Advances in Applied Mathematics 应用数学进展, 2022, 11(2), 836-844
Published Online February 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.112089

Hans汉斯

加性乘积风险率模型在聚类失效时间数据中的应用

康泽宇，亢芳圆

北京信息科技大学理学院，北京

收稿日期：2022年1月23日；录用日期：2022年2月21日；发布日期：2022年2月28日

摘　要

论文对聚类失效时间数据建立了加性－乘积风险率函数模型，根据模型创建了估计方程，应用牛顿迭代法求得参数向量的估计，并给出估计量的相合性和渐近正态性的证明。通过数值模拟验证了估计量在有限样本下的表现，最后分析了慢性肉芽肿病风险率函数的影响因素和因素的影响方式。

关键词

聚类失效时间，估计方程，乘积加性风险率函数，数值模拟，数据分析

Application of Additive Multiplicative Hazard Rate Model in Clustered Failure Time Data

Zeyu Kang, Fangyuan Kang

School of Applied Science, Beijing Information Science & Technology University, Beijing

Received: Jan. 23rd, 2022; accepted: Feb. 21st, 2022; published: Feb. 28th, 2022

Abstract

In this paper, an additive multiplicative hazard rate function model is established for clustered failure time data, an estimation equation is created according to the model, the estimation of parameter vector is obtained by Newton iterative method, and the consistency and asymptotic normality of estimators are proved. The performance of the estimator under limited samples is verified by numerical simulation. Finally, the influencing factors and ways of the hazard rate function of chronic granulomatosis are analyzed.

文章引用：康泽宇，亢芳圆．加性乘积风险率模型在聚类失效时间数据中的应用[J]．应用数学进展，2022，11(2): 836-844. DOI: 10.12677/aam.2022.112089

Keywords

Clustered Failure Time, Estimation Equation, Additive Multiplivative Hazard Rate Function, Numerical Simulation, Data Analysis

1. 引言

生存分析在生物医学，工程中应用广泛。比如我们经常关心患某种疾病的病人的生存时间，工业产品的使用寿命，以及广义的生存时间，如某个感兴趣事件发生的时间。生存时间研究方法有非参数分析，参数分析方法和半参数分析方法。半参数分析方法的灵活性，稳健性和有效性介于其它两种方法之间，所以是最常用的一种方法。在生存分析数据的收集中，常常会遇到聚类数据，也就是个体之间不是完全独立的。比如数据是来自于各个医疗中心，或者是以家庭为单位收集的，还有的数据是一个个体的事件的重复发生。来自同一医疗中心的个案或同一家庭的个案或同一个体的多次复发数据，叫做成簇数据或聚类数据。同一个聚类的数据是有关系的，不同的类之间相互独立。

聚类失效时间的研究方法有边际建模方法和混合效应建模方法。在边际建模中，对同一个类的每个个体建立边际模型，建模时，对个体之间的关系不作任何假定，而在模拟部分，对同一类内个体间的关系进行一些假定[1] [2] [3]。在混合效应模型中，对同一类的个体用一个共同的脆弱变量来说明个体之间的相关性[4] [5]。以上文献大部分围绕着Cox模型[6]和加性风险模型。实际上加性乘积模型具有更强的灵活性。文献[7]提出一个广义加性乘积风险模型，以后的学者在此基础上进行进一步推广。文献[8]对带有辅助生存信息的加性乘积模型进行了统计推断。文献[9]对复发事件和终止事件进行联合分析，假设终止事件的风险率函数服从加性乘积模型。

加性乘积模型在聚类失效时间中的应用比较少，这篇论文是将加性乘积模型应用到聚类失效时间中。分析影响慢性肉芽肿病风险率函数的因素，以及影响方式。论文第一部分介绍了模型形式，估计方法和渐近理论证明，第二部分进行了数值模拟，验证提出的方法，第三部分对实际数据进行分析，第四部分是论文的总结和展望。

2. 模型和估计过程

2.1. 模型

首先把全部对象根据某个属性聚成n个簇。设第i个类中有n_i个对象，其中$i=1,2,\cdots,n$，并用T_{ij}表示第i个簇中第j个对象的失效时间，C_{ij}为右删失时间，$X_{ij}=\min\left(T_{ij},C_{ij}\right)$表示观察到的生存时间，$\delta_{ij}=I\left(T_{ij}\le C_{ij}\right)$表示第$i$类中第$j$个对象是否删失，$Z_{ij}$，$W_{ij}$为协变量向量。$\tau$表示试验终止时间。

