# 带离散辅助协变量的 AFT 模型的 B-J 估计

郭丽莎[1],金凌辉[2,3,*]

(1 中南民族大学 数学与统计学学院,武汉 430074;2 武汉科技大学 城市学院,武汉 430083;
3 武汉大学 数学与统计学院,武汉 430072)

**摘 要** 为处理协变量随机缺失的 AFT 模型的参数估计问题,首先利用离散辅助协变量对缺失的协变量进行了插补,再结合 Buckley-James 方法提出了带辅助信息的 AFT 模型的一种参数估计方法. 此方法作为 B-J 估计在不完全协变量情形下的一个推广,无须指定模型误差项的分布,在应用上有一定的便利性. 数据模拟表明: 此方法具有较好的估计效果.

**关键词** AFT 模型; 辅助协变量; B-J 估计; 生存分析

**中图分类号** O212.2 **文献标识码** A **文章编号** 1672-4321(2017)04-0146-03

## B-J Estimator of AFT Model with Discrete Auxiliary Covariates

*Guo Lisha*[1], *Jin Linghui*[2,3]

(1 College of Mathematics and Statistics, South-Central University for Nationalities, Wuhan 430074, China;
2 City College of Wuhan University of Science and Technology, Wuhan 430083, China;
3 College of Mathematics and Statistics, Wuhan University, Wuhan 430072, China)

**Abstract** To handle the parametric estimate problem of AFT model when covariates missing at random, we used the discrete auxiliary covariates to impute the missing covariates, then utilized the estimate method of Buckley-James to get the estimator of unknown parameter. The distribution of the error is unspecified, so our method is convenient in practical application. The results of simulation showed that our method has good effect in estimation.

**Keywords** AFT model; auxiliary covariates; B-J estimator; survival analysis

在临床医学研究中,由于财力所限或技术原因,有些目标协变量信息很难搜集或存在测量误差. 然而,通常我们可以比较容易获得与目标协变量有着较高关联度的辅助协变量,从而利用这些辅助协变量进行统计推断,这样的问题称之为辅助协变量问题. 关于这个问题已经有了很多重要的研究成果,如文献[1,2]研究了 Cox 模型的辅助协变量问题,文献[3,4]探讨了加法危险率模型的辅助协变量问题. 相对于上述两种模型,加速失效时间模型(AFT Model)的协变量对失效时间,即响应变量的解释更为直接,因此也有着广泛应用. 在这里我们尝试探讨 AFT 模型的辅助协变量问题.

## 1 AFT 模型和 B-J 估计

假定有一个由 $n$ 个个体构成的随机样本,令 $T_i$ 和 $C_i(i=1,2,\cdots,n)$ 分别为第 $i$ 个个体的失效时间和删失时间,$X_i$ 为相应的 $p$ 维协变量. 一般地,假定在给定 $X_i$ 的条件下,$T_i$ 和 $C_i$ 独立,于是我们能观测到的数据为 $\{\tilde{T}_i,\delta_i,X_i\}$,其中 $\tilde{T}_i=\min(T_i,C_i)$,$\delta_i=I(T_i\leqslant C_i)$,这里 $I(\cdot)$ 为示性函数. 记 $Y_i=\log(T_i)$,则 AFT 模型的形式为: $Y_i=\log(T_i)=\beta^{\mathrm{T}}X_i+\varepsilon_i$,$i=1,2,\cdots,n$. 这里 $\beta$ 是 $p$ 维待估参数,$\varepsilon_i$ 为独立同分布的误差项,但在接下来所讨论的方法中其分布函数 $F$ 无须特别指定,仅要求其均值为零且方差有限. 对于这个模型,Buckley 和 James[5,6] 给出了一种基于最小二乘的参数估计方法,称为 B-J 估计,具体如下:

首先注意到,若响应变量没有删失,回归参数的估计应是最小化 $n^{-1}\sum_{i=1}^{n}(Y_i-\beta^{\mathrm{T}}X_i)^2$ 的结果,由此可以得到 $\beta$ 的估计方程:

$$\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\beta^{\mathrm{T}}X_i)=0. \qquad (1)$$

**收稿日期** 2017-06-10 * **通讯作者** 金凌辉,研究方向: 生物统计,E-mail: jinlinghui163@163.com
**作者简介** 郭丽莎(1980-),女,讲师,博士,研究方向: 生物统计,E-mail: lsgscuec@hotmail.com
**基金项目** 全国统计科学研究项目(2014LY102);湖北省教育厅科研计划项目(B2017427);中央高校基本科研业务费专项资金资助项目(CZW15066);武汉科技大学城市学院重点科研项目(2016CYZDKY003)

这里 $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$. 为处理 $Y_i$ 的删失,Buckley 和 James 提出用 $Y_i^* = Y_i\delta_i + E[Y_i \mid Y_i > \log(C_i)](1-\delta_i)$ 代替 $Y_i$,容易验证 $Y_i^*$ 满足估计方程(1). 然而,$Y_i^*$ 中所包含的条件数学期望未知,于是 Buckley 和 James 用其估计:

$$\hat{E}[Y_i \mid Y_i > \log(C_i)] = \beta^{\mathrm{T}}X_i + \frac{\int_{e_i(\beta)}^{\infty} u\mathrm{d}\hat{F}_\beta(u)}{1-\hat{F}_\beta(e_i(\beta))}, \quad (2)$$

来代替,其中 $e_i(\beta) = \bar{Y}_i - \beta^{\mathrm{T}}X_i$,$\bar{Y}_i = \log(T_i)$,而 $\hat{F}_\beta$ 则是 $F$ 基于$\{e_i(\beta),\delta_i\}$ 的 K-M 估计,于是可以得到 $Y_i^*$ 的估计为 $\hat{Y}_i^* = Y_i\delta_i + \hat{E}[Y_i \mid Y_i > \log(C_i)](1-\delta_i)$.

定义估计方程:

$$U(\beta,b) = \sum_{i=1}^{n}(X_i - \bar{X})(\hat{Y}_i^*(b) - \beta^{\mathrm{T}}X_i) = 0, \quad (3)$$

或:

$$U(\beta,b) = \sum_{i=1}^{n}(X_i - \bar{X})\{\hat{Y}_i^*(b) - \bar{Y}_i^*(b) - \beta^{\mathrm{T}}(X_i - \bar{X})\} = 0, \quad (4)$$

其中 $\bar{Y}^*(b) = n^{-1}\sum_{i=1}^{n}\hat{Y}_i^*(b)$,于是未知参数 $\beta$ 的 B-J 估计是方程 $U(\beta,\beta) = 0$ 的根. 文献[7]证明了这个估计的渐近性质.

## 2　带辅助协变量的 B-J 估计

若协变量不能完全观测,我们希望能利用可以观测到的辅助协变量对未知参数进行更有效的估计. 一般地,把观测队列中目标协变量和相应的辅助协变量都能观测到的子集称为确信集(Validation Set),记为 $V$;而把仅能观测到辅助协变量、观测不到目标协变量的子集称为非确信集(Non-validation Set),记为 $\bar{V}$. 令 $A_i$ 为相应于协变量 $X_i$ 的辅助协变量,这里假定 $A_i$ 是离散的,其取值及相应的概率为 $Pr(A = a_k) = p_{a_k}$,$k = 1, 2, \cdots, q$,这里 $\sum_{k=1}^{q} p_{a_k} = 1$. 于是 $V$ 中的样本为$\{\tilde{Y}_i,\delta_i,X_i,A_i\}$,$\bar{V}$中的样本为$\{\tilde{Y}_i,\delta_i,A_i\}$. 如果第 $i$ 个个体属于 $V$,则其对应的观测值显然满足模型(1),若第 $i$ 个个体属于 $\bar{V}$,此时协变量 $X_i$ 不能观测,仅能观测到相应的辅助协变量 $A_i$,借助于文献[1]的思想,我们用 $\tilde{X}_i = E[X_i \mid A_i]$ 来对缺失的协变量 $X_i$ 进行插补,不难验证 $\tilde{X}_i$ 依然保持 AFT 模型的结构. 进一步,定义示性变量 $\eta_i = I(i \in V)$,从而有 $X_i^* = X_i\eta_i + \tilde{X}_i(1-\eta_i)$ 满足模型(1).

