威布尔脆弱性比例危险模型的等级似然估计和应用*

王宁宁[1]，徐淑一[2]
（1. 广州医科大学公共卫生学院，广东 广州 510182；
2. 中山大学岭南学院，广东 广州 510275）

摘 要：针对包含右删失的纵列生存数据，建立了威布尔分布的脆弱性半参数比例危险模型，为避免边际似然的积分运算，采用等级似然方法估计协变量系数和随机效应的实现值，采用调整的轮廓等级似然方法估计随机效应分布参数。并且通过模拟研究将威布尔脆弱性模型和对数正态脆弱性模型以及伽马脆弱性模型作了比较，发现威布尔分布最适合半参数比例危险模型的脆弱性因子的分布设定。最后，将建立的模型用于分析肾病感染数据和烧伤病人皮肤移植数据，结果表明威布尔脆弱性模型都给出了不弱于伽马脆弱性模型和对数正态脆弱性模型的估计结果，而且威布尔脆弱性模型有助于诊断异质性问题。

关键词：脆弱性模型；威布尔分布；等级似然；比例危险；右删失
中图分类号：O212 **文献标志码**：A **文章编号**：0529－6579（2014）05－0039－08

The Estimation and Application for Cox Model with Weibull Frailty Via Hierarchical-likelihood

WANG Ningning[1], *XU Shuyi*[2]
（1. Public Health School, Guangzhou Medical University, Guangzhou 510182, China;
2. Lingnan (University) College, Sun Yat-sen University, Guangzhou 510275, China）

Abstract: The proportional hazard model with Weibull frailty is used for analyzing right-censoring longitudinal survival data. Complex marginal integrals are avoided by using Hierarchal-likelihood to estimate covariate coefficients and prediction of random effects, the adjusted profile hierarchical-likelihood is taken to estimate the parameters of random factors'distribution. Simulation studies of comparing with other frailties including lognormal frailty and Gamma frailty are provided and the result indicates that the Weibull frailty performs very well. The examples for analyzing two real data sets are also presented, and the results indicate that the Weibull frailty model is at least as good as lognormal and Gamma frailty model, and the Weibull frailty model is more helpful to diagnose the heterogeneity.

Key words: frailty model; Weibull distribution; hierarchal-likelihood; proportional hazard; right censor

Cox[1]比例危险模型自从 1972 年提出以来一直是占据生存数据分析的主要模型，在面对同一个体的重复观测时（或者数据按照不同来源分组时），对生存数据的建模需要考虑个体的异质性影响。Keyfitz 和 Littman 发现忽略个体的异质性会导致对协变量系数的估计有偏。Hougarrd 与 Aalen 也发现，忽略个体的异质性时，估计出的相对危险率都有偏高的倾向，Lancaster 对失业持续时间建模时，发现模型中忽略脆弱性因子会导致低估协变量效应。Pickles 等[2]对脆弱性模型的回顾性总结中表

* **收稿日期**：2013－11－08
基金项目：广州医科大学科研基金资助项目（L135021）；中山大学青年教师培育资助项目（10000－3161146）
作者简介：王宁宁（1977 年生），男；**研究方向**：生存分析及其应用；**通讯作者**：徐淑一；E-mail：lnsxsy@mail.sysu.edu.cn

明，一般来说，模型中忽略异质性的影响会使得协变量系数的估计向 0 这个方向接近。将脆弱性（Frailty）因子引入模型就是考虑到这种不可忽略的异质性。脆弱性模型是 Cox 提出的比例危险模型的一个推广，脆弱性模型允许同一个体的重复观测数据之间或者同一个组内的个体之间具有相关性，广泛应用于多元生存数据分析中。

经典的脆弱性比例危险模型 $h(t_{ij}|v_i) = h_0(t_{ij})\exp\{x_{ij}\beta + v_i\} = h_0(t_{ij})\exp\{x_{ij}\beta\}u_i$，下标 ij 表示第 i 个体的第 j 次重复观测。除了参数向量 β 之外，v 的分布函数 $F(v)$ 以及基本危险率 h_0 都是未知的。在很多脆弱性比例危险模型的应用中，基本危险率通常认为是不规则的形式，通过非参数方法估计。有很多学者对基本危险率是非参数设定的脆弱性模型的估计进行了研究，McGilchrist 和 Aisbett 以及 McGilchrist 研究了对数正态脆弱性模型，利用 Cox 的部分似然（Partial Likelihood）方法估计参数，并且使用这种模型来分析肾感染的数据，但是没有处理数据打结的情况。近年来的研究主要集中于非正态脆弱性模型的估计，比如，出于数学上处理的方便，Klein 及 Nielsen 等建立了伽马（Gamma）脆弱性模型的 EM 算法。Hougaard 建议了三种脆弱性因子的分布：伽马分布、逆高斯（Inverse Gaussian）分布和正稳定（Positive stable）分布，Hougaard[3] 详细讨论了脆弱性因子的分布的选择。EM 算法在估计带有不可观测的协变量模型是很有用的技术，见 Mclachlan 等[4]。EM 算法的收敛对初始值的选择和停止计算的规则非常敏感，而且 EM 算法需要脆弱性因子在给定观测数据下的条件期望，除了伽马分布，逆高斯分布等几种特定的分布，一般情况下，无法写出解析表达式，而是需要数值算法，而且参数估计的方差也不能直接获得（Louis，Jamshidian 等[5]）。

