生存资料回归模型分析——生存资料参数回归模型分析基础

刘红伟[1],张甜甜[1],刘媛媛[1],李长平[1,2],胡良平[2,3*]

(1. 天津医科大学公共卫生学院,天津　300070;
2. 世界中医药学会联合会临床科研统计学专业委员会,北京　100029;
3. 军事科学院研究生院,北京　100850
*通信作者:胡良平,E-mail:lphu927@163.com)

【摘要】 本文目的是介绍生存资料参数回归模型有关的基础知识。首先,介绍了构建三个常见的生存资料参数回归模型的基本原理,包括指数分布回归模型、Weibull分布回归模型和Log-logistic分布回归模型;其次,介绍了基于图示法判断生存时间服从何种概率分布的方法;最后,介绍了基于最大似然估计法求解参数回归模型中的参数和两个参数回归模型拟合优度的比较。得到如下结论:①当资料中的生存时间服从特定概率分布时,应选用相应的参数回归模型;②图示法可用于粗略判断生存时间服从何种概率分布;③似然比检验可用于包含不同参数数目的两个参数回归模型之间拟合优度的比较。

【关键词】 生存分析;参数回归;拟合优度检验;似然比检验;Weibull分布;Log-logistic分布

中图分类号:R195.1　　文献标识码:A　　doi:10.11886/scjsws20200106004

Regression analysis of the parametric models for the survival data——the basis of the parametric regression model analysis of the survival data

Liu Hongwei[1], *Zhang Tiantian*[1], *Liu Yuanyuan*[1], *Li Changping*[1,2], *Hu Liangping*[2,3*]

(1. *School of Public Health*, *Tianjin Medical University*, *Tianjin*　300070, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies*, *Beijing* 100029, *China*;
3. *Graduate School*, *Academy of Military Sciences PLA China*, *Beijing*　100850, *China*
**Corresponding author*: Hu Liangping, E-mail: lphu927@163.com)

【**Abstract**】 The purpose of this paper was to introduce the fundamental knowledge of the parametric models for the survival data. Firstly, the basic principles of three common parametric models for the survival data were given, including exponential distribution regression model, Weibull distribution regression model and Log-logistic distribution regression model. Secondly, the method of judging the probability distribution of survival time based on graphs was introduced. Finally, the solution for the parameters in the parametric regression model based on maximum likelihood estimation and the comparison between the goodness of fit of the two parametric regression models was introduced. The results of the article showed that: ① when the data followed the specific distribution, the corresponding parametric regression model should be applied; ② graphical methods could be use to judge rouphly the survival time followed what kind of the probability distribution; ③ likelihood ratio test could be used to compare the effects of the goodness of fit between two parametric regression models with different numbers of parameters.

【**Keywords**】 Survival analysis; Parametric regression; Goodness-of-fit tests; Likelihood-ratio tests; Weibull distribution; Log-logistic distribution

在分析多个因素对生存时间的影响时,人们通常希望像一般的回归分析一样,能建立生存时间(因变量)随危险因素(自变量或协变量)变化的回归方程,以便对危险因素的作用大小有一个全面的了解和掌握,并根据危险因素的不同取值对生存率(或危险率)进行预测。能实现此目的的生存分析方法有Cox模型回归分析和参数模型回归分析。当生存时间的准确分布无法获得时,可采用Cox模型回归分析[1],此模型在形式上与参数模型相似,但对模型中各参数进行估计时不依赖于特定分布的假设,所以又称半参数模型。当然,在可以通过图示法或统计检验法得到待分析的生存时间服从某特定分布的参数模型时,如指数分布回归模型或Weibull分布回归模型,可采用生存资料的参数模型

基金项目:国家自然科学基金项目(项目名称:贝叶斯生存分析方法在肝细胞癌肝移植患者预后预测中的应用研究,项目编号:81803333)

