文章编号：2095-2163(2023)02-0114-05 中图分类号：O212.1 文献标志码：A

区间删失数据下比例风险模型的概率填补方法

朱 婷，苟洪山，李 荣
（贵州民族大学 数据科学与信息工程学院，贵阳 550025）

摘 要：在生存分析中，填补是删失数据下比例风险模型回归分析的重要方法。在区间删失数据下，单点填补的方法存在参数估计不稳定，本文在混合插值的基础上提出了一种概率填补方法，该方法利用删失数据集的深层信息，依概率进行有效填补。经模拟和实证分析论证，在区间删失数据下比例风险模型回归分析中，概率填补方法比目前单点填补方法的参数估计更稳定。

关键词：区间删失数据；比例风险模型；概率填补；参数估计

A probability imputation method for proportional hazards model with interval-censored data

ZHU Ting, GOU Hongshan, LI Rong
(School of Date Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, China.)

【Abstract】 In survival analysis, imputation is an important method for regression analysis of proportional hazards model under censored data. Under interval censored data, the single-point imputation method has the shortcoming of parameter instability. In this paper, we propose a novel probability imputation method. This method uses the deep information of the censored data set to effectively impute according to the probability. After simulation and empirical analysis and demonstration, in the proportional hazards model regression analysis under the interval censored data, the probability imputation method is better than the current single point method. Parameter estimates for padding methods are more stable.

【Key words】 interval-censored data; proportional hazards model; probability imputation; parameter estimation

0 引 言

生存分析是研究生存现象和响应时间数据及其统计规律的。在实际研究中，由于各种各样的条件限制，无法观测出准确的数据，只知道这个数据大于、小于某个值或在两个值之间，这样的不完全数据在生存分析中被称为删失数据。常见的删失类型有区间删失、右删失和左删失，其中区间删失是一种常见的删失类型。

Cox[1]提出了 cox 比例风险模型，对于右删失数据而言 cox 比例风险模型能直接适用，是生存分析中最重要的模型之一，cox 比例风险模型有参数部分即参数所在的指数函数，还有非参数部分即基底风险函数。cox 比例风险模型对右删失数据进行分析时，可将非参数部分抵消掉，从而借助偏似然函数进行参数估计，但分析区间删失数据时无法抵消非参数部分，增加了估计回归参数的难度；Finkelstein[2]首次在区间删失数据下对 cox 比例风险模型的基底风险函数和回归参数运用牛顿-拉夫森算法（Newton-Raphson 算法）进行估计；Goggins 等[3]采用期望最大化算法（EM 算法）对区间删失数据下 cox 比例风险模型中的参数进行估计；Betendky 等[4]采用局部似然的方法对区间删失数据下 cox 比例风险模型进行拟合；这些都是最大似然法的思想。另一种思想是填补的思想，就是将区间删失数据填补后转换为右删失数据再进行参数估计，填补分为单点填补和多重填补，Pan[5]将多重填补法运用于区间删失数据下 cox 比例风险模型的参数估计中，多重填补法需要将 cox 比例风险模型的非参数部分估计出来，这也失去了 cox 比例风险模型不依赖非

作者简介：朱 婷（1997-），女，硕士研究生，主要研究方向：统计模型与统计计算；苟洪山（1995-），男，硕士研究生，主要研究方向：统计建模与模式识别；李 荣（1980-），女，硕士，副教授，主要研究方向：设备可靠性分析与统计建模。
通讯作者：李 荣 Email：lirongjiewu@126.com

收稿日期：2022-06-17

参数部分的优势；Sun[6]将单点填补法运用于区间删失数据下 cox 比例风险模型的参数估计中；Sun 等[7]提出将左端点填补法运用于区间删失数据下 cox 比例风险模型的参数估计中，并证明了可行性。

