doi：10.13413/j.cnki.jdxblxb.2023503

# 现状数据下带测量误差的半参数加速风险模型的估计

裴宜凡，赵 波，王纯杰
（长春工业大学 数学与统计学院，长春 130012）

**摘要**：基于现状数据提出一种带有测量误差的半参数加速风险回归模型. 首先，用 I 样条近似未知累积基线风险函数，并基于 Sieve 极大似然估计方法获得模型的参数估计；其次，用模拟外推方法修正协变量的测量误差带来的估计误差；再次，通过数值模拟验证该方法的有效性以及忽略协变量误差对估计的影响；最后将该方法应用到心脑血管病死亡率的研究中获得心脑血管病发死亡的风险函数估计. 实验结果表明，该方法有效.

**关键词**：加速风险模型；模拟外推；测量误差；极大似然估计；I 样条
**中图分类号**：O212 **文献标志码**：A **文章编号**：1671-5489(2024)05-1122-07

# Estimation for Semi-parametric Accelerated Hazard Model with Measurement Error under Current Status Data

PEI Yifan, ZHAO Bo, WANG Chunjie
(*School of Mathematics and Statistics, Changchun University of Technology, Changchun* 130012, *China*)

**Abstract**: We proposed a semi-parametric accelerated hazard regression model with measurement errors based on the current status data. Firstly, the unknown baseline cumulative hazard function was approximated by using I-spline, and parameter estimates of the model were obtained based on Sieve maximum likelihood estimation method. Secondly, a simulation extrapolation method was used to correct estimation error caused by measurement errors in covariates. Thirdly, the numerical simulations were carried out to verify the effectiveness of the proposed method as well as the impact of ignoring measurement error in covariates. Finally, the proposed method was applied to study cardiovascular and cerebrovascular disease mortality, we obtained estimation of hazard function for cardiovascular and cerebrovascular disease mortality. The experimental results show that the proposed method is effective.

**Keywords**: accelerated hazard model; simulation extrapolation; measurement error; maximum likelihood estimate; I-spline

仅已知个体在当前状态下感兴趣事件是否发生，确切的发生时间不能被观测到，这时得到的数据

**收稿日期**：2023-11-29.

**第一作者简介**：裴宜凡(1998—)，女，汉族，硕士研究生，从事生存分析的研究，E-mail：pyf19980622@163.com. **通信作者简介**：王纯杰(1978—)，女，汉族，博士，教授，从事生存分析、Copula 理论、高频数据和工业统计数据分析的研究，E-mail：wangchunjie@ccut.edu.cn.

**基金项目**：国家自然科学基金(批准号：12271060；12301332)、国家自然科学基金天元基金(批准号：12226416)、吉林省科技厅重大专项研究项目(批准号：20210301038GX；20220301031GX)和吉林省科技厅重点研发项目(批准号：20230204078YY).

即为现状数据[1-4]. 例如，在对心脑血管病死亡率的研究中，病人准确的死亡时间不能被准确观察到，只知道个体在随访期间内是否发生死亡，这类数据即为经典的现状数据，在这项研究中选取 6 个变量对死亡率进行预测，其中血液中血小板数量和血清钠的水平经常存在测量误差，如果忽略测量误差的影响可能会使估计结果存在偏差，因此针对带有测量误差的现状数据建立模型并提出误差修正方法非常必要.

近年来，关于对协变量带有测量误差的模型研究已有许多结果，特别是对在右删失数据下带测量误差的加速失效时间模型的研究. 例如：He 等[5]用模拟外推方法对参数加速失效时间模型进行了分析，并将所提出的方法应用到冠心病死亡率的研究中；Yu 等[6]使用回归校正[7]的方法对带测量误差的协变量进行修正，并将所提出的方法应用到抑郁症死亡率的研究中；在实际应用中，随机误差项的分布形式通常是未知的，Zhang 等[8]在文献[5]的基础上将参数加速失效时间模型推广到了半参数加速失效时间模型；Chen[9]在长度偏差数据下对带有测量误差的加速失效时间模型进行了估计，用模拟外推方法对测量误差进行修正，并对其理论性质进行了证明. 此外，在其他模型假设下讨论带测量误差数据的分析方法也取得了一些研究成果. 例如：Cao 等[10]对协变量带有误差的 Cox 比例风险模型进行研究，使用回归校正的方法对带有误差的协变量进行误差修正；Pan 等[11]对具有时依协变量的部分线性 Cox 比例风险模型进行研究，使用 Bayes 平滑方法对带有误差的协变量进行修正；Ghapani 等[12]提出了一种双参数加权混合估计方法对线性测量误差模型的参数进行估计，并对其理论性质进行了证明；Dong 等[13]提出了一个半参数估计方法对存在测量误差的变系数模型进行估计.

