·基础理论与方法·

零频数过多模型在亚健康状态研究中的应用

徐涛 李卫 陈涛

【导读】 探讨零频数过多(ZI)模型在亚健康症状数研究中的应用。应用Stata 11.0软件拟合ZI模型分析亚健康症状数的危险因素,并用α系数、Vuong检验、O检验、似然比拟合优度检验比较ZI模型与传统负二项回归模型、Poisson回归模型的拟合效果。$\alpha=0.939$, Vuong检验$Z=32.08$, $P<0.0001$,表明此数据的零频数过多。亚健康症状数的$\bar{x}=2.90$, $s=3.85$,过度离散统计量$O=308.011$, $P<0.001$, $s^2>\bar{x}$,表明存在过度离散。从4个模型中的拟合优度看,零频数过多的负二项回归(ZINB)模型log likelihood最大,AIC最小,说明ZINB模型的拟合效果最佳。当计数资料中出现过多的零频数时(如亚健康症状数资料),应用ZINB模型能够获得最佳的拟合效果。

【关键词】 亚健康; 零频数过多模型; 负二项回归; Poisson回归

Application of zero-inflated models on regression analysis of count data: a study of sub-health symptoms

XU Tao[1,2], *LI Wei*[1], *CHEN Tao*[1]. 1 *Cardiovascular Institute and Fuwai Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing* 100037, China; 2 *Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences and School of Basic Medicine, Peking Union Medical College*

Corresponding author: LI Wei, Email: liwei0325@yahoo.com.cn

【Introduction】 To explore the goodness of fit about the zero-inflated (ZI) models in analyzing data related to sub-health symptoms in which the counts are non-negative integers. ZI models are conducted with Stata 11.0. The coefficient of α, Vuong test, O test and likelihood test are used to compare the goodness of fit for ZI models with the common used models such as passion model, negative binomial model. When α is 0.939, and the Z statistic of Vuong test is 32.08, $P<0.0001$, which shows that there are too many zeros. The mean number of sub-health symptoms is 2.90, $s=3.85$, $O=308.011$, $P<0.001$, $s^2>\bar{x}$, indicating that the data are over-dispersed. In addition, the optimum goodness of fit is found in zero-inflated negative binomial (ZINB) model with the largest log likelihood and the smallest AIC. ZINB seems the optimal model to study those over-dispersed count data with too many zeros.

【Key words】 Sub-health; Zero-inflated model; Negative binomial regression; Poisson regression

在医学研究中,经常遇到零频数过多(zero-inflated)的数据资料,而此种资料的处理是统计学研究的一个重要问题。数据离散型资料,其观察值常为零或正整数,如亚健康症状数。对于亚健康的研究多数通过健康量表的形式来完成,如国内多采用陈青山等[1]制定的亚健康评价量表评价受试者的健康状况,该量表有18项症状组成。研究者经常是将在1年时间内持续1个月以上出现所列18种症状中1项以上者定义为正处于亚健康状态。该分析实际上是将受试者的亚健康状态作为一个二分类变量来处理,通过亚健康现患情况的粗率得以反映[2-4]。二分类变量应用简便,但受试者可能只有1项亚健康症状,也可能同时有几项亚健康症状,因此亚健康现患率仅仅考虑了处于亚健康状态的人数,不能同时考虑到所拥有的亚健康症状数,这便损失了很多有用的信息,并导致不准确的结论。或以亚健康症状数作为连续性指标进行线性回归分析,但是亚健康症状数呈明显的正偏态分布、过度离散,且有很大比例的受试者症状数为0,不适于作为连续性指标拟合线性回归模型。

正是考虑到这些局限性及资料的分布特征,笔

DOI: 10.3760/cma.j.issn.0254-6450.2011.02.018

作者单位: 100037 北京协和医学院中国医学科学院阜外心血管病医院(徐涛、李卫、陈涛); 中国医学科学院基础医学研究所北京协和医学院基础学院(徐涛)

