第 46 卷第 3 期　　应用数学学报　　Vol. 46 No. 3
2023 年 5 月　　ACTA MATHEMATICAE APPLICATAE SINICA　　May, 2023

加速失效时间模型下现状数据的Jackknife模型平均*

赵 慧　　刘斌霞†　　董庆凯

(中南财经政法大学统计与数学学院, 武汉 430070)

(E-mail: lbx@stu.zuel.edu.cn)

张新雨

(中国科学院数学与系统科学研究院, 北京 100190)

摘 要　文章研究了响应变量为现状数据的情况下, 加速失效时间模型的Jackknife模型平均方法. 首先对数据进行合理的无偏变换, 进而得到回归参数的最小二乘估计. 然后引入删一交叉验证准则来选取候选模型的权重, 并在一定正则性条件下, 建立对应模型平均估计量的渐近最优性. 此外, 数值模拟表明, 与现有的其他模型平均和模型选择方法相比, 本文所提出的方法在预测上表现更佳. 最后将所提方法应用于尼日利亚儿童死亡率的数据进行实证研究, 进一步验证了所提方法的优良性质.

关键词　现状数据; 加速失效时间模型; 无偏变换; 模型平均; 渐近最优

MR(2023)主题分类　62N01; 62N02; 62F12

中图分类　O212.1

1 引言

模型选择和模型平均是两种常用的处理模型不确定性及提高预测准确性的方法. 其中, 模型选择方法试图从一系列候选模型集合中选出一个最优模型, 比较流行的方法包括: 赤池信息准则(AIC), 贝叶斯信息准则(BIC), Mallows' Cp 准则等. 不同于模型选择方法, 模型平均方法不选定某个最优模型, 而是给所有候选模型赋予一定权重从而得到模型平均估计. 模型平均方法不丢弃任何一个候选模型, 从而也不丢失原始数据信息. 因此在模型预测方面往往更加稳健和精确, 是模型选择方法的有效替代和补充.

本文 2022 年 10 月 13 日收到. 2023 年 3 月 21 日收到修改稿.
*国家自然科学基金(批准号: 12171483)资助项目.
†通讯作者.

模型平均方法近年来受到了学者们的广泛关注, 主要分为贝叶斯模型平均和频率模型平均两大方向, 其中频率模型平均方法更受关注. Buckland等[1]提出了SAIC(Smoothed AIC)、SBIC(Smoothed BIC)模型平均, Claeskens和Hjort[2]发展了适用于研究不同感兴趣变量的SFIC(Smoothed FIC)模型平均估计. Hansen[3]提出了基于Mallows准则的模型平均(MMA)方法, Wan等[4]将MMA方法拓展到了连续的权重集合以及非嵌套线性模型中. Hansen和Racine[5]提出了Jackknife模型平均(JMA)方法, Zhang等[6]将JMA方法拓展到了非对角的误差协方差结构以及滞后的相依数据中, 改进了只适用于独立同分布样本的情形. Liu和Ouki[7]给出了异方差稳健的模型平均方法. 还有一些研究关注更为复杂的半参数模型的模型平均方法, 如Li等[8]研究了变系数模型的模型平均方法; Zhang和Wang[9]研究了部分线性模型下的Mallows模型平均; Zhu等[10]和Hu等[11]给出了部分线性变系数模型下的模型平均方法.

加速失效时间模型(Accelerated failure time model, 以下记为AFT模型)是生存分析中一类重要模型. 该模型结构简单、解释性强, 实际应用广泛. 现状数据(也称I 型区间删失数据)是生存分析领域中一种常见的删失数据类型, 在临床试验、流行病学、经济学和社会学研究中都有出现. 此类数据的特点是感兴趣事件的确切发生时间无法被准确观测, 研究人员只能得到观测时间和示性函数的信息. 当实际数据中存在有多个协变量的情况时, 就会存在多个候选模型, 而使用组合多个候选模型的模型平均方法往往可以提高预测的精确度. 目前关于删失数据模型平均方法的相关研究还处于起步阶段, 且主要是针对随机右删失数据. 如:Hjort和Claeskens[12]研究了Cox回归模型的Focused信息准则(FIC) 以及模型平均估计. 孙志猛等[13]研究了响应变量随机右删失时, 基于Focused信息准则的线性模型的模型选择和模型平均. 孙志猛[14]继续将该准则拓展到线性分位数模型中并得到了拓展的Focused信息准则(E-FIC), 有效解决了同时估计多个兴趣参数的问题. Sun等[15]和吕晓玲等[16](2018)分别研究了响应变量随机右删失时, 部分线性分位数回归模型和部分线性变系数分位数回归模型的频率模型平均估计方法. He等[17]基于泛函鞅残差过程研究了高维情形下右删失数据Cox回归模型的JMA方法. Yan等[18]研究了高维线性模型下, 响应变量为右删失时的JMA方法. Li等[19]研究了右删失数据下, 非参数比例风险可加模型的模型平均方法. Liang等[20]研究了响应变量随机右删失时线性回归模型的MMA方法.

通过对现有文献的梳理发现, 目前关于删失数据的研究大多基于模型局部误设定的假设, 而基于模型全局误设定框架的最优模型平均方法的研究也主要集中于随机右删失数据. 而现状数据也是很常见且重要的一类删失数据, 其数据结构与右删失数据有所不同, 适用于右删失数据的模型平均方法并不完全适用于现状数据. 而无论是在模型局部误设定或是全局误设定的框架下对现状数据的模型平均方法都少有研究. 为了填补这方面的理论空缺, 并进一步拓展模型平均方法的应用

范围, 本文研究加速失效时间模型下现状数据的JMA方法, 给出候选模型权重的选取方法, 并在一定正则条件下建立所提模型平均估计量的渐近最优性. 最后通过数值模拟和实际数据来展示所提方法的性能表现, 并给出相关理论结果的证明.

2　数据和模型

假设有n个独立观测个体, T_i和C_i分别表示个体i的失效时间和观测时间, 且二者相互独立. 对于现状数据, T_i的精确值无法观测, 只知道T_i出现在C_i之前或之后. 也就是说, 事件失效时间要么被左删失, 要么被右删失, 只能观测到$\{(C_i,\delta_i), i=1,2,\cdots,n\}$, 其中$\delta_i = I(T_i \le C_i)$. 令$Y_i = \log T_i$, $V_i = \log C_i$. 考虑如下AFT模型

$$Y_i = \mu_i + \varepsilon_i = X_i^T\beta + \varepsilon_i = \sum_{p=1}^{\infty} x_{ip}\beta_p + \varepsilon_i, \qquad i = 1,2,\cdots,n, \tag{1}$$

这里的$X_i = (x_{i1}, x_{i2}, \cdots)^T$是可数的无穷维随机向量, $\beta = (\beta_1, \beta_2, \cdots)^T$代表未知回归系数向量, ε_i是随机误差项, 满足$\mathrm{E}(\varepsilon_i|X_i) = 0$, $\mathrm{Var}(\varepsilon_i|X_i) = \sigma_i^2$. 因为$Y_i$无法观测到, 可参考[21]的做法, 对$Y_i$采用如下变换. 令

$$Y_i^* = \varphi_1(V_i)\delta_i + \varphi_2(V_i)(1-\delta_i), \qquad i = 1,2,\cdots,n,$$

其中$V_1, V_2, \cdots, V_n$是独立同分布的随机变量, 其密度函数$g(\cdot)$已知. φ_1 和φ_2是满足

$$\begin{cases} \varphi_2(v) = \varphi_1(v) + g^{-1}(v) \\ \mathrm{E}\varphi_1(V_i) = 0 \end{cases}$$

的连续函数. 为满足变换的无偏性, 我们在此假设观测时间$C_i > 1$, 容易证明Y_i^*是Y_i的无偏变换, 即$\mathrm{E}Y_i^* = \mathrm{E}Y_i$(详细证明过程见第7节).

