·基础理论与方法·

稳健Poisson和log-binomial的GEE模型应用于非独立数据的研究

周舒冬 郜艳晖 李丽霞 张敏 杨翌 陈跃

【导读】 探讨流行病学资料中非独立数据的*RR*/患病率比(*PR*)的合适估计方法。采用计算机模拟实验和实例分析观察稳健Poisson-GEE和log-binomial-GEE模型的适用性并进行比较。结果表明log-binomial-GEE模型与稳健Poisson-GEE模型的收敛率基本均为100%,两模型估计各参数的平均值均与真值接近;在类内聚集性变小或类别数增加时,两模型估计各参数的95%*CI*覆盖率均有所提高;稳健Poisson-GEE模型对参数估计的稳健性较好,应用到实例时可正确评价暴露对结局的影响。稳健Poisson和log-binomial的GEE模型很少存在收敛问题,且有较高的准确率,可用于流行病学资料中非独立数据的*RR*/*PR*值估计。

【关键词】 稳健Poisson回归; log-binomial模型; 非独立; 广义估计方程

A simulation/case study under the use of robust Poisson and log-binomial model with generalized estimating equation models regarding non-independent data

Zhou Shudong[1], *Gao Yanhui*[1], *Li Lixia*[1], *Zhang Min*[1], *Yang Yi*[1], *Chen Yue*[2]. 1 *Guangdong Key Laboratory of Molecular Epidemiology, Department of Epidemiology and Biostatistics, School of Public Health, Guangdong Pharmaceutical University, Guangzhou* 510310, China; 2 *Department of Epidemiology and Community Medicine, University of Ottawa*, Canada

Corresponding author: Gao Yanhui, Email: gao_yanhui@163.com

This work was supported by a grant from the Science Foundation of Guangdong Province (No. 10151022401000018).

【Introduction】 To explore the appropriate method in estimating relative risk (*RR*)/prevalence ratio (*PR*) related to non-independent datasets. The simulation datasets generated by computer and case study were analyzed by two generalized estimating equation (GEE) models to investigate and compare the related applicability. Both convergence effects of log- binomial-GEE model and Robust Poisson-GEE model were almost 100%. The estimation results of the two GEE models were both closer to the true value. 95%*CI* coverage of the two GEE models increased along with the reduction of class aggregation or the increase of the number of categories. Robust-Poisson-GEE model seemed to be more stable and steady than the log-binomial-GEE. The two GEE models could correctly evaluate the effects of exposure on the outcome in the case study. Rarely, there appeared problems on convergence of Robust Poisson or log-binomial-GEE model, and the accuracy was high. Both models could be used to estimate the *RR*/*PR* on non-independent epidemiological data.

【Key words】 Robust Poisson regression; Log-binomial model; Non-independent; Generalized estimating equation

医学研究中纵向观察、多阶段抽样资料的设计并不满足独立性条件,如在统计分析时忽略数据间的非独立性,可能增加统计推论时犯Ⅰ类错误的概率;由于未估计群内聚集性以及评价不同群内反应变量差异是否与协变量有关,导致信息的删失,而这些信息正是对研究者决策和评价所需。此时可考虑使用广义估计方程(GEE)模型进行参数估计,其作业相关矩阵有助于解释观察值之间的相关性[1],且参数估计稳定。此外测量暴露对结局的影响是许多科研工作的主要目的之一,实际工作中常用logistic回归模型计算*OR*值作为*RR*值的估计并给予同样解释。但当研究结局为非稀有事件(发病率或患病率较高)时,*OR*值可严重高估暴露因素对结局的影响[2],此时可用稳健Poisson模型或log-binomial模型

DOI: 10.3760/cma.j.issn.0254-6450.2014.04.024

基金项目:广东省自然科学基金(10151022401000018)

作者单位:510310 广州,广东药学院公共卫生学院流行病与卫生统计学系 广东省分子流行病学重点实验室(周舒冬、郜艳晖、李丽霞、张敏、杨翌);加拿大渥太华大学流行病与社区医学系(陈跃)

