非配对设计多值名义资料多水平多重Logistic回归模型

李长平[1,2*]，张甜甜[1]，宋德胜[1]，胡良平[2,3]

（1. 天津医科大学公共卫生学院卫生统计学教研室，天津 300070；
2. 世界中医药学会联合会临床科研统计学专业委员会，北京 100029；
3. 军事科学院研究生院，北京 100850
*通信作者：李长平，E-mail：1067181059@qq.com）

【摘要】 本文目的是介绍非配对设计多值名义资料多水平多重logistic回归模型的构建与求解方法。首先介绍了有关的基本概念，涉及“多值名义结果变量”“分层或多水平数据结构”和“扩展的多重logistic回归模型”；其次，呈现了一个具有二水平结构的横断面调查资料，该资料涉及多个影响因素和一个多值有序的结果变量（在本文中，将其视为多值名义结果变量）；最后，借助SAS中的两个过程（即GLIMMIX和NLMIXED）对给定的资料进行统计分析，即构建和求解“非配对设计多值名义资料多水平多重logistic回归模型”，并对相关结果进行比较和解释。

【关键词】 多值名义资料；多水平；SAS软件；多重logistic回归分析
中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20191119006

Multilevel multiple Logistic regression analysis with the multi-value nominal data collected from the unpaired design

Li Changping[1,2*], *Zhang Tiantian*[1], *Song Desheng*[1], *Hu Liangping*[2,3]

（1. *Department of Health Statistics, School of Public Health, Tianjin Medical University, Tianjin* 300070, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*;
3. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*
**Corresponding author: Li Changping, E-mail: 1067181059@qq.com*）

【Abstract】 The purpose of the paper was to introduce the construction and solution of multilevel multiple logistic regression analysis with the multi-value nominal data collected from the unpaired design. Firstly, the basic concepts related to “multi-value nominal outcome variables” “stratified or multilevel data structures” and “generalized multiple logistic regression models” were introduced. Secondly, a cross-sectional survey data with two-level was presented. The data contained many independent variables and a multi-value ordinal outcome variable (in this paper, it was treated as a multi-value nominal outcome variable). Finally, statistical analysis of data was performed by two procedures (GLIMMIX and NLMIXED) in the SAS software. Construction and solution of multilevel multiple logistic regression analysis with the multi-value nominal data collected from the unpaired design was preformed and the related output results were compared and explained.

【Keywords】 Multi-value nominal data; Multilevel; SAS software; Multiple logistic regression analysis

1 基本概念

当反应变量为无序离散型结局变量且具有3个或更多类别时，这样的结局变量称为多值名义结果变量。在众多因素中，欲探讨哪些因素对多值名义结果变量具有统计学意义的影响，可采用多值名义资料多重logistic回归模型进行分析。该模型是二值资料多重logistic回归模型的扩展。如本刊上一期科研方法专题已发表的文章[1]所述，多水平数据具有非独立性，需要采用能处理多水平数据的多水平模型。

对于一个具有M个类别的名义结果变量的多水平资料，可采用广义线性混合效应模型进行分析。该模型是二值结果变量多水平logistic回归模型的扩展。该模型将构建（M-1）个logistic回归模型，且估计（M-1）组参数[2]。

2 问题与数据结构

【例1】以美国国家毒品滥用研究所开展的一项以社区为基础的艾滋病干预研究数据[1]为例。共收集20个调查点，包含9 824名静脉注射吸毒者的基

项目基金：国家高技术研究发展计划课题资助（2015AA020102）

本情况。收集的信息包括受检者年龄、性别、种族、受教育程度以及所在地区的HIV流行程度和调查前30天内毒品的注射次数等，试研究各因素对静脉注射吸毒程度的影响。部分结果见表1。

表1 9 824名静脉注射吸毒者基础情况

调查点编号（site）	受检者编号（id）	性别（sex）	种族（race）	受教育程度（educ）	年龄（岁）（age）	中心化年龄（cenage）	HIV流行程度（regionl）	静脉注射吸毒程度（inject）
10	101234	0	1	1	32	-7.79	0	3
10	101235	1	1	1	35	-4.79	0	2
…	…	…	…	…	…	…	…	…
28	288045	0	1	1	42	2.21	1	1
28	288046	0	1	0	43	3.21	1	3
…	…	…	…	…	…	…	…	…
32	321563	1	1	0	44	4.21	0	2
32	321566	0	0	0	35	-4.79	0	2

