Annals of the Institute of Statistical Mathematics (2021) 73:535–553
https://doi.org/10.1007/s10463-020-00759-y

Model averaging for linear models with responses missing at random

Yuting Wei[1] · Qihua Wang[2,3] · Wei Liu[4]

Received: 27 August 2019 / Revised: 8 May 2020 / Published online: 1 July 2020

Abstract
In this paper, a model averaging approach is developed for the linear regression models with response missing at random. It is shown that the proposed method is asymptotically optimal in the sense of achieving the lowest possible squared error. A Monte Carlo study is conducted to investigate the finite sample performance of our proposal by comparing with some related methods, and the simulation results favor the proposed method. Moreover, a real data analysis is given to illustrate the practical application of our proposal.

Keywords Missing responses · Missing at random · Model averaging · Asymptotic optimality

1 Introduction

Due to the complication of reality, the true model, from which data were generated, is hard to be known. Thus, in most cases, one may not be able to guess the true model. As pointed out by the maxim formulated by G. E. P. Box "All models are wrong, but some are useful," see, e.g., Claeskens and Hjort (2008). From this

✉ Qihua Wang
qhwang@amss.ac.cn

Yuting Wei
ytwei@mail.ustc.edu.cn

Wei Liu
liuwei@mathstat.yorku.ca

[1] Department of Statistics and Finance, University of Science and Technology of China, Hefei 230026, China

[2] School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018, Zhejiang, China

[3] Acacdemy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[4] Department of Mathematics and Statistics, York University, Toronto M3J 1P3, Canada

perspective, model selection seems to be an inappropriate strategy since it selects one model and thus loses the useful information contained in the others. As discussed in Yuan and Yang (2005) and Zhang et al. (2012), unlike mode selection, model averaging incorporates the information contained in each candidate model by combining parameter estimates across the set of candidate models with appropriate weights and thus produces results that are more robust and potentially with smaller risk than that obtained by model selection. Over the past decade, model averaging has received a substantial amount of attention. Various model averaging procedures have been proposed. See, e.g., Buckland et al. (1997), Yang (2001), Yuan and Yang (2005), Hansen (2007), Wan et al. (2010), Hansen and Racine (2012), Liu and Okui (2013), Zhang et al. (2013), Zhang et al. (2016) among others. However, these methods are proposed with data observed completely and cannot be applied to the case of missing data directly. Therefore, in this paper, we consider model averaging in the presence of missing data.

Missing data occur commonly in market research surveys, socioeconomic investigations, medical studies and other scientific experiments. There are three missing data mechanisms: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) (Little and Rubin 2002). If the missingness does not depend on any variables, then the mechanism is called MCAR. If the missingness depends on observable variables, but not on missing variables, then the mechanism is called MAR. The mechanism is called NMAR if the missingness depends on the missing variables. MCAR is the simplest but also the most unrealistic. MAR is more complex and more reasonable in practice than MCAR. Although MAR is less natural than NMAR, it has been found to yield more accurate predictions of the missing values than methods based on NMAR mechanism in some empirical settings (Claeskens and Hjort 2008). Also, MAR can be explained reasonably in practice (Little and Rubin 2002). In the literature, most of research has focused on the case of MAR. In this article, we consider model averaging with MAR.

As far as we know, there is little work on the development of model averaging methods in the presence of missing data. Schomaker et al. (2010) considered two model average estimation approaches under the case where the covariates are missing at random, while the response values are fully observed. Their first approach modified the exponential AIC weight of Buckland et al. (1997) based on the weighted AIC (Hens et al. 2006) and then utilized the modified weights to combine estimates from different candidate models. The weighted AIC is a modification of the traditional AIC and is based on reweighting the complete observations by their inverse selection probabilities. Their second approach filled in the unobserved values by existing imputation techniques and then, based on the resultant complete data, averaged over estimates from different candidate models with conventional model average weights. Unfortunately, Schomaker et al. (2010) did not analyze the theoretical properties of their proposed model average estimates. A possible reason is that each candidate model is a parametric specification of the true model, and thus, the model average estimates under imputation could be too complicate to analyze. Later, Zhang (2013) considered a simpler situation where each candidate model is a linear regression model and the covariates are missing with missing mechanism MCAR. Under this simpler situation, Zhang (2013)

used Mallows model averaging (MMA) of Hansen (2007) by combining least squared estimates, derived by CC method, from different candidate models. And they showed that the resultant model average estimator is asymptotically optimal in the sense that its squared loss is asymptotically identical to that of the infeasible best possible model average estimator. Based on the focused information criterion (FIC) of Hjort and Claeskens (2003), Sun et al. (2014) developed a model average estimation scheme for linear regression models with responses missing at random under the local misspecification framework. The data generating process considered in their article is a linear regression model

$$Y = \mathrm{X}\beta + \epsilon, \tag{1}$$

where Y is a scalar response, $\mathrm{X} = (X_1, X_2, \ldots, X_p)$ is the covariable vector, β is the vector of unknown parameters and ϵ is the random error of the model. In their article, β is separated into two parts, $\beta = (\ddot{\beta}^\top, \bar{\beta}^\top)^\top$, where $\top$ represents transpose. $\ddot{\beta}$ is a $\ddot{p} \times 1$ vector corresponding to the covariates which are surely included in the true model, while $\bar{\beta}$ is a $(p - \ddot{p}) \times 1$ vector corresponding to the covariates which may be potentially included in the true model. Every candidate model is a linear model using all the covariates corresponding to $\ddot{\beta}$ and a subset of the covariates corresponding to $\bar{\beta}$. They considered a local misspecification framework in which the true value of β is $\beta_0 = (\ddot{\beta}_0^\top, \eta^\top/\sqrt{n})^\top$ where η is a $(p - \ddot{p}) \times 1$ vector. According to (1), it is easy to see that the largest model, including all the covariates, is actually the true model. Moreover, according to β_0, we know that all the candidate models get closer to the largest model as the sample size increases. Apparently, the local misspecification framework requires a great deal of knowledge about the true model and thus is somehow unrealistic due to the complex practice. Besides, this framework, introduced by Hjort and Claeskens (2003), is designed for facilitating the analysis of an estimator's asymptotic behavior, but not for prediction.

In this article, we consider model averaging for linear regression models with responses missing at random without the local misspecification framework. It is shown that our proposal is asymptotically optimal under certain conditions. The asymptotic optimality is an important theoretical property of model average estimators and has been studied widely in the field of model averaging. To the best of our knowledge, the current paper is the first work to develop model averaging approaches with missing responses without the local misspecification framework.

The presentation of this paper goes as follows. In Sect. 2, we describe the model framework and develop our model averaging procedure. In Sect. 3, we present asymptotic optimality of our proposed method. A simulation study is conducted in Sect. 4, and a real data analysis is given in Sect. 5. A discussion is made in Sect. 6, and all the technical details are given in "Appendix."

2 Model framework and estimation procedure

2.1 Model framework

Let $\{Y_i, \mathrm{X}_i\}_{i=1}^n$ be a random sample from (Y, X) where Y is a scalar response and X is the vector of covariates. Consider the following data generating process

$$Y_i = \mu_i + e_i, \quad \mu_i = E(Y_i|\mathrm{X}_i), \quad E(e_i|\mathrm{X}_i) = 0, \quad Var(e_i|\mathrm{X}_i) = \sigma^2. \tag{2}$$

A collection of linear regression models is considered. The number of these linear regression models is M_n. Under the mth model, we have

$$Y_i = \sum_{j=1}^{k_m} X_{ij(m)}\theta_{j(m)} + e_i, \tag{3}$$

where $X_{ij(m)}$ is the $j(m)$th element of X_i and $\theta_{j(m)}$ is the corresponding regression coefficient. In this article, we consider the case where some Y-values in a sample of size n are missing and all X-values are observed completely; that is, the data consist of incomplete observations $\{(\mathrm{X}_i, Y_i, \delta_i) : i = 1, 2, \ldots, n\}$ generated from model (2), where $\delta_i = 1$ if Y_i is observed, $\delta_i = 0$ otherwise. Throughout this paper, we assume that Y is missing at random; that is,

$$P(\delta = 1|Y, \mathrm{X}) = P(\delta = 1|\mathrm{X}) := \pi(\mathrm{X}), \tag{4}$$

and further, we assume that the selection probability function $\pi(\mathrm{X})$ is bounded away from zero. Our goal is to find an asymptotically optimal model average estimator for the conditional mean $\mu = (\mu_1, \mu_2, \ldots, \mu_n)^\top$ with responses missing at random.

