事件史分析方法介绍

米 红 曾昭磐

摘要 事件史分析方法是近年来国际学术界发展最为迅速的小样本分析技术。本文从事件史分析方法的源流、概念及方法等方面对此作了比较全面的介绍，并举例说明了事件史分析方法的应用。

作者 米红，男，1996年毕业于西安交通大学人口与经济研究所，获人口系统工程专业博士学位，现任厦门大学人口研究所讲师。(厦门市 361005)

曾昭磐，1960年毕业于厦门大学数学系，现任厦门大学人口研究所所长，教授。(厦门市 361005)

一、事件史分析(Event History Analysis)方法的源流

在社会科学的各个分支领域，有很多极为有趣的事件及其形成这些事件的原因。如：犯罪学家研究犯罪、逮捕、牺牲等事件；经济学家在工作和就业的研究中，注意的焦点则是工作变换、提升、解雇和退休等事件；政治学家则关注暴乱、革命、政府的和平更迭等事件；人口学家则关注出生、死亡、结婚、离婚和迁移等事件，医学家则关心病人访问医生的次数等等。

在上述的每个例子中，每个事件包含一些在特定时点上的定性的变化。因此，一般不能用"事件"来普通的描述一些定性变量变化的过程。因为在这些变量的变化之中，往往包含了一些尖锐的突变。

由于事件是按照时间变化来分类的，因此，研究和认识这些事件及其产生的原因的最好的方法是收集描述这些事件的数据。在其最简单的形式中，即当事件集是以个体的集合出现时，它们往往呈现一种纵向的记录。例如，一项人口调查中的结婚的日期记录就是一个典型的事件史数据集。而如果调查的目的是研究事件发生的原因，那么事件史也应当包括描述其变化的一些解释性变量的数据。

事件史是研究事件发生原因的基础，且具有两个特征：截断(censoring)和随时间变化的解释变量(explanatory variable)——即：能产生利用诸如多元回归分析等统计学方法来描述问题的随机变量。在过去的20年里，一些具有创新性的方法与技术首先在一些特殊的事件史数据的研究得到了应用。因此，事件史方法并非是一种单一的方法，而是一些相关的不断完善的方法与技术的集合体，并在许多领域里由很多深奥和复杂的方法和技术发展起来的。这同时也是引起某些混乱的原因，因为有时事件史中的一些相同的概念经常以不同的方式来表达。因此，有必要对事件史分析方法发展的历史源流作一回顾。

事件史分析方法的源流主要有三个，其一是生物医学统计方法的生存分析(survival analysis)学派；其二是可靠性技术的"失效时间分析"(failure time analysis)学派；其三是数理社会学中的马尔科夫(Markov)学派。

从人口学中可以得知，生命表方法是最早出现的，也是最著名的，应用最广泛的分析方法。自18世纪以来在社会学及生物医学等领域被广泛的采用。直到本世纪50年代末60年代初逐渐被更为先进的方法——事件史分析方法所替代。在生物医学中，用来处理其实际问题的方法又可称为生存数据的分析方法，的确，大多数事件史方法的文献是用生存分析和生命分析的名称而写成的。例如，在实验室进行的动物实验中，通过让动物吃不同剂量的、毒性不一的药物，实验人员可以观察到在不同的过程中动物是如何存活的。在这里，

"事件"就是特指动物的死亡,"截断"的出现则是因为该实验通常是在所有动物死亡以前而结束。生物统计学家已经写了大量的文献说明如何分析这些数据的有效的方法,这些方法已经变成为研究与分析癌症病人存活的数据的主要工具。其中,由该领域中所发展起来的最有影响的回归方法——部分似然估计方法[1]成为事件史分析方法中的重要的核心技术之一。

可靠性分析技术是工程上研究与分析机器和电子元件故障数据所发展起来的重要方法。该领域里所发展起来失效时间分析技术,在过去的十几年里已经被有效的应用于事件史分析之中,并形成了新的流派。

另外,自本世纪60—70年代以来,马尔科夫过程理论在社会科学的应用有了长足的进步和发展,这一方法的转折点是Tuma[2]将解释变量引入连续时间的马尔科夫模型,这可以说是沟通生物统计和工程学方法的一座桥梁。其中,用马尔科夫模型研究社会科学现象变化过程中不同状态的个体事件的分布构成了事件史分析中的重要方法。

