生存分析的应用误区*

李 强　徐 刚　陈丽梅

【摘　要】随着调查数据的日渐丰富和统计软件易用性的不断提高，生存分析在人口学和社会学研究领域中得到广泛的应用。但对这两个研究领域的SSCI和CSSCI来源期刊中应用生存分析的文献进行考察后发现，目前在生存分析应用中存在一些误区：(1)仅考虑生存数据的右删截特征，忽略截平等特征，导致错误的似然函数、参数误估和模型拟合差等问题。右删截虽然是生存数据最常见的特征，但不是唯一的特征。判断生存数据的特征要综合考虑事件的起点时间、历险时间及观测时间；(2)忽视Cox比例风险模型的等比例假设检验；(3)把时变变量误用为非时变变量而造成估计偏差；(4)死亡风险分析中，错将观测期当作历险时间，实际上，死亡风险的历险时间应该是年龄。此外，中文文献中对生存分析的专有名词翻译不一致。导致这些问题的主要原因是使用者没有透彻理解该方法。生存分析的规范和正确应用要求使用者准确理解和把握重要的理论和方法细节。

【关键词】生存分析　事件史分析　删截　截平　历险时间

【作　者】李 强　华东师范大学中国现代城市研究中心暨社会发展学院人口研究所，副教授；徐 刚　华东师范大学社会发展学院人口研究所，硕士研究生；陈丽梅　华东师范大学社会发展学院人口研究所，博士。

一、引　言

生存分析(也称事件史分析)研究事件发生的历险时间及其影响因素。1662年，英国的格兰特(Graunt)创制了第一张死亡生命表，标志着人口学的诞生，也标志着生存分析的诞生。生存分析能揭示事件发生风险随历险时间的变化模式，也能探究事物的因果关系(Guo，2010)，还可以有效分析不完全的历险时间数据(Klein等，2003)。因此，在调查数据的日益丰富和统计软件易用性的推动下，生存分析在包括人口学、社会学、经济学、临床医学、流行病学和工程学等在内的众多学科和领域得到广泛应用，为人类探知纷繁的社会现象、复杂的个人行为等提供了便捷的方法。人口学和社会学中的死亡、婚

* 本文为国家自然科学基金青年项目“中国老年人健康预期寿命的区域差异、影响因素与对策研究”(编号：71503082)的阶段性成果。

姻、生育、迁移、教育、就业、退休、犯罪、性行为等研究都使用了生存分析方法(Gampe,2010;Long等,1993;Massey等,1997;Sweeney,2002;Whitbeck等,1999;潘光辉,2017)。

然而,生存分析在很多应用中并没有被规范和正确地使用。例如,生存分析对不完全数据的有效处理,不仅包括要处理右删截数据,必要时还要处理左删截、区间删截和截平的情况(Klein等,2003)。在实际应用过程中很多应用往往只考虑右删截,而忽略其他不完全数据类型,这样会引致错误的似然函数、有偏的参数估计和较差的模型拟合度,得到的结论也可能失之毫厘、谬以千里(Thiébaut等,2004;李强、张震,2009)。再如,Cox比例风险模型在应用前需要检验能否满足等比例假设,如果不满足,使用Cox模型便有可能导致偏误(Xue等,2017)。在大部分应用Cox模型的文献中,使用者并没有检验等比例假设。

生存分析误用的主要原因在于使用者没有正确理解该方法。生存分析具有一些不同于一般回归分析的特点,如删截和截平数据特征、历险时间和观测时间的区分、时变变量等,这些都要求使用者对方法的准确把握。目前主流的统计分析软件,如Stata、SPSS、SAS和R等都提供了生存分析软件包,这使生存分析的实践应用变得简便和容易,然而,生存分析的基本要求并没有因软件易用性的提高而降低。即使软件运行正常,也能给出一些统计结果,但并不表明生存分析得到了正确的应用。为了更好地促进生存分析的规范和正确使用,本文在简要回顾生存分析方法的基础上,考察了"Web of Science"(SSCI)和"中文社会科学引文索引"(CSSCI)来源期刊中人口学和社会学领域应用生存分析的文献,总结出生存分析使用中存在的一些问题和误解,并给出相应的解决办法和建议。

二、生存分析方法简介

(一)生存分析的基本函数

假设T表示历险时间,是一个非负的随机变量。T的分布主要通过4个函数描述,分别为生存函数、风险函数、概率密度函数和平均剩余寿命。只要知道其中一个函数,就可以推导出其他3个函数,可以说,这几个函数在刻画事件发生风险时是等价的(郭志刚,2001)。