注：“性别”一列中，1=男性，0=女性；“种族”一列中：1=黑人，0=白人；“受教育程度”一列中，1=高中及以上，0=高中以下；“HIV流行程度”一列中，1=HIV高流行区，0=HIV低流行区；“静脉注射吸毒程度”一列中，1=重度（30天内注射吸毒100次以上），2=中度（30天内注射吸毒31～100次），3=轻度（30天内注射吸毒1～30次）；中心化年龄=年龄-平均年龄

3 非配对设计多值名义资料多水平多重logistic回归模型的构建原理

3.1 多值名义资料logistic回归模型

当名义结局变量y具有M个类别时，即m=1，2，…，M，可用式（1）表示多值名义资料logistic回归模型：

$$\log\left[\frac{P(y=m)}{P(y=M)}\right]=\alpha_m+\sum_{k=1}^{K}\beta_{1k}x_k \quad (1)$$

假设结局变量有M个类别，则会拟合（M-1）个logistic回归模型，会产生（M-1）个截距和（M-1）组斜率估计值。在这（M-1）个logistic回归模型中，每个对数发生比都是多个结局中的一个类别与参考类别进行比较。表达式如下：

$$\log\left[\frac{P_r(y=1)}{P_r(y=M)}\right]=\alpha_1+\sum_{k=1}^{K}\beta_{1k}x_k$$

……

$$\log\left[\frac{p_r(y=m)}{P_r(y=M)}\right]=\alpha_m+\sum_{k=1}^{K}\beta_{mk}x_k \quad (2)$$

$$\log\left[\frac{p_r(y=(M-1))}{p_r(y=M)}\right]=\alpha_{(M-1)}+\sum_{k=1}^{K}\beta_{(M-1)k}x_k$$

在此前的模型中，假定以第M个类别为参照类别。由于$P_r(y=1)+P_r(y=2)+\cdots+P_r(y=M)=1$，则：

$$P_r(y=1)+P_r(y=2)+\cdots+P_r(y=M)$$
$$=P_r(y=M)\left[1+\sum_{m=1}^{M-1}\exp\left(\alpha_m+\sum_{k=1}^{K}\beta_{mk}x_k\right)\right]=1 \quad (3)$$

由此可得下式：

$$P_r(y=M)=\frac{1}{1+\sum_{m=1}^{M-1}\exp\left(\alpha_m+\sum_{k=1}^{K}\beta_{mk}x_k\right)} \quad (4)$$

3.2 多值名义资料多水平logistic回归模型

当结局变量为多值名义变量且具有M个类别时，反应变量取第m类值的多值名义多水平logistic回归模型可用广义线性混合效应模型表达[2-3]，其表达式为：

$$\log\left[\frac{p_r(y=m)}{p_r(y=M)}\right]=\eta_m=\mathrm{X}\beta_m+\mathrm{Z}U_m \quad (5)$$

在式（5）中，m=1，2，…，M；设计矩阵X的固定效应向量为β_m；设计矩阵Z的随机效应向量为U_m。该模型假定随机效应服从正态分布，其均数为零，方差/协方差为矩阵G[即μ~N(0，G)]。当结局变量为具有3个类别的名义变量时，其多值名义资料多水平logistic回归模型有两个logistic回归模型：

$$\log\left[\frac{P_r(y=1)}{P_r(y=3)}\right]=\eta_1=\mathrm{X}\beta_1+\mathrm{Z}U_1 \quad (6)$$

$$\log\left[\frac{p_r(y=2)}{P_r(y=3)}\right]=\eta_2=\mathrm{X}\beta_2+\mathrm{Z}U_2 \quad (7)$$

其中，我们将结局变量取类别3作为参照类别，且同时拟合两个logistic回归模型，估计两组固定效应（β_1和β_2）和两组随机效应（U_1和U_2）。根据式（6）和式（7），结局变量为各类别的条件概率模型为：

$$P_r(y_{ij}=1|\mu_1)=\frac{\exp(\eta_1)}{1+\sum_{m=1}^{3}\exp(\eta_m)} \quad (8)$$

$$P_r(y_{ij}=2|\mu_2)=\frac{\exp(\eta_2)}{1+\sum_{m=1}^{3}\exp(\eta_m)} \quad (9)$$

$$P_r(y_{ij}=3|\mu_3)=\frac{1}{1+\sum_{m=1}^{3}\exp(\eta_m)} \quad (10)$$

4 基于SAS分析实例

4.1 分析与解答

例1中，若将响应变量静脉注射吸毒程度作为多值名义变量，试研究各因素对静脉注射吸毒程度的影响。该数据中个体可以看作1水平单位，调查点（site）看作2水平单位。可进行多值名义资料多水平多重logistic回归模型的构建。基于此，结局变量吸毒程度作为具有三个类别的名义变量，且随机系数为截距和水平1解释变量race的斜率，而其他水平1协变量，如sex、cenage、educ等都为有固定效应的自变量。另外，水平2协变量只有一个，即region。其模型表达式为：