T_{ij}的风险率函数的定义如下：

$$\lambda_{ij}\left(t\right)=\lim_{\Delta t\to 0}\frac{P\left(t\le T_{ij}<t+\Delta t\middle| T_{ij}\ge t\right)}{\Delta t}$$

对风险率函数$\lambda_{ij}\left(t\right)$建立加性乘积模型，给定协变量$Z_{ij}$，$W_{ij}$的情况下，假设风险率函数满足：

$$\lambda_{ij}\left(t\middle|Z_{ij},W_{ij}\right)=\lambda_0\left(t\right)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta_0}+W_{ij}^{\mathrm{T}}\gamma_0 \tag{1}$$

式中：$\lambda_0\left(t\right)$是基准风险率函数；β_0是协变量Z_{ij}的效应，表示对风险率函数的乘积效应；γ_0是是协变量W_{ij}的效应，表示对风险率函数的加性效应。

2.2. 模型估计过程

引入潜在的计数过程$N_{ij}^{*}\left(t\right)=I\left(T_{ij}\le t\right)$和观测到的计数过程$N_{ij}\left(t\right)=\delta_{ij}I\left(T_{ij}\le t\right)$以及风险过程$Y_{ij}\left(t\right)=I\left(T_{ij}>t\right)\delta_{ij}$，根据风险率函数模型和计数过程的定义，得

$$E\left(\mathrm{d}N_{ij}\left(t\right)\middle|Z_{ij},W_{ij}\right)=Y_{ij}\left(t\right)\left(\lambda_0\left(t\right)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta_0}+W_{ij}^{\mathrm{T}}\gamma_0\right)\mathrm{d}t$$

进一步得到以下的零均值过程：$\mathrm{d}M_{ij}\left(t\right)=\mathrm{d}N_{ij}\left(t\right)-Y_{ij}\left(t\right)\left[\lambda_0\left(t\right)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta_0}+W_{ij}^{\mathrm{T}}\gamma_0\right]\mathrm{d}t$。

根据文献[10]中的广义估计方程的思想，我们建立以下的估计方程：

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^t\mathrm{d}N_{ij}\left(u\right)-Y_{ij}\left(u\right)\left[\lambda_0\left(u\right)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}+W_{ij}^{\mathrm{T}}\gamma\right]\mathrm{d}u=0 \tag{1}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau Z_{ij}\left[\mathrm{d}N_{ij}\left(t\right)-Y_{ij}\left(t\right)\left[\lambda_0\left(t\right)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}+W_{ij}^{\mathrm{T}}\gamma\right]\mathrm{d}t\right]=0 \tag{2}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau W_{ij}\left[\mathrm{d}N_{ij}\left(t\right)-Y_{ij}\left(t\right)\left[\lambda_0\left(t\right)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}+W_{ij}^{\mathrm{T}}\gamma\right]\mathrm{d}t\right]=0 \tag{3}$$

由(1)式可得：

$$\Lambda_0\left(t\right)=\int_0^t\lambda_0\left(u\right)\mathrm{d}u=\int_0^t\frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\mathrm{d}N_{ij}\left(u\right)-Y_{ij}\left(u\right)W_{ij}^{\mathrm{T}}\gamma\mathrm{d}u}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}Y_{ij}\left(u\right)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}},$$

将$\Lambda_0\left(t\right)$代入式(2)进一步计算得

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau\left(Z_{ij}-\bar{Z}\right)\left(\mathrm{d}N_{ij}\left(t\right)-Y_{ij}\left(t\right)W_{ij}^{\mathrm{T}}\gamma\mathrm{d}t\right)=0 \tag{4}$$

其中$\bar{Z}=\dfrac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}Z_{ij}Y_{ij}\left(t\right)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}Y_{ij}\left(t\right)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}}$。