回归分析直接拟合之，所得结果将更加准确[2]。

1 生存资料参数回归模型概述

生存资料参数回归模型分析的一个重要内容是模型拟合或分布拟合。描述生存时间分布的模型通常有指数分布、Weibull分布、Log-logistic分布、对数正态分布、广义Gamma分布模型等。在生存分析研究中，常用概率密度函数$f(t)$、生存函数$S(t)$和风险函数（或称危险率函数）$h(t)$来描述生存过程，这三种函数在数学推导上是等价的[3]，如果给定其中一种函数，另外两种函数即可推导得出，它们的关系如下：

(1) 若 $f(t)$ 已知，则 $S(t)=\int_t^{\infty} f(u)du$，$h(t)=\frac{-d[S(t)]/dt}{S(t)}$；

(2) 若 $S(t)$ 已知，则 $h(t)=\frac{-d[S(t)]/dt}{S(t)}$，$f(t)=\frac{-d[S(t)]}{dt}$；

(3) 若 $h(t)$ 已知，则 $S(t)=\exp(-\int_0^t h(u)du)$，$f(t)=h(t)S(t)$。

对实际的生存数据进行分布拟合时，可用上述模型分别进行拟合，根据拟合优度检验的结果选择适当的模型。有时，对于一批生存数据，事先不知道生存时间的确切分布，也难以判断何种模型最合适，许多研究者一般直接采用非参数或半参数回归模型。但是，如果已知一批数据确实符合某特定的参数回归模型，由于非参数或半参数方法的精度一般低于参数方法，此时，宜选用相应的参数回归模型。由于篇幅所限，本文主要介绍指数分布回归模型、Weibull分布回归模型和Log-logistic分布回归模型。

2 常用的三种参数回归模型介绍

2.1 指数分布回归模型

2.1.1 仅以生存时间t为自变量的指数分布回归模型

指数分布回归模型是最简单的生存资料参数回归模型，在任何时间点上的风险函数为一常数，风险函数的大小不受生存时间长短的影响，即“无记忆性”。设生存时间服从指数分布，则生存时间变量的概率密度函数为：

$$f(t)=\begin{cases}\lambda e^{-\lambda t} & (t\geqslant 0,\lambda>0)\\ 0 & (t<0)\end{cases} \tag{1}$$

分布函数为：

$$F(t)=1-e^{-\lambda t} \quad (t\geqslant 0) \tag{2}$$

生存函数为：

$$S(t)=1-F(t)=e^{-\lambda t} \tag{3}$$

风险函数为：

$$h(t)=f(t)/S(t)=\lambda \quad (\lambda>0,t\geqslant 0) \tag{4}$$

λ为常数，与时间无关，代表指数分布回归模型的风险率，决定了生存率下降的快慢。风险率越大，生存率下降越快，生存时间越短；风险率越小，生存率下降越慢，生存时间越长。

指数分布回归模型自变量既满足比例风险（Proportional Hazard，PH）假设，也满足加速失效时间（Accelerated Failure Time，AFT）假设。PH假设要求一个人的风险与任何其他人的风险成正比，且比例为一个常数，与时间无关；AFT假设要求对于任一固定生存概率，不同个体间生存时间比值为一个常数，这个常数称为加速因子（Accelerated factor）。因此PH假设中预测变量（即自变量或影响因素）对个体发生风险的影响成比例，而AFT假设中预测变量对个体生存时间的影响成比例[4]。

2.1.2 基于生存时间t添加其他自变量的指数分布回归模型

在上面的四个模型表达式中，生存时间t为自变量，其因变量分别为f(t)、F(t)、S(t)和h(t)，它们分别为密度函数、分布函数、生存函数和风险函数。但是，在处理实际的生存资料时，研究者希望考察除时间t之外的其他许多自变量或协变量对前述提及的四个因变量的影响，于是，统计学家将模型中的“重要参数”视为除时间t之外的其他许多自变量或协变量的函数（基于数学上处理方便角度考量，选取“指数函数”形式）。这样，就建立起因变量依赖包括生存时间t在内并包含其他自变量的回归模型。为简便起见，下面的例子中只包含了一个叫做“TRT”的“新自变量”。