通信作者:郜艳晖, Email: gao_yanhui@163.com

计算*RR*/患病率比(*PR*),其对点估计和区间估计的解释均比logistic回归模型的*OR*更为合理[3]。

针对流行病学资料中非独立数据的*RR*/*PR*值的估计,本研究利用计算机模拟和实例研究,探讨非独立数据中基于稳健Poisson和log-binomial的GEE模型应用并进行比较。

基本原理

稳健Poisson回归及log-binomial的GEE模型原理:

设y_{ki}和$X_{ki}=(x_{ki1},x_{ki2},\cdots,x_{kiP})^T$分别为第$k(k=1,2,\cdots,K)$类内第$i(i=1,2,\cdots,n_K)$个个体的二分类结局变量和$P\times1$维解释变量向量,稳健Poisson和log-binomial模型形式:

$$\log(\hat{p}_{ki})=\beta_0+\sum_{p=1}^{P}\beta_p x_{pki} \quad (1)$$

式中$\hat{p}_{ki}=\Pr(y_{ki}=1\mid X_{ki})$,回归系数$\beta_p$的涵义表示当控制其他自变量后,第$p$个自变量$x_p$每变化一个单位时边际效应平均值的相应变化。

当误差分布指定为Poisson分布时,可构建稳健Poisson-GEE模型,由于Poisson分布的方差等于均数,应用到二项分布资料易出现过度离散(overdispersion),导致较宽的置信区间。因此应引入Huber的稳健方差:

$$Var(\hat{\beta})=A^{-1}BA^{-1} \quad (2)$$

其中

$$A=\sum_{k=1}^{K}\sum_{i=1}^{n_k}X_{ki}X_{ki}^T\hat{p}_{ki}$$

$$B=\sum_{k=1}^{K}\Big[\sum_{i=1}^{n_k}X_{ki}(y_{ki}-\hat{p}_{ki})\Big]\Big[\sum_{i=1}^{n_k}(y_{ki}-\hat{p}_{ki})X_{ki}^T\Big] \quad (3)$$

式(1)中,当误差分布指定为二项分布时可构建log-binomial-GEE模型,该模型在参数估计过程中有限制条件,且不易收敛,此时可用COPY算法解决模型收敛的问题[4]。

最后根据一致性估计方程理论[5],参数$\beta=(\beta_0,\beta_1,\cdots,\beta_P)^T$的得分(score)方程为

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k}X_{ki}\Big[y_{ki}-\exp(\beta_0+\sum_{p=1}^{P}\beta_p x_{kip})\Big]=0 \quad (4)$$

式(4)的解即为参数β的一致估计。

虽然GEE模型用于解决观察值间的非独立性时要求指定作业相关矩阵,但Liang和Zeger[6]指出,只要模型本身正确,即使模型中作业相关矩阵指定错误,所得模型的固定参数估计仍然一致,且当二水平单位数较多时,GEE中方差的估计也渐近无偏;不过正确的指定可提高估计效率。

实例分析

1. 模拟研究:

(1)方法及参数设置:考虑到流行病学研究中非独立数据的特点,即①以常见的二分类变量和连续型变量作为协变量,构造X_1(1=暴露;0=非暴露)和X_2(连续型协变量),Y为二分类因变量(1=结局发生;0=结局未发生)。②假设数据集中有N($N=20$或50)个群,每个群包括N_k个个体(当群数$N=20$时,群内个体数N_k由均匀分布$U(0,50)$随机产生,$N_k=25$;群数$N=50$时,群内个体数N_k由均匀分布$U(0,20)$随机产生,$N_k=10$[7,8]。③实际工作中非独立数据的类内相关系数常介于0.01和0.15之间,故假设群内个体具有相似性,并利用随机效应模型生成非独立数据(假设随机效应参数来自均数为0,方差为0.1或0.2的两种正态分布情况,其大小依赖于暴露因素和协变量的效应)。④考虑基线患病率为0.3和0.15两种情况,暴露因素X_1的*PR*值分别为1.5和2.0,暴露率分别为0.2和0.5,协变量X_2的回归系数分别为0.18、0.10、0.36和0.20[5]。共可组成8种参数情况(表1)。