将$\Lambda_0\left(t\right)$代入式(3)进一步计算得

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau\left(W_{ij}-\bar{W}\right)\left(\mathrm{d}N_{ij}\left(t\right)-Y_{ij}\left(t\right)W_{ij}^{\mathrm{T}}\gamma\mathrm{d}t\right)=0 \tag{5}$$

其中$\bar{W}=\dfrac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}W_{ij}Y_{ij}\left(t\right)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}Y_{ij}\left(t\right)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}}$。

接下来，我们应用牛顿迭代法，对方程(4)和(5)进行求解，得到β,γ的估计。应用牛顿迭代法时，需要求出方程(4)和(5)左边表达式对β,γ的导函数。记：方程(4)的左边表达式为$U_1\left(\beta,\gamma\right)$，方程(5)的左边表

达式为$U_2(\beta,\gamma)$。

$$\frac{\partial U_1(\beta,\gamma)}{\partial \beta^T}=-\int_0^\tau\left(\frac{\sum_{i=1}^n\sum_{j=1}^{n_i}Z_{ij}Z_{ij}^{\mathrm{T}}Y_{ij}\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}}{\sum_{i=1}^n\sum_{j=1}^{n_i}Y_{ij}(t)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}}-\frac{\sum_{i=1}^n\sum_{j=1}^{n_i}Z_{ij}Y_{ij}(t)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}\sum_{i=1}^n\sum_{j=1}^{n_i}Z_{ij}^{\mathrm{T}}Y_{ij}(t)\mathrm{e}^{Z_{ij}^{\mathrm{T}}}\beta}{\left(\sum_{i=1}^n\sum_{j=1}^{n_i}Y_{ij}(t)\mathrm{e}^{Z_{ij}^{T}\beta}\right)^2}\right)\sum_{i=1}^n\sum_{j=1}^{n_i}\left(Y_{ij}(t)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}\lambda_0(t)\mathrm{d}t\right)$$

$$=-\int_0^\tau\sum_{i=1}^n\sum_{j=1}^{n_i}Z_{ij}Y_{ij}(t)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}\left(Z_{ij}-\bar{Z}\right)^{\mathrm{T}}\lambda_0(t)\mathrm{d}t+o_p(n)$$

$$=-\int_0^\tau\sum_{i=1}^n\sum_{j=1}^{n_i}\left(Z_{ij}-\bar{Z}\right)Y_{ij}(t)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}Z_{ij}^{\mathrm{T}}\lambda_0(t)\mathrm{d}t+o_p(n)$$

类似的，可以计算出：$\frac{\partial U_1(\beta,\gamma)}{\partial \gamma^T}=-\int_0^\tau\sum_{i=1}^n\sum_{j=1}^{n_i}\left(Z_{ij}-\bar{Z}\right)Y_{ij}W_{ij}^{\mathrm{T}}\mathrm{d}t$，

$$\frac{\partial U_2(\beta,\gamma)}{\partial \beta^T}=-\sum_{i=1}^n\sum_{j=1}^{n_i}\int_0^\tau\left(W_{ij}-\bar{W}\right)Y_{ij}(t)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}Z_{ij}^{\mathrm{T}}\lambda_0(t)\mathrm{d}t+o_p(n)$$

$$\frac{\partial U_2(\beta,\gamma)}{\partial \gamma^T}=-\int_0^\tau\sum_{i=1}^n\sum_{j=1}^{n_i}\left(W_{ij}-\bar{W}\right)Y_{ij}W_{ij}^{\mathrm{T}}\mathrm{d}t$$

把这四项写成矩阵形式，并记：$\hat{A}=\begin{pmatrix}\sum_{i=1}^n\sum_{j=1}^{n_i}\int_0^\tau\left(Z_{ij}-\bar{Z}\right)Y_{ij}(t)\left(\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}Z_{ij}^{\mathrm{T}}\lambda_0(t)\mathrm{d}t,W_{ij}^{\mathrm{T}}\mathrm{d}t\right)\\ \sum_{i=1}^n\sum_{j=1}^{n_i}\int_0^\tau\left(W_{ij}-\bar{W}\right)Y_{ij}(t)\left(\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}Z_{ij}^{\mathrm{T}}\lambda_0(t)\mathrm{d}t,W_{ij}^{\mathrm{T}}\mathrm{d}t\right)\end{pmatrix}$