2.2 Estimation procedure

First of all, we illustrate that μ can be estimated by the existing heteroscedasticity-robust C_p (HRC_p) approach of Liu and Okui (2013) which is developed without missing data. We first consider the case where the selection probability function $\pi(\mathrm{X})$ is known and postpone the discussion of estimation of $\pi(\mathrm{X})$ at the end of this section. Under this case, the dataset $\{(Z_{\pi,i}, \mathrm{X}_i) : i = 1, 2, \ldots, n\}$ is complete where $Z_{\pi,i} = \{\pi(\mathrm{X}_i)\}^{-1}\delta_i Y_i$, $i = 1, 2, \ldots, n$. By (2) and the MAR assumption, we have

$$\begin{aligned} &Z_{\pi,i} = \mu_i + e_{\pi,i}, \quad \mu_i = E(Z_{\pi,i}|\mathrm{X}_i), \quad E(e_{\pi,i}|\mathrm{X}_i) = 0, \\ &\mathrm{Var}(e_{\pi,i}|\mathrm{X}_i) = \sigma_{\pi,i}^2, \quad \sigma_{\pi,i}^2 = \left[\{\pi(\mathrm{X}_i)\}^{-1} - 1\right]\mu_i^2 + \{\pi(\mathrm{X}_i)\}^{-1}\sigma^2. \end{aligned} \tag{5}$$

And by (3), under the mth model, we have

$$Z_{\pi,i} = \sum_{j=1}^{k_m} X_{ij(m)}\theta_{j(m)} + e_{\pi,i}. \tag{6}$$

These results indicate that μ can be estimated by applying HRC_p method on $\{(Z_{\pi,i}, \mathrm{X}_i) : i = 1, 2, \ldots, n\}$ because HRC_p method is a model averaging method

developed for linear regression models with heteroscedastic errors in the absence of missing data.

In what follows, we present the details of applying HRC_p approach to get a model average estimator of μ. Let $Z_\pi = (Z_{\pi,1}, Z_{\pi,2}, \ldots, Z_{\pi,n})^\top$, $\boldsymbol{X}_{(m)}$ be an $n \times k_m$ matrix with ijth element $X_{ij(m)}$ and $e_\pi = (e_{\pi,1}, e_{\pi,2}, \ldots, e_{\pi,n})^\top$. It is assumed that $\boldsymbol{X}_{(m)}$ is of full column rank. The matrix form of (6) is $Z_\pi = \boldsymbol{X}_{(m)}\theta_{(m)} + e_\pi$, where $\theta_{(m)}$ is the vector of regression coefficients. Accordingly, the least squares estimator of $\theta_{(m)}$ is $\hat{\theta}_{(\pi,m)} = (\boldsymbol{X}_{(m)}^\top \boldsymbol{X}_{(m)})^{-1}\boldsymbol{X}_{(m)}^\top Z_\pi$. And the corresponding estimator of μ is

$$\hat{\mu}_{(\pi,m)} = \boldsymbol{X}_{(m)}\hat{\theta}_{(\pi,m)} = \mathbf{P}_{(m)}Z_\pi, \quad \mathbf{P}_{(m)} = \boldsymbol{X}_{(m)}(\boldsymbol{X}_{(m)}^\top \boldsymbol{X}_{(m)})^{-1}\boldsymbol{X}_{(m)}^\top. \tag{7}$$

The corresponding model average estimator of μ is

$$\hat{\mu}_\pi(\omega) = \sum_{m=1}^{M_n} \omega_m \hat{\mu}_{(\pi,m)} = \sum_{m=1}^{M_n} \omega_m \mathbf{P}_{(m)} Z_\pi = \mathbf{P}(\omega) Z_\pi, \tag{8}$$

where $\mathbf{P}(\omega) = \sum_{m=1}^{M_n} \omega_m \mathbf{P}_{(m)}$ and $\omega = (\omega_1, \omega_2, \ldots, \omega_{M_n})^\top$ is a weight vector in

$$\mathscr{H}_n = \left\{ \omega \in [0,1]^{M_n} : \sum_{m=1}^{M_n} \omega_m = 1 \right\}. \tag{9}$$

Taking advantage of HRC_p approach, we get the following model averaging criterion:

$$C_\pi(\omega) = \|Z_\pi - \hat{\mu}_\pi(\omega)\|^2 + 2\sum_{i=1}^{n} \hat{e}_{\pi,i}^2 \mathbf{P}_{ii}(\omega), \tag{10}$$

where $\hat{e}_{\pi,i}$ is a preliminary estimate of $e_{\pi,i}$ and $\mathbf{P}_{ii}(\omega)$ is the ith diagonal element of $\mathbf{P}(\omega)$. As what Liu and Okui (2013) recommended, we take $\hat{e}_\pi = (\hat{e}_{\pi,1}, \hat{e}_{\pi,2}, \ldots, \hat{e}_{\pi,n})^\top$ as

$$\hat{e}_\pi = \sqrt{\frac{n}{n-k_u}}(\mathbf{I}_n - \mathbf{P}_u)Z_\pi, \quad \mathbf{P}_u = \boldsymbol{X}_u(\boldsymbol{X}_u^\top \boldsymbol{X}_u)^- \boldsymbol{X}_u^\top, \tag{11}$$

where $\mathbf{I}_n$ is an $n \times n$ identity matrix, $(\boldsymbol{X}_u^\top \boldsymbol{X}_u)^-$ denotes a g-inverse of $\boldsymbol{X}_u^\top \boldsymbol{X}_u$, k_u is the rank of $\boldsymbol{X}_u$, $\boldsymbol{X}_u$ is the matrix whose ith row is the ith observation of X_u for $i = 1, 2, \ldots, n$, and X_u is the random vector consisting of the covariates that have been used in the candidate models.

At last, we discuss the estimation of the selection probability function $\pi(\mathrm{X})$. We assume a parametric model $\tilde{\pi}(\mathrm{X};\alpha)$ for $\pi(\mathrm{X})$ with α being the unknown model parameter vector. Denote $\hat{\alpha}_n$ as the maximum likelihood estimate (MLE) of α. And then write $\hat{\pi}(\mathrm{X}_i) = \tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)$ for $i = 1, 2, \ldots, n$. In what follows, a Greek letter subscripted by $\hat{\pi}$ represents that it is derived by replacing $\{\pi(\mathrm{X}_i) : i = 1, 2, \ldots, n\}$ in its corresponding estimator with $\{\hat{\pi}(\mathrm{X}_i) : i = 1, 2, \ldots, n\}$, respectively. For example, $Z_{\hat{\pi}}$ is derived by replacing $\{\pi(\mathrm{X}_i) : i = 1, 2, \ldots, n\}$ in Z_π with $\{\hat{\pi}(\mathrm{X}_i) : i = 1, 2, \ldots, n\}$, respectively. With $\{\hat{\pi}(\mathrm{X}_i) : i = 1, 2, \ldots, n\}$, $C_\pi(\omega)$ in (10) becomes

$$C_{\hat{\pi}}(\omega) = \|Z_{\hat{\pi}} - \hat{\mu}_{\hat{\pi}}(\omega)\|^2 + 2\sum_{i=1}^{n} \hat{e}_{\hat{\pi},i}^2 \mathbf{P}_{ii}(\omega). \tag{12}$$

Let $\hat{\omega}_c$ be the minimizer of $C_{\hat{\pi}}(\omega)$ among $\mathscr{H}_n$; that is,

$$\hat{\omega}_c = \arg\min_{\omega \in \mathscr{H}_n} C_{\hat{\pi}}(\omega). \tag{13}$$

Then the corresponding model average estimator of μ is $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_c)$.