通信作者: 李卫, Email: liwei0325@yahoo.com.cn

者曾用Poisson回归模型和负二项回归模型处理亚健康状态[5]，研究发现负二项分布的拟合优度较好，模型的Value/DF比值接近1，而且Poisson分布要求均数和方差相等，但是亚健康症状数的方差远大于均数，这种过度离散性将会低估回归参数的标准误。通过引入Gamma分布的误差项来构建负二项分布，以减小这种由于过度离散性产生的不利影响。故负二项分布比Poisson分布具有更好的适用性。

但是近半数的受试者的亚健康症状数为零，这些过多零频数的存在使得传统的负二项回归和Poisson回归模型的拟合效果受到影响，因为根据Poisson分布和负二项分布的概率密度函数所获得的零频数的概率均低于实际情况，在模型拟合的过程中，如果忽略这些零的存在，对参数的估计就会产生偏差。当样本中零频数过多，概率函数单调下降时，即$P(X=0)>P(X=1)>P(X=2)$等，零频数过多模型（ZI模型）是合适的[6-8]。本文将探讨零频数过多的Poisson回归（ZIP）模型和零频数过多的负二项回归（ZINB）模型在亚健康状态研究中的应用，并与传统的Poisson回归和负二项回归模型进行拟合效果比较，为零频数过多的医学资料的分析研究提供理论依据。

基本原理

ZI模型最初应用于经济统计领域[9]。Cheung[10]认为在医学领域中可将ZI模型估计一个两阶段的疾病进程，在开始阶段受试者并不处于风险之中，因此有0个事件，而某些协变量的影响使其成为有风险的人群，观察结果服从Poisson分布或负二项分布。

ZI模型可以同时考虑资料过度离散和零频数过多的问题。假设数据分为两部分：第一部分来源于那些从未可能发生事件的个体概率p_i（logit部分，常用的连接函数为logit函数），其结果解释和一般的logistic回归一样，说明协变量是否影响事件发生；第二部分为在Poisson或负二项分布下没有发生事件的个体概率$1-p_i$（Poisson或负二项部分），其结果解释与传统的Poisson或负二项回归一样，说明协变量影响事件发生多少的问题。两部分的零频数分别称为多余的零和抽样的零。

1. ZI模型：为了解释过多的零频数，在Poisson回归模型和负二项回归模型的基础上进行改进，建立ZI模型，使模型具备处理亚健康症状数多为0的能力。ZIP模型被定义为两种分布形式的复合体，以包含多余的零频数，该模型假定

$$P(y_i \mid X_i)=\begin{cases}\pi_i+(1-\pi_i)e^{-\mu} & y_i=0 \quad \text{logit 部分}\\ (1-\pi_i)\dfrac{e^{-\mu}\mu^{y_i}}{y_i!} & y_i\geqslant 1 \quad \text{Poisson 部分}\end{cases}$$

在ZIP模型条件下，y_i的条件期望和方差为

$$E(y)=\mu(1-p_i),\ \mathrm{Var}(y)=E(y)(1+\mu p_i)$$

ZINB模型也分为两部分，概率密度函数为

$$P(y_i \mid X_i)=\begin{cases}\pi_i+(1-\pi_i)\left(\dfrac{1}{1+\alpha\mu}\right)^{\alpha^{-1}} & y_i=0 \quad \text{logit 部分}\\ (1-\pi_i)\dfrac{\Gamma(y_i+\alpha^{-1})(\alpha\mu)^{y_i}}{y_i!\Gamma(\alpha^{-1})(1+\alpha\mu)^{(y_i+\alpha^{-1})}} & y_i\geqslant 1 \quad \text{负二项部分}\end{cases}$$

π_i为多余的0的概率。在ZINB模型条件下，y_i的条件期望和方差为

$$E(y)=\mu(1-p_i),\ \mathrm{Var}(y)=E(y)[1+\mu(p_i+\alpha)]$$

用极大似然法对ZI模型的参数进行估计及可信区间的检验。

2. 模型的选择：Poisson回归模型、负二项回归模型、ZIP模型和ZINB模型均用Stata 11.0软件进行拟合，并选用α系数、O检验、Vuong检验和比分检验（score test）进行过度离散性和零频数是否过多的判断，并用似然比检验和AIC对模型的拟合优度进行判定，以选择最优的模型。