表1　模拟研究固定效应参数设置情况

参数设置	β_0	β_1	β_2	p
1	log(0.3)	log(1.5)	0.18	0.2
2	log(0.3)	log(1.5)	0.18	0.5
3	log(0.3)	log(2.0)	0.10	0.2
4	log(0.3)	log(2.0)	0.10	0.5
5	log(0.15)	log(1.5)	0.36	0.2
6	log(0.15)	log(1.5)	0.36	0.5
7	log(0.15)	log(2.0)	0.20	0.2
8	log(0.15)	log(2.0)	0.20	0.5

在表1各参数设置情况下,本研究共模拟2(类别数为20和50)×2(随机效应方差0.1和0.2)×8(固定效应参数设置情况1~8)=32种情况。每种情况改变随机数的种子可产生不相关的1 000个模拟数据集,每个数据集均利用稳健Poisson-GEE模型和log-binomial-GEE模型进行参数估计,最后观察每种情况下两模型平均的收敛率和参数估计值及其95%*CI*覆盖率。

(2)结果:32种模拟条件下两种GEE模型参数估计的收敛率几乎均为100%,且两种模型对各设定条件下的参数估计结果与真值均较接近,受类别数或随机效应方差的影响不大。

各种模拟条件下,稳健Poisson-GEE和log-

binomial-GEE的覆盖率为71%～95%，相同模拟条件下两类模型的覆盖率非常接近，差别甚微，但log-binomial-GEE模型覆盖率的变动稍大于稳健Poisson-GEE模型。和随机效应方差为0.1相比，随机效应方差为0.2时，稳健Poisson-GEE和log-binomial-GEE模型对各参数估计的95%*CI*覆盖率均有轻微下降；类别数为50时，两类模型的95%*CI*覆盖率均有轻微增大（表2）。

2. 实例分析：数据源自欧洲社会调查（www.europeansocialsurvey.org），观察2010年26个欧洲国家49 024名居民生活满意度情况（表3），并研究自我健康评价及家庭收支情况对生活满意度的影响。在26国居民中，对生活持满意态度者约占44.83%（21 979/49 024），故研究结局为非稀有事件；另一方面考虑到各国居民的生活满意度可能在本国内具有相似性，数据不满足独立性，因此利用log-binomial-GEE和稳健Poisson-GEE模型进行影响因素分析。

表4显示稳健Poisson-GEE模型、log-binomial-GEE模型估计各因素对居民生活满意度影响的*PR*值及其95%*CI*，两模型均可顺利收敛，调整性别和年龄后，自我健康评价和家庭经济状况对生活满意度的影响均有统计学意义，即自我健康评价越满意或家庭经济状况越好，则生活满意程度越高。稳健Poisson-GEE模型与log-binomial-GEE模型对*PR*值的估计结果非常接近，且稳健Poisson-GEE模型估计的95%*CI*稍宽于log-binomial-GEE模型。利用logistic模型估计*OR*值，与两个GEE模型相比，logistic模型估计解释变量的*OR*值均高于*PR*值。

讨　论

本文主要探讨流行病学调查统计分析时常忽略的两个重要问题。一是对非罕见事件宜直接采用*PR*值或*RR*值描述暴露对结局的影响，二是分析非独立资料时需考虑数据的层次结构特征，为此通过模拟研究了基于稳健Poisson回归和log-binomial模型的GEE模型，在不同参数设置和背景下其适用性及存在问题。模拟结果显示，两类模型的GEE收敛效果均较好；在模拟实验条件下，两类GEE模型的参数估计均值与真值均接近，表明两类模型参数估计的准确性均较高；但模拟结果也显示，在类别数较少时，两种GEE模型95%*CI*的覆盖率也稍低于类别