其中$\theta=\left(\beta^{\mathrm{T}},\gamma^{\mathrm{T}}\right)^{\mathrm{T}}=\begin{pmatrix}\beta\\ \gamma\end{pmatrix}$，

$$\theta_0=\left(\beta_0^{\mathrm{T}},\gamma_0^{\mathrm{T}}\right)^{\mathrm{T}}=\begin{pmatrix}\beta_0\\ \gamma_0\end{pmatrix},$$

$$A(\theta)=-E\left\{\int_0^\tau\begin{pmatrix}Z_{ij}-\bar{z}\\ W_{ij}-\bar{w}\end{pmatrix}\left(\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}Z_{ij}^{\mathrm{T}}\lambda_0(t)\mathrm{d}t,W_{ij}^{\mathrm{T}}\mathrm{d}t\right)Y_{ij}(t)\right\}$$

其中$\bar{z},\bar{w}$是$\bar{Z},\bar{W}$的极限，由大数定律可知$\frac{1}{n}\hat{A}(\theta)\xrightarrow{p}A(\theta)$。

2.3. 估计量的渐近性质

首先给出一些正则条件。

(C1) 协变量Z_{ij},W_{ij}, $j=1,2,\cdots,n_i;i=1,\cdots,n$.是有界的；

(C2) n组观测$\left\{Z_{ij},W_{ij},T_{ij}\wedge C_{ij},\delta_{ij},Y_{ij}(\cdot),N_{ij}(\cdot),j=1,\cdots,n_i\right\}$， $i=1,\cdots,n$.是相互独立的；

(C3) $A(\theta)$在θ_0的邻域Ⓗ上可逆。

定理 1：在以上正则条件下，方程(4)，(5)的解在θ_0邻域Ⓗ内存在且唯一，将这个解记作：$\hat{\theta}=\begin{pmatrix}\hat{\beta}\\ \hat{\gamma}\end{pmatrix}$，且$\hat{\theta}$是$\theta_0$的相和估计。

证明：记

$$\mathrm{d}M_{ij}(t)=\mathrm{d}N_{ij}(t)-Y_{ij}(t)\left(\lambda_0(t)\exp\left(Z_{ij}^{\mathrm{T}}\beta\right)+W_{ij}^{\mathrm{T}}\gamma\right)\mathrm{d}t,$$

$$U(\theta)=\begin{pmatrix}U_1(\theta)\\U_2(\theta)\end{pmatrix}=\begin{pmatrix}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^{\tau}\left(Z_{ij}-\bar{Z}\right)\mathrm{d}M_{ij}(t)\\\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^{\tau}\left(W_{ij}-\bar{W}\right)\mathrm{d}M_{ij}(t)\end{pmatrix}。$$

可以证明$\frac{1}{n}U\left(\theta_0\xrightarrow{p}0\right)$。由于$\frac{\partial\frac{1}{n}U(\theta)}{\partial\theta^{\mathrm{T}}}=\frac{1}{n}\hat{A}(\theta)$，再根据大数定律，以及$\hat{A}(\theta)$关于$\theta$的连续性，可知在$\theta_0$邻域内，$\frac{1}{n}\hat{A}(\theta)$依概率一致收敛到$A(\theta)$，由假设：$A(\theta)$在$\theta_0$邻域内可逆，这样当$n$充分大的时候，$\frac{1}{n}\hat{A}(\theta)$在$\theta_0$邻域内可逆。由逆函数定理(Rudin, 1976) [11]，可以证明$U(\theta)$在θ_0邻域内存在唯一解$\hat{\theta}$。再由中值定理，在θ_0邻域内存在一个点θ^*使得，$\frac{1}{n}U\left(\hat{\theta}\right)-\frac{1}{n}U(\theta_0)=\frac{1}{n}A\left(\theta^*\right)\left(\hat{\theta}-\theta_0\right)$，当$n$充分大时，等号左边依概率收敛于0，等号右边的$\frac{1}{n}A\left(\theta^*\right)$可逆，因此当$n$充分大时$\hat{\theta}\xrightarrow{p}\theta_0$。

