doi：10.13413/j.cnki.jdxblxb.2024058

带碎片协变量右删失数据的模型平均方法

王淑影，周丽芳，程云飞
（长春工业大学 数学与统计学院，长春 130012）

摘要：考虑在带有碎片协变量的右删失数据下比例风险模型的模型平均问题，先利用极大似然估计方法对模型中的参数进行估计，再采用基于信息准则的模型平均方法选取权重. 模拟结果表明，模型平均方法相比于模型选择方法预测精度更高. 并通过乳腺癌实例分析验证了该方法的优越性和可行性.

关键词：右删失数据；碎片协变量；模型平均；比例风险模型；信息准则
中图分类号：O212 **文献标志码**：A **文章编号**：1671-5489(2024)05-1091-11

Model Averaging Method for Right-Censored Data with Fragmentary Covariates

WANG Shuying, ZHOU Lifang, CHENG Yunfei
(*School of Mathematics and Statistics, Changchun University of Technology, Changchun* 130012, *China*)

Abstract: We considered the model averaging problem of the proportional hazard model in the right-censored data with fragmentary covariates. We first used the maximum likelihood estimation method to estimate the parameters in the model, and then used the model averaging method based on the information criterion to select the weights. The simulation results show that the model averaging method has higher prediction accuracy than the model selection method, and the superiority and feasibility of the proposed method are verified by the analysis of breast cancer examples.

Keywords: right-censored data; fragmentary covariate; model averaging; proportional hazard model; information criterion

在生存分析中，由于获得生存数据的实验设计、观测时间的局限，以及观测对象在进入或退出实验时个体差异等因素的影响，使得所关注的事件通常不能获得精确的观测时间，这类数据称为删失数据. 其中右删失数据是指在进行随访中，只能获取到个体的起始时间，无法准确观测到事件终点的时间，即个体生存时间未知，只已知大于观察时间. 比例风险模型(proportional hazards (PH) model)[1]是右删失数据回归问题中的一种常见模型，它可以同时考虑多种因素对个体生存时间的影响，且不同受试者组的危险率成比例，与时间无关，因此得到广泛关注[2-5].

在传统统计建模中，存在模型不确定性的问题，处理该类问题目前常用的方法是模型选择，通常先利用如 AIC(Akaike information criterion)和 BIC(Bayesian information criterion)等准则从候选模型

收稿日期：2024-02-07.

第一作者简介：王淑影(1990—)，女，汉族，博士，副教授，从事生物统计和数理统计的研究，E-mail：wangshuying0601@163.com. **通信作者简介**：周丽芳(1998—)，女，汉族，硕士，从事生物统计和模型平均的研究，E-mail：zhoulif1008@163.com.

基金项目：国家自然科学基金数学天元基金(批准号：12226416)、国家自然科学基金面上项目(批准号：12271060)和吉林省自然科学基金优秀青年基金(批准号：20230101371JC).

集中选出预测误差最小的单个模型，再对单个模型进行一系列的统计推断，但模型选择过程中存在不确定性，会严重影响建模的科学性和稳健性，降低预测精度. 为克服模型选择方法的不足，减少有用信息的遗失，一种有效的解决方法是模型平均，模型平均主要包括 Bayes 模型平均(Bayesian model averaging，BMA)和频率模型平均(frequentist model averaging，FMA)[6]. 目前，Bayes 模型平均方法已得到广泛关注，但其模型假设十分复杂，并难以从理论上证明其渐近性质，因此越来越多的研究者开始关注频率模型平均. 例如：Buckland 等[7]在基于 AIC 和 BIC 信息准则的基础上，提出了光滑的 AIC(S-AIC)和光滑的 BIC(S-BIC)模型平均方法；Hjort 等[8]考虑了建模偏差，在极大似然估计的框架下证明了频率模型平均的渐近性；Hansen[9]提出了基于 Mallows 准则的权重选择方法，从组合嵌套模型中获得最小二乘估计值；Deng 等[10]引入了一个新的模型选择标准，即 FIC(focused information criterion)；朱容等[11]研究了部分函数线性模型的模型平均方法，提出了该模型下最优权重的选择准则，并证明了模型平均估计量的渐近最优性.

上述模型平均方法均假设个体协变量都是完全观测到的，而近年来碎片数据应用越来越广泛，其主要特征是并非每个个体都有相同的协变量. 这种碎片数据在统计学中也称为分块缺失数据[12]. 处理这类数据最简单的方法是删除所有具有缺失值的样本，但这会丢弃大量有用的信息并极大减少分析中的样本数量. 因此，研究者们提出了各种插补方法，通过可用数据估计缺失值[13-15]，Lin 等[16]提出了迭代最小二乘估计(ILSE)，用于估计有个体特定缺失模式和高比例缺失数据的回归系数. Fang 等[17]研究表明，碎片数据中并非所有抽样对象都有相同的预测变量，并提出了一种基于频率模型平均的新方法；Yuan 等[18]基于这类碎片数据为经典 Mallows 模型平均(MMA)中的 Mallows 准则引入了偏差，提出了一种新的 Mallows 模型平均方法，并将该方法从线性回归模型推广到广义线性模型[19]. 基于上述研究结果，本文考虑在带有碎片协变量右删失数据的框架下，使用基于信息准则的模型平均方法对比例风险模型进行统计推断，以避免选择单一模型产生的误差，并为带碎片协变量删失数据的分析开辟一个新思路.

1 数据、模型及似然函数

本文主要考虑带有碎片协变量的右删失数据，其示例列于表 1. 随机样本由 n 个受试者组成，T 为生存时间，在生存分析中假设删失时间为 C，T 和 C 是独立的连续随机变量. 记 T_i 为个体 i 的生存时间，C_i 为个体 i 的删失时间. 个体观测时间 $\widetilde{T}_i=\min\{T_i,C_i\}$，$\delta=I(T_i\leqslant C_i)$是示性变量，$\delta=1$ 表示精确观测，否则为右删失. $D=\{X_j,\ j=1,2,\cdots,p\}$表示协变量集. 响应指标 $R=\{1,2,\cdots,K\}$，其中 $R=k$ $(k=1,2,\cdots,K)$表示可观察到协变量$\{X_j,\ j\in\Delta_k\}$，Δ_k 是 $D=\{1,2,\cdots,p\}$的子集，K 是所有响应变量类型的个数.

表 1 带碎片协变量右删失数据示例

Table 1 Examples of right-censored data with fragmentary covariates

个体	T	δ	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	R
1	*	*	*	*	*	*	*	*	*	*	*	1
2	*	*	*	*	*	*	*	*	*	*	*	1
3	*	*	*	*	*	*	*					2
4	*	*	*	*	*	*	*					2
5	*	*	*	*	*	*	*					2
6	*	*	*	*	*			*	*	*	*	3
7	*	*	*	*	*			*	*	*	*	3
8	*	*	*	*	*			*	*	*	*	3
9	*	*	*	*	*			*	*	*	*	3
10	*	*	*	*	*	*						4
11	*	*	*	*	*	*						4
12	*	*	*	*	*	*						4

注：* 表示可观测到的数据.