加速风险模型是基于加速失效时间模型推广得到的一种半参数模型[14-19]，因其灵活性而备受关注. 在特殊情形下，该模型可退化为比例风险模型、比例优势模型和加速失效模型. 因此，本文在加速风险模型下讨论协变量带有测量误差的现状数据的半参数回归分析问题，利用 I 样条近似未知基准累积基线风险函数，同时引入模拟外推方法修正测量误差导致的估计误差. 数值模拟结果验证了所提方法的有效性.

## 1　数据及模型

设 $T$ 为个体的失效时间，$C$ 为个体的观测时间. 当 $T\leqslant C$ 时只知在 $C$ 时刻已发生感兴趣事件，具体的发生时间观测不到，这时便出现了左删失；当 $T>C$ 时只知在 $C$ 时刻未发生感兴趣事件，具体的发生时间不能准确观测到，这时便出现了右删失. 通常这类只知个体当前所处状态的数据称为现状数据. 设 $\boldsymbol{X}$ 是受测量误差影响的 $q_1$ 维协变量，$\boldsymbol{Z}$ 是不受测量误差影响的 $q_2$ 维协变量，$\delta$ 为一个示性函数，当 $T\leqslant C$ 时，$\delta=1$，当 $T>C$ 时，$\delta=0$. 则数据可表示为 $(C,\delta,\boldsymbol{X},\boldsymbol{Z})$.

当给定协变量时，加速风险模型可表示为

$$\Lambda(t\,|\,\boldsymbol{X},\boldsymbol{Z})=\Lambda_0(\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}+\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}\}t),$$

其中 $\Lambda_0(\cdot)$ 为分布形式未知的累积基线风险函数，$\boldsymbol{X}$ 和 $\boldsymbol{Z}$ 为观测到的协变量，$\boldsymbol{\beta}$ 和 $\boldsymbol{\alpha}$ 为待估参数，$\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}+\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}\}$ 为加速因子. 当 $\Lambda_0(s)=\exp\{s\}$ 时，加速风险模型可退化为比例风险模型. 加速风险模型的生存函数可表示为

$$S(t\,|\,\boldsymbol{X},\boldsymbol{Z})=\exp\{-\Lambda(t\,|\,\boldsymbol{X},\boldsymbol{Z})\}=\exp\{-\Lambda_0(\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}+\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}\}t)\}.$$

在实际应用中的一些变量通常带有测量误差，本文假设带误差的协变量满足经典测量误差模型

$$\boldsymbol{W}=\boldsymbol{X}+\boldsymbol{e},$$

其中 $\boldsymbol{W}$ 为协变量的实际观测值，$\boldsymbol{e}$ 为测量误差，$\boldsymbol{e}$ 与 $\boldsymbol{X}$ 和 $\boldsymbol{W}$ 相互独立，$E(\boldsymbol{e})=0$，$\mathrm{Var}(\boldsymbol{e})=\Sigma_e$.

## 2　极大似然估计

假设 $(C_i,\delta_i,\boldsymbol{X}_i,\boldsymbol{Z}_i)$ 是 $(C,\delta,\boldsymbol{X},\boldsymbol{Z})$ 的第 $i(i=1,2,\cdots,n)$ 个观测样本，假设删失时间 $C_i$ 与失效时间 $T_i$ 相互独立，则似然函数可通过下式给出：

$$L_n=\prod_{i=1}^{n}[(1-S(C_i\,|\,\boldsymbol{X}_i,\boldsymbol{Z}_i))^{\delta_i}S(C_i\,|\,\boldsymbol{X}_i,\boldsymbol{Z}_i)^{1-\delta_i}].$$

将似然函数取对数，可写为

$$\log L_n=\sum_{i=1}^{n}[\delta_i\log(1-\exp\{-\Lambda_0(\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i+\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}_i\}C_i)\})-(1-\delta_i)(\Lambda_0(\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i+\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}_i\}C_i))].$$

由于 $\Lambda_0(\cdot)$ 为分布形式未知的累积基线风险函数，直接对似然函数求极大值较困难，因此本文通过构建一个有限维的线性函数去逼近分布形式未知的无限维累积基线风险函数. 目前常用的方法有B样条、I样条、M样条、Bernstein多项式等近似方法[20]，本文根据累积基线风险函数单调非递减的性质以及I样条的单调性，使用文献[21]中的I条对累积基线风险函数进行逼近，只需控制样条系数大于0便可保证累积基线风险函数估计量的非负性和单调性. 因此未知函数可近似为

$$\Lambda_0(\cdot)\approx\Lambda_n(\cdot)=\sum_{k=1}^{K}\gamma_kI_k(\cdot),$$

其中 $K=m+p+1$，$m$ 为样条基函数内部节点的个数，通常取百分位点作为样条节点的位置，$p$ 为样条基函数的自由度. 本文参考文献[21]，选择三次样条($p=3$)，5个内部节点数($m=5$)，则在样条近似的情况下可得如下对数似然：

$$\log L_n=\sum_{i=1}^{n}\left[\delta_i\log\left(1-\exp\left\{-\left(\sum_{k=1}^{K}\gamma_kI_k(\tau_i^*)\right)\right\}\right)-(1-\delta_i)\left(\sum_{k=1}^{K}\gamma_kI_k(\tau_i^*)\right)\right],$$