二、事件史分析方法的内容概述

1. 事件史分析中的主要概念

重复发生的和不重复发生的事件

事件是事件史分析中最基本的概念,不重复发生的事件则是指只能发生一次的事件;重复发生的事件则是指可以发生两次以上的事件。

如:出生与死亡事件,就属于不重复发生的事件;而工作变换和结婚等在人的一生中可以出现多次的事件,则属于重复性事件。关于重复性事件的模型往往要比不重复事件的模型更为复杂,因此,也更能引出许多复杂的统计问题。然而,掌握不重复事件的分析方法是理解和掌握重复事件模型方法的基础。

单一种类事件和多种类事件

单一种类事件是指影响某一类事件发生的原因是相同的,多种类事件则是指影响某一类事件发生的原因是不同的。

在单一种类事件分析中,用几乎相同的分析方法可以方便的处理所有的事件。例如,在工作变换的研究中,要区分自愿终止工作和非自愿终止工作是非常困难的,也是不必要的。若不考虑这一区别,则可用单一种类事件史分析方法来分析这类事件的变化。然而,在诸如癌症等疾病治疗有效性的研究中,区分因癌症或其它病因导致的死亡显然是很重要的。为了适应不同类型的事件史的研究,类似于人口学中的多减量生命表一样,该领域中发展了一种被称为"完全风险"(competing risks)的方法。然而,由于多种类型事件的引入会导致更为复杂的统计方法的使用,而在某些实际问题中,要完全区分事件的类型有时是不可能的。因此,在实际问题的分析中,这些方法的应用会受到一些局限。

截断(censoring)和解释变量(explanatory vairable)

截断是指所观察的个体在观察的时期内并没有发生变化,但因其它原因而退出观察,或直到观察结束时仍未发生变化,但因研究停止而被中断的个体状态。一般来说,被截断的个体过多,则分析的结果会有较大的偏差。解释变量是指描述事件发生的原因和变化过程的相互独立的随机变量,这是运用多元统计方法建立事件史分析模型的基础。

风险率(hazard rate)

风险率,也称风险概率,是事件史分析方法的核心概率。它描述了所观察的平均每一个体所发生该事件的概率。在离散时间模型中,风险率是指每一个体将在某一特定的时点发生事件的概率,也同时表示了平均每一个体在该时刻所处的风险。在连续时间模型中,设P(t,t+s)是每一个体在时间(t,t+s)中发生事件的风险概率,显然,当s=1时,即为离散时间的风险概率。而当s趋近于零时,定义连续时间的风险概率为h(t),则有:

$$h(t) = \lim_{s \to 0} P(t, t+s)/s \qquad (1)$$

它描述了个体瞬时发生的风险概率,其值可正可负,也可为常数。(这一概念同生命表中死亡力的概念类

似)。若设 P(t)为一个体在时刻 t 之后发生的风险概率,则有: $h(t) = -\frac{P'(t)}{P(t)}$(限于篇幅,证明略)(1)'。

2. 事件史分析的研究方法

关于事件史分析的研究方法可分为两类,一类为离散型的,另一类为连续型的。在这两类方法中,又分别可分为参数或非参数的方法。非参数方法是通过对事件随时间的分布作出可能的假设;参数方法则是用诸如指数分布,威布尔分布,Gompertz 分布等来假设事件之间的关系和事件随时间变化的关系。联系这两种方法的一个重要的桥梁是可以描述半参数或部分参数的 Cox 的比例风险模型[3]。本文限于篇幅,将仅对单一类型和不重复事件史分析方法作一介绍,对于多类型和重复事件史分析模型和方法将另外述及。

离散型事件史分析模型及计算方法

尽管事件是随时间连续变化的,但实际上,时间总是以离散的单位来测量的,当这些时间单位非常小时,通常可以将其作为连续型变量来处理。另一方面,当时间单位很大时,如以月,年,十年为单位时,则更适合用离散时间模型方法来分析,由于离散时间模型方法特别容易理解和应用,因此它可以作为事件史分析基本理论的导论。

下面要引入单一类型不重复事件的离散时间模型方法,所描述的方法也适用于其它许多的情况,而且也可以推广到不同类型事件的重复事件的分析之中。

首先要讨论风险率是如何依赖于解释变量而发生变化的。不仿设 P(t)为一个体在时刻 t 发生的风险概率。又假定有 n 个相互独立解释变量是 $x_1(t), x_2(t), \cdots, X_n(t)$。