表2 不同模拟条件下两类模型参数估计值的95%*CI*覆盖率（%）

参数设置	参数	真值	*K*=50				*K*=20			
			随机效应方差0.1		随机效应方差0.2		随机效应方差0.1		随机效应方差0.2	
			模型1	模型2	模型1	模型2	模型1	模型2	模型1	模型2
1	β_0	0.3	90.9	89.9	87.5	86.7	88.6	88.4	87.3	88.1
	β_1	1.5	94.5	94.1	90.2	87.0	92.4	91.5	88.3	86.2
	β_2	0.18	91.6	90.5	89.7	88.1	91.7	90.3	90.0	88.3
2	β_0	0.3	93.7	93.1	91.4	90.4	89.7	89.3	91.0	90.0
	β_1	1.5	92.5	91.5	90.9	89.0	91.9	89.8	90.3	88.7
	β_2	0.18	89.8	86.1	82.0	73.1	86.0	82.9	79.0	71.7
3	β_0	0.3	91.0	91.6	89.4	88.7	91.3	91.3	87.7	87.8
	β_1	2.0	91.6	90.4	88.3	86.2	91.0	89.9	84.7	82.8
	β_2	0.1	91.9	90.4	91.0	88.6	89.3	87.5	91.0	86.8
4	β_0	0.3	93.4	93.1	92.4	92.8	92.1	92.4	90.4	90.2
	β_1	2.0	93.8	93.4	85.6	84.7	90.6	91.0	85.3	84.4
	β_2	0.1	87.4	84.0	78.4	71.8	82.5	76.8	79.4	75.3
5	β_0	0.15	89.9	89.8	88.1	87.9	92.4	92.1	88.6	88.3
	β_1	1.5	91.6	90.2	92.9	92.5	90.9	91.0	90.4	90.3
	β_2	0.36	93.0	90.2	93.4	90.1	90.9	89.4	88.4	86.8
6	β_0	0.15	93.3	93.4	90.5	90.0	92.8	92.4	88.5	89.0
	β_1	1.5	92.8	92.3	93.4	92.5	89.4	89.7	91.2	90.1
	β_2	0.36	92.7	88.9	87.6	79.0	90.4	85.5	85.2	76.1
7	β_0	0.15	93.7	94.5	88.7	88.4	89.2	89.6	87.1	86.7
	β_1	2.0	94.2	94.4	94.5	94.3	92.6	92.5	93.3	93.4
	β_2	0.2	92.8	92.6	91.1	91.9	90.9	92.0	91.5	92.9
8	β_0	0.15	92.2	92.1	88.5	89.0	88.8	88.8	89.5	89.5
	β_1	2.0	92.8	92.8	91.7	91.1	90.9	90.7	92.4	91.9
	β_2	0.2	95.2	95.3	90.7	89.7	92.8	93.1	87.6	88.1

注：模型1和模型2分别是稳健Poisson-GEE和log-binomial-GEE（COPY算法）模型

表3　2010年26个欧洲国家49 024名居民不同因素状态下生活满意度分布

因素	满意	不满意
性别		
男	10 249(45.86)	12 101(54.14)
女	11 730(43.98)	14 944(56.02)
自我健康评价*		
较差	934(19.04)	3 972(80.96)
一般	4 507(33.21)	9 065(66.79)
良好	16 520(54.18)	13 969(45.82)
家庭经济状况		
很艰难	805(15.11)	4 521(84.89)
艰难	2 667(24.83)	8 073(75.17)
一般	9 874(46.87)	11 193(53.13)
舒适	8 633(72.60)	3 258(27.40)

注：*数据有缺失

数较多时，结果与Lu等[9]的结论一致，主要是由于GEE在类别数较少时可低估参数的方差，这可能增加Ⅰ类错误的风险。同时两类模型相比，稳健Poisson-GEE模型对参数估计的结果较log-binomial-GEE模型更稳健。

在“欧洲社会调查”的实例分析中，对生活持满意态度者的频率高达44.83%，由于各国居民对生活满意度的评价具有群内相似性，个体间的研究结局并不独立，因此本文选用稳健Poisson-GEE和log-binomial-GEE模型，与普通logistic回归相比，该两GEE模型均未高估自我健康评价和家庭经济状况对生活满意度的影响；而且稳健Poisson-GEE模型估计的*PR*值及其95% *CI*范围稍宽于log-binomial-GEE模型，故该模型的Ⅰ类错误概率更低。