其中 $\tau_i^*=\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{W}_i+\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}_i\}C_i$，$\boldsymbol{W}_i=\boldsymbol{X}_i+\boldsymbol{e}_i$.

使用极大似然估计方法对感兴趣参数求解，在用极大似然估计时首先通过对感兴趣参数求偏导得到得分方程，然后对得分方程求根进而得到感兴趣参数的估计值. 通过对似然函数求偏导得到得分方程：

$$\frac{\partial\log L_n}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{n}\left[\delta_i\left(\frac{\exp\{-A_i\}D_i\tau_i^*\boldsymbol{W}_i}{1-\exp\{-A_i\}}\right)-(1-\delta_i)D_i\tau_i^*\boldsymbol{W}_i\right],$$

$$\frac{\partial\log L_n}{\partial\boldsymbol{\alpha}}=\sum_{i=1}^{n}\left[\delta_i\left(\frac{\exp\{-A_i\}D_i\tau_i^*\boldsymbol{Z}_i}{1-\exp\{-A_i\}}\right)-(1-\delta_i)D_i\tau_i^*\boldsymbol{Z}_i\right],$$

$$\frac{\partial\log L_n}{\partial\boldsymbol{\gamma}_k}=\sum_{i=1}^{n}\left[\delta_i\left(\frac{\exp\{-A_i\}I_k(\tau_i^*)}{1-\exp\{-A_i\}}\right)-(1-\delta_i)I_k(\tau_i^*)\right],$$

其中：$A_i=\sum_{k=1}^{K}\gamma_kI_k(\tau_i^*)$；$D_i=\sum_{k=1}^{K}\gamma_kI_k'(\tau_i^*)$，$I_k'(\cdot)$为 $I_k(\cdot)$的导数.

对上述得分方程求根，即可得到感兴趣参数 $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}$ 的估计值. 由于协变量 $\boldsymbol{W}_i$ 带有测量误差，直接对得分方程进行求根得到的参数估计值可能有偏差，因此需要选择合适的方法对带误差的协变量进行误差校正. 常用的误差校正方法有回归校正[7]和模拟外推两种方法，由于本文的模型形式较复杂，因此使用模拟外推方法对误差进行校正，分为以下3个步骤.

1) 模拟步.

在该步使用重抽样的方法对带有测量误差的协变量进行重新测量得到一组新的协变量，得到的新协变量与之前带测量误差的协变量有以下关系：

$$\boldsymbol{W}_{ig}(h)=\boldsymbol{W}_i+(h\Sigma_e)^{1/2}\boldsymbol{u}_{\mathrm{ig}},$$

其中：随机变量 $\boldsymbol{u}_{ig}\sim N(0,1)$，$i=1,2,\cdots,n$，$g=1,2,\cdots,G$，$\boldsymbol{u}_{ig}$ 与观测数据独立同分布；$h\in H=\{h_1,\cdots,h_M\}$，$h$ 的取值范围为$[0,2]$；$G$ 和 $M$ 为给定的正整数.

2) 估计步.

对于给定的 $h$ 和 $g$，用估计步得到的 $\boldsymbol{W}_{ig}(h)$ 去代替实际观测到的 $\boldsymbol{W}_i$，得到新样本数据 $\{C_i,\delta_i,\boldsymbol{W}_{ig}(h),\boldsymbol{Z}_i\}_{i=1}^{n}$，将得到的新样本数据代入似然函数中，则似然函数可写为以下形式：

$$\log L_n=\sum_{i=1}^{n}\left[\delta_i\log\left(1-\exp\left\{-\left(\sum_{k=1}^{K}\gamma_kI_k(\tilde{\tau}_i)\right)\right\}\right)-(1-\delta_i)\left(\sum_{k=1}^{K}\gamma_kI_k(\tilde{\tau}_i)\right)\right],$$

其中 $\tilde{\tau}_i=\exp\{\boldsymbol{\beta}_g^{\mathrm{T}}(h)\boldsymbol{W}_{ig}(h)+\boldsymbol{\alpha}_g^{\mathrm{T}}(h)\boldsymbol{Z}_i\}C_i$. 当 $h$ 给定时，对每个 $g$ 都可得到 $\boldsymbol{\theta}=(\boldsymbol{\beta}^{\mathrm{T}},\boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}$ 的估计值，进而得到一列关于 $\boldsymbol{\theta}$ 的估计值，然后将所得到的估计值取均值得到 $\hat{\boldsymbol{\theta}}(h)=\frac{1}{G}\sum_{g=1}^{G}\hat{\boldsymbol{\theta}}_g(h)$. 当 $h$ 取遍 $H$ 中给

定的数时即可得到$\{h,\hat{\boldsymbol{\theta}}(h),h\in H\}$.