表 1 中有 $K=4$ 种响应模式，令 D_i 为受试者 i 观察到的协变量集合，则

$$D_1=D_2=\Delta_1=\{X_1,X_2,\cdots,X_9\}=\{1,2,\cdots,9\},$$
$$D_3=D_4=D_5=\Delta_2=\{X_1,X_2,X_3,X_4,X_5\}=\{1,2,3,4,5\},$$
$$D_6=D_7=D_8=D_9=\Delta_3=\{X_1,X_2,\cdots,X_9\}=\{1,2,3,6,7,8,9\},$$

以此类推. $\xi_k=\{i:D_i=\Delta_k\}$表示具有响应模式 $R=k$ 的个体集，因此$\{1,2,\cdots,n\}=\bigcup_{k=1}^{K}\xi_k$，且当 $k\neq l$ 时，$\xi_k\cap\xi_l=\varnothing$. $S_k=\{i:D_i\supseteq\Delta_k\}$表示可用协变量 Δ_k 的个体集，表 1 中，

$$\xi_1=\{1,2\},\quad S_1=\{1,2\},\quad \xi_2=\{3,4,5\},\quad S_2=\{1,2,3,4,5\},\quad \xi_3=\{6,7,8,9\},$$
$$S_3=\{1,2,6,7,8,9\},\quad \xi_4=\{10,11,12\},\quad S_4=\{1,2,3,4,5,10,11,12\}.$$

本文考虑带碎片协变量右删失数据下的比例风险模型：

$$h(t\mid\boldsymbol{X}_i)=h_0(t)\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i\},\tag{1}$$

其中 $i=1,2,\cdots,n$，$h_0(t)$表示任意的基准风险率，$h(t\mid\boldsymbol{X}_i)$表示第 i 个个体 t 时刻的风险率，$\boldsymbol{\beta}$ 为 p 维未知参数向量.

假设事件发生时间内不存在“结”，将患者的生存时间按增长的顺序排列：$t_1<t_2<\cdots<t_n$，定义时间 t_i 时的风险集 $R(t_i)$为$\{j:t_j\geqslant t_i\}$，它表示在 t_i 时刻前仍处于研究中所有个体的集合，d_i 表示在 t_i 时刻失效的个体数. 则模型(1)对应的生存函数为

$$S(t\mid\boldsymbol{X})=\{S_0(t)\}^{\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i\}},\tag{2}$$

其中 $S_0(t)=\prod_{t_i\leqslant t}\left\{1-\dfrac{d_i}{\sum_{j\in R(t_i)}\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_j\}}\right\}$.

进一步，若 $R(t_i)$中的某个个体在 t_i 时刻死亡，则具有协变量 $\boldsymbol{X}_i$ 的个体在 t_i 时刻死亡的条件概率为

$$p=\frac{h(t_i\mid\boldsymbol{X}_i)}{\sum_{j\in R(t_i)}h(t_i\mid\boldsymbol{X}_j)}=\frac{h_0(t_i)\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i\}}{\sum_{j\in R(t_i)}h_0(t_i)\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_j\}}=\frac{\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i\}}{\sum_{j\in R(t_i)}\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_j\}},\tag{3}$$

右删失数据下的偏似然函数为

$$L(\boldsymbol{\beta})=\prod_{i=1}^{n}\left(\frac{\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i\}}{\sum_{j\in R(t_i)}\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_j\}}\right)^{\delta_i},\tag{4}$$

对数似然函数为

$$\ln L(\boldsymbol{\beta})=\sum_{i=1}^{n}\left[\delta_i\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i-\delta_i\ln\Big(\sum_{j\in R(t_i)}\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_j\}\Big)\right],\tag{5}$$

其中 $\boldsymbol{\beta}$ 为未知参数. 对 $\boldsymbol{\beta}$ 求偏导数：

$$\frac{\partial(\ln L(\boldsymbol{\beta}))}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{n}\delta_i\left(\boldsymbol{X}_i-\frac{\sum_{j\in R(t_i)}\boldsymbol{X}_j\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_j\}}{\sum_{j\in R(t_i)}\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_j\}}\right),\tag{6}$$

由非线性方程$\dfrac{\partial(\ln L(\boldsymbol{\beta}))}{\partial\boldsymbol{\beta}}=0$ 可得 $\boldsymbol{\beta}$ 的极大似然估计.

下面基于 S-AIC(smoothed AIC)和 S-BIC(smoothed BIC)的模型平均方法[8]分别对包含所有协变量的个体($R=1$)和包含部分协变量的个体($R>1$)进行预测.

2　对 $R=1$ 的个体进行预测

在带有碎片协变量的右删失数据 $\Omega=\{(x_{ij},r_i,\delta_i),\ i=1,2,\cdots,n,\ j\in D_i\}$下对包含所有协变量的个体($R=1$)进行预测，其中 x_{ij}，r_i 分别表示对变量 $\boldsymbol{X}_i$ 和响应模式 R 的观测值，D_i 表示第 i 个个体观察到的协变量集合. 对 $R=1$ 的个体，考虑 K 个候选模型(即所有响应变量类型的个数)，第 k 个候选模型为

$$h(t_k \mid \boldsymbol{X}_{i,k}) = h_0(t_k)\exp\{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{X}_{i,k}\}, \tag{7}$$

其中 $k=1,2,\cdots,K$ 表示候选模型的个数，$\boldsymbol{\beta}_k$ 表示第 k 个候选模型中协变量的回归系数，$\boldsymbol{X}_{i,k}=(x_{ij}: i\in S_k, j\in\Delta_k)\in\mathbb{R}^{n_k\times p_k}$，且 $n_k=|S_k|$ 表示某种类型响应变量可用的个体数，$p_k=|\Delta_k|$ 表示某种类型响应变量包含的协变量个数. 则第 k 个候选模型下的对数似然函数为

$$\ln L_k(\boldsymbol{\beta}_k) = \sum_{i=1}^{n_k}\left[\delta_{i,k}\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{X}_{i,k} - \delta_{i,k}\ln\Big(\sum_{j\in R(t_{i,k})}\exp\{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{X}_{j,k}\}\Big)\right]. \tag{8}$$

再通过极大化对数似然函数获得 $\hat{\boldsymbol{\beta}}_k=(\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\beta}}_2,\cdots,\hat{\boldsymbol{\beta}}_{p_k})$，得到每个候选模型下的对数似然函数估计值为

$$\hat{l}_k = \sum_{i\in\xi_1}\left[\delta_{i,k}\hat{\boldsymbol{\beta}}_k^{\mathrm{T}}\boldsymbol{x}_{i,k} - \delta_{i,k}\ln\Big(\sum_{j\in R(t_{i,k})}\exp\{\hat{\boldsymbol{\beta}}_k^{\mathrm{T}}\boldsymbol{x}_{j,k}\}\Big)\right], \tag{9}$$

其中 $\boldsymbol{x}_{i,k}=(x_{ij}: i\in\xi_1, j\in\Delta_k)$.