1. 生存函数$S(t)$是个体到t时还未经历该事件的概率(事件在t时之后发生)。其定义为:$S(t)=P(T>t)$。当T是连续变量时,$S(t)$是非增的连续函数。$F(t)$为T的分布函数,$F(t)=P(T\leqslant t)$,$S(t)=P(T>t)=1-P(T\leqslant t)=1-F(t)$。

2. 风险函数$h(t)$描述在t时事件仍未发生的个体在下一刻经历事件的风险率,其定义为:$h(t)=\lim\limits_{\Delta t\to 0}\dfrac{P(t\leqslant T<t+\Delta t\,|\,T\geqslant t)}{\Delta t}$。相应的累积风险函数$H(t)$的定义为:$H(t)=\int_0^t h(x)dx=-\ln[S(t)]$。这样,$S(t)=\exp[-H(t)]=\exp[-\int_0^t h(x)dx]$。于是,$h(t)=f(t)/S(t)$。

3. 概率密度函数 $f(t)$ 是 t 时事件发生的非条件概率，其定义为：$f(t)=-\frac{dS(t)}{dt}$。

4. 平均剩余时间 $mrl(t)$ 是事件发生的平均预期时间，如人口学中的平均预期寿命。其定义为：$mrl(t)=E(T-t|T>t)$。其连续形式为：$mrl(t)=\frac{\int_t^{\infty}(x-t)f(x)dx}{S(t)}=\frac{\int_t^{\infty}S(x)dx}{S(t)}$。

（二）历险时间与数据特征

历险时间也称为存活时间或失效时间，是指从历险起点开始到经历事件的时间。比如，从出生到死亡的生命历程，育龄妇女生育一孩到生育二孩的间隔。

以历险时间为区分标准，可以将生存分析数据分为完全数据和不完全数据（Klein等，2003）。在观测期内，如果所有的样本都经历了事件，并且该事件的经历过程是已知的，这样就得到了完全数据。例如，对果蝇的死亡率进行研究，在实验室条件下，可以观测到果蝇从出生到死亡的全过程，可以得到所有果蝇的存活时间信息。但在非实验室条件下，特别是人口学和社会学的数据中，研究者很难获得完全数据，大部分调查数据都是不完全数据。不完全数据有删截（Censoring）和截平（Truncation）两种基本类型（Klein等，2003）。宽泛地讲，删截是指我们知道一些个体的事件发生在某时间区间，但无法获知其确切的时间。例如，在65岁及以上老年人死亡风险的研究中，在调查结束时，有些老人依然健在，虽然他们会在以调查结束为起点、右端为无限的时间区间内死亡，但我们无从得知其确切的死亡时间，这部分老人是删截样本，更准确地说是右删截。对于这些老人来说，其死亡没有发生在观测期，因此他们对死亡事件发生数没有影响；但他们在观测期内一直保持存活状态，所以他们增加了观测期内的历险人年数。死亡率是死亡数除以历险人年数，因此删截样本会影响死亡率的分母，而不是分子。认识到这一点有助于理解后文的似然函数构造。

截平是指只有事件发生时间是在给定观测期内的个体才会被观测到的情况。以65岁及以上老年人的死亡风险为例，只有那些存活到65岁的个体才可能进入观测，这个过程中就得到了截平数据，严格说是左截平。一位64岁死亡的老人，虽然是一个死亡事件，但对该研究来说，并不贡献任何信息，也不会进入调查范围。只有满足一定条件的个体才能进入观测范围，截平是与条件概率相联系的。

根据观测时间与个体事件时间的关系，删截和截平可以细分为左、右和中间三类。三类删截数据包括：（1）右删截。如前所述，被访者在观测期内未经历事件，但由于时间和成本的考虑或调查设计的需要，观测已经结束。研究者不知道这些被访者何时将会经历该事件，但可以肯定他们会在将来的某个时点经历该事件。在很多跟踪调查中，右删截的情况极为常见，因为在某一期调查结束时，总会有一些个体未经历事件。（2）左删截。被访者在某一时点开始进入观测，在此之前该被访者已经经历所研究的事件，但无

从得知准确的事件发生时间。这种情况多出现在医学领域中，例如，在艾滋病研究中，由于艾滋病具有长短不一的潜伏期，当患者在出现症状后前往就医时，患者和医生都难以确定患者感染艾滋病毒的具体时间。（3）区间删截。被访者在某个时间段内经历事件，却不知道具体发生时间。一般而言，这个时间段的起点是左删截端点，终点是右删截端点，这就是区间删截。比较常见的例子是发病前的疼痛这个事件，患者不记得疼痛发生的时间，可能的发生时间是医院两次临床检查之间，这就是区间删截（彭非、王伟，2004）。