作为粗略的估计,可以认为 P(t)是 $x_1(t), x_2(t), \cdots, x_n(t)$的线性方程,即:

$$P(t) = a + b_1x_1(t) + b_2x_2(t) + \cdots\cdots + b_nx_n(t) \qquad (2)$$

$$t = 1, 2, \cdots\cdots 5$$

然而,这里要注意的一个问题是,由于 P(t)是一个概率,因此,它不应大于 1 和小于零。但上式的右端显然可以是任意实数。因此,这一模型是不可能用来进行预测的,且带来计算和解释的复杂性。但这一问题可以通过对 P(t)进行 logit 变换避免,即:

$$log(P(t)/(1 - P(t))) = a + b_1x_1(t) + b_2x_2(t) + \cdots + b_nx_n(t) \qquad (3)$$

当 P(t)在 0、1 之间变化时,方程的左边从负无穷大到正无穷大。尽管还有其它一些形式的变换,但 logit 变换是最方便的和熟悉的形式。回归系数 $b_i(i=1, \cdots\cdots n)$则表示当 $x_i(i=1\cdots\cdots n)$分别变化一个单位时,logit 所变化的数量。

在上述离散模型中,通过让参数 a 随时间而变化,这样,可以得到描述离散型事件史分析方法的核心模型,即:

$$log(P(t)/(1 - P(t))) = a(t) + b_1x_1(t) + b_2x_2(t) + \cdots + b_nx_n(t) \qquad (4)$$

而关于参数 $a(t), b_1, b_2, \cdots, b_n$ 的估计将通过极大似然估计的方法来进行(限于篇幅,从略)。

连续型事件史分析模型及计算方法

虽然离散时间模型方法应用较广泛,但大多数事件史分析方法却是连续型的模型方法。仿前述离散型变量的分析可以得到描述连续型事件史分析的比例风险模型(proportional harzards model)为:

$$logh(t) = a(t) + b_1x_1 + b_2x_2 + \cdots + b_nx_n + u \qquad (5)$$

这里 h(t)是连续时间风险率,$x_i(i=1, \cdots n)$是相互独立的解释变量,$a, b_1, b_2, \cdots, b_n, u$ 是一个独立于 $x_i(i=1, \cdots n)$的随机分布项。

对于给定 u 以不同的分布,可以得到连续型风险率将适合诸如指数(exponential)分布,威布尔(weibull)分布,冈帕兹(Gompertz)分布,伽马(gamma)分布等的情形。

之所以将模型(5)称为比例风险模型是因为在该模型中在任意时点上的两个个体所处的风险概率之比$(h_i(t)/h_j(t) = c)$c 仅依赖于解释变量,而与时间无关。(本文因限于篇幅,不再证明)

关于部分似然估计方法(partial likelihood method)的数学细节见附录。这里要谈一些该方法需要注意的问题。部分似然估计方法实际上依赖于所建立的似然函数中所引用的数据的情况,这些数据在比例风险模型

中可分为两类，第一类是决定回归系数 $b_i(i=1,\cdots,n)$的部分，第二类是决定回归系数 $b_i(i=1,\cdots,n)$和 $a(t)$的部分。特别是，这第一类因素仅依赖于事件发生的顺序，而不是事件发生的具体时间。

关于其结果的估计是渐近无偏的，并服从正态分布。但由于该方法忽视了发生事件的确切时间，所以其结果有时并非是完全有效的。但是这种无效性通常是很小的，以至于没有必要担心。[4]关于在比例风险模型下利用部分似然估计参数可以通过一些统计软件包如：SAS、SPSS、BMDP 等来进行。这里不再赘述。

三、事件史案例分析

让我们从一个实际例子中考察一下离散型事件史分析方法的应用。这是一个包含 200 个男性生物化学家的样本，他们分别在本世纪 50 年代末期和 60 年代初期获得博士学位，并且在这期间的某一时刻毕业后在大学里作助理教授。从他们作助理教授开始，他们被观察的最大时间是 5 年。如表 1 所示。

表 1　200 名生物化学家工作变换的分布

年	工作变换人数	风险个数	风险率估计
1	11	200	.055
2	25	189	.132
3	10	164	.061
4	13	154	.084
5	12	141	.085
>5	129		
合计	200	848	