实践中的建模策略是首先基于信息标准从 K 个候选模型中选择一个最合适的模型，然后使用选择的模型推断潜在的生存函数. 在模型选择中，最常用的是 AIC 和 BIC 准则，其表达式分别为

$$\mathrm{AIC}_k = -2\log\hat{l}_k + 2M_k, \tag{10}$$

$$\mathrm{BIC}_k = -2\log\hat{l}_k + M_k\log n_k, \tag{11}$$

其中 $\hat{l}_k$ 表示模型 k 的极大似然函数估计值，M_k 表示模型 k 中未知参数的个数. 根据 AIC_k 值和 BIC_k 值分别对所有模型进行排序，最小的 AIC_k 值和 BIC_k 值所对应的模型即为最优模型.

使用模型选择方法可能会遗失一些重要信息，导致模型预测结果不准确，为解决该问题，本文选用模型平均的方法得到参数估计值，然后采用基于 S-AIC 和 S-BIC 的模型平均方法计算组合权重：

$$\omega_k = \frac{\exp\{-x\mathrm{IC}_k/2\}}{\sum_k\exp\{-x\mathrm{IC}_k/2\}}, \tag{12}$$

其中 k 表示第 k 个候选模型，ω_k 是模型平均中第 k 个候选模型的权重，xIC 表示 AIC 或 BIC. 设 $\boldsymbol{\omega}=(\omega_1,\cdots,\omega_K)^{\mathrm{T}}$ 是 K 个模型的权重向量，并限制在如下集合中：

$$v = \Big\{\boldsymbol{\omega}\in[0,1]^K: 0\leqslant\omega_k\leqslant 1, \sum_{k=1}^{K}\omega_k = 1\Big\}, \tag{13}$$

其中 v 表示 K 个模型的权重向量集合. 则模型平均后参数 $\boldsymbol{\beta}$ 的估计值为

$$\hat{\boldsymbol{\beta}}_{\mathrm{MA}} = \sum_{k=1}^{K}\omega_k\hat{\boldsymbol{\beta}}_k. \tag{14}$$

3 对 $R>1$ 的个体进行预测

下面考虑对包含部分协变量的个体($R=l$)进行预测，即 $D^*=\Delta_l$，此时可用的协变量为 $\{\boldsymbol{X}_j, j\in\Delta_l\}$，并将不属于 Δ_l 的协变量除外，在基于 $\{\boldsymbol{X}_j, j\in\Delta_l\}$ 协变量的基础上进行模型平均.

为验证本文方法，考虑对表 1 中 $R=2$ 的个体进行预测，此时可用的协变量为 $D^*=\{X_1,X_2,X_3,X_4,X_5\}$，模型平均过程中将不属于 Δ_2 的协变量除外，因此产生一个新的碎片数据 $\Omega^{(2)}$，如表 2 所示. 表 2 中，

$$D_1^{(2)}=D_2^{(2)}=D_3^{(2)}=D_4^{(2)}=D_5^{(2)}=\{1,2,3,4,5\},$$

$$D_6^{(2)}=D_7^{(2)}=D_8^{(2)}=D_9^{(2)}=\{1,2,3\},$$

$$D_{10}^{(2)}=D_{11}^{(2)}=D_{12}^{(2)}=\{1,2,3,4\},\quad K^{(2)}=3,$$

$$\Delta_1^{(2)}=\{1,2,3,4,5\},\quad \Delta_2^{(2)}=\{1,2,3\},\quad \Delta_3^{(2)}=\{1,2,3,4\}.$$

因此当对 $R=2$ 的个体进行预测时，考虑 $K^{(2)}=3$ 个候选模型，第 k 个候选模型可用的协变量为 $\{x_{ij}: i\in S_k^{(2)}, j\in\Delta_k^{(2)}\}$，其中

$$S_1^{(2)}=\{1,2,3,4,5\},\quad S_2^{(2)}=\{1,2,3,4,5,6,7,8,9\},\quad S_3^{(2)}=\{1,2,3,4,5,10,11,12\}.$$

下面在给定碎片数据 $\Omega^{(l)}=\{(x_{ij}^{(l)},r_i^{(l)},\delta_i^{(l)}), i=1,2,\cdots,n, j\in D_i^{(l)}\}$ 下对 $R=l$ 的个体进行预测，其中 $D_i^{(l)}$ 表示第 i 个个体所观察到协变量 Δ_l 的集合. 对 $R=l$ 的个体，考虑 $K^{(l)}$ 个候选模型，第 k 个候选模型所用的协变量为 $\boldsymbol{X}_k^{(l)}=(x_{ij}^{(l)}: i\in S_k^{(l)}, j\in\Delta_k^{(l)})\in\mathbb{R}^{n_k^{(l)}\times p_k^{(l)}}$，其中 $S_k^{(l)}=\{i: D_i^{(l)}\supseteq\Delta_k^{(l)}\}$，

$n_k^{(l)}=|S_k^{(l)}|$ 表示某种类型的响应变量可用的个体数，$p_k^{(l)}=|\Delta_k^{(l)}|$ 表示某种类型的响应变量包含的协变量个数. 先用极大似然法得到第 k 个候选模型的参数估计值 $\hat{\boldsymbol{\beta}}_k^{(l)}$，然后针对 $R=l$ 的个体得到各候选模型下的对数似然函数值为

$$\hat{l}_k^{(l)}=\sum_{i\in\xi_l}\Big[\delta_{i,k}^{(l)}(\hat{\boldsymbol{\beta}}_k^{(l)})^{\mathrm{T}}\boldsymbol{x}_{i,k}^{(l)}-\delta_{i,k}^{(l)}\log\Big(\sum_{j\in R(t_{i,k}^{(l)})}\exp\{(\hat{\boldsymbol{\beta}}_k^{(l)})^{\mathrm{T}}\boldsymbol{x}_{j,k}^{(l)}\}\Big)\Big],\tag{15}$$

其中 $\boldsymbol{x}_{i,k}^{(l)}=(x_{ij}^{(l)}:i\in\xi_l,\ j\in\Delta_k^{(l)})$.

