非配对设计多值名义资料一水平多重Logistic回归分析

巩晓文[1]，李长平[1,2]，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：胡良平，E-mail：lphu812@sina.com）

【摘要】 本文目的是介绍非配对设计多值名义资料一水平多重logistic回归分析的基本原理、建模策略及注意事项。结合实例，应用SAS 9.4构建未经变量筛选和经变量筛选的多值名义资料多重logistic回归模型。通过回归分析的计算结果可知，同一变量的回归系数在不同logit函数中存在代数关系。多值名义多重logistic回归分析可以用来处理结果变量为多值名义变量的回归建模问题，并可以结合SAS实现对自变量的筛选，以获得简洁的回归模型。

【关键词】 多值名义资料；logistic回归分析；自变量筛选；设计矩阵

中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20191119004

One-level multiple Logistic regression analysis of the nominal data collected from the unpaired design

Gong Xiaowen[1], *Li Changping*[1,2], *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: Hu Liangping, E-mail: lphu812@sina.com*)

【Abstract】 The paper introduced the basic principle, modeling strategy and key points of the one-level multiple logistic regression analysis of the multi-value nominal data collected from the unpaired design. In order to analyze the influencing factors of the choice of treatment for patients with non-ST-segment elevation myocardial infarction (NSTEMI) of the real example, and to predict the appropriate treatment according to the important characteristics, it was built that the one-level multiple nominal logistic regression model with and without variable selection by using SAS 9.4 software. The regression results showed that the regression coefficients of the same variable had algebraic relations in different logit functions. Multiple nominal logistic regression analysis could deal with the regression problem of the multi-value nominal data and with the help of SAS software. We could establish a concise model by filtering the insignificant independent variables.

【Keywords】 Multi-value nominal data; Logistic regression analysis; Variables selection; Design matrix

医学研究中的资料常涉及结局变量为多值名义变量的资料。多值名义资料的特点是结果变量的多种取值之间没有内在等级和数量大小之分[1]。故在对此类数据进行统计建模时，二值和多值有序多重logistic回归模型均不适用。Andson于1972年提出了多值名义logistic回归模型，以解决结局变量为多值名义变量的回归分析问题[2]。根据医学研究设计类型，研究可分为配对设计和非配对设计，前者一般是指病例对照研究中根据病例组的重要特征匹配对照而收集数据；后者是没有经过匹配便可收集数据，多见于横断面研究或队列研究。根据数据层级关系，又可将研究分为“一水平”和“多水平”研究。本文着重讨论非配对设计多值名义资料一水平多重logistic回归分析，并结合实例，采用SAS 9.4予以实现。

1 构建多值名义资料一水平多重logistic回归模型的基本原理

1.1 三值名义资料logistic回归模型

首先对三值名义资料logistic回归模型加以说明[3]。假设某一事件可能有A、B、C三种情况，P_{iA}=个体i发生事件A的概率；P_{iB}=个体i发生事件B的概率；P_{iC}=个体i发生事件C的概率。假定问题中涉及4个协变量（即自变量），个体i的设计矩阵所对应的向量

项目基金：国家高技术研究发展计划课题资助（2015AA020102）

可以用$x_i=[1\ x_{i1}\ x_{i2}\ x_{i3}\ x_{i4}]'$来表示(向量中第1个分量"1"与回归模型中"截距"相对应)。不妨沿用二分类logistic回归分析的思路构建模型:

对于A有:$\log\left(\frac{P_{iA}}{1-P_{iA}}\right)=\beta_A x_i$

对于B有:$\log\left(\frac{P_{iB}}{1-P_{iB}}\right)=\beta_B x_i$

对于C有:$\log\left(\frac{P_{iC}}{1-P_{iC}}\right)=\beta_C x_i$

上述思想本质上是做了三个传统的logistic回归模型,当计算A发生的概率时,将B、C合并,其余同理。然而,尽管这样可以估计出A、B、C三种事件发生各自的概率,但忽略了一个重要前提条件,即:对于任意个体i而言,其约束条件为$P_{iA}+P_{iB}+P_{iC}=1$。上述模型无法从理论上保证该约束条件成立。因此,不妨考虑选择一个类别为参考(如选择C为参考),来计算其他类别相对于该参考类别的概率:

$$\log\left(\frac{P_{iA}}{P_{iC}}\right)=\beta_A x_i,\qquad \log\left(\frac{P_{iB}}{P_{iC}}\right)=\beta_B x_i$$

那么,$\log\left[\frac{\left(P_{iA}/P_{iC}\right)}{\left(P_{iB}/P_{iC}\right)}\right]=\log\left(\frac{P_{iA}}{P_{iB}}\right)$,

令:$\log\left(\frac{P_{iA}}{P_{iB}}\right)=\beta_C x_i$

根据对数函数性质可以发现有如下关系:

$$\log\left(\frac{P_{iA}}{P_{iB}}\right)=\log\left(\frac{P_{iA}}{P_{iC}}\right)-\log\left(\frac{P_{iB}}{P_{iC}}\right)$$

即:$\beta_C x_i=\beta_A x_i-\beta_B x_i$,亦即:$\beta_C=\beta_A-\beta_B$。因此,只要估计出$\beta_A$、$\beta_B$和$\beta_C$中任意两者,即可求出第三个。同理,对于$K(K\geqslant 3)$分类的情形,只要估计出$(K-1)$个方程中的参数即可。求解$P_{iA}$、$P_{iB}$、$P_{iC}$,可得:

$$P_{iA}=\frac{e^{\beta_A x_i}}{1+e^{\beta_A x_i}+e^{\beta_B x_i}}$$

$$P_{iB}=\frac{e^{\beta_B x_i}}{1+e^{\beta_A x_i}+e^{\beta_B x_i}}$$

$$P_{iC}=\frac{1}{1+e^{\beta_A x_i}+e^{\beta_B x_i}}$$

由此可见,对于任意个体i,其A、B、C事件发生的概率之和恒等于1。

1.2　模型的一般形式

上文介绍了三分类的多重logistic回归模型,现在推广到J分类的多重logistic回归模型[4]。同理,用$k(k=1,2,\cdots,J-1,J)$表示类别。P_{ik}表示第i个个体分到第k类的概率。模型为:

$$\log\left(\frac{P_{ik}}{P_{iJ}}\right)=\beta_k x_i,\ k=1,2,\cdots,J-1,J。$$

其中x_i代表第i个个体的协变量向量,β_k代表第k类相对于第J类的回归系数向量。经转换可得:

$$P_{ik}=\frac{e^{\beta_k x_i}}{1+\sum_{k=1}^{J-1}e^{\beta_k x_i}},\ k=1,2,\cdots,J-1,J。$$

因为所有J类的概率之和必须为1,所以第J类的概率为:$P_{iJ}=\frac{1}{1+\sum_{k=1}^{J-1}e^{\beta_k x_i}}$

2　基于SAS的实例分析

2.1　未经变量筛选的多值名义资料多重logistic回归分析

2.1.1　问题与数据

研究某医院非ST段抬高型心肌梗死后血运重建治疗方式的影响因素。目前血运重建的治疗方式主要有药物治疗、经冠状动脉介入(PCI)治疗和冠状动脉搭桥(CABG)。共收集1 293例患者的资料(因数据过多,此处从略),包括治疗方式、年龄、性别、是否吸烟、是否饮酒、是否患高血压、是否患糖尿病、是否患脑卒中、是否患高脂血症、是否患陈旧性心肌梗死、是否曾行PCI手术、是否曾行CABG手术及入院时的KILLIP分级,详细编码方式见表1。如果研究者想观察每一个变量对回归结果的影响,则无需进行变量筛选,而将所有的变量都纳入模型即可。