将指数分布回归模型应用到42例白血病患者数据中[4]，其中21例患者接受了治疗，另外21例患者使用了安慰剂。结局为白血病患者的生存时间，预测变量（即自变量）是TRT，取值(0,1)，1代表接受了治疗，0代表未接受治疗。

基于PH假设，包含预测变量的指数分布回归

模型表达式(此处特指风险函数)为:

$$h(t)=\lambda=\frac{1}{\exp(\beta_0+\beta_1 TRT)} \tag{5}$$

其中 $h(t)$ 为个体风险大小,TRT 表示是否接受治疗,则治疗组相对于非治疗组的风险比(Hazard ratio)为:

$$HR(TRT=1\,vs\,TRT=0)=\frac{1/\exp(\beta_0+\beta_1)}{1/\exp(\beta_0)}$$

$$=\frac{1}{\exp(\beta_1)} \tag{6}$$

若 $\beta_1>0$,则 $\frac{1}{\exp(\beta_1)}<1$,相对于非治疗组,治疗组发生死亡的风险更小;若 $\beta_1<0$,则 $\frac{1}{\exp(\beta_1)}>1$,相对于非治疗组,治疗组发生死亡的风险更大。

基于AFT假设,包含预测变量的指数分布回归模型表达式为:

$$t=[-\ln S(t)]\times\exp(\alpha_0+\alpha_1 TRT) \tag{7}$$

上式中,t 代表个体生存时间,$S(t)$ 代表个体生存函数。对任一固定生存概率 $S(t)=q$,治疗组相对于非治疗组的加速因子(Acceleration factor)γ 为:

$$\gamma(TRT=1\,vs\,TRT=0)=\frac{[-\ln(q)]\times\exp(\alpha_0+\alpha_1)}{[-\ln(q)]\times\exp(\alpha_0)}$$

$$=\exp(\alpha_1) \tag{8}$$

若 $\alpha_1>0$,则 $\exp(\alpha_1)>1$,表明相对于安慰剂,治疗对于生存时间有正向促进作用,延长了患者的生存时间;若 $\alpha_1<0$,则 $\exp(\alpha_1)<1$,表明相对于安慰剂,治疗对于生存时间有反向抑制作用,缩短了患者的生存时间。

2.2 Weibull分布回归模型

2.2.1 仅以生存时间t为自变量的Weibull分布回归模型

Weibull分布回归模型是最广泛使用的参数回归模型,由瑞典科学家 Waloddi Weibull 提出。Weibull分布是指数分布的一种推广形式,应用更广泛。设生存时间服从Weibull分布,则对应的概率密度函数为:

$$f(t)=\begin{cases}\lambda\gamma(\lambda t)^{\gamma-1}\exp[-(\lambda t)^{\gamma}] & (t\geqslant 0,\gamma>0,\lambda>0)\\ 0 & (t<0)\end{cases} \tag{9}$$

生存函数为:

$$S(t)=\exp[-(\lambda t)^{\gamma}] \tag{10}$$

风险函数为:

$$h(t)=\lambda\gamma(\lambda t)^{\gamma-1} \tag{11}$$

λ 是尺度参数,γ 是形状参数,决定函数图像的形状。若 $\gamma>1$,风险随着时间的增加而增加;若 $\gamma=1$,风险为常数,则Weibull分布变为指数分布;若 $\gamma<1$,风险随着时间的增加而减小。

2.2.2 基于生存时间t添加其他自变量的Weibull分布回归模型

基于PH假设,包含预测变量的Weibull分布回归模型表达式(此处特指风险函数)为:

$$h(t)=\frac{\gamma t^{\gamma-1}}{\exp[\gamma(\beta_0+\beta_1 TRT)]} \tag{12}$$

基于AFT假设,包含预测变量的Weibull分布回归模型表达式为:

$$t=[-\ln S(t)]^{1/\gamma}\times\exp(\alpha_0+\alpha_1 TRT) \tag{13}$$

风险比(Hazard ratio)和加速因子(Acceleration factor)的求解和解释同指数分布类似,此处不再赘述。

2.3 Log-logistic分布回归模型

2.3.1 仅以生存时间t为自变量的Log-logistic分布回归模型

在生存分析中,Log-logistic分布用于描述事件的发生率,例如诊断或治疗后的肿瘤患者的死亡率。设生存时间服从Log-logistic分布,则对应的概率密度函数为:

$$f(t)=\begin{cases}\dfrac{\lambda\gamma(\lambda t)^{\gamma-1}}{[1+(\lambda t)^{\gamma}]^2} & (t\geqslant 0,\gamma>0,\lambda>0)\\ 0 & (t<0)\end{cases} \tag{14}$$