目前国外已提供多种基于*RR*/*PR*的模型及参数估计方法[10]，并逐步推广到具有非独立特征的医学研究中，基于稳健Poisson回归和log-binomial的广义估计方程模型已得到多数学者的认同，且两模型均可利用SAS软件的proc genmod过程实现，数据结构和程序指定简便，易于应用。Yelland等[7]通过大型模拟研究比较后也提出log-binomial-GEE模型偶尔存在收敛问题，推荐非独立数据中的*RR*/*PR*估计可选择稳健Poisson模型的GEE方法。

表4　不同模型估计各因素对居民生活满意度影响的*PR*值和*OR*值及其95% *CI*

固定效应	稳健Poisson-GEE *PR*值(95%*CI*)	log-binomial-GEE *PR*值(95%*CI*)	logistic *OR*值(95%*CI*)
自我健康评价			
不满意	1.000	1.000	1.000
一般	1.464(1.349～1.590)	1.463(1.354～1.581)	1.683(1.547～1.832)
满意	2.065(1.825～2.337)	2.028(1.814～2.268)	3.320(3.056～3.607)
家庭经济状况			
很艰难	1.000	1.000	1.000
艰难	1.551(1.349～1.782)	1.551(1.350～1.782)	1.717(1.572～1.875)
一般	2.749(2.177～3.472)	2.740(2.169～3.462)	4.225(3.896～4.582)
舒适	4.032(3.111～5.225)	3.993(3.080～5.176)	11.675(10.704～12.734)

注：所有模型均调整了性别和年龄(连续型变量)

参考文献

[1] Rao KQ. Development and application of health statistics method [M]. Vol. 2. Beijing: People's Medical Publishing House, 2008. (in Chinese)
饶克勤. 卫生统计方法与应用进展[M]. 第2卷. 北京：人民卫生出版社，2008.

[2] Str O, Mberg U. Prevalence odds ratio v prevalence ratio[J]. Occup Environ Med, 1994, 51(2): 143-144.

[3] Lee J, Chia KS. Estimation of prevalence rate ratios for cross sectional data: an example in occupational epidemiology[J]. Br J Ind Med, 1993, 50(9): 861-862.

[4] Deddens JA, Petersen MR, Lei X. Estimation of prevalence ratios when PROC GENMOD does not converge[C]. Proceedings of the 28th Annual SAS Users Group International Conference, Seattle, Washington, 2003: 270-280.

[5] Zou GY, Donner A. Extension of the modified Poisson regression model to prospective studies with correlated binary data[J]. Stat Methods Med Res, 2013, 22(6): 661-670.

[6] Liang KY, Zeger SL. Longitudinal data analysis using generlaized linear models[J]. Biometrika, 1986, 73: 13, 222.

[7] Yelland LN, Salter AB, Ryan P. Relative risk estimation in cluster randomized trials: a comparison of generalized estimating equation methods[J]. Int J BIOST, 2011, 7(1): 1-26.

[8] Yu B, Wang Z. Estimating relative risks for common outcome using PROC NLP[J]. Comput Methods Programs Biomed, 2008, 90(2): 179-186.

[9] Lu B, Preisser JS, Qaqish BF, et al. A comparison of two bias-corrected covariance estimators for generalized estimating equations[J]. Biometrics, 2007, 63(3): 935-941.

[10] Gao YH, Zhou SD, Li LX, et al. Statistical methods on the estimation of relative risk or prevalence ratio [J]. Chin J Epidemiol, 2013, 34(9): 935-939. (in Chinese)
郜艳晖，周舒冬，李丽霞，等. 基于相对危险度/患病率比的模型及参数估计方法研究进展[J]. 中华流行病学杂志，2013，34(9)：935-939.

(收稿日期：2013-10-18)
(本文编辑：张林东)