由 Lee 等[6] 提出的带随机效应的广义线性模型的等级似然（Hierarchicacl Likelihood）估计方法，允许随机效应因子是任何分布，等级似然方法近年来开始应用于生存数据的混合模型估计，展现了良好的应用前景。Ha 等[7] 将等级似然估计方法用于脆弱模型的估计，估计了随机效应（脆弱性因子）呈正态分布和 Gamma 分布的情形，并证明了在给定随机效应分布参数的情况下，最大化边际似然得到的协变量系数估计和最大化等级似然得到的协变量系数估计是一致的。自此，等级似然估计方法在纵列数据的生存分析与混合模型方面的应用研究进一步展开，如 Ha 等[8-9]、徐淑一等[10] 的研究。近年来学术界关于等级似然方法的应用研究仍在继续，如 Lee 等[11] 使用等级似然方法估计预测疾病测绘时相对风险的的区间，模拟显示其不弱于贝叶斯方法。Noh 等[12] 使用等级似然建立非线性混合效应模型处理纵向数据。Noh 等[13] 使用等级似然方法提供了一种减弱对数据缺失机制不正确假定的影响并提供了实例和模拟研究。Lee 等[14] 把等级似然函数方法用于缺失数据的建模并应用于厚尾分布的纵向数据分析。Wu 等[15] 使用等级似然解决因子分析模型中二元相应变量的情形，使用简单而且高效。王宁宁等将该方法推广到 AR（1）相依结构的脆弱性模型研究。

本文考虑 Cox 比例危险的脆弱性模型，允许数据打结以及删失的情况下，假设脆弱性因子服从威布尔分布。对固定参数的估计和随机效应实现的预测采用联合最大化扩展似然方法，实际上扩展似然即为等级似然，但是等级似然要求特定的随机效应尺度，见 Lee 和 Nelder 等；而随机效应分布参数则采用调整的轮廓等级似然进行估计。本文推导了模型的等级似然函数和估计过程，并将不同分布的脆弱性模型做对比研究。

1 随机效应是威布尔分布的半参数比例危险模型的估计

威布尔分布广泛应用于生存时间以及可靠性试验的统计分析中，Peto 和 Lee 认为在一定条件下失效时间应该是威布尔分布。威布尔分布的危险率函数是 $h(t) = \gamma t^{\gamma-1}$（尺度参数设为 1），当形状参数 γ 大于 1 时，危险率随时间而增加；当 γ 等于 1 时，危险率函数是常数；当 γ 小于 1 时，危险率是减函数。威布尔分布灵活的危险率形式使得它非常适合作为生存时间的模型，在生命科学领域，已经有许多的统计学家采用威布尔分布来分析数据。由于威布尔分布的危险率函数的简单性和灵活性，威布尔分布在多元生存分析中也非常有用。Kimber 和 Crowder 在对异质性建模时使用威布尔分布作为生存分布，Lancaster 在考虑异质性的持续数据的比例危险模型中采用威布尔分布。

令 T_{ij}（$i = 1,2,\cdots,q$，$j = 1,2,\cdots,n_i$）是第 i 个个体（组）的第 j 次重复观测，假定给定 $U_i = u_i$ 的的条件下，数据对（T_{ij}, C_{ij}）是条件独立的，而且 T_{ij} 与 C_{ij} 条件独立；并且给定 $U_i = u_i$，$\{C_{ij}, j = 1,2,\cdots,n_i\}$ 不含有 u_i 的信息。C_{ij} 为删失时间，设 T_{ij} 的条件危险率为比例危险形式，U_i 之间独立同分布，设其服从威布尔分布，考虑危险率函数

$$h_i(t|u_i) = h_0(t)\exp\{x_i\beta\}u_i\exp(-E(\log u_i)) \tag{1}$$

u_i 服从威布尔分布，其密度函数为

$$f(u_i) = \lambda\rho(\lambda u_i)^{\rho-1}\exp\{-(\lambda u_i)^{\rho}\} \tag{2}$$

对脆弱性因子 u_i 进行变换，使 u_i 的对数为零均值，令 $u_i = e^{w_i}$，那么 $w_i = \log(u_i)$ 具有极值分布

$$f(w_i) = \frac{1}{b}\exp\left\{\frac{w_i-\mu}{b} - \exp\left(\frac{w_i-\mu}{b}\right)\right\} \tag{3}$$