3 Theoretical properties

In this section, we present the theoretical results of this paper which demonstrate the asymptotic optimality of $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_c)$. For ease of expression, let us first introduce some notations. Let $\mathscr{A}_x$ be the support sets of X and Θ_α be the parameter space of α. Let $\ell(\alpha) = E[\delta \log \tilde{\pi}(\mathrm{X};\alpha) + (1-\delta)\log\{1 - \tilde{\pi}(\mathrm{X};\alpha)\}]$. Define the loss function and the risk function of $\hat{\mu}_\pi(\omega)$ as

$$L_\pi(\omega) = \|\mu - \hat{\mu}_\pi(\omega)\|^2, \quad R_\pi(\omega) = E\{L_\pi(\omega)|\boldsymbol{X}\}, \tag{14}$$

respectively, where $\|\cdot\|$ is the Euclidean norm and $\boldsymbol{X}$ is the matrix whose ith row is X_i. Let $\xi_\pi = \inf_{\omega \in \mathscr{H}_n} R_\pi(\omega)$ and ω_m^0 be a $M_n \times 1$ vector whose mth element is one, while the other elements are zeros. The asymptotic optimality of $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_c)$ requires the following conditions where all the limiting processes discussed here and throughout the paper are with respect to $n \to \infty$.

(C1) Θ_α is bounded and closed. $\ell(\alpha)$ has a unique maximum at α_0 in Θ_α, and α_0 is an inner point of Θ_α. $\tilde{\pi}(\mathrm{X};\alpha)$ is twice continuously differentiable with respect to α. $E\left[\frac{\partial \tilde{\pi}(\mathrm{X};\alpha)}{\partial \alpha}\frac{\partial \tilde{\pi}(\mathrm{X};\alpha)}{\partial \alpha^T}|_{\alpha=\alpha_0}\right]$ is positive definite. Besides, $\tilde{\pi}(\mathrm{X};\alpha)$ is bounded away from 0. Interchange of difference and integration of $\tilde{\pi}(\mathrm{X};\alpha)$ is valid for first and second derivatives with respect to α. For all α's in a neighborhood of α_0, $\max_{1\le i\le n} \|\frac{\partial \tilde{\pi}(\mathrm{X}_i;\alpha)}{\partial \alpha}\| = O_p(1)$.

(C2) For some integer $1 \le G < \infty$, $\max_{1\le i\le n} E(e_i^{4G}|\mathrm{X}_i) \le C_e$, a.s. and $\max_{1\le i\le n} |\mu_i| \le C_\mu$, a.s., where C_μ and C_e are two constants.

(C3) For the integer G in (C2), $M_n \xi_\pi^{-2G} \sum_{m=1}^{M_n} \left\{R_\pi\left(\omega_m^o\right)\right\}^G \xrightarrow{\text{a.s.}} 0$.

(C4) $\max_{1\le m\le M_n} \max_{1\le i\le n} \mathbf{P}_{(m),ii} = O\left(n^{-1/2}\right)$, a.s. where $\mathbf{P}_{(m),ii}$ is the ith diagonal element of $\mathbf{P}_{(m)}$.

(C5) $n^{-1}k_u^2 \xrightarrow{\text{a.s.}} O(1)$, where k_u is the rank of $\boldsymbol{X}_u$.

(C6) $n\xi_\pi^{-2} \longrightarrow 0$.

The following theorem states the asymptotic optimality of $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_c)$.

Theorem 1 *If Conditions (C1)–(C6) are satisfied, then*

$$\frac{L_{\hat{\pi}}(\hat{\omega}_c)}{\inf_{\omega\in\mathscr{H}_n} L_{\hat{\pi}}(\omega)} \xrightarrow{p} 1. \tag{15}$$

Theorem 1 states that the selected weight vector, $\hat{\omega}_c$, yields a squared error that is asymptotically identical to that of the infeasible optimal weight vector. This implies the asymptotic optimality of $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_c)$.

Remark 1 In Condition (C1), the part before the last sentence is required for the consistency and asymptotic normality of the MLE $\hat{\alpha}_n$ and is obtained based on White (1982). The last sentence in Condition (C1) imposes some restrictions on $\partial\tilde{\pi}(\mathrm{X};\alpha)/\partial\alpha$ and is satisfied when the other part of Condition (C1) holds and $\mathscr{A}_x$ is bounded and closed.

Remark 2 Clearly, Condition (C2) is satisfied when $e_i \sim N(0,\sigma^2)$ and $E(Y|\mathrm{X})$ is bounded. Condition (C3) is actually Assumption 2.3 of Liu and Okui (2013) for the fully observed dataset $\{(Z_{\pi,i},\mathrm{X}_i) : i = 1,2,\ldots,n\}$. Such a condition is commonly used in the model averaging literature such as Wan et al. (2010), Zhang et al. (2013) and Gao et al. (2019). In particular, Wan et al. (2010) has explained Condition (C3) in detail and provided two explicit examples that Condition (C3) holds. Condition (C3) indicates that M_n is allowed to be fixed or go to infinity.

Remark 3 Condition (C4) is the same as Assumption 2.4 of Liu and Okui (2013). This condition imposes some restrictions on the element of $\boldsymbol{X}_{(m)}$. As what Liu and Okui (2013) pointed out, Condition (C4) excludes peculiar models, such as a model that contains a dummy variable on some single observation. Condition (C6) is similar to the third part of (A7) in Zhang et al. (2014) and Condition C.3 in Zhang et al. (2016). It can be shown that Condition (C6) holds if Condition (C3) holds and there exists an $m \in \{1,2,\ldots,M_n\}$ such that $n\{R_\pi(\omega_m^o)\}^{-1} = O(1)$, a.s.. In fact, Condition (C6) holds under Example 1 given in Wan et al. (2010).

Remark 4 Condition (C5) forbids the rank of $\boldsymbol{X}_u$ to grow faster than $n^{1/2}$. Similar conditions can be found in the existing literature, such as condition (12) in Wan et al. (2010) and condition (9) in Gao et al. (2019). Apparently, according to the definition of $\boldsymbol{X}_u$ given below (11), Condition (C5) indirectly requires that

$$\max_{1\le m\le M_n} nk_m^2 = O(1),$$

which indicates that the number of covariates used in each candidate models should be moderate. Moreover, a large $\max_{1\le m\le M_n} k_m$ always means a large M_n which gives rise to a heavy computation burden of our proposal. Fortunately, several dimension reduction methods, such as Ding and Wang (2011) and Wang and Li (2018), have been developed with missing response at random. Therefore, we suggest using one of these dimension reduction methods before implementing our proposal if $\max_{1\le m\le M_n} k_m$ is large. The theoretical properties as well as the finite sample performance related to this suggestion are not analyzed here since they are outside the

scope of this article. However, they are worthy of investigation, no doubt, and we leave them for future study.

