非配对设计多值有序资料一水平多重Logistic回归分析

凤思苑[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有序资料一水平多重logistic回归模型的构建与求解方法。本文详细介绍了构建累积logistic回归模型的原理和具体方法，并结合实例介绍如何使用SAS软件中的LOGISTIC过程来拟合此回归模型，并对逐步回归法的输出结果进行了解释；其次讨论了有关构建累积logistic回归模型的过程中自变量筛选、模型评价以及拟合模型时需注意的问题。

【关键词】 多值有序因变量；累积logistic回归分析；自变量筛选；模型评价；SAS实现
中图分类号：R195.1 文献标识码：A doi：10.11886/j.issn.1007-3256.2019.05.003

One-level multiple Logistic regression analysis with the multi-value ordered data collected from the unpaired design

Feng Siyuan[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 purpose of this paper was to introduce the construction and solution of one-level multiple logistic regression models for unpaired design multi-value ordered data. This paper introduced the principle and methods of cumulative logistic regression model in detail, and introduced how to use the LOGISTIC procedure of SAS software to fit the regression model, and explained the results of the screening independent variables by using stepwise method. In addition, the paper discussed the problems that should be paid attention to in the process of constructing the cumulative logistic regression model, such as independent variable selection, model evaluation and model fitting.

【**Keywords**】 Multi-value ordered dependent variables; Cumulative logistic regression analysis; Independent variable selection; Model evaluation; SAS realization

生物医学研究中最常见的问题之一就是探究各种影响因素（自变量X）与“是否发病”或“健康状况”（因变量Y）之间的关系。当结局变量是多值有序变量（如治疗结局为治愈、好转、显效和无效等）时，常用的线性回归模型就不再适合了。本文将结合实例介绍如何使用SAS实现非配对设计多值有序资料一水平多重logistic回归分析，其中一水平主要是指受试对象不具有层级结构，即满足研究样本随机来自同一个总体（即认为受试对象在变量之间关系上具有“同质性”）。

项目基金：国家高技术研究发展计划课题资助（2015AA020102）

1 基本概念

1.1 模型定义

多值有序logistic回归模型不同于二分类logistic回归模型，它是基于累积概率构建累积logistic回归模型。假设结局变量Y有J个有序分类，其自然结局顺序表示为$Y=1,2,\cdots,J$，每个分类结局对应的发生概率为$\pi_1,\pi_2,\cdots,\pi_j$，则其有序分类≤m的累计发生概率表示为$P(Y\leqslant m)=\pi_1+\pi_2+\cdots+\pi_m$。因此，可以通过指定累积概率$P(Y\leqslant m)$的阈值将整个结局变量$Y$的$J$个有序分类从指定的阈值点截断，使之成为二分类结局。设有P个自变量记为$X=(x_1,x_2,\cdots,x_p)$表示相应的影响因素。由此定义累积logit $P(Y\leqslant m)$函数：

$$\text{logit}\left[P\left(y\leqslant m\right)\right]=\log\left(\frac{p\left(y\leqslant m|x\right)}{1-p\left(y\leqslant m|x\right)}\right)=\log\frac{p\left(y\leqslant m|x\right)}{p\left(y>m|x\right)}$$
$$=\log\frac{\pi_1+\pi_2+\ldots+\pi_m}{\pi_{m+1}+\pi_{m+2}+\ldots+\pi_j}$$

该累积logit $P(Y\leqslant m)$函数是两个累积概率比的对数值，这两个累积概率分别表示结局变量Y的取值小于等于结局分类m与大于分类m的可能性大小[1-2]。因为结果Y共有J个有序分类，故最多可以写成$J-1$个累积logit函数。

$$\begin{cases}\text{logit}\left[P\left(y\leqslant 1\right)\right]=\log\left(\frac{p\left(y\leqslant 1|x\right)}{1-p\left(y\leqslant 1|x\right)}\right)=\log\frac{p\left(y\leqslant 1|x\right)}{p\left(y>1|x\right)}=\log\frac{\pi_1}{\pi_2+\pi_3+\ldots+\pi_j}\\ \cdots\\ \text{logit}\left[P\left(y\leqslant \text{j}-1\right)\right]=\log\left(\frac{p\left(y\leqslant j-1|x\right)}{1-p\left(y\leqslant j-1|x\right)}\right)=\log\frac{p\left(y\leqslant j-1|x\right)}{p\left(y>j-1|x\right)}=\log\frac{\pi_1+\pi_2+\ldots+\pi_{j-1}}{\pi_j}\end{cases}$$

