零膨胀有序 logit 模型和传统模型在零膨胀等级资料分析中的比较研究*

复旦大学公共卫生学院流行病学教研室,公共卫生安全教育部重点实验室(200032)
黎 怡 张 欣 樊 虹 徐艺耘 吴 声 张铁军△

【提 要】 **目的** 对零膨胀有序 logit 模型展开应用并和传统模型进行比较,探讨其优势和局限性。**方法** 对各自变量进行单因素分析并计算相应的对数似然值、AIC 及 BIC 等模型评价指标,根据这些指标结果依次纳入各自变量并组成不同的自变量集。分别使用有序 logit 模型、广义有序 logit 模型和零膨胀有序 logit 模型基于不同自变量集进行数据拟合,通过比较不同情况下各模型的评价指标及参数估计结果对模型进行评价。**结果** 根据模型评价指标,零膨胀有序 logit 模型对数据的拟合效果优于其他两种模型。零膨胀有序 logit 模型对不同自变量集进行拟合时,其参数估计结果高度一致。不同模型的参数估计结果存在一定的差异。**结论** 零膨胀有序 logit 模型在应用时有着较好的性能。单因素分析的模型评价指标可以用来帮助筛选使得模型性能最优的自变量集。

【关键词】 零膨胀有序 logit 模型 等级资料 Stata 软件
【中图分类号】 R181.2 【文献标识码】 A **DOI** 10.11783/j.issn.1002-3674.2024.05.008

Comparison on the Effects of the Zero-inflated Ordered Logit Model and Traditional Models for Zero Inflation Hierarchical Data

Li Yi, Zhang Xin, Fan Hong, et al (*Department of Epidemiology, School of Public Health, Fudan University, Shanghai* 200032)

【**Abstract**】 **Objective** Apply the Zero-inflated ordered logit model and explore its advantages and limitations by comparing it with the traditional models. **Methods** Univariate analysis of each independent variable was performed, and the corresponding model evaluation indicators such as log likelihood, AIC, and BIC were calculated. According to the results of these indicators, the respective variables were incorporated in turn and formed different sets of independent variables. The ordered logit model, generalized ordered logit model, and Zero-inflated ordered logit model were used to fit the data based on these different sets of independent variables, respectively. The models were evaluated by comparing the model evaluation indicators and parameter estimation results. **Results** According to the model evaluation indicators, the Zero-inflated ordered logit model performed better than the traditional models. The parameter estimation results of the Zero-inflated ordered logit model based on different sets of independent variables are highly consistent. There are differences in the parameter estimation results of different models. **Conclusion** The Zero-inflated ordered logit model has good performance in application. Model evaluation indicators from univariate analyses can be used to help filter the set of independent variables that make the model's performance optimal.

【**Key words**】 Zero-inflated ordered logit model; Hierarchical data; Stata software

等级资料是临床和公共卫生研究中经常遇到的一类数据,常用的分析方法为有序 logit 模型或广义有序 logit 模型。等级资料中响应变量的零值往往代表着样本不具备某个特征,剩余的非零值代表该特征水平的增加。当资料涉及异常行为、症状或副作用时,由于这些事件的发生率较低,可能造成异常多的“零”数据存在,称为零膨胀现象。零膨胀数据中的零值比例远高于传统模型的预期,若继续采用传统模型进行处理,得出的结论可能偏于真实情况。

零膨胀模型考虑到了对数据中零观测值的处理,在具有零膨胀问题资料的分析上有着独到的优势。零膨胀模型最初由 Lambert[1] 开发,目的是解决计数资料中零过多的问题。通过 Kelley 和 Anderson[2] 开发的零膨胀有序 logit 模型(Zero-inflated ordered logit model, ZIOL),这一模型框架已经扩展到等级资料。目前,用于处理计数资料的零膨胀模型已经得到广泛应用[3-4],而对于 ZIOL 模型的研究和应用尚且较少。因此,本文旨在对 ZIOL 模型展开应用,并通过和传统模型进行比较,探讨其优势和局限性。