3) 外推步.

该步需要通过外推函数去寻求给定 $h$ 与估计步中得到的估计值$\hat{\boldsymbol{\theta}}(h)$之间的关系，定义外推函数为 $f(h)=a+bh+ch^2$. 基于给定的 $h$ 与估计出的$\hat{\boldsymbol{\theta}}(h)$可得目标函数 $\sum\limits_{h\in H}\{\hat{\boldsymbol{\theta}}(h)-f(h)\}^2$. 极小化该目标函数，即可得外推函数参数的估计值. 最后根据模拟外推方法[5]可知，当 $h=-1$ 时，即可得到修正后的参数估计值 $\hat{\boldsymbol{\theta}}_{\mathrm{SIMEX}}=f(-1)$.

下面讨论估计的大样本性质. 设 $\boldsymbol{\theta}_0$ 表示模型参数 $\boldsymbol{\theta}$ 的真值，根据文献[5,9]中参数估计的理论性质可知，当样本容量 $n\to\infty$时，模拟外推修正后获得的参数估计值 $\hat{\boldsymbol{\theta}}_{\mathrm{SIMEX}}$具有相合性和渐近正态性，即 $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathrm{SIMEX}}-\boldsymbol{\theta}_0)\to N(0,\Omega^{-1})$. 在实际分析中对 $\Omega$ 进行计算较复杂，因此本文采用 Bootstrap 抽样方法[22]近似计算参数估计值的方差. 假设$\{C^j,\delta^j,\boldsymbol{W}^j,\boldsymbol{Z}^j\}(j=1,2,\cdots,J)$表示样本$\{C,\delta,\boldsymbol{W},\boldsymbol{Z}\}$中有放回抽取容量为 $n$ 的样本，参数估计值 $\hat{\boldsymbol{\theta}}^{(j)}_{\mathrm{SIMEX}}(j=1,2,\cdots,J)$表示第 $j$ 个样本下获得的模拟外推修正估计，则 $\hat{\boldsymbol{\theta}}_{\mathrm{SIMEX}}$的方差可表示为

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}_{\mathrm{SIMEX}})=\frac{1}{J-1}\sum_{j=1}^{J}\left(\hat{\boldsymbol{\theta}}^{(j)}_{\mathrm{SIMEX}}-\frac{1}{J}\sum_{j=1}^{J}\hat{\boldsymbol{\theta}}^{(j)}_{\mathrm{SIMEX}}\right)^{\otimes 2},$$

其中 $\boldsymbol{a}^{\otimes 2}=\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$.

## 3　模拟研究

下面进行数值模拟以验证本文方法的有效性. 考虑不带误差的协变量服从概率为 0.5 的二项分布 $B(0.5)$，带有误差的协变量 $\boldsymbol{X}$ 分别服从标准正态分布 $N(0,1)$，取值范围为 0～3.5 的均匀分布 $U(0,3.5)$，均值为 1 的指数分布 $\mathrm{Exp}\{1\}$，协变量的误差服从正态分布 $\boldsymbol{e}\sim N(0,\sigma_e^2)$，$\boldsymbol{e}$ 的标准差为 $\sigma_e=0.3$或 $\sigma_e=0.4$，可靠性比由 $\sigma_x^2/(\sigma_x^2+\sigma_e^2)$给出，本文可靠性比为 0.9 和 0.85，删失时间 $C\sim\mathrm{Exp}(10)$. 失效时间 $T$ 由以下模型生成：

$$F(t\,|\,\boldsymbol{X},\boldsymbol{Z})=1-\exp\{-\Lambda_0(\exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}+\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}\}t)\},$$

其中 $\Lambda_0(s)=0.25s$. $\delta$ 为一个示性函数，当 $T\leqslant C$ 时 $\delta=1$，当 $T>C$ 时 $\delta=0$.