表1　变量说明表

研究因素	变量名	编码
治疗方式	Trt	0=药物治疗,1=PCI,2=CABG
年龄	Age	连续性变量
性别	Gender	0=女,1=男
吸烟	Smoking	0=否,1=是
饮酒	Drinking	0=否,1=是
高血压	Hypertension	0=否,1=是
糖尿病	Diabetes	0=否,1=是
脑卒中	Stroke	0=否,1=是
高脂血症	Hyperlipemia	0=否,1=是
陈旧心肌梗死	HoMI	0=否,1=是
PCI史	PCI	0=否,1=是
CABG史	CABG	0=否,1=是
KILLIP分级	KILLIP	1=1级,2=2级,3=3级

2.1.2　分析过程

首先需要创建SAS数据集nstemi(因篇幅所限,

此处从略）。多值名义资料多重 logistic 回归分析可以调用 LOGISTIC、SURVEYLOGISTIC、CATMOD 或 GLIMMIX 过程实现，鉴于 LOGISTIC 过程比较常用，故本例调用 LOGISTIC 过程实现[5]。

```
proc logistic data= nstemi;
class Gender Smoking Drinking Hypertension
Diabetes Stroke Hyperlipemia HoMI PCI CABG;
model Trt(ref="0") = Age Gender Smoking Drinking Hypertension
Diabetes Stroke Hyperlipemia HoMI PCI CABG KILLIP / link=glogit;
model Trt(ref="2") = Age Gender Smoking Drinking Hypertension
Diabetes Stroke Hyperlipemia HoMI PCI CABG KILLIP / link=glogit;
run;
```

【程序说明】使用名为 nstemi 的数据集。调用 LOGISTIC 过程。Class 语句指定后面的变量为分类变量。在 model 语句中使用 link=glogit 选项要求使用多值名义资料的多重 logistic 回归分析；如果省略该语句，那么系统将构建累计 logit 模型，即认为 Trt 是有序变量。此外，用 ref="0"指定以药物治疗作为参考类别，即估计 PCI vs. 药物治疗和 CABG vs. 药物治疗的结果。为了验证 $\beta_C = \beta_A - \beta_B$，又设定 ref="2"来获得 PCI vs. CABG 的结果。读者可根据研究需求选择合适的参考类别。

2.1.3 主要输出结果及解读

主要输出结果见表 2。

表 2 多值名义资料回归分析结果

变量	PCI vs. 药物治疗			CABG vs. 药物治疗			PCI vs. CABG		
	β	*S. E.*	*P*	β	*S. E.*	*P*	β	*S. E.*	*P*
Age	−0.0340	0.0065	<0.0001	−0.0170	0.0112	0.1281	−0.0170	0.0106	0.1090
Gender	0.1572	0.0791	0.0469	0.1666	0.1370	0.2240	−0.0094	0.1331	0.9436
Smoking	−0.1252	0.0765	0.1018	−0.0645	0.1319	0.6249	−0.0607	0.1264	0.6310
Drinking	0.1134	0.0840	0.1769	0.1612	0.1475	0.2746	−0.0478	0.1389	0.7309
Hypertension	0.1162	0.0718	0.1055	0.0240	0.1266	0.8495	0.0922	0.1198	0.4414
Diabetes	−0.0830	0.0704	0.2386	−0.3094	0.1165	0.0079	0.2265	0.1108	0.0410
Stroke	0.1252	0.0760	0.0996	−0.0154	0.1288	0.9049	0.1406	0.1255	0.2626
Hyperlipemia	−0.0786	0.0743	0.2898	−0.2517	0.1407	0.0737	0.1731	0.1374	0.2078
HoMI	0.4280	0.0963	<0.0001	−0.0540	0.1605	0.7362	0.4820	0.1615	0.0028
PCI	−0.2161	0.0998	0.0304	0.2931	0.1935	0.1297	−0.5092	0.1901	0.0074
CABG	0.1762	0.1352	0.1926	0.8640	0.3767	0.0218	−0.6878	0.3768	0.0680
KILLIP	−0.6773	0.1420	<0.0001	−0.8922	0.2953	0.0025	0.2149	0.3024	0.4772