此时，

$$E(w_i) = \mu - \gamma b, \operatorname{var}(w_i) = \frac{\pi^2}{6}b^2$$

上式中 $b = \frac{1}{\rho} > 0, \mu = -\log(\lambda), -\infty < \mu < \infty$，$\gamma = 0.572\,2\cdots$ 是 Euler 常数（Kalbfleisch 和 Prentice），再考虑变换 $v_i = w_i - E(w_i)$，则

$$f(v_i, b) = \frac{1}{b}\exp\left\{-\gamma + \frac{v_i}{b} - \exp\left(-\gamma + \frac{v_i}{b}\right)\right\} \tag{4}$$

其中，$E(v_i) = 0, \operatorname{var}(v_i) = \frac{\pi^2}{6}b^2$，由于常数部分可以被基本危险率吸收，则 Cox 比例危险模型实际为

$$h_i(t|u_i) = h_0(t)\exp\{x_i\beta + v_i\} \tag{5}$$

于是，等级似然函数可以写成

$$h = \sum_{i,j} l_{1ij} + \sum_i l_{2i} = l_1 + l_2 \tag{6}$$

其中

$$l_{1ij} = \delta_{ij}\{\log\lambda_0(y_{ij}) + \eta_{ij}\} - \{\Lambda_0(y_{ij})\exp(\eta_{ij})\} \tag{7}$$

$$l_{2i} = -q\ln b - q\gamma + \frac{1}{b}\sum_{i=1}^{q} v_i - e^{-\gamma}\sum_{i=1}^{q}\exp\left(\frac{v_i}{b}\right) \tag{8}$$

（7）式中 $\eta_{ij} = \exp(x_i\beta + v_i)$，等级似然函数 $h = \log\{\prod_{ij} f(T_{ij}, \beta; x_{ij}|v_i) f(v_i)\}$，其中 $f(T_{ij}, \beta; x_{ij}|v_i) f(v_i)$ 为 T_{ij} 和随机效应（v_i）的联合密度函数。可以将等级似然看作是关于 β 和 v 的联合似然函数，v 是不可观测的随机效应的实现，也就是说可以将 v 当成一个固定参数来处理，因为，给定 y 的观测值，随机效应 v 的实现值就是一个固定的常数，关于 β 和 v 最大化等级似然可以得到 β 和 v 的估计，这样就避免了高维积分的计算，而且得到随机效应的估计为最优无偏预测（Lee and Nelder）。对基本危险率的估计，采用 Breslow 的最大似然估计结果，将 T_{ij} 从小到大排序得到：$T_1, T_2, \cdots, T_K$，其中 $K = \sum_{i=1}^{q} n_i$，$H_0(t)$ 与 $h_0(t)$ 的非参数极大似然估计为

$$\hat{H}_0(t) = \sum_{k; y_{(k)} \leqslant t} \frac{d_{(k)}}{\sum_{ij \in R(y_{(k)})}\exp(\eta_{ij})},$$

$$\hat{h}_0(t) = \sum_{k; y_{(k)} \leqslant t}\left\{\frac{d_{(k)}}{\sum_{ij \in R(y_{(k)})}\exp(\eta_{ij})}\right\}$$

上式中 $d_{(k)}$ 为个体在时间 T_k 发生感兴趣事件的数目，$R(y_{(k)})$ 表示在时间 T_k 的危险集。将 $\hat{H}_0(t)$ 和 $\hat{h}_0(t)$ 代入到等级似然函数 h 中，得 $h^* = h|_{\hat{\Lambda}_0(t)}$，去掉与参数无关的项，可以得到

$$h^* \propto \sum_{ij}\{\delta_{ij}(x'_{ij}\beta + v_i)\} - \sum_k d(k)\ln\{\sum_{ij \in y_{(k)}}\exp(\eta_{ij})\} + \sum_i l_{2i} \tag{9}$$

对 h^* 关于 β, v 一起最大化，可以得到 β 和 v 的估计，估计过程采用 Newton-Raphson 迭代。

对随机效应分布参数 b 的估计，Lee 和 Nelder 建议使用最大化调整的轮廓似然估计，调整的轮廓似然为

$$h_A^* = h^* + \frac{1}{2}\log\{\det(2\pi J^{-1})\},$$

$$J = \frac{-\partial^2 h^*}{\partial(\beta, v) \otimes \partial(\beta, v)'} \tag{10}$$

h_A^* 是非线性函数，要估计 b，仍需要采用迭代算法，在精确计算出 h_A^* 对 b 的一、二阶导数之后，可以调用 SAS 的最优化模块完成。其中 J^{-1} 为估计量的方差协方差阵，我们用 h_A^* 关于 b 的二阶导数的相反数的逆作为 b 的估计的方差的近似。

2　各种随机效应模型的比较和模拟研究

为了进一步研究不同的脆弱性模型，下面对随机效应的不同分布生成模拟数据，采用不同的脆弱性模型进行估计。为了进行比较，也把带有随机效应生成的生存数据用一般的 Cox 比例危险模型进行估计并和随机效应模型估计结果进行比较，用以检查忽略随机效应对模型估计的影响，模型设为

$$h(T_{ij}|u_i) = h_0(T_{ij})\exp(x'_{ij}\beta)u_i = h_0(T_{ij})\exp(x'_{ij}\beta + v_i) \tag{11}$$