模拟计算分别产生了 200 个和 400 个独立同分布的样本，Bootstrap 次数设为 50 次，并进行 500 次循环. 在模拟外推中，$G=30$，$M=5$，$H=\{0,0.5,1,1.5,2\}$，在外推步中使用二次外推函数. 用 NAIVE 表示不考虑测量误差的估计结果；SIMEX 表示使用模拟外推的估计方法对带有误差的协变量进行误差修正后的估计结果. 用 BIAS 表示估计结果的偏差；SSE 表示样本估计值的标准差；SEE 表示用 Bootstrap 方法得到的标准差估计的平均值；CP 表示待估参数的 95%经验覆盖率. 表 1 为协变量 $\boldsymbol{X}$ 服从标准正态分布时参数的估计结果.

**表 1　协变量 $\boldsymbol{X}$ 服从标准正态分布的估计结果**

**Table 1　Estimation results of covariate $\boldsymbol{X}$ following standard normal distribution**

| 样本量 | 可靠 | 待估 | SIMEX | | | | NAIVE | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 性比 | 参数 | BIAS | SSE | SEE | CP/% | BIAS | SSE | SEE | CP/% |
| 200 | 0.90 | $\beta=1$ | 0.011 5 | 0.255 6 | 0.322 3 | 97.1 | −0.066 9 | 0.191 5 | 0.202 4 | 91.5 |
| | | $\alpha=1$ | 0.038 5 | 0.411 2 | 0.491 9 | 95.8 | 0.026 2 | 0.324 4 | 0.338 9 | 94.9 |
| | 0.85 | $\beta=1$ | 0.023 6 | 0.268 4 | 0.335 6 | 97.4 | −0.121 9 | 0.197 8 | 0.208 8 | 92.2 |
| | | $\alpha=1$ | 0.054 1 | 0.444 9 | 0.578 2 | 95.8 | −0.038 5 | 0.338 5 | 0.356 3 | 95.2 |
| 400 | 0.90 | $\beta=1$ | −0.002 0 | 0.158 6 | 0.191 9 | 97.2 | −0.059 1 | 0.121 6 | 0.130 1 | 89.0 |
| | | $\alpha=1$ | −0.014 5 | 0.288 7 | 0.328 9 | 96.4 | −0.011 5 | 0.225 2 | 0.227 7 | 93.8 |
| | 0.85 | $\beta=1$ | −0.003 2 | 0.280 8 | 0.348 5 | 97.0 | −0.121 8 | 0.191 2 | 0.205 3 | 84.3 |
| | | $\alpha=1$ | −0.026 0 | 0.494 3 | 0.605 0 | 96.3 | 0.023 0 | 0.341 5 | 0.398 3 | 94.8 |

由表 1 可见，与未修正的估计结果相比，使用 SIMEX 方法修正后的估计结果在大多数情况下都有更小的偏差和更好的 CP，且随着样本量的增大，估计的偏差也随之减小，估计结果具有相合性. 当

增加测量误差的幅度时，测量误差的影响也随之增大.

表2为协变量 $\boldsymbol{X}$ 服从均匀分布时参数的估计结果. 由表2可见：协变量服从均匀分布和协变量服从指数分布性质相同；对带误差的协变量进行修正后的偏差小于没有修正前的偏差，且随着样本量的增大，估计值的偏差有减小的趋势；样本估计值的标准差与标准差估计的平均值之间相差较小，且修正后的CP值也较好.

**表2 协变量 $X$ 服从均匀分布的估计结果**

**Table 2 Estimation results of covariate $X$ following uniform distribution**

| 样本量 | 可靠 | 待估 | SIMEX | | | | NAIVE | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 性比 | 参数 | BIAS | SSE | SEE | CP/% | BIAS | SSE | SEE | CP/% |
| 200 | 0.90 | $\beta=0.5$ | −0.001 4 | 0.225 8 | 0.255 7 | 96.2 | −0.040 8 | 0.174 1 | 0.178 6 | 92.6 |
| | | $\alpha=0.5$ | 0.019 7 | 0.419 5 | 0.576 9 | 95.8 | 0.020 9 | 0.356 2 | 0.411 4 | 93.4 |
| | 0.85 | $\beta=0.5$ | 0.009 9 | 0.222 7 | 0.258 2 | 97.0 | −0.056 9 | 0.165 7 | 0.170 9 | 90.7 |
| | | $\alpha=0.5$ | 0.012 0 | 0.431 5 | 0.502 3 | 96.3 | 0.014 8 | 0.357 9 | 0.395 5 | 93.7 |
| 400 | 0.90 | $\beta=0.5$ | −0.009 0 | 0.133 5 | 0.159 4 | 96.6 | −0.048 5 | 0.109 4 | 0.114 9 | 90.6 |
| | | $\alpha=0.5$ | −0.000 8 | 0.256 5 | 0.305 5 | 97.1 | 0.002 9 | 0.228 7 | 0.238 2 | 94.4 |
| | 0.85 | $\beta=0.5$ | −0.021 9 | 0.142 9 | 0.164 9 | 95.9 | −0.078 7 | 0.105 3 | 0.111 8 | 86.2 |
| | | $\alpha=0.5$ | 0.012 6 | 0.277 6 | 0.315 1 | 97.0 | 0.020 4 | 0.232 1 | 0.238 9 | 93.7 |