$$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right)=\beta_{0j}+\beta_{1j}race_{ij}+\beta_2 sex_{ij}+\beta_3 cenage_{ij}+\beta_4 educ_{ij}+e_{ij} \quad (11)$$

$$\beta_{0j}=\gamma_{00}+\gamma_{01}region_j+\mu_{0j}$$

$$\beta_{1j}=\gamma_{10}+\gamma_{11}region_j+\mu_{1j}$$

因此，

$$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right)=\gamma_{00}+\gamma_{01}region_j+\gamma_{10}race_{ij}+\gamma_{11}region_j*race_{ij}+\beta_2 sex_{ij}+\beta_3 cenage_{ij}+\beta_4 educ_{ij}+(\mu_{0j}+\mu_{1j}*race_{ij}+e_{ij}) \quad (12)$$

这里，同时对两个logistic回归模型进行分析。SAS程序如下：

```
proc import datafile= 'C:\Users\Administrator\Desktop\d.xlsx' out=d dbms=xlsx replace;  /*1 数据导入*/
run;
proc glimmix data=d method=rspl;  /*2 进行多值名义资料多水平logistic回归*/
class site inject;
model inject=region|race sex cenage educ/sdist=multi link=glogit ddfm=bw;
random int race/subject=site group=inject;
nloptionstech=nrridg;
run;
proc nlmixed data=d;  /*3进一步拟合非线性混合模型*/
parms
ga0=-0.9 ga1_region=1.4 ga2rac=-0.75 ga3reg_rac=0.32 ga4sex=-0.16 ga5cenage=0.02 ga6educ=0.05 v_u0a=0.81 v_u1a=0.18
gb0=-0.08 gb1_region=0.96 gb2rac=-0.69 gb3reg_rac=0.10
gb4sex=0.01 gb5cenAGE=0.02 GB6EDUC=0.05 V_U0B=0.40 V_U1B=0.08;
eta1=ga0+ga1_region*region+ga2rac*race+ga3reg_rac*region*race+ga4sex*sex+ga5cenage*cenage+ga6educ*educ+(u0a+u1a*race);
eta2=gb0+gb1_region*region+gb2rac*race+gb3reg_rac*region*race+gb4sex*sex+gb5cenage*cenage+gb6educ*educ+(u0b+u1b*race);
if inject=3 then p=1/(1+exp(eta1)+exp(eta2));
else if inject=1 then p=exp(eta1)/(1+exp(eta1)+exp(eta2));
else if inject=2 then p=exp(eta2)/(1+exp(eta1)+exp(eta2));
ll=log(p);
model inject~general(ll);
random u0a u1a u0b u1b~normal([0,0,0,0],[v_u0a,0,v_u1a,0,0,v_u0b,0,0,0,v_u1b])
subject=site;
run;
```

【说明】程序共分为3步，第1步是导入数据集，后面是分析过程，分别使用的是GLIMMIX过程和NLMIXED过程。

在GLIMMIX过程中，method选项指定广义线性混合模型的参数估计方法是虚拟残差似然法。Class语句指定了site和inject为分类变量。Model语句等号前指定inject为因变量，等号后指定模型中的自变量。S指定显示固定效应的解，dist指定响应变量的分布类型为多项式分布，link指定连接函数。由于结局变量为多值名义变量，故此处指定连接函数为广义logit函数，ddfm选项指定分配自由度的方式，参数值bw表示如果固定效应在一个对象内发生变化，分配对象内自由度，否则分配对象间自由度。Random语句指定了随机效应为截距和race。Subject选项指定了模型中的对象，group指定协方差参数的分组。Nloptions语句中的tech指定了使用Newton-Raphson岭稳定优化法进行非线性参数估计的优化，以解决在某些分布中默认的二元准牛顿算法存在的收敛问题。

NLMIXED过程中parms语句设定了一些系数的初始值，这些初始值来自于GLIMMIX的参数估计值。Model语句中“~”之前是因变量，后面是因变量服从的分布以及相关参数。General(ll)指定model语句之前使用编程语句构造的广义对数似然函数。

Random 语句指定随机效应以及它的分布。Subject 选项定义随机效应的唯一识别变量。

4.2 主要输出结果及解释

Dimensions

G-side Cov. Parameters	6
Columns in X	14
Columns in Z per Subject	6
Subjects (Blocks in V)	20
Max Obs per Subject	1311