其中，$i = 1, 2, \cdots, q, j = 1, 2, \cdots, n_i$，也就是说假定 q 个病人，每个病人有 n_i 个重复观察；或者观察数据按照其自然状态有 q 个组，每个组内有 n_i 个个体。

我们先生成正态分布随机效应的生存数据，数据生成过程允许右删失。设 u_i 服从对数正态分布，那么 v_i 就服从正态分布，分布密度函数为 $f(v_i, a) = \frac{1}{\sqrt{2\pi a}}\exp\left(-\frac{v_i^2}{2a}\right)$。为了模拟方便，假定基本危

险率 $h_0(T_{ij}) = 1$，首先生成正态分布的随机效应数据 v_i，设方差 $a = 1$，只有一个协变量 x，协变量 x 一半取值为 1，一半取值为 0，协变量通过随机效应模型（11）影响危险函数，给定随机效应的情况下，则可以由参数是 $\exp(X_{ij}\beta + v_i)$ 的指数分布生成事件发生时间，假设 β 的真值是 -2，之所以将协变量系数设置为负数，是因为若设其为正数，生成的生存时间太小，很不方便比较；设删失时间来自参数为 $\theta_{ij} = (\frac{\rho}{1-\rho})\lambda_{ij}$ 的指数分布，其中 ρ 为删失率。设删失率分别为 0.1、0.3、0.5 三种情形生成数据，分别进行一般 Cox 模型、正态脆弱性模型、伽马脆弱性模型、威布尔脆弱性模型进行估计，模拟进行 200 次。结果列在表 1 中。

表 1 正态随机效应生成的生存数据的估计结果[1)]

Table 1 Simulation results under lognormal data generation process

删失率	模型	回归系数 β				方差参数 a			
		MEAN(绝对误差)	SD	SE	95% CP	MEAN	SD	SE	95% CP
0.1	Cox	−1.346 9(0.653 1)	0.200 6	0.173 2	0.081	–	–	–	–
	LNM	−2.010 7(0.010 7)	0.201 6	0.203 5	0.947	1.065 1	0.441 8	0.347 6	0.862
	GAM	−1.998 6(0.001 4)	0.199 7	0.204 4	0.943	1.442 9	0.585 8	0.431 7	–
	WEL	−2.033 6(0.033 6)	0.201 4	0.205 9	0.952	1.049 4	0.205 6	0.144 5	–
0.3	Cox	−1.333 9(0.666 1)	0.215 5	0.195 3	0.105	–	–	–	–
	LNM	−1.951 3(0.049 0)	0.216 8	0.227 1	0.964	1.025 5	0.423 2	0.336 2	0.868
	GAM	−1.937 0(0.063 0)	0.215 9	0.228 2	0.959	1.529 4	0.636 2	0.464 5	–
	WEL	−1.987 6(0.012 4)	0.216 3	0.230 5	0.959	1.019 8	0.180 5	0.141 5	–
0.5	Cox	−1.374 5(0.625 5)	0.261 4	0.234 2	0.264	–	–	–	–
	LNM	−1.919 6(0.080 4)	0.280 7	0.266 6	0.917	0.946 6	0.415 1	0.314 2	0.805
	GAM	−1.896 6(0.103 4)	0.275 7	0.267 5	0.902	1.718 7	0.689 4	0.538 8	–
	WEL	−1.988 4(0.011 6)	0.295 5	0.271 2	0.929	0.971 3	0.158 7	0.135 6	–

1）Cox 是指 Cox 比例危险模型，LNM 指对数正态脆弱性模型，GAM 指伽马脆弱性模型，WEL 指威布尔脆弱性模型；MEAN 指 200 次模拟估计的平均值，SD 是 200 次模拟估计的标准离差，SE 是指平均的标准差，95% CP 为 95% 的置信区间包含参数真值的比例。下面表 2、表 3 相同

由表 1 可以发现，删失率为 0.1 时，三种随机效应模型的估计结果差别不大，尤其是伽马脆弱性模型的估计最为精确，绝对误差为 0.001 4，似乎最为合适的脆弱性模型应该是伽马脆弱性模型。删失率的增加到 0.3 时，不再是伽马脆弱性模型估计最精确，而是威布尔脆弱性模型估计最精确，绝对误差为 0.012 4，而此时伽马脆弱性模型最差。删失率增加到 0.5 时，这种变化趋势已经很明显了，此时威布尔脆弱性估计最精确，绝对误差仅为 0.011 6，而且和相对低的删失率情况下估计结果差别很小，而其它两种脆弱性模型误差都比较大，尤其伽马脆弱性精确度很差。从这三种脆弱性模型的估计结果来看，假如真实的随机效应是对数正态的，估计时分布的误设带来的误差不是很明显，然而，当删失率提高时，威布尔脆弱性模型却能大大改善回归系数的估计。进一步地，我们发现，模拟估计的标准差和平均的 SE 基本一致，说明用等级似然函数关于参数的二阶导数矩阵地相反数地逆作为估计的方差也很精确。