由表 1 可以看出，这些事件是以离散时间来记录的，以年为单位。要想区别这些事件是由于他们自愿的还是不自愿的发生是不可能的。因此，我们把这些事件作为单一类(同类)事件来处理。表 1 表示了生物化学家们在 5 年中的每一年中变换工作的次数。在总数为 200 个个体的样本中，有 129 个生物化学家在所观察的时期里没有变换工作，因此，可以将这些个体看作为截断事件。

现在，让我们来估计反映一年期工作变换的风险概率和基于选择的 5 个解释变量的回归模型。其中的两个变量描述了初始就业机构的特征，即为：机构的威信(x_1)和获得的联邦基金(x_2)，根据对它们的计算，可以认为它们都是常量。其它三个变量则按年描述了生物化学家个体的情况：出版文章的累积数($x_3(t)$)；被其他学者所引用的学术著作数($x_4(t)$)；学术级别($x_5(t)$)——若是副教授则定为 1，若是助理教授则定为 0。(虽然在观察时期的起始时刻，他们都是助理教授，但有些人很快就得到了提升)。这样，我们就不仅有了常量，而且有了时变的解释变量。

对于这 200 个生物化学家的样本来说，他们在第一年里变换工作的风险就是整个样本在那一年里的风险集。在那一年里，仅有 11 人变换了工作，则这 11 人在第二年里将不再有风险率。(也许他们中的人在第二年仍会变换工作，但由于其概率很小，这里不再考虑这种情形)因此，在每一年结束时，由于事件的总量在减少，因此，风险集合的数量也在减少。

例如，在表 1 中，可以看到风险集合的数量从第 1 年的 200 下降到第 5 年的 141。在本例中，第一次变换工作的概率对那些还没有变换工作的人构成了风险。

若假设风险率随年而变化，但每一年的每一个体的风险率均相同，则可以容易估计出其风险率：例如，在第二年里，在风险集合为 189 个个体中，有 25 个个体变换了工作。则其风险率为 25/189 = 0.132。其它年度风险率的估计可以类似的得到，如表 1 所示。从表 1 看到，并没有风险(关于工作变换)明显的增加或减少的变化趋势。但应注意的是，因为风险集合的个数在严格减少，即使是工作变换的个数在减少，其风险率也可能在增加。例如，虽然第一年的工作变换个数要小于第一年，但第三年的风险率要大于第一年的风险率。

下面将根据离散模型(6)来估计参数 $b_i(i=1,\cdots 5)$和 $a(t)$的 5 个值。

$$log(P(t)/(1-P(t))) = a(t) + \sum_{i=1}^{2} b_i x_i + \sum_{i=3}^{5} b_i x_i(t) \qquad (6)$$

其中 a(t)代表 5 个不同的常量,这些常量分别代表了每一年被观察到的量,将在模型中被估计出来。

实际上,整个过程可以描述如下:对于处于风险下的每个个体的时间单位,可以产生一个分离的观察记录。在本例中,可以人年作为所观察的时间单位。这样,每个在第一年里改变工作的人就贡献了一人年数。那些在第 3 年里改变工作的人,每人则贡献了 3 人年数。而截断的个体——那些在第 5 年里仍保持工作场所不变的人,每人贡献 5 人年数。这样,在样本为 200 个生物化学家的案例中,一共有 848 个人年数。从表 1 还可以清楚的看到,这一数值是 5 年中每一年风险数的累积。

对每一人年来说,若在该年度改变工作则记为 1,否则记为 0。其它的解释变量被分配给在每个人年应承担的数值。最后的步骤是将 848 个人年数引入这一样本,利用极大似然估计的方法对 logit 模型进行估计。(利用 SPSS 和 SAS 软件进行)对模型(6)中的回归系数的估计和统计检验如表 2 所示。

表 2　用 Logit 模型预测工作变换概率的估计　　848 人年数

解释变量数	模型参数		解释变量数	模型参数	
	b	t		b	t
所在机构	.056	.26	第 1 年	-.96	-2.11*
学术权威			第 2 年	.025	.06
基　金	-.078	-2.47*	第 3 年	-.74	-1.60
出 版 物	.023	.79	第 4 年	-.18	-.42
引用著作	.0069	2.33**	常量(>5)	2.35	
任职级别	-1.6	-3.12**	对数似然函数	226.25	

* 置信度为 0.05 的双边检验

** 置信度为 0.01 的双边检验

四、结论

尽管事件史分析方法在处理小样本问题上取得了很大的成功,但该方法也存在着两个重要的限制。其一是风险关于时间的变化在模型的估计时被忽略掉了;其二是比例风险模型及其相关的部分似然估计方法并不包括代表未观察到的异质性的随机扰动项。事实上,已经证明[5]部分似然方法并不包含这样的扰动项的分析和估计。但若仅研究不重复事件时,这一限制的影响是很小的。