4 A Monte Carlo study

In this section, a Monte Carlo study with two designs was conducted to investigate the finite sample performance of our proposed method. And for a better analysis, we also considered three intuitive methods as our proposal's competitors. The first one is the classical MMA approach of Hansen (2007) with CC analysis which just ignores all the individuals with missingness. And we termed this method as CC-MMA. The second one is the classical adaptive regression by mixing (ARM) method of Yang (2001) with CC analysis and is termed as CC-ARM. In order to implement CC-ARM, we need to specify a probability density function for e_i in (2). And we used the true probability density function of e_i in our simulation studies. Besides, CC-ARM needs to randomly permute the order of the observations several times. And we set the number of permutations for CC-ARM to be 100. The last one is the MMA approach with missing data replaced by imputed values and is termed as IM-MMA. If Y_i is missing, then its imputed value is given by

$$X_{ui}(\boldsymbol{X}_u^\top \boldsymbol{\Delta} \boldsymbol{X}_u)^- \boldsymbol{X}_u^\top \boldsymbol{\Delta} \mathrm{Y},$$

where X_{ui} is the ith row of $\boldsymbol{X}_u$ defined below (11), $\boldsymbol{\Delta}$ is an $n \times n$ diagonal matrix whose ith diagonal element is δ_i and $\mathrm{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$. Besides, we considered the infeasible complete data-based MMA approach as the "gold standard" to see how much loss of efficiency there is for a method in the presence of missing data. And we termed this method as CD-MMA. For easy of illustration, we termed our proposal as M-HRC_p. The details and results of the Monte Carlo study are given below.

Design 1 In this design, we considered the case where the number of candidate models, M_n, is fixed. The data generating process was

$$Y_i = \mu_i + e_i, \quad \mu_i = X_{i1}\theta_1 + X_{i2}^2\theta_2 + X_{i3}\theta_3 + X_{i4}\theta_4,$$

for $i = 1, 2, \ldots, n$, where $X_{i1} = 1$ is the intercept, while $\{X_{i2}, X_{i3}, X_{i4}\}$ are independent standard normal random variables, $\theta = (0.3, 0.6, 0.3, 0.3)^\top$, e_i is the random error generated from $N(0, \sigma^2)$ and the parameter σ^2 was determined by the population R^2. Following Hansen (2007), the population R^2 is defined as

$$R^2 = \frac{var(Y_i) - var(e_i)}{var(Y_i)} = \frac{2\theta_2^2 + \theta_3^2 + \theta_4^2}{2\theta_2^2 + \theta_3^2 + \theta_4^2 + \sigma^2}. \tag{16}$$

The parameter σ^2 was chosen to let the population R^2 vary on a grid between 0.1 and 0.9. The selection probability function was

$$\pi(\mathrm{X}_i) = \Phi(X_{i1}\alpha_1 + X_{i2}\alpha_2), \tag{17}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The following two settings of $\alpha = (\alpha_1, \alpha_2)^\top$ were taken into consideration,

$$\text{Case 1} : \alpha = (0.6, 0.5)^\top, \quad \text{Case 2} : \alpha = (0, 0.5)^\top.$$

The corresponding average missing rates are approximated 30% and 50%, respectively. All the candidate models include the intercept term and were constructed by varying combinations of $\{X_2, X_3, X_4\}$. As a result, $M_n = 2^3$. The sample size were taken to be $n = 100$ and $n = 200$, respectively. The parametric model $\tilde{\pi}(\mathrm{X};\alpha)$ used in our proposal was taken to be (17). Following Hansen (2007), we used the following mean squared error (MSE) to assess the finite sample performances of estimators:

$$\frac{1}{D}\sum_{d=1}^{D} \|\{\hat{\mu}(\omega)\}^{(d)} - \mu^{(d)}\|^2,$$

where $\mu^{(d)}$ is the conditional mean in the dth trial, $\{\hat{\mu}(\omega)\}^{(d)}$ is the model average estimator of $\mu^{(d)}$ and D is the number of simulation trials. For better comparison, we reported the ratios of the MSEs that are computed with the MSE of the infeasible CD-MMA approach as the denominator. The number of simulation trials was 1000.

Figure 1 plots the MSE ratio against the population R^2 for a variety of combinations of sample sizes and missing rates. The dotted, dashed, dash dotted and solid lines correspond to the curves of CC-MMA, CC-ARM, IM-MMA and M-HRC_p, respectively. Figure 1 shows that our proposal, M-HRC_p, achieves a lower MSE ratio than CC-MMA and IM-MMA for all the combinations of sample sizes, missing rates and the population R^2 considered. This is consistent with our expectation since

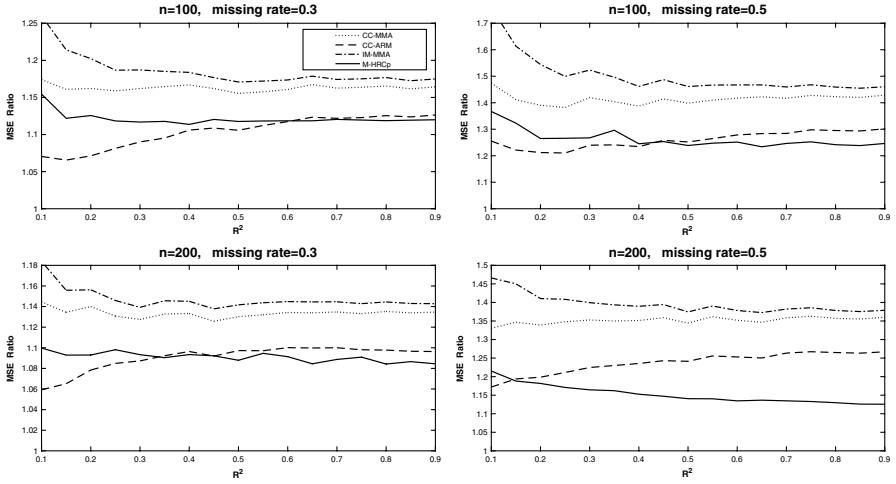

Fig. 1 The dotted, dashed, dash dotted and solid lines are the MSE ratio curves of CC-MMA, CC-ARM, IM-MMA and the proposed M-HRC_p, respectively. M_n is fixed

the asymptotically optimality of the classical MMA approach fails when CC analysis or IM-MMA is used. Comparing to CC-ARM, our proposed M-HRC_p has a better performance in the majority of the combinations of sample sizes, missing rates and the population R^2 considered, especially when R^2 is moderate or large.

Following a referee's suggestion, we took the strategy taken by Zhang et al. (2016) to numerically demonstrate Theorem 1. Concretely, we calculated the means of

$$\text{LR} = \frac{L_{\hat{\pi}}(\hat{\omega}_c)}{\inf_{\omega \in \mathscr{H}_n} L_{\hat{\pi}}(\omega)}, \tag{18}$$

based on 1000 replications under different sample sizes with $R^2 = 0.5$. And the mean curves are displayed in Fig. 2.

As shown in Fig. 2, the mean of LR decreases and approaches to 1 as the sample size increases for the two missing rates considered. Such a result confirms the asymptotic optimality of $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_c)$ stated in Theorem 1.

Design 2 In this design, we considered the case where M_n is allowed to grow with n. The data generating process was

$$\tilde{Y}_i = \mu_i + \tilde{\mu}_i + e_i, \quad \tilde{\mu}_i = c \cdot \sum_{j=1}^{J} j^{-2} \tilde{X}_{ij},$$

where μ_i and e_i are the same as that in Design 1, $\tilde{X}_{ij}$'s are independent standard normal random variables, $J = 10^3$ and $c = 0.25$. The population R^2 is now equal to

