mathematics

MDPI

Article

Model Averaging for Accelerated Failure Time Models with Missing Censoring Indicators

Longbiao Liao and Jinghao Liu *

Department of Statistics and Data Science, School of Economics, Jinan University, Guangzhou 510632, China; lbl023@stu2022.jnu.edu.cn

* Correspondence: jinghao123@stu2021.jnu.edu.cn

Abstract: Model averaging has become a crucial statistical methodology, especially in situations where numerous models vie to elucidate a phenomenon. Over the past two decades, there has been substantial advancement in the theory of model averaging. However, a gap remains in the field regarding model averaging in the presence of missing censoring indicators. Therefore, in this paper, we present a new model-averaging method for accelerated failure time models with right censored data when censoring indicators are missing. The model-averaging weights are determined by minimizing the Mallows criterion. Under mild conditions, the calculated weights exhibit asymptotic optimality, leading to the model-averaging estimator achieving the lowest squared error asymptotically. Monte Carlo simulations demonstrate that the method proposed in this paper has lower mean squared errors compared to other model-selection and model-averaging methods. Finally, we conducted an empirical analysis using the real-world Acute Myeloid Leukemia (AML) dataset. The results of the empirical analysis demonstrate that the method proposed in this paper outperforms existing approaches in terms of predictive performance.

Keywords: model averaging; accelerated failure time model; censoring indicator

MSC: 62D10; 62N01; 62N02

Citation: Liao, L.; Liu, J. Model Averaging for Accelerated Failure Time Models with Missing Censoring Indicators. *Mathematics* **2024**, *12*, 641. https://doi.org/10.3390/math12050641

Academic Editor: Jiancang Zhuang

Received: 23 January 2024
Revised: 16 February 2024
Accepted: 20 February 2024
Published: 22 February 2024

1. Introduction

In some practical scenarios, we often need to select useful models from a candidate model set. A popular approach to address this issue is model selection. Methods such as the Akaike Information Criterion (AIC) [1], Mallows' Cp [2] and Bayesian Information Criterion (BIC) [3] are designed to identify the best model. However, in cases where a single model does not receive strong support from the data, these model-selection methods may overlook valuable information from other candidate models, leading to issues of model-selection uncertainty and bias [4].

To tackle these challenges and enhance prediction accuracy, several model-averaging techniques have been developed to leverage all information from the candidate models. Taking inspiration from AIC and BIC, Buckland et al. [5] proposed smoothed AIC (SAIC) and smoothed BIC (SBIC) methods based on AIC and BIC, respectively. Hansen [6] introduced the Mallows model-averaging (MMA) estimator, obtaining weights through the minimization of Mallow's Cp criterion. The MMA estimator asymptotically attains the minimum squared error among the model-averaging estimators in its class. Subsequently, Wan et al. [7] relaxed the constraints of Hansen [6], allowing for non-nested candidate models and continuous weights. In practical applications, many datasets exhibit heteroscedasticity. Therefore, it is essential to explore model-averaging methods tailored for heteroscedastic settings. Firstly, Hensen and Racine [8] proposed Jackknife model averaging (JMA), which determines weights by minimizing a cross-validation criterion. JMA significantly reduces Mean Squared Error (MSE) compared to MMA when errors are heteroscedastic. Secondly, Liu and Okui [9] modified the MMA method proposed by Hensen [6] to make it suitable for

heteroscedastic scenarios. Furthermore, Zhao et al. [10] extended [6]'s work by estimating the covariance matrix based on the weighted average of squared residuals corresponding to all candidate models. This approach improves the model average estimator under heteroskedasticity settings.

In survival analysis, the accelerated failure time (AFT) model provides a straightforward description of how covariates directly impact survival time and has consequently garnered widespread attention. There are several parameter-estimation methods for the Accelerated Failure Time (AFT) model, including Miller's estimator [11], Buckley–James estimator [12], Koul–Susarla–Van Ryzin (KSV) estimator [13] and WLS estimator [14]. However, all these methods assume that the censoring indicator is observable. Therefore, Wang and Dinse [15] improved the KSV estimator to make it adaptable to situations where the censoring indicator is missing.

Under practical conditions, it is common to encounter situations where only the observed time is available and it is uncertain whether the event of interest has occurred. In such cases, data suffer from missingness in the censoring indicator. For example, in a clinical trial for lung cancer, a patient may die for unknown reasons and while the survival time is observed, it is uncertain whether the patient died specifically due to lung cancer. This situation leads to missingness in the censoring indicator. Previous studies have mainly addressed the issue of missingness in the censoring indicator under a specific model. Research on model averaging for right-censored data typically assumes that the censoring indicator is observable. Therefore, this paper adopts the inverse probability weighting method proposed by [15] to construct the response variable. Through appropriate weight-selection criteria, weights are chosen to build the model-averaged estimator for the accelerated failure time model. It significantly enhances the predictive performance of the model and mitigates the bias introduced by the selection of a single model. Compared to previous research, this paper makes two main contributions: First, it introduces a novel model-averaging method for the case of missingness in the censoring indicator. Second, the paper allows for heteroscedasticity and employs model-averaging techniques to estimate variance.

The remaining sections of this paper are organized as follows. In Section 2, we commence by introducing the notation and progressively delineate the methodology and associated theoretical properties of the proposed model-averaging approach. In Section 3, we report the Monte Carlo simulation results. In Section 4, we assess the predictive performance of the proposed model-averaging method against other approaches using the real-world Acute Myeloid Leukemia (AML) dataset. In Section 5, we provide a comprehensive summary of the entire paper and suggest future research directions in this area. All theorem proofs will be presented in Appendix A.

2. Methodology and Theoretical Property

We denote $Y = \log(T) = (Y_1, \cdots, Y_n)'$, $C = \log(V) = (C_1, \cdots, C_n)'$, where T represents the survival time and V denotes the censored time. $X = (X_1', X_2', \ldots, X_n')'$ denotes the covariate matrix for n independent observations, where $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. The accelerated failure time model can be expressed as follows:

$$Y_i = \mu_i + e_i = \sum_{j=1}^{p} \beta_j x_{ij} + e_i, \quad (i = 1, \ldots, n), \tag{1}$$

where e_i is the random error with $E(e_i|X_i) = 0$ and $E(e_i^2|X_i) = \sigma_i^2$.

We assume that there are M candidate models in the candidate model set. Where the mth candidate model contains p_m covariates. Following [7], these candidate model forms are non-nested. The mth candidate model is

$$Y_{mi} = \sum_{j=1}^{p_m} \beta_j x_{ij} + e_{mi}, \quad (i = 1, \ldots, n), \tag{2}$$

for $m = 1, \cdots, M$. The matrix form of (2) is

$$Y_m = X_m \beta_m + e_m, \tag{3}$$

where X_m is an $n \times p_m$ dimensional full column-rank matrix, $Y_m = (Y_{m1}, \cdots, Y_{mn})'$, $\beta_m = (\beta_1, \cdots, \beta_{p_m})'$, $e_m = (e_{m1}, \ldots, e_{mn})'$.

In the case of right censored data, the response variable Y_i might be censored, making it unobservable. We only observe (Z_i, X_i, δ_i), where $Z_i = min(Y_i, C_i)$ and the censoring indicator $\delta_i = I(Y_i \leq C_i)$. Define a missingness indicator ξ_i which is 1 if δ_i is observed and is 0 otherwise. When the censoring indicators are missing, the observed data are $\{Z_i, X_i, \xi_i, \xi_i \delta_i\}$. For simplicity, we set $U_i = (Z_i, X_i)'$. In this paper, similar to [15], we assume the missing mechanism for δ to be:

$$P(\xi = 1|Z, X, \delta) = P(\xi = 1|Z).$$

This assumption is more stringent than the missing at random (MAR) condition yet less restrictive than the assumption of missing completely at random (MCAR).