只用删失区间的一侧端点填补会丧失部分信息，导致估计结果出现较大偏差，安玉洁[8]提出了混合填补方法，在一定条件下左端点、右端点和中点都会成为部分删失区间的填补值，效果更稳定。本文运用聚类中心的思想，在混合填补法的基础上，提出一种概率填补方法，即利用删失区间的信息也利用未删失数据信息，通过迭代的方式减少填补偏差，不依赖 cox 比例风险模型中非参数部分的估计。

1　模　型

n 个观察对象，对于第 i 个观察对象，T_i 表示生存时间，设存在一个用 $X_i, i = 1,2,\cdots,n$，表示的协变量向量，假设 T_i 满足 cox 比例风险模型：

$$\lambda(t \mid X_i) = \lambda_0(t)^{\exp(X\beta')} \tag{1}$$

其中，λ_0 是未知的基底风险函数，$\boldsymbol{\beta}'$ 是回归参数向量。

因为区间删失，不能直接观测到 T_i，仅仅知道在一个删失区间 (L_i, R_i) 内，右删失时是 $R_i = \infty$，左删失时是 $L_i = -\infty$，左删失和右删失都可以视为区间删失的一种特殊类型。通常假设 T_i 独立于删失机制，为了区别左、右删失，将有限 (L_i, R_i) 的观测称为有限区间删失。

基于 n 个观测值 $(L_1, R_1, X_1), \cdots, (L_n, R_n, X_n)$，本文的最终目的是估计回归参数 β。

2　概率填补法

有限区间删失可以将删失区间中的真实生存时间视为缺失，如果用确切的时间点替换每个有限的删失区间，如用左端点替换、右端点替换和中点替换，就可以使用常规方法来分析填补数据。混合插值考虑部分删失区间用左端点填补、右端点填补和中点填补，用这种混合填补的方式减少信息的损失，使最终的参数估计结果更加理想，但是混合填补法填补的值虽然使用了删失区间所含的信息，但是未挖掘其深层次信息。为了挖掘删失区间中更深层次的信息，本文提出了概率填补方法，利用了删失区间的深度信息和未删失数据所含信息。

深度信息是指使用改进的 K-means 算法提取混合填补法填补的时间和真实时间的聚类信息。概率填补方法是从深度信息中依概率选取删失区间的代表元作为迭代算法前进方向，并且通过迭代的方式优化填补时间，从而使得参数估计结果更好。具体步骤如下：

(1)采用混合填补法将区间删失数据转换为右删失数据；

(2)对右删失数据集中非右删失的数据提取多个代表元，并判断每个删失区间有几个代表元；

(3)产生可能的右删失数据 $\{T_{(i+1)j}, \delta_j, X_j\}$。对于有限删失区间 (L_j, R_j)，判断每个有限删失区间 (L_j, R_j) 中的代表元，若删失区间中只有一个代表元 V_1 或者删失区间中没有代表元，则删失区间前一次填补的值不变，即 $T_{(i+1)j} = T_{ij}, \delta_j = 1$；若删失区间不止一个代表元，则先从一个均匀分布中随机取出一个值 α，选择最大代表概率的代表元 V_2 即聚类数目最多的聚类中心，将判断点 $V_2 + \alpha$ 和判断点 $V_2 - \alpha$ 与前一次填补值作比较，若删失区间前一次填补的值 $T_{ij} > V_2 + \alpha$，则再从一个均匀分布中随机取出一个值 ε，使得 $T_{(i+1)j} = T_i - \varepsilon, \delta_j = 1$，删失区间前一次填补值 $T_{ij} < V_2 - \alpha$，则 $T_{(i+1)j} = T_i + \varepsilon, \delta_j = 1$；若填补值与判断点相等，则填补值不变。对于 $R_j = \infty$ 时，令 $T_{(i+1)j} = L_j$，且 $\delta_j = 0, \delta_j$ 是删失函数，$\delta_j = 0$ 表删失，$\delta_j = 1$ 表未删失，$j = 1,2,\cdots,n$，下标 i 表示第 i 次填补；