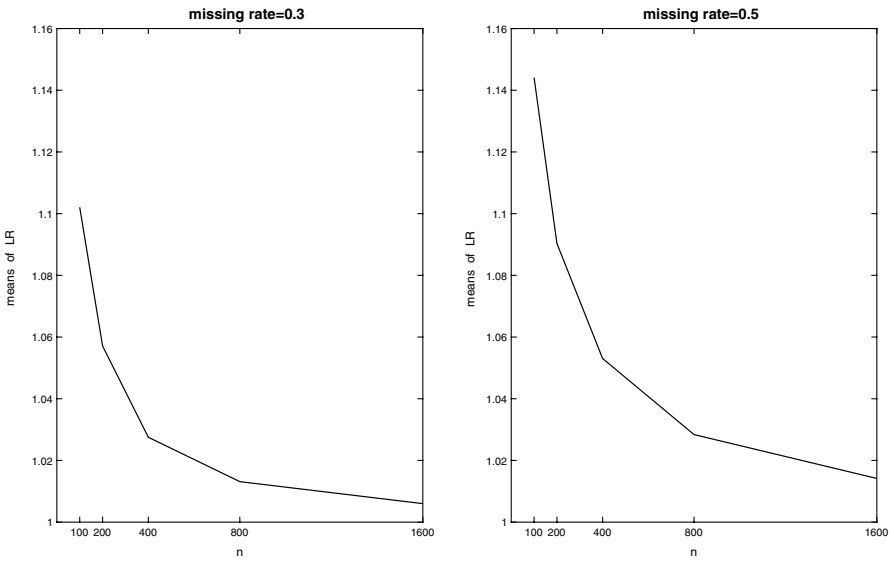

Fig. 2 Assessing the asymptotic optimality of $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_c)$. M_n is fixed

$$\frac{2\theta_2^2+\theta_3^2+\theta_4^2+c^2\sum_{j=1}^{J}j^{-4}}{2\theta_2^2+\theta_3^2+\theta_4^2+c^2\sum_{j=1}^{J}j^{-4}+\sigma^2}.$$

And the parameter σ^2 was also chosen to let R^2 vary on a grid between 0.1 and 0.9. The selection probability function as well as its parametric model assumption was the same as that in Design 1. The candidate models are strictly nested with the mth linear regression model using the first m covariates in $\{X_{i1},\ldots,X_{i4},\tilde{X}_{i1},\tilde{X}_{i2},\ldots,\tilde{X}_{iJ}\}$. And the number of candidate models, M_n, was set to be the nearest integer from $1.5n^{1/2}$. The sample size was considered to be $n=100,200,400$ and 800, so that $M_n=15,21,30$ and 42, respectively. The results are shown in Fig. 3.

The results displayed in Fig. 3 show a similar pattern to that in Fig. 1. Figure 3 shows that our proposed M-HRC_p performs better than CC-MMA and IM-MMA for all the combinations of sample sizes and missing rates considered. Moreover, the

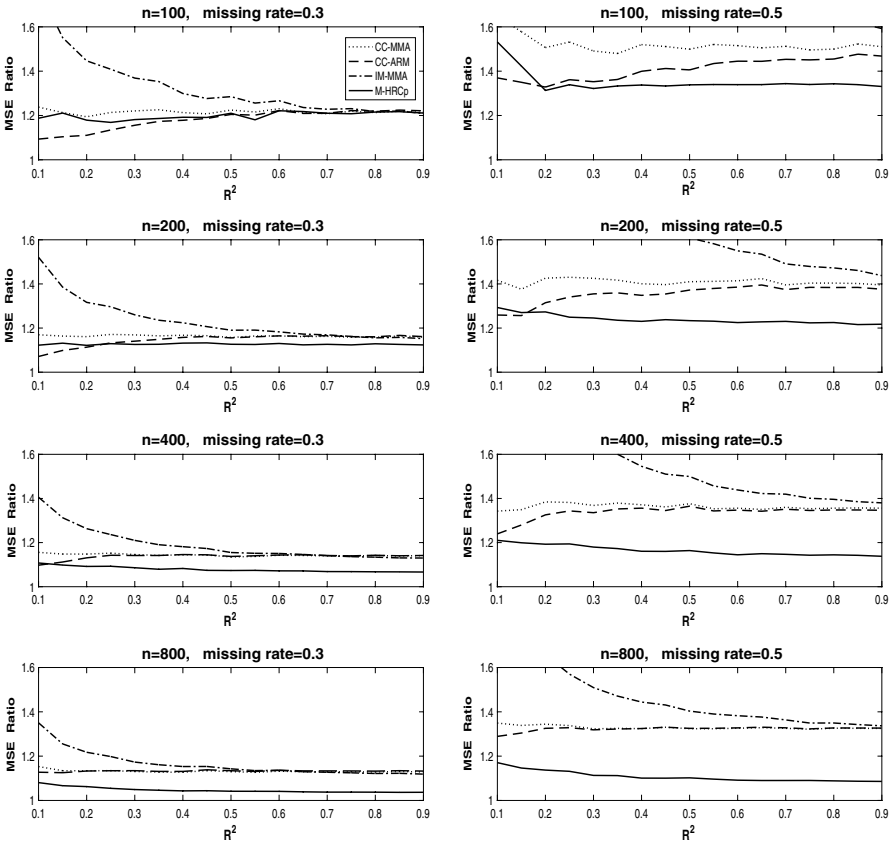

Fig. 3 The dotted, dashed, dash dotted and solid lines are the MSE ratio curves of CC-MMA, CC-ARM, IM-MMA and the proposed M-HRC_p, respectively. M_n grows with n

performance of our proposal is better than that of CC-ARM in most of the cases. What else can be seen is that the superiority of our proposal over its competitors gets more prominent as the sample size or the missing rate increases.

For the purpose of assessing the result stated in Theorem 1 when M_n diverges to infinity, we also calculated the means of LR in (18) based on 1000 replications under different sample sizes with $R^2 = 0.5$. The results are shown in Fig. 4. It is seen that the mean of LR decreases and gets closer to 1 as n increases for the two missing rates considered. This numerically confirms Theorem 1.

5 Real data analysis

In this section, a real data analysis is conducted to analyze the practical performance of our proposal. We analyzed the PM2.5 data taken at Beijing Olympic Sports Center in July 2015. This dataset can be obtained from the Machine Learning Repository at the University of California Irvine (https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data). We took the PM2.5 concentration as the response variable Y and the following attributes as covariates: PM10 concentration (X_1), CO concentration (X_2), dew point temperature (X_3), temperature (X_4), SO2 concentration (X_5), NO2 concentration (X_6), O3 concentration (X_7), precipitation (X_8), wind speed (X_9) and pressure (X_{10}). The original data contain 744 sample points of which 35 are missing either in response or covariates. For the consideration of simplicity and the fact that 35 is much smaller than 744, we removed these 35 sample points. Besides, in order to eliminate the influence of the scale, we

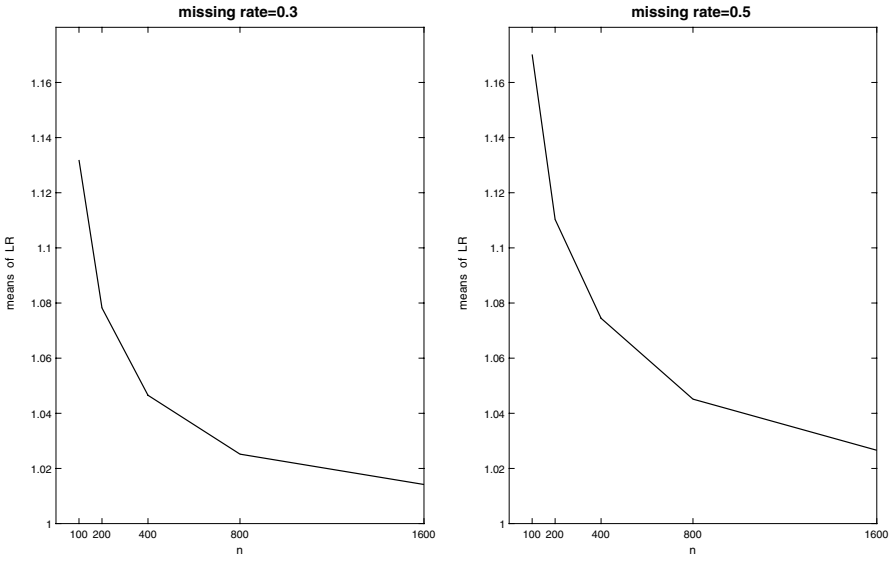

Fig. 4 Assessing the asymptotic optimality of $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_c)$. M_n grows with n

centralized and standardized each covariate by its sample mean and sample standard error, respectively.

