·基础理论与方法·

log-binomial模型估计的患病比及其应用

叶荣　郜艳晖　杨翌　陈跃

【导读】 探讨有和无连续协变量时的log-binomial模型估计患病比的统计方法及其应用。文中使用禁烟立法态度与吸烟关联的研究实例，比较log-binomial模型估计的现患比(*PR*)与logistic回归模型估计的优势比(*OR*)。当模型中无连续协变量时，采用最大似然估计拟合log-binomial模型；当因含有连续协变量导致模型不收敛时，则采用COPY方法估计*PR*。分别估计男、女禁烟立法态度与吸烟的关联。由于女性吸烟率低，用*PR*与*OR*所估计的关联结果相似。而男性吸烟率较高，*OR*值明显大于*PR*。当年龄作为连续协变量纳入模型时，导致log-binomial模型不收敛，采用COPY方法解决此问题。所有分析均在SAS软件中实现。结论：当患病率较高时，*PR*比*OR*更好地测量了暴露与疾病的关联。文中给出log-binomial回归模型和COPY方法估计*PR*的SAS程序。

【关键词】 现患比；Log-binomial模型；COPY方法

Using log-binomial model for estimating the prevalence ratio

YE Rong[1], *GAO Yan-hui*[1], *YANG Yi*[1], *CHEN Yue*[2]. 1 *Department of Epidemiology and Health Statistics, Guangdong Pharmaceutical University, Guangzhou* 510310, China; 2 *Department of Epidemiology and Community Medicine, University of Ottawa*, Canada

Corresponding author: *GAO Yan-hui*, Email: gao_yanhui@163.com

【Introduction】 To estimate the prevalence ratios, using a log-binomial model with or without continuous covariates. Prevalence ratios for individuals' attitude towards smoking-ban legislation associated with smoking status, estimated by using a log-binomial model were compared with odds ratios estimated by logistic regression model. In the log-binomial modeling, maximum likelihood method was used when there were no continuous covariates and COPY approach was used if the model did not converge, for example due to the existence of continuous covariates. We examined the association between individuals' attitude towards smoking-ban legislation and smoking status in men and women. Prevalence ratio and odds ratio estimation provided similar results for the association in women since smoking was not common. In men however, the odds ratio estimates were markedly larger than the prevalence ratios due to a higher prevalence of outcome. The log-binomial model did not converge when age was included as a continuous covariate and COPY method was used to deal with the situation. All analysis was performed by SAS. Prevalence ratio seemed to better measure the association than odds ratio when prevalence is high. SAS programs were provided to calculate the prevalence ratios with or without continuous covariates in the log-binomial regression analysis.

【Key words】 Prevalence ratio; Log-binomial model; COPY algorithm

在流行病学研究中，优势比(*OR*)常作为相对危险度(*RR*)的估计值来描述暴露(干预)与疾病(结局)的关联强度。在横断面研究中，现患优势比(prevalence odds ratio, POR)也被广泛应用。但当疾病患病率较大(>10%)时，若仍用*OR*描述关联强度，则会高估暴露与疾病的关联[1-3]。Thompson等[4]建议横断面研究中应采用现患比(*PR*)描述暴露与疾病的关联强度。*PR*是暴露(干预)与非暴露(未干预)者患病概率的比值，可通过拟合log-binomial模型[5-7]得到*PR*的最大似然估计(maximum likelihood estimation, MLE)。但当存在连续型自变量时，用现有统计软件常常出现log-binomial模型不能收敛的情形。本文主要介绍log-binomial模型估计*PR*，以及模型不收敛时的COPY方法[8]和加权log-binomial模型[9]。

模型与方法

采用log-binomial模型可直接估计*PR*。log-binomial模型的因变量*Y*服从二项分布，且因变量($Y=1$)概率的对数与自变量呈线性关系：

$$P(Y=1 \mid X_1, X_2, \cdots\cdots) = e^{X\beta} \quad (1)$$

式(1)中，$X\beta=(\beta_0+\beta_1X_1+\beta_2X_2+\cdots\cdots)$。$\beta$表示在控制其他自变量时，自变量$X$与$Y$关联的回归系数，$PR=exp(\beta)$。由于因变量$Y$的概率介于0和1之间，所以log-binomial模型利用MLE估计参数β时需加一