Koul et al. [13] introduced a method that involves synthetic data for constructing linear regression models. Wang and Dinse [15] extended [13]'s method to address the situation where censoring indicators are missing. In our work, we follow the approach proposed by [15] to construct a response in the form of inverse probability weighting, specifically:

$$Y_{Wi} = \frac{\frac{\xi_i \delta_i}{\pi(Z_i)} + (1 - \frac{\xi_i}{\pi(Z_i)}) m(U_i)}{1 - G_n(Z_i)} Z_i, \tag{4}$$

where $\pi(z) = E(\xi|Z = z)$, $m(u) = E(\delta|U = u)$. $G_n(\cdot)$ represents the cumulative distribution function of C . It is easy to observe that under the missing data mechanism in this paper:

$$E(Y_{Wi}|X_i) = \mu_i = X_i \beta.$$

Similar to Equation (2), we have:

$$Y_{Wi} = \sum_{j=1}^{p_m} \beta_j x_{ij} + e_{Wi}, \quad (i = 1, \ldots, n), \tag{5}$$

where $E(e_{Wi}|X_i) = 0$, $\sigma^2_{Wi} = var(e_{Wi}|X_i)$. This is expressed in matrix form as:

$$Y_W = X_m \beta_m + \boldsymbol{e}_W, \tag{6}$$

where $Y_W = (Y_{W1}, \cdots, Y_{Wn})'$, $e_W = (e_{W1}, \cdots, e_{Wn})'$. And then the weighted least squares estimator of β_m:

$$\widehat{\beta}_m = (X'_m D X_m)^{-1} X'_m D Y_W, \tag{7}$$

where $D = diag\{\frac{1}{\sigma^2_{W1}}, \cdots, \frac{1}{\sigma^2_{Wn}}\}$.

Let $\mu_{mi} = E(Y_{Wi}|X_i)$; subsequently, the estimation for the mth candidate model $\boldsymbol{\mu}_m = (\mu_{m1}, \cdots, \mu_{mn})'$ is given by:

$$\widehat{\boldsymbol{\mu}}_m = X_m \widehat{\beta}_m = X_m (X'_m D X_m)^{-1} X'_m D Y_W = P_m Y_W, \tag{8}$$

where $P_m = X_m (X'_m D X_m)^{-1} X'_m D$. Denote weight vector $\boldsymbol{w} = (w_1, \cdots, w_M)^T$, belonging to the set $\mathcal{H}_M = \{\boldsymbol{w} \in [0, 1]^M : \sum_{m=1}^M w_m = 1\}$. The model-averaging estimator of $\boldsymbol{\mu}$ is defined as follows:

$$\widehat{\boldsymbol{\mu}}_{G_n}(\boldsymbol{w}) = \sum_{m=1}^{M} w_m X_m \widehat{\beta}_m = \sum_{m=1}^{M} w_m X_m (X'_m D X_m)^{-1} X'_m D Y_W = P(w) Y_W, \tag{9}$$

for any $w \in \mathcal{H}_M$, where $P(w) = \sum_{m=1}^{M} w_m X_m (X_m' D X_m)^{-1} X_m' D$.

Define the square loss function $L_{G_n}(\boldsymbol{w}) = \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{w})\|^2$, where $\|\cdot\|$ denotes the Euclidean norm. Then the risk function is defined as:

$$R_{G_n}(\boldsymbol{w}) = E(L_{G_n}(\boldsymbol{w})) = \|P(\boldsymbol{w})\boldsymbol{\mu} - \boldsymbol{\mu}\|^2 + tr\{P(\boldsymbol{w})\Omega P'(\boldsymbol{w})\}, \tag{10}$$

where $\Omega = diag\{\sigma_{W1}^2, \cdots, \sigma_{Wn}^2\}$. The derivation of (10) is as follows:

$$\begin{aligned} R_{G_n}(\boldsymbol{w}) &= E[L_{G_n}(\boldsymbol{w})] \\ &= E[(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{w}))'(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{w}))] \\ &= E[\boldsymbol{u}'\boldsymbol{\mu} - 2\boldsymbol{u}'P(\boldsymbol{w})Y_W + Y_W'P'(\boldsymbol{w})P(\boldsymbol{w})Y_W] \\ &= \boldsymbol{u}'\boldsymbol{\mu} - 2\boldsymbol{u}'P(\boldsymbol{w})\boldsymbol{\mu} + \boldsymbol{u}'P'(\boldsymbol{w})P(\boldsymbol{w})\boldsymbol{\mu} + tr(P'(\boldsymbol{w})\Omega P(\boldsymbol{w})) \\ &= (P(\boldsymbol{w})\boldsymbol{\mu} - \boldsymbol{\mu})'(P(\boldsymbol{w})\boldsymbol{\mu} - \boldsymbol{\mu}) + tr(P(\boldsymbol{w})\Omega P'(\boldsymbol{w})). \end{aligned} \tag{11}$$

Regarding the choice of weights, a natural approach is to minimize the risk function to obtain the optimal weights. However, as shown in Equation (11), we recognize that the risk function includes the unknowns $\boldsymbol{\mu}$, which makes it infeasible to directly minimize the risk function to obtain the optimal weights. Therefore, we replace $\boldsymbol{\mu}$ with Y_W and seek an unbiased estimator of the risk function as the criterion for weight selection.

Define the criterion for weight selection as

$$C_{G_n}(\boldsymbol{w}) = \|Y_W - \hat{\boldsymbol{\mu}}(\boldsymbol{w})\|^2 + 2tr\{P(\boldsymbol{w})\Omega\}. \tag{12}$$

It is not difficult to observe that $E(C_{G_n}(\boldsymbol{w})) = R_{G_n}(\boldsymbol{w}) + \sum_{i=1}^{n} \sigma_{Wi}^2$. By disregarding a term that is independent of $\boldsymbol{w}$, $C_{G_n}(\boldsymbol{w})$ serves as an unbiased estimator of the risk function .

In practice, $m(\cdot)$, $\pi(\cdot)$ and $G_n(\cdot)$ are usually unknown; therefore, we need to estimate them. Firstly regarding the estimation of $m(u)$, it is usually estimated by the Logit model. Suppose $m(u)$ is estimated by the parametric model $m_0(u;\theta)$, where $m_0(u;\theta) = \frac{e^{U\theta}}{1+e^{U\theta}}$. By the maximum likelihood estimation method, we can obtain the parameter estimate $\widehat{\theta}_n$ for the parameter θ. $\pi(z)$ usually can be estimated nonparametrically by

$$\widehat{\pi}_n(z) = \sum_{i=1}^{n} \xi_i W\left(\frac{z - Z_i}{b_n}\right) / \sum_{i=1}^{n} W\left(\frac{z - Z_i}{b_n}\right),$$

where $W(\cdot)$ is a kernel function and b_n is a bandwidth sequence. Next, we define $u(z) = E(\delta | Z = z)$, $u(z)$ estimated nonparametrically by

$$\widehat{u}_n(z) = \frac{\sum_{i=1}^{n}\left(\delta_i \frac{\xi_i}{\widehat{\pi}_n(Z_i)} K\left(\frac{z - Z_i}{h_n}\right)\right)}{\sum_{i=1}^{n}\left(\frac{\xi_i}{\widehat{\pi}_n(Z_i)} K\left(\frac{z - Z_i}{h_n}\right)\right)},$$

where $K(\cdot)$ is a kernel function and h_n is a bandwidth sequence. We adopt the following estimator of $G_n(z)$:

$$\widehat{G}_n(z) = 1 - \prod_{i:Z_i \le z} \left(\frac{n - R_i}{n - R_i + 1}\right)^{1 - \widehat{u}_n(Z_i)},$$

where R_i denotes the rank of Z_i .

Next, replacing $m(\cdot)$, $\pi(\cdot)$ and $G_n(\cdot)$ with $m_0(\cdot,\cdot)$, $\widehat{\pi}_n(\cdot)$ and $\widehat{G}_n(\cdot)$, we have:

$$\widehat{Y}_{Wi} = \frac{\frac{\xi_i \delta_i}{\widehat{\pi}_n(Z_i)} + (1 - \frac{\xi_i}{\widehat{\pi}_n(Z_i)}) m_0(U_i, \widehat{\theta}_n)}{1 - \widehat{G}_n(Z_i)} Z_i.$$

And the corresponding weight selection criterion is as follows:

$$C_{\widehat{G}_n}(\boldsymbol{w}) = \|\hat{Y}_W - \hat{\boldsymbol{\mu}}_{\widehat{\boldsymbol{G}}_{\boldsymbol{n}}}(\boldsymbol{w})\|^2 + 2trace\{P(\boldsymbol{w})\Omega\}, \tag{13}$$

where $\widehat{Y}_W = (\widehat{Y}_{W1}, \cdots, \widehat{Y}_{Wn})$. The weights for minimizing $C_{\widehat{G}_n}(\boldsymbol{w})$ are given by:

$$\widetilde{\boldsymbol{w}} = \arg \min_{w \in \mathcal{H}_M} C_{\widehat{G}_n}(\boldsymbol{w}). \tag{14}$$

Then, we enumerate the necessary regularity conditions for the asymptotic optimality.