Note that the number of all possible candidate linear models is $2^{10} - 1$ which is large and thus time-consuming. To reduce the computation burden, we considered the forward/backward (FW/BW) procedure descried in Jiang et al. (2015) to generate a sequence of candidate linear models. We randomly chose n_1 sample points as training data and then took the remaining sample points as test data. Note that the sample points we used are observed completely. Therefore, in order to apply our proposal, we selected some sample points from training data and treated their observations in Y as missing values. Such a selection was conducted according to the following MAR assumption:

$$P(\delta = 1|Y, \mathrm{X}) = \Phi(0.5X_1 + 0.5X_2 + 0.5X_3),$$

where $\mathrm{X} = (X_1, X_2, \ldots, X_{10})$. The corresponding average missing rates are approximated 50%. Similar to Zhu et al. (2019), we took the following normalized mean squared prediction error (NMSPE) as our performance metric:

$$\mathrm{NMSPE} = \frac{\sum_{i=n_1+1}^{n} \left\{Y_i - \hat{\mu}_i(\hat{\omega})\right\}^2}{\sum_{i=n_1+1}^{n} \left\{Y_i - \hat{\mu}_i(\hat{\omega}_{cd})\right\}^2},$$

where $\hat{\mu}_i(\hat{\omega})$ is obtained by a monitored method, while $\hat{\mu}_i(\hat{\omega}_{cd})$ is obtained by the "gold standard" CD-MMA method. We considered n_1 to be 200, 400 and 600. Considering the simulation results presented in Sect. 4, in this section, we only took CC-MMA and CC-ARM as our proposal's competitors. The parametric model $\tilde{\pi}(\mathrm{X};\alpha)$ used in our proposal was taken to be the true model. And the probability density function of e_i used in CC-ARM was taken to be a normal density function. Besides, the number of permutations for CC-ARM was set to be 100. The results are shown in Table 1.

Table 1 Mean, median and standard deviation (SD) of NMSPE based on 500 replications

Method	M-HRC_p	CC-MMA	CC-ARM
$n_1 = 200$			
Mean	1.2094	1.3138	1.2215
Median	1.157	1.2249	1.1893
SD	0.2509	0.3618	0.2035
$n_1 = 400$			
Mean	1.1354	1.3175	1.2825
Median	1.1007	1.2784	1.2547
SD	0.1461	0.2065	0.1706
$n_1 = 600$			
Mean	1.1115	1.3164	1.2929
Median	1.0882	1.2744	1.2602
SD	0.1352	0.2271	0.2099

Table 1 shows the mean, median and standard deviation (SD) of NMSPE based on 500 replications. From this table, we can see that our proposed M-HRC_p achieves a lower mean as well as a lower median of NMSPE than its competitors for all considered sample sizes. Moreover, both mean and median of NMSPE of our proposal get closer to 1 as n_1 increases, while this is not the case for CC-MMA and CC-ARM. The results indicate that our proposal is preferable than the two intuitive methods, CC-MMA and CC-ARM, for handling the problem considered in this paper.

6 Discussion

In this paper, a model average method is proposed for linear models with responses missing at random. It is shown that our proposed method is asymptotically optimal in the sense of achieving the lowest possible squared error. The simulation results favor our method comparing with three intuitive methods: CC-MMA, CC-ARM and IM-MMA.

We have focused on the linear regression model in this article. Apparently, extending the idea of our proposal to more complex models is meaningful and thus warrants future researches. Note that the ARM method of Yang (2001) is capable of combining estimates from different models. And these models can be linear regression models, generalized linear regression models, additive models and so on. Therefore, using the idea of the ARM method may be a successful way to generalize the application of our proposal to more complex models. However, it is a very challenging research topic and needs further investigation.

In this paper, the missing data mechanism is assumed to be MAR. As we mentioned in Sect. 1, NMAR is a more natural and more complex missing data mechanism than MAR. To the best of our knowledge, there is no work in the field of model averaging with data not missing at random. It is an interesting but challenging topic to develop mode averaging method for linear models with responses not missing at random in the future.

It should be pointed out that our proposal requires the parametric model $\tilde{\pi}(\mathrm{X};\alpha)$ to be correctly specified. This arouses the interest of a referee in the question of how the misspecification of the selection probability function affects the performance of our proposal. To address this question, we have investigated the performance of our proposal with $\tilde{\pi}(\mathrm{X};\alpha)$ being misspecified through a simulation study. Specific details of this simulation study are available upon request from the authors. The simulation results indicate that our proposal still performs well if $\tilde{\pi}(\mathrm{X};\alpha)$ deviates slightly from the true model, while its power fails if $\tilde{\pi}(\mathrm{X};\alpha)$ deviates greatly from the true mode. This, to some extent, gives the readers some senses of the aforementioned question from an empirical standpoint. Addressing the aforementioned question from a theoretical standpoint is clearly very difficult and needs further investigation. However, we think it is more meaningful to put our effort into developing a model average method that is robust against the misspecification of the selection probability function. Here, "robust" means that the model average method is asymptotically optimal even if $\tilde{\pi}(\mathrm{X};\alpha)$ is misspecified.

Acknowledgements Wang's research was supported by the National Natural Science Foundation of China (General program 11871460, Key program 11331011 and program for Creative Research Group in China 61621003), a grant from the Key Lab of Random Complex Structure and Data Science, CAS.

Appendix

We use c to denote a generic positive constant that could take different values in different occasions. Before presenting the proof of Theorem 1, let us first present a lemma which is required for the proof of Theorem 1.

Lemma 1 *Provided that Conditions (C1) and (C2) hold, we have*

$$\|Z_{\hat{\pi}} - Z_{\pi}\|^2 = O_p(1). \tag{19}$$

Proof of Lemma 1 According to the definition of Z_{π} given at the beginning of Section 2.2 and Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}\|Z_{\hat{\pi}} - Z_{\pi}\|^2 &= \sum_{i=1}^{n} \left\{ \frac{\delta_i}{\hat{\pi}(\mathrm{X}_i)} - \frac{\delta_i}{\pi(\mathrm{X}_i)} \right\}^2 \cdot Y_i^2 \le \sum_{i=1}^{n} \left\{ \frac{1}{\hat{\pi}(\mathrm{X}_i)} - \frac{1}{\pi(\mathrm{X}_i)} \right\}^2 \cdot (\mu_i + e_i)^2 \\ &\le \left\{ \sqrt{n} \cdot \max_{1\le i\le n} \left| \frac{1}{\hat{\pi}(\mathrm{X}_i)} - \frac{1}{\pi(\mathrm{X}_i)} \right| \right\}^2 \cdot c \cdot \left(\frac{1}{n}\mu^{\top}\mu + \frac{1}{n}e^{\top}e \right).\end{aligned}$$

By (C2), it is easy to see that $n^{-1}\mu^{\top}\mu = O_p(1)$. And by the law of larger numbers, $n^{-1}e^{\top}e = O_p(1)$. These results imply that (19) holds if the following holds:

$$\sqrt{n} \cdot \max_{1\le i\le n} \left| \frac{1}{\hat{\pi}(\mathrm{X}_i)} - \frac{1}{\pi(\mathrm{X}_i)} \right| = O_p(1). \tag{20}$$

In what follows, we present the proof of (20) which completes the proof of Lemma 1.