DOI: 10.3760/cma.j.issn.0254-6450.2010.05.024

作者单位：510310 广州，广东药学院公共卫生学院流行病学与卫生统计学系(叶荣、郜艳晖、杨翌)；加拿大渥太华大学流行病学与社会医学系(陈跃)

通信作者：郜艳晖，Email: gao_yanhui@163.com

个限制条件，即$X\beta \leqslant 0$。log-binomial模型可用SAS/STAT软件中的PROC GENMOD程序实现，见程序1。

程序1 用log-binomial模型估计*PR*

```
PROC GENMOD DATA=ORIGINAL;
ODS SELECT ParameterEstimates;
ODS OUTPUT ParameterEstimates=PARA0;
MODEL Y=X1 X2 … Xk/D=BIN LINK
      =LOG INTERCEPT=-1 LRCI;
DATA PARA;
SET PARA0;
PR=EXP(ESTIMATE);
LPR=EXP(LOWERLRCI);
UPR=EXP(UPPERLRCI);
KEEP PARAMETER PR LPR UPR;
PROC PRINT DATA=PARA;
RUN;
```

当存在连续自变量时，MLE估计的参数通常在参数所限制范围的边界上，MLE得不到似然函数导数为零的极大值点，导致模型不能收敛。Deddens等[8]提出可先对原始数据集调整扩充后再拟合log-binomial模型，即可得到回归系数和*PR*的近似MLE。这种对原始数据集调整扩充的方法称为COPY方法。

COPY方法的步骤是先将原始数据集重复复制，扩大(c−1)倍，再将原始数据集中的因变量值0−1互换生成1个新的数据集，之后把扩大(c−1)倍的数据集与互换因变量值的数据集合并，最终得到COPY数据集，用于拟合log-binomial模型。可以看到，COPY方法的本质是增大样本含量(c倍于原始数据集)并尽可能多地[(c−1)/c]保留原始数据信息，得到近似的最大似然估计。利用COPY的数据集拟合log-binomial模型，原始数据的信息占(c−1)/c。c越大，MLE的参数估计偏倚越小，但参数估计耗时越长，特别是当原始数据为大样本时，使用COPY算法将面对非常大的计算量。通常至少c≥100。

Lumley等[9]在2006年的工作论文里提出可利用weighted log-binomial模型实现COPY算法。即将原始数据集和因变量0−1互换后的数据集合并，并定义原始数据集中观测的权重为(c−1)/c，而因变量0−1互换后的数据集的权重为1/c，再拟合log-binomial模型。可以看到，用weighted log-binomial模型实现COPY方法，被分析的数据集大小只是原始数据集的2倍，因此大大减少了计算工作量。

目前SAS/STAT软件中的PROC GENMOD程序提供了weight语句，可实现weighted log-binomial模型，因此可利用weight语句很方便地实现COPY方法，而不用预先生成c倍于原始数据集的COPY数据集。以c=1000为例，可利用程序2进行log-binomial模型的COPY方法。

程序2 log-binomial模型不收敛时的COPY方法

```
DATA ONE;SET ORIGINAL;W=0.999;
DATA TWO;SET ORIGINAL;Y=1-Y;W=0.001;
DATA THREE; SET ONE TWO;
PROC GENMOD DATA=THREE;
ODS SELECT ParameterEstimates;
ODS OUTPUT ParameterEstimates=PARA0;
WEIGHT=W;
MODEL Y=X1 X2 … Xk/D=BIN LINK
      =LOG INTERCEPT=-1 LRCI;
DATA PARA;
SET PARA0;
PR=EXP(ESTIMATE);
LPR=EXP(LOWERLRCI);
UPR=EXP(UPPERLRCI);
KEEP PARAMETER PR LPR UPR;
PROC PRINT DATA=PARA;
RUN;
```

实例分析

实例来自2009年“广州市禁烟立法基线调查”资料。研究禁烟立法态度(赞成、无所谓和反对)与吸烟之间是否存在关联。表1显示不同禁烟立法态度者的吸烟情况。男性和女性的总吸烟率分别为39.92%和2.20%。男性赞成禁烟立法者的吸烟率最低(33.85%)，反对禁烟立法者的吸烟率最高(81.71%)。女性也表现相同的趋势。