原理和方法

ZIOL 模型认为数据中的零观测值来源于两部分:一部分是数据中存在某些特殊结构而产生的“结构零”,即个体本身没有发生所调查现象的特质,这些个体对应于一个始终为零的总体;另一部分是考虑到抽样的不确定性和测量误差而产生的“抽样零”,指个体有发生所调查现象的特质,但在数据收集期间并未表现出来。“抽样零”和数据中的非零观测值共同构成了一个服从多项式分布的总体。

ZIOL 模型属于两部分模型:首先使用 logit 模型区分响应变量的零值为“结构零”还是“抽样零”,该过程所生成的定量关系式被称为膨胀方程(inflation equa-

*基金项目:国家自然科学基金(81772170);国家重点研发计划重点专项(2017YFC0211704)
△通信作者:张铁军,E-mail: tjzhang@ shmu.edu.cn

tion)。接着将“抽样零”和剩余非零个体纳入有序 logit 模型以进行决定响应变量等级的回归,相应的定量关系式被称为强度方程(intensity equation)。两个方程可以有不同的协变量集。ZIOL 模型开发过程如下:

记响应变量为 Y,Y 等级编码为 0,1,2,…,H。如果第 j 个个体属于“抽样零”,则设 $s_j=1$;若个体属于“结构零”,则 $s_j=0$。$s_j=1$ 的概率为:

$$\Pr(s_j=1\mid z_j)=F(z_j\gamma) \tag{1}$$

z_j 是自变量组成的向量,γ 是 z_j 的系数,$F(\,)$ 是 logistic 函数:$F(x)=\dfrac{e^x}{1+e^x}$。

在 $s_j=1$ 的条件下,使用有序 logit 模型对响应变量等级 $\tilde{y}_j$ 进行建模。相应的概率由下式给出:

$$\Pr(\tilde{y}_j=h\mid s_j=1,x_j)=F(\kappa_h-x_j\beta)-F(\kappa_{h-1}-x_j\beta)\quad h=1,2,\cdots,H \tag{2}$$

x_j 是确定 $\tilde{y}_j$ 的自变量组成的向量(可以不同于 z_j),β 是 x_j 的系数,切割点 κ_h 是要估计的边界参数(服从 $\kappa_{-1}=-\infty$,$\kappa_H=+\infty$)。

观察到的响应变量 $y_j=s_j\tilde{y}_j$。因此,当①$s_j=0$ 或②$s_j=1$ 且 $\tilde{y}_j=0$ 时,则会出现“零值”。当 $s_j=1$ 且 $\tilde{y}_j>0$ 时,才能观察到 $y_j\neq0$。

综上,响应变量 Y 的分布由下式给出:

$$\begin{aligned}\Pr(Y)&=\begin{cases}\Pr(y_j=0\mid z_j,x_j)\\ \Pr(y_j=h\mid z_j,x_j)\ h=1,2,\cdots,H\end{cases}\\&=\begin{cases}\Pr(s_j=0\mid z_j)+\Pr(s_j=1\mid z_j)\Pr(\tilde{y}_j=0\mid s_j=1,x_j)\\ \Pr(s_j=1\mid z_j)\Pr(\tilde{y}_j=h\mid s_j=1,x_j)\ h=1,2,\cdots,H\end{cases}\end{aligned} \tag{3}$$

将(1),(2)代入(3)式,得:

$$\begin{aligned}\Pr(Y)&=\begin{cases}\Pr(y_j=0\mid z_j,x_j)\\ \Pr(y_j=h\mid z_j,x_j)\ h=1,2,\cdots,H-1\\ \Pr(y_j=H\mid z_j,x_j)\end{cases}\\&=\begin{cases}\{1-F(z_j\gamma)\}+F(z_j\gamma)F(\kappa0-x_j\beta)\\ F(z_j\gamma)\{F(\kappa_h-x_j\beta)-F(\kappa_{h-1}-x_j\beta)\}\ h=1,2,\cdots,H-1\\ F(z_j\gamma)\{1-F(\kappa_{H-1}-x_j\beta)\}\end{cases}\end{aligned} \tag{4}$$