表 2　对 $R=2$ 的个体进行预测时所用的数据示例

Table 2　Examples of data used to predict individuals with $R=2$

个体	T	δ	X_1	X_2	X_3	X_4	X_5	R	$R^{(2)}$
1	*	*	*	*	*	*	*	1	1
2	*	*	*	*	*	*	*	1	1
3	*	*	*	*	*	*	*	2	1
4	*	*	*	*	*	*	*	2	1
5	*	*	*	*	*	*	*	2	1
6	*	*	*	*	*			3	2
7	*	*	*	*	*			3	2
8	*	*	*	*	*			3	2
9	*	*	*	*	*			3	2
10	*	*	*	*	*	*		4	3
11	*	*	*	*	*	*		4	3
12	*	*	*	*	*	*		4	3

注：* 表示可观测到的数据.

下面采用基于 S-AIC 和 S-BIC 的模型平均方法计算组合权重：

$$\omega_k^{(l)}=\frac{\exp\{-x\mathrm{IC}_k^{(l)}/2\}}{\sum_k\exp\{-x\mathrm{IC}_k^{(l)}/2\}},\tag{16}$$

其中 $\omega_k^{(l)}$ 表示对 $R=l$ 的个体进行预测时第 k 个候选模型的权重，k 表示第 k 个候选模型，xIC 表示 AIC 或 BIC. 设 $\boldsymbol{\omega}^{(l)}=(\omega_1^{(l)},\cdots,\omega_K^{(l)})^{\mathrm{T}}$ 是 K 个模型的权重向量，并限制在如下集合中：

$$v^{(l)}=\Big\{\boldsymbol{\omega}^{(l)}\in[0,1]^K:0\leqslant\omega_k^{(l)}\leqslant1,\ \sum_{k=1}^K\omega_k^{(l)}=1\Big\}.\tag{17}$$

则模型平均后参数 $\boldsymbol{\beta}^{(l)}$ 的估计值为

$$\hat{\boldsymbol{\beta}}_{\mathrm{MA}}^{(l)}=\sum_{k=1}^K\omega_k^{(l)}\hat{\boldsymbol{\beta}}_k^{(l)}.\tag{18}$$

4　模拟研究

下面用模拟研究验证模型平均方法的有效性，并将其与模型选择方法进行比较. 数据 T_i 由以下模型生成：

$$h(t\,|\,\boldsymbol{x}_i)=h_0(t)\exp\{\boldsymbol{\beta}_1^{\mathrm{T}}x_{i1}+\boldsymbol{\beta}_2^{\mathrm{T}}x_{i2}+\cdots+\boldsymbol{\beta}_i^{\mathrm{T}}x_{ij}\},\tag{19}$$

其中 $i\in\xi_k$，$n_k=|\xi_k|$ 表示某种类型响应变量包含的个体数，基准风险函数设为 $h_0(t)=0.04t$，$j\in\Delta_k$，$p_k=|\Delta_k|$ 表示某种类型响应变量包含的协变量数，这里考虑共有 8 个协变量，即 $j=8$，且变量 $(x_{i1},x_{i2},\cdots,x_{ij})$ 由 $E(x_{ij})=0$，$\mathrm{Var}(x_{ij})=1$ 的标准正态分布生成. 本文考虑 3 种情形，在每种情形下 8 个协变量均分成 4 组，则 $K=8$，并对每种类型的个体分别进行预测. 令 C 服从均匀分布 $U\sim(0,c)$，其中 c 控制删失率，在样本量为 300 和 600 的条件下令删失比例分别约为 10% 和 45%，模拟循环 1 000 次.

情形 1）仅包含 1 个必选协变量. 第一组包含的协变量仅有变量 X_1，并始终可用，第二组包含的协变量为 X_2,X_3，第三组包含的协变量为 X_4,X_5，第四组包含的协变量为 X_6,X_7,X_8. 当 $X_2<0.3$，$X_4<0.3$，$X_6<0.3$ 时，第二、三、四组包含的协变量分别可用. 必选协变量的参数真值均设为 0.4，

其他候选协变量的参数真值均设为0.1.

情形2) 包含2个必选协变量. 第s组包含的协变量为$X_{2(s-1)+1}\sim X_{2s}$($s=1,2,3,4$),第一组协变量始终可用. 当$X_3<0.3$,$X_5<0.3$,$X_7<0.3$时,第二、三、四组包含的协变量分别可用. 必选协变量的参数真值均设为0.4,其他候选协变量的参数真值均设为0.1.

情形3) 包含3个必选协变量. 第一组包含的协变量为X_1,X_2,X_3,并始终可用,第二组包含的协变量为X_4,X_5,第三组包含的协变量为X_6,X_7,第四组包含的协变量为X_8. 当$X_4<0.3$,$X_6<0.3$,$X_8<0.3$时,第二、三、四组包含的协变量分别可用. 必选协变量的参数真值均设为0.4,其他候选协变量的参数真值设为0.1,0.1,0.1,-0.1,0.2.

在模拟研究中,先分别计算每种情形下各候选模型的AIC和BIC值,并利用S-AIC和S-BIC法计算每种情形下各候选模型的权重,再计算模拟循环1 000次后不同方法感兴趣指标的均方根误差(root mean squared error, RMSE),即

$$\mathrm{RMSE}^{(l)}=\Big[(1\ 000)^{-1}\sum_{j=1}^{1\ 000}\Big(\sum_{k=1}^{K^{(l)}}\hat{\omega}_{kj}^{(l)}\hat{\vartheta}_{kj}^{(l)}-\vartheta_{kj}^{(l)}\Big)^2\Big]^{1/2},\tag{20}$$

其中$K^{(l)}$表示对第l个个体进行预测时候选模型的个数,$\hat{\vartheta}_{kj}^{(l)}$表示第$l$个个体第$j$次循环中感兴趣参数的估计值,$\vartheta_{kj}^{(l)}$表示第$l$个个体第$j$次循环中感兴趣参数的真值,$\hat{\omega}_{kj}^{(l)}$表示对第$l$个个体进行预测时第$k$个候选模型第$j$次循环的权重.

本文主要考虑两个指标:估计协变量参数的欧氏距离$\|\alpha\|=\sqrt{\beta_1^2+\beta_2^2\cdots+\beta_8^2}$和当协变量为0.2、时间为5时的生存概率$S(\cdot)=\{S_0(t)\}^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i)}$. 最终模拟结果分别列于表3~表5.