然而,在实际运用时,我们很难得到 $\tilde{X}_i$ 的值,不过可以用其经验估计:

$$\hat{\tilde{X}}_i = \frac{\sum_{j\in V} I(A_j = A_i)X_j}{\sum_{j\in V} I(A_j = A_i)}$$

来代替,因此 $X_i^*$ 的估计应为 $\hat{X}_i^* = X_i\eta_i + \hat{\tilde{X}}_i(1-\eta_i)$,从而得到一个"估计的"估计方程:

$$\hat{U}(\beta,b) = \sum_{i=1}^{n}(\hat{X}_i^* - \bar{\hat{X}}^*)\{\hat{Y}_i^*(b) - \bar{Y}_i^*(b) - \beta^{\mathrm{T}}(\hat{X}_i^* - \bar{\hat{X}}^*)\} = 0. \quad (5)$$

由于 $\hat{U}(\beta,\beta)$ 关于 $\beta$ 既不连续也不单调,因此求 $\beta$ 的估计值并不容易,通常使用迭代算法. 首先可在估计方程 $\hat{U}(\beta,b)$ 中给定一个初始值 $b$ 去求解 $\beta$,从而得到 $\beta = L(b)$,其中:

$$L(b) = \{\sum_{i=1}^{n}(\hat{X}_i^* - \bar{\hat{X}}^*)^{\otimes 2}\}^{-1}\{\sum_{i=1}^{n}(\hat{X}_i^* - \bar{\hat{X}}^*)\cdot[\hat{Y}_i^*(b) - \bar{Y}_i^*(b)]\},$$

这里符号"$\otimes$"表示对任意向量 $a$,有 $a^{\otimes 0} = 1$,$a^{\otimes 1} = a$,$a^{\otimes 2} = aa^{\mathrm{T}}$. 于是可以得到一个迭代运算 $\hat{\beta}_{(m)} = L(\hat{\beta}_{(m-1)})$,其中 $m \geqslant 1$,若 $\|\hat{\beta}_{(m)} - \hat{\beta}_{(m-1)}\|$ 小于某个指定的数,我们便得到最终的估计值 $\hat{\beta}_E = \hat{\beta}_{(m)}$. 然而,这个迭代也有可能出现不收敛的情形,它们可能在两个值之间摆动,Buckley 和 James 建议取这些估计值的平均值作为 $\beta$ 的估计值;另外初始值的选取也很重要,一般地,若初始值具有相合性和渐近正态性,则 $\hat{\beta}_{(m)}$ 也会有相合性和渐近正态性,关于初始值的选取可参见文献[8].

## 3　数值模拟

为检验所提出的方法的估计效果,利用 R 软件进行数值模拟. 以下用 $\hat{\beta}_E$ 表示用我们的方法得到的未知参数估计值;$\hat{\beta}_V$ 表示基于确信集的参数估计值,即只使用能完全观测到协变量的那一部分个体的信息得到的 B-J 估计;$\hat{\beta}_N$ 为用 Naive 方法估计的结果,即直接用辅助协变量代替缺失协变量所得到的 B-J 估计.