类似的，截平数据也分为三类，但比较常见的有两类：（1）左截平。个体只有满足某个条件（通常高于某个门槛条件），才能进入观测。如前所述，在研究 65 岁及以上老人的死亡率时，只关注那些存活到 65 岁及以上的老人。但被纳入调查的老人并不都是整齐划一的 65 岁，因为 65 岁只是一个进入观测的门槛值，只要满足这个条件的老人就会在观测范围。因此，调查中还会有很多 65 岁以上的老人，如 70 岁、80 岁甚至百岁老人。由于研究内容是 65 岁及以上老人的死亡率，所以，65 岁老人可视为是“晚了 65 年”才被调查，70 岁老人可视为是“晚了 70 年”才被调查。这也就是左截平通常也被称为“延迟进入时间”的原因。（2）右截平。与左截平相反，右截平是指事件在给定时间区间内（通常是小于某个条件）发生才能被观测到。例如，根据死亡记录来研究 110 岁以上老年人的死亡率，老年人在达到某一个给定年龄（如 125 岁）前死亡才会被观测到（Gampe，2010）。使用复学学生数据研究辍学风险，只有在调查时点前复学的学生才能被观测到，调查之后复学的学生就不在观测范围内（常保宁，2010）。

从图中可以看到前文所述的几种情况：（1）样本 A 是左截平，因为其在观测开始前已经历险一段时间，且在观测期内经历事件发生；（2）样本 B 是左截平和右截平，只在 $[t_1, t_2]$ 时间窗口内经历事件才能被观测到。（3）样本 C 是左删截的情况，在进入观测 t_1 前已经经历该事件，事件发生的时间不清楚（或者已经忘记）；（4）样本 D 是右删截，在观测结束时还未经历该事件；（5）被完整观测的只有样本 E；（6）样本 F 是左截平和右删截，因为其在观测开始前已经历险一段时间，在观测结束时还未经历事件发生。在很多的跟踪调查中，左截平—右删截组合数据很常见，包括的样本类型如 A、D、E、F 所示。

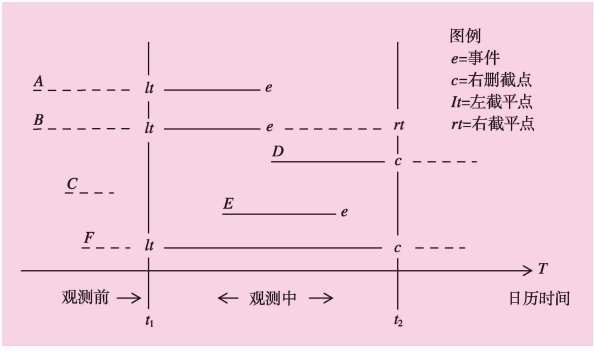

图　各类数据特征的事件史

注：水平轴表示日历时间，用 T 表示。观测期一般是一段有限的时间，始于 t_1，终于 t_2。e 表示事件在观测期内发生，c 表示样本右删截，lt 是指在观测起始 t_1 左截平，rt 是指在观测终点 t_2 右截平。

在实证研究中，对于这类存在删截和截平的数据，传统的分析方法如多元回归往往会将这些不完全但有部分信

息的数据删除，导致分析结果出现偏差。生存分析则可以通过构造风险集来充分利用不完全信息，有效处理删截和截平数据，这也是生存分析的优势之一（Guo，2010）。

（三）似然函数的构造

基于不同的数据特征，为了进行推断和估计参数，需要构造相应的似然函数。构造似然函数的一个关键假设是历险时间、截平时间和删截时间是相互独立的。构造函数时需要明确每一个样本的数据特征。

如果是完全数据，那么这个观测的似然函数就近似等于在该时点 t 的密度函数。如果是右删截数据，即事件的发生时间大于删截时间，那么似然函数为生存函数。如果是左删截数据，事件已经发生，似然函数为累积分布函数。如果是区间删截数据，事件在某一时间段内发生，那么似然函数就是事件在该时间段内的发生概率。对于截平数据，似然函数需要用适当的条件概率（Klein 等，2003）。

用生存分析的基本函数表示为：完全数据 $f(t)$、右删截观测 $S(C_r)$、左删截观测 $1-S(C_l)$、区间删截观测 $[S(L)-S(R)]$、左截平观测 $f(t)/S(Y_L)$、右截平观测 $f(t)/[1-S(Y_R)]$、区间截平观测 $f(t)/[S(Y_L)-S(Y_R)]$。其中，C_r 为固定删截时间，表示右删截；C_l 为开始观测的删截时间，表示左删截；L 为左删截端点；R 为右删截端点；Y_L 为左截平事件发生的时间点，Y_R 为右截平事件发生的时间点。