接下来我们生成伽马随机效应的生存数据，设 $f(u_i,\alpha) = \frac{\alpha^\alpha}{\Gamma(\alpha)}u_i^{\alpha-1}e^{-\alpha u_i}$ 为随机效应 u_i 的分布，$E(u_i) = 1, \mathrm{var}(u_i) = \frac{1}{\alpha}$，我们仍设基本危险率 $h_0(T_{ij}) = 1$，设 $a = 1$。估计结果列在表 2 中。

表 2　伽马随机效应生成的生存数据的估计结果

Table 2　Simulation results under gamma data generation process

删失率	模型	回归系数 β				方差参数 a			
		MEAN(绝对误差)	SD	SE	95% CP	MEAN	SD	SE	95% CP
0.1	Cox	−1.588 4(0.411 6)	0.208 1	0.188 0	0.402	–	–	–	–
	LNM	−2.039 2(0.039 2)	0.206 9	0.207 2	0.945	1.016 9	0.567 1	0.332 3	0.708
	GAM	−2.001 7(0.001 7)	0.198 5	0.206 1	0.964	1.709 4	1.159 2	0.528 2	0.735
	WEL	−2.066 6(0.066 6)	0.208 6	0.208 5	0.927	0.876 2	0.243 8	0.114 1	0.564
0.3	Cox	−1.555 7(0.443 0)	0.202 4	0.212 0	0.418	–	–	–	–
	LNM	−1.963 8(0.036 2)	0.217 9	0.231 5	0.959	0.962 5	0.536 5	0.316 2	0.731
	GAM	−1.928 5(0.071 5)	0.215 0	0.230 5	0.942	1.717 3	0.828 9	0.531 4	0.827
	WEL	−1.995 8(0.004 2)	0.228 0	0.233 2	0.929	0.882 7	0.218 8	0.116 5	0.634
0.5	Cox	−1.566 9(0.431 1)	0.260 5	0.251 0	0.555	–	–	–	–
	LNM	−1.943 9(0.056 1)	0.283 3	0.270 7	0.926	1.055 5	0.619 7	0.349 5	0.730
	GAM	−1.911 0(0.089 0)	0.274 0	0.268 5	0.912	1.581 6	0.789 7	0.493 7	0.867
	WEL	−1.944 6(0.055 4)	0.274 2	0.268 7	0.932	0.790 8	0.170 9	0.104 7	0.475

由表 2 的结果可以看出，删失率为 0.1 时，三种随机效应模型估计结果差不多，绝对误差都比较小，伽马模型估计的最精确，但是删失率增加到 0.3，再增加到 0.5 的时候，伽马脆弱性模型估计结果不如另外两种模型精确，这种变化趋势很明显，尤其删失率比较高时，比如说 0.5 时，威布尔脆弱性模型估计最为精确，对数正态模型次之。

我们考虑威布尔随机效应分布的生存数据的模拟估计，设随机效应 v_i 分布为 $f(v_i,b) = \frac{1}{b}\exp\{-\gamma + \frac{v_i}{b} - \exp(-\gamma + \frac{v_i}{b})\}$，设基本危险率 $h_0(T_{ij}) = 1, b = 1$，估计结果列在表 3 中。为估计方便，估计过程中采用了再参数化，令 $\delta = \frac{1}{b}$，估计 δ 的值，因为 b 的真值为 1，因此，δ 的真值仍为 1。模拟结果表明，删失率比较低时，威布尔随机效应的生存数据估计结果都很精确，如表 3 所示，删失率为 0.1，威布尔脆弱性模型估计的参数精度达到绝对误差仅为 0.004 2，意味着相对误差仅为 0.21%，另外两种脆弱性模型估计结果也不错，但不如威布尔脆弱性模型。删失率提高到 0.3 再到 0.5 时，威布尔脆弱性模型估计协变量参数的相对误差都在 3.5% 以内，绝对误差在 0.07 之内，伽马模型和对数正态模型估计的精度都差不多，随删失率的变化也差不多，在删失率为 0.5 时，伽马模型和对数正态模型估计的精度比威布尔模型低一半左右。