注：(1) 因为$V_1, V_2, \cdots, V_n$均可观测, $g(v)$即使未知, 也可以采用诸如核密度方法进行估计. (2) [21]针对现状数据提出了无偏变换方法, 但却不适用于AFT模型中的对数变换; [22]在Remark2.2中考虑了区间II型删失下, 对T_i取对数后的无偏变换, 对本文的研究有一定的借鉴作用. 但他们的方法中需要确定满足特定条件的V的某种函数$M(v)$, 虽然这种设定在数学理论上是成立的, 但在实践中当V的分布未知时很难操作, 可行性不高. (3) 本文假设$C_i > 1$虽然有一定局限性, 但对于一些实际问题当$P(C_i > T_i)$较大时是可以满足的, 或者可以通过坐标轴平移来实现.

3　Jackknife模型平均方法

考虑K_n个候选模型, 其中第k $(k = 1,2,\cdots,K_n)$个候选模型M_k定义为

$$Y_i = X_{k,i}^T\beta_k + b_{k,i} + \varepsilon_i, \qquad i = 1,2,\cdots n,$$

其中$X_{k,i}$为解释变量X_i的p_k维子向量, β_k为对应的回归系数向量, $b_{k,i}=\mu_i-X_{k,i}^T\beta_k$表示第$k$个候选模型的近似误差, ε_i同模型(1)中的假设.

基于上面提出的无偏变换, 可得到第k个候选模型的回归系数β_k的最小二乘估计为$\hat{\beta}_k=(X_k^TX_k)^{-1}X_k^TY^*$, 对应可得到$\mu$的一系列估计量为$\hat{\mu}=\{\hat{\mu}_1,\cdots,\hat{\mu}_{K_n}\}$. 其中第$k$个候选模型中$\mu$的估计量可记为$\hat{\mu}_k=X_k\hat{\beta}_k=P_kY^*$, 这里的$P_k=X_k(X_k^TX_k)^{-1}X_k^T$.

设$\mathbf{w}=(w_1,\cdots,w_{K_n})^T$是连续集合上的权重向量. 令

$$\mathcal{H}_n\triangleq\Big\{\mathbf{w}\in[0,1]^{K_n}:\sum_{k=1}^{K_n}w_k=1\Big\},$$

则$\mu=E(Y|X)$的模型平均估计为

$$\hat{\mu}(\mathbf{w})=\sum_{k=1}^{K_n}w_k\hat{\mu}_k=\sum_{k=1}^{K_n}w_kP_kY^*=P(\mathbf{w})Y^*,$$

这里的$P(\mathbf{w})=\sum\limits_{k=1}^{K_n}w_kP_k$为对应的加权帽子矩阵.

下面采用删一交叉验证法[5]来选取候选模型的权重. 首先, 设由删一交叉验证法得到的第k个候选模型预测值为$\tilde{\mu}_k=(\tilde{\mu}_k^{(-1)},\tilde{\mu}_k^{(-2)},\cdots,\tilde{\mu}_k^{(-n)})^T$, 其中$\tilde{\mu}_k^{(-i)}$表示从第$k$个候选模型中删除第$i$个观测值$(V_i,X_i,\delta_i)$所得到的预测值. 根据Li(1987)可得$\tilde{\mu}_k=\tilde{P}_kY^*$. 这里$\tilde{P}_k$是满足$\tilde{P}_k=D_k(P_k-I_n)+I_n$ 的光滑矩阵, D_k是对角线元素为$(1-p_{ii}^k)^{-1}$的对角阵, p_{ii}^k 是矩阵P_k的第i个对角元. 对应的删一预测值为

$$\tilde{\mu}(\mathbf{w})=\sum_{k=1}^{K_n}w_k\tilde{\mu}_k=\tilde{\mu}\mathbf{w}=\tilde{P}(\mathbf{w})Y^*,$$

其中$\tilde{\mu}=(\tilde{\mu}_1,\tilde{\mu}_2,\cdots,\tilde{\mu}_{K_n})$, $\tilde{P}(\mathbf{w})=\sum\limits_{k=1}^{K_n}w_k\tilde{P}_k$. 第$k$个候选模型对应的残差向量为$\tilde{e}_k=D_k\hat{e}_k=D_k(Y^*-P_kY^*)$, 进而得到Jackknife残差向量$\tilde{\mathbf{e}}(\mathbf{w})=Y^*-\tilde{\mu}(\mathbf{w})=\sum\limits_{k=1}^{K_n}w_k\tilde{e}_k=\tilde{\mathbf{e}}\mathbf{w}$, 其中$\tilde{\mathbf{e}}=(\tilde{e}_1,\tilde{e}_2,\cdots,\tilde{e}_{K_n})$, 则相应的交叉验证准则为

$$CV_n(\mathbf{w})=\|Y^*-\hat{\mu}(\mathbf{w})\|^{\mathbf{2}}=\tilde{\mathbf{e}}(\mathbf{w})^T\tilde{\mathbf{e}}(\mathbf{w})=\mathbf{w}^T\tilde{\mathbf{e}}^T\tilde{\mathbf{e}}\mathbf{w},\tag{2}$$

最小化上述准则得到权重向量$\hat{\mathbf{w}}=\arg\min\limits_{\mathbf{w}\in\mathcal{H}_n}CV_n(\mathbf{w})$, 最终定义$\mu$的Jackknife模型平均估计量为$\hat{\mu}(\hat{\mathbf{w}})=\hat{\mu}\hat{\mathbf{w}}$.