(4)收敛标准是每 5 次概率填补法填补后得到数据集的参数估计值两两相减求平均值小于 0.01 或 i 大于 50 次则停止迭代，最终收敛的 $\hat{\beta}_i$ 是参数估计值。

在步骤(1)中，当协变量 X_i 取有限多个值时，要先对区间删失数据分类。为了方便一个协变量取值为 0 和 1。首先按协变量取值对区间删失数据分两类，分别求区间右端点均值 u，标准差 sd，设判断点 $A_1 = u - sd$ 和 $A_2 = u + sd$。当协变量 X_i 是连续性变量时，则直接计算所有删失区间的均值和标准差，以求得判断点 A_1 和 A_2。若对应的删失区间落在判断点 A_1 的左边，则用右端点代替真实的时间；若落在判断点 A_2 的右边，则用区间的左端点来代替；若删失区间与判断点区间 (A_1, A_2) 有交点，则取删失区间的中点作真实时间。

概率填补法在混合填补的基础上进行填补，混合填补依赖于删失区间得到的判断点，当删失区间很大或者数量较少时，得到的信息就不够代表整个数据集的信息，得到的判断点也不够有效。而概率填补法解决了这两个问题，提取代表元的方法是对数据集中非右删失的数据考虑了数据集中不同类型

的数据,得到的结果更能反映数据集的信息;在模拟中,设每个类中的组内方差小于 0.1,选出 K 个聚类中心, K 由组内方差决定,聚类中心就是代表元。

概率填补法没有固定判断点和填补值,允许这两个值在一个有效范围内波动,判断点和填补值的有效范围通常围绕非右删失数据的标准差和删失区间的长度取值,希望使填补值有更多的可能性去在删失区间内接近真实值,以求最后的估计结果更加准确。如:在模拟中根据数据集中的时间点,从均匀分布 $U(0.05,0.15)$ 中随机取出一个值 α ,得到判断点 $V_2+\alpha$ 和 $V_2-\alpha$ 判断如何填补,再从均匀分布 $U(0,0.05)$ 中随机取出一个值 ε 以调整填补值的大小,以这种方式为单点填补方法增加可变性,得到填补数据集,用填补后的右删失数据集 $\{T_{(i+1)j},\delta_j,X_j\}$ 去拟合 cox 模型,得到参数估计。

3 模拟研究

通过比较概率填补法和混合填补法、左端点填补法、右端点填补法、中点填补法这 5 种填补方法的填补性能,证明了概率填补方法能有效提升参数估计的效果。

为了验证概率填补方法在不同情况下所填补的数据在 cox 模型中都能估计出较好的参数值,分别在含有一定比例准确生存时间的数据集和不含准确生存时间的数据集中设置不同的样本量和删失率进行实验。在模拟数据集中的回归参数真值 $\beta=1$, 生存分布为威布尔分布,其中形状参数 $\alpha=2$, 尺度参数 $\lambda=1$。

数据集 a(含有一定比例准确生存时间的数据集)有准确生存数据、右删失数据和区间删失数据,通过在准确的生存时间中制造右删失,右删失比例 $N\%$ 可通过调整删失变量 $F\sim U(0,c)$ 中 c 的大小得到, c 由模拟迭代计算得出,生成需要的删失时间点 F_j。 设生存时间 T_j 和 F_j 相互独立,若 $T_j<F_j,\delta_j=1$ 表未删失; $T_j\geqslant F_j,\delta_j=0$ 表删失,则观测的生存时间 $t_j=\min(T_j,F_j)$。 在已经生成右删失时间的数据集上,取没有删失的数据再产生区间删失数据,为了得到不同的删失区间长度,令两次检查的时间间隔 l 相等, 假定有 $k+1$ 次检查, 从均匀分布 $U(0,0.7)$ 中产生一个间隔 l, 在均匀分布 $U(l-0.05,l+0.05)$ 产生第一个删失区间的右端点 e_j。为模拟实际情况,设删失区间的左端点不为 0,所以第一个删失区间的左端点 e_{0j} 从均匀分布 $U(0,0.001)$ 取出,从 $(e_{0j},e_j),(e_j,e_j+l),\cdots(e_j+(k-1)*l,e_j+k*l)$ 中选择一个区间 (L_j,R_j), 使得 T_j 满足 $L_j<T_j<R_j$。 重复以上步骤得到需要的删失区间,并将没有删失的时间按区间删失比例替换成删失区间。

