中华流行病学杂志1995年8月第16卷第4期

·方法评价·

GLIM 在流行病学研究资料分析中的应用

——队列研究中 Poisson 回归模型的配合

项永兵　高玉堂　邓　杰　金　凡　阮志贤　秦德霖

GLIM[1,2]是基于广义线性模型理论[3]所设计的一个用于数据统计分析的小型计算机软件。以GLIM3.77版为例，除具有一般软件的通用功能(如数据描述、多变量回归模型等)外，用户尚可以拟合“自定义模型”。而且 GLIM 系统中所配置的宏指令库，其功能也相当丰富，并可以不断更新。在流行病学研究资料分析中，GLIM 有着广泛的应用。它可以很方便地用于配合 Poisson 模型、非条件 logistic 模型、各种参数回归模型等。除此之外，亦可用于拟合生存数据分析中的 Cox 回归模型等。本文主要介绍 Poisson 回归模型及其在队列随访资料分析中的应用。Poisson 回归模型在流行病学研究中的应用主要有两个方面，一是描述流行病学，二是分析流行病学。例如，用 Poisson 回归模型分析长期积累的疾病(如肿瘤)发病、死亡资料，可以定量评价年龄、时期和队列因素在疾病发病或死亡发生过程中的作用，即所谓的 APC 模型[4,5]。而在分析流行病学研究中，Poisson 回归模型主要用于队列随访或前瞻性研究资料的分析[6~11]。亦可用于生存随访资料的分析[11]。本文的目的在于介绍如何借助于 GLIM 软件来实现 Poisson 回归模型的拟合，并给出一个队列随访资料分析的实例。

统计背景

一、数据结构：队列随访研究的资料通常按分层与暴露因素整理成下述 $J\times K$ 列联表形式，见表1。其中用 d_{jk} 表示病例数，n_{jk} 表示处于 j 层和暴露水平数为 k 的人年数。列联表各个格子中的发病率估计值为：$r_{jk}=d_{jk}/n_{jk}$

表1　队列随访资料的数据格式(以发病为例)

层数(j)		暴露水平数(k)				合计
		1	2	…	k	
1	发病数	d_{11}	d_{12}	…	d_{1k}	D_1
	人年数	n_{11}	n_{12}	…	n_{1k}	N_1
2	发病数	d_{J1}	d_{J2}	…	d_{1k}	D_2
	人年数	n_{21}	n_{22}	…	n_{2k}	N_2
·		·	·	…	·	·
·		·	·	…	·	·
·		·	·	…	·	·
J	发病数	d_{J1}	d_{J2}	…	d_{Jk}	D_J
	人年数	n_{J1}	n_{J2}	…	n_{Jk}	N_J
合计	发病数	o_1	o_2	…	o_k	o_+
	人年数	n_{+1}	n_{+2}	…	n_{+k}	$N_+=n_{++}$

二、模型[5-11]：假定 d_{jk} 是服从于均数和方差为 $\mu_{jk}=n_{jk}\cdot\lambda_{jk}$ 的 *Poisson* 分布的独立变量，其中 λ_{jk} 为未知的发病概率，需用样本数据的 r_{jk} 来估计。那么，如果暴露因素在分层因素的各个层次中的相对危险度估计值(与基准水平比)是一个常数，则 λ_{jk} 与所研究因素之间的关系可以用下述相乘作用方式的回归模型来表达

$$\lambda_{jk}=exp(\alpha_j+\beta x_{jk}) \quad 或 \quad log(\lambda_{jk})=\alpha_j+\beta x_{jk} \tag{1}$$

式中 α_j 为无效参数，代表的是分层因素的作用，$\beta=(\beta_1\cdots\beta_p)$ 为 p 维待估回归系数向量，代表的是暴露因素的作用。上式是 Poisson 回归模型中最常见的一种模式。

三、利用 GLIM 配合：首先是数据输入，可以从键盘直接录入，或以文本文件的形式由程序读入。数据中包含的内容是各因素的水平数及相应的 d_{jk} 和 n_{jk}。观察单位数(units)将是各因素水平数的乘积。其次，用 FACTOR 指定各个因素的水平数，并且最低水平将被默认为基准水平，形成(0,1)指示变量(即 dummy 变量)。然后用 YVARIATE 定义 d_{jk} 为模型的因变量，连接函数取对数(LINK L)。问题是人年数的处理。由(1)式可得

$$\log\mu_{jk}=\log(n_{jk})+\alpha_j+\beta x_{jk} \tag{2}$$

作者单位：上海市肿瘤研究所流行病学研究室　200032

可见模型中含有一常数项 $\log(n_{jk})$。在 GLIM 中，把它作为分枝(offset)处理[1,2,4,5,10]。另一种处理办法是指定 λ_{jk}($\hat{\lambda}_{jk}=r_{jk}$)作为模型的因变量，而把 n_{jk}作为权重(weight)处理[1,2,8~10]。参数的估计用迭代加权最小二乘法[1~3,8~10]。