表 3　威布尔随机效应生成的生存数据的估计结果

Table 3　Simulation results under Weibull data generation process

删失率	模型	回归系数 β				方差参数 δ			
		MEAN(绝对误差)	SD	SE	95% CP	MEAN	SD	SE	95% CP
0.1	Cox	−1.095 6(0.904 4)	0.214 1	0.162 9	0.009	–	–	–	–
	LNM	−1.976 0(0.024 0)	0.211 5	0.202 2	0.932	1.613 7	0.730 5	0.525 2	0.886
	GAM	−1.980 2(0.019 8)	0.212 0	0.204 1	0.936	1.128 5	0.421 5	0.328 7	0.936
	WEL	−1.995 8(0.004 2)	0.213 4	0.204 9	0.935	0.978 7	0.184 5	0.139 0	0.857
0.3	Cox	−1.088 7(0.911 3)	0.224 6	0.184 5	0.023	0	0	0	0
	LNM	−1.920 6(0.079 4)	0.240 8	0.225 3	0.905	1.534 3	0.701 2	0.500 6	0.909
	GAM	−1.925 4(0.074 6)	0.240 3	0.227 8	0.909	1.169 1	0.459 4	0.344 1	0.959
	WEL	−1.946 6(0.053 4)	0.241 8	0.228 5	0.922	0.962 5	0.170 6	0.136 8	0.844
0.5	Cox	−1.115 7(0.884 3)	0.279 1	0.221 2	0.082	–	–	–	–
	LNM	−1.860 5(0.139 5)	0.287 0	0.264 0	0.872	1.417 6	0.725 8	0.466 0	0.849
	GAM	−1.864 9(0.134 1)	0.286 1	0.267 5	0.899	1.324 5	0.637 4	0.403 1	0.940
	WEL	−1.930 5(0.069 5)	0.288 7	0.270 1	0.930	0.915 8	0.154 3	0.130 9	0.816

上述三种分布的随机效应生存数据的模拟估计还显示：如果不采用随机效应模型，协变量系数的估计将会严重偏离，整体上均呈现高估的趋势，绝对值上是低估，这说明，如果模型中忽略异质性（Frailty）的影响，会使得估计结果存在很大谬误。模拟还发现，即使采用了错误的脆弱性模型，也比不用脆弱性模型估计结果好得多，这说明，实际中，对纵列生存数据建模和估计，要非常小心数据是否存在个体或者组别的异质性，以及注意模型是否忽略了重要因素而导致估计不准确。在实际中，我们推荐使用威布尔脆弱性模型。

模拟研究还发现，用不正确的脆弱性模型以及随着删失率的增加，估计也呈现高估趋势（协变量参数在绝对之上是低估趋势）。究其原因，我们认为是由于随机效应的存在，一方面增加了参数估计的不确定性，另一方面，由于随机效应问题，使得生存数据的波动存在于两个方面，一是协变量的影响，二是随机效应的影响，如果随机效应高估，那么协变量参数会呈现低估现象。观察表 1 到表 3 的模拟结果，我们发现，当随机效应方差低估时，协变量系数会高估；当随机效应方差高估时，协变量系数则低估。

威布尔随机效应生成的生存数据，用威布尔脆弱性模型估计的随机效应分布参数也相当精确，δ 的真实值为 1，删失率为 0.1、0.3、0.5 时估计的参数分别为 0.978 69（0.184 45）、0.962 53（0.170 6）、0.915 83（0.154 32），括号内为模拟估计的标准差，可以看出，他们也是显著的，而且，对威布尔随机效应模拟的数据，用调整的轮廓等级似然函数的二阶导数估计方差，也是比较精确的，因为估计的标准差和用二阶导数计算的平均的 SE 差别不大，仅有一点高估。其它两种脆弱性模型的随机效应分布参数的估计也很精确，但是估计量的方差用调整的轮廓似然的二阶导数计算则偏误较大。由于不同随机效应分布的参数涵义不一致，不同脆弱性模型随机效应参数的直接比较没有意义，但是按照和数据生成过程一致的脆弱性模型估计的随机效应分布参数与真值非常接近。

模拟还显示，对带有随机效应的生存数据，不能忽略随机效应的存在，否则，Cox 模型会有很大偏误。对不同随机效应的生存数据，当然理论上是能够使用真实分布的脆弱性模型估计会很好，但是模拟研究发现，对正态分布、伽马分布、威布尔分布这三种随机效应的生存数据来说，威布尔脆弱性模型估计结果显示出很大的优势，尤其当删失率提高时，这种优势最为明显。那么，我们可以得出这样的结论，当数据存在较大比例的删失时，不妨采用威布尔脆弱性模型。至于为何威布尔脆弱性模型显示出这么明显的优势，还有待于进一步研究。本文本文模拟采用的样本容量为 200，而实际上样本容量为 50 的情况下，本文的模型也给出了非常类似的结果，这表明等级似然估计方法应用于脆弱性模型估计的时候对于小样本情形有很好的近似。