定理 2：在正则条件下$\hat{\theta}$渐近服从$(p+q)$维的正态分布，$\sqrt{n}\left(\hat{\theta}-\theta_0\right)$渐近服从$N_{p+q}(0,\Sigma)$，其中

$$\Sigma=A^{-1}(\theta_0)E\left(\xi_i^{\otimes 2}\right)A^{-1}(\theta_0)^{\mathrm{T}}。$$

证明：根据函数中值定理，有

$$\frac{1}{\sqrt{n}}U\left(\hat{\theta}\right)-\frac{1}{\sqrt{n}}U(\theta_0)=\frac{1}{n}\hat{A}(\theta_0)\left(\sqrt{n}\left(\hat{\theta}-\theta_0\right)\right)+O_p(1)，\quad\sqrt{n}\left(\hat{\theta}-\theta_0\right)=-\left(\frac{1}{n}\hat{A}(\theta_0)\right)^{-1}\frac{1}{\sqrt{n}}U(\theta_0)+O_p(1)，$$

因为

$$U(\theta_0)=\sum_{i=1}^{n}\sum_{j=1}^{n_i}\begin{pmatrix}\int_0^{\tau}\left(Z_{ij}-\bar{Z}\right)\mathrm{d}M_{ij}(t)\\\int_0^{\tau}\left(W_{ij}-\bar{W}\right)\mathrm{d}M_{ij}(t)\end{pmatrix}\overset{记作}{=}\sum_{i=1}^{n}\xi_i+O_p\left(\sqrt{n}\right)，$$

其中，$\xi_i=\sum_{j=1}^{n_i}\begin{pmatrix}\int_0^{\tau}\left(Z_{ij}-\bar{z}\right)\mathrm{d}M_{ij}(t)\\\int_0^{\tau}\left(W_{ij}-\bar{w}\right)\mathrm{d}M_{ij}(t)\end{pmatrix}$是均值为0的随机变量，而且$\xi_1,\xi_2,\cdots,\xi_n$相互独立。所以由中心极限定理$U(\theta_0)$渐近服从正态分布$N_{p+q}\left(0,nE\left(\xi_i^{\otimes 2}\right)\right)$。前面已有$\sqrt{n}\left(\hat{\theta}-\theta_0\right)=-\sqrt{n}\hat{A}^{-1}(\theta_0)U(\theta_0)+O_p(1)$再结合$\frac{1}{n}\hat{A}(\theta_0)\xrightarrow{p}A(\theta_0)$，可得$\sqrt{n}\left(\hat{\theta}-\theta_0\right)$渐近服从多元正态分布$N_{p+q}\left(0,A^{-1}(\theta_0)E\left(\xi_i^{\otimes 2}\right)A^{-1}(\theta_0)^{\mathrm{T}}\right)$。

3. 数值模拟

首先，生成$n(n=100,200,400)$个独立的类，每一个类的大小为2。协变量为Z_{ij}和W_{ij}分别服从参数为0.5的二项分布和区间(0, 1)的均匀分布，设失效时间的风险率函数模型为$\lambda_{ij}\left(t\middle|Z_{ij},W_{ij}\right)=\lambda_0(t)\mathrm{e}^{\left(Z_{ij}^{\mathrm{T}}\beta_0\right)}+W_{ij}^{\mathrm{T}}\gamma_0$，设$\lambda_0(t)=0.5$，$(\beta_0,\gamma_0)$分别取(0.5, 0.5)，(1, 0.5)，(0.5, 1)。取试验截止时间为$\tau=2$，删失时间C取区间$(0,v)$上的均匀分布，v可以变化使得删失控制在30%~70%之间。首先生成基准失效时间$T_{ij0}=-\log\left(1-\Phi\left(A_i+B_{ij}\right)\right)$，其中$A_i$服从均值为0，方差为$\rho$的正态分布，$B_{ij}$服从均值为0，方差为$1-\rho$的正态分布。参数$\rho$表示不同类间的异质性，$1-\rho$则表示同一类内不同对象的异质性，其中$\rho$取三种情况分别为0.25，0.5，0.75。该模拟重复进行500次。上述条件下的模拟结果如表1，表2，表3所示，其中EST，SE，ESE和CP分别代表参数估计的均值，样本标准差，标准差的渐近估计，以及正态近似条件下，β_0的95%置信区间下的覆盖率。从表1~3中的EST，SE，ESE和CP值都可以看出，当样本量变大时参数估计的均值越接近真值，随着样本