Covariance Parameter Estimates

Cov Parm	Subject	Group	Estimate	Standard Error
Intercept	site	inject 1	0.8052	0.3067
race	site	inject 1	0.1882	0.1035
Intercept	site	inject 2	0.3958	0.1480
race	site	inject 2	0.0768	0.0567
Intercept	site	inject 3	0.0000	-
race	site	inject 3	0.0000	-

Solutions for Fixed Effects

effect	inject	estimate	error	df	t value	Pr>|t|
Intercept	1	-0.9002	0.2517	10	-3.58	0.0050
Intercept	2	-0.0773	0.1785	10	-0.43	0.6740
region	1	1.4047	0.5187	10	2.71	0.0220
region	2	0.9622	0.3808	10	2.53	0.0300
race	1	-0.7523	0.1536	10	-4.90	0.0006
race	2	-0.6928	0.1035	10	-6.69	<0.0001
region*race	1	0.3218	0.3426	10	0.94	0.3696
region*race	2	0.0990	0.2607	10	0.38	0.7121
sex	1	-0.1638	0.0673	9800	-2.44	0.0149
sex	2	0.0102	0.0556	9800	0.18	0.8553
cenage	1	0.0155	0.0042	10	3.71	0.0041
cenage	2	0.0235	0.0033	10	7.10	<0.0001
educ	1	0.0526	0.0615	9800	0.85	0.3927
educ	2	0.0513	0.0491	9800	1.04	0.2961

以上 SAS 结果是 GLIMMIX 的输出结果。SAS 输出的 Dimensions 显示模型有 6 个随机效应，而 Convariance Parameter Estimates 仅显示了 4 个随机效应，其原因是随机效应不适用于参照组。由于 GLIMMIX 过程不提供随机截距的统计显著性检验，因此我们将采用 NLMIXED 过程确认其统计学显著性。

SAS 输出的 Solutions for Fixed Effects 部分列出 14 个固定参数估计值，即每个 logistic 回归模型有 7 个参数估计值。例如，Intercept1 和 Intercept2 分别是第 1 个和第 2 个 logit 的截距系数估计值。控制协变量之后，类别 1 发生几率小于类别 3 发生几率：优势比为 $\exp(-0.9002)=0.41$（$P=0.0022$）；发生类别 2 与类别 3 的几率没有统计学差异：优势比为 $\exp(-0.0773)=0.93$（$P=0.6700$）。两个 logit 中，region 均有统计学差异，类别 1 与类别 3 比较，发生比率为 $\exp(1.4047)=4.07$（$P=0.0068$）；类别 2 与类别 3 比较，发生比率为 $\exp(0.9622)=2.67$（$P=0.0115$）。表明高 HIV 流行区的静脉注射吸毒者更易成为重、中度吸毒者。值得注意的是，多水平 logistic 回归模型采取了多次 logit 变换，因此解释协变量效应时具体的 logit 应与效应一一对应。region 和 race 之间的跨层交互作用在两个 logit 中均无统计学意义（P 值分别为 0.3696 和 0.7121）。

Specifications

Data Set	WORK. D
Dependent Variable	inject
Distribution for Dependent Variable	General
Random Effects	u0a u1a u0b u1b
Distribution for Random Effects	Normal
Subject Variable	site
Optimization Technique	Dual Quasi-Newton
Integration Method	Adaptive Gaussian Quadrature

Dimensions

Observations Used	9824
Observations Not Used	0
Total Observations	9824
Subjects	20
Max Obs per Subject	1311
Parameters	18
Quadrature Points	1