3 应用实例

下面对文献上常见的两个纵列生存数据的例子用前面讨论的各种脆弱性模型进行估计，一个是肾病感染数据，另一个是烧伤病人皮肤移植数据。

肾病感染数据来自于 McGilchrist 和 Aisbett，总共有 38 个肾病患者做便携式肾透析，记录下在导管插入处发生感染的时间，从导管插入开始时记录时间，每个患者均被记录了两次发生感染的时间，如果某次感染发生时，导管已经由于感染或其它原因被移除，则这次感染时间被记录为删失。研究中主要考虑五个协变量的影响：年龄，用 x_1 表示；性别，用 x_2 表示，$x_2 = 1$ 为女性，$x_2 = 0$ 为男性；是否为血球性肾炎用 x_3 表示，$x_3 = 1$ 表示是血球性肾炎，否则 $x_3 = 0$；是否为急性肾炎，用 x_4 表示，$x_4 = 1$ 表示是急性肾炎，否则 $x_4 = 0$；是否为多囊肾病，用 x_5 表示，$x_5 = 1$ 表示是多囊肾病，否则 $x_5 = 0$ 否则。考虑每个病人的异质性，建立脆弱性模型为

$$H(t_{ij} \mid x_1, x_2, x_3, x_4, x_5; u_i) = h_0(t_{ij}) \cdot \exp(\beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{5ij} + \beta_5 x_{5ij} + v_i),$$
$$i = 1, 2, \cdots, 38, \quad j = 1, 2$$

每个病人有一个异质性因子 v_i，以此来体现每个病人两次感染之间的联系。对肾感染数据，采用一般 Cox 比例危险模型、对数正态脆弱性模型、伽马脆弱性模型、以及威布尔脆弱性模型进行估计，结果列在表 4 中，可以看出，性别比较显著，是个危险因素，这说明男性比女性更容易发生感染，这个结果与 McGilchrist 和 Aisbett 使用的边际似然估计的结果一致。不同脆弱性模型估计结果有一定的差别，而且都显示存在显著的随机效应，也就是说不同病人除了性别以及疾病类型之外，还有明显的个体差异。

表 4　肾感染数据的各种模型估计结果

Table 4　Results of kidney data

	Cox model	Lgnormal frailty	Gamma frailty	Weibull frailty
β_1	0.001 151	0.002 378	0.001 406	0.000 248
(sd)	(0.010 946)	(0.015 997)	(0.013 9739)	(0.020 720)
(p_value)	(0.916 285)	(0.881 828)	(0.919 863)	(0.990 431)
β_2	−1.442 410	−1.727 616	−1.687 322	−2.068 904
(sd)	(0.354 167)	(0.489 967)	(0.453 123)	(0.622 434)
(p_value)	(4.65E−05)	(0.000 422)	(0.000 196)	(0.000 888)
β_3	0.136 183	0.262 867	0.254 705	0.469 168
(sd)	(0.404 791)	(0.582 816)	(0.517 804)	(0.744 803)
(p_value)	(0.736 548)	(0.651 969)	(0.622 794)	(0.528 746)
β_4	0.526 266	0.694 958	0.673 683	0.932 979
(sd)	(0.414 204)	(0.595 189)	(0.529 888)	(0.755 450)
(p_value)	(0.203 889)	(0.242 958)	(0.203 598)	(0.216 831)
β_5	−1.354 020	−1.008 496	−0.853 446	−0.335 677
(sd)	(0.626 798)	(0.860 729)	(0.842 229)	(1.079 952)
(p_value)	(0.030 756)	(0.241 327)	(0.310 908)	(0.755 933)
a		0.691 952	2.759 466	0.878 514
(sd)	–	(0.208 790)	(0.843 115)	(0.103 461)
(p_value)		(0.000 919)	(0.001 064)	(0)

但是 Klein 的伽马边际似然方法给出的结果和一般 Cox 模型结果类似，而且利用 Klein 的检验说明该例不存在个体异质性。通过上一章的模拟研究，我们发现用调整的等级轮廓似然的二阶导数估计随机效应分布参数估计值的方差会偏低，但偏低并不是很严重，本例的估计结果显示存在个体异质性。Klein 的异质性检验方法是个很一般的检验，功效往往很弱，经常检验不出异质性。Yus T. BOENG 在他的博士论文研究中对参数基本危险率的威布尔脆弱性模型估计结果也证实存在个体异质性，而且提出了针对威布尔分布的基本危险率的得分检验证实存在异质性。根据上一节的模拟研究结论，当数据存在适量删失时，建议采用威布尔 Frailty 模型。

皮肤移植数据来自 Batchelor 和 hackett，研究了 16 个严重烧伤的病人，由于每一个病人多块皮肤需要移植，研究每块皮肤的治疗效果，有的病人只有一块皮肤移植，而有的病人有多块，这是一个非平衡的观测数据，不同病人做皮肤移植的块数不同，最多的有四块，因此同一个体可以有多个生存时间，适合建立脆弱性模型研究。这个数据只考虑了一个协变量因素，就是人白细胞抗原（human leukocyte antigen，HLA）的匹配好坏，模型设为：$H(t_{ij}\mid x;u_i) = h_0(t_{ij})\exp(\beta \times \mathrm{HLA}_{ij} + v_i)$，HLA 为 1 表示匹配的好，0 表示匹配的不好，$i = 1,2,\cdots,16$，$j = 1,\cdots,n_i$。采用一般 Cox 比例危险模型和三种脆弱性模型进行估计，结果列在表 5 中。