Recalling that $\hat{\pi}(\mathrm{X}) = \tilde{\pi}(\mathrm{X};\hat{\alpha}_n)$, we apply Taylor expansion to $\{\hat{\pi}(\mathrm{X}_i)\}^{-1}$ around the true value α_0 and then obtain

$$\begin{aligned}&\sqrt{n} \cdot \max_{1\le i\le n} \left| \frac{1}{\hat{\pi}(\mathrm{X}_i)} - \frac{1}{\pi(\mathrm{X}_i)} \right| \\ &= \sqrt{n} \cdot \max_{1\le i\le n} \left| \left\{ \frac{1}{\{\tilde{\pi}(\mathrm{X}_i;\alpha)\}^2} \cdot \frac{\partial \tilde{\pi}(\mathrm{X}_i;\alpha)}{\partial \alpha^{\top}} \right\}\Big|_{\alpha=\alpha_{n,\mathrm{X}_i}} \cdot (\hat{\alpha}_n - \alpha_0) \right| \\ &\le c \cdot \max_{1\le i\le n} \left\| \frac{\partial \tilde{\pi}(\mathrm{X}_i;\alpha)}{\partial \alpha}\Big|_{\alpha=\alpha_{n,\mathrm{X}_i}} \right\| \cdot \sqrt{n}\|\hat{\alpha}_n - \alpha_0\|,\end{aligned}$$

where the last inequality is due to (C1) and Cauchy–Schwarz inequality, and $\alpha_{(n,\mathrm{X}_i)}$ is a vector between $\hat{\alpha}_n$ and α_0. Since $\hat{\alpha}_n$ is MLE, by (C1) and a standard argument we have $\sqrt{n}\|\hat{\alpha}_n - \alpha_0\| = O_p(1)$. Because of the consistency of $\hat{\alpha}_n$ and (C1), we have $\max_{1\le i\le n} \|\frac{\partial \tilde{\pi}(\mathrm{X}_i;\alpha)}{\partial \alpha}|_{\alpha=\alpha_{n,\mathrm{X}_i}}\| = O_p(1)$. These results imply (20). □

Proof of Theorem 1 Let $\bar{\mathbf{P}}(\omega)$ be an $n \times n$ diagonal matrix whose ith diagonal element is $\mathbf{P}_{ii}(\omega)$, the ith diagonal element of $\mathbf{P}(\omega)$ in (8). Recalling $C_{\hat{\pi}}(\omega)$ in (12), after some careful calculations, we have $C_{\hat{\pi}}(\omega) = L_{\hat{\pi}}(\omega) + 2a_n(\omega) + \|Z_{\hat{\pi}} - \mu\|^2$ where

$$a_n(\omega) = (Z_{\hat{\pi}} - Z_\pi)^\top \{\mu - \hat{\mu}_{\hat{\pi}}(\omega)\} + e_\pi^\top \{\mu - \hat{\mu}_{\hat{\pi}}(\omega)\} + \hat{e}_{\hat{\pi}}^\top \bar{\mathbf{P}}(\omega)\hat{e}_{\hat{\pi}}. \tag{21}$$

Thus, by (13), it is readily seen that

$$\hat{\omega}_c = \arg\min_{\omega \in \mathscr{H}_n} \{L_{\hat{\pi}}(\omega) + 2a_n(\omega)\}.$$

Accordingly, from the proof of Theorem $1'$ in Wan et al. (2010), Theorem 1 is valid if the following holds:

$$\sup_{\omega \in \mathscr{H}_n} \left| \frac{L_{\hat{\pi}}(\omega)}{R_\pi(\omega)} - 1 \right| = o_p(1), \tag{22}$$

$$\sup_{\omega \in \mathscr{H}_n} \left| \frac{a_n(\omega)}{R_\pi(\omega)} \right| = o_p(1), \tag{23}$$

where $R_\pi(\omega)$ is the risk function defined in (14). By (8), (14) and Cauchy–Schwarz inequality, we have

$$\begin{aligned}
&\left| \frac{L_{\hat{\pi}}(\omega)}{R_\pi(\omega)} - 1 \right| \\
&\quad = \left| \frac{\|\mu - \hat{\mu}_{\hat{\pi}}(\omega)\|^2}{R_\pi(\omega)} - 1 \right| = \left| \frac{\|\mu - \hat{\mu}_\pi(\omega) + \hat{\mu}_\pi(\omega) - \hat{\mu}_{\hat{\pi}}(\omega)\|^2}{R_\pi(\omega)} - 1 \right| \\
&\quad \le \left| \frac{L_\pi(\omega)}{R_\pi(\omega)} - 1 \right| + \frac{2\{L_\pi(\omega)\}^{1/2} \cdot \|\hat{\mu}_\pi(\omega) - \hat{\mu}_{\hat{\pi}}(\omega)\|}{R_\pi(\omega)} + \frac{\|\hat{\mu}_\pi(\omega) - \hat{\mu}_{\hat{\pi}}(\omega)\|^2}{R_\pi(\omega)}, \\
&\|\hat{\mu}_\pi(\omega) - \hat{\mu}_{\hat{\pi}}(\omega)\|^2 = \|\mathbf{P}(\omega)Z_\pi - \mathbf{P}(\omega)Z_{\hat{\pi}}\|^2 \le \|Z_{\hat{\pi}} - Z_\pi\|^2.
\end{aligned}$$

Accordingly, to prove (22), it suffices to prove

$$\sup_{\omega \in \mathscr{H}_n} \left| \frac{L_\pi(\omega)}{R_\pi(\omega)} - 1 \right| = o_p(1), \tag{24}$$

$$\sup_{\omega \in \mathscr{H}_n} \frac{\|Z_{\hat{\pi}} - Z_\pi\|^2}{R_\pi(\omega)} = o_p(1). \tag{25}$$

Let $\tilde{p} = \sup_{\omega \in \mathscr{H}_n} \max_{1 \le i \le n} \mathbf{P}_{ii}(\omega)$. Then it is easy to verify that

$$\tilde{p} = O_p\left(\frac{1}{\sqrt{n}}\right), \tag{26}$$

by (C4). By (8), (11) and Cauchy–Schwarz inequality, we have

$$
\begin{aligned}
&\left|e_\pi^\top\{\mu-\hat{\mu}_{\hat{\pi}}(\omega)\}+\hat{e}_{\hat{\pi}}^\top\bar{\mathbf{P}}(\omega)\hat{e}_{\hat{\pi}}\right| \\
&\quad=\left|e_\pi^\top\{\mu-\hat{\mu}_\pi(\omega)+\hat{\mu}_\pi(\omega)-\hat{\mu}_{\hat{\pi}}(\omega)\}+(\hat{e}_{\hat{\pi}}-\hat{e}_\pi+\hat{e}_\pi)^\top\bar{\mathbf{P}}(\omega)(\hat{e}_{\hat{\pi}}-\hat{e}_\pi+\hat{e}_\pi)\right| \\
&\quad=|e_\pi^\top\{\mu-\mathbf{P}(\omega)\mu-\mathbf{P}(\omega)e_\pi\}+e_\pi^\top\{\hat{\mu}_\pi(\omega)-\hat{\mu}_{\hat{\pi}}(\omega)\} \\
&\qquad+(\hat{e}_{\hat{\pi}}-\hat{e}_\pi)^\top\bar{\mathbf{P}}(\omega)(\hat{e}_{\hat{\pi}}-\hat{e}_\pi)+2(\hat{e}_{\hat{\pi}}-\hat{e}_\pi)^\top\bar{\mathbf{P}}(\omega)\hat{e}_\pi+\hat{e}_\pi^\top\bar{\mathbf{P}}(\omega)\hat{e}_\pi| \\
&\quad\le|e_\pi^\top\mathbf{A}(\omega)\mu|+|e_\pi^\top\mathbf{P}(\omega)e_\pi-\mathrm{tr}\{\mathbf{\Omega}_\pi\mathbf{P}(\omega)\}|+|e_\pi^\top\mathbf{P}(\omega)(Z_{\hat{\pi}}-Z_\pi)| \\
&\qquad+\frac{n}{n-k_u}\cdot(Z_{\hat{\pi}}-Z_\pi)^\top(\mathbf{I}_n-\mathbf{P}_u)\bar{\mathbf{P}}(\omega)(\mathbf{I}_n-\mathbf{P}_u)(Z_{\hat{\pi}}-Z_\pi) \\
&\qquad+\frac{2n}{n-k_u}\cdot\left|(Z_{\hat{\pi}}-Z_\pi)^\top(\mathbf{I}_n-\mathbf{P}_u)\bar{\mathbf{P}}(\omega)(\mathbf{I}_n-\mathbf{P}_u)Z_\pi\right| \\
&\qquad+|\hat{e}_\pi^\top\bar{\mathbf{P}}(\omega)\hat{e}_\pi-\mathrm{tr}\{\mathbf{\Omega}_\pi\mathbf{P}(\omega)\}| \\
&\quad\le|e_\pi^\top\mathbf{A}(\omega)\mu|+|e_\pi^\top\mathbf{P}(\omega)e_\pi-\mathrm{tr}\{\mathbf{\Omega}_\pi\mathbf{P}(\omega)\}|+\|\mathbf{P}(\omega)e_\pi\|\cdot\|Z_{\hat{\pi}}-Z_\pi\| \\
&\qquad+\frac{n}{n-k_u}\cdot\tilde{p}\cdot\|Z_{\hat{\pi}}-Z_\pi\|^2+\frac{2n}{n-k_u}\cdot\tilde{p}\cdot\|Z_{\hat{\pi}}-Z_\pi\|\cdot\|Z_\pi\| \\
&\qquad+|\hat{e}_\pi^\top\bar{\mathbf{P}}(\omega)\hat{e}_\pi-\mathrm{tr}\{\mathbf{\Omega}_\pi\mathbf{P}(\omega)\}|,
\end{aligned}
$$