过度离散性的判断可以用α系数和O检验。α为负二项分布的离散参数，用来衡量分布的离散程度，$\alpha=0$时，负二项分布即为Poisson分布，$\alpha\neq 0$时，说明事件的发生不是随机独立的，α值越大，分布的方差与均数的比值越大，当α显著>0时，表示数据过度离散。

O检验是一种基于均数和方差的进行过度离散性检验的方法，O统计量服从标准正态分布。

$$O=\frac{\sqrt{\frac{n-1}{2}}(s^2-\bar{x})}{\bar{x}}$$

零频数是否过多可以用Vuong检验来进行判定，即

$$V=\frac{\sqrt{n}\left[(1/n)\sum_{i=1}^{n}m_i\right]}{\sqrt{(1/n)\sum_{i=1}^{n}(m_i-\bar{m})^2}}=\frac{\sqrt{Nm}}{S_m}$$

式中$m_i=\log[f_1(y_i)/f_2(y_i)]$，$S_m$为标准差；$f_1$、$f_2$为ZI模型和传统模型的概率密度函数。$V$统计量服从标准正态分布，$V>1.96$则选择模型1，即ZI模型；$V<1.96$则选择模型2，即传统模型（Poisson回归和负二

项回归)。

似然比检验是以似然函数为基础所做的统计检验,极大似然法是对参数θ作点估计的方法之一,似然函数$L(\theta)=\prod_{i=1}^{n}P(X_i)$。似然比检验的统计量是两个模型的最大对数似然估计值之差的负二倍,$\chi^2_D=-2(\ln L_1-\ln L_2)$,在样本量较大的情况下,该统计量服从自由度为$(\alpha-j)$的$\chi^2$分布。$-2\ln L$越小,表明模型的拟合效果越好。AIC也可用于统计模型的拟合优度检验,$AIC=-2\ln(L)+2\alpha$。L是似然函数,α是可估计的自由参数的个数。对于同一组数据,AIC越小,表明模型的拟合效果越好。

实例分析

样本来源于在黑龙江省进行的大规模横断面调查,该项调查采用分层二阶段整群抽样的原则。以省作为主层,主层下分为城市和农村2个分层,每个分层各整群抽取若干个村或社区,以抽到的社区和村所有无心、肺、肝、脑、肾等主要脏器疾病者的10~80岁人群作为调查对象。亚健康采用陈青山等[1]制定的亚健康评价量表,该量表有6个维度、18项症状组成,即躯体症状5项(疲劳、头痛或头晕、耳鸣、肩或腿麻木僵硬、咽部有异物感)、心理症状6项(心烦意乱、孤独感、注意力不集中、焦虑、多梦和休息不好、记忆力差)、活力3项(活力减退、对周围事物不感兴趣、情绪差)、社会适应能力2项(工作觉得吃力、同事关系紧张)、免疫力1项(容易感冒或其他疾病)和到医院看病1项(感觉有病但未确诊)。在1年时间内持续1个月以上出现所列18种症状中任何一项者即认为该受试者具有1项亚健康症状,本文以受试者所具有的症状数作为研究指标。受试者的主要相关信息包括年龄、性别、职业(脑力和体力)、吸烟、饮酒、民族、婚姻状况、肥胖、血压高。共调查11 227名受试者。

对亚健康症状数的过度离散度检验,$\bar{x}=2.90$,$s=3.85$,过度离散统计量$O=308.011$,$P<0.001$,$s^2>\bar{x}$,表明存在过度离散。由图1可见,频数分布极度正偏态,其中43.3%的受试者(4863人)无阳性亚健康症状。无症状受试者平均年龄(25.3±17.9)岁显著低于有症状者(40.3±18.9)岁,$P<0.001$,其他情况见表1。