(C1) Let $S(t) = 1 - (1 - F(t)(1 - G_n(t)))$ and $\tau_H = \inf\{t : S(t) = 1\}$, where $F(t)$ is the cumulative distribution function of Y_i . Assume that $1 - G_n(\tau_H -) > 0$.
(C2) There exists a positive constant k such that $\max_{1 \le i \le n} |\mu_i| \le k$.
(C3) Denote $\xi_n = \inf_{\mathbf{w} \in \mathcal{H}_M} R_{G_n}(\boldsymbol{w})$ and w_m^0 is an $M \times 1$ unit vector in which the mth element is 1 and the others are 0. For some integer $1 \le J < \infty$ and some positive constant k such that $E(e_i^{4J}) \le k < \infty$, assume

$$M\xi_n^{-2J} \sum_{m=1}^{M} \left\{ R_{G_n}\left(\boldsymbol{w}_m^0\right) \right\}^J \to 0.$$

(C4) There exists $\epsilon > 0$ such that $\inf e_{Wi}^2 > \epsilon$, $i = 1, \cdots, n$.
(C5) $m(\cdot)$ and $\pi(\cdot)$ are bounded.
(C6) $nh_n \to \infty$ and $nh_n^2 \to 0$.
(C7) Let $\tilde{p} = \max_m p_m$, ρ_{ii}^m denote the ith diagonal element of P_m. There exists a constant c such that $\left|\rho_{ii}^{(m)}\right| \le cn^{-1}p_m$.

Condition (C1) is utilized in [16] and it ensures that $1 - G_n(t)$ is not equal to 0. Condition (C2) mandates that the conditional expectation of μ_i remains within bounded limits, in line with assumptions seen in prior research, including [7,17]. Condition (C3) is a requirement commonly found in model-averaging literature (e.g., [7,18]). Condition (C4) mandates the non-degeneracy of the covariance matrix Ω as $n \to \infty$. Similar assumptions can also be found in [9,10]. Similar to [15], Conditions (C5) and (C6) impose constraints on the bounds of $m(\cdot)$, $\pi(\cdot)$ and bandwidth, respectively. Condition (C7) is frequently employed in the analysis of the asymptotic optimality of cross-validation methods, as seen in prior works like [8].

Theorem 1. *Under Conditions (C1) to (C6),*

$$\frac{L_{\widehat{G}_n}(\widetilde{\boldsymbol{w}})}{\inf\limits_{\boldsymbol{w} \in \mathcal{H}_M} L_{\widehat{G}_n}(\boldsymbol{w})} \xrightarrow{p} 1.$$

Theorem 1 establishes the asymptotic optimality of the model-averaging procedure employing weights $\widetilde{\boldsymbol{w}}$, as its squared loss converges to that of the infeasible best possible model average estimator.

In most cases, Ω is unknown and needs to be estimated. We estimate Ω using residuals derived from the model-averaging process: $\widehat{\boldsymbol{e}}(\boldsymbol{w}) = \widehat{Y}_W - \widehat{\boldsymbol{\mu}}(\boldsymbol{w}) = \{\widehat{e}_{W1}(\boldsymbol{w}), \cdots, \widehat{e}_{Wn}(\boldsymbol{w})\}'$. Specifically, the estimator of Ω is

$$\widehat{\Omega}(\boldsymbol{w}) = diag\{\widehat{\sigma}_{W1}^2(\boldsymbol{w}), \cdots, \widehat{\sigma}_{Wn}^2(\boldsymbol{w})\}, \tag{15}$$

where $\widehat{\sigma}_{Wi}^2 = var(\widehat{e}_{Wi})$.

In the existing literature on model averaging, most estimates of variance are predominantly derived from the largest candidate model, as exemplified by works such as [6,16]. In contrast, our approach, following [10], leverages information from all candidate models for estimation rather than relying on a single model. Such an estimation method is more robust. Replacing Ω by $\widehat{\Omega}(\boldsymbol{w})$ in (13), $C(\boldsymbol{w})$ becomes

$$\widehat{C}_{\widehat{G}_n}(\boldsymbol{w}) = \|\widehat{Y}_W - \widehat{\boldsymbol{\mu}}_{\widehat{\boldsymbol{G}}_{\boldsymbol{n}}}(\boldsymbol{w})\|^2 + 2trace\{P(\boldsymbol{w})\widehat{\Omega}(\boldsymbol{w})\}. \tag{16}$$

The weights that minimize $\widehat{C}_{\widehat{G}_n}(\boldsymbol{w})$ are as follows:

$$\widehat{\boldsymbol{w}} = \arg\min_{w \in \mathcal{H}_M} \widehat{C}_{\widehat{G}_n}(\boldsymbol{w}). \tag{17}$$

This weight selection criterion $\widehat{C}_{\widehat{G}_n}(\boldsymbol{w})$ is a cubic function of w.

Theorem 2. *Under Conditions (C1) to (C7),*

$$\frac{L_{\widehat{G}_n}(\widehat{\boldsymbol{w}})}{\inf\limits_{\boldsymbol{w} \in \mathcal{H}_M} L_{\widehat{G}_n}(\boldsymbol{w})} \xrightarrow{p} 1.$$

3. Simulation

In the simulation study, we generate data from the accelerated failure time (AFT) model, $\log(T_i) = Y_i = \sum_{j=1}^{1000} \beta_j x_{ij} + e_i$, where $\beta_j = 1/j^2$; the observations of $X_i = (x_{i1}, x_{i2}, \cdots, x_{i1000})$ are generated from a multivariate normal distribution with zero mean and covariance matrix $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = 0.5^{|i-j|}$. The errors e_i follow normal distribution $N(0, \gamma^2(x_{i2}^4 + 0.01))$. By varying the value of γ, we allow R^2 to range from 0.1 to 0.9. This variance specification closely resembles that of [8]. However, we introduce a small constant, 0.01, to ensure that the variances remain strictly positive. The censoring time C_i is generated from $N(C_0, 7)$. By varying the value of C_0, we achieve censoring rates (CRs) of approximately 20%, 40%. We set sample sizes n = 150, 300. Here, our model configuration is set in a nested form, meaning the first m models include the first m regressors. The number of candidate models M was set to be $\lceil 3n^{1/3} \rceil$, where $\lceil x \rceil$ denote the smallest integer greater than x.

Based on the missing mechanism described in this paper, we assume that the probability of missing censoring indicators, denoted as $1 - \pi(z)$, is determined via a logistic model: $\log\{\frac{\pi(z)}{1-\pi(z)}\} = \theta_1 + \theta_2 z$. Following [15], we employed the uniform kernel function $W(x) = \frac{1}{2}$ for $|x| \leq 1$ and $W(x) = 0$ otherwise. Additionally, we used the biweight kernel function $K(x) = \frac{15}{16}(1 - 2x^2 + x^4)$ for $|x| \leq 1$ and $K(x) = 0$ otherwise. The bandwidths were $b_n = h_n = n^{-\frac{1}{3}} \max(Z)$. We estimated $m(u)$ under the logistic model: $\log\{\frac{m(u)}{1-m(u)}\} = \gamma_1 + \gamma_2 z + \gamma_3 x$. As highlighted by [19], when the data on δ are completely (or quasi-completely) separated, the maximum likelihood estimate of $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ does not exist. In our simulation setup, the number of covariates significantly exceeds the sample size. Therefore, we employ the lasso method to estimate the parameters.

We compare the proposed Model-Averaging method for the Missing Censoring Indicators in the Heteroscedastic setting (HCIMA) with other classical model-selection and model-averaging methods in this article. Brief descriptions of these methods are provided below:

- The model-selection methods rely on AIC and BIC, where the AIC and BIC criterion for the mth model are defined as follows:

  $$\mathrm{AIC}(m) = \log\left(\widehat{\sigma}^2_{\widehat{G}_n m}\right) + 2n^{-1}\,\mathrm{tr}(P_m),$$

  and

  $$\mathrm{BIC}(m) = \log\left(\widehat{\sigma}^2_{\widehat{G}_n m}\right) + n^{-1}\,\mathrm{tr}(P_m) log(n),$$

  where $\widehat{\sigma}^2_{(\widehat{G}_n m)} = n^{-1}\left\|\widehat{Y}_W - \widehat{\mu}_{\widehat{G}_n m}\right\|^2$.
- Model methods based on SAIC and SBIC: The weights for the mth candidate model are given by:

  $$w^{AIC}_{(m)} = \exp(-AIC_m/2) / \sum_{m=1}^{M} \exp(-AIC_m/2),$$

  $$w^{BIC}_{(m)} = \exp(BIC_m/2) / \sum_{m=1}^{M} \exp(-BIC_m/2),$$

where $AIC_m = AIC(m) - min(AIC)$, $BIC_m = BIC(m) - min(BIC)$.

