中华流行病学杂志1994年8月第15卷第4期

相对危险度回归模型

项永兵　高玉堂　金　凡

流行病学研究资料分析中无论是队列研究、病例对照研究或是生存随访研究，数据分析的最后阶段都是配合一些多变量分析模型，如Logistic回归模型、Cox回归模型等。由这些多变量回归模型分析的结果，我们获得的一个重要的指标就是相对危险度。因此，近年来文献中多有“相对危险度回归模型”的提法〔1～11〕，并在此题目下讨论Logistic回归模型或Cox回归模型等多变量分析方法。笔者将就相对危险度回归模型方面的研究情况作一介绍。

问题的提出及模型的建立

用Z（u）表示研究个体在某一观察时间单位（u）里的协变量（covariate），则〔Z（u），0≤u<t〕表示研究个体在时刻t之前所具有的全部协变量特征。假定T为研究个体的失效（failure）时间，就理论上足够大或无限大的研究人群，研究个体的失效时间在时间轴上是一连续性变量。则研究人群的瞬时失效率（instantaneous failure rate）为：

$$\lambda[t;Z(u),\ 0\leqslant u<t]=\lim_{\Delta\to 0} pr[t\leqslant T<t+\Delta \mid Z(u),\ 0\leqslant u<t;T\geqslant t]\Delta^{-1}, \quad (1)$$

此式即我们通常所称的危险率（hazard rate）或死亡力（force of mortality）。相对危险度则是指具有协变量〔Z（u），0≤u<t〕的研究人群的危险率与其具有标准或基准协变量〔Z_0（u），0≤u<t〕状态时的危险率的比。通常情况下，标准或基准协变量〔Z o（u），0≤u<t〕总是选择暴露史阴性、治疗中的对照组等等。引入相对危险度概念后，(1)式又可表达为：

$$\lambda[t;Z(u),\ 0\leqslant u<t]=\lambda_o(t)RR[t;Z(u),\ 0\leqslant u<t] \quad (2)$$

式中$\lambda_0(t)=\lambda[t;Z_0(u),\ 0\leqslant u<t]$，RR（·）即本文所讨论的相对危险度。所以相对危险度回归模型就是通过配合RR（·）和λ（·）来描述研究人群中某种事件（如发病、死亡等）的发生，即失效与其所具有的协变量特征之间的关系。模型一般取下面的参数形式：

$$RR[t;Z(u),\ 0\leqslant u<t]=\gamma[X(t),\beta], \quad (3)$$

式中，λ（·）根据不同的调查研究及资料的性质取不同的函数形式，通常为exp（·）或1+（·），即危险或预后因素对危险率的作用取相乘或相加作用方式。在标准或基准协变量情形下，γ（0）=1。X（t）=〔X_1（t），…，Xp（t）〕是包含了协变量〔Z(u)，0≤u<t〕与危险率之间某种函数关系的回归向量。在协变量为〔Z_0（u），0≤u<t〕时，X（t）=0。β则是待估的回归系数列向量。分层分析是流行病学研究资料分析中常用的技术，相应地可以获得分层的相对危险度回归模型，即

$$RR_s[t;Z(u),\ 0\leqslant u<t]=\lambda[t;Z(u),\ 0\leqslant u<t]/\lambda_o(t)=\gamma[X(t),B_s], \quad (2)$$

式中$\lambda_{os}(t)=\lambda[Z_{os}(u),\ 0\leqslant u<t]$，$S\in\{0,\cdots\cdots,q\}$，〔$Z_{os}$(u)，0≤u<t〕为S层的基准协变量，q为层数。

模型（3）、（4）实际上是在Cox回归模型〔11～13〕研究的启发下所提出的更一般的形式。通过对模型稍作改动，既可用于队列研究、生存分析，又可用于病例对照研究资料的配合或分析。不过这些模型是在研究对象为人群总体的假设前提下推导的，但队列研究、生存随访研究等的研究对象是在人群总体中独立、随机地选取的一个有代表性的样本。因此，Prentice等〔1～4〕根据研究对象的抽样性质，进一步讨论了研究样本的相对危险度回归模型。同时由于队列研究、生存随访研究中难免有个体失访或非研究疾病所致的死亡等情况发生，亦即存在截尾个体。因此，用Yi（t）=0表示个体在时刻t时已经发生截尾或失效，Yi（t）=1则表示个体仍处于观察队列中。然后根据不同的流行病学调查研究再建立相应的相对危险度回归模型。

队列研究及生存随访研究

在失效与截尾相互独立的假设条件下，用下式描

本文作者单位：上海市肿瘤研究所流行病学研究室
200032

述观察队列中研究个体的危险率：

$$\lambda_i(t)=Y_i(t)\lambda[t;Z_i(u);0\leqslant u<t],\ i=1,\ \cdots\cdots,\ n,$$