表3 情形1)下AIC,BIC,S-AIC,S-BIC方法的RMSE

Table 3 RMSE of AIC, BIC, S-AIC and S-BIC methods in case 1)

指标	个体R	$n=300$,删失比10%				$n=300$,删失比45%			
		S-AIC	S-BIC	AIC	BIC	S-AIC	S-BIC	AIC	BIC
$\|\alpha\|$	1	0.108 9*	0.114 6	0.115 4	0.115 4	0.119 1	0.119 0*	0.119 9	0.119 9
	2	0.083 9*	0.089 0	0.090 5	0.090 6	0.092 9*	0.098 0	0.099 9	0.099 7
	3	0.093 0*	0.097 6	0.098 4	0.098 6	0.100 7*	0.104 9	0.105 7	0.105 9
	4	0.092 7*	0.097 6	0.098 4	0.098 6	0.099 3*	0.104 7	0.105 7	0.105 9
	5	0.074 3*	0.075 5	0.076 7	0.076 6	0.088 0*	0.088 8	0.089 8	0.089 8
	6	0.073 2*	0.075 3	0.076 5	0.076 6	0.087 3*	0.088 7	0.089 9	0.089 8
	7	0.081 0*	0.082 8	0.083 5	0.083 2	0.092 8*	0.093 9	0.094 5	0.094 2
$S(\cdot)$	1	0.093 2*	0.096 5	0.096 8	0.096 9	0.093 3*	0.096 8	0.097 2	0.097 3
	2	0.059 9*	0.063 6	0.064 4	0.064 6	0.060 6*	0.064 1	0.065 0	0.065 1
	3	0.067 5*	0.070 5	0.071 1	0.071 1	0.068 2*	0.071 1	0.071 8	0.071 7
	4	0.066 8*	0.069 7	0.070 2	0.070 3	0.067 5*	0.070 3	0.070 9	0.070 9
	5	0.033 1*	0.034 5	0.035 7	0.035 7	0.033 9*	0.035 1	0.036 2	0.036 1
	6	0.033 6*	0.035 4	0.036 4	0.036 4	0.034 7*	0.036 2	0.037 2	0.037 1
	7	0.040 3*	0.041 2	0.041 6	0.041 6	0.040 7*	0.041 8	0.042 0	0.042 0
指标	个体R	$n=600$,删失比10%				$n=600$,删失比45%			
		S-AIC	S-BIC	AIC	BIC	S-AIC	S-BIC	AIC	BIC
$\|\alpha\|$	1	0.105 2*	0.109 4	0.109 6	0.109 7	0.103 8*	0.109 7	0.109 9	0.110 1
	2	0.075 7*	0.080 7	0.081 2	0.081 3	0.077 7*	0.083 3	0.084 0	0.084 2
	3	0.086 4*	0.090 4	0.090 5	0.090 7	0.087 9*	0.092 2	0.092 5	0.092 6
	4	0.086 4*	0.090 3	0.090 6	0.090 7	0.087 9*	0.092 2	0.092 6	0.092 6
	5	0.060 7*	0.062 9	0.063 7	0.063 6	0.066 3*	0.068 4	0.069 1	0.069 1
	6	0.061 1*	0.063 1	0.063 5	0.063 6	0.066 4*	0.068 5	0.068 8	0.069 1
	7	0.070 2*	0.072 1	0.072 1	0.072 2	0.074 3*	0.076 1	0.076 2	0.076 2

续表 3

Continued to table 3

指标	个体 R	$n=600$，删失比 10%				$n=600$，删失比 45%			
		S-AIC	S-BIC	AIC	BIC	S-AIC	S-BIC	AIC	BIC
$S(\cdot)$	1	0.093 5*	0.097 0	0.097 1	0.097 3	0.093 4*	0.097 3	0.097 5	0.097 5
	2	0.059 3*	0.063 5	0.063 9	0.064 0	0.059 4*	0.063 8	0.064 0	0.064 4
	3	0.067 6*	0.070 6	0.070 7	0.070 9	0.067 6*	0.070 8	0.071 0	0.071 1
	4	0.067 4*	0.070 4	0.070 6	0.070 7	0.067 4*	0.070 5	0.070 6	0.070 8
	5	0.032 2*	0.034 3	0.034 9	0.034 9	0.032 3*	0.034 4	0.035 0	0.035 0
	6	0.031 3*	0.033 6	0.034 1	0.034 1	0.031 9*	0.034 2	0.034 7	0.034 7
	7	0.039 3*	0.040 5	0.040 6	0.040 6	0.039 7*	0.041 0	0.041 1	0.041 1

注：标 * 数字表示最优估计结果.

表 4　情形 2）下 AIC，BIC，S-AIC，S-BIC 方法的 RMSE

Table 4　RMSE of AIC，BIC，S-AIC and S-BIC methods in case 2）

指标	个体 R	$n=300$，删失比 10%				$n=300$，删失比 45%			
		S-AIC	S-BIC	AIC	BIC	S-AIC	S-BIC	AIC	BIC
$\|\alpha\|$	1	0.090 1*	0.094 8	0.095 6	0.095 6	0.096 7*	0.101 3	0.102 3	0.102 3
	2	0.080 9*	0.084 3	0.085 6	0.085 4	0.091 4*	0.093 6	0.095 3	0.094 7
	3	0.080 7*	0.084 3	0.085 2	0.085 4	0.091 4*	0.093 7	0.094 7	0.094 7
	4	0.080 8*	0.084 3	0.085 5	0.085 4	0.090 5*	0.093 5	0.094 5	0.094 7
	5	0.075 4*	0.076 0	0.077 4	0.077 4	0.089 1*	0.089 1	0.089 4	0.089 4
	6	0.075 8	0.076 6	0.077 2	0.077 2	0.087 9*	0.088 7	0.089 4	0.089 4
	7	0.075 1	0.076 4	0.077 3	0.077 2	0.086 2*	0.088 9	0.089 4	0.089 4
$S(\cdot)$	1	0.090 5*	0.094 5	0.095 1	0.095 2	0.091 2*	0.095 8	0.096 6	0.096 6
	2	0.062 5*	0.066 2	0.067 2	0.067 3	0.062 8*	0.066 6	0.067 7	0.067 7
	3	0.091 6*	0.095 0	0.095 7	0.095 9	0.091 9*	0.095 1	0.096 1	0.096 0
	4	0.060 6*	0.066 3	0.067 8	0.067 9	0.060 6*	0.066 2	0.067 9	0.067 9
	5	0.034 5*	0.035 7	0.037 2	0.037 1	0.035 9*	0.036 9	0.038 3	0.038 3
	6	0.034 8*	0.036 4	0.037 3	0.037 3	0.035 2*	0.036 9	0.037 9	0.037 9
	7	0.034 3*	0.036 1	0.037 0	0.037 1	0.037 0*	0.038 4	0.039 1	0.039 1
指标	个体 R	$n=600$，删失比 10%				$n=600$，删失比 45%			
		S-AIC	S-BIC	AIC	BIC	S-AIC	S-BIC	AIC	BIC
$\|\alpha\|$	1	0.078 3*	0.082 0	0.082 2	0.082 3	0.082 5*	0.086 8	0.084 5	0.087 2
	2	0.065 3*	0.068 8	0.069 2	0.069 3	0.072 2*	0.075 8	0.075 9	0.076 3
	3	0.065 2*	0.068 8	0.069 3	0.069 3	0.072 3*	0.075 8	0.076 2	0.076 3
	4	0.065 5*	0.068 8	0.069 1	0.069 3	0.072 0*	0.075 8	0.076 4	0.076 3
	5	0.055 8*	0.057 2	0.057 9	0.057 9	0.065 8*	0.066 9	0.067 5	0.067 5
	6	0.056 4*	0.057 5	0.057 9	0.057 9	0.066 4*	0.067 2	0.067 9	0.067 5
	7	0.056 2*	0.057 5	0.057 7	0.057 9	0.065 5*	0.067 1	0.067 6	0.067 5
$S(\cdot)$	1	0.091 0*	0.094 5	0.094 6	0.094 7	0.091 2*	0.095 3	0.096 0	0.095 7
	2	0.062 2*	0.066 4	0.066 7	0.066 9	0.062 2*	0.066 6	0.067 0	0.067 2
	3	0.090 5*	0.094 2	0.094 4	0.094 7	0.091 1*	0.094 9	0.095 1	0.095 3
	4	0.059 6*	0.065 9	0.066 6	0.066 8	0.059 6*	0.066 0	0.066 7	0.066 8
	5	0.033 6*	0.035 6	0.036 4	0.036 4	0.033 9*	0.035 7	0.036 5	0.036 5
	6	0.033 0*	0.035 2	0.035 6	0.035 7	0.032 9*	0.035 2	0.035 7	0.035 7
	7	0.033 1*	0.035 2	0.035 7	0.035 7	0.033 8*	0.036 1	0.036 5	0.036 5