接下来讨论$\hat{\mu}(\hat{\mathbf{w}})$的渐近性质. 首先定义如下平方误差损失函数

$$L_n(\mathbf{w})=(\mu-\hat{\mu}(\mathbf{w}))^T(\mu-\hat{\mu}(\mathbf{w})),$$

则对应的风险函数为

$$R_n(\mathbf{w})=E(L_n(\mathbf{w})|X)=\|A(\mathbf{w})\mu\|^2+\mathrm{tr}\{P(\mathbf{w})\Omega P^T(\mathbf{w})\},$$

其中$\Omega=\mathrm{diag}(\sigma_1^2,\cdots,\sigma_n^2)$, $A(\mathbf{w})=I_n-P(\mathbf{w})$, 定义$\xi_n=\inf\limits_{\mathbf{w}\in\mathcal{H}_\mathbf{n}}R_n(\mathbf{w})$, $\varepsilon_i^*=Y_i^*-\mu_i$, $r_k=\mathrm{rank}(X_k)$, $\bar{r}=\max\limits_{1\le k\le K_n}r_k$. 令$\mathbf{w}_k^0$是第$k$个元素为1, 其他元素为0的权重向量, 即模型选择的权重. 要建立所提JMA估计量的渐近性质, 需要如下正则性条件:

C1: 存在正整数G及正常数κ, 使得$E[(\varepsilon_i^*)^{4G}|X_i]\le\kappa<\infty$, 且$K_n\xi_n^{-2G}\sum\limits_{k=1}^{K_n}\{R_n(\mathbf{w}_k^0)\}^G\to0$几乎处处成立.

C2: $\bar{r}=o(n)$, 且$\bar{r}\xi_n^{-1}\to0$几乎处处成立.

C3: $n^{-1}\|\mu\|^2=O(1)$几乎处处成立.

C4: 存在常数$c>0$使得$p_{ii}^k\le cn^{-1}r_k$几乎处处成立.

C5: T与C相互独立, $P(C>T)>0$, $V=\log C$是具有连续密度函数$g(v)$的非负随机变量, 且$g(v)>0$.

注　条件C1的前半部分对随机误差项的矩进行限制, 控制了数据产生过程中的噪声; 后半部分与[6]中的条件13类似, [4]对该条件进行了详细讨论. 该条件限制了候选模型个数随样本量的增长速率, 其成立的必要条件为$\xi_n\to\infty$. 这表明不存在具有零偏差的简单近似模型, 即所有的近似模型都是误设定的. 条件C2与[6]中的条件22类似, 表示允许r_k随样本量的增加而增加, 但限制其增长速率. 条件C3关注μ_i^2的均值, 是一个常见且合理的假设. 只要$|\mu_i|<\infty$ 对于任意的$i=1,2,\cdots,n$成立, 该条件即可满足. 条件C4常用于交叉验证方法中渐近最优性的研究, 表示没有任何占主导地位的候选模型, 可见[6,23,24]. 条件C5是现状数据中的一些比较容易满足的条件.

下面给出本文的主要理论结果, 证明过程见附录.

定理 1　在条件C1–C5成立的情况下, 前文提出的JMA估计量是渐近最优的, 也即

$$\frac{L_n(\hat{\mathbf{w}})}{\inf\limits_{\mathbf{w}\in\mathcal{H}_\mathbf{n}}L_n(\mathbf{w})}\xrightarrow{p}1. \tag{3}$$

4　模拟研究

下面通过数值模拟来评估本文提出的Jackknife模型平均方法的表现. 这里选用基于AIC和BIC准则的两种模型选择方法以及光滑AIC(SAIC)、光滑BIC(SBIC)两种模型平均方法, 还有等权重以及最大候选模型共六种方法进行对比. 第k个候选模型所对应的AIC值为$AIC_k=\log(\hat{\sigma}_k^2)+2n^{-1}\mathrm{tr}\{P_k\}$, BIC值为$BIC_k=\log(\hat{\sigma}_k^2)+n^{-1}\log(n)\mathrm{tr}\{P_k\}$, 其中$\hat{\sigma}_k^2=n^{-1}\|Y^*-\hat{\mu}_k\|^2$. 分别选取AIC值和BIC值最小的候选模型作为模型选择中的最优模型, 而第k个候选模型所对应的SAIC和SBIC模型平均方法的候选模型权重分别为

$$w_k^{\mathrm{SAIC}}=\frac{\exp(-AIC_k/2)}{\sum\limits_{k=1}^{K_n}\exp(-AIC_k/2)},\qquad w_k^{\mathrm{SBIC}}=\frac{\exp(-BIC_k/2)}{\sum\limits_{k=1}^{K_n}\exp(-BIC_k/2)}.$$

假设响应变量Y来自AFT模型$Y_i = X_i^T\beta + \varepsilon_i,\ i = 1, 2, \cdots, n$, 其中$X_i = (x_{i1}, \cdots, x_{ip})$ $(p = 200)$服从均值0, 协差阵为$\Sigma=(\rho^{|l-k|})_{1\le l,k\le 200}, \rho = 0.5$的多元正态分布. 分别考虑回归系数真值为$\beta_j = 1/j^2$及$\beta_j = \sqrt{2}/j^2 (j = 1, 2, \cdots, p)$的情形. 误差$\varepsilon_i \sim N(0, \eta^2 x_{i2}^2)$, 通过改变$\eta$的取值使得$R^2 = \mathrm{var}(\mu_1, \cdots, \mu_n)/\mathrm{var}(Y_1, \cdots, Y_n)$在0.1–0.9范围内变化, 其中$\mathrm{var}(\cdot)$表示样本方差. 取$V_i \sim \exp(\lambda)$, 其中$\lambda = 0.25$. 考虑$Y$的如下两种无偏变换形式

$$P1: \begin{cases} \varphi_1(v_i) = 0, \\ \varphi_2(v_i) = \dfrac{1}{\lambda} e^{\lambda v}, \end{cases} \qquad P2: \begin{cases} \varphi_1(v_i) = v - \dfrac{1}{\lambda}, \\ \varphi_2(v_i) = v - \dfrac{1}{\lambda} + \dfrac{1}{\lambda} e^{\lambda v}. \end{cases}$$

关于候选模型, 我们考虑了两种不同的设置: (1)候选模型嵌套: 即$M_1 \subset M_2 \cdots \subset M_{K_n}$; 嵌套情形下又分别考虑两种情况: 候选模型个数固定, 如$K_n = 20$, 以及候选模型个数随样本量变化$(K_n = [3n^{1/3}])$; (2)候选模型非嵌套, 每个候选模型中至少包含一个协变量. 本文考虑用前5个协变量$(X_j,\ j = 1, \cdots, 5)$构造候选模型, 则得到候选模型个数为$K_n = 2^5 - 1 = 31$. 评测标准采用标准化均方误差(NMSE)

$$\mathrm{NMSE} = \frac{1}{D}\sum_{d=1}^{D} \|\hat{\mu}^d - \mu^d\|^2 / \min_{\mathrm{mse}},$$

其中D表示循环次数(这里取$D = 200$), $\hat{\mu}^d$表示第d次循环中μ的估计量, $\min\limits_{\mathrm{mse}}$表示七种方法中对应最小的MSE值.