生存函数为:

$$S(t)=\frac{1}{1+(\lambda t)^{\gamma}} \tag{15}$$

风险函数为:

$$h(t)=\frac{\lambda\gamma(\lambda t)^{\gamma-1}}{1+(\lambda t)^{\gamma}} \tag{16}$$

γ 是形状参数,当 $\gamma\leqslant 1$ 时,风险值随着时间增加而下降;若 $\gamma>1$,则风险值先增加后减小,风险函数图像为"单峰"。Log-logistic分布回归模型要求自变量满足AFT假设,不满足PH假设,但满足比例优比(Proportional odds,PO)假设,PO假设要求生存优势比(Survival odds ratio)随着时间的变化保持不变,即生存优势比为常数,其中生存优势比

(Survival odds ratio, SOR)定义为两个个体的生存比值的比值,表达式为:

$$SOR = \frac{S_1(t)/[1 - S_1(t)]}{S_2(t)/[1 - S_2(t)]} \quad (17)$$

其中,$S_1(t)$是个体1的生存函数,$S_2(t)$是个体2的生存函数,若SOR和时间无关,则Log-logistic分布回归模型自变量满足PO假设。

2.3.2 基于生存时间t添加其他自变量的Log-logistic分布回归模型

以42例白血病患者数据中的白细胞计数变量(white blood cell count, WBCCAT)为例,WBCCAT变量取值1和2,1代表中位数,2代表最大值。基于PO假设,包含预测变量的Log-logistic分布回归模型表达式为:

$$\frac{1-S(t)}{S(t)}=(\lambda t)^{\gamma}=\frac{t^{\gamma}}{\exp[\gamma(\beta_0+\beta_1 WBCCAT)]} \quad (18)$$

其中,变量WBCCAT的失效比为:

$$OR(WBCCAT=2\ vs\ WBCCAT=1)$$
$$=\frac{t^{\gamma}/\exp[\gamma(\beta_0+2\beta_1)]}{t^{\gamma}/\exp[\gamma(\beta_0+1\beta_1)]}=\frac{1}{\exp(\gamma\beta_1)} \quad (19)$$

若$\beta_1>0$,则WBCCAT取值1的结局风险更大。基于AFT假设,包含预测变量的Log-logistic分布回归模型表达式为:

$$t=[\frac{1}{S(t)}-1]^{1/\gamma}\times\exp(\alpha_0+\alpha_1 TRT) \quad (20)$$

加速因子的求解与解释与指数分布回归模型类似,此处不再赘述。

2.4 其他概率分布下参数回归模型简介

2.4.1 对数正态分布回归模型

对数正态分布参数回归模型定义为时间变量的对数遵从正态分布,其概率密度函数为:

$$f(t)=\frac{1}{t\sigma\sqrt{2\pi}}\exp[-\frac{1}{2\sigma^2}(\ln t-\mu)^2]\ (t>0,\sigma>0) \quad (21)$$

生存函数为:

$$S(t)=1-\Phi[\frac{\ln t-\mu}{\sigma}] \quad (22)$$

风险函数为:

$$h(t)=\frac{\frac{1}{t\sigma\sqrt{2\pi}}\exp[-(\frac{\ln at}{\sigma\sqrt{2}})^2]}{1-\Phi(\frac{\ln at}{\sigma})} \quad (23)$$

其中,$a=\exp(-\mu)$,$\Phi(x)$为标准正态分布函数。对数正态分布的形状与Log-logistic分布形状接近,不同的是对数正态分布模型要求自变量满足AFT假设,但不满足PO假设。