量增加样本标准差和标准差的渐近估计越来越接近，CP 也越来来越近似于 95%，可以得到本文提出的估计方法模拟效果良好，在其他参数设置下，得到结果是差不多的。

Table 1. Estimation for β_0 = (0.5, 0.5)
表 1. β_0 为(0.5, 0.5)的估计

n	ρ	EST	SE	ESE	CP	EST	SE	ESE	CP
100	0.25	0.535	0.286	0.280	0.962	0.528	0.277	0.262	0.944
	0.5	0.534	0.287	0.276	0.956	0.503	0.276	0.263	0.942
	0.75	0.542	0.272	0.277	0.968	0.515	0.274	0.263	0.942
200	0.25	0.514	0.198	0.185	0.942	0.509	0.189	0.185	0.938
	0.5	0.519	0.189	0.187	0.950	0.518	0.182	0.186	0.958
	0.75	0.515	0.193	0.186	0.950	0.506	0.195	0.188	0.958
400	0.25	0.505	0.151	0.148	0.960	0.487	0.154	0.151	0.944
	0.5	0.509	0.160	0.151	0.950	0.506	0.153	0.151	0.956
	0.75	0.505	0.154	0.152	0.954	0.511	0.150	0.152	0.954

Table 2. Estimation for β_0 = (1, 0.5)
表 2. β_0 为(1, 0.5)的估计

n	ρ	EST	SE	ESE	CP	EST	SE	ESE	CP
100	0.25	1.044	0.372	0.380	0.939	0.500	0.354	0.354	0.955
	0.5	1.034	0.359	0.373	0.959	0.502	0.350	0.355	0.953
	0.75	1.035	0.370	0.377	0.947	0.488	0.347	0.351	0.951
200	0.25	1.033	0.257	0.250	0.948	0.497	0.263	0.250	0.920
	0.5	1.050	0.284	0.256	0.952	0.506	0.252	0.248	0.940
	0.75	1.035	0.260	0.252	0.940	0.498	0.251	0.249	0.946
400	0.25	1.004	0.175	0.168	0.954	0.494	0.176	0.175	0.952
	0.5	1.014	0.169	0.168	0.960	0.496	0.181	0.176	0.940
	0.75	1.041	0.191	0.176	0.950	0.519	0.173	0.175	0.948

Table 3. Estimation for β_0 = (0.5, 1)
表 3. β_0 为(0.5, 1)的估计

n	ρ	EST	SE	ESE	CP	EST	SE	ESE	CP
100	0.25	0.533	0.375	0.395	0.970	1.013	0.349	0.364	0.955
	0.5	0.529	0.379	0.403	0.974	1.007	0.387	0.371	0.945
	0.75	0.535	0.371	0.405	0.965	1.007	0.390	0.368	0.931
200	0.25	0.531	0.257	0.263	0.978	1.005	0.258	0.259	0.960
	0.5	0.523	0.249	0.261	0.966	1.013	0.247	0.257	0.962
	0.75	0.507	0.252	0.258	0.960	0.988	0.247	0.259	0.968
400	0.25	0.499	0.173	0.177	0.958	0.992	0.180	0.182	0.958
	0.5	0.525	0.177	0.178	0.966	0.997	0.183	0.184	0.954
	0.75	0.514	0.183	0.180	0.966	1.004	0.184	0.185	0.952

4. 应用

在这部分我们采用提出的方法分析一个临床医学数据[12] gamma 干扰素治疗慢性肉芽肿性疾病的研究。数据共有 128 名慢性肉芽肿性疾病患者，被随机分为两个组，gamma 干扰素组和安慰剂组，每个病例的数据给出了患者发生重复严重感染的时间，在进行中期分析时，65 名安慰剂患者中的 20 名和 63 名 gamma 干扰素患者中的 7 名都至少经历过一次严重感染。在这里，我们把每一个个体视作一个簇，每个个体的各次复发的间隔时间视为簇中的成员。