2.有序 logit 模型和广义有序 logit 模型

有序 logit 模型要求自变量系数在各有序等级中始终保持一致,即参数值不随响应变量等级的变化而变化,结果也只输出一组自变量的系数[5]。但是,现实中参数固定的假设通常难以成立。因此,在应用有序 logit 模型时,须对自变量系数相等的假设进行检验,即平行性检验[6]。本研究采用 Brant 检验来判断数据是否满足有序 logit 模型的前提条件,$P<0.05$ 表明不满足。广义有序 logit 模型是在有序 logit 模型基础上的改进模型,具有有序 logit 模型因变量特有的有序性,一般是在数据不满足平行性检验时应用[7]。和有序 logit 模型相比,广义有序 logit 模型兼顾自变量对因变量的影响,随着潜变量阈值的变化而不同,适用性更强[8]。广义有序 logit 模型设定可由以下公式表示:

$$\Pr(Y_i>j)=g(X_i\beta_j)=\frac{\exp(\alpha_j+X_i\beta_j)}{1+\exp(\alpha_j+X_i\beta_j)} \tag{5}$$

式中:Y_i 为可观测的刻度变量,即响应变量;j 为响应变量等级;X_i 是自变量向量;β_j 是第 j 水平等级的自变量回归系数向量;α_j 是第 j 水平等级的常数项。若 Y_i 有 4 个等级,则对于 $j=1$,则将 $Y_i=1$ 与 $Y_i=2$、3、4 进行比较;对于 $j=2$,则将 $Y_i=1$、2 与 $Y_i=3$、4 进行比较;对于 $j=3$,则将 $Y_i=1$、2、3 与 $Y_i=4$ 进行比较。因此,模型最终将会输出 3 组参数。如果在每个类别区间内对应的 β_j 相等,则广义有序 logit 模型可以简化为有序 logit 模型。

3.模型比较

采用对数似然值(log likelihood of full model, Model LL)[9]、赤池信息准则(Akaike information criterion, AIC)[10] 和贝叶斯信息准则(Bayesian information criterion, BIC)[11] 来判断和比较模型优劣。Model LL 适用于纳入相同预测变量的不同模型之间的比较,值越高表明模型对数据集的拟合程度越高。AIC 和 BIC 则通过引入惩罚项,从信息的视角兼顾衡量了模型的解释能力和模型的复杂度,避免模型过于复杂,从而造成过拟合现象。信息准则值越小,代表模型越优。AIC 及 BIC 的计算公式如下:

$$AIC=-2\ln(L)+2K \tag{6}$$

$$BIC=\ln(n)K-2\ln(L) \tag{7}$$

其中,$\ln(L)$ 为对数似然值,对应于本研究中的 Model LL;K 为参数数量;n 为样本量。一般而言,当纳入的变量增加时,对数似然值也会增大,从而使 AIC 或 BIC 变小,然而纳入的变量数量过多会导致 K 过大,其带来的影响超过对数似然值增长的幅度,导致 AIC 或者 BIC 增大。由公式可知,BIC 考虑了样本含量,惩罚项比 AIC 的大,因此其对参数数量的变化更敏感。

实例分析

研究数据来源于 1988—1994 年美国第三次全国健康和营养调查(the third national health and nutrition examination survey, NHANES Ⅲ)(https://wwwn.cdc.gov/nchs/nhanes/nhanes3/Default.aspx)。NHANES Ⅲ 数据库中包含了 20050 例成年(年龄≥17)人口数据,排除 8597 例性伴数信息缺失的样本及 1034 例其他信息缺失的样本,最终纳入 10419 名研究对象。