表 2 为 PCI vs. 药物治疗、CABG vs. 药物治疗和 PCI vs. CABG（设定 ref="2"计算得到）的回归结果。由表 2 可知，$\beta_C = \beta_A - \beta_B$。以年龄为例，PCI vs. 药物治疗的回归系数为−0.0340，CABG vs. 药物治疗的回归系数为−0.0170，通过 SAS 结果可以验证 PCI vs. CABG 的回归系数为−0.0340−(−0.0170)=−0.0170。回归分析结果表明：相对于药物治疗而言，年龄越小、KILLIP 分级越低（理由是其系数为负值，KILLP 分级越低表示病情越轻）、男性、有心肌梗死史、既往未做过 PCI 的患者更倾向于选择 PCI 治疗；相对于药物治疗而言，年龄越小、没有糖尿病、既往做过 CABG 且 KILLIP 分级越低的患者更倾向于选择 CABG 治疗。

2.2 经变量筛选的多值名义资料多重 logistic 回归分析

2.2.1 问题与数据

在实际工作中，为了使模型简洁或避免变量之间的共线性，在构建回归模型时往往需要进行变量筛选。常见的变量筛选策略包括向前法、向后法、逐步法和最优子集法。仍然沿用前面的实例，如果研究者想要建立一个预测模型，并能快速判断应该选择的治疗方式，可采用以下 SAS 程序。

2.2.2 SAS 程序

```
proc logistic data= nstemi;
class Gender Smoking Drinking Hypertension
Diabetes Stroke Hyperlipemia HoMI PCI CABG;
model Trt (ref="0") = Age Gender Smoking Drinking Hypertension
Diabetes Stroke Hyperlipemia HoMI PCI CABG KILLIP / link=glogit selection=stepwise SLENTRY=0.05 SLSTAY=0.05;
run;
```

【程序说明】与上文稍有不同的是，在 model 语句中添加了 selection=stepwise 选项用来指定逐步法筛选自变量，除了 stepwise 之外，还可以选择 forward（向

前法）、backward（向后法）和 score（最优子集法）。SLENTRY=0.05 指定效应进入模型的得分卡方显著性水平为 0.05。SLSTAY=0.05 指定在向后消除步中，效应保留在模型中的显著性水平为 0.05。

2.2.3　主要输出结果及解释

经过逐步法筛选后，模型得以精简，最终有四个变量纳入模型，分别为 Age、HoMI、PCI 和 KILLIP。相对于药物治疗来说，年龄越小、有心肌梗死史、既往没有做 PCI 且 KILLIP 分级越低的患者越容易选择 PCI 治疗；相对于药物治疗来说，年龄越小且 KILLIP 分级越低的患者越容易选择 CABG 治疗；相对于 CABG 来说，年龄越小、有心肌梗死史且既往没有做 PCI 的患者越容易选择 PCI 治疗。见表 3。

表 3　多值名义资料多重 logistic 回归分析结果

变　量	PCI vs 药物治疗			CABG vs 药物治疗			PCI vs CABG		
	β	*S. E.*	*P*	β	*S. E.*	*P*	β	*S. E.*	*P*
Intercept	3.96	0.4154	<0.0001	0.75	0.7206	0.2993	3.2139	0.6826	<0.0001
Age	−0.04	0.0059	<0.0001	−0.02	0.0101	0.0319	−0.0211	0.0095	0.0272
HoMI	0.42	0.0935	<0.0001	0.01	0.1531	0.9676	0.4135	0.1542	0.0073
PCI	−0.22	0.0988	0.0259	0.23	0.1885	0.2263	−0.4481	0.1848	0.0153
KILLIP	−0.66	0.1396	<0.0001	−0.77	0.2872	0.0070	0.1133	0.2951	0.7010