表3为协变量 $\boldsymbol{X}$ 服从指数分布时参数的估计结果. 由表3可见：对带误差的协变量进行误差修正后减少了参数估计值的偏差，与表1和表2具有相同的性质；随着样本量的增大，估计的偏差随之减少，随着误差幅度的增大，估计的偏差也随之增大，且修正后比修正前的CP值更接近95%.

**表3 协变量 $X$ 服从指数分布的估计结果**

**Table 3 Estimation results of covariate $X$ following exponential distribution**

| 样本量 | 可靠 | 待估 | SIMEX | | | | NAIVE | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 性比 | 参数 | BIAS | SSE | SEE | CP/% | BIAS | SSE | SEE | CP/% |
| 200 | 0.90 | $\beta=1$ | 0.010 8 | 0.403 6 | 0.469 1 | 92.5 | −0.176 8 | 0.257 3 | 0.276 9 | 81.2 |
| | | $\alpha=1$ | 0.001 8 | 0.453 0 | 0.514 1 | 95.8 | −0.011 6 | 0.374 2 | 0.443 7 | 93.9 |
| | 0.85 | $\beta=1$ | −0.054 9 | 0.399 8 | 0.449 3 | 93.0 | −0.272 6 | 0.239 9 | 0.248 1 | 72.1 |
| | | $\alpha=1$ | 0.017 3 | 0.480 7 | 0.617 0 | 95.8 | 0.019 6 | 0.389 5 | 0.496 1 | 96.7 |
| 400 | 0.90 | $\beta=1$ | 0.000 9 | 0.292 9 | 0.298 9 | 94.3 | −0.177 0 | 0.189 8 | 0.183 9 | 77.6 |
| | | $\alpha=1$ | 0.053 2 | 0.322 1 | 0.365 0 | 96.3 | 0.045 8 | 0.286 1 | 0.285 7 | 93.3 |
| | 0.85 | $\beta=1$ | −0.078 0 | 0.246 8 | 0.281 5 | 95.0 | −0.295 4 | 0.153 5 | 0.163 8 | 52.1 |
| | | $\alpha=1$ | 0.034 7 | 0.330 1 | 0.378 3 | 95.7 | 0.025 2 | 0.280 8 | 0.285 2 | 93.9 |

上述模拟结果表明，协变量在不同分布形式下有相同的性质，证明了本文方法的有效性. 由模拟结果可知，测量误差对参数的估计效果有较大影响，若不对测量误差进行修正，则估计结果将会产生巨大偏差. 因此对带有误差的协变量进行误差修正十分必要.

## 4 实例分析

下面将本文提出的半参数加速风险模型应用于心脑血管病死亡率的研究中，所用数据集来源于巴基斯坦费萨拉巴德心脏病学研究所和费萨拉巴德联合医院[23]. 该数据集是为研究哪些指标对患心脑血管病病人的死亡率有影响，在这项研究中个体的死亡时间不能被直接观察到，只知道个体在随访期间是否发生死亡，因此该数据为现状数据. 该数据集共收集了299个样本，用6个变量对死亡时间进行预测，其中 $T$ 表示死亡时间，$C$ 表示观察时间，$x_1$ 表示血液中血小板数量，$x_2$ 表示血液中血清钠水平，$z_1$ 表示是否有高血压，$z_2$ 表示是否吸烟，$z_3$ 表示是否患有糖尿病，$z_4$ 表示个体的性别. 在血液中血小板数量和血清钠水平的测量中，由于仪器设备的因素，常会存在一定程度的测量误差. 因此，用本文方法对带有误差的协变量进行误差修正. 为获得模型的估计结果，设模拟外推常数 $G=30$，$M=5$，$H=\{0,0.5,1,1.5,2\}$. 根据文献[8]，将测量误差的标准差设为 $\sigma_1,\sigma_2=0.25,0.5,0.75$.