累积logit函数还可以用线性函数形式表示如下：

$$\begin{cases}\text{logit}\left[P\left(y\leqslant 1\right)\right]=\beta_{10}+\beta_{11}x_1+\ldots+\beta_{1p}x_p\\ \cdots\\ \text{logit}\left[P\left(y\leqslant \text{j}-1\right)\right]=\beta_{j-1,0}+\beta_{j-1,1}x_1+\ldots+\beta_{j-1,p}x_p\end{cases}$$

上述模型就是累积logistic回归模型。为了进一步简化该模型，假定对于所有$J-1$个累积logit函数，各个自变量X_i所对应的系数β_i假设都是等同的，即每个累积logit函数相同自变量X_i都有相同的系数β_i以及不同的截距β_{j0}。在此假设条件下，$J-1$个累积logit函数的回归线其实是相互平行的，只是截距β_{j0}不同，该假设被称为平行假设。满足平行假设的模型简化后为：

$$\text{logit}\left[P\left(y\leqslant j\right)\right]=\beta_{j0}+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p=\beta_{j0}+\sum_{i=1}^{p}\beta_ix_i$$

该简化后的模型称为成比例比数比累积logit回归模型，该模型和一般累积logistic回归模型一样，至多有$J-1$个方程形式，即同样有$J-1$个截距，但是p个自变量的回归系数在不同方程中分别相同[3]。该模型对应的概率模型形式为：

$$P\left(y\leqslant j\right)=\frac{exp(\beta_{j0}+\beta_1x_1+\ldots+\beta_px_p)}{1+exp(\beta_{j0}+\beta_1x_1+\ldots+\beta_px_p)}=\frac{1}{1+exp\left[-(\beta_{j0}+\beta_1x_1+\ldots+\beta_px_p)\right]}$$

通过上述公式，可获得结局Y取类别为j时的概率：

$$\begin{cases}P_1=P\left(y\leqslant 1\right)=\frac{exp(\beta_{10}+\beta_1x_1+\ldots+\beta_px_p)}{1+exp(\beta_{10}+\beta_1x_1+\ldots+\beta_px_p)}=\frac{1}{1+exp\left[-(\beta_{10}+\beta_1x_1+\ldots+\beta_px_p)\right]}\\ \cdots\\ P_j=P\left(y\leqslant j\right)-P\left(y\leqslant j-1\right)=\frac{1}{1+exp\left[-(\beta_{j,0}+\beta_1x_1+\ldots+\beta_px_p)\right]}-\frac{1}{1+exp\left[-(\beta_{j-1,0}+\beta_1x_1+\ldots+\beta_px_p)\right]}\end{cases}$$

1.2 参数估计

多值有序资料的logistic回归分析的参数估计和结局为二分类的logistic回归分析相似，都可以用极大似然的方法估计[4]。对于n个独立观察对象的样本，第i个观察对象X_i出现$Y=j$分类结局的概率记为$P_j=P(Y=j\mid X_i)$，它是累积概率函数的差，即$P_j=P(Y=j\mid X_i)=P(Y\leqslant j\mid X_i)-P(Y\leqslant j-1\mid X_i)$。由此构建的似然函数$L$为：

$$L=\prod_{i=1}^{n}(P_{i1}^{y_{i1}}P_{i2}^{y_{i2}}\ldots P_{ij}^{y_{ij}})=\prod_{i=1}^{n}\prod_{j=1}^{J}(P_{ij}^{y_{ij}})=\prod_{i=1}^{n}\prod_{j=1}^{J}[P(Y\leqslant j|X_i)-P(Y\leqslant j-1|X_i)]^{y_{ij}}$$

式中y_{ij}表示第i个观察对象的结局变量Y分类为j等级时所对应的编码，它满足$\sum_{j=1}^{J}y_{ij}=1$，而该观测实际只可能对应一个等级结局，故而只有某个y_{ij}取值为1，其余皆为0。相应的对数似然函数如下：

$$LL=\sum_{i=1}^{n}\sum_{j=1}^{j}y_{ij}\ln(P_{ij})=\sum_{i=1}^{n}\sum_{j=1}^{J}y_{ij}\ln[\text{logit}(P_{ij})-\text{logit}(P_{i,j-1})]$$