数据集 b(不含准确生存时间的数据集)中只有右删失数据和区间删失数据。首先,在完整数据集中制造 $N\%$ 的右删失数据,然后剩余 $1-N\%$ 数据制造为区间删失数据(区间删失数据生成过程和数据集 a 相同)。

所有的估计结果均由 200 次独立模拟获得,每一次模拟都计算 5 种填补方法在同一数据集中的估计结果,5 种填补方法分别为本文所提概率填补方法(PIA),混合填补法(MIA),左端点填补法(LEPIA),右端点填补法(REPIA)和中点填补法(MPIA);其次,在偏差($Bias$)、平均绝对误差(MAE)、均方根误差($RMSE$)、标准离差(SSE)这 4 类定量误差结果下比较 REPIA 与 MPIA、LEPIA、MPIA 和 REPIA 在不同样本量和不同删失率下填补后的参数估计效果, $Bias$ 是参数估计值的平均值和参数真值的差值。

平均绝对误差 (MAE), 公式(2):

$$MAE=\frac{1}{M}\sum_{m}^{M}|\hat{\beta}_m-\hat{\beta}_{m0}| \tag{2}$$

均方根误差 ($RMSE$),公式(3):

$$RMSE=\sqrt{\frac{1}{M}\sum_{m}^{M}(\hat{\beta}_m-\hat{\beta}_{m0})^2} \tag{3}$$

标准离差 (SSE) , 公式(4):

$$SSE=\sqrt{\frac{1}{M}\sum_{m}^{M}(\hat{\beta}_m-\bar{\hat{\beta}}_m)^2} \tag{4}$$

其中, $\hat{\beta}_m$ 为第 m 次模拟删失数据的参数估计值; $\hat{\beta}_{m0}$ 为第 m 次模拟完整真实数据的参数估计值; $\bar{\hat{\beta}}_m$ 为 m 次模拟删失数据的参数估计值的平均值。

在数据集 a 中固定右删失率为 20%时,不同样本量、不同区间删失率下 5 种填补方法的 4 类误差结果见表 1 和表 2。由表 1 和表 2 的实验结果可知,在数据集 a 中不同删失率下时, 概率填补法填补的数据的参数估计误差相比于混合填补法、左端点填补法、右端点填补法和中点填补法所填补的数据的参数估计误差有所减小,而且在样本量 200,不同删失率下,概率填补法的标准离差总是小于其他方法的标准离差,说明随着样本量增加概率填补法比其他填补方法效果更好、更稳定。

表 1　数据集 a 中样本量 200 时 5 种填补方法的 4 类误差结果

Tab. 1　Error of four types of five imputations when the sample size is 200 in data set a

区间删失率	方法	*RMSE*	*SSE*	*MAE*	*Bias*
30%	PIA	0.075 314 57	0.149 151 6	0.058 476 85	−0.010 061 50
	MIA	0.076 557 47	0.150 995 5	0.058 792 45	−0.006 372 15
	LEPIA	0.083 109 01	0.164 927 7	0.067 548 99	−0.035 070 81
	REPIA	0.110 574 60	0.166 052 5	0.091 346 21	−0.058 528 30
	MPIA	0.076 684 68	0.164 213 3	0.060 220 81	−0.017 459 95
50%	PIA	0.089 018 36	0.175 811 4	0.072 130 23	−0.024 460 98
	MIA	0.091 958 67	0.177 423 5	0.073 653 50	−0.019 566 70
	LEPIA	0.098 690 04	0.182 512 8	0.078 725 01	−0.065 722 71
	REPIA	0.127 043 30	0.185 561 4	0.105 642 89	−0.086 393 59
	MPIA	0.089 943 08	0.187 100 9	0.073 180 34	−0.036 985 47