四、模型拟合优度：主要有两个统计量：卡方统计量(残差平方和)和偏差度，即下述(3)、(4)式

$$x^2=\sum_{j=1}^{J}\sum_{k=1}^{K}(d_{jk}-\hat{d}_{jk})^2/\hat{d}_{jk} \qquad (3)$$

$$G^2=2\{\sum_{j=1}^{J}\sum_{k=1}^{K}[d_{jk}\log(d_{jk}/\hat{d}_{jk})+(\hat{d}_{jk}-d_{jk})]\} \qquad (4)$$

自由度为人年数不为零的观察单位数减于模型中的参数个数，再根据卡方分布来确定 P 值。此外，根据两个模型的偏差度及自由度的改变，对选进或剔除出模型的变量进行假设检验。进一步可以做残差分析及回归诊断等[1,2,8~10]。

实例应用

笔者利用在上海地区所开展的一项大型前瞻性研究[12]的部分数据为例，说明 Poisson 回归模型在队列随访资料分析中的应用。以市区的男性队列为例，所考虑的的因素为年龄、吸烟。其中，年龄为分层因素，吸烟为研究者所感兴趣的暴露因素。资料整理成表1所示的列联表形式。年龄从20岁以下组开始，5岁为一组，至80岁以上组，分成13个层次。至于吸烟因素，首先看它分为不吸烟与吸烟二水平暴露的情形。模型的拟合过程及拟合优度检验见表2，即所谓的 Poisson ANOVA 表[1,2,9,10]。从模型的拟合优度检验来看，模型4拟合良好。从模型2与3的拟合情况来看，无论是年龄，还是吸烟因素都不足以描述该数据资料所存在的变异。模型4可以说是拟合良好的模型。在该模型的基础上，可以对参数进行估计。如吸烟因素的回归系统和其标准误为1.4095 (0.1561)。其相对危险度及其95%置信区间的估计值为4.09 (95%CI：3.01～5.56)。其次，当吸烟因素为多级水平暴露时，亦可很方便地拟合 Poisson 回归模型。例如吸烟分为六个水平：不吸烟、≤6支/天、7～14支/天、15～24支/天、25～34支/天、≥35支/天。Poisson ANOVA 分析结果列于表3。与表2的结果相比，模型4的拟合似乎更好一些。基于该模型的参数估计值及有关统计量列于表4。表中同时列出了粗相对危险度的估计值。结果表明吸烟是男性肺癌发病重要的危险因素。从吸烟量上看，存在着明显的剂量反应关系。

若要控制更多的混杂因素，可以在表2或表3所分析的资料基础上引进其它可能的混杂因素(如职业因素、文化经济程度、居住小环境等)，再对参数进行估计。例如以吸烟为二水平暴露时为例，再引进性别变量。模型拟合情况见表5。模型2、3、4相当于单因素分析的模型，模型5、6、7为两两相互调整的模型，模型8为三因素相互调整的模型，模型9、10是在模型8的基础上引进一次交互作用项的模型。还可以进一步引进高次交互作用项。从结果来看，年龄与吸烟、性别与吸烟交互作用项对模型没有贡献。在模型8的

表2 Poisson ANOVA 表：上海市区男性肺癌发病数据拟合 Poisson 模型*

模型	模型项	参数个数	偏差度 G^2	自由度 df	P 值
1	总均数	1	594.84	25	<0.001
2	年龄	13	120.16	13	<0.001
3	吸烟	2	470.51	24	<0.001
4	年龄＋吸烟	14	18.40	12	0.104
5	饱和模型#	26	0.00	—	—

* 吸烟因素为二水平情形（不吸烟与吸烟）

\# 在模型4的基础上引入年龄×吸烟交互作用项。

表3 Poisson ANOVA 表：上海市区男性肺癌发病数据拟合 Poisson 模型*

模型	模型项	参数个数	偏差度 G^2	自由度 df	P 值
1	总均数	1	658.57	77	<0.001
2	年龄	13	183.89	65	<0.001
3	吸烟	6	450.78	72	<0.001
4	年龄＋吸烟	18	47.59	60	0.877
5	饱和模型#	78	0.00	—	—