假设数据可以由加性－乘积模型拟合，即 $\lambda_{ij}\left(t\middle|Z_{ij},W_{ij}\right)=\lambda_0(t)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\beta}+W_{ij}^{\mathrm{T}}\gamma$，定义 Z_{ij} 为治疗代码，如果病人接受 gamma 干扰素治疗则取 1，如果病人在安慰剂组则取 2，W_{ij} 为规范化后的个体年龄。

通过本文的方法可以得到的 β,γ 的估计分别为：$\hat{\beta}=0.7827$，$\hat{\gamma}=-0.0023$，标准差分别是 0.2242 和 0.0011，检验 P 值分别为 2.3996×10^{-4} 和 0.01689，根据 $\hat{\beta}$ 的值的大小，能够得出当 $Z_{ij}=1$ 时比 $Z_{ij}=2$ 基准风险率函数低，由此可见用 gamma 干扰素治疗的风险比安慰剂组的风险更低，而且 P 值小于 0.05，说明 gamma 干扰素治疗能显著降低感染慢性肉芽肿病的风险率函数，根据 $\hat{\gamma}$ 的值的大小以及相应的 P 值，可以得出年龄越大感染慢性肉芽肿病的风险显著地越低。具体地，固定年龄时，干扰素治疗组的风险率在时刻 t 时比安慰剂组的风险率绝对数值降低 $\hat{\lambda}_0(t)\times2.5972$，固定治疗组别时，年龄每增加 10 岁，风险率的绝对数值降低 4.8×10^{-4}。另外我们给出累积的基准风险率函数的估计 $\int_0^t\hat{\lambda}_0(t)\mathrm{d}t$，由它可以得到 $\hat{\lambda}_0(t)$ 在每个时刻的近似估计，见图 1 (横坐标为所有个体的复发事件的间隔时间 t，单位为天)。

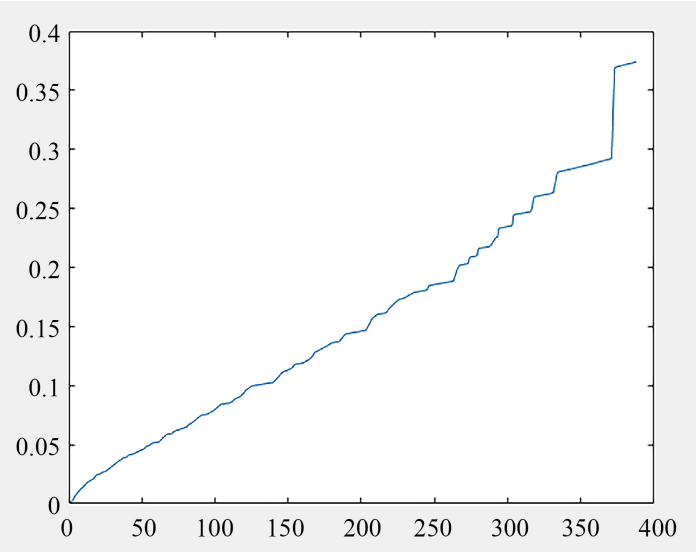

Figure 1. Estimation for $\int_0^t\hat{\lambda}_0(t)\mathrm{d}t$

图 1. $\int_0^t\hat{\lambda}_0(t)\mathrm{d}t$ 对时间的函数图像

为了对模型的拟合优度进行检验，我们对残差进行分析：我们来绘制残差 $\hat{r}_{ij}$ 关于个体复发 $i=1,2,...,128,\ j=1,2,...,n_i$ 的散点图，见图 2 (横坐标为所有个体的复发的标识，128 个个体共有 203 个复发)，其中

$$\hat{r}_{ij}=\int_0^{\tau}\mathrm{d}N_{ij}(t)-Y_{ij}(t)\left[\hat{\lambda}_0(t)\mathrm{e}^{Z_{ij}^{\mathrm{T}}\hat{\beta}}+W_{ij}^{\mathrm{T}}\hat{\gamma}\right]\mathrm{d}t$$