where $\mathbf{A}(\omega)=\mathbf{I}_n-\mathbf{P}(\omega)$ and $\mathbf{\Omega}_\pi$ is an $n\times n$ diagonal matrix whose ith diagonal element is $\sigma_{\pi,i}^2$ in (5). This together with (21), (C5), Lemma 1, (26) and Cauchy–Schwarz inequality proves that (23) holds if (24), (25) and the following hold:

$$\sup_{\omega\in\mathscr{H}_n}\frac{|e_\pi^\top\mathbf{A}(\omega)\mu|}{R_\pi(\omega)}=o_p(1), \tag{27}$$

$$\sup_{\omega\in\mathscr{H}_n}\frac{|e_\pi^\top\mathbf{P}(\omega)e_\pi-\mathrm{tr}\{\mathbf{\Omega}_\pi\mathbf{P}(\omega)\}|}{R_\pi(\omega)}=o_p(1), \tag{28}$$

$$\sup_{\omega\in\mathscr{H}_n}\frac{|\hat{e}_\pi^\top\bar{\mathbf{P}}(\omega)\hat{e}_\pi-\mathrm{tr}\{\mathbf{\Omega}_\pi\mathbf{P}(\omega)\}|}{R_\pi(\omega)}=o_p(1), \tag{29}$$

$$\sup_{\omega\in\mathscr{H}_n}\frac{\|\mathbf{P}(\omega)e_\pi\|}{R_\pi(\omega)}=o_p(1), \tag{30}$$

$$\sup_{\omega\in\mathscr{H}_n}\frac{\tilde{p}\cdot\|Z_\pi\|}{R_\pi(\omega)}=o_p(1). \tag{31}$$

Under Conditions (C1)–(C5), it can be shown that the assumptions of Theorem 2.2 of Liu and Okui (2013) are satisfied for the dataset $\{(Z_{\pi,i},\mathrm{X}_i):i=1,2,\ldots,n\}$. Therefore, according to the proof of Theorem 2.2 of Liu and Okui (2013), we know that (24) and (27)–(29) are satisfied. In what follows, we present the proofs of (25), (30) and (31) which complete the proof of Theorem 1.

By (5), (C1), (C2), (C6), (26), Lemma 1, the law of larger numbers and Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\sup_{\omega\in\mathscr{H}_n}\frac{\|Z_{\hat{\pi}}-Z_{\pi}\|^2}{R_{\pi}(\omega)} &\le \xi_{\pi}^{-1}\cdot\|Z_{\hat{\pi}}-Z_{\pi}\|^2 = o_p(1),\\
\sup_{\omega\in\mathscr{H}_n}\frac{\|\mathbf{P}(\omega)e_{\pi}\|}{R_{\pi}(\omega)} &\le \xi_{\pi}^{-1}\|e_{\pi}\| = \left\{n\xi_{\pi}^{-2}\cdot\frac{1}{n}\|e_{\pi}\|^2\right\}^{1/2}\\
&\le c\cdot\left[n\xi_{\pi}^{-2}\cdot\left\{\frac{1}{n}\mu^{\top}\mu+\frac{1}{n}e^{\top}e\right\}\right]^{1/2} = o_p(1),\\
\sup_{\omega\in\mathscr{H}_n}\frac{\tilde{p}\cdot\|Z_{\pi}\|}{R_{\pi}(\omega)} &\le c\cdot\xi_{\pi}^{-1}\tilde{p}\cdot\{\mu^{\top}\mu+e^{\top}e\}\\
&= c\cdot\sqrt{n}\xi_{\pi}^{-1}\cdot\sqrt{n\tilde{p}}\cdot\left\{\frac{1}{n}\mu^{\top}\mu+\frac{1}{n}e^{\top}e\right\} = o_p(1).
\end{aligned}$$

These results indicate (25), (30) and (31). □

References

Buckland, S. T., Burnham, K. P., Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, *53*(2), 603–618.

Claeskens, G., Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge: Cambridge University Press.

Ding, X., Wang, Q. (2011). Fusion-refinement procedure for dimension reduction with missing response at random. *Journal of the American Statistical Association*, *106*(495), 1193–1207.

Gao, Y., Zhang, X., Wang, S., Chong, T. T.-L., Zou, G. (2019). Frequentist model averaging for threshold models. *Annals of the Institute of Statistical Mathematics*, *71*(2), 275–306.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, *75*(4), 1175–1189.

Hansen, B. E., Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, *167*(1), 38–46.

Hens, N., Aerts, M., Molenberghs, G. (2006). Model selection for incomplete and design-based samples. *Statistics in Medicine*, *25*(14), 2502–2520.

Hjort, N. L., Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, *98*(464), 879–899.

Jiang, J., Nguyen, T., Rao, J. S. (2015). The E-MS algorithm: Model selection with incomplete data. *Journal of the American Statistical Association*, *110*(511), 1136–1147.

Little, R. J . A., Rubin, D . B. (2002). *Statistical analysis with missing data*, 2nd ed. Hoboken, NJ: Wiley.

Liu, Q., Okui, R. (2013). Heteroscedasticity-robust Cp model averaging. *The Econometrics Journal*, *16*(3), 463–472.

Schomaker, M., Wan, A. T., Heumann, C. (2010). Frequentist model averaging with missing observations. *Computational Statistics and Data Analysis*, *54*(12), 3336–3347.

Sun, Z., Su, Z., Ma, J. (2014). Focused vector information criterion model selection and model averaging regression with missing response. *Metrika*, *77*(3), 415–432.

Wan, A. T., Zhang, X., Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics*, *156*(2), 277–283.

Wang, Q., Li, Y. (2018). How to make model-free feature screening approaches for full data applicable to the case of missing response? *Scandinavian Journal of Statistics*, *45*(2), 324–346.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*: *Journal of the Econometric Society*, *50*(1), 1–25.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association*, *96*(454), 574–588.

Yuan, Z., Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association*, *100*(472), 1202–1214.

Zhang, X. (2013). Model averaging with covariates that are missing completely at random. *Economics Letters*, *121*(3), 360–363.

Zhang, X., Wan, A. T., Zhou, S. Z. (2012). Focused information criteria, model selection, and model averaging in a Tobit model with a nonzero threshold. *Journal of Business and Economic Statistics*, *30*(1), 132–142.

Zhang, X., Wan, A. T., Zou, G. (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics*, *174*(2), 82–94.

Zhang, X., Zou, G., Liang, H. (2014). Model averaging and weight choice in linear mixed-effects models. *Biometrika*, *101*(1), 205–218.

Zhang, X., Yu, D., Zou, G., Liang, H. (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association*, *111*(516), 1775–1790.

Zhu, R., Wan, A. T., Zhang, X., Zou, G. (2019). A Mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association*, *114*(526), 882–892.