2 实例分析

冠状动脉旁路移植术（Coronary artery bypass grafting，CABG）是目前治疗冠心病最有效方法之一，但往往会存在术后静脉移植血管病变，从而降低血管通畅率并引起患者缺血症状的复发。为了研究引起术后血管狭窄可能的影响因素，随机选择207名CABG术后超过一年的患者，评价性别、桥龄、原位靶血管病变支数、冠心病类型、左室射血分数、左室舒张末期内径、α-羟丁酸脱氢酶、极低密度脂蛋白、脂蛋白a和同型半胱氨酸对血管狭窄程度的影响。其中结局变量血管狭窄程度分为无狭窄（DS=1），部分狭窄（DS=2）和完全狭窄（DS=3）三个等级。见表1。

表 1 多值有序 logistic 回归分析的数据表

ID	Sex	ql	NLV[a]	HDT[b]	LVEF	LVDED	HBDB	VLDL	LPa	Hcy	DS
1	2	7	3	2	1	57	203	0.49	12.6	12.7	3
2	2	9	3	1	1	50	144	0.49	52.0	10.9	2
3	1	9	3	1	0	54	105	0.51	29.6	21.7	3
…	…	…	…	…	…	…	…	…	…	…	…
205	1	5	2	1	0	49	121	0.91	16.3	11.6	1
206	1	6	3	2	1	56	463	0.69	13.0	13.6	3
207	1	16	3	2	1	58	131	0.40	27.0	13.1	3

注：sex=性别；ql=桥龄；NLV=原位靶血管病变支数；HDT=冠心病类型；LVEF=左室射血分数；HBDB=α_羟丁酸脱氢酶；VLDL=极低密度脂蛋白；LPa=脂蛋白 a；Hcy=同型半胱氨酸；DS=血管狭窄程度；a=1、2、3 分别代表病变血管支数；b=1 为稳定性心绞痛，2 为急性冠状动脉综合征

2.1 SAS 程序

```
data DS;
input Sex ql NLV HDT LVEF LVDED HBDH VLDL LPa Hcy DS @@;
cards;
2 7 3 2 1 57 203 0.49 12.6 12.7 3
2 9 3 1 1 50 144 0.49 52.0 10.9 2
1 9 3 1 0 54 105 0.51 29.6 21.7 3
…
1 5 2 1 0 49 121 0.91 16.3 11.6 1
1 6 3 2 1 56 463 0.69 13.0 13.6 3
1 16 3 2 1 58 131 0.40 27.0 13.1 3
;
proc logistic data=DS;
class NLV (ref="3");
model DS= Sex ql NLV HDT LVEF LVDED HBDH VLDL LPa Hcy /selection=stepwise sle=0.10 sls=0.15;
run;
proc logistic data=DS;
class NLV (ref=first);
model DS= Sex…Hcy /selection=forward sle=0.10;
run;
proc logistic data=DS;
class NLV (ref="3");
model DS= Sex …Hcy /selection=backward sls=0.15;
run;
```

【说明】首先建立临时数据集 DS，依次输入变量性别、桥龄、原位靶血管病变支数、冠心病类型、左室射血分数、左室舒张末期内径、α_羟丁酸脱氢酶、极低密度脂蛋白、脂蛋白 a 和同型半胱氨酸。接着调用 LOGISTIC 过程完成累积回归模型的分析。其中 class 语句为分类变量 NLV 创建哑变量，选项 ref="3"是以变量的第三个水平为对照实现哑变量赋值；Model 语句中因变量为 DS，其余变量为自变量。选项 selection= stepwise 表示变量筛选采用逐步回归方法，选项 sle 为选入自变量的显著性水平，选项 sls 为剔除自变量的显著性水平。

接下来依次调用第二、第三个 LOGISTIC 过程，采用的变量筛选分别为向前（forward）、向后（backward）回归方法。

【说明】在左栏的 SAS 程序中，第 2 和第 3 个“model 语句”中省略号部分的内容与第 1 个“model 语句”中相应位置上的变量相同；在实际使用时，最好取“sls=0.05”。