在构造似然函数时，可以将上述各函数形式组合起来：

$$L \propto \prod_{i \in D} f(t_i) \prod_{i \in R} S(C_r) \prod_{i \in L} [1-S(C_l)] \prod_{i \in I} [S(L_i)-S(R_i)] \qquad (1)$$

式（1）中，D 代表事件发生观测集；R 代表右删截观测集；L 代表左删截观测集；I 代表区间删截观测集。

对于左截平数据，若截平时间 (Y_{Li}, Y_{Ri}) 与事件发生时间独立，可用 $f(t_i)/[S(Y_{Li})-S(Y_{Ri})]$ 和 $S(C_i)/[S(Y_{Li})-S(Y_{Ri})]$ 分别代替上式中的 $f(t_i)$ 和 $S(C_i)$。对于右截平数据，似然函数的形式为：

$$L \propto \prod_i f(Y_i)/[1-S(Y_i)] \qquad (2)$$

（四）几种常用的生存分析模型

生存分析模型包括参数模型、非参数模型和半参数模型。参数模型主要是事件的历险时间遵循某种统计分布，如韦伯分布、Gompertz 分布、指数分布等。非参数模型主要指历险时间不遵循某种统计分布，影响因素也是非参数形式。常见的非参数模型包括生命表分析（大样本时使用）和 Kaplan-Meier 估计（小样本时使用）。半参数模型指事件的历险时间是非参数形式的，而历险时间的影响因素是参数形式的，其中最常用的为 Cox 半参数比例风险模型（Cox，1972）。

如果定义协变量为：$X=(x_{j1},\cdots,x_{jp})^t$，$h(t|X)$是$t$时点影响因素为$X$的风险函数，Cox模型的基本形式为：

$$h(t|X)=h_0(t)c(\beta^tX) \tag{3}$$

式(3)中，$h_0(t)$为基准风险函数，$\beta=(\beta_1,\cdots,\beta_p)^t$是协变量的系数向量，$c(\beta^tX)$是协变量的参数模型。由于$h(t|X)$是正数，$c(\beta^tX)$一般采用指数形式：$c(\beta^tX)=\exp(\beta^tX)=\exp\left(\sum_{k=1}^{p}\beta_kX_k\right)$。因此，$h(t|X)=h_0(t)\exp(\beta^tX)=h_0(t)\exp\left(\sum_{k=1}^{p}\beta_kX_k\right)$。

如果两个样本的协变量向量分别为X和X^*，那么其风险比为：

$$\frac{h(t|X)}{h(t|X^*)}=\frac{h_0(t)\exp\left(\sum_{k=1}^{p}\beta_kX_k\right)}{h_0(t)\exp\left(\sum_{k=1}^{p}\beta_kX_k^*\right)}=\exp\left[\sum_{k=1}^{p}\beta_k(X_k-X_k^*)\right] \tag{4}$$

式(4)的风险比中，在分子、分母中都有基准风险率，所以被消去。由于基准风险率刻画了风险率变化的时间模式，被消去后，上述风险比成为不随时间变化的常数，协变量的影响只是成比例提高或降低整个的基准风险率，所以Cox模型也称比例模型，风险比也称相对风险，相对于参照类(如性别变量，可以把其中一个设为参照类)，具有某种特征的个体面临的相对风险。如果协变量是单变量，且取值为0-1时，风险比简化为$\frac{h(t|X)}{h(t|X^*)}=\exp(\beta_1)$。

Cox模型一个重要假设是不同变量值对应的个体具有相同形式的基准风险率，所以需要进行等比例检验。目前检验的方法主要有Schoenfeld残差检验、Cox-Snell残差检验、Martingale残差检验、画图检验、时变变量检验5种(Klein等，2003；Xue等，2017)。前3种是正式的检验法，在统计软件中可以实现。画图检验主要是绘制分类变量的生存曲线或者累积风险函数图，如果各类别的生存曲线或者累积风险函数曲线不交叉，基本可以判定风险比满足等比例假设。如果样本不满足等比例假设，可以使用其他生存分析方法，如离散时间事件史模型。如果模型中有时变变量，可以尝试建立分层的Cox模型，然后比较分层和一般的Cox模型的拟合度，如果一般Cox模型的拟合好，说明风险比是等比例的，如果分层模型的拟合好，则说明风险比不是等比例的，需要拟合分层Cox模型。