表1–表3分别给出了P1变换模式下各方法在不同样本量和各种R^2设置下所得到的NMSE的中位数和均值(P2 变换模式下结果类似, 这里省略), 其中JMA为Jackknife模型平均方法, SAIC为光滑AIC模型平均方法, SBIC为光滑BIC模型平均方法, EW表示等权重, LM表示最大候选模型, 每一行最优的值加黑表示.

从表1–表3可看出, 不论候选模型是否嵌套, JMA方法对应的NMSE都是七种方法中最小的. 候选模型嵌套情形下, 除JMA方法之外, SAIC, SBIC, EW 方法表现较为接近, 而LM方法的表现最差. 候选模型非嵌套的情况下, JMA方法表现最好, 而EW和LM方法随着样本量的增大表现逐渐接近, AIC和BIC方法则表现最差.

此外, 为了使模拟结果更加全面且更具有说服力, 我们还考虑了回归系数为$\beta_j = \sqrt{2}/j^2$的情况, 这里只给出了P1变换模式下候选模型嵌套且个数变化时的预测结果(表4). 从表4可看出, 在该种模拟设置下, 大多数情况下JMA方法的预测效果依然最好, 在不是最优的情况时表现也还不错. 我们还考虑了V的密度函数$g(v)$未知时, P1变换模式下候选模型嵌套且个数变化时的预测结果(表5). 表5同样表明JMA方法的估计效果是最好的. 进一步我们考虑了误差ε服从自由度为3的t分布时, 各个方法所对应的MSE 值(表6). 从表6也可看出JMA方法的预测表现在大多数情况下仍是最好的, 这进一步验证了所提方法的优良性. 我们还通过折线图来展示了NMSE的变化. 图1到图5也显示了JMA方法相较于其他几种方法有一定优势, 且表现比较稳定.

表 1. 嵌套情况下候选模型个数固定时, P1变换模式的NMSE($\beta_j = 1/j^2$)

样本量	R^2	方法	JMA	SAIC	SBIC	AIC	BIC	EW	LM
		Median	**1.000**	1.137	1.126	1.286	1.256	1.104	1.705
	0.2	Mean	**1.000**	1.096	1.083	1.185	1.157	1.104	1.600
		Median	**1.000**	1.171	1.150	1.343	1.236	1.174	1.766
100	0.4	Mean	**1.000**	1.119	1.107	1.224	1.177	1.111	1.603
		Median	**1.000**	1.123	1.109	1.301	1.234	1.083	1.708
	0.8	Mean	**1.000**	1.096	1.083	1.208	1.154	1.106	1.622
		Median	**1.000**	1.019	1.014	1.089	1.084	1.026	1.351
	0.2	Mean	**1.000**	1.027	1.023	1.084	1.075	1.032	1.317
		Median	**1.000**	1.017	1.011	1.075	1.041	1.030	1.284
200	0.4	Mean	**1.000**	1.028	1.023	1.096	1.081	1.030	1.272
		Median	**1.000**	1.024	1.021	1.095	1.061	1.019	1.317
	0.8	Mean	**1.000**	1.025	1.021	1.078	1.071	1.032	1.309

表 2. 嵌套情况下候选模型个数变化时, P1变换模式的NMSE($\beta_j = 1/j^2$)

样本量	R^2	方法	JMA	SAIC	SBIC	AIC	BIC	EW	LM
		Median	**1.000**	1.029	1.022	1.109	1.091	1.018	1.456
	0.2	Mean	**1.000**	1.037	1.031	1.116	1.101	1.035	1.382
		Median	**1.000**	1.065	1.058	1.197	1.170	1.067	1.430
100	0.4	Mean	**1.000**	1.053	1.047	1.155	1.123	1.042	1.403
		Median	**1.000**	1.035	1.028	1.136	1.115	1.038	1.505
	0.8	Mean	**1.000**	1.050	1.042	1.153	1.127	1.046	1.435
		Median	**1.000**	1.046	1.043	1.103	1.098	1.041	1.305
	0.2	Mean	**1.000**	1.018	1.014	1.080	1.069	1.016	1.261
		Median	**1.000**	1.068	1.064	1.144	1.137	1.039	1.325
200	0.4	Mean	**1.000**	1.023	1.019	1.105	1.084	1.016	1.274
		Median	**1.000**	1.007	1.005	1.060	1.058	1.016	1.279
	0.8	Mean	**1.000**	1.022	1.018	1.090	1.076	1.018	1.270

表 3. 非嵌套情形下P1变换模式对应的NMSE($\beta_j = 1/j^2$)

样本量	R^2	方法	JMA	SAIC	SBIC	AIC	BIC	EW	LM
		Median	**1.000**	1.716	1.719	2.266	2.243	1.273	1.809
	0.2	Mean	**1.000**	1.689	1.689	2.002	2.004	1.253	1.816
		Median	**1.000**	1.702	1.706	2.214	2.238	1.261	1.802
100	0.4	Mean	**1.000**	1.625	1.623	1.955	1.955	1.238	1.744
		Median	**1.000**	1.726	1.724	2.276	2.220	1.291	1.874
	0.8	Mean	**1.000**	1.587	1.587	1.953	1.925	1.240	1.775
		Median	**1.000**	1.809	1.809	2.331	2.337	1.409	1.396
	0.2	Mean	**1.000**	1.764	1.766	2.040	2.042	1.365	1.381
		Median	**1.000**	1.873	1.883	2.423	2.423	1.402	1.406
200	0.4	Mean	**1.000**	1.759	1.759	2.074	2.072	1.360	1.360
		Median	**1.000**	1.825	1.827	2.443	2.443	1.402	1.396
	0.8	Mean	**1.000**	1.805	1.805	2.114	2.135	1.378	1.386

表 4. 嵌套情况下候选模型个数变化时, P1变换模式的NMSE($\beta_j = \sqrt{2}/j^2$)

样本量	R^2	方法	JMA	SAIC	SBIC	AIC	BIC	EW	LM
		Median	1.013	1.004	**1.000**	1.139	1.085	1.005	1.388
	0.2	Mean	**1.000**	1.009	1.004	1.088	1.069	1.013	1.286
		Median	1.031	1.043	1.040	1.187	1.166	**1.000**	1.278
100	0.4	Mean	**1.000**	1.010	1.005	1.108	1.098	1.002	1.261
		Median	**1.000**	1.036	1.030	1.088	1.072	1.010	1.347
	0.8	Mean	**1.000**	1.013	1.008	1.084	1.075	1.011	1.287
		Median	1.014	1.002	**1.000**	1.085	1.060	1.003	1.231
	0.2	Mean	**1.000**	1.006	1.005	1.061	1.047	1.004	1.172
		Median	1.013	1.010	1.008	1.096	1.096	**1.000**	1.262
200	0.4	Mean	**1.000**	1.009	1.007	1.067	1.053	1.001	1.191
		Median	**1.000**	1.009	1.006	1.069	1.061	1.010	1.187
	0.8	Mean	**1.000**	1.009	1.006	1.073	1.067	1.010	1.204