按照以下方法生成数据: 二维协变量 $X_i$ 的两个分量独立地由 $[0,4]$ 上的均匀分布产生,随机误差 $\varepsilon_i$ 则由标准正态分布生成,失效时间通过 $Y_i = \log(T_i) = \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$ 得到. 删失时间的对数 $\log(C_i)$ 来自 $[0,c]$ 上的均匀分布,其中 $c$ 用于调节删失比例. 我们从发生概率为 $\rho$ 的伯努利分布生成示性变量 $\eta$,从而得到确信集所占的比例. 为生成辅助协变量,对 $X_i$ 的第一个分量 $X_{i1}$,首先生成随机变量 $W_{i1} = X_{i1} + e$,其中 $e \sim N(0, \sigma_e^2)$,参数 $\sigma_e$ 用于调节 $X_{i1}$ 和 $W_{i1}$ 之间的关联程度. 于是,辅助协变量 $A_i$ 的第一个分量 $A_{i1}$ 则由 $W_{i1}$ 落入区间 $(-\infty, Q_{i1}]$,$(Q_{i1}, Q_{i2}]$,$(Q_{i2}, Q_{i3}]$ 或$(Q_{i3}, +\infty]$,相应地

取值1,2,3和4,这里 $Q_{i1}$、$Q_{i2}$ 和 $Q_{i3}$ 分别为 $W_{i1}$ 的1/4,1/2和3/4分位数;用同样的方法,可以得到 $A_i$ 的第二个分量 $A_{i2}$. 由文献[8]可知,基于秩方法(rank-based method)所得到的完整协变量下AFT模型的参数估计值具有相合性和渐近正态性,因此也选取这个估计值作为迭代初值.

把参数的真值设置为 $\beta = (\beta_1, \beta_2) = (\ln 1.5, \ln 2) = (0.405, 0.693)$,样本量 $n$ 设置为200,重复计算1000次. 表1为删失率取0.2,$\rho$ 分别取0.7和0.5以及 $\sigma_e$ 设为0.2和0.6时的参数估计结果. 其中Est为1000次参数估计的平均值,SD为估计值的平均样本标准差,SE是估计的标准差的平均值,其结果通过50次重抽样的bootstrap方法计算所得. 95% CP表示置信度为95%置信区间覆盖参数真值的覆盖率.

表1 $n = 200$,删失率为20%的模拟结果

Tab.1 Simulation results for $n = 200$ and censoring rate is 20%

| $\rho$ | $\sigma_e$ | $\hat{\beta}$ | $\beta_1$ | | | | $\beta_2$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Est | SD | SE | 95% CP | Est | SD | SE | 95% CP |
| 0.7 | | $\hat{\beta}_V$ | 0.407 | 0.064 | 0.064 | 0.943 | 0.687 | 0.066 | 0.064 | 0.935 |
| | 0.2 | $\hat{\beta}_N$ | 0.368 | 0.059 | 0.057 | 0.897 | 0.647 | 0.059 | 0.057 | 0.857 |
| | 0.6 | $\hat{\beta}_N$ | 0.361 | 0.062 | 0.059 | 0.862 | 0.645 | 0.062 | 0.059 | 0.857 |
| | 0.2 | $\hat{\beta}_E$ | 0.402 | 0.056 | 0.055 | 0.951 | 0.691 | 0.057 | 0.056 | 0.955 |
| | 0.6 | $\hat{\beta}_E$ | 0.394 | 0.060 | 0.059 | 0.939 | 0.693 | 0.060 | 0.059 | 0.952 |
| 0.5 | | $\hat{\beta}_V$ | 0.407 | 0.079 | 0.076 | 0.952 | 0.686 | 0.079 | 0.077 | 0.942 |
| | 0.2 | $\hat{\beta}_N$ | 0.351 | 0.061 | 0.059 | 0.837 | 0.623 | 0.060 | 0.058 | 0.765 |
| | 0.6 | $\hat{\beta}_N$ | 0.340 | 0.065 | 0.062 | 0.805 | 0.621 | 0.064 | 0.062 | 0.770 |
| | 0.2 | $\hat{\beta}_E$ | 0.401 | 0.057 | 0.057 | 0.948 | 0.692 | 0.058 | 0.058 | 0.951 |
| | 0.6 | $\hat{\beta}_E$ | 0.389 | 0.063 | 0.063 | 0.939 | 0.695 | 0.064 | 0.063 | 0.943 |