表 2　数据集 a 中样本量 100 时 5 种填补方法的 4 类误差结果

Tab. 2　Error of four types of five imputations when the sample size is 100 in data set a

区间删失率	方法	*RMSE*	*SSE*	*MAE*	*Bias*
30%	PIA	0.113 884 6	0.265 200 1	0.093 161 88	0.009 999 891
	MIA	0.115 646 1	0.267 144 8	0.093 755 26	0.014 790 260
	LEPIA	0.113 806 7	0.272 921 4	0.093 998 65	−0.004 676 539
	REPIA	0.170 132 0	0.272 988 7	0.135 357 97	−0.056 756 630
	MPIA	0.115 607 1	0.272 611 3	0.094 350 17	0.003 575 320
50%	PIA	0.121 116 3	0.252 647 9	0.096 233 59	0.005 061 865
	MIA	0.121 934 8	0.256 997 8	0.096 865 60	0.013 623 730
	LEPIA	0.146 855 8	0.257 983 8	0.120 716 13	−0.031 805 800
	REPIA	0.170 578 0	0.267 304 0	0.134 303 43	−0.058 872 170
	MPIA	0.122 405 2	0.261 928 3	0.097 073 30	−0.003 581 620

在数据集 b 中右删失率为 70%和 50%，区间删失率为 30%和 50%时，不同样本量下 5 种填补方法的 4 类误差结果见表 3 和表 4。对比表 3、表 4 的结果可知，数据集 b 中相同条件时概率填补法始终比混合填补法效果更好，与左端点填补、中点填补和右端点填补总体上性能相近。

模拟实验结果可以说明，概率填补法的填补效果比混合填补法效果更好，并且与左端点填补、中点填补和右端点填补相比，概率填补法的填补性能总体上较好。

表 3　数据集 b 中样本量 200 时 5 种填补方法的 4 类误差结果

Tab. 3　Error of four types of five imputations when the sample size is 200 in data set b

区间删失率	方法	*RMSE*	*SSE*	*MAE*	*Bias*
30%	PIA	0.166 343 2	0.223 327 8	0.130 228 9	0.082 143 06
	MIA	0.176 717 3	0.229 426 1	0.139 520 7	0.093 703 11
	LEPIA	0.166 348 0	0.226 596 6	0.132 322 8	−0.044 680 62
	REPIA	0.165 508 1	0.235 149 8	0.135 771 8	0.032 304 53
	MPIA	0.156 035 2	0.230 314 1	0.124 937 7	0.057 277 15
50%	PIA	0.174 690 5	0.221 212 0	0.141 319 5	0.125 232 90
	MIA	0.183 155 4	0.226 663 3	0.148 752 9	0.138 098 40
	LEPIA	0.179 446 6	0.230 601 1	0.142 998 3	−0.010 703 21
	REPIA	0.160 462 3	0.228 435 4	0.127 257 1	0.083 232 14
	MPIA	0.166 849 1	0.232 491 8	0.134 693 3	0.101 288 10

表 4 数据集 b 中样本量 100 时 5 种填补方法的 4 类误差结果

Tab. 4 Error of four types of five imputations when the sample size is 100 in data set b

区间删失率	方法	*RMSE*	*SSE*	*MAE*	*Bias*
30%	PIA	0.348 058 5	0.390 835 9	0.263 570 5	0.148 836
	MIA	0.366 271 3	0.404 786 5	0.274 908 1	0.168 145 7
	LEPIA	0.349 841 5	0.396 519 1	0.277 859	−0.038 572 76
	REPIA	0.376 503 9	0.420 578 1	0.289 634 5	0.170 254 8
	MPIA	0.345 062 9	0.402 351 0	0.261 435 5	0.129 207 3
50%	PIA	0.252 452 4	0.333 545 9	0.191 974 5	0.120 817 6
	MIA	0.261 703 1	0.336 928 4	0.198 088 0	0.137 957 9
	LEPIA	0.249 273 5	0.330 159 3	0.192 636 5	−0.005 431 8
	REPIA	0.255 929 9	0.331 551 9	0.193 933 3	0.082 272 4
	MPIA	0.243 773 5	0.331 026 2	0.186 043 8	0.101 227 5