纳入的变量包括人口统计学特征(性别、年龄、种

族/民族、婚姻状况、受教育年限、贫困收入比、职业、地区)及被调查者过去一年的性伴数。贫困收入比(poverty income ratio, PIR)是衡量收入水平的指标,该指数的值越大,表明家庭经济状况越好。因变量“多性伴行为”基于研究对象自我报告的过去一年的性伴数,将其划分为 4 个等级(1:性伴数为 0~1;2:性伴数为 2~5;3:性伴数为 6~10;4:性伴数大于 10)。等级 1 表示“过去一年没有发生多性伴行为”;等级越高表明多性伴行为的严重程度越高。所有数据分析均采用 Stata 软件 17.0 版本。绘图采用 R 软件 4.1.3 版本。

研究对象过去一年性伴数的取值分布如图 1 所示。结果显示,报告没有多性伴行为的比例为 83.03%,表明数据是典型的零膨胀数据。

首先针对各自变量进行单因素分析,并计算相应的模型评价指标。如图 2 所示,“婚姻状况”变量对应的各指标值最优。根据各变量的模型评价指标结果大小,按照顺序从左往右进行变量累加并构建不同的自变量集,之后分别采用有序 logit 模型、广义有序 logit 模型和 ZIOL 模型对不同自变量集进行数据拟合。针对不同自变量集的各模型评价指标如图 3 所示。

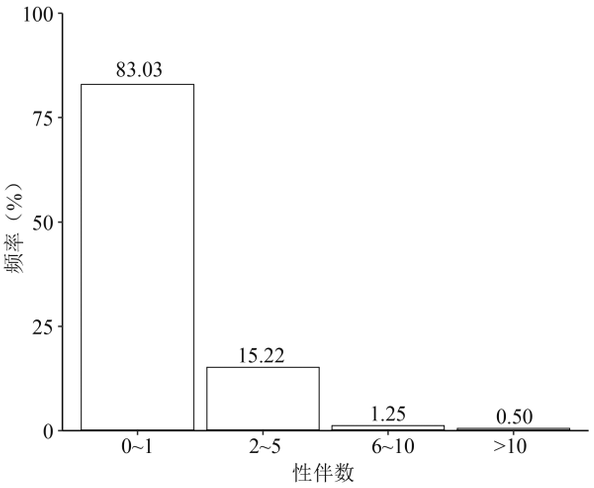
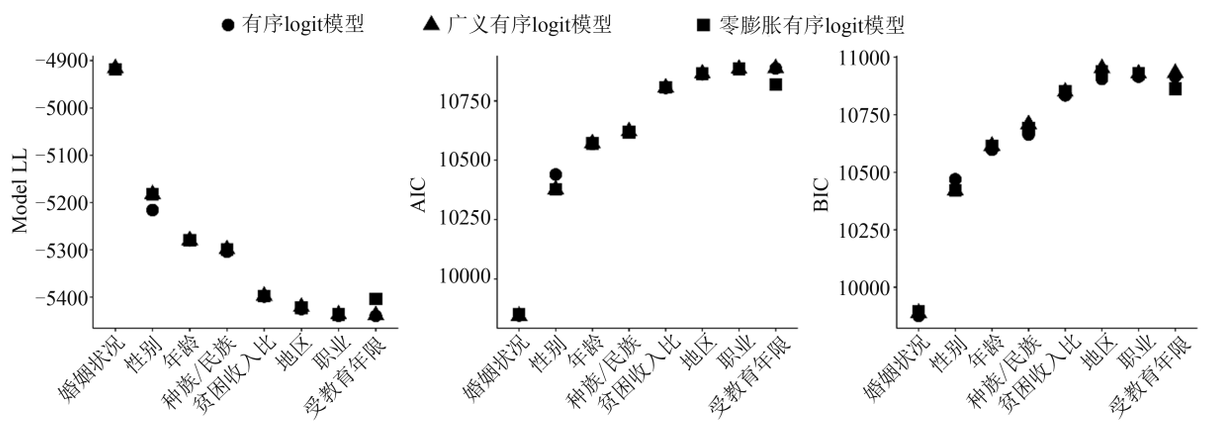