- Additionally, we compare our approach with the method that estimates the variance using the maximum candidate model (MCIMA). And the specifics of variance estimation and weight selection in their approach are as follows:

$$\widehat{\boldsymbol{\sigma}}_{\widehat{G}_n} = (\widehat{\sigma}_{\widehat{G}_n 1}, \cdots, \widehat{\sigma}_{\widehat{G}_n n})^T = \sqrt{\frac{n}{n-M}}(I - P_M)\widehat{Y}_W,$$

$$\widehat{C}_n(\boldsymbol{w}) = \|\widehat{Y}_W - \widehat{\boldsymbol{\mu}}(\boldsymbol{w})\|^2 + trace\{P(\boldsymbol{w})\widehat{\Omega}\},$$

where $\widehat{\Omega} = diag\{\widehat{\sigma}^2_{\widehat{G}1}, \cdots, \widehat{\sigma}^2_{\widehat{G}n}\}$.

In the simulation, we utilize the Mean Squared Error (MSE) to evaluate the performance of various methods, where the MSE is defined as $\frac{1}{n}\left\|\widehat{\mu}_{\widehat{G}_n} - \mu\right\|^2$. We present the mean of MSEs from 500 replications.

Figures 1 and 2, respectively, show the Mean Squared Error (MSE) values for various methods across 500 repetitions under different censored rates and sample sizes, with missing rates of 20% and 40%. In terms of Mean Squared Error (MSE), our proposed HCIMA method outperforms other approaches. Additionally, the MCIMA method performs better than existing methods in all cases except for when compared to HCIMA. Furthermore, it is evident that SAIC and SBIC outperform their respective AIC and BIC counterparts, further highlighting the advantages of model-averaging methods.

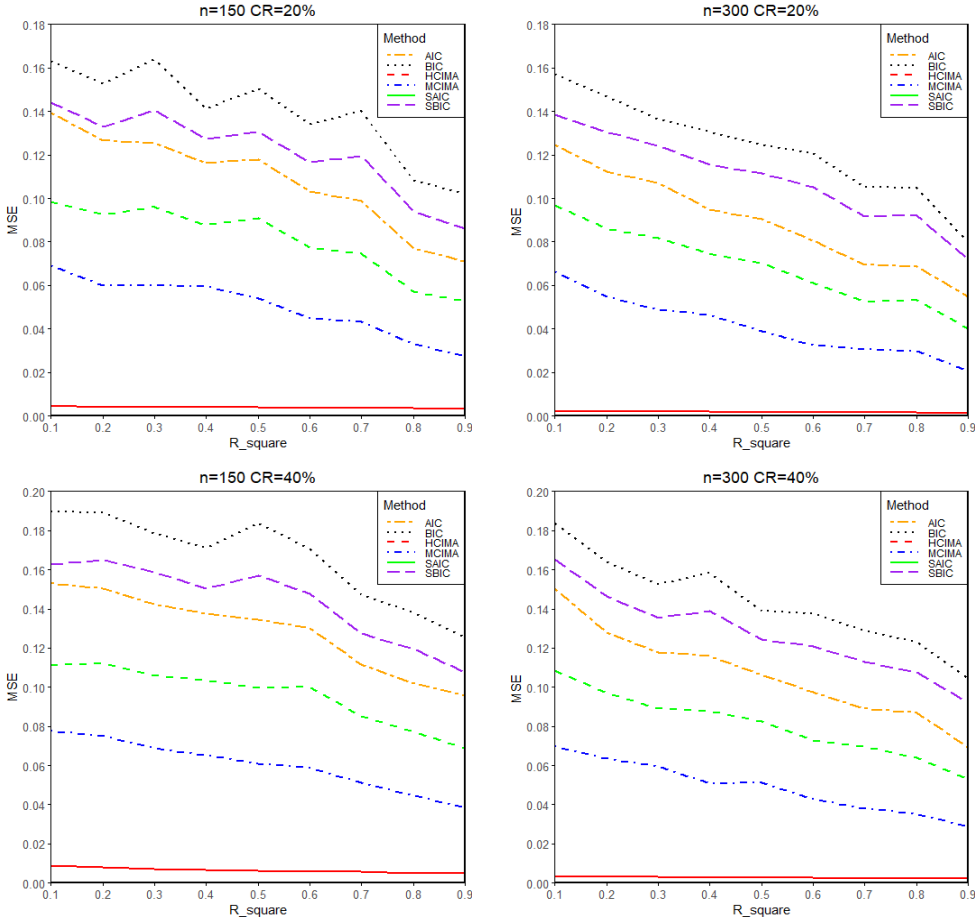

Figure 1. Mean Squared Errors (MSEs) of various methods under different sample sizes and censor rates at MR = 20%.

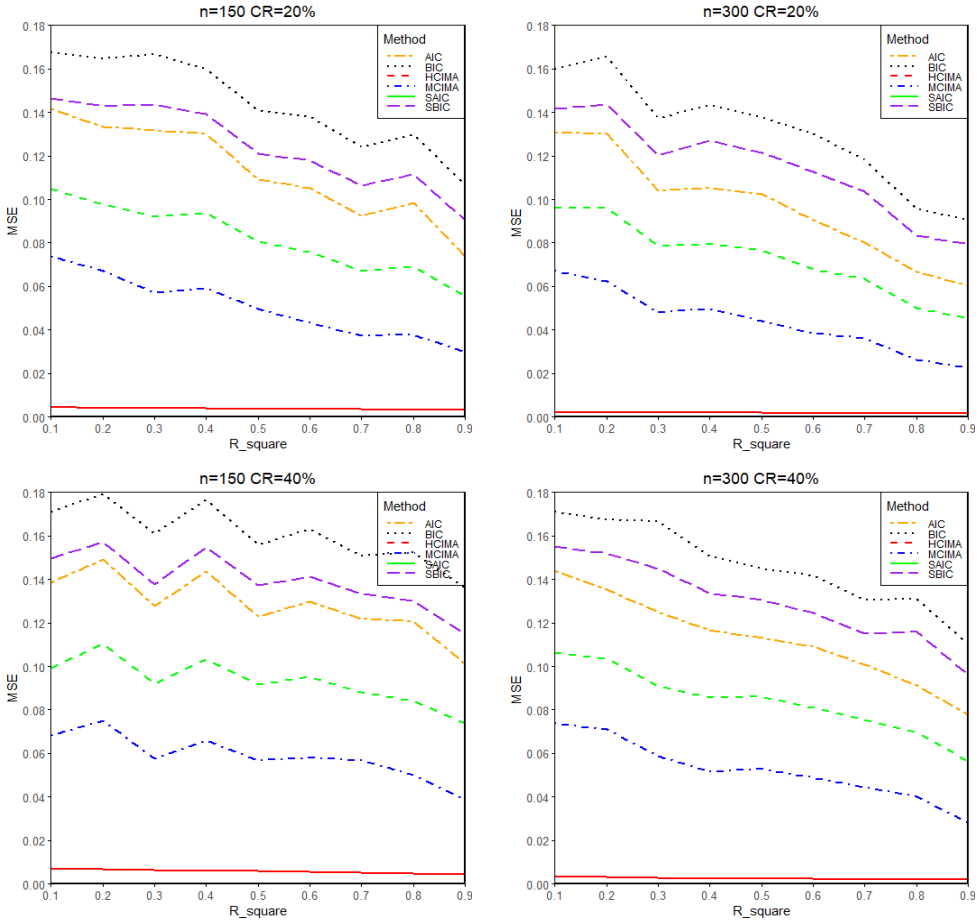

Figure 2. Mean Squared Errors (MSEs) of various methods under different sample sizes and censor rates at MR = 40%.

Comparing Figures 1 and 2, it is observed that the MSE at MR = 20% is slightly higher than at MR = 40%. The reason for this occurrence is that when $\xi_i = 1, \delta_i = 0$, the signs of Y_{Wi} and Z_i are opposite. As MR increases, the occurrence of the $\xi_i = 1, \delta_i = 0$ situation decreases. Although this result may seem counterintuitive, it does not affect the performance of the method proposed in this paper, which still keeps its advantages in this case.

4. Real Data Analysis

In this section, we assess the predictive performance of our proposed HCIMA method using the real Acute Myeloid Leukemia (AML) dataset. This dataset contains 672 samples, including 97 variables such as patient age, survival time, gender, race, mutation count, etc. For more specific information about this dataset, we refer the reader to https://www.cbioportal.org/study/clinicalData?id=aml_ohsu_2018 (accessed on 13 December 2023).

We selected ten variables for analysis: Cause Of Death, Age, Sex, Overall Survival Status, Overall Survival Months (Survival Time), Number of Cumulative Treatment Stages, Cumulative Treatment Regimen Count, Mutation Count, Platelet Count and WBC (White Bloodcell Count). After removing rows with missing values, we retained a total of 396 samples. We treat samples with unknown causes of death as missing censoring indicators. Among these 396 samples, 76 have unknown causes of death and 167 samples are still alive after the clinical trial ends. Therefore, the missing rate is approximately 19% and the censoring rate is 42%. We focus on the impact of seven variables, excluding "Cause Of Death" and "Overall Survival Status" on Survival Time. Therefore, we can construct $2^7 - 1 = 127$ non-nested candidate models.