4 实例分析

应用概率填补方法,在 Sun(2006)的数据集Ⅱ上填补区间删失数据,此数据集中的数据是对来自 5 个研究中心的 368 名患者进行 HIV-1 的感染观察,研究目的是比较未接受因子Ⅷ浓缩物的患者和接受低剂量因子Ⅷ浓缩物的患者之间 HIV-1 的感染风险。在这项研究中,患者的 HIV-1 感染时间只有区间删失数据,不含准确生存时间,未接受因子Ⅷ浓缩物的患者人数为 236 人,接受低剂量因子Ⅷ浓缩物的患者人数为 132 人。对于无剂量组的患者,定义协变量为 0,否则为 1,并假设 HIV-1 感染时间服从 cox 比例风险模型。为了进行比较还对数据采用了最大似然法(MLE)进行估计还有左端点填补法、混合填补法进行填补后估计,其结果见表 5。

表 5 实例数据集中 4 种方法的估计结果

Tab. 5 Estimation results of the four methods in the example dataset

方法	β 估计值	标准误差
PIA	1.842 0	0.220 0
MLE	1.864 4	0.219 4
LEPIA	1.824 4	0.220 2
MIA	1.846 6	0.220 1

表 5 中的结果表明:在不同的方法下,其 β 的估计值都比较接近,且标准误差相比左端点填补法和混合填补法的标准误差是较低的,说明概率填补方法在实际的区间删失数据上填补的数据是有效的。

5 结束语

本文讨论了区间删失数据下比例风险模型的参数估计,许多学者为此提出了不同的方法,其中大多数都涉及未知基底函数的估计。单点填补法将区间删失数据问题转换为右删失数据的问题,避开了未知基底函数的估计,但是一般情况下,当风险函数在很大范围内变化或者删失区间很宽时,使用单点填补法估计会出现偏差较大或者不稳定的情况。例如上述的左、右端点填补和中点填补。而概率填补法主要优点是只涉及回归参数的估计,且估计总体上较为稳定。在样本量较大,含有一定比例准确时间的区间删失数据集下,概率填补法提取的信息更加准确,所以参数估计有更好的效果。模拟和实证分析表明,这种方法是可行和有效的。

参考文献

[1] COX D R. Regression Models and Life-Tables [J]. Journal of the Royal Statistical Society, 1972, 34(2): 187-220.

[2] FINKELSTEIN D.M. A proportional hazards model for interval-censored failure time data [J]. Biometrics, 1986, 42(4): 845-854.

[3] GOGGINS W B, FINKELSTEIN D M, SCHOENFELD D A. A Markov Chain Monte Carlo EM Algorithm for Analyzing Interval-Censored Data under the Cox Proportional Hazards Model [J]. Biometrics, 1998, 54(4): 1498-1507.

[4] BETENSKY R A, LINDSEY J C, RYAN L M. A local likelihood proportional hazards model for interval censored data[J]. Statistics in Medicine, 2002, 21(2): 263-275.

[5] WEI P. A Multiple Imputation Approach to Cox Regression with Interval-Censored Data[J]. Blackwell Publishing Ltd, 2000, 56(1): 199-203.

[6] SUN.J. The statistical analysis of interval-censored failure time data[M]. Berlin:Springer,2006.

[7] SUN J, FENG Y, ZHAO H. Simple estimation procedures for regression analysis of interval-censored failure time data under the proportional hazards model[J]. Lifetime Data Analysis, 2013, 21(1): 138-155.

[8] 安玉洁. 基于区间删失数据的比例风险模型中的参数估计[D]. 华中师范大学, 2013.