残差图可看出，残差随机分布在[−2, 2]中，因此不拒绝给出的模型。

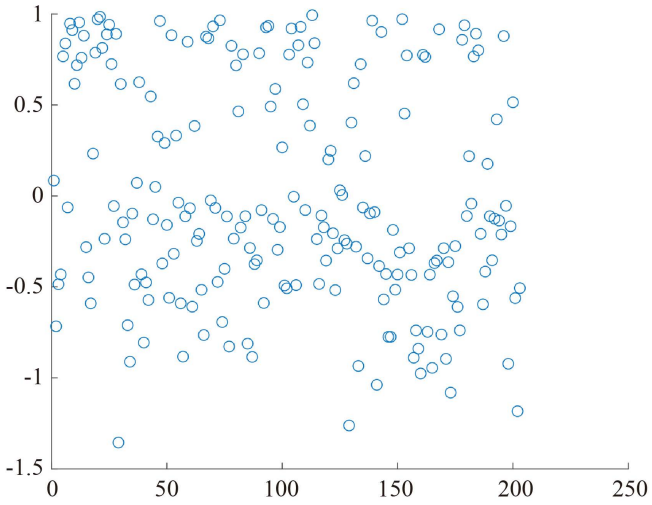

Figure 2. Scatter of residual vs. individual
图 2. 残差对个体的散点图

5. 总结

在这篇论文中，我们对聚类失效时间建立了边际加性乘积风险率函数模型，应用估计方程的思想进行统计推断，在计算中，我们使用了牛顿迭代法，得到参数的估计，应用大数定律证明估计量的相合性，应用中心极限定理证明估计量的渐近正态分布，并给出估计量的渐近方差，然后进行数值模拟验证提出的方法，模拟结果表现较好，最后将提出的模型和方法应用于实际数据，找到 gamma 干扰素治疗和年龄对慢性肉芽肿病的影响方式。未来我们将考虑在模型中加入随机效应项表示同一类之间的相关性，利用的数据信息更充分，相应的估计方法也更具有挑战性。

参考文献

[1] Cai, T., Wei, L. and Wilcox, M. (2000) Semiparametric Regression Analysis for Clustered Failure Time Data. *Biometrika*, **87**, 867-878. https://doi.org/10.1093/biomet/87.4.867

[2] Yin, G. and Cai, J. (2004) Additive Hazards Model with Multivariate Failure Time Data. *Biometrika*, **91**, 801-818. https://doi.org/10.1093/biomet/91.4.801

[3] Lu, S. and Wang, M. (2005) Marginal Analysis for Clustered Failure Time Data. *Lifetime Data Analysis*, **11**, 61-79. https://doi.org/10.1007/s10985-004-5640-6

[4] Cai, T., Cheng, S.C. and Wei, L.J. (2002) Semiparametric Mixed-Effects Models for Clustered Failure Time Data. *Journal of the American Statistical Association*, **97**, 514-522. https://doi.org/10.1198/016214502760047041

[5] Cai, J.W. and Zeng, D.L. (2011) Additive Mixed Effect Model for Clustered Failure Time Data. *Biometrics*, **67**, 1340-1351. https://doi.org/10.1111/j.1541-0420.2011.01590.x

[6] Cox, D.R. (1972) Regression Model Sandlife Tables (with Discussion). *Journal of the Royal Statistical Society*, *Series B*, **34**, 187-220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

[7] Martinussen, T. and Scheike, T.H. (2002) A Flexible Additive Multiplicative Hazard Model. *Biometrika*, **89**, 283-298. https://doi.org/10.1093/biomet/89.2.283

[8] Shang, W.P. and Wang, X. (2017) The Generalized Moment Estimation of the Additive-Multiplicative Hazard Model with Auxiliary Survival Information. *Computational Statistics & Data Analysis*, **112**, 154-169. https://doi.org/10.1016/j.csda.2017.03.013

[9] Han, M., Sun, L.Q., Liu, Y.T. and Zhu, J. (2018) Joint Analysis of Recurrent Event Data with Additive-Multiplicative

Hazards Model for the Terminal Event Time. *Metrika*, **81**, 523-547. https://doi.org/10.1007/s00184-018-0654-3

[10] Liang, K.Y. and Zeger, S.L. (1986) Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, **73**, 13-22. https://doi.org/10.1093/biomet/73.1.13

[11] Fleming, T.R. and Harrington, D.P. (1991) Counting Processes and Survival Analysis. Wiley, New York.

[12] Rudin, W.B. (1976) Principles of Mathematical Analysis. 3rd Edition, McGfraw-Hill, New York.