表 5　皮肤移植数据的各种模型估计结果

Table 5　Results of skin data

	Cox model	Lgnormal frailty	Gamma frailty	Weibull frailty
β	−1.028 587	−1.353 218	−1.167 288	−1.386 303
(sd)	(0.426 426)	(0.492 116)	(0.464 853)	(0.507 327)
(p_ value)	(0.015 860)	(0.005 963)	(0.012 036)	(0.006 284)
a		1.444 386 4	1.785 55	0.759 679
(sd)	–	(0.566 530)	(0.702 561)	(0.124 693)
(p_ value)		(0.010 787)	(0.011 038)	(1.112 2E−9)

我们发现，HLA 的影响是显著的，是影响皮肤移植生存的重要因素，不同脆弱性模型估计结果差别很小，对数正态脆弱性模型和威布尔脆弱性模型估计结果比较接近，伽马脆弱性模型和 Cox 模型估计结果比较接近，而且都显示了存在显著的随机效应，尤其是威布尔模型，显示个体存在明显的异质性。

4 结 论

本文主要考察了并非指数族分布的威布尔脆弱性模型的建模，使用等级似然方法对其进行估计。结果显示，等级似然方法可以很好给出威布尔脆弱性模型的估计结果，调整的轮廓等级似然可以用来估计得到精确的随机效应的参数。模拟结果显示在纵列持续数据分析和应用中，对带有随机效应的生存数据，不能忽略随机效应的存在，否则，Cox 模型会有很大偏误。对不同随机效应的生存数据，当然理论上是能够使用真实分布的脆弱性模型估计会很好，但是模拟研究发现，对正态分布、伽马分布、威布尔分布这三种随机效应的生存数据来说，威布尔脆弱性模型估计结果显示出很大的优势，尤其当删失率提高时，这种优势最为明显。本文模拟和应用研究均表明，等级似然估计在实际应用中对于中等以下规模（样本容量 50 以下）的样本数据具有很好的稳健性。因此，除了伽玛脆弱性模型和对数正态脆弱性模型之外，威布尔脆弱性模型是也是一个非常值得应用的模型。

参考文献：

[1] COX D R. Regression models in life tables (with discussion) [J]. J Roy Statist Soc: Ser B, 1972, 34: 187 – 220.

[2] PICKLES R, CROUCHLEY. A comparison of frailty models for multivariate survival data [J]. Statistics in Medicine, 1995, 14(13): 1447 – 1461.

[3] HOUGAARD P. Frailty models for survival data [J]. Lifetime data analysis, 1995, 1(3): 255 – 273.

[4] MCLACHLAN G, PEEL D. Mixtures of factor analyzers [J]. Finite Mixture Models, 2000: 238 – 256.

[5] JAMSHIDIAN M, JENNRICH R I. Acceleration of the EM algorithm by using quasi-Newton methods [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1997, 59(3): 569 – 587.

[6] LEE Y, NELDER J A. Hierarchical generalized linear models [J]. Journal of the Royal Statistical Society: Series B (Methodological), 1996: 619 – 678.

[7] HA IL DO, LEE Y, SONG J K. Hierarchical likelihood approach for frailty models [J]. Biometrika, 2001, 88(1): 233 – 233.

[8] HA IL DO, LEE Y, SONG J K. Hierarchical-likelihood approach for mixed linear models with censored data [J]. Lifetime data analysis, 2002, 8(2): 163 – 176.

[9] HA IL DO, LEE Y. Multilevel mixed linear models for survival data [J]. Lifetime Data Analysis, 2005, 11(1): 131 – 142.

[10] 徐淑一,王宁宁. 竞争风险下纵列数据的随机效应建模和估计[J]. 中山大学学报:自然科学版, 2007, 46(1): 7 – 10.

[11] LEE Y, JANG M, LEE W. Prediction interval for disease mapping using hierarchical likelihood [J]. Computational Statistics, 2011, 26(1): 159 – 179.

[12] NOH M, LEE Y, KENWARD M G. Robust estimation of dropout models using hierarchical likelihood [J]. Journal of Statistical Computation and Simulation, 2011, 81(6); 693 – 706.

[13] NOH M, WU L, LEE Y. Hierarchical likelihood methods for nonlinear and generalized linear mixed models with missing data and measurement errors in covariates [J]. Journal of Multivariate Analysis, 2012, 109: 42 – 51.

[14] LEE D, LEE Y, PAIK M C, et al. Robust inference using hierarchical likelihood approach for heavy-tailed longitudinal outcomes with missing data: An alternative to inverse probability weighted generalized estimating equations [J]. Computational Statistics & Data Analysis, 2013, 59: 171 – 179.

[15] WU J, BENTLER P M. Application of H-likelihood to factor analysis models with binary response data [J]. Journal of Multivariate Analysis, 2012, 106: 72 – 79.