2.4.2 Gompertz分布回归模型

Gompertz分布回归模型定义为生存时间服从Gompertz分布,其概率密度函数为:

$$f(t)=\exp[(\lambda+\gamma t)-\frac{1}{\gamma}(e^{\lambda+\gamma t}-e^{\lambda})]\ (t>0) \quad (24)$$

生存函数为:

$$S(t)=\exp[-\frac{e^{\lambda}}{\gamma}(e^{\gamma t}-1)] \quad (25)$$

风险函数为:

$$h(t)=\exp(\lambda+\gamma t) \quad (26)$$

Gompertz模型自变量不满足AFT假设,但回归模型和Cox模型相似。

2.4.3 广义Gamma分布回归模型

广义Gamma分布回归模型有三个参数,形状有更大的灵活性。设生存时间服从广义Gamma分布,则对应的概率密度函数为:

$$f(t)=\begin{cases}\frac{\gamma\lambda^{k}t^{\gamma k-1}\exp(-\lambda t^{\gamma})}{\Gamma(k)} & (t\geq 0,\gamma>0,\lambda>0,k>0)\\ 0 & (t<0)\end{cases} \quad (27)$$

生存函数为:

$$S(t)=1-I[\lambda t^{\gamma},k] \quad (28)$$

风险函数为:

$$h(t)=\frac{f(t)}{S(t)} \quad (29)$$

其中,$I(s,\gamma)=\frac{1}{\Gamma(\gamma)}\int_0^s x^{\gamma-1}e^{-x}dx$称为不完全Gamma函数,因为风险函数形式复杂,这里用关系式表示。当$\gamma=1$时,分布是指数分布,当$\lambda=\frac{1}{2}$和$\gamma=\frac{1}{2}\nu$,分布是v个自由度的χ^2分布,此处v是整数。

3 基于图示法判断生存时间的概率分布

3.1 基于图示法判断生存时间服从指数分布的方法

指数分布回归模型根据$S(t)=e^{-\lambda t}$可以得出$\log[S(t)]=-\lambda t$,可绘制出$\log[S(t)]$对t的*Kaplan-Meier*图,若图像是经过原点的一条直线,表明数据符合指数分布回归模型,斜率是$-\lambda$的估计值。

3.2 基于图示法判断生存时间服从Weibull分布的方法

Weibull分布回归模型根据$S(t)=\exp[-(\lambda t)^{\gamma}]$，可以得到$\ln[-\ln S(t)]=\gamma\ln\lambda+\gamma\ln t$，即$\ln[-\ln S(t)]$和$\ln t$是线性关系，斜率为$\gamma$。可绘制出$\ln[-\ln S(t)]$对$\ln t$的*Kaplan－Meier*图，若图像是一条直线，表明数据符合Weibull分布回归模型，斜率为γ，截距为$\gamma\ln\lambda$。

3.3 基于图示法判断生存时间服从Log-logistic分布的方法

在Log-logistic分布回归模型中，失效比值（Failure odds）是生存比值（Survival odds）的倒数，失效比值为$\dfrac{1-S(t)}{S(t)}=\dfrac{\dfrac{(\lambda t)^{\gamma}}{1+(\lambda t)^{\gamma}}}{\dfrac{1}{1+(\lambda t)^{\gamma}}}=(\lambda t)^{\gamma}$，对失效比值取ln得到：

$$\ln(failure\ odds)=\ln[(\lambda t)^{\gamma}]=\gamma\ln\lambda+\gamma\ln t \qquad (30)$$

即失效比值的对数和$\ln t$是线性关系，斜率为γ，截距为$\gamma\ln\lambda$。绘制$\ln[\dfrac{1-S(t)}{S(t)}]$关于$\ln t$的*Kaplan－Meier*图，若图像是一条直线，表明数据适合Log-logistic分布回归模型。

4 基于最大似然估计法求解参数回归模型中的参数

参数回归模型中回归系数可以通过求极大似然函数最大值的方法得到。参数回归模型的似然函数是观测数据和未知参数的一个函数，等于每个个体似然值相乘，其中似然函数的形式和结局变量的概率密度函数有关。在生存分析中似然函数和普通的似然函数区别在于数据包含删失数据，删失数据类型一般包括左删失、右删失和区间删失。个体失效时间与似然值之间的关系见表1。