表 5. 嵌套情况下候选模型个数变化时, P1变换模式的NMSE($g(v)$未知, $\beta_j = 1/j^2$)

样本量	R^2	方法	JMA	SAIC	SBIC	AIC	BIC	EW	LM
		Median	**1.000**	1.100	1.090	1.232	1.156	1.129	1.649
	0.2	Mean	**1.000**	1.088	1.079	1.197	1.172	1.084	1.565
		Median	**1.000**	1.048	1.039	1.132	1.084	1.050	1.509
100	0.4	Mean	**1.000**	1.045	1.039	1.136	1.103	1.056	1.475
		Median	**1.000**	1.071	1.060	1.196	1.166	1.086	1.555
	0.8	Mean	**1.000**	1.078	1.069	1.173	1.137	1.078	1.535
		Median	**1.000**	1.077	1.067	1.148	1.122	1.065	1.405
	0.2	Mean	**1.000**	1.053	1.049	1.143	1.126	1.044	1.383
		Median	**1.000**	1.044	1.039	1.183	1.119	1.057	1.405
200	0.4	Mean	**1.000**	1.037	1.032	1.116	1.087	1.037	1.389
		Median	**1.000**	1.095	1.082	1.156	1.154	1.082	1.424
	0.8	Mean	**1.000**	1.045	1.040	1.143	1.122	1.045	1.393

表 6. 随机误差ε服从自由度为3的t分布时对应的MSE

样本量	方法	JMA	SAIC	SBIC	AIC	BIC	EW	LM
	Median	**0.522**	0.623	0.613	0.701	0.661	0.625	0.937
100	Mean	**0.568**	0.652	0.646	0.724	0.690	0.660	1.024
	Optimal rate	**0.620**	0.015	0.075	0.200	0.050	0.040	0.000
	Median	**0.461**	0.499	0.498	0.538	0.509	0.500	0.737
200	Mean	**0.480**	0.533	0.529	0.590	0.553	0.541	0.796
	Optimal rate	**0.525**	0.005	0.115	0.225	0.065	0.065	0.000

注: Optimal rate表示最优率, 指D次循环中该方法对应的MSE最小的比例.

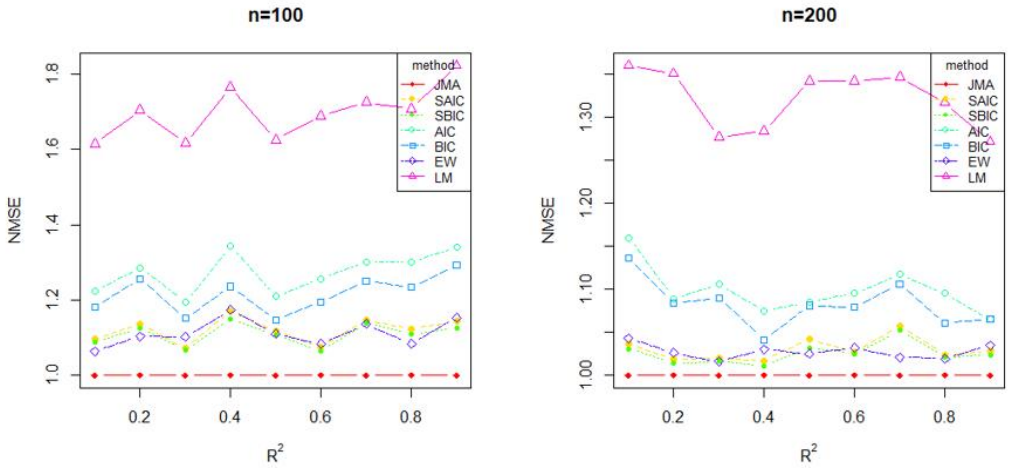

图 1. 嵌套情况下候选模型个数固定时, P1变换模式下的NMSE($\beta_j = 1/j^2$)

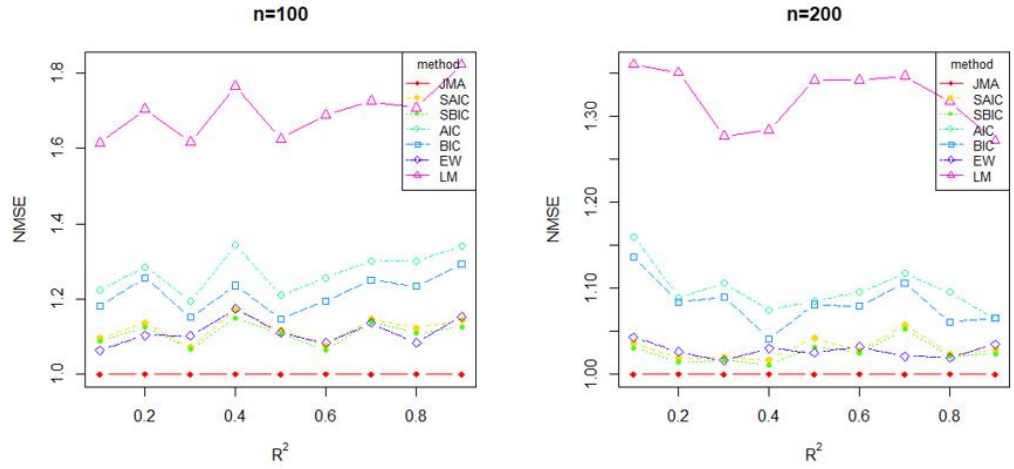

图 2. 嵌套情况下候选模型个数变化时, P1变换模式下的NMSE($\beta_j = 1/j^2$)

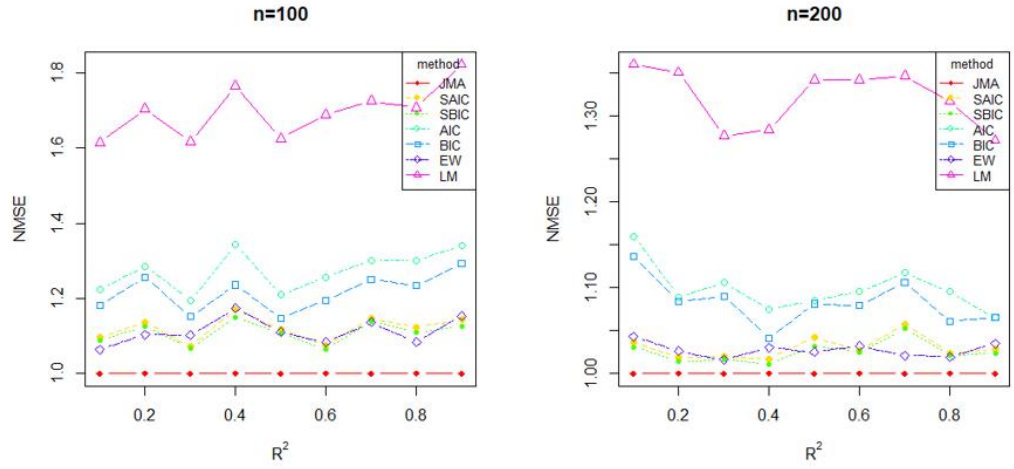

图 3. 非嵌套情形下P1变换模式对应的NMSE($\beta_j = 1/j^2$)