式中λ即为前面（1）式所定义的人群危险率。于是队列中研究个体i在时刻t时发生失效的条件概率为：[1～4,11～13]

$$\lambda_i(t)/\lambda_r(t)=Y_i(t)\lambda_o(t)\gamma[X_i(t)\beta]/\sum_{r=1}^{n}Y_r(t)\lambda_o(t)\gamma[X_r(t)\beta],\ \text{即}\ Y_i(t)\gamma[X_i(t)\beta]/\sum_{r=1}^{n}Y_r(t)\gamma[X_r(t)\beta], \quad (5)$$

r为队列中随访时间大于等于t的研究个体所组成的危险集（risk set）。根据偏似然函数理论[11～13]，则整个队列的偏似然函数为：

$$L(\beta)=\prod_{r=1}^{n}\{\gamma_{ii}/\sum_{r=1}^{n}\gamma_{ri}\}^{\delta_i}, \quad (6)$$

式中 $\gamma_{ii}=Y_i(t_i)\gamma[X_i(t_i)\beta]$，$\gamma_{ri}=Y_r(t_i)\gamma[X_r(t_i)\beta]$，$\delta_i$为截尾指示变量（$\delta_i=0$为截尾，$\delta_i=1$非截尾）。有了样本的偏似然函数，即可以对相对危险度回归模型中的回归系数β进行估计和假设检验[1～4,11～13]。同样，亦可获得分层相对危险度回归模型的偏似然函数[2,4]。

上述模型同样适合于生存随访研究。不过Crowley和Breslow[5]曾根据Cox的研究[11～13]，针对生存数据分析给出了一个相对危险度回归模型：

$$\lambda[t,\ Z(t)]=\lambda_o(t)\gamma[Z(t),\beta], \quad (7)$$

式中$\lambda[t,\ Z(t)]$为具有某协变量特征Z的个体在时刻t时的危险率，$\lambda o(t)$为任意基准危险率。可以看出，此式与（3）式是等价的。就Cox模型而言，$\gamma[Z(t),\beta]$取$\exp[Z\beta]$的形式。此外，还有一些作者对Cox模型这种形式的相对危险度回归模型进行了一些拓展，即建立分组Cox模型[14]。

病例对照研究

病例对照研究中的病例和对照是人群总体中的病例和非病例的一个代表性样本，因而不能获得上述相对危险度指标，不过可以估计比数比指标，即

$$OR[t;Z(u),0\leqslant u<t,\Delta]=\frac{pr[t\leqslant T<t+\Delta\mid Z(u),0\leqslant u<t;T\geqslant t]/pr[T\geqslant t+\Delta\mid Z(u),0\leqslant u<t;T\geqslant t]}{pr[t\leqslant T<t+\Delta\mid Z_o(u),0\leqslant u<t;T\geqslant t]/pr[T\geqslant t+\Delta\mid Z_o(u),0\leqslant u<t;T\geqslant t]}. \quad (8)$$

它描述的是在时间区间（t，t+△）内样本中研究个体在具有不同协变量特征时的失效率的比数比。其与研究个体所具有的协变量之间的关系，即相对危险度回归模型为：

$$OR[t;Z(u),0\leqslant u<t,\Delta]=\gamma[X(t),\beta]. \quad (9)$$

结合（8）式，可以推导出（9）式的另一种形式：

$$pr[t\leqslant T<t+\Delta\mid Z(u);0\leqslant u<t;T>t]=\frac{\lambda_o(t)\gamma[X(t),\beta]}{1+\lambda_o(t)\gamma[X(t),\beta]}, \quad (10)$$

式中$\lambda o(t)$即为（8）式的分母部分。（10）式即是流行病学研究资料分析中为大家广泛应用的logistic回归模型。在实际应用中$\lambda o(t)$常取$\exp(\alpha)$的形式，$\gamma[X(t),\beta]$则为$\exp[X(t),\beta]$。根据前面（4）式的形式，同样可获得分层情形下的回归模型。（10）式不能够处理含有截尾数据的资料，而(3)式则要灵活的多，能够充分利用所有研究个体数据信息。模型（3）在配对病例对照研究资料分析中的应用见于Prentice和Breslow的研究[2,4,15]，他们对样本的条件似然函数的建立和参数的统计推断进行了较详细的讨论。