表1是不同个体的失效时间和对应的似然值，其中$f(t)$是个体生存时间的概率密度函数。若该研究共有表1中的5人，则总的似然函数为：

$$L=f(2)\times\int_8^{\infty}f(t)dt\times f(6)\times\int_0^2 f(t)dt\times\int_4^8 f(t)dt \qquad (31)$$

未知参数的解可以通过最大化似然函数得到，最大化似然函数的过程通常是将$\ln(L)$对各参数求偏导数，并令偏导数为零，从而获得所谓的“正规方程组”，然后求解此方程组便可获得各参数的估计值，即：

表1 个体失效时间和似然值

个体ID	时　间	似然值
1	t=2	$f(2)$
2	t>8（右删失）	$\int_8^{\infty}f(t)dt$
3	t=6	$f(6)$
4	t<8（左删失）	$\int_0^2 f(t)dt$
5	4<t<8（区间删失）	$\int_4^8 f(t)dt$

$$\frac{\partial Ln(L)}{\partial\beta_j}=0 \quad (j=1,2,...,N) \qquad (32)$$

5 两个参数回归模型拟合优度的比较

对同一个生存资料拟合了两个包含参数数目不同的参数回归模型后，需要比较它们之中哪一个更好，称为“拟合优度检验”，通常可采用“似然比检验”。比较嵌套模型之间的拟合效果可以采用似然比检验[5]，似然比统计量的公式为：

$$\chi_v^2=-2\log L_q-(-2\log L_{q+v}) \qquad (33)$$

式中χ_v^2服从自由度为v的χ^2分布，$-2\log L_q$和$-2\log L_{q+v}$分别为含有q和$q+v$个参数的回归模型的对数似然函数值。

一般来说，一个回归模型对应的“-2log(L)”的数值间接反映了模型对生存资料的拟合效果。当有两个回归模型时，若所含参数数目相同，此值越小越好；若所含参数数目不相同，含参数数目多的回归模型的“-2log(L)”的数值必须明显小于含参数数目少的回归模型的“-2log(L)”的数值（即上述检验结果为$P<0.05$），则应选取含参数数目多的回归模型。否则，应选择含参数数目少的回归模型。

6 讨论与小结

6.1 讨论

在生存资料参数回归模型中，图示法帮助我们选择合适的概率分布，拟合优度检验帮助我们确定嵌套模型中的最优模型，两种方法结合提供了一个有效的模型选择方法。

相对于非参数和半参数回归模型而言，参数回归模型的结果精确度要高一些，但是，目前暂无非常精准的方法判定待分析的生存资料中的生存时间究竟服从何种概率分布，这可能是生存资料参数

回归模型在实际使用中比较受限的根本原因。

6.2 小结

本文比较详细地介绍了三种常见的概率分布回归模型的构建、求解和拟合优度的比较方法；扼要地介绍了其他几种并非常用但很重要的概率分布回归模型。在实际应用过程中，应首先采用图示法判断资料中的生存时间是否符合特定的概率分布，然后拟合对应的参数回归模型，采用最大似然法求解参数，通过拟合优度的比较，最后选择出最优的参数回归模型。

参考文献

[1] George B，Seals S，Aban I. Survival analysis and regression models[J]. J Nucl Cardiol，2014，21(4)：686-694.
[2] 胡良平. SAS常用统计分析教程[M]. 2版. 北京：电子工业出版社，2015：533-537.
[3] 陈兵，骆福添. 生存分析中的回归模型[J]. 中国卫生统计，2006，23(4)：462-465.
[4] Kleinbaum DG，Klein M. Survival analysis：a self-learning text[M]. 3rd Edition. New York：Springer Science Business Media，2012：289-351.
[5] 王建文. 生存分析参数回归模型拟合及其SAS实现[D]. 太原：山西医科大学，2008.

（收稿日期：2020-01-06）
（本文编辑：陈 霞）