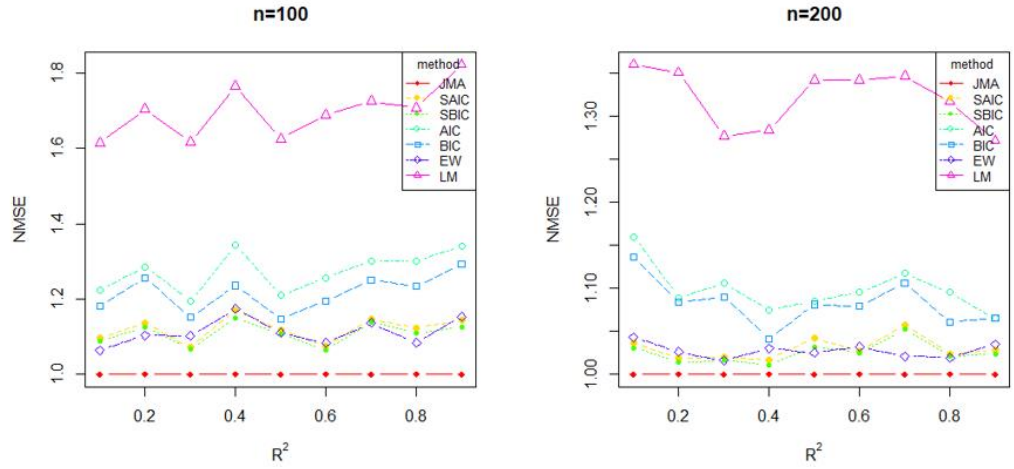

图 4. 嵌套情况下候选模型个数变化时, P1变换模式下的NMSE($\beta_j = \sqrt{2}/j^2$)

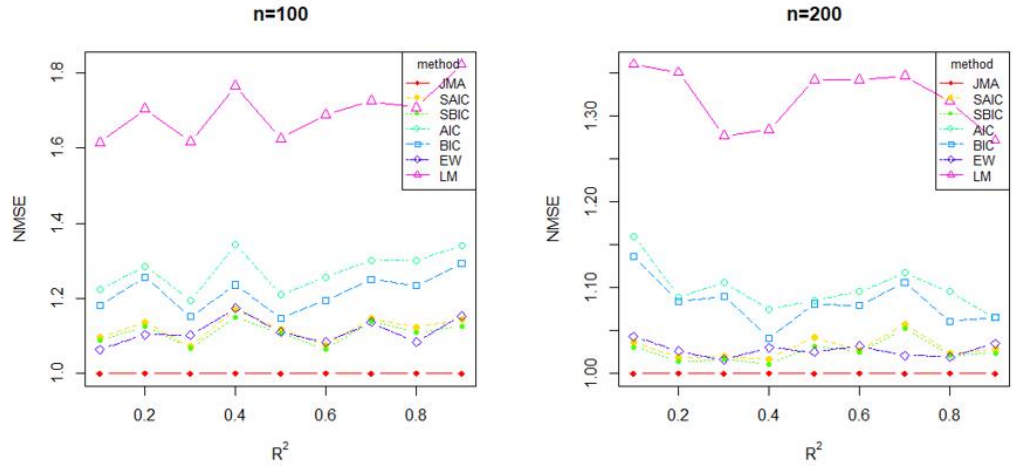

图 5. 嵌套情况下候选模型个数变化时, P1变换模式下的NMSE($g(v)$未知, $\beta_j = 1/j^2$)

5　实例分析

下面用本文提出的模型平均方法来分析2003年尼日利亚人口与健康调查中(NDHS)收集的尼日利亚儿童死亡率数据, [25,26]对该数据进行过研究. 该数据主要关注育龄女性(13–49岁)及其子女健康状况, 其中关于儿童生存时间的信息主要来自于对其母亲的调查访问. 而由于遗忘等原因, 会出现无法获取儿童确切生存时间(以天为单位)的情况. 该数据中若儿童死亡发生在出生之后的前两个月, 则生存

时间可以确切观测到; 而其他超过两个月的生存时间都可以看作是区间I型删失数据. 因为数据中有117 个儿童在出生后的前两个月就死亡了, 可得到确切的生存时间, 因此不在本文研究范围内. 进一步删掉包含缺失值的个体后, 最终得到的数据中包含$n = 5482$个观测.

本文关注的响应变量为儿童的生存时间. 变量Death代表示性函数δ, $Death = 1$表示该个体在研究持续的时间内死亡, Death=0表示该个体在研究截止后依然是存活的. 考虑4个协变量: $X_1 = 1$表示孩子在医院出生, $X_1 = 0$ 表示不在医院出生; $X_2 = 1$ 表示男孩, $X_2 = 0$表示女孩; $X_3 = 1$ 表示母亲接受过高等教育, $X_3 = 0$ 表示母亲没有接受过高等教育; $X_4 = 1$ 表示家庭居住在城市, $X_4 = 0$ 表示家庭不在城市. 在分析中, 我们根据预测变量的排序结果考虑嵌套模型平均方法. 将实际数据分为训练集和验证集, 按照[60%n], [70%n], [80%n], [90%n]的标准来选取训练集样本量n_0, 剩余$n_{test} = n - n_0$的样本用来做验证集. 首先用训练集来进行估计, 然后带入验证集中进行预测和评估, 循环计算200次来消除随机误差的影响. 最后, 用归一化的均方预测误差(NMSPE)来评估各个方法的预测表现(指用每种方法对应的MSPE除以最小的MSPE, 对应最小MSPE的方法值为1).

从表7的结果可以看出, JMA方法对应的NMSPE是七种方法中最小的, 而等权重、SAIC和SBIC方法的结果相近, AIC和BIC模型选择方法的结果接近, 最大候选模型的方法表现最差. 该结论进一步验证了所提模型平均方法可以提高预测精度.

表 7. 实际数据下七种方法所对应的NMSPE值

训练样本	方法	JMA	SAIC	SBIC	AIC	BIC	EW	LM
	Median	**1.000**	1.007	1.007	1.020	1.019	1.007	1.251
[60%n]	Mean	**1.000**	1.007	1.008	1.021	1.020	1.008	1.250
	Median	**1.000**	1.008	1.008	1.023	1.022	1.008	1.249
[70%n]	Mean	**1.000**	1.008	1.008	1.023	1.022	1.008	1.250
	Median	**1.000**	1.008	1.008	1.027	1.024	1.008	1.248
[80%n]	Mean	**1.000**	1.008	1.008	1.028	1.026	1.008	1.249
	Median	**1.000**	1.008	1.008	1.040	1.036	1.008	1.250
[90%n]	Mean	**1.000**	1.010	1.010	1.042	1.039	1.010	1.249

6 结论与展望

本文主要研究了加速失效时间模型下现状数据的Jackknife模型平均方法, 首先对现状数据进行合理的无偏变换, 然后引入删一交叉验证准则来选取候选模型的权重. 并证明了在一定的正则性条件下, 所提模型平均估计量有渐近最优性质. 最后通过数值模拟和实证分析说明了所提方法的优良性.