表1 禁烟立法态度与吸烟的频数分布

禁烟立法态度	男性		女性	
	吸烟	不吸烟	吸烟	不吸烟
赞成	593(33.85)	1159(66.15)	43(1.65)	2565(98.35)
无所谓	114(56.44)	88(43.56)	8(7.55)	98(92.45)
反对	143(81.71)	32(18.29)	10(17.54)	47(82.46)
合计	850(39.92)	1279(60.08)	61(2.20)	2710(97.80)

注：括号外数据为人数，括号内数据为构成比(%)

分别拟合log-binomial模型和logistic模型，估计男性与女性禁烟立法态度与吸烟关联的*PR*和*OR*(表2)。可以看到，男性中对禁烟态度持无所谓态度者的吸烟率是持赞成态度者的1.667(95%*CI*：1.443～1.902)倍；持反对态度者的吸烟率是持赞成态度者的2.414(95%*CI*：2.184～2.649)倍。而*OR*严重高估禁烟态度和吸烟的关联，分别是2.532(95%*CI*：1.885～3.401)和8.734(95%*CI*：5.879～12.975)。女性资料中高估程度远低于男性。

考虑到年龄是可能的混杂因素，因此把年龄也纳入模型。因为年龄为连续变量，结果导致男性年龄、禁烟立法态度与吸烟关联的log-binomial模型不收敛。因而我们采用了COPY方法，取c=1000，定义原始数据集的权重为0.999，因变量0-1互换数据集的权重为0.001，拟合weighted log-binomial模型，最终收敛(表3)。可以看到，COPY方法可以解决自变量为连续变量时导致的模型不收敛的问题。此外，在多

变量回归模型里，*OR*仍然高估关联的强度，尤其对患病率高的资料（如男性吸烟率为39.92%），关联越强，高估越严重。

表2　禁烟立法态度与吸烟关联的*PR*与*OR*

禁烟立法态度	男性		女性	
	*PR*值(95%*CI*)	*OR*值(95%*CI*)	*PR*值(95%*CI*)	*OR*值(95%*CI*)
赞成	1.000	1.000	1.000	1.000
无所谓	1.667(1.443～1.902)	2.532(1.885～3.401)	4.577(2.036～8.945)	4.869(2.230～10.635)
反对	2.414(2.184～2.649)	8.734(5.879～12.975)	10.641(5.280～19.174)	12.692(6.018～26.765)

表3　用COPY方法估计年龄、禁烟立法态度与吸烟关联的*PR*

自变量	男性		女性[a]	
	*PR*值(95%*CI*)	*OR*值(95%*CI*)[b]	*PR*值(95%*CI*)	*OR*值(95%*CI*)[b]
年龄	1.004(1.002～1.005)	1.009(1.004～1.014)	1.017(1.002～1.032)	1.018(1.002～1.033)
禁烟立法态度				
赞成	1.000	1.000	1.000	1.000
无所谓	1.695(1.468～1.932)	2.627(1.952～3.535)	4.557(2.032～8.878)	4.906(2.240～10.744)
反对	2.419(2.209～2.642)	8.974(6.033～13.348)	10.870(5.416～19.464)	13.086(6.174～27.736)

注：[a]女性资料的log-binomial模型收敛；[b] *OR*值为logistic回归模型估计的结果

讨　论

在流行病学的横断面研究中，*OR*值仍被广泛应用，Thompson建议用*PR*代替*OR*估计暴露与疾病的关联强度。与*PR*相比，当所研究的疾病患病率较高（>10%）时，*OR*会高估暴露与疾病的关联强度。本文通过禁烟立法态度和吸烟关联的实例说明该问题。和女性相比，男性有更高的吸烟率（分别为2.20%和39.92%）。结果显示，*OR*严重地高估了男性禁烟立法态度和吸烟关联的强度，在多变量回归模型中也是如此。