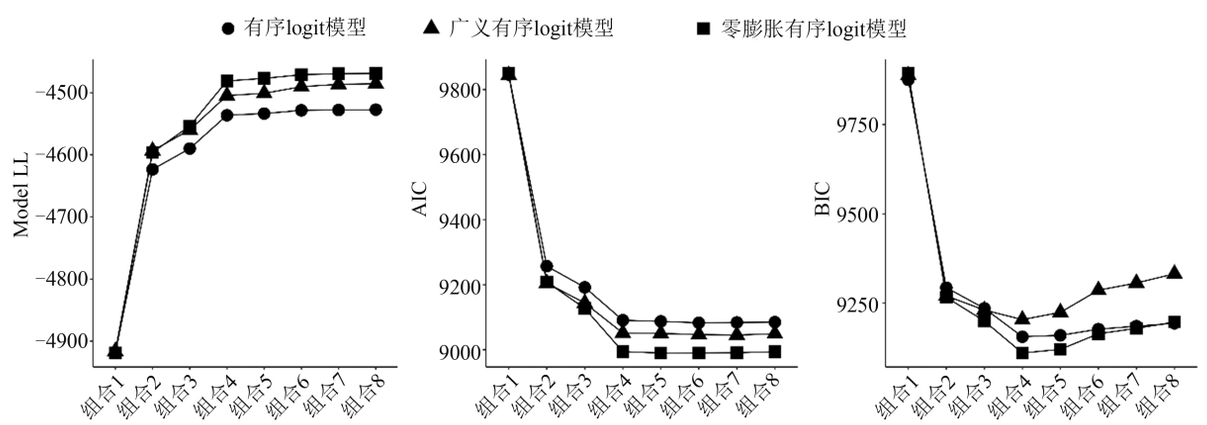
图 1 过去一年中调查对象性伴数分布情况

根据 Model LL 和 AIC 指标,ZIOL 模型对原始数据的拟合效果总是优于有序 logit 模型和广义有序 logit 模型(图 3)。对于 BIC 指标,当纳入的变量数目较多时,各模型的 BIC 值增大,其中 ZIOL 模型的 BIC 值增长幅度大于有序 logit 模型。基于变量组合 8 的结果显示,ZIOL 模型的 BIC 值稍高于有序 logit 模型(图 3)。