此外, 本文仅仅关注了AFT模型下现状数据的模型平均, 也可将所提方法拓展

到其他半参数模型, 比如: 加性风险模型、部分线性AFT模型、部分线性Cox模型、半参数变换模型等. 还可以考虑协变量是高维的情形, 以及将现状数据拓展到区间II型删失数据等等, 这些都是值得进一步研究的问题.

7 定理证明

首先证明无偏变换的合理性. 为了书写过程中的方便, 我们在证明过程中省去了下角标. 当条件C5成立时, 易得

$$\begin{aligned}
E(Y^*) &= \int_{-\infty}^{\infty}\int_{y}^{\infty}\varphi_1(v)\mathrm{d}G(v)\mathrm{d}F(y)+\int_{-\infty}^{\infty}\int_{0}^{y}\varphi_2(v)\mathrm{d}G(v)\mathrm{d}F(y)\\
&= \int_{-\infty}^{\infty}\int_{y}^{\infty}\varphi_1(v)g(v)\mathrm{d}v\mathrm{d}F(y)+\int_{-\infty}^{\infty}\int_{0}^{y}\varphi_2(v)g(v)\mathrm{d}v\mathrm{d}F(y)\\
&= \int_{-\infty}^{\infty}\int_{y}^{\infty}\varphi_1(v)g(v)\mathrm{d}v\mathrm{d}F(y)+\int_{-\infty}^{\infty}\int_{0}^{y}\varphi_1(v)g(v)\mathrm{d}v\mathrm{d}F(y)+\int_{-\infty}^{\infty}\int_{0}^{y}\mathrm{d}v\mathrm{d}F(y)\\
&= \int_{-\infty}^{\infty}\int_{0}^{\infty}\varphi_1\left(v\right)g\left(v\right)\mathrm{d}v\mathrm{d}F\left(y\right)+\int_{-\infty}^{\infty}\int_{0}^{y}\mathrm{d}v\mathrm{d}F(y)\\
&= 0+\int_{-\infty}^{\infty}y\mathrm{d}F(y)=E(Y).
\end{aligned}$$

其中$g(v)$是V的密度函数, $G(v)$是V的分布函数, $F(y)$是Y的分布函数. 这个结论说明在对响应变量T取对数后, 对应的无偏变换依然成立. 证毕.

定理1的证明主要参考[6], 本文与已有研究的主要区别是响应变量的观测是删失的, 而在模型(1)假设下, 对响应变量Y的无偏变换并不影响证明中所需条件的成立. 因此本文给出简要的证明过程, 与已有研究相似的部分不再赘述.

定义$\lambda_{\max}(\cdot)$为矩阵的最大奇异值, P_k对称且幂等, 可得到

$$\lambda_{\max}(\Omega)=O(1),\qquad \lambda_{\max}(P_k)\le 1,\qquad \sup_{\mathbf{w}\in\mathcal{H}_{\mathbf{n}}}\lambda_{\max}(A(\mathbf{w}))\le 1+1=2,$$

根据$r_k=\mathrm{rank}(X_k)$, 可得到$\mathrm{trace}(P_k)=r_k$, $\mathrm{trace}(P_k^TP_k)\le r_k$. 根据条件C4可得$\max\limits_{1\le k\le K_n}\max\limits_{1\le i\le n}p_{ii}^k\le cn^{-1}\max\limits_{1\le k\le K_n}r_k\le cn^{-1}\bar{r}$. 定义$Q_k$为$n\times n$维的对角矩阵, $\lambda_{\max}(Q_k)=\max\limits_{1\le i\le n}(Q_{ii}^k)=\max\limits_{1\le i\le n}\frac{p_{ii}^k}{1-p_{ii}^k}$, 根据条件C2可得$\sup\limits_{\mathbf{w}\in\mathcal{H}_{\mathbf{n}}}\lambda_{\max}(V(\mathbf{w}))\le O_p(\frac{\bar{r}}{n})$.

已知$D_k=\frac{1}{1-p_{ii}^k}$, 则$D_k-I_n=\frac{1}{1-p_{ii}^k}-\frac{1-p_{ii}^k}{1-p_{ii}^k}=\frac{p_{ii}^k}{1-p_{ii}^k}$, 可得到$D_k=I_n+Q_k$. 经计算可得$\tilde{P}_k=P_k-Q_kA_k$, $\tilde{P}\left(\mathbf{w}\right)=P\left(\mathbf{w}\right)-Q\left(\mathbf{w}\right)A\left(\mathbf{w}\right)$. 定义$V_k=Q_kA_k$, $V\left(\mathbf{w}\right)=\sum\limits_{k=1}^{K_n}w_kV_k$, $A\left(\mathbf{w}\right)=\sum\limits_{k=1}^{K_n}w_kA_k$, 则Jackknife准则(2)可改写为

$$CV_n\left(\mathbf{w}\right)=Y^{*T}A(\mathbf{w})^TA\left(\mathbf{w}\right)Y^*+Y^{*T}D\left(\mathbf{w}\right)Y^*,$$

其中$D(\mathbf{w}) = A(\mathbf{w})^T V(\mathbf{w}) + V(\mathbf{w})^T A(\mathbf{w}) + V(\mathbf{w})^T V(\mathbf{w})$. 进一步可得

$$\begin{aligned} CV_n(\mathbf{w}) =& L_n(\mathbf{w}) + \|\varepsilon^*\|^2 + 2\mu^T A(\mathbf{w})\varepsilon^* - 2\varepsilon^{*T} P(\mathbf{w})\varepsilon^* \\ &+ \mu^T D(\mathbf{w})\mu + \varepsilon^{*T} D(\mathbf{w})\varepsilon^* + 2\mu^T D(\mathbf{w})\varepsilon^*, \end{aligned}$$