离散时间事件史模型是常用的一种模型(Allison，1982)。在很多纵向追踪调查中，有些事件在某次追踪调查中发生了，但没有具体的发生时间，不能使用连续形式的生存分析模型。这种情况下，就可以使用离散时间事件史模型。该模型将生存分析的思路和常用的logistic模型结合起来，模型的基本形式与一般的Logistic模型一样，参数解释也

相同，还可以包含时变变量，而且这个模型不要求检验等比例假设，常用的统计软件都可以用于分析这个模型。

三、生存分析在应用中的常见问题

本研究在 Web of Science SSCI 数据库的 Demography 和 Sociology 学科下使用“Survival Analysis”和“Event History Analysis”检索到 1 123 篇文章，使用中国知网 CSSCI 数据库在人口学和社会学学科下使用“生存分析”和“事件史分析”检索到 36 篇文章[①]。从研究内容来看，婚姻、生育、死亡、迁移、教育、就业、职位晋升及退休、贫困、性行为和组织行为等都有涉及。这些文献中常见的应用误区主要有以下 5 点。

（一）仅考虑生存数据中的右删截特征，忽略其他数据特征

文献中使用的数据均是不完全数据，涉及的删截和截平的组合有很多种。右删截几乎在每篇文献中都存在，是生存分析研究中最常见的数据特征。左截平与右删截是最常见的组合，但大部分使用者没有明确指出这种组合，或者只提到右删截。还有其他多种组合，如右截平与右删截组合，左、右截平和右删截组合等。使用者大都知道生存分析能够处理不完全数据，但他们认为不完全数据仅指右删截数据，很少考虑截平等特征。很多使用者没有提到任何生存数据特征。

使用纵向追踪数据研究事件史，如死亡，很多数据不是从出生就开始观测，而是从某一个特定的年龄开始观测。例如，CLHLS 数据调查 65 岁及以上的老年人，这些老人推迟了至少 65 年才进入观测，那么该数据就具有左截平的特征，到调查结束时，还有部分老人存活，存活老人的数据又具有右删截的特点。在使用中国老年健康影响因素跟踪调查（CLHLS）数据或类似数据研究死亡风险的文献中，大部分指出右删截特征，仅有很少文章指出左截平特征（Molitoris，2017；李强、张震，2009）；也有一些文献完全没有指出任何数据特征，还有文献将左截平误认为是左删截（位秀平、吴瑞君，2015）。例如，研究离婚，如果纵向追踪数据不是从结婚开始，而是从某一个时点开始，如 Panel Study of Income Dynamics（PSID），很多样本在调查开始前已经结婚，这部分样本就具有左截平的特征，到调查结束时，很多样本没有离婚，这部分样本又具有右删截的特点（Guo，1993）。在使用这个数据或其他类似数据研究离婚风险的分析中，大部分都忽略了左截平特征，仅提到右删截，或者完全忽略数据特征（Oppenheimer，2003；Lopoo 等，2005）。

如果使用回顾性数据研究事件发生史，如辍学，不同的调查设计会呈现不同的数据特征。例如，使用“中国家庭追踪调查”数据的成人问卷（16 岁及以上的人群）研究“撤点并校”对中小学生辍学的影响（潘光辉，2017）。“撤点并校”政策的发生和执行是在1995～

① 本研究没有包含生命表分析，主要关注现代统计分析中的生存分析方法。

2012 年，研究者只对这段时间内发生的辍学感兴趣，所以数据具有左截平和右截平的特点。如果有的样本没有经历过辍学，那么其数据具有右删截的特征。但研究者只考虑右删截，忽略左、右截平。还有研究使用复学学生的数据研究辍学（常宝宁，2010），数据具有右截平的特征，因为只有在调查时点前复学的学生才能被观测到，调查之后复学的学生不在观测内。而且只研究 4～9 年级学生的辍学，这个数据又具有左截平的特征，因为辍学是从开始上学就经历的风险，不是从 4 年级才开始经历辍学风险。没有辍学的学生具有右删截的特征。也就是说，这个研究的生存数据具有左、右截平和右删截的特征，但研究者只考虑了右删截。

正确理解生存分析数据的特征是构造似然函数、进行统计推断和参数估计的前提，如果理解有误，会导致整个过程的偏差。截平对应的似然函数是条件概率函数，如左截平的似然函数是 $f(t_i)/S(Y_{Li})$，分子是概率密度函数，分母是生存函数。非截平数据对应的似然函数是 $f(t_i)$。这两种似然函数截然不同。似然函数是参数估计和模型拟合度估计的基础，错误的似然函数会增大参数估计偏差、降低拟合优度。例如，忽略左截平会低估事件发生风险，因为存活到进入观测的被访者一般会有较低的事件发生风险（Guo，1993）。忽略左截平对协变量系数的估计偏差在不同的事件研究中不同，如在离婚研究中会高估协变量的系数（Guo，1993），在死亡风险研究中则不会出现明显的偏差，但拟合优度较差（李强、张震，2009）。截平对模型的影响其实是样本选择性的问题，理解了这一点有助于评估忽略截平带来的估计误差。