注:各变量基于有序 logit 模型的评价指标结果的大小从左到右进行排序,越往右表明对应变量的有序 logit 模型表现越差。

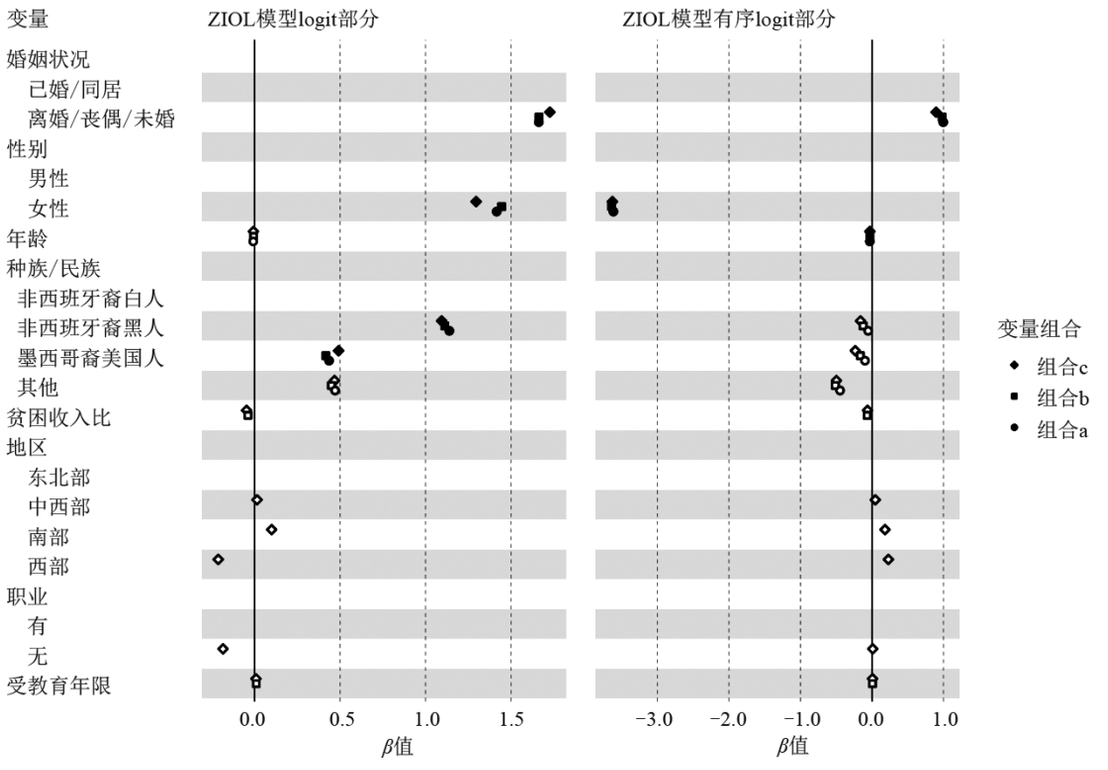
图 2 单因素分析中各模型评价指标结果



注:各组合所包含的变量基于图 2 的结果,按顺序进行变量累加。即组合 1:婚姻状况;组合 2:婚姻状况、性别;组合 3:婚姻状况、性别、年龄;组合 4:婚姻状况、性别、年龄、种族/民族;组合 5:婚姻状况、性别、年龄、种族/民族、贫困收入比;组合 6:婚姻状况、性别、年龄、种族/民族、贫困收入比、地区;组合 7:婚姻状况、性别、年龄、种族/民族、贫困收入比、地区、职业;组合 8:婚姻状况、性别、年龄、种族/民族、贫困收入比、地区、职业、受教育年限。

图 3 不同自变量组合下各模型的拟合效果比较

本研究还比较了基于不同变量集的 ZIOL 模型的参数估计值，结果显示很高的一致性（图 4）。其中，基于组合 c（包含所有自变量）的 ZIOL 模型及其他两种模型的参数估计结果如表 1 所示。为了提高可解释性，计算对应的优势比（odds ratio，OR）。



注：根据 ZIOL 模型评价指标结果的大小选择不同的自变量组合（参考图 2）。组合 a：婚姻状况、性别、年龄、种族/民族；组合 b：婚姻状况、性别、年龄、种族/民族、贫困收入比、受教育年限；组合 c：婚姻状况、性别、年龄、种族/民族、贫困收入比、受教育年限、地区、职业。图中散点实心表示相应自变量与因变量之间的关联有统计学意义（$P<0.05$），空心则表示无统计学意义。