We randomly select data from n_0 samples as the training dataset, while the remaining $n_1 = n - n_0$ samples are used as the testing dataset. We set the training dataset size to 50%, 60%, 70% and 80% of the total dataset size, respectively. Following [16,20], we employed the normalized mean squared prediction error (NMSPE) as the performance metric:

$$\text{NMSPE} = \frac{\sum_{i=n_0+1}^{n}\left(\widehat{Y}_{Wi} - \widehat{\mu}_i\right)^2}{\min_{m=1,2,\cdots,M}\sum_{i=n_0+1}^{n}\left(\widehat{Y}_{Wi} - \widehat{\mu}_{mi}\right)^2},$$

where $\widehat{\mu}_i$ represents the predicted value and $\widehat{\mu}_{mi}$ denotes the value of $\widehat{\mu}$ for the mth model.

We calculate the mean, the standard deviation and the optimal rate of each method over these 1000 repetitions. Specifically, the optimal rate refers to the frequency at which the minimum value is achieved across these 1000 repetitions.

Table 1 displays the mean, optimal rates and standard deviations of NMSPE for each method over 1000 repetitions. Consistent with the simulation results, the HCIMA method exhibits the lowest average NMSPE and standard deviation and the highest optimal rate. The MCIMA method also performs well, ranking second after HCIMA. This indicates that the proposed model-averaging methods in this paper demonstrate superior predictive performance compared to other approaches.

Table 1. The mean, optimal rate and standard deviation of NMSPE.

	Method	AIC	SAIC	BIC	SBIC	MCIMA	HCIMA
50%	Mean	1.3628	1.3370	1.3517	1.3345	1.2765	**1.2165**
	Standard deviation	0.5283	0.5060	0.5123	0.4970	0.4039	**0.3500**
	Optimal rate	0.084	0.137	0.042	0.093	0.306	**0.338**
60%	Mean	1.3663	1.3388	1.3556	1.3404	1.2651	**1.1800**
	Standard deviation	0.5504	0.5166	0.5151	0.5068	0.4343	**0.3066**
	Optimal rate	0.094	0.119	0.049	0.091	0.288	**0.359**
70%	Mean	1.3347	1.3213	1.3361	1.3259	1.2451	**1.1766**
	Standard deviation	0.5324	0.5232	0.5288	0.5257	0.3433	**0.2966**
	Optimal rate	0.097	0.140	0.057	0.079	0.259	**0.368**
80%	Mean	1.2794	1.2619	1.2828	1.2628	1.2034	**1.1504**
	Standard deviation	0.4865	0.4714	0.4861	0.4777	0.2941	**0.2030**
	Optimal rate	0.083	0.165	0.063	0.129	0.240	**0.320**

5. Discussion

To address the uncertainty in model selection and enhance predictive accuracy, this paper proposes a novel model-averaging approach for the accelerated failure time model with missing indicators. Moreover, we establish asymptotic optimality under certain mild conditions. In Monte Carlo simulations, the method proposed in this paper exhibits lower mean squared errors compared to other model-selection and model-averaging methods. Empirical results demonstrate that the proposed method has a lower NMSPE compared to other approaches, indicating its superior predictive performance. This further underscores the applicability of the proposed method to real-life data scenarios with missing censoring indicators.

In this paper, we introduce the inverse probability weighted form of response variable proposed in [15]. The primary advantage of this form of response variable lies in its double robustness, making it less susceptible to the impact of model misspecification (if $\pi(\cdot)$ or $m(\cdot)$ is misspecified). However, as mentioned in [15], its drawback, compared to synthetic response [13], regression calibration and imputation [15], is a larger variance. Yet, in practical scenarios, the harm caused by model misspecification often outweighs the harm of higher variance. Therefore, in our work, we follow the recommendation of [15] to use

the inverse probability weighted form of the response variable. A future research direction is to further enhance this response variable for better applicability in the context of missing censoring indicators.

As far as we know, there is currently very limited research on model averaging for missing censoring indicators. Therefore, there are still many questions that deserve further investigation. There is potential for extending our approach to high-dimensional data in terms of data and in terms of models, exploration into partial linear models, generalized linear models and other extensions could be pursued.

Author Contributions: Conceptualization, L.L.; Methodology, L.L.; Writing—review and editing, L.L.; Software, J.L.; Data curation, J.L.; Writing—original draft, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Acknowledgments: We would like to thank the reviewers and editors for their careful reading and constructive comments.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

In this appendix, we provide the proofs for Theorems 1 and 2. To facilitate the presentation, we begin with several lemmas.

Lemma A1. *Under Conditions (C1) to (C3), there exists a positive constant* c_1 *such that*

$$\max_{1\le i\le n} E\left(e_{W,i}^{4J} \mid X_i\right) \le c_1,$$

where J is given in Condition (C3).

Lemma A1 is consistent with Lemma 6.1 in [16] and under our specific conditions, the proof technique for Lemma 1 is the same as the proof technique for Lemma 6.1 in [16].

Lemma A2. *Under Conditions (C1) to (C3),*

$$\left|\frac{L_{G_n}(\boldsymbol{w})}{R_{G_n}(\boldsymbol{w})} - 1\right| = o_p(1).$$

Under our specific conditions, we can prove this lemma using the same techniques as the proof of (A.3) in [7]. Therefore, we omit the proof here.

Lemma A3. *Under Conditions (C1) to (C5), as* $n \to \infty$ *, we have*

$$\left\|\widehat{Y}_W - Y_W\right\|^2 = o_p(1).$$

Proof of Lemma A3.

$$\begin{aligned}
\left\|\widehat{Y}_W - Y_W\right\|^2 &= \sum_{i=1}^{n}\{\frac{\frac{\xi_i\delta_i}{\pi(Z_i)} + (1-\frac{\xi_i}{\pi(Z_i)})m(U_i)}{1-G_n(Z_i)} - \frac{\frac{\xi_i\delta_i}{\widehat{\pi}_n(Z_i)} + (1-\frac{\xi_i}{\widehat{\pi}_n(Z_i)})m_0(U_i,\widehat{\theta}_n)}{1-\widehat{G}_n(Z_i)}\}^2 Z_i \\
&\le K\sum_{i=1}^{n}\{\frac{1}{1-G_n(Z_i)} - \frac{1}{1-\widehat{G}_n(Z_i)}\}^2 Z_i \\
&\le C\left\{n^{1/2}\max_{1\le i\le n}\left|\frac{1}{1-\widehat{G}_n(Z_i)} - \frac{1}{1-G_n(Z_i)}\right|\right\}^2\left(\frac{1}{n}\mu^T\mu + \frac{1}{n}e_W^T e_W\right),
\end{aligned}$$

where K is a constant. By Condition (C2), we have $\frac{1}{n}\mu^T\mu = O_p(1)$ and $\frac{1}{n}e_W^T e_W = O_p(1)$. According to [15], $\widehat{G}_n(Z_i) - G_n(Z_i) = o_p(1)$. Combined with Condition (C1), we have:

$$\left|\frac{\widehat{G}_n(z) - G_n(z)}{1 - G_n(z)}\right| = o_p(1),$$

and

$$\left|\frac{\widehat{G}_n(z) - G_n(z)}{1 - \widehat{G}_n(z)}\right| = o_p(1).$$

Similar to the proof of Lemma 6.2 in [16], we have:

$$n^{1/2}\max_{1\le i\le n}\left|\frac{1}{1-\widehat{G}_n(Z_i)} - \frac{1}{1-G_n(Z_i)}\right| = o_p(1).$$

Furthermore, we can obtain

$$\|\hat{Y}_W - Y_W\|^2 = o_p(1).$$

□

With the three lemmas mentioned above, we can now proceed to prove Theorem 1.