（二）应用 Cox 模型忽略等比例假设检验

Cox 模型因其能对删截和截平数据在历险时间分布未知的情况下进行处理而成为应用最广泛的生存分析模型。但 Cox 模型假定任何两个样本的风险比是等比例的，使用模型需要检验等比例假设是否被满足。绝大部分应用 Cox 模型的文献没有进行等比例假设的检验。等比例假设没有被满足，会导致参数估计错误，模型拟合度差（Xue 等，2017）。在某些情况下，得益于 Cox 模型本身的灵活性和稳健性，违反等比例假设依然能够得到一个近似估算。

（三）将时变变量当做非时变变量分析

生存分析在动态分析方面的优势不仅体现在研究事件发生（因变量）的动态变化，而且还体现在能够在动态变化中找出自变量和因变量的关系（郭志刚，2001）。生存分析的因变量的动态变化是指不同个体的事件发生的不同时间刻画出事件发生的动态变化。自变量的动态变化是自变量在观测期内的不同时点上的观测值的变化，即时变变量。有了这类变量，研究者就可以更准确地在动态变化中找出自变量和因变量的关系。例如，在老年人的死亡风险分析中，老年人的健康状况就是时变变量，随着老年人的健康状况的变化，老年人的死亡风险也在变化。与时变变量相对应的是不随时间变化的变

量，即非时变变量，如性别和出生地就是非时变变量。

现有文献中使用回顾性数据的研究中包含时变变量的文献较少。有些研究中自变量的性质应该是时变变量，如影响流动人口生育行为的经济和社会因素（梁同贵，2016）。但是，这些研究在分析中并没有将这些变量处理为时变变量。原因可能是：（1）回顾性数据中这些变量仅有一个时点的数据，在分析中无法将这些变量处理成时变变量；（2）回顾性数据中这些变量也是回溯性的，有明确的变化时间，可以处理成时变变量，但是作者没有处理。

一般而言，如果在分析中将时变变量当作非时变变量处理，会造成该变量的相对风险的估计偏差，而且观测窗口越长，偏差越大（van Walraven 等，2004；Austin 等，2006）。例如，Austin 等（2006）通过模拟得出，当某种治疗的真实效应为 0（死亡风险比为 1），忽略这种治疗的时变特性会导致死亡风险比被低估 4%～27%，观测期越长，低估越大。因此，在数据条件允许的情况下，研究者要尽量充分利用数据信息，将时变变量的变化纳入模型。在数据条件不允许的情况下，研究者也要通过定性描述指出将时变变量当作非时变变量分析可能带来的偏差。如研究躯体功能对老年人的死亡风险的影响时，老年人的躯体功能是一个时变变量，会随着年龄的增长而衰退。如果在研究中只使用基期的躯体功能分析其对死亡风险的影响，也就是说将躯体功能作为非时变变量处理，可能会高估躯体功能较好的老年人的死亡风险，因为他们在死亡之前躯体功能可能已经转差。有时也可能低估躯体功能较差的老年人的死亡风险，如小部分躯体功能较差的老年人可能会好转，经过这种转化，其死亡风险可能会降低。由于老年人的躯体功能由好转差的概率远大于由差转好的概率，所以总体来讲，将躯体功能视为非时变变量会高估躯体功能的影响效度。观测窗口越长，老年人的躯体功能转换的概率越大，高估的偏差也会越大。

（四）死亡风险模型中错将观测期视为历险时间

生存分析中的一个核心概念是历险时间。这里的时间是相对于事件而言的。例如，如果“死亡”是所关注的事件，那么从出生开始面临死亡风险的时间即年龄是历险时间；如果“生育二孩”是所关注的事件，那么历险时间就是从母亲给第一个孩子断奶、可能受孕开始计算的时间区间；离婚的历险时间则应从结婚开始计算。

历险时间的概念看似简单，但在实际研究中往往会出现混淆。在现有的文献中，如果研究事件不是死亡，大部分的文献对历险时间的理解是正确的。当研究事件是死亡时，对历险时间的理解就容易出现错误。死亡风险的历险时间是年龄，很多文献却使用观测期作为历险时间，基期的年龄作为协变量。中英文文献均出现了这种错误的建模。