从表1中可得到如下结论:(1)对于两个待估参数,用我们的方法和用基于确信集的方法实质上都是无偏的,但用Naive方法得到的结果却是有偏的;(2)因为我们的方法使用的样本信息比基于确信集的方法多,所以很自然我们的方法比后者更有效($\hat{\beta}_E$ 的SD和SE都比 $\hat{\beta}_V$ 相应的值小);(3)当辅助协变量与主要协变量的关联度较低时($\sigma_\varepsilon$ 较大),用Naive方法和用我们的方法所得到的估计精确度都会降低,但是实际上这种偏差会随着样本量的增大而减小. 为此,将上述模拟计算中的样本量调整为 $n = 50$ 和 $n = 500$ 以比较估计偏差的变化,比较结果见表2.

表2 不同样本量下95% CP的比较结果

Tab.2 Comparison results of 95% CP under different sample size

| $\hat{\beta}$ | $\rho$ | $\sigma_e$ | $n = 50$ | $n = 200$ | $n = 500$ |
|---|---|---|---|---|---|
| $\hat{\beta}_{1E}$ | 0.5 | 0.2 | 0.926 | 0.948 | 0.951 |
| | | 0.6 | 0.905 | 0.939 | 0.944 |
| $\hat{\beta}_{2E}$ | 0.5 | 0.2 | 0.932 | 0.951 | 0.953 |
| | | 0.6 | 0.909 | 0.943 | 0.949 |

这里 $\hat{\beta}_{1E}$ 和 $\hat{\beta}_{2E}$ 分别为用我们的方法得到的 $\beta_1$,$\beta_2$ 的估计值,通过比较不难发现,随着样本量的增加,置信度为95%的置信区间覆盖参数真值的覆盖率会随之上升,也就是说估计的偏差会逐渐减小.

## 4 结语

本文基于Buckley和James的方法,给出了带离散辅助协变量的AFT模型的B-J估计. 数据模拟的结果表明:我们所提出的方法有着较好的估计效果. 这种方法相对于传统的基于似然函数的方法而言,不需要对误差项的分布进行指定,在使用上有一定的便利性. 另外,为充分利用辅助信息,也可考虑使用连续的辅助协变量,但这方面的已有结果主要还是集中在COX模型上,关于AFT模型还有许多方面值得进一步研究.

## 参 考 文 献


[1] Zhou H, Pepe M S. Auxiliary covariate data in failure time regression analysis [J]. Biometrika, 1995, 82(1): 139-149.

[2] Liu Y, Zhou H, Cai J. Estimated pseudopartial-likelihood method for correlated failure time data with auxiliary covariates [J]. Biometrics, 2009, 65: 1184-1193.

[3] Kulich M, Lin D. Additive hazards regression with covariate measurement error [J]. Journal of the American Statistical Association, 2000, 95: 238-248.

[4] Shi X, Liu Y, Wu Y. Continuous auxiliary covariate in additive hazards regression for survival data [J]. Journal of Systems Science & Complexity, 2014, 26(6): 1247-1262.

[5] Buckley J, James I. Linear regression with censored data [J]. Biometrika, 1979, 66(3): 429-436.

[6] James I, Smith J. Consistency results for linear regression with censored data [J]. Annals of Statistics, 1984, 12(2): 145-146.

[7] Lai T, Ying Z. Large sample theory of a modified Buckley-James estimator for regression analysis with censored data [J]. Annals of Statistics, 1991, 19(3): 1370-1402.

[8] Jin Z, Lin D, Ying Z. On least-squares regression with censored datas [J]. Biometrika, 2006, 93(1): 147-161.