Proof of Theorem 1. Fist, we note that

$$\begin{aligned}
C_{\widehat{G}_n}(\boldsymbol{w}) &= \|\widehat{Y}_W - \widehat{\boldsymbol{\mu}}_{\widehat{G}_n}(\boldsymbol{w})\|^2 + 2trace\{P(\boldsymbol{w})\Omega\} \\
&= \|\widehat{Y}_W - \mu + \mu - \widehat{\boldsymbol{\mu}}_{\widehat{G}_n}(\boldsymbol{w})\|^2 + 2trace\{P(\boldsymbol{w})\Omega\} \\
&= \|\widehat{Y}_W - \mu\|^2 + \|\mu - \widehat{\boldsymbol{\mu}}_{\widehat{G}_n}(\boldsymbol{w})\|^2 + 2(\widehat{Y}_W - \mu)'(\mu - \widehat{\boldsymbol{\mu}}_{\widehat{G}_n}(\boldsymbol{w})) + 2trace\{P(\boldsymbol{w})\Omega\} \\
&= \|e_W\|^2 + L_{\widehat{G}_n}(\boldsymbol{w}) + 2e'_W(\mu - P(\boldsymbol{w})\mu + P(\boldsymbol{w})\mu - \widehat{\boldsymbol{\mu}}_{\widehat{G}_n}(\boldsymbol{w})) + 2trace\{P(\boldsymbol{w})\Omega\} \\
&= L_{\widehat{G}_n}(\boldsymbol{w}) + 2e'_W(I - P(\boldsymbol{w}))\mu + 2trace(P(\boldsymbol{w})\Omega) - 2e'_W P(\boldsymbol{w})e_W + \|e_W\|^2.
\end{aligned}$$

Following [7], except for a term unrelated to w, to prove Theorem 1, we only need to verify

$$\sup_{\boldsymbol{w}\in\mathcal{H}_M}\frac{|e'_W(I - P(\boldsymbol{w}))\mu|}{R_{G_n}(\boldsymbol{w})} = o_p(1), \tag{A1}$$

$$\sup_{\mathbf{w}\in\mathcal{H}_M}\frac{|trace(P(\boldsymbol{w})\Omega) - 2e'_W P(\boldsymbol{w})e_W|}{R_{G_n}(\boldsymbol{w})} = o_p(1), \tag{A2}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}_M}|\frac{L_{\widehat{G}_n}(\boldsymbol{w})}{R_{G_n}(\boldsymbol{w})} - 1| = o_p(1). \tag{A3}$$

We begin by proving Equation (A1). As per Equation (11), we can ascertain that:

$$R_{G_n}(\boldsymbol{w}_m^0) \ge \|P(\boldsymbol{w}_{\boldsymbol{m}}^{\mathbf{0}})\mu - \mu\|^2, \tag{A4}$$

$$R_{G_n}(\boldsymbol{w}_m^0) \ge trace(P(\boldsymbol{w}_{\boldsymbol{m}}^{\mathbf{0}})\Omega P^T(\boldsymbol{w}_{\boldsymbol{m}}^{\mathbf{0}})). \tag{A5}$$

Furthermore, we denote the maximum eigenvalue of matrix A as $\lambda_{max}(A)$; since P_m is an idempotent matrix, we have:

$$\lambda_{\max}(P_m) = 1, \tag{A6}$$

$$\lambda_{\max}\{P(w)\} \le \sum_{m=1}^{M} w_m\lambda_{\max}\{P_m\} \le 1. \tag{A7}$$

According to the proof of Theorem 1 in [21], we have:

$$\lim_{n\to\infty} \sup_{\boldsymbol{w}\in\mathcal{H}_M} \lambda_{\max}(P(\boldsymbol{w})P(\boldsymbol{w})') < \infty. \tag{A8}$$

Applying the triangle inequality, Bonferroni's inequality, Chebyshev's inequality and Theorem 2 of [22], we can conclude, for any $\tau > 0$,

$$\begin{aligned}
&P\left\{\sup_{\boldsymbol{w}\in\mathcal{H}_M} \frac{|e_W'(I-P(\boldsymbol{w}))\mu|}{R_{G_n}(\boldsymbol{w})} > \tau\right\} \\
&\le P\left\{\sup_{\boldsymbol{w}\in H_M} \sum_{m=1}^{M} w_m \left|e_W'(I-P_m)\mu\right| > \tau\xi_n\right\} \\
&= P\left\{\max_{1\le m\le M} \left|e_W'(I-P_m)\mu\right| > \tau\xi_n\right\} \\
&= P\left\{\left\{\left|\left\langle e_W, A\left(w_1^0\right)\mu\right\rangle\right| > \tau\xi_n\right\} \cup \cdots \cup \left\{\left|\left\langle e_W, A\left(w_M^0\right)\mu\right\rangle\right| > \tau\xi_n\right\}\right\} \\
&\le \sum_{m=1}^{M} P\left\{\left|\left\langle e_W, A\left(w_m^0\right)\mu\right\rangle\right| > \tau\xi_n\right\} \\
&\le \sum_{m=1}^{M} E\left\{\frac{\left\langle e_W, A\left(w_m^0\right)\mu\right\rangle^{2J}}{\tau^{2J}\xi_n^{2J}}\right\} \\
&\le C_1\tau^{-2J}\xi_n^{-2J} \sum_{m=1}^{M} \left\|\Omega(2J)^{1/2} A\left(\boldsymbol{w}_m^0\right)\mu\right\|^{2J},
\end{aligned}$$

where $\langle,\rangle$ represents an inner product, $A(\boldsymbol{w}) = I - P(\boldsymbol{w})$. C_1 is a constant, $\Omega(2J) = \mathrm{diag}(\gamma_1^2(2J),\ldots,\gamma_n^2(2J))$ and $\gamma_i^2(2J) = E(e_{Wi}^{2J}|X_i)^{1/2J}$. By Lemma A1, $\gamma_i^2(2J) < \infty$; thus, $\lambda_{max}(\Omega(2J))^J = O(1)$. Hence, combining this with Equation (A4), we have:

$$\begin{aligned}
&P\left\{\sup_{\boldsymbol{w}\in H_M} |\langle e_W, A(\boldsymbol{w})\mu\rangle| / R_{G_n}(\boldsymbol{w}) > \tau\right\} \\
&\le C_1\tau^{-2J}\xi_n^{-2J}\lambda_{max}(\Omega(2J))^J \sum_{m=1}^{M} \left\|A\left(\boldsymbol{w}_m^0\right)\mu\right\|^{2J} \\
&\le C_1'\tau^{-2J}\xi_n^{-2J} \sum_{m=1}^{M} \left\|A\left(\boldsymbol{w}_m^0\right)\mu\right\|^{2J} \\
&\le C_1'\tau^{-2J}\xi_n^{-2J} \sum_{m=1}^{M} \left(R_{G_n}\left(\boldsymbol{w}_m^0\right)\right)^J.
\end{aligned}$$

And together with condition (C3), we can prove Equation (A1). Next, we will prove (A2). Similar to the proof of Equation (A1), we have:

$$\begin{aligned}
&P\left\{\sup_{\boldsymbol{w}\in\mathcal{H}_M} |\,\mathrm{trace}[\Omega P(\boldsymbol{w})] - \langle e_W, P(\boldsymbol{w})e_W\rangle| / R_{G_n}(\boldsymbol{w}) > \tau\right\} \\
&\le \sum_{m=1}^{M} P\left\{\left|\mathrm{trace}\left[\Omega P\left(\boldsymbol{w}_m^0\right)\right] - \left\langle e_W, P\left(\boldsymbol{w}_m^0\right)e_W\right\rangle\right| > \tau\xi_n\right\} \\
&\le \sum_{m=1}^{M} E\left\{\frac{\left[\mathrm{trace}\left[\Omega P\left(\boldsymbol{w}_m^0\right)\right] - \left\langle e_W, P\left(w_m^0\right)e_W\right\rangle\right]^{2J}}{\tau^{2J}\xi_n^{2J}}\right\} \\
&\le C_2\tau^{-2J}\xi_n^{-2J} \sum_{m=1}^{M} \left\{\mathrm{tr}\left[P\left(\boldsymbol{w}_m^0\right)'\Omega(4J)P\left(\boldsymbol{w}_m^0\right)\right]\right\}^J,
\end{aligned}$$

where C_2 is a constant, $\Omega(4J) = \mathrm{diag}\left(\gamma_1^2(4J), \ldots, \gamma_n^2(4J)\right)$ and $\gamma_i^2(4J) = E(e_{Wi}^{4J}|X_i)^{1/4J}$. By Lemma A1, $\gamma_i^2(4J) < \infty$; thus, $\lambda_{max}(\Omega(4J))^J = O(1)$. Hence, combining Equation (A5) and condition (C3), we have:

$$\begin{aligned}
&P\left\{\sup_{w\in\mathcal{H}_M} |\operatorname{trace}[\Omega P(w)] - \langle e_W, P(w)e_W\rangle| / R_{G_n}(w) > \tau\right\} \\
\leq& C_2\tau^{-2N}\xi_n^{-2N}\lambda_{\max}[\Omega(4N)]^N \sum_{m=1}^{M} \operatorname{tr}\left[P\left(w_m^0\right)' P\left(w_m^0\right)\right] \\
\leq& C_2\tau^{-2J}\xi_n^{-2J}\left(\inf_i e_{Wi}^2\right)^{-J} \sum_{m=1}^{M}\left\{\inf_i e_{Wi}^2 trace(P^2(w_m^0))\right\}^J \\
\leq& C_3\tau^{-2J}\xi_n^{-2J} \sum_{m=1}^{M}\left[R_{G_n}\left(w_m^0\right)\right]^J = o(1).
\end{aligned}$$