以老年死亡风险研究为例，这类研究大多基于左截平和右删截的调查数据。使用者

会把死亡历险时间错误地设定为从老年人被调查开始到死亡(事件发生时间)或调查结束时间(即右删截)。事实上,无论年龄大小、无论何时被纳入调查,任何被访者从出生时就面临死亡风险,历险时间应该从出生开始计算,即被访者的年龄。以年龄作为历险时间的基准风险函数就是年龄别死亡率。似然函数的形式是$f(t)/S(Y_L)$,是条件概率,作为条件的分母是生存函数。当以观测期为死亡风险的历险时间时,每个个体进入观测的时间就是历险开始的时间,设为 0,似然函数的形式是$f(t)$。似然函数是估计参数和似然值的基础,只有正确的似然函数,才能估计出正确的参数值和似然值。

已经有文献致力于讨论死亡风险的历险时间,这些研究指出,由于死亡风险服从指数形式的分布,将观测期作为历险时间的模型估算的自变量系数的偏差实际上很小,但是这样的建模在理论上、逻辑上、似然函数的构建方面是不合理的,模型对数据的拟合优度也比较差(Korn 等,1997;Thiébaut 等,2004;李强、张震,2009)。

(五)中文文献中的专有名词翻译不一致,容易造成误解

由表可知,生存分析的专有名词译法多样,容易造成误解。Censoring 和 Truncation 是两个关键的专有名词。Censoring 有 7 种译法,Truncation 仅在一篇文献中提到, 翻译为"截平"(李强、张震,2009)。我们从其他文献中发现的 Truncation 的翻译有"截尾"和"截断"。"截尾"和"截断"的翻译和 Censoring 的翻译重复了,也反映了概念理解的混乱。

表　生存分析专有名词翻译

英语原词	中文翻译						
Censoring	删截	删失	截尾	删节	截删	截取	截断
Truncation	截平	截尾	截断				
Kaplan–Meier estimator	Kaplan–Meier 法		乘积极限法		卡普兰迈耶法		
Cox Proportional Hazards Model	Cox 比例风险模型		持续时间模型		Cox 等比例风险模型		
Hazards	风险率		危险率				

Kaplan–Meier estimator 通常不翻译直接引用如"Kaplan–Meier 法",音译"卡普兰迈耶法"和意译"乘积极限法"各出现 1 次。绝大部分研究者将 Cox Proportional Hazards Model 译为"Cox 比例风险模型",此外,还有"持续时间模型"、"Cox 等比例风险模型"。Hazards 通常被译为"风险率"或"危险率"。

四、结　语

本文简要介绍了生存分析方法,系统地梳理了删截和截平这两类不完全数据类型,指出了一些文章在应用生存分析时存在的误区。总的来看,目前生存分析的应用存在的主要问题包括以下几点:

第一,许多研究考虑到了不完全数据中的右删截特征,但却忽略截平等特征,导致

建模错误，构造似然函数时忽略了条件概率，从而高估似然值，模型的拟合优度差，参数估计也出现偏误。右删截比较容易识别，而截平等特征识别不易，如果使用者对起始时间、历险时间和观测时间等重要概念的理解和把握不够准确和深刻，很难正确识别截平等特征。识别数据特征的较好的方法是画出历险时间、起始时间和观测时间，然后绘出各类型样本的事件史历程，这样数据的特征就一目了然。

第二，忽视 Cox 比例风险模型的等比例检验。大部分应用 Cox 模型的文献中没有做等比例检验。本文在模型介绍中列出了目前常用的检验等比例假设的方法，这些方法在目前流行的统计软件中都可以实现。当不满足等比例假设时，可以根据实际情况使用其他模型，如离散时间事件史分析或分层 Cox 模型。

第三，自变量中的时变变量被当作非时变变量分析，这可能是数据条件不允许，也可能是使用者忽略时变变量。这种误用会导致估计偏差，而且观测时间越长，偏差越大（van Walraven 等，2004；Austin 等，2006）。

第四，死亡风险分析中错将观测期作为历险时间，实际上，年龄才是死亡风险的历险时间。由于死亡风险服从指数形式的分布，所以以观测期为历险时间不会导致自变量系数的估计偏差，但这样的建模在理论上、逻辑上、似然函数的构造上均不合理，也会降低模型拟合数据的拟合度（Korn 等，1997；Thiebaut 等，2004；李强、张震，2009）。