表 4 列出了数据集中心脑血管病的分析结果，其中 EST 表示回归系数的估计值，SE 表示标准误差，$p$ 值为检验回归系数等于 0 的最小显著性水平. 由表 4 可见，当测量误差较小时，所有变量的参数估计值均大于 0 且对应的 $p$ 值都小于 0.05，表明这些变量对心脑血管病死亡率都具有显著性的正向影响. 在一定范围内，随着血液中血小板数量的增加或血液中血清钠水平的升高均会增加心脑血管病死亡的风险，因此控制血液中血小板数量和血清钠水平有助于降低心脑血管病死亡的风险；男性患有心脑血管病的死亡率高于女性；患有高血压、吸烟以及患有糖尿病都会增加心脑血管病死亡的风险，因此对患有心脑血管病的人群需要做到健康饮食，并进行适当的体育锻炼，进而降低个体患高血压和糖尿病的风险，吸烟人群也需要尽量戒烟，进而降低死亡的风险.

**表 4　心脑血管病分析结果**

**Table 4　Analysis results of cardiovascular and cerebrovascular diseases**

| 回归系数 | $\sigma_1$ | $\sigma_2$ | SMIEX | | | NAIVE | | |
|---|---|---|---|---|---|---|---|---|
| | | | EST | SE | $p$ 值 | EST | SE | $p$ 值 |
| $\beta_1$ | 0.25 | 0.25 | 4.231 | 0.305 | $<0.001$ | 2.458 | 0.382 | $<0.001$ |
| | 0.50 | 0.50 | 1.541 | 0.293 | $<0.001$ | | | |
| | 0.75 | 0.75 | 1.963 | 0.302 | $<0.001$ | | | |
| $\beta_2$ | 0.25 | 0.25 | 2.048 | 0.295 | $<0.001$ | 4.130 | 0.334 | $<0.001$ |
| | 0.50 | 0.50 | 0.995 | 0.298 | $<0.001$ | | | |
| | 0.75 | 0.75 | 1.433 | 0.122 | $<0.001$ | | | |
| $\alpha_1$ | 0.25 | 0.25 | 3.461 | 0.263 | $<0.001$ | 3.081 | 0.410 | $<0.001$ |
| | 0.50 | 0.50 | 4.262 | 0.786 | $<0.001$ | | | |
| | 0.75 | 0.75 | 2.529 | 0.644 | $<0.001$ | | | |
| $\alpha_2$ | 0.25 | 0.25 | 4.415 | 0.409 | $<0.001$ | 3.895 | 0.494 | $<0.001$ |
| | 0.50 | 0.50 | 2.364 | 0.536 | $<0.001$ | | | |
| | 0.75 | 0.75 | 0.321 | 0.683 | 0.638 | | | |
| $\alpha_3$ | 0.25 | 0.25 | 1.991 | 0.419 | $<0.001$ | 2.977 | 0.499 | $<0.001$ |
| | 0.50 | 0.50 | 3.069 | 0.416 | $<0.001$ | | | |
| | 0.75 | 0.75 | 1.796 | 0.826 | 0.296 | | | |
| $\alpha_4$ | 0.25 | 0.25 | 1.178 | 0.259 | $<0.001$ | 3.583 | 0.326 | $<0.001$ |
| | 0.50 | 0.50 | 3.975 | 0.507 | $<0.001$ | | | |
| | 0.75 | 0.75 | 1.216 | 0.614 | 0.048 | | | |

当测量误差增大时，系数的估计方向与测量误差较小时的估计方向一致，即随着血液中血小板数量的增高或血液中血清钠水平的升高均会增加心脑血管病死亡的风险，因此患有心脑血管病的病人需要定期到医院进行检查或者通过服用一些药物将这两个变量控制在一定范围内，从而降低死亡的风险. 男性患有心脑血管病死亡的风险高于女性. 患有高血压将会增加心脑血管病死亡的风险，因此对患有心脑血管病的人群需要关注自身血压的情况，进而降低死亡风险. 但随着测量误差的增大，吸烟和患有糖尿病对应的 $p$ 值大于 0.05，表明在实际应用中当血液中血小板数量和血液中血清钠水平存在较大误差时，会导致吸烟和患有糖尿病对心脑血管病死亡率的影响在统计学意义上是不显著的，因此在实例数据分析中对带有测量误差的协变量进行误差修正非常必要.

综上所述，本文在现状数据下，对协变量带有测量误差的加速风险模型进行了分析，其中随机误差项的分布形式未知. 先使用Ⅰ样条对分布形式未知的累积基线风险函数进行逼近，再使用模拟外推方法对带误差的协变量进行修正，最后将本文方法应用到心脑血管病的研究中. 数值模拟结果表明，使用模拟外推方法对带误差的变量进行修正可以减小参数估计值的偏差.

## 参　考　文　献

[1] PUNEKAR S R, VELCHETI V, NEEL B G, et al. The Current State of the Art and Future Trends in RAS-Targeted Cancer Therapies [J]. Nature Reviews Clinical Oncology, 2022, 19(10): 637-655.