总而言之,事件史分析方法是一门正在发展的应用学科,其存在的问题将在未来的实践中得到不断的修正与完善。

*　　*　　*　　*

附录:部分似然估计方法介绍

部分似然估计方法的第一步像极大似然估计方法一样,建立依赖于未知参数和观察值的似然函数;第二步是寻找使该方程极大化的参数值。然而,普通的似然函数是样本中所有个体的似然函数,而部分似然估计是在观察中所发生的事件所构成的似然函数。即:

$$PL = \prod_{k=1}^{k} L_k \qquad (A1)$$

这里,PL 是部分似然函数,K 是样本中所发生事件的总数。

为了搞清楚 L_k 的构成,让我们仍然考虑表 A1 中的例子,这是一个有 10 个个体的样本,但只有 5 个事件被观察到,其它 5 个事件是截断的,三个是在 12 时截断的,因为研究是在那时终止。第 4 个和第 6 个事件分别是在 5 时和 9 时截断的。这两个截断事件发生的原因也许是由于个体的死亡,或从研究的样本中离开,或是不适于作进一步考察的个体。

表 A1　部分似然估计计算举例

i	t_i	k	L_k
1	2	1	$e^{bx1}/(e^{bx1}+e^{bx2}+\cdots+e^{bx10})$
2	4	2	$e^{bx2}/(e^{bx1}+e^{bx2}+\cdots+e^{bx10})$
3	5	3	$e^{bx3}/(e^{bx1}+e^{bx2}+\cdots+e^{bx10})$
4	5*		
5	6	4	$e^{bx5}/(e^{bx1}+e^{bx2}+\cdots+e^{bx10})$
6	9		
7	11	5	$e^{bx7}/(e^{bx1}+e^{bx2}+\cdots+e^{bx10})$
8	12*		
9	12*		
10	12*		

注：i 表示个体，t_i 第 i 个个体事件发生或截断的时刻，k 表示事件数

* 表示截断情况

为方便起见，这些被观察的个体按照事件先后出现的或被截断的时刻来排列的。如表 A1 所示。第一个发生的事件是个体 1 在时刻 2 发生的。而事实上，样本中的 10 个个体在时刻 2 都处于发生事件的风险之中。则个体 1 在时刻 2 发生事件的概率为：

$$L_1 = \frac{h_1(2)}{h_1(2)+\cdots+h_1(2)} \tag{A2}$$

假设风险函数服从指数分布，则有：

$$h_i = \exp(a(t)+bx_i) = \exp(a(t))exp(bx_i) \tag{A3}$$

代入(A2)并化简得：

$$L_1 = \frac{exp(bx_1)}{\exp(bx_1)+\cdots+exp(bx_{10})} \tag{A4}$$

同理可得：

$$L_2 = \frac{exp(bx_2)}{\exp(bx_2)+\cdots+exp(bx_{10})} \tag{A5}$$

其它诸如 L_3, L_4, L_5 的计算结果如表 A1 所示。

需要注意的是，每个 L_i 的值并不依赖于第 i 个事件的出现的确切时刻。其意义只是表示在第 i－1 个事件之后和第 i＋1 个事件之前出现，并表示在任何时刻出现第 i 次事件的概率都是相同的。只有事件出现的顺序将影响部分似然函数的变化。

这样，将式(Ai)(i＝2,3,4,5,6)代入式 A(1)之中，在取对数，就得到 LogPL 的表达式。一旦对数部分似然函数建立以后，就可以同极大似然方法一样对未知参数进行估计了。(本文限于篇幅，不再赘述)

参考文献：

1 3 Cox, D. R. (1972)"Regression models and life tabel." Journal of the Royal statistical Society, Series B 34:187 202.

2 Tuma(1976) "Rewards, resources and the rate of mobility: a nonstationary multivariate stochastic model." American Sociological Review 41:338 360.

4 Efron, B(1977)"The efficiency of Cox's likelihood function for censored data." Journal of the Amrican Statistical Association 72:557－565.

5 Tuma, N. B. (1982)"Nonparametric and partially parametric approaches to event history analysis," PP. 1 60 in S. Leinhardt (ed.) Sociology Methodology 1982. San Francisco: jossey－Bass.