此处$\|\varepsilon^*\|^2$与$\mathbf{w}$无关. 因此要证定理1成立, 只需证明当$n \to \infty$时, 如下式子成立

$$\sup_{\mathbf{w}\in\mathcal{H}_\mathbf{n}} |\mu^T D(\mathbf{w})\mu| / R_n(\mathbf{w}) \xrightarrow{p} 0, \tag{4}$$

$$\sup_{\mathbf{w}\in\mathcal{H}_\mathbf{n}} |\varepsilon^{*T} D(\mathbf{w})\varepsilon^*| / R_n(\mathbf{w}) \xrightarrow{p} 0, \tag{5}$$

$$\sup_{\mathbf{w}\in\mathcal{H}_\mathbf{n}} |\mu^T D(\mathbf{w})\varepsilon^*| / R_n(\mathbf{w}) \xrightarrow{p} 0, \tag{6}$$

$$\sup_{\mathbf{w}\in\mathcal{H}_\mathbf{n}} |\mu^T A(\mathbf{w})\varepsilon^*| / R_n(\mathbf{w}) \xrightarrow{p} 0, \tag{7}$$

$$\sup_{\mathbf{w}\in\mathcal{H}_\mathbf{n}} |\varepsilon^{*T} P(\mathbf{w})\varepsilon^*| / R_n(\mathbf{w}) \xrightarrow{p} 0, \tag{8}$$

$$\sup_{\mathbf{w}\in\mathcal{H}_\mathbf{n}} \left| \frac{L_n(\mathbf{w})}{R_n(\mathbf{w})} - 1 \right| \xrightarrow{p} 0. \tag{9}$$

根据条件C2可得 $\sup\limits_{\mathbf{w}\in\mathcal{H}_\mathbf{n}} \lambda_{\max}(D(\mathbf{w})) \le O_p(\frac{\bar{r}}{n})$, 接着根据条件C1–C4并参考[6]可证公式(4)–(9)成立, 即可证明定理1成立, 具体证明过程此处省略. 证毕.

参 考 文 献

[1] Buckland S, Burnham K, Augustin N. Model selection: an integral part of inference. *Biometrics*, 1997, 53(2): 603–618

[2] Claeskens G, Hjort N L. The focused information criterion. *Journal of the American Statistical Association*, 2003, 98(464): 900–916

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

[4] Wan A T K, Zhang X Y, Zou G H. Least squares model averaging by mallows criterion. *Journal of Econometrics*, 2010, 156(2): 277–283

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

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

[7] Liu Q F, Okui R. Heteroskedasticity-robust Cp model averaging. *Econometrics Journal*, 2013, 16: 463–472

[8] Li C, Li Q, Racine J, Zhang D Q. Optimal model averaging of varying coefficient models. *Statistica Sinica*, 2018, 28(4): 2795–2809

[9] Zhang X Y, Wang W D. Optimal model averaging estimation for partially linear models. *Statistica Sinica*, 2019, 29(2): 693–718

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

[11] Hu G Z, Cheng W H, Zeng J. Model averaging by jackknife criterion for varying-coefficient partially linear models. *Communications in Statistics-Theory and Methods*, 2020, 49(11): 2671–2689

[12] Hjort N L, Claeskens G. Focused information criteria and model averaging for the cox hazard regression model. *Journal of the American Statistical Association*, 2006, 101(476): 1449–1464

[13] 孙志猛, 马景义, 苏治. 响应变量删失情况下线性模型的FIC模型选择和模型平均. 中国科学:数学, 2013, 43(07): 647–661
(Sun Z M, Ma J Y, Su Z. FIC model selection and model averaging for linear model with censored response. *Scientia Sinica(Mathematica)*, 2013, 43(07): 647–661)

[14] 孙志猛. 删失分位数回归模型基于扩展兴趣信息准则的平均估计. 中国科学:数学, 2014, 44(08): 857–874
(Sun Z M. Extended focused information criterion for censored quantile regression model and averaging estimation. *Scientia Sinica(Mathematica)*, 2014, 44(08): 857–874)

[15] Sun Z M, Sun L Q, Lu X L, Zhu J, Li Y Z. Frequentist model averaging estimation for the censored partial linear quantile regression model. *Journal of Statistical Planning and Inference*, 2017, 189: 1–15

[16] 吕晓玲, 王小宁, 孙志猛. 删失分位数变系数回归模型的FIC模型平均估计. 系统科学与数学, 2018, 38(07): 4–21
(Lu X L, Wang X N, Sun Z M. FIC based model averaging for the censored quantile varying coefficient regression model. *Journal of Systems Science and Mathematical Sciences*, 2018, 38(07): 4–21)

[17] He B H, Liu Y Y, Wu Y S, Yin G S, Zhao X Q. Functional martingale residual process for high-dimensional cox regression with model averaging. *Journal of Machine Learning Research*, 2020, 21(207): 1–37

[18] Yan X D, Wang H N, Wang W, Xie J H, Ren Y Y, Wang X J. Optimal model averaging forecasting in high-dimensional survival analysis. *International Journal of Forecasting*, 2021, 37(3): 1147–1155

[19] Li J L, Yu T H, Lv J, Lee M L T. Semiparametric model averaging prediction for lifetime data via hazards regression. *Journal of the Royal Statistical Society:Series C(Applied Statistics)*, 2021, 70(5): 1187–1209

[20] Liang Z Q, Chen X L, Zhou Y Q. Mallows model averaging estimation for linear regression model with right censored data. *Acta Mathematicae Applicatae Sinica, English Series*, 2022, 38(1): 5-23

[21] Zheng Z K. A class of estimators of the mean survival time from interval censored data with application to linear regression. *Applied Mathematics*, 2008, 23(4): 377–390

[22] Deng W L, Tian Y, Lv Q P. Parametric estimator of linear model with interval-censored Data. *Communications in Statistics-Simulation and Computation*, 2012, 41(10): 1794–1804

[23] Li K C. Asymptotic optimality for Cp, CL, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics*, 1987, 15(3): 958–975

[24] Andrews D. Asymptotic optimality of generalized CL, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics*, 1991, 47(3): 359–377

[25] Kneib T. Mixed model-based inference in geoadditive hazard regression for interval-censored survival times. *Computational Statistics and Data Analysis*, 2006, 51(2): 777–792

[26] Zhao H, Wu Q W, Li G, Sun J G. Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. *Journal of the American Statistical Association*, 2020, 115(529): 204–216

The Jackknife Model Averaging of Accelerated Failure Time Model with Current Status Data

ZHAO Hui LIU Binxia† DONG Qingkai

(*College of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan* 430070, *China*)

(†E-mail: lbx@stu.zuel.edu.cn)

ZHANG Xinyu

(*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing* 100190, *China*)

Abstract This paper studies the jackknife model averaging method of the accelerated failure time model with current status data. Firstly, through the unbiased transformation, we can obtain the LSE of regression parameters. Then the delete-one cross validation criterion is introduced to select the weights of candidate models, and under some regularity conditions, the asymptotic optimality of the model averaging estimator is established. Numerous simulation results show that the proposed method is superior to other existing model averaging and model selection methods in terms of prediction performance. Finally, we applied the proposed method to the NDHS data, and the real data also verify the excellent properties of the proposed method.

Keywords current status data; accelerated failure time model; unbiased transformation; model averaging; asymptotic optimal

MR(2023) Subject Classification 62N01; 62N02; 62F12

Chinese Library Classification O212.1