注：标 * 数字表示最优估计结果.

表5 情形3)下AIC,BIC,S-AIC,S-BIC方法的RMSE

Table 5 RMSE of AIC,BIC,S-AIC and S-BIC methods in case 3)

指标	个体 R	$n=300$，删失比10%				$n=300$，删失比45%			
		S-AIC	S-BIC	AIC	BIC	S-AIC	S-BIC	AIC	BIC
$\|\alpha\|$	1	0.099 9*	0.105 1	0.107 8	0.107 6	0.102 3*	0.107 1	0.111 3	0.110 0
	2	0.085 2*	0.088 9	0.089 9	0.090 2	0.093 8*	0.095 6	0.096 7	0.096 7
	3	0.091 2*	0.094 8	0.098 6	0.098 6	0.098 0*	0.099 7	0.105 6	0.102 7
	4	0.091 2*	0.095 0	0.099 2	0.098 7	0.096 8*	0.099 2	0.102 7	0.102 7
	5	0.081 4*	0.081 2	0.083 5	0.083 4	0.091 2*	0.091 7	0.092 5	0.092 2
	6	0.081 4*	0.082 5	0.083 7	0.083 5	0.091 6*	0.091 7	0.092 1	0.092 1
	7	0.086 4*	0.087 2	0.091 4	0.090 7	0.095 0	0.094 4*	0.098 5	0.097 7
$S(\cdot)$	1	0.094 0*	0.098 8	0.100 9	0.101 7	0.094 1*	0.099 4	0.101 7	0.102 5
	2	0.055 0*	0.057 0	0.058 1	0.057 7	0.056 3*	0.058 2	0.059 4	0.058 8
	3	0.078 2*	0.084 1	0.088 3	0.089 0	0.078 2*	0.083 9	0.088 3	0.088 9
	4	0.063 4*	0.068 2	0.071 8	0.072 8	0.065 0*	0.069 6	0.073 9	0.074 1
	5	0.038 8*	0.039 3	0.041 5	0.041 4	0.039 2*	0.039 9	0.041 6	0.041 5
	6	0.025 1*	0.025 1	0.025 5	0.025 3	0.028 8	0.028 5*	0.028 6	0.028 6
	7	0.029 6*	0.032 3	0.038 3	0.038 2	0.032 1*	0.033 7	0.039 9	0.038 9
指标	个体 R	$n=600$，删失比10%				$n=600$，删失比45%			
		S-AIC	S-BIC	AIC	BIC	S-AIC	S-BIC	AIC	BIC
$\|\alpha\|$	1	0.090 6*	0.095 7	0.097 0	0.097 2	0.091 1*	0.096 6	0.098 0	0.098 3
	2	0.071 1*	0.074 6	0.075 2	0.075 1	0.076 7*	0.079 4	0.080 2	0.079 9
	3	0.077 1*	0.082 7	0.085 1	0.085 8	0.081 5*	0.086 0	0.088 9	0.088 6
	4	0.076 4*	0.082 4	0.084 7	0.085 6	0.080 9*	0.085 7	0.088 1	0.088 6
	5	0.063 4*	0.064 5	0.065 3	0.065 2	0.070 9*	0.071 8	0.072 5	0.072 6
	6	0.063 2*	0.064 8	0.065 1	0.065 2	0.071 4*	0.072 3	0.072 7	0.072 6
	7	0.068 0*	0.071 1	0.074 8	0.075 1	0.075 4*	0.077 1	0.079 8	0.080 2
$S(\cdot)$	1	0.095 2*	0.101 0	0.102 6	0.102 6	0.094 7*	0.100 7	0.102 3	0.102 8
	2	0.054 1*	0.056 8	0.057 8	0.057 2	0.054 4*	0.057 1	0.057 9	0.057 6
	3	0.078 8*	0.085 5	0.087 4	0.088 9	0.077 7*	0.085 3	0.087 9	0.089 2
	4	0.062 3*	0.068 5	0.070 6	0.071 8	0.062 8*	0.069 4	0.072 1	0.073 0
	5	0.035 3*	0.037 1	0.038 1	0.038 1	0.035 8*	0.037 4	0.038 4	0.038 4
	6	0.021 3*	0.022 1	0.022 3	0.022 2	0.021 3*	0.022 1	0.022 3	0.022 3
	7	0.026 7*	0.031 1	0.036 0	0.036 1	0.028 4*	0.032 8	0.037 6	0.037 6

注：标*数字表示最优估计结果.

表3～表5分别列出了4种方法S-AIC,S-BIC,AIC,BIC对不同类型个体的预测结果，由于对3种情形中第8种类型的个体进行预测时，候选模型的个数均仅有一个，无法体现模型平均方法的优势，因此在模拟中不考虑对这类个体的预测. 其中，S-AIC和S-BIC为模型平均方法，AIC和BIC为模型选择方法. 由表3～表5可得以下结论：

1) 无论哪种情形，基于S-AIC和S-BIC模型平均方法的RMSE值在大多数情况下都比模型选择方法的RMSE值小，说明S-AIC和S-BIC的模型平均方法优于基于AIC和BIC的模型选择方法.