以亚健康症状数为应变量,以性别、年龄、婚姻状况、职业、睡眠质量、睡眠时间、吸烟、饮酒、血压高和肥胖为自变量,用Stata 11.0软件分别拟合Poisson

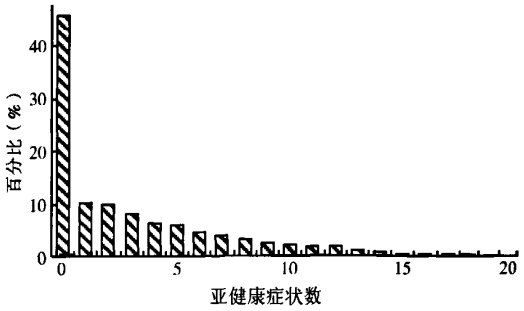

图1 亚健康症状数分布

表1 受试者的基本特征

症状	症状数=0		症状数≥1		χ^2值	P值
	人数	构成比(%)	人数	构成比(%)		
性别					109.224	<0.001
男	2265	49.2	2341	50.8		
女	2598	39.2	4023	60.8		
婚姻状况					1743.452	<0.001
已婚	1362	25.4	3991	74.6		
离婚或丧偶	116	20.0	465	80.0		
单身	3385	64.0	1908	36.0		
职业					999.410	<0.001
体力劳动	1041	24.4	3222	75.6		
脑力劳动	3822	54.9	3142	45.1		
吸烟					119.942	<0.001
是	566	31.6	1227	78.4		
否	4297	45.5	5137	54.5		
饮酒					196.992	<0.001
是	612	29.5	1461	70.5		
否	4251	46.4	4903	53.6		
肥胖					45.498	<0.001
是	379	33.8	741	66.2		
否	4484	44.4	5623	55.6		
血压高					270.774	<0.001
是	559	27.1	1504	72.9		
否	4304	47.0	4860	53.0		

回归模型、负二项回归模型、ZIP模型和ZINB模型,获得各模型的拟合优度指标见表2。从4个模型的拟合效果看,ZINB模型log likelihood最大,AIC最小,说明ZINB的拟合效果最佳。此外,ZINB似然比检验$\chi^2=6342.65$,$P<0.0001$;在ZINB模型中,离散系数$\alpha=1.013$,95%CI:0.965~1.063。可以认为离散系数α显著>0,数据过度离散。Vuong检验的统计量$Z=31.93$,$P<0.0001$,表明此数据的零频数过多,远远超出了传统负二项分布的估计范围。各项拟合效果的检验结果表明,ZINB模型是进行亚健康症状数研究的最优模型。

表3为ZINB模型的结果。ZINB的结果包括负二项部分和logit部分,负二项部分是针对资料中亚

表2　各模型的拟合指标

模型	log likelihood	AIC
Poisson回归	-34 879.937	69 781.87
负二项回归	-23 698.153	47 418.31
ZIP	-25 342.065	50 706.13
ZINB	-22 170.741	44 363.48

健康症状数>0的计数建立负二项回归模型，结果解释与传统的负二项回归模型一致；logit部分对数据中的零频数建立模型，结果解释与传统的logistic回归模型相似。从logit部分可见，年龄、职业、民族、婚姻状况和肥胖是否具有亚健康症状的影响因素，从回归系数来看，年龄越高、脑力劳动、民族为朝鲜族及肥胖是出现亚健康症状的危险因素，而单身者不容易受到亚健康症状的危害。此外，由负二项部分发现，年龄、性别、脑力劳动、饮酒和婚姻状况是亚健康症状数多少的影响因素，低年龄、女性、脑力劳动、不饮酒、离异或丧偶的受试者，其具有的亚健康症状数目越多。

同时用传统的负二项回归模型和logistic回归模型对亚健康症状数和是否具有亚健康症状两个因变量分别进行模型拟合，由表4结果看，在传统的负二项回归模型中亚健康症状数的影响因素为年龄、女性、脑力劳动、吸烟、饮酒、朝鲜族和单身；而logistic回归模型显示年龄、女性、血压高、脑力劳动、吸烟、饮酒、朝鲜族和单身是否具有亚健康症状的危险因素。