针对病例对照研究，除了Prentice等所讨论的相对危险度回归模型，还有一些研究者在该研究领域内作过一些探讨。例如Thomas[6] 1981年曾提出一个相对危险度回归模型。在形式上比（9）或（10）式直观或具体，通过引入一个中间参数η，使模型包含了相加和相乘两种方式，即

$$\ln RR(Z,\beta,\eta)=\eta Z^t\beta+(1-\eta)\ln(1+Z^t\beta). \quad (11)$$

当$\eta=0$时，上述模型取相加方式，而$\eta=1$则为相乘模型[10]。Breslow和Storer[7]在1985年提出了一个用于病例对照研究的比数比回归模型：

$$\ln OR(Z,\beta,\eta)=\begin{cases}\dfrac{(1+Z^t\beta)^\eta-1}{\eta}, & \eta\neq 0\\ \ln(1+Z^t\beta), & \eta=0\end{cases} \quad (12)$$

η的意义同前。在他们的文章里还给出了GLIM宏指令，借助于GLIM软件即可实现上述模型的拟合。Guerreo和Johnson[8] 1982年提出的相对危险度回归模型为：

$$\ln RR(Z,\beta,\eta)=\begin{cases}\dfrac{\ln(1+\eta Z^t\beta)}{\eta}, & \eta\neq 0\\ Z^t\beta, & \eta=0\end{cases} \quad (13)$$

η的意义与前两者正好相反。Barlow[9]对分层相对危险度回归模型中的相乘、相加和指数相加模型进行了讨论，并给出了它们在配对病例对照研究和队列研究中的应用实例。

结　语

在过去的十年中，流行病学研究中的多变量统计分析方法的发展和应用非常迅速〔16〕。相对危险度回归模型则是从广义或一般性的角度对流行病学研究中一些常用的统计模型进行了概括和总结提高。在众多研究者中，以Prentice等〔1～4〕提出的相对危险度回归模型更具有一般性，其他一些研究者则对此进行了一些补充和讨论。其基本点是从具有某种协变量特征的研究人群的失效或危险率与其处于基准协变量状态时的失效或危险率的比出发，研究该比值与个体所具有的协变量特征之间所存在的某种函数关系。这是近年来在流行病学研究中学术气氛活跃、内容相当丰富的一个领域，可以说是现代流行病学研究中最引人注目的统计分析方法。所以，其在理论和实用性方面的研究正方兴未艾。如不少作者对考虑了死因竞争风险时模型的建立及参数推断〔17,18〕、小样本情形下的参数统计推断〔4〕、模型的拟合优度检验或分析〔19〕、时间区间分组长短对参数估计的影响〔2〕、时变协变量存在时模型的应用和参数的统计推断〔4〕、关于回归诊断〔1,20〕等具体情况或问题上进行过论述。

参考文献

1 Prentice RL, Kalbfleisch JD. Hazard rate models with covariates. Biometrics, 1979, 35：25.

2 Prentice RL, Farewell VT. Relative risk and odd ratio regression. Aun Rev Public Health, 1986, 7：35.

3 Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. Biometrika, 1982, 69：331.

4 Moolgavkar SH. Prentice RL. Modern statistical methods in chronic disease epidemiology. New York：John Wiley & Sons, 1986, 50.

5 Crowley J, Breslow N. Statistical analysis of survival data. Ann Rev Public Health, 1984, 5：385.

6 Thomas DC. General relative risk models for survival time and matched case-control analysis. Biometrics, 1981, 37：673.

7 Breslow NE, Storer BE. General relative risk functions for case control studies. Am J Epidemiol, 1985, 122：149.

8 Guerreo VM, Johnson RA. Use of the Box-Cox transformation with binary response models. Biometrika, 1982, 69：309.

9 Barlow WE. General relative risk models in stratified epidemiologic studies. Appl Statist, 1985, 34：246.

10 Moolgavkar SH, Venzon DJ. General relative risk regression model for epidemiologic studies. Am J Epidemiol, 1937, 126：949.

11 Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. New York：John Wiley & Sons, 1980.70-142.

12 Cox DR. Regression models and life tables (with discussion). J R Statist Soc B, 1972, 34：187.

13 Cox DR. Partial likelihood. Biometrika, 1975, 62：269.

14 Aranda-Ordaz FJ. An extension of the proportional-hazards model for grouped data. Biometrics, 1983, 39：109.

15 Prentice RL, Breslow NE. Retorspective studies and failure time models. Biometrika, 1978, 65：153.

16 Gail MH. A bibliography and comments on the use of statistical models in epidemiology in the 1980s. Stat Med, 1991, 10：1819.

17 Prentice RL, Kalbfleish SD, Peterson AV, et al. The analysis of failure times in the presence of competing risks. Biometrics, 1978, 34：541.

18 Lubin JH. Case-control methods in the presense of multiple failure times and competing risks. Biometrics, 1985, 41：49.

19 Schoenfeld D. Chi-squared goodnese of fit tests for the proportional regression hazards model. Biometrika 1980, 67：145.

20 Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika, 1982, 69：239.

（收稿：1993-05-28　修回：1993-07-21）