Next, we will prove Equation (A3). Note that

$$\begin{aligned}
&\left|\frac{L_{\widehat{G}_n}(w)}{R_{G_n}(w)} - 1\right| \\
=&\left|\frac{L_{G_n}(w)}{R_{G_n}(w)} - 1 + \frac{L_{\widehat{G}_n}(w) - L_{G_n}(w)}{R_{G_n}(w)}\right| \\
\leq&\left|\frac{L_{G_n}(w)}{R_{G_n}(w)} - 1\right| + \left|\frac{\left\|\mu - \widehat{\mu}_{\widehat{G}_n}(w)\right\|^2 - \|\mu - \widehat{\mu}(w)\|^2}{R_{G_n}(w)}\right|.
\end{aligned} \tag{A9}$$

From Lemma A2, we know that $|\frac{L_{G_n}(w)}{R_{G_n}(w)} - 1| = o_p(1)$, Therefore, to prove (A3), it is sufficient to verify that the second part of Equation (A9) converges to 0 in probability.

$$\begin{aligned}
&\left|\frac{\left\|\mu - \widehat{\mu}_{\widehat{G}_n}(w)\right\|^2 - \|\mu - \widehat{\mu}(w)\|^2}{R_{G_n}(w)}\right| \\
=&\left|\frac{2(\mu - \widehat{\mu}(w))'\left(\widehat{\mu}(w) - \widehat{\mu}_{\widehat{G}_n}(w)\right) + \left\|\widehat{\mu}(w) - \widehat{\mu}_{\widehat{G}_n}(w)\right\|^2}{R_{G_n}(w)}\right| \\
\leq&\frac{2\{L_{G_n}(w)\}^{1/2}\left\|P(w)\left(\widehat{Y}_W - Y_W\right)\right\|}{R_{G_n}(w)} + \frac{\left\|P(w)\left(\widehat{Y}_W - Y_W\right)\right\|^2}{R_{G_n}(w)}.
\end{aligned} \tag{A10}$$

According to Lemma A3, we have:

$$\left\|P(w)\left(\widehat{Y}_W - Y_W\right)\right\|^2 \leq \lambda_{\max}(P(w))\left\|\widehat{Y}_W - Y_W\right\|^2 = O_p(1).$$

Combining this with Lemma A3, we can conclude that (A9) is of $o_p(1)$, which establishes the proof for (A3). □

Proof of Theorem 2. It is evident from Equations (13) and (16) that:

$$\widehat{C}_{\hat{G}_n}(w) = C_{\hat{G}_n}(\mathbf{w}) + 2\operatorname{trace}\{P(w)\widehat{\mathbf{\Omega}}(w)\} - 2\operatorname{trace}\{P(w)\mathbf{\Omega}\}.$$

In conjunction with Theorem 1, it is evident that to prove Theorem 2, we only need to establish:

$$\sup_{\boldsymbol{w}\in\mathcal{H}_{\mathcal{M}}}\left[|\operatorname{trace}\{P(\boldsymbol{w})\widehat{\Omega}(\boldsymbol{w})\}-\operatorname{trace}\{P(\boldsymbol{w})\Omega\}|/R_{G_n}(\boldsymbol{w})\right]=o_p(1). \tag{A11}$$

We denote $Q_m=\operatorname{diag}\left(\rho_{11}^{(m)},\ldots,\rho_{nn}^{(m)}\right)$ and $Q(\boldsymbol{w})=\sum_{m=1}^{M}w_sQ_m$. According to Lemma A1, we have:

$$\lambda_{max}(\Omega)=O(1). \tag{A12}$$

Considering the definition of $\widehat{\Omega}(\boldsymbol{w})$, and employing proof techniques similar to [10,23], we obtain:

$$\begin{aligned}
&\sup_{\boldsymbol{w}\in\mathcal{H}_{\mathcal{M}}}\left[|\operatorname{trace}\{P(\boldsymbol{w})\widehat{\Omega}(\boldsymbol{w})\}-\operatorname{trace}\{P(\boldsymbol{w})\Omega\}|/R_{G_n}(\boldsymbol{w})\right]\\
=&\sup_{\boldsymbol{w}\in\mathcal{H}_{\mathcal{M}}}\left[|\ \{\widehat{Y}_W-P(\boldsymbol{w})\widehat{Y}_W\}'Q(\boldsymbol{w})\{\widehat{Y}_W-P(\boldsymbol{w})\widehat{Y}_W\}-\operatorname{trace}\{Q(\boldsymbol{w})\Omega\}\ |\ /R_{G_n}(\boldsymbol{w})\right]\\
=&\sup_{\boldsymbol{w}\in\mathcal{H}_{\mathcal{M}}}\left[|\ \{e_W+\mu-P(\boldsymbol{w})\widehat{Y}_W\}'Q(\boldsymbol{w})\{e_W+\mu-P(\boldsymbol{w})\widehat{Y}_W\}-\operatorname{trace}\{Q(\boldsymbol{w})\Omega\}\ |\ /R_{G_n}(\boldsymbol{w})\right]\\
\le&\sup_{\boldsymbol{w}\in\mathcal{H}_{\mathcal{M}}}\left[\left|e_W'Q(\boldsymbol{w})e_W-\operatorname{trace}\{Q(\boldsymbol{w})\Omega\}\right|/R_{G_n}(\boldsymbol{w})\right]\\
&+2\sup_{\boldsymbol{w}\in\mathcal{H}_{\mathcal{M}}}\left[\left|e_W'Q(\boldsymbol{w})\{P(\boldsymbol{w})\widehat{Y}_W-\mu\}\right|/R_{G_n}(\boldsymbol{w})\right]\\
&+\sup_{\boldsymbol{w}\in\mathcal{H}_{\mathcal{M}}}\left[\left|\{P(\boldsymbol{w})\widehat{Y}_W-\mu\}'Q(\boldsymbol{w})\{P(\boldsymbol{w})\widehat{Y}_W-\mu\}\right|/R_{G_n}(\boldsymbol{w})\right]\\
\le&\sup_{\boldsymbol{w}\in\mathcal{H}_{\mathcal{M}}}\left[\left|e_W'Q(\boldsymbol{w})e_W-\operatorname{trace}\{Q(\boldsymbol{w})\Omega\}\right|/R_{G_n}(\boldsymbol{w})\right]\\
&+2\sup_{\boldsymbol{w}\in\mathcal{H}_{\mathcal{M}}}\left[\left|e_W'Q(\boldsymbol{w})\{P(\boldsymbol{w})\mu-\mu\}\right|/R_{G_n}(\boldsymbol{w})\right]\\
&+2\sup_{\boldsymbol{w}\in\mathcal{H}_{\mathcal{M}}}\left[\left|e_W'Q(\boldsymbol{w})P(\boldsymbol{w})e_W-\operatorname{trace}\{Q(\boldsymbol{w})P(\boldsymbol{w})\Omega\}\right|/R_{G_n}(\boldsymbol{w})\right]\\
&+2\sup_{\boldsymbol{w}\in\mathcal{H}_{\mathcal{M}}}\left[|\operatorname{trace}\{Q(\boldsymbol{w})P(\boldsymbol{w})\Omega\}|/R_{G_n}(\boldsymbol{w})\right]\\
&+\sup_{\boldsymbol{w}\in\mathcal{H}_{\mathcal{M}}}\left[\left|\{P(\boldsymbol{w})\hat{Y}_W-\mu\}'Q(\boldsymbol{w})\{P(\boldsymbol{w})\hat{Y}_W-\mu\}\right|/R_{G_n}(\boldsymbol{w})\right]\\
\equiv&T_1+T_2+T_3+T_4+T_5.
\end{aligned}$$

Let $\rho=\max_m\max_i\rho_{ii}^{(m)}$. According to condition (C7), we have:

$$\rho=O\left(n^{-1}\tilde{p}\right). \tag{A13}$$

Given the definition of $R_{G_n}(\boldsymbol{w})$ and condition (C4), the following equation holds:

$$\begin{aligned}
R_{G_n}\left(\boldsymbol{w}_m^0\right)\ge\operatorname{trace}\left\{\mathbf{P}_m\boldsymbol{\Omega}P_m^T\right\}\ge\epsilon\operatorname{trace}(P_m)=\epsilon p_m,\\
\xi_n\to\infty\quad\text{and}\quad M\xi_n^{-2J}=o(1).
\end{aligned} \tag{A14}$$