讨　论

ZI模型在医学领域的应用还不太广泛，但是零频数过多的计数资料确是医学研究中常遇到的资料类型，如本文所研究的亚健康症状数，还有尿失禁次数、日常生活功能受损次数、流产次数及临床试验中

表3　ZINB模型的参数估计及假设检验

指标	负二项部分					logit部分				
	系数	Z	P>Z	95%CI		系数	Z	P>Z	95%CI	
年龄	-0.003	-2.610	0.009	-0.006	-0.001	-0.436	-20.860	0.000	-0.477	-0.395
性别(女)	0.280	8.160	0.000	0.213	0.348	-0.050	-0.530	0.596	-0.235	0.135
血压高	-0.004	-0.120	0.902	-0.070	0.062	-0.118	-0.440	0.660	-0.642	0.406
脑力劳动	0.076	2.500	0.012	0.016	0.136	-1.527	-4.320	0.000	-2.219	-0.834
吸烟	-0.037	-0.920	0.360	-0.117	0.042	-0.096	-0.450	0.653	-0.515	0.323
饮酒	-0.098	-2.640	0.008	-0.171	-0.025	-0.039	-0.190	0.848	-0.433	0.356
朝鲜族	0.034	1.010	0.311	-0.031	0.099	-2.253	-11.670	0.000	-2.632	-1.874
其他民族	0.052	0.770	0.444	-0.080	0.184	0.032	0.160	0.871	-0.358	0.423
离异或丧偶	0.200	3.920	0.000	0.100	0.300	-13.106	-0.010	0.994	-3206.1	3179.9
单身	0.076	1.630	0.103	-0.015	0.167	2.175	2.730	0.006	0.615	3.734
肥胖	0.013	0.320	0.752	-0.066	0.092	-0.436	-20.860	0.000	-0.477	-0.395
常数项	1.113	11.980	0.000	0.931	1.295	-0.050	-0.530	0.596	-0.235	0.135
α	1.013			0.965	1.063					

表4　传统负二项回归模型和logistic回归模型的参数估计及假设检验

指标	负二项模型					logistic回归模型				
	系数	Z	P>Z	95%CI		系数	Z	P>Z	95%CI	
年龄	0.004	2.250	0.025	0.000	0.007	0.015	6.090	0.000	0.010	0.020
性别(女)	0.373	10.270	0.000	0.302	0.444	0.461	9.090	0.000	0.361	0.560
血压高	-0.082	-1.920	0.055	-0.166	0.002	-0.168	-2.480	0.013	-0.300	-0.035
脑力劳动	0.200	5.410	0.000	0.128	0.273	0.392	6.830	0.000	0.280	0.505
吸烟	-0.192	-3.960	0.000	-0.287	-0.097	-0.235	-3.220	0.001	-0.378	-0.092
饮酒	-0.245	-5.520	0.000	-0.332	-0.158	-0.451	-6.680	0.000	-0.584	-0.319
朝鲜族	0.278	6.900	0.000	0.199	0.356	0.428	6.810	0.000	0.305	0.552
其他民族	0.094	1.270	0.205	-0.051	0.238	0.165	1.570	0.116	-0.041	0.372
离异或丧偶	0.083	1.240	0.216	-0.049	0.215	-0.002	-0.020	0.985	-0.226	0.222
单身	-0.427	-7.990	0.000	-0.532	-0.322	-0.880	-10.770	0.000	-1.040	-0.720
肥胖	-0.029	-0.570	0.566	-0.126	0.069	-0.065	-0.870	0.384	-0.213	0.082
常数项	0.956	8.120	0.000	0.725	1.186	0.191	1.050	0.292	-0.164	0.546
α	1.933			1.859	2.010					

不良事件发生数目等。现在绝大多数研究者还是用传统的Poisson回归、负二项回归，甚至以logistic回归模型来处理此类数据，这些不恰当的模型很难得到符合实际的研究结果。