2) 由模拟结果可见，S-AIC比S-BIC的RMSE值普遍小，表明基于S-AIC的模型平均方法优于基于S-BIC的模型平均方法.

3) 当样本量不变，增加删失比时，基于模型平均和模型选择方法两种指标下的RMSE值在大多数情况下有一定幅度的增大. 而当删失比不变，增加样本量时，上述4种方法下各指标的RMSE值基本都逐渐减小，表明参数估计值更接近真值，不同类型个体的生存概率预测值也更接近真实的生存概率值，说明这4种方法随着样本量的增加估计效果都变得更好，估计精度均有提高.

模拟结果表明，无论在哪种情形下，基于S-AIC和S-BIC的模型平均方法均比基于AIC和BIC的

模型选择方法更有优势，这主要是因为模型平均方法考虑了所有协变量的信息，避免了模型选择过程带来的不确定性.

5　实例分析

下面利用 Schumacher 等[20]分析的乳腺癌数据集进行实例分析，以进一步验证模型平均方法相比于模型选择方法的优越性. 该数据来自一项原发性乳腺癌实验：从 1984 年 7 月到 1989 年 12 月，德国乳腺癌研究组(German breast cancer study group，GBSG)招募了 686 例原发性淋巴结阳性乳腺癌患者，以研究乳腺癌的治疗和临床试验中的重要预后因素.

该数据集可在 R 软件 survival 包中找到，原始数据集中共有 686 名患者，10 个变量分别为 rfstime(患者的生存时间)、status(生存状态)、age(年龄)、meno(更年期状态，0 表示更年期前，1 表示更年期后)、tsize(肿瘤大小/mm)、tgrade(肿瘤水平因子，水平 1＜水平 2＜水平 3)、pnodes(正节点个数)、pgr(孕酮受体个数)、er(雌激素受体个数)、horTH(是否进行激素治疗). 本文考虑其中 8 个变量(age，meno，tsize，tgrade，pnodes，pgr，er，horTH)预后因素的相对重要性. 结合文献[20]的研究发现，pnodes，pgr，horTH 3 个变量对乳腺癌疾病有重要影响，因此本文将上述变量作为必选协变量，并对连续性变量(age，tsize，pnodes，pgr，er)进行标准化处理. 同时，考虑到原始数据集中不存在任何缺失值，因此设置随机数种子并选择随机删除部分数据使其变得碎片化，缺失率为 40%，最终产生 $2^5=32$ 个候选模型.

本文主要对 $D=$(age，meno，tsize，tgrade，pnodes，pgr，er，horTH)的个体进行分析，分别用 S-AIC，S-BIC，AIC 和 BIC 方法计算每个变量的系数估计值及其置信水平为 95%的置信区间，由于在模型选择过程中 AIC 和 BIC 均选择了同一个候选模型，因此将这两种方法的结果合并在同一列中，所得结果列于表 6.

表 6　乳腺癌数据集的估计及置信区间

Table 6　Estimation and confidence intervals of breast cancer dataset

变量	估计	S-AIC	S-BIC	AIC/BIC
age	估计值(标准差)	−0.015 8(0.039 5)	−0.008 2(0.142 0)	—
	95%置信区间	(−0.093 2，0.061 6)	(−0.036 0，0.019 6)	—
meno	估计值(标准差)	0.046 3(0.061 0)	0.021 1(0.021 5)	—
	95%置信区间	(−0.073 2，0.165 8)	(−0.073 2，0.165 8)	—
tsize	估计值(标准差)	0.010 3(0.033 5)	0.005 3(0.012 5)	—
	95%置信区间	(−0.055 4，0.079 5)	(−0.019 3，0.029 8)	—
tgrade	估计值(标准差)	0.034 9(0.021 4)	0.013 8(0.007 2)	—
	95%置信区间	(−0.007 0，0.076 7)	(−0.000 3，0.027 9)	—
pnodes	估计值(标准差)	0.314 8(0.057 3)	0.309 0(0.044 7)	0.306 2(0.036 6)
	95%置信区间	(0.202 5，0.427 2)	(0.221 4，0.396 6)	(0.234 5，0.377 9)
pgr	估计值(标准差)	−0.619 7(0.191 0)	−0.565 0(0.144 5)	−0.531 4(0.115 4)
	95%置信区间	(−0.994 2，0.245 3)	(−0.848 3，−0.282 8)	(−0.757 6，−0.305 2)
er	估计值(标准差)	−0.004 3(0.002 7)	−0.000 4(0.011 7)	—
	95%置信区间	(−0.058 5，0.049 9)	(−0.023 4，0.022 5)	—
horTH	估计值(标准差)	−0.274 9(0.181 2)	−0.302 5(0.147 5)	−0.328 0(0.124 9)
	95%置信区间	(−0.630 1，0.080 3)	(−0.591 7，−0.013 4)	(−0.572 7，−0.083 2)

注："—"表示 AIC，BIC 不选择该变量.

由表 6 可见：变量 pnodes 系数估计值较大，表明正节点数与乳腺癌的发生呈正相关，即正节点数越多，患乳腺癌的风险越大；变量 pgr 和 horTH 的系数估计值为负值，表明孕酮受体个数越多患乳腺癌的可能性越小，接受激素治疗也可降低患乳腺癌的风险. 这与 Sauerbrei 等[21]的研究结果一致. 此外，Sauerbrei 等[21]和 Royston 等[22]的研究充分肯定了变量 age，meno，tsize，tgrade，er 对患者患乳腺癌风险的影响. Sauerbrei 等[21]指出 40 岁前，患者年龄越小，患乳腺癌的风险越高，同时，肿瘤越大，

肿瘤水平越高，患乳腺癌的可能性也越大. Royston 等[22]研究表明，患者处在更年期或者雌激素受体个数越少时患乳腺癌的风险也越高. 而在基于 AIC 和 BIC 的模型选择过程中均不包括变量 age，meno，tsize，tgrade，er，说明模型选择过程中遗失了重要信息的影响.

为比较 4 种方法 AIC，BIC，S-AIC 和 S-BIC 的预测性能，本文从每种类型的个体中分别依次随机抽取 75%，80%，85%的个体，将其组合成训练数据进行模型拟合，再利用极大似然估计方法对未知参数进行估计. 将剩余个体作为测试数据进行预测，其中训练集数据的样本量设为 n_0，所占比例为 π，则测试集的样本量为 $n-n_0$，所占比例为 $1-\pi$，n 为整体样本量. 然后，使用基于信息准则的模型选择和模型平均方法对 $D=$(age，meno，tsize，tgrade，pnodes，pgr，er，horTH)的个体生存概率进行预测，对该过程循环 500 次，并计算生存概率预测值的均值、中位数和标准差，结果列于表 7.

