0.0533)	7.24 (0.0685)	5.70 (0.0308)

correct under the MAR assumption and therefore one can expect (2) to have a comparable (if not better) finite sample performance in MAR scenarios.

5. Discussion

We have proposed kernel-type estimators of a regression function $m(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}]$, with $\mathbf{x} \in \mathbb{R}^d$ and $Y \in \mathbb{R}$, when the response variable Y may be missing according to a *not-missing-at-random* (NMAR) setup, where the underlying missing probability mechanism can depend on both the predictor $\mathbf{X}$ and the response Y. Our proposed estimator is based on a particular representation of $m(\mathbf{x})$ in terms of four associated conditional expectations that can be estimated nonparametrically. We have established the convergence properties of our estimators in general L_p norms. An application of our results to the problem of classification is also studied.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by National Science Foundation, United States grant DMS-2310504 of M. Mojirsheibani.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2024.110246.

References

Devroye, L., Györfi, L., Lugosi, G., 1996. A Probabilistic Theory of Pattern Recognition. Springer-Verlag, New York.

Devroye, L., Krzyzak, A., 1989. An equivalence theorem for L_1 convergence of kernel regression estimate. J. Statist. Plann. Inference 23, 71–82.

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

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

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

Morikawa, K., Kim, J.K., Kano, Y., 2017. Semiparametric maximum likelihood estimation with data missing not at random. Canad. J. Statist. 45, 393–409.

Morikawa, K., Kim, J.K., 2018. A note on the equivalence of two semiparametric estimation methods for nonignorable nonresponse. Stat. & Probab. Lett 140, 1–6.

Morikawa, K., Kano, Y., 2018. Identification problem of transition models for repeated measurement data with nonignorable missing values. J. Multivariate Anal. 165, 216–230.

Fang, F., Zhao, J., Shao, J., 2018. Imputation-based adjusted score equations in generalized linear models with nonignorable missing covariate values. Statist. Sinica 28, 1677–1701.

O'Brien, J., Gunawardena, H., Paulo, J., Chen, X., Ibrahim, J., Gygi, S., Qaqish, B., 2018. The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. Ann. Appl. Statist 12, 2075–2095.

Maity, A., Pradhan, V., Das, U., 2019. Bias reduction in logistic regression with missing responses when the missing data mechanism is nonignorable. Amer. Statist. 73, 340–349.

Sadinle, M., Reiter, J., 2019. Sequentially additive nonignorable missing data modelling using auxiliary marginal information. Biometrika 106, 889–911.

Chen, X., Diao, G., Qin, J., 2020. Pseudo likelihood-based estimation and testing of missingness mechanism function in nonignorable missing data problems. Scand. J. Stat. 47, 1377–1400.

Liu, Z., Yau, C.Y., 2021. Fitting time series models for longitudinal surveys with nonignorable missing data. J. Statist. Plann. Inference 214, 1–12.

Mojirsheibani, M., 2022. On the maximal deviation of kernel regression estimators with MNAR response variables. Statist. Papers 63, 1677–1705.

Molenberghs, G., Kenward, M., 2007. Missing Data in Clinical Studies. Wiley, New York.

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

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

Guo, X., Song, Y., Zhu, L., 2019. Model checking for general linear regression with nonignorable missing response. Comput. Statist. Data Anal. 138, 1–12.

Bindele, H., Zhao, Y., 2018. Rank-based estimating equation with non-ignorable missing responses via empirical likelihood. Statist. Sinica 28, 1787–1820.

Li, T., Xie, F., Feng, X., Ibrahim, J., Zhu, H., 2018. Functional linear regression models for nonignorable missing scalar responses. Statist. Sinica 28, 1867–1886.

van der Vaart, A., Wellner, J., 1996. Weak Convergence and Empirical Processes with Applications To Statistics. Springer, New York.

Audibert, J.Y., Tsybakov, A.B., 2007. Fast learning rates for plug-in classifiers under the margin condition. Ann. Statist. 35, 608–633.

Mammen, E., Tsybakov, A.B., 1999. Smooth discriminant analysis. Ann. Statist. 27, 1808–1829.

Kohler, M., Krzyzak, A., 2007. On the rate of convergence of local averaging plug-in classification rules under a margin condition. IEEE Trans. Inform. Theory 53, 1735–1742.

Döring, M., Györfi, L., Walk, H., 2016. Exact rate of convergence of kernel-based classification rule, challenges in computational statistics and data mining. Stud. Comput. Intell. 605, 71–91.

Meier, L., van de Geer, S., Bühlmann, P., 2009. High-dimensional additive modeling. Ann. Statist. 37, 3779–3821.

Racine, J., Li, Q., 2004. Nonparametric estimation of regression functions with both categorical and continuous data. J. Econometrics 119, 99–130.

Racine, J., Hayfield, T., 2008. Nonparametric econometrics: The np package. J. Stat. Softw. 27, 1–32.