负二项分布和Poisson分布是最常用于计数资料的两种分布形式，而负二项分布能够更好地处理分布中的异质性，以解决数据过度离散的问题。然而当资料中含有过多的零频数时，该两种方法都不能研究这种不规则的分布形式。ZI模型是探讨零频数过多计数资料的最优模型，同时对二项分布和负二项分布或Poisson分布部分拟合模型，并综合考虑是否发生及发生对结局的影响。实际上是将疾病进程看作两个阶段，第一阶段是某些因素使受试者处于某种风险之中，第二阶段探讨处于风险之中的受试者发生强度的影响因素。两个阶段可能受到不同因素或不同机制的影响[6]。本文正是基于这个原则得到了两个回归模型，分别解释发生亚健康症状及其症状发生多少两方面的影响因素。如果把亚健康仅仅作为一个二分类变量进行logistic回归分析，则“血压高”和“单身”也是一个危险因素；传统的负二项回归分析结果认为“吸烟”和“朝鲜族”也能影响到亚健康症状数的多少。而这些变量在ZINB模型中，如果只用传统的回归模型获得假阳性结果，并未显示对亚健康症状数有显著影响。

总之，各种拟合优度检验及对过度离散和零频数过多的检验结果，均显示ZINB模型在处理亚健康症状数方面的优势，为ZINB模型在零频数过多的计数资料应用上提供了理论依据。

参 考 文 献

[1] Chen QS, Wang SY, Jing CX, et al. Evaluation on diagnostic criterion of sub-health with Delphi method. Chin J Public Health, 2003, 19(12): 1467-1468. (in Chinese)
陈青山，王声湧，荆春霞，等. 应用Delphi法评价亚健康的诊断标准. 中国公共卫生，2003，19(12)：1467-1468.

[2] Nie XL, Xue Q, Lai MH, et al. Sub-health and its influential factors among civil servants in taxation department. Chin J Public Health, 2010, 26(5): 634-635. (in Chinese)
聂晓莉，薛琪，赖名慧，等. 税务部门公务员亚健康现况及影响因素分析. 中国公共卫生，2010，26(5)：634-635.

[3] Fan CX, Wang SY, Zhu L, et al. Study on the prevalence and risk factors of "sub-health" status in college and university staff in Guangdong province. Chin J Epidemiol, 2003, 24(9): 774-777. (in Chinese)
范存欣，王声湧，朱丽，等. 广东省高校教工亚健康现况及危险因素分析. 中华流行病学杂志，2003，24(9)：774-777.

[4] Zhu Li, Wang SY, Fan CX, et al. Logistic regression analysis of risk factors on sub-health of young teachers in institution of higher learning. Chin J Public Health, 2003, 19(5): 595-596. (in Chinese)
朱丽，王声湧，范存欣，等. 高校青年教师亚健康危险因素logistic回归分析. 中国公共卫生，2003，19(5)：595-596.

[5] Xu T, Han SM, Liu JT, et al. Comparison of sub-health status between Tiban people and Han people. Chin J Med, 2009, 89(38): 2671-2674. (in Chinese)
徐涛，韩少梅，刘军廷，等. 四川省汉族与藏族人群亚健康状态现患率和症状数的比较. 中华医学杂志，2009，89(38)：2671-2674.

[6] Solinas G, Campus G, Maida C, et al. What statistical method should be used to evaluate risα factors associated with dmfs index? Evidence from the National Pathfinder Survey of 4-year-old Italian Children. Community Dent Oral Epidemiol, 2009, 37(6): 539-546.

[7] Aipnis V, Midthune D, Bucaman DW, et al. Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. Biometrics, 2009, 65(4): 1003-1010.

[8] Wang H, Heitjan DF. Modeling heaping in self-reported cigarette counts. Stat Med, 2008, 27(19): 3789-3804.

[9] Mullahy J. Specification and testing of some modified count data models. J Economet, 1986, 33(3): 341-365.

[10] Cheung YB. Zero-inflated models for regression analysis of count data: a study of growth and development. Stat Med, 2002, 21(10): 1461-1469.

(收稿日期：2010-09-21)
(本文编辑：张林东)