表 7 乳腺癌数据集生存概率预测值

Table 7 Predicted values of survival probability of breast cancer dataset

π	指标	S-AIC	S-BIC	AIC	BIC
75%	均值	0.561 8	0.554 3	0.561 5	0.546 9
	中位数	0.568 8	0.562 2	0.571 1	0.552 5
	标准差	0.134 3	0.136 9	0.137 1	0.138 8
80%	均值	0.559 9	0.552 7	0.555 1	0.545 4
	中位数	0.563 0	0.557 3	0.560 2	0.548 9
	标准差	0.150 9	0.153 4	0.156 6	0.156 3
85%	均值	0.563 1	0.558 5	0.551 5	0.543 7
	中位数	0.565 6	0.561 6	0.557 7	0.549 7
	标准差	0.173 2	0.176 1	0.183 4	0.182 2

由表 7 可见，随着训练集样本量的增加，基于模型平均方法预测得到的患者生存概率值标准差低于基于模型选择方法得到的生存概率值标准差，其中基于 S-AIC 的模型平均方法小于基于 S-BIC 的模型平均方法预测得到的生存概率值标准差，表明基于 S-AIC 的模型平均方法预测结果更稳健. 这主要是因为模型平均方法考虑了所有变量信息，而模型选择方法只考虑了部分变量信息. 因此，基于 S-AIC 和 S-BIC 的模型平均方法估计稳健性更强.

综上，本文主要研究了带有碎片协变量的右删失数据下比例风险模型的模型平均方法，先使用极大似然估计法对模型中的未知参数进行估计，再使用基于信息准则的模型平均方法对候选模型的权重进行计算. 模拟和实例研究结果表明，基于 S-AIC 和 S-BIC 的模型平均方法普遍优于基于 S-AIC 和 S-BIC 的模型选择方法，其中 S-AIC 方法估计效果更好，这主要是因为模型平均考虑了所有变量的信息，而模型选择只考虑了部分协变量的影响. 在传统的模型平均方法中，一般在样本量相同的情况下根据协变量的个数构建候选模型，而本文提出的方法各候选模型在参数估计时使用不同的样本量，且候选模型的个数依赖响应变量类型的个数，同时对每种类型的个体分别进行预测，当个体类型数较多时，可能使计算更复杂.

参 考 文 献

[1] COX D R. Regression Models and Life-Tables [J]. Journal of the Royal Statistical Society: Series B (Methodological), 1972, 34(2): 187-202.

[2] GENTLEMAN R, CROWLEY J. Local Full Likelihood Estimation for the Proportional Hazards Model [J]. Biometrics, 1991, 47(4): 1283-1296.

[3] GU C. Penalized Likelihood Hazard Estimation: A General Procedure [J]. Statistica Sinica, 1996, 6(4): 861-876.

[4] FAN J Q, GIJBELS I, KING M. Local Likelihood and Local Partial Likelihood in Hazard Regression [J]. The Annals of Statistics, 1997, 25(4): 1661-1690.

[5] SLEEPER L A, HARRINGTON D P. Regression Splines in the Cox Model with Application to Covariate Effects in Liver Disease [J]. Journal of the American Statistical Association, 1990, 85: 941-949.

[6] 张新雨，邹国华. 模型平均方法及其在预测中的应用 [J]. 统计研究，2011，28(6)：97-102. (ZHANG X Y，ZOU G H. Model Averaging Method and Its Application in Forecast [J]. Statistical Research，2011，28(6)：97-102.)

[7] BUCKLAND S T，BURNHAM K P，AUGUSTIN N H. Model Selection：An Integral Part of Inference [J]. Biometrics，1997，53(2)：603-618.

[8] HJORT N L，CLAESKENS G. Frequentist Model Average Estimators [J]. Journal of the American Statistical Association，2003，98：879-899.

[9] HANSEN B E. Least Squares Model Averaging [J]. Econometrica，2007，75(4)：1175-1189.

[10] DENG G H，LIANG H. Model Averaging for Semiparametric Additive Partial Linear Models [J]. Science China Mathematics，2010，53(5)：1363-1376.

[11] 朱容，邹国华，张新雨. 部分函数线性模型的模型平均方法 [J]. 系统科学与数学，2018，38(7)：777-800. (ZHU R，ZOU G H，ZHANG X Y. Optimal Model Averaging Estimation for Partial Functional Linear Models [J]. Journal of Systems Science and Mathematical Sciences，2018，38(7)：777-800.)

[12] FANG F，BAO S L. FragmGAN：Generative Adversarial Nets for Fragmentary Data Imputation and Prediction [J]. Statistical Theory and Related Fields，2024，8(1)：15-28.

[13] SCHNEIDER T. Analysis of Incomplete Climate Data：Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values [J]. Journal of Climate，2001，14(5)：853-871.

[14] SCHOTT J M，BARTLETT J W，BARNES J，et al. Reduced Sample Sizes for Atrophy Outcomes in Alzheimer's Disease Trials：Baseline Adjustment [J]. Neurobiology of Aging，2010，31(8)：1452-1462.

[15] ZHU X F，ZHANG S C，JIN Z，et al. Missing Value Estimation for Mixed-Attribute Data Sets [J]. IEEE Transactions on Knowledge and Data Engineering，2010，23(1)：110-121.

[16] LIN H Z，LIU W，LAN W. Regression Analysis with Individual-Specific Patterns of Missing Covariates [J]. Journal of Business & Economic Statistics，2021，39(1)：179-188.

[17] FANG F，LAN W，TONG J J，et al. Model Averaging for Prediction with Fragmentary Data [J]. Journal of Business & Economic Statistics，2019，37(3)：517-527.

[18] YUAN C X，FANG F，NI L. Mallows Model Averaging with Effective Model Size in Fragmentary Data Prediction [J]. Computational Statistics & Data Analysis，2022，173(9)：107497-1-107497-18.

[19] YUAN C X，WU Y，FANG F. Model Averaging for Generalized Linear Models in Fragmentary Data Prediction [J]. Statistical Theory and Related Fields，2022，6(4)：344-352.

[20] SCHUMACHER M，BASTERT G，BOJAR H，et al. Randomized 2×2 Trial Evaluating Hormonal Treatment and the Duration of Chemotherapy in Node-Positive Breast Cancer Patients. German Breast Cancer Study Group [J]. Journal of Clinical Oncology，1994，12(10)：2086-2093.

[21] SAUERBREI W，ROYSTON P，BOJAR H，et al. Modelling the Effects of Standard Prognostic Factors in Node-Positive Breast Cancer [J]. British Journal of Cancer，1999，79(11/12)：1752-1760.

[22] ROYSTON P，ALTMAN D G. External Validation of a Cox Prognostic Model：Principles and Methods [J]. BMC Medical Research Methodology，2013，13：31-1-31-15.

(责任编辑：李　琦)