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
ga0	−0. 9051	0. 2513	16	−3. 60	0. 0024	−1. 4378	−0. 3723	0. 1378
ga1_region	1. 4004	0. 5156	16	2. 72	0. 0153	0. 3074	2. 4934	−0. 0752
ga2rac	−0. 7495	0. 1500	16	−5. 00	0. 0001	−1. 0674	−0. 4316	−0. 3412
ga3reg_rac	0. 3208	0. 3280	16	0. 98	0. 3426	−0. 3745	1. 0162	0. 0083
ga4sex	−0. 1671	0. 0633	16	−2. 48	0. 0246	−0. 3098	−0. 0243	−0. 4161
ga5cenage	0. 0156	0. 0042	16	3. 73	0. 0018	0. 0067	0. 0244	−0. 0118
ga6educ	0. 0531	0. 0616	16	0. 86	0. 4010	−0. 0774	0. 1836	−0. 2059
v_u0a	0. 7951	0. 3105	16	2. 56	0. 0210	0. 1368	1. 4534	0. 7212
v_u1a	0. 1519	0. 0821	16	1. 85	0. 0828	−0. 0221	0. 3259	−1. 1240
gb0	−0. 0777	0. 1731	16	−0. 45	0. 6596	−0. 4446	0. 2892	−0. 1434
gb1_region	0. 9618	0. 3736	16	2. 57	0. 0204	0. 1699	1. 7537	−0. 1325
gb2rac	−0. 6939	0. 0994	16	−6. 98	<0. 0001	−0. 9047	−0. 4831	−0. 2743
gb3reg_rac	0. 0930	0. 2656	16	0. 35	0. 7309	−0. 4701	0. 6560	0. 1943
gb4sex	0. 0091	0. 0557	16	0. 16	0. 8729	−0. 1090	0. 1271	−0. 0491
gb5cenAGE	0. 0236	0. 0033	16	7. 13	<0. 0001	0. 0166	0. 0307	−0. 0208
GB6EDUC	0. 0529	0. 0492	16	1. 08	0. 2978	−0. 0513	0. 1572	0. 7521
V_U0B	0. 3670	0. 1335	16	2. 75	0. 0143	0. 0839	0. 6502	0. 5416
V_U1B	0. 0595	0. 0481	16	1. 24	0. 2333	−0. 0423	0. 1614	0. 1928

以上是 NLMIXED 过程的输出结果。在 NLMIXED 过程的 PARMS 语句中，带有前缀 GA 和 GB 的参数分别是第 1 个和第 2 个 logistic 回归模型的系数，该回归系数的设定来源于 GLIMMIX 过程的结果。Eta1 和 eta2 分别是两个 logit 函数的线性预测指标。根据公式(7)(8)(9)，IF THEN 语句设定结局变量的类比概率。

Random 语句设定 4 个随机效应的方差/协方差矩阵，即 logit1 的 u0a、u1a 和 logit2 的 u0b、u1b。随机效应的均数设定为零，方差/协方差是根据 G 矩阵的下三角定义的，其对角线上的元素为方差，对角线以外的元素为协方差。结果显示，logit1 和 logit2 的随机截距方差均有统计学意义（v_u0a=0. 8，P=0. 0210；V_U0B=0. 37，P=0. 0143）；但变量 race 的随机斜率方差无统计学意义（v_u1a=0. 15，P=0. 0827；V_U1B=0. 06，P=0. 2333）。结果表明，种族对成为重度注射吸毒类别的几率的效应不随各项目实施点显著变化。以上两个过程的输出结果显示，NLMIXED 和 GLIMMIX 两个过程对固定效应的估算结果相近。

【结论】对结局变量而言，HIV 流行程度、种族、年龄和性别的影响都是不可忽视的。至于每个影

响因素各水平对结果影响的差异情况，需结合每个因素的参照类别和回归系数的正负号，方可给出具体的解释。

5 讨论与小结

对于响应变量为多值名义变量的资料，一般采用广义 logistic 回归模型。但是该模型要求观测结局相互独立。当资料为多层的多值名义资料时，考虑到数据之间的聚集性，应当采用多水平回归模型进行分析，这样可以使个体的随机误差更纯[3-4]。

我们采用 GLIMMIX 和 NLMIXED 两个过程来构建模型：前者构建模型的速度快且用法简单，但在模型比较时通常不适用，且没有提供随机效应的假设检验，不能采用 t 检验计算相应的 P 值；而 NLMIXED 过程可以提供真实的对数似然值，并提供随机效应假设检验的结果，也可以通过似然比检验对嵌套模型的拟合效果进行比较，但用法复杂，需设置模型参数的初始值。因此，一般以 GLIMMIX 过程得到的参数估计值作为 NLMIXED 过程的模型参数初始值，最后以 NLMIXED 过程的结果为准。

参考文献

[1] 刘红伟，张甜甜，李长平，等. 非配对设计二值资料多水平多重 Logistic 回归分析[J]. 四川精神卫生，2019，32(5)：390-394.

[2] 王济川，谢海义，姜宝法. 多层统计分析模型——方法与应用[M]. 北京：高等教育出版社，2008：160-168.

[3] 胡良平，王琪. 定性资料统计分析及应用[M]. 北京：电子工业出版社，2016：198-238.

[4] 胡良平. 面向问题的统计学——(2)多因素设计与线性模型分析[M]. 北京：人民卫生出版社，2012：482-494，518-526，610-617.

（收稿日期：2019-11-19）

（本文编辑：吴俊林）