图 4　基于不同自变量组合的零膨胀有序 logit 模型参数估计结果比较

表 1　不同模型 *OR* 值比较

	有序 logit 模型	广义有序 logit 模型			零膨胀有序 logit 模型	
		等级 1 vs 2、3、4	等级 1、2 vs 3、4	等级 1、2、3 vs 4	logit	有序 logit
婚姻状况						
已婚/同居	1.00	1.00	1.00	1.00	1.00	1.00
离婚/丧偶/未婚	5.48**	5.43**	5.13**	14.67**	5.62**	2.45**
性别						
男性	1.00	1.00	1.00	1.00	1.00	1.00
女性	0.24**	0.25**	0.04**	0.04**	3.65*	0.03**
年龄（岁）	0.98**	0.98**	0.98*	0.98	0.99	0.97**
种族/民族						
非西班牙裔白人	1.00	1.00	1.00	1.00	1.00	1.00
非西班牙裔黑人	1.87**	1.83**	2.49**	3.19*	2.99**	0.85
墨西哥裔美国人	1.26*	1.27*	1.29	1.54	1.63**	0.79
其他	1.04	1.02	1.82	2.33	1.60	0.61
贫困收入比	0.95*	0.95*	0.93	0.84	0.95	0.94
地区						
东北部	1.00	1.00	1.00	1.00	1.00	1.00
中西部	1.08	1.08	0.94	1.54	1.01	1.05
南部	1.22*	1.21*	1.39	2.32	1.10	1.20
西部	0.99	0.96	1.70	1.65	0.81	1.26
职业						
有	1.00	1.00	1.00	1.00	1.00	1.00
无	0.94	0.91	1.32	0.91	0.83	1.01
受教育年限	1.01	1.01	1.00	1.03	1.01	1.01

*：$P<0.05$；**：$P<0.001$

讨　论

零膨胀现象是研究中常见的一类问题,过多零值所带来的潜在影响不可忽视。本研究比较了纳入不同自变量集情况下 ZIOL 模型和两种传统模型对零膨胀等级资料的拟合效果。总体上,ZIOL 模型对数据的拟合效果优于其他两种传统模型,反映了其在处理零膨胀等级资料时稳定优良的性能。

本研究基于单因素分析中各个变量的模型评价指标的优劣将变量依次纳入模型组成不同的自变量集,并分别计算了相应的模型评价指标。这种方法类似于传统的向前逐步回归,即每次选择一个自变量进入模型,计算 AIC 或 BIC,然后再次引入另一个自变量并计算相应指标。若新变量和原始变量共线性很低并且对于最终的预测具有一定的贡献度,模型拟合优度上升。若新变量的贡献度很小,不足以覆盖由于参数数量增大导致模型复杂度增高所带来的影响,反而会导致模型性能降低。图 3 对这个过程进行了可视化。值得注意的是,当变量纳入增多(尤其是对最终预测不具显著贡献度的变量),ZIOL 模型 BIC 值的增长幅度超过了有序 logit 模型,甚至出现了 ZIOL 模型 BIC 值高于有序 logit 模型的情况。这可能是由于 BIC 指标受模型复杂度的影响较大。在相同自变量的情况下,ZIOL 模型估计的参数数量接近于有序 logit 模型的两倍,因此随着更多变量的引入,ZIOL 模型复杂度的增加程度高于有序 logit 模型,从而导致 ZIOL 模型的 BIC 值上升较快。尽管基于“组合 8”的 ZIOL 模型 BIC 指标值稍高于有序 logit 模型,但从 Model LL 和 AIC 指标来看,ZIOL 模型的拟合效果依然优于其他两种模型。

对于 ZIOL 模型来说,根据模型评价指标的变化趋势,本研究中具有显著预测作用的变量为“婚姻状况”、“性别”、“年龄”及“种族/民族”,ZIOL 模型参数估计结果及其一致性也印证了这点,表明在考虑了零膨胀的影响后,这 4 个变量可以有效预测响应变量。不同于有序 logit 模型,目前相应的操作命令暂不支持 ZIOL 模型基于逐步回归方法对相关变量进行筛选。在实际分析中,可以根据具体目标纳入合适的自变量,若是以构建预测模型为目的,可以参考本研究中的方法,即基于单因素分析中各个变量的模型评价指标,筛选出使模型拟合优度最佳的变量集。