From (A6), (A11), (A13), the Chebyshev inequality and Theorem 2 of [22], for any $\tau>0$, there exist constants c_1 and c_2 such that:

$$\begin{aligned}
\mathrm{P}(T_1 > \tau) &\le \sum_{m=1}^{M} \mathrm{P}\left[\left|e_W' Q_m e_W - \mathrm{trace}(Q_m\Omega)\right| > \tau\xi_n\right] \\
&\le \tau^{-2J}\xi_n^{-2J} \sum_{m=1}^{M} E\left\{e_W^T Q_m e_W - \mathrm{trace}(Q_m\Omega)\right\}^{2J} \\
&\le c_1\tau^{-2J}\xi_n^{-2J} \sum_{m=1}^{M} \{\mathrm{trace}\{\Omega^{1/2}(4J)Q_m\Omega(4J)Q_m\Omega^{1/2}(4J)\}\}^J \\
&\le c_1\tau^{-2J}\xi_n^{-2J} M\lambda_{\max}^{2J}(\Omega(4J)) \max_{1\le m\le M}\{trace(Q_m)\}^J \\
&= \xi_n^{-2J} M\left\{O\left(n^{-1}\widetilde{p}^2\right)\right\}^J = o(1),
\end{aligned} \tag{A15}$$

$$\begin{aligned}
\mathrm{P}(T_3/2 > \tau) &\le \sum_{m=1}^{M} \mathrm{P}\left\{\left| e_W' Q_m P_m e_W - \mathrm{trace}(Q_m P_m\Omega)\right| > \tau\xi_n\right\} \\
&\le \tau^{-2J}\xi_n^{-2J} \sum_{m=1}^{M} E\left[e_W' Q_m P_m e_W - \mathrm{trace}(Q_m P_m\Omega)\right]^{2J} \\
&\le c_2\tau^{-2J}\xi_n^{-2J} \sum_{m=1}^{M} \mathrm{trace}\{\Omega^{1/2}(4J)Q_m P_m\Omega(4J)P_m^T Q_m\Omega^{1/2}(4J)\}^J \\
&\le c_2\tau^{-2J}\xi_n^{-2J} M\lambda_{\max}^{2J}(\Omega(4J))\lambda_{\max}^{J}\left(P_m P_m^T\right) \max_{1\le m\le M}\{trace(Q_m^2)\}^J \\
&= \xi_n^{-2J} M\left\{O\left(n^{-1}\widetilde{p}^2\right)\right\}^J = o(1),
\end{aligned} \tag{A16}$$

$$\begin{aligned}
T_2/2 &\le \sup_{\boldsymbol{w}\in\mathbb{H}_{\mathbb{M}}} \left\{\|e_W\|^2\rho^2\|P(\boldsymbol{w})\mu - \mu\|^2/R_{G_n}^2(\mathbf{w})\right\}^{1/2} \\
&\le \|e_W\|\rho\xi_n^{-1/2} = \xi_n^{-1/2}O\left(n^{-1/2}\widetilde{p}\right) = o(1),
\end{aligned} \tag{A17}$$

$$\begin{aligned}
T_4/2 &\le \xi_n^{-1}\rho\lambda_{\max}(\Omega) \sup_{\mathbf{w}\in\mathcal{H}_{\mathcal{M}}} [\mathrm{trace}\{P(\mathbf{w})\}] \\
&\le \xi_n^{-1}\rho\lambda_{\max}(\Omega) \max_m\{\mathrm{trace}(P_m)\} \\
&\le \xi_n^{-1}\rho\lambda_{\max}(\Omega) \max_m\{\lambda_{\max}(P_m)\} \max_m\{\mathrm{rank}(P_m)\} \\
&= \xi_n^{-1}O\left(n^{-1}\widetilde{p}^2\right) = \xi_n^{-1}O\left(n^{-1}\widetilde{p}^2\right) = o(1),
\end{aligned} \tag{A18}$$

$$\begin{aligned}
T_5 &\le \rho \sup_{\boldsymbol{w}\in\mathcal{H}_{\mathcal{M}}} \left[\{P(\boldsymbol{w})\hat{Y}_W - \mu\}^T\{P(\boldsymbol{w})\hat{Y}_W - \mu\}/R_{G_n}(\boldsymbol{w})\right] \\
&= \rho \sup_{\boldsymbol{w}\in\mathcal{H}_{\mathcal{M}}} [L_{G_n}(\boldsymbol{w})/R_{G_n}(\boldsymbol{w})] = O\left(n^{-1}\widetilde{p}\right).
\end{aligned} \tag{A19}$$

Therefore, combining (A15)–(A19), along with Condition (C7), it is clear that Theorem 2 holds. □

References

1. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. In *Second International Symposium on Information Theory*; Petrov, B., Csáki, F., Eds.; Akadémiai Kiadó: Budapest, Hungary, 1973; pp. 267–281.
2. Mallows, C.L. Some Comments on Cp. *Technometrics* **1973**, *15*, 661–675.
3. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 15–18. [CrossRef]

4. Hjort, N.L.; Claeskens, G. Frequentist Model Average Estimators. *J. Am. Stat. Assoc.* **2003**, *98*, 879–899. [CrossRef]
5. Buckland, S.T.; Burnham, K.P.; Augustin, N.H. Model selection: An integral part of inference. *Biometrics* **1997**, *53*, 603–618. [CrossRef]
6. Hansen, B.E. Least squares model averaging. *Econometrica* **2007**, *75*, 1175–1189. [CrossRef]
7. Wan, A.T.; Zhang, X.; Zou, G. Least squares model averaging by Mallows criterion. *J. Econom.* **2010**, *156*, 277–283. [CrossRef]
8. Hansen, B.E.; Racine, J.S. Jackknife model averaging. *J. Econom.* **2012**, *167*, 38–46. [CrossRef]
9. Liu, Q.; Okui, R. Heteroscedasticity-robust Cp model averaging. *Econom. J.* **2013**, *16*, 463–472. [CrossRef]
10. Zhao, S.; Zhang, X.; Gao, Y. Model averaging with averaging covariance matrix. *Econom. Lett.* **2016**, *145*, 214–217. [CrossRef]
11. Miller, R. Least square regression with censored data. *Biometrika* **1976**, *63*, 449–464. [CrossRef]
12. Buckley, J.; James, I. Linear regression with censored data. *Biometrika* **1979**, *66*, 429–436. [CrossRef]
13. Koul, H.; Susarla, V.; Van Ryzin, J. Regression analysis with randomly right-censored data. *Ann. Stat.* **1981**, *9*, 1276–1288. [CrossRef]
14. He, S.; Huang, X. Central limit theorem of linear regression model under right censorship. *Sci. China Ser. A-Math.* **2003**, *46*, 600–610. [CrossRef]
15. Wang, Q.; Dinse, G.E. Linear regression analysis of survival data with missing censoring indicators. *Lifetime Data Anal.* **2011**, *17*, 256–279. [CrossRef]
16. Liang, Z.Q.; Chen, X.L.; Zhou, Y.Q. Mallows model averaging estimation for linear regression model with right censored data. *Acta Math. Appl. Sin. Engl. Ser.* **2022**, *38*, 5–23. [CrossRef]
17. Wei, Y.; Wang, Q.; Liu, W. Model averaging for linear models with responses missing at random. *Ann. Inst. Stat. Math.* **2020**, *73*, 535–553. [CrossRef]
18. Liu, Q.; Okui, R.; Yoshimura, A. Generalized least squares model averaging. *Econom. Rev.* **2016**, *35*, 1692–1752. [CrossRef]
19. Albert, A.; Anderson, J.A. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **1984**, *71*, 1–10. [CrossRef]
20. Zhu, R.; Wan, A.T.; Zhang, X.; Zou, G. A Mallows-type model averaging estimator for the varyingcoecient partially linear model. *J. Am. Stat. Assoc.* **2019**, *114*, 882–892. [CrossRef]
21. Dong, Q.K.; Liu, B.X.; Zhao, H. Weighted least squares model averaging for accelerated failure time models. *Comput. Stat. Data Anal.* **2023**, *184*, 107743. [CrossRef]
22. Whittle, P. Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **1960**, *5*, 302–305. [CrossRef]
23. Qiu, Y.; Wang, W.; Xie, T.; Yu, J.; Zhang, X. Boosting Store Sales Through Machine Learning-Informed Promotional Decisions. 2023. Available online: http://www.mysmu.edu/faculty/yujun/Research/Maml_sales.pdf (accessed on 10 February 2024).