[2] LIU T Q, YUAN X H, SUN J G. Weighted Rank Estimation of Nonparametric Transformation Models with

Case-1 and Case-2 Interval-Censored Failure Time Data [J]. Journal of Nonparametric Statistics, 2021, 33(2): 225-248.

[3] ABOUBAKAR M, KELLIL M, ROUX P. A Review of IoT Network Management: Current Status and Perspectives [J]. Journal of King Saud University: Computer and Information Sciences, 2022, 34(7): 4163-4176.

[4] ZAWACKI-RICHTER O. The Current State and Impact of Covid-19 on Digital Higher Education in Germany [J]. Human Behavior and Emerging Technologies, 2021, 3(1): 218-226.

[5] HE W Q, YI G Y, XIONG J. Accelerated Failure Time Models with Covariates Subject to Measurement Error [J]. Statistics in Medicine, 2007, 26: 4817-4832.

[6] YU M G, NAN B. Regression Calibration in Semiparametric Accelerated Failure Time Models [J]. Biometrics, 2010, 66(2): 405-414.

[7] WANG C Y, HSU L, FENG Z D, et al. Regression Calibration in Failure Time Regression [J]. Biometrics, 1997, 53(1): 131-145.

[8] ZHANG J J, HE W Q, LI H F. A Semiparametric Approach for Accelerated Failure Time Models with Covariates Subject to Measurement Error [J]. Communications in Statistics: Simulation and Computation, 2014, 43(2): 329-341.

[9] CHEN L P. Semiparametric Estimation for the Accelerated Failure Time Model with Length-Biased Sampling and Covariate Measurement Error [J]. Stat, 2018, 7(1): e209-1-e209-8.

[10] CAO Z Q, WONG M Y. Approximate Profile Likelihood Estimation for Cox Regression with Covariate Measurement Error [J]. Statistics in Medicine, 2022, 41(5): 910-931.

[11] PAN A Q, SONG X, HUANG H W. Bayesian Analysis for Partly Linear Cox Model with Measurement Error and Time-Varying Covariate Effect [J]. Statistics in Medicine, 2022, 41(23): 4666-4681.

[12] GHAPANI F, BABADI B. Two Parameter Weighted Mixed Estimator in Linear Measurement Error Models [J]. Communications in Statistics: Simulation and Computation, 2022, 51(12): 6936-6946.

[13] DONG H, OTSU T, TAYLOR L. Estimation of Varying Coefficient Models with Measurement Error [J]. Journal of Econometrics, 2022, 230(2): 388-415.

[14] LIN X Y. A Bayesian Semiparametric Accelerated Failure Time Model for Arbitrarily Censored Data with Covariates Subject to Measurement Error [J]. Communications in Statistics: Simulation and Computation, 2017, 46(1): 747-756.

[15] SHEN J S, LI Z N, YU H J, et al. Semiparametric Bayesian Inference for Accelerated Failure Time Models with Errors-in-Covariates and Doubly Censored Data [J]. Journal of Systems Science and Complexity, 2017, 30(5): 1189-1205.

[16] YI G Y, HE W Q. Methods for Bivariate Survival Data with Mismeasured Covariates under an Accelerated Failure Time Model [J]. Communications in Statistics: Theory and Methods, 2006, 35(7/8/9): 1539-1554.

[17] USOL'TSEVA O S. A Consistent Estimator in the Accelerated Failure Time Model with Censored Observations and Measurement Errors [J]. Theory of Probability and Mathematical Statistics, 2011, 82: 161-169.

[18] LI Y, LIANG M X, MAO L, et al. Robust Estimation and Variable Selection for the Accelerated Failure Time Model [J]. Statistics in Medicine, 2021, 40(20): 4473-4491.

[19] DONG X X, KONG L, WAHED A S. Accelerated Failure Time Model for Case-Cohort Design with Longitudinal Covariates Subject to Measurement Error and Detection Limits [J]. Statistics in Medicine, 2016, 35(8): 1327-1339.

[20] PERPEROGLOU A, SAUERBREI W, ABRAHAMOWICZ M, et al. A Review of Spline Function Procedures in R [J/OL]. BMC Medical Research Methodology, (2019-03-06)[2024-03-10]. https://doi.org/10.1186/s12874-019-0666-3.

[21] RAMSAY J O. Monotone Regression Splines in Action [J]. Statistical Science, 1988, 3(4): 425-441.

[22] JOHNSON R W. An Introduction to the Bootstrap [J]. Teaching Statistics, 2001, 23(2): 49-54.

[23] CHICCO D, JURMAN G. Machine Learning Can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone [J/OL]. BMC Medical Informatics and Decision Making, (2020-02-03)[2023-11-20]. https://doi.org/10.1186/s12911-020-1023-5.

(责任编辑:赵立芹)