Log-binomial模型可用于估计*PR*，得到它的最大似然估计。但是当模型中有连续型协变量时，常规的统计软件（SAS、STATA等）中MLE方法估计参数时常出现模型不收敛。Petersen和Deddens[10]提出可以用COPY方法估计*PR*。COPY方法的基本原理类似于处理有"0"频数的四格表资料时，可以将"0"用一个很小的值取代，从而将无解模型变得有解，得到近似估计。一般情况下，COPY方法除了对异常值或者模型的错误指定敏感外，都能得到*PR*的近似无偏估计量。对某些复杂模型不能收敛的问题，有学者曾提出数据扩增（data inflation）的方法[11]，即将原始数据简单的复制k倍，点估计可直接获得，区间估计由扩增数据区间的$k^{1/2}$倍获得。数据扩增不改变原始数据的信息，但有时解决模型不收敛的效果不佳。尽管模型不收敛的原因很多，但COPY方法也为小样本数据或拟合其他复杂模型时的收敛问题提供了一条解决的思路。本研究的log-binomial模型，常因自变量为连续变量导致模型不收敛，如本文中男性禁烟立法态度与吸烟关联的模型包含了年龄变量后模型不收敛，但使用COPY方法后有效地解决了该问题。

除本文介绍的log-binomial模型和COPY方法估计*PR*外，还有学者提出可用Poisson回归模型[12]、Cox比例风险模型[3,13]估计*PR*。Barros[12]和Petersen[10,14]曾讨论各种模型的优缺点，结果显示Poisson回归模型所估计的参数标准误较大，可用稳健Poisson回归模型解决该问题；无论Poisson回归模型，还是稳健Poisson回归模型均可能出现不合理的概率估计值（如概率>1）。Cox比例风险模型不能估计截距项，因而不能估计出患病概率[12]。通常log-binomial模型估计*PR*具有较高的功效、较小的标准误且所估计的患病概率不会>1，结果也更容易解释。

总之，在流行病学的横断面研究中，可用*PR*代替*OR*描述暴露与疾病的关联强度，特别当疾病患病率较高时。采用log-binomial模型可得到*PR*的最大似然估计；当模型不收敛时，可用COPY方法估计*PR*。

参 考 文 献

[1] Stromberg U. Prevalence odds ratio *v* prevalence ratio. Occup Environ Med, 1994, 51(2): 143-144.

[2] Axelson O, Fredricksson M, Ekberg K. Use of the prevalence ratio *v* the prevalence odds ratio as a measure of risk in cross sectional studies. Occup Environ Med, 1994, 51(8): 574.

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

[4] Thompson ML, Myers JE, Kriebel D. Prevalence odds ratio or prevalence ratio in the analysis of cross sectional data: what is to be done? Occup Environ Med, 1998, 55: 272-277.

[5] Wacholder S. Binomial regression in GLIM, estimating risk ratios and risk difference. Am J Epidemiol, 1986, 123: 174-184.

[6] Zocchetti C, Consonni D, Bertazzi P. RE: Estimation of prevalence rate ratios from cross-sectional data (letter). Int J Epidemiol, 1995, 24: 1064-1065.

[7] Skov T, Deddens J, Petersen MR, et al. Prevalence proportion ratios: estimation and hypothesis testing. Int J Epidemiol, 1998, 27: 91-95.

[8] Deddens JA, Petersen MR, Lei X. Estimation of prevalence ratios when proc genmod does not converge. Proceedings of the 28th Annual SAS Users Group International Conference. Cary, NC: SAS Institute Inc, 2003: 270.

[9] Lumley T, Kronmal R, Ma S. Relative risk regression in medical research: models, contrasts, estimators, and algorithms. UW Biostatistics Working Paper Series, 2006: 293. http://www.bepress.com/uwbiostat/paper293.

[10] Petersen MR, Deddens JA. Approaches for estimating prevalence ratios. Occup Environ Med, 2008, 65: 501-506.

[11] Burton PR, Tiller KJ, Gurrin LC, et al. Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling. Genetic Epidemiol, 1999, 17: 118-140.

[12] Barros AJ, Hirakata VN. Alternative for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Med Res Methodol, 2003, 3: 21.

[13] Lee J. Odds ratio or relative risk for cross-sectional data? Int J Epidemiol, 1994, 23: 201-203.

[14] Petersen MR, Deddens JA. A comparison of two methods for estimating prevalence ratios. BMC Med Res Methodol, 2008, 8: 9.

（收稿日期：2009-10-30）
（本文编辑：张林东）