根据表 3 的回归系数，计算每一类的概率：

$$P_{PCI}=\frac{e^{3.96-0.04*Age+0.42*HoMI-0.22*PCI-0.66*KILLIP}}{1+e^{3.96-0.04*Age+0.42*HoMI-0.22*PCI-0.66*KILLIP}+e^{0.745-0.02Age+0.01*HoMI-0.23*PCI-0.77*KILLIP}}$$

$$P_{CABG}=\frac{e^{0.745-0.02Age+0.01*HoMI-0.23*PCI-0.77*KILLIP}}{1+e^{3.96-0.04*Age+0.42*HoMI-0.22*PCI-0.66*KILLIP}+e^{0.745-0.02Age+0.01*HoMI-0.23*PCI-0.77*KILLIP}}$$

$$P_{drug\ therapy}=\frac{1}{1+e^{3.96-0.04*Age+0.42*HoMI-0.22*PCI-0.66*KILLIP}+e^{0.745-0.02Age+0.01*HoMI-0.23*PCI-0.77*KILLIP}}$$

3　讨论与小结

Logistic 回归分析是医学领域常用的回归分析方法，传统的 logistic 回归分析是以二分类变量为结局变量。但在现实研究中，疾病种类、治疗方案等往往存在多种类别。在病例对照研究中，有一个对照组、两个或多个病例组；或者有一个病例组、两个或者多个对照组。以上情况涉及没有等级关系的多分类结果，如果对每两类结果都采用传统的 logistic 回归分析，可能会增加一类错误的概率。因此，多值名义资料多重 logistic 回归分析应运而生。

多分类结果 logistic 回归系数的解释与传统二分类的 logistic 回归分析相似，不过需要明确所选择的参照类别是哪一类，以免在结果解释时发生混淆，因为同一变量在不同 logit 函数的效应往往不同。传统的 logistic 回归模型估计系数在大多数情况下与多值名义 logistic 回归分析结果相近[6]。因此，可以将传统的 logistic 回归分析用于变量筛选，最后将各自筛选出的变量并集用于多值名义 logistic 回归分析中。随着 SAS 软件的发展，目前可以通过逐步法、向前法、向后法和最优子集法自动实现变量筛选，而不必手动筛选变量。当然，读者也可结合临床实际选择不同的变量筛选策略。

此外，在拟合多值名义资料多重 logistic 回归分析时应注意以下问题：变量间是否存在共线性问题、样本量不宜过小、变量间是否有交互作用、哑变量设置是否合理。当遇到异常值时，应慎重考虑，并做敏感性分析[7]。

参考文献

[1] 李长平，胡良平. 定性资料的数据结构与分析方法概述[J]. 四川精神卫生，2019，32(4)：289-296.

[2] 刘世良. 多项 Logistic 回归分析及其应用[J]. 南华大学学报（医学版），1989(2)：198-201.

[3] Allison PD. Logistic regression using SAS®：theory and application[M].2nd Edition. Cary，NC：SAS Institute Inc，2012：139-166.

[4] Agresti A. Categorical data analysis[M].2nd Edition. New York：John Wiley & Sons，2002：137-172

[5] SAS Institute Inc . SAS/Stat 9.4 user's guide[M]. Cary，NC：SAS Institute Inc，2016：4163-4400.

[6] 冯国双，刘德平. 医学研究中的 logistic 回归分析及 SAS 实现[M]. 北京：北京大学医学出版社，2015：97-158.

[7] 胡良平. 提高回归模型拟合优度的策略（Ⅰ）——哑变量变换与其他变量变换[J]. 四川精神卫生，2019，32(1)：1-8.

（收稿日期：2019-11-19）

（本文编辑：陈　霞）