定量分析方法在人口学和社会学的应用越来越广泛，用户友好的统计软件使定量分析方法的应用更加便利，同时也降低了相关统计知识的门槛，这反而增加了定量分析方法不规范、不正确使用的风险。本文对应用生存分析的文献考察发现，在数据允许的情况下，研究者已经有意识地使用更先进和更合适的方法分析数据，研究驱动是应用生存分析的主要原因。然而，生存分析的应用还不够规范和正确。本文发现的 4 个问题是目前生存分析应用中比较突出的问题，希望通过这样的分析能引起使用者、评阅人和读者的关注，推动生存分析的规范和正确应用。

参考文献：

1. 常宝宁（2010）：《免费政策实施后儿童辍学问题实证研究——基于 COX 比例风险模型的分析》，《青年研究》，第 6 期。
2. 郭志刚（2001）：《历时研究与事件史分析》，《中国人口科学》，第 1 期。
3. 李强、张震（2009）：《生存分析中时间变量的选择》，《中国人口科学》，第 6 期。
4. 梁同贵（2016）：《乡城流动人口的生育间隔及其影响因素——以上海市为例》，《人口与经济》，第 5 期。
5. 潘光辉（2017）：《“撤点并校”、家庭背景与入学机会》，《社会》，第 3 期。
6. 彭非、王伟（2004）：《生存分析》，中国人民大学出版社。
7. 位秀平、吴瑞君（2015）：《中国老年人的躯体功能对死亡风险的影响》，《人口与经济》，第 2 期。

8. Allison, P.D.(1982), Discrete-time Methods for the Analysis of Event Histories. *Sociological Methodology*. 13: 61-98.

9. Austin, P.C., Mamdani, M.M., Van Walraven, C., & Tu, J.V.(2006), Quantifying the Impact of Survivor Treatment Bias in Observational Studies. *Journal of Evaluation in Clinical Practice*. 12(6): 601-612.

10. Cox, D.R.(1972), Regression Models and Life Tables. *Journal of Royal Statistical Society*. 34(2): 187-202.

11. Gampe, J.(2010), Human Mortality Beyond Age 110. In Maier, H. et al.(eds) *Supercentenarians*. Springer, Berlin, Heidelberg.

12. Guo, G.(1993), Event-history Analysis for Left-truncated Data. *Sociological Methodology*. 23: 217-243.

13. Guo, S.(2010), *Survival Analysis*. Oxford University Press.

14. Klein, P. J., & Moeschberger, M.L.(2003), *Survival Analysis: Techniques for Censored and Truncated Data* (second edition). Springer-Verlag New York. Inc.

15. Korn, E.L., Graubard, B. I., & Midthune, D.(1997), Time-to-event Analysis of Longitudinal Follow-up of a Survey: Choice of the Time-scale. *American Journal of Epidemiology*. 145(1): 72-80.

16. Long, J.S., Allison, P.D., & McGinnis, R.(1993), Rank Advancement in Academic Careers: Sex Differences and the Effects of Productivity. *American Sociological Review*. 58(5): 703-722.

17. Lopoo, L. M., & Western, B.(2005), Incarceration and the Formation and Stability of Marital Unions. *Journal of Marriage and Family*. 67(3): 721-734.

18. Massey, D.S., & Espinosa, K.E.(1997), What's Driving Mexico-US Migration? A Theoretical, Empirical, and Policy Analysis. *American Journal of Sociology*. 102(4): 939-999.

19. Molitoris, J.(2017), Disparities in Death: Inequality in Cause-specific Infant and Child Mortality in Stockholm, 1878-1926. *Demographic Research*. 36: 455-500.

20. Oppenheimer, V.K.(2003), Cohabiting and Marriage During Young Men's Career-development Process. *Demography*. 40(1): 127-149.

21. Sweeney, M.M.(2002), Two Decades of Family Change: The Shifting Economic Foundations of Marriage. *American Sociological Review*. 67(1): 132-147.

22. Thiébaut, A.C., & Bénichou, J.(2004), Choice of Time-scale in Cox's Model Analysis of Epidemiologic Cohort Data: A Simulation Study. *Statistics in Medicine*. 23(24): 3803-3820.

23. van Walraven, C., Davis, D., Forster, A.J., & Wells, G.A.(2004), Time-Dependent Bias was Common in Survival Analyses Published in Leading Clinical Journals. *Journal of Clinical Epidemiology*. 57(7): 672-682.

24. Whitbeck, L.B., Yoder, K.A., Hoyt, D.R., & Conger, R.D.(1999), Early Adolescent Sexual Activity: A Developmental Study. *Journal of Marriage and Family*. 61(4): 934-946.

25. Xue, Y., & Schifano, E.D.(2017), Diagnostics for the Cox Model. *Communications for Statistical Applications and Methods*. 24(6): 583-604.

(责任编辑:李玉柱)