* 吸烟因素为多级暴露水平情形(不吸烟、≤5、6～14、15～24、25～34、≥34支/天)。

\# 同表2。

表4 上海市区男性吸烟对肺癌发病的相对危险度估计

吸烟*	粗相对危险度	调整相对危险度**
不吸烟	1.0	1.0
≤6	2.09	2.70(1.47～4.96)#
7—14	2.36	2.45(1.64～3.64)#
15—24	7.29	5.24(3.77～7.28)#
24—34	16.16	9.06(5.34～15.38)#
35＋	11.50	7.05(3.20～15.56)#

* 吸烟：支/天；* * 在表3模型4的基础上估计，括号内为95%置信区间(95%*CI*)；# $P<0.01$，有高度统计意义。

表5 Poisson ANOVA 表：上海市区肺癌发病数据拟合 Poisson 模型*

模型	模型项	参数个数	偏差度 G^2	自由度 DF	P 值
1	总均数	1	884.67	51	<0.001
2	年　龄	13	287.29	39	<0.001
3	性　别	2	810.10	50	<0.001
4	吸　烟	2	602.40	50	<0.001
5	年龄＋性别	14	196.99	38	<0.001
6	年龄＋吸烟	14	63.16	38	=0.006
7	性别＋吸烟	3	602.40	49	<0.001
8	年龄＋性别＋吸烟	15	49.75	37	=0.079
9	＋年龄×吸烟#	27	34.66	25	=0.095
10	＋性别×吸烟#	16	49.28	36	=0.069

*吸烟为二水平暴露情形；#在模型8基础上引入的一次交互作用项。

基础上，吸烟因素的参数估计为1.4825(0.1289)，相对危险度及其95%置信区间的估计值为4.40（95% CI：3.4151～5.6793)。性别（男＝2，女＝1）因素的参数估计为0.4888（0.1298)，相对危险度及其95%置信区间为1.6304（95%CI：1.2461～2.1332)。

结　语

笔者对 Poisson 回归模型在队列随访资料分析中的应用进行了系统讨论。主要就如何借助于 GLIM 软件配合 Poisson 回归模型进行了详细介绍。文中给出了吸烟与肺癌关系研究的队列随访资料分析实例。结果表明，男性吸烟与非吸烟者相比，其发生肺癌的相对危险度约为4，且有高度统计意义，并有明显的剂量－反应关系存在。

很显然，多变量回归模型分析具有很多的优势。例如，可以控制更多的混杂因素，来研究所感兴趣的暴露因素与疾病发生之间的关系，并给出定量估计；无论是连续性变量，还是等级变量都可以很方便引进模型；同时对因素间的交互作用项的测度也很方便。此外，利用 Poisson 回归模型分析队列随访资料时还有一大优势，就是无论是采用外部对照，还是内对照参比，同样可借助于 GLIM 来配合，并获得相应的参数估计值[9,10]。

参 考 文 献

1 Baker RJ, Nelder JA. The GLIM system: release 3.77 manual. Oxford: Numerical Algorithms Group. 1985.

2 Aitkin M, Anderson D, Francis B, Hinde J. Statistical modelling in GLIM. Oxford: Oxford University Press, 1989.

3 McCullagh P, Nelder JA. Generalized linear models. London: Chapman and Hall, 1983.

4 Kupper LL, Janis JM, Karmous A, et al. Statistical age－period－cohort analysis: a review and critique. J Chron Dis, 1985, 38: 811.

5 Clayton D, Schifflers E. Models for temporal variation in cancer rates. Ⅰ. age－period－cohort models. Stat Med, 1987, 6: 469.

6 Pearce N, Checkoway H. A simple computer program for denerating person－time data in cohort studies involving time－related factors. Am J Epidemiol, 1987, 125: 1085.

7 Selmer R. A comparison of Poisson regression model fitted to multiway summary tables and Cox's survival model using data from a blood pressure screening in the city of Bergen, Norway. Stat Med, 1990, 9: 1157.

8 Frome EL. The analysis of rates using Poisson regression models. Biometrics, 1983, 39: 665.

9 Frome EL, Checkoway H. Use of Poisson regression models in estimating incidence rates and ratios. Am J Epidemiol, 1985, 121: 309.

10 Breslow NE, Day NE. Statistical methods in cancer research. Vol Ⅱ. The design and analysis of cohort studies. Lyon: IARC, 1987.

11 Holford TR. The analysis of rates and of survivorship using log－linear models. Biometrics, 1980, 36: 299.

12 邓杰，高玉堂，汪钟贤，等. 吸烟、大气污染与肺癌的关系－上海市21万成年居民的前瞻性研究. 肿瘤，1992，12：258.

（收稿：1994-10-22　修回：1994-12-29）