本研究还比较了不同模型的参数估计结果,结果表明 ZIOL 模型与其他两种模型存在着不一致。例如,本研究中 ZIOL 模型显示女性更容易成为发生多性伴行为的“易感者”,只是在那些易发生多性伴行为的人中,女性的严重程度相比于男性更低。此外,有序 logit 模型认为经济状况是多性伴行为的影响因素,而 ZIOL 模型的结果提示无统计学意义,与其他研究一致[12]。由于大多数研究在探讨多性伴行为时只是单纯的将其作为二分类变量处理,并且尚缺乏 ZIOL 模型在此领域的应用,因此对于本研究结果的解释有待进一步探索。然而,从模型的评价指标可以看出,ZIOL 模型的表现优于有序 logit 模型及广义有序 logit 模型,表明其可能更接近真实情况。

ZIOL 模型也有一定的局限性。ZIOL 模型属于一维零膨胀模型,适用于单项目零膨胀等级资料,而无法处理彼此相关的多项目零膨胀等级资料。针对此问题,目前已有研究在其基础上进行改进并创建出多维零膨胀分级反应模型,以适应此类情况[13]。此外,研究[2]指出样本量大小对模型拟合和结果误差具有一定影响,ZIOL 模型对于小样本量数据的应用效果尚需进一步探索。

除了多性伴行为外,研究工作中还存在许多具有零膨胀问题的等级资料,例如药物的不良反应等。随着零膨胀现象越来越受到关注,适用于等级资料的零膨胀模型有着广阔的应用前景。本研究对 ZIOL 模型展开了初步应用和探讨,可为将来该模型的应用提供一定的参考价值。

参 考 文 献

[1] Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing[J]. Technometrics, 1992, 34(1): 1-14.

[2] Kelley ME, Anderson SJ. Zero inflation in ordinal data: incorporating susceptibility to response through the use of a mixture model[J]. Statistics in Medicine, 2008, 27(18): 3674-3688.

[3] 吴学福, 刘振球, 吴明山, 等. 零膨胀计数资料几种模型方法的比较研究[J]. 中国卫生统计, 2020, 37(3): 331-334.

[4] 刘振球, 严琼, 左佳鹭, 等. 零膨胀计数数据回归模型的选择与比较及 R 语言的实现[J]. 中国卫生统计, 2018, 35(2): 310-312.

[5] Mckelvey RD, Zavoina W. A statistical model for the analysis of ordinal level dependent variables[J]. Journal of Mathematical Sociology, 1975, 4(1): 103-120.

[6] 赵晋芳, 范月玲, 曾平, 等. 多分类有序 logit 模型资料平行线假设及检验方法[J]. 中国卫生统计, 2009, 26(1): 11-13.

[7] Williams R. Generalized ordered logit/partial proportional odds models for ordinal dependent variables[J]. The Stata Journal, 2006, 6(1): 58-82.

[8] Williams R. Understanding and interpreting generalized ordered logit models[J]. The Journal of Mathematical Sociology, 2016, 40(1): 7-20.

[9] Myung IJ. Tutorial on maximum likelihood estimation[J]. Journal of Mathematical Psychology, 2003, 47(1): 90-100.

[10] Akaike H. A new look at the statistical model identification[J]. IEEE transactions on automatic control, 1974, 19(6): 716-723.

[11] Schwarz G. Estimating the dimension of a model[J]. The Annals of Statistics, 1978, 461-464.

[12] Ali MM, Merdad L, Bellizzi S. Socioeconomic variations in risky sexual behavior among adolescents in 14 sub-Saharan Africa countries who report ever having had sex[J]. International Journal for Equity in Health, 2021, 20(1): 11.

[13] Magnus BE, Garnier-villarreal M. A Multidimensional Zero-Inflated Graded Response Model for Ordinal Symptom Data[J]. Psychological Methods, 2022, 27(2): 261-279.

(责任编辑:邓　妍)