2.2 结果解释

LOGISTIC 过程输出结果的第一部分为模型总体的相关信息，所分析的数据集是临时数据集 DS，响应变量为血管狭窄程度 DS，采用的模型方法为 cumulative logit（累积 logit），模型优化的技术为 Fisher's scoring。结果变量共有三个水平，各自的例数分别为 53、27 和 117。其次该模型是以结局排序较低的取值为对比的基础，即以“y=1”为参照水平，也就是以血管无狭窄组为基础（即对照组）建模。

LOGISTIC 过程输出结果的第二部分输出了自变量筛选的过程，包括每次模型拟合后拟合统计量、整个模型检验以及平行线假设的结果。此实例中逐步法进行自变量的筛选过程共四步，由于篇幅原因，不做过多展示。逐步筛选法的筛选结果显示，最终自变量 LVDED、HDT、LVN、QL 进入了回归方程。平行线假设的检验结果为 χ^2=9.4233，P=0.0933>0.05，说明资料满足平行线假设。

LOGISTIC 过程输出结果的第三部分主要输出参数估计的结果：

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	7.7648	1.8673	17.2919	<0.0001
Intercept	2	1	8.7607	1.8896	21.4956	<0.0001
QL		1	−0.0868	0.0434	4.1839	0.0408
HDT		1	−1.1481	0.4464	6.6142	0.0101
NLV	1	1	0.7861	0.5909	1.1798	0.1834
NLV	2	1	0.9094	0.3292	7.6330	0.0057
LVDED		1	−0.1370	0.0372	13.5907	0.0002

在累积logistic回归模型中，截距项有多个，其个数为因变量的水平数减1。本例中因变量水平数为3，因此包含2个截距项。如果用P_1、P_2、P_3分别表示血管无狭窄、部分狭窄、完全狭窄的概率，则回归方程如下：

$$P_1=\frac{e^{7.7648-0.0868x_1-1.1481x_2+0.7861x_{31}+0.9094x_{32}-0.1370x_4}}{1+e^{7.7648-0.0868x_1-1.1481x_2+0.7861x_{31}+0.9094x_{32}-0.1370x_4}}$$

$$P_1+P_2=\frac{e^{8.7607-0.0868x_1-1.1481x_2+0.7861x_{31}+0.9094x_{32}-0.1370x_4}}{1+e^{8.7607-0.0868x_1-1.1481x_2+0.7861x_{31}+0.9094x_{32}-0.1370x_4}}$$

$$P_3=1-P_1-P_2=\frac{1}{1+e^{8.7607-0.0868x_1-1.1481x_2+0.7861x_{31}+0.9094x_{32}-0.1370x_4}}$$

式中x_1、x_2、x_{31}、x_{32}和x_4分别为自变量QL、HDT、NLV(1 VS 3)、NLV(2 vs 3)和LVDED。此外，本例中筛选出自变量对应的P值均<0.05，表明自变量的回归系数的估计值与0之间的差异均有统计学意义。其中QL回归系数估计值小于0，说明自变量桥龄取值越大，血管出现无狭窄的概率P_1越低，血管出现完全狭窄的概率P_3越大。QL的*OR*估计值为0.917，95%置信区间为(0.844，0.996)。其他变量的结果：

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
QL	0.917	0.844	0.996
HDT	0.317	0.132	0.761
NLV	2.195	0.689	6.988
NLV	2.483	1.302	4.733
LVDED	0.872	0.811	0.094

本文在筛选变量时除了逐步法以外，还采用了前进法和后退法。虽然变量筛选的具体过程不同，但最终纳入的变量以及相关的最大似然估计结果与逐步法相同，此处不做重复展示。

专业结论：桥龄(QL)、心脏病类型、原位靶血管病变支数和左室舒张末期内径与CABG术后血管再狭窄程度有关，而与其他变量无关。*OR*的点估计和置信区间结果显示桥龄越大、心脏病类型为急性冠状动脉综合征以及左室舒张末期内径越大，则血管无狭窄的可能性越低；原位靶血管病变支数2支相对于3支而言，血管出现无狭窄的可能性越高。

3 讨 论

本文主要采用了LOGISTIC过程对多值有序资料拟合累积logistic回归模型，在变量筛选方面分别选用了常用的逐步、向前和向后三种方法，结果表明三种方法最后纳入了相同的自变量，参数的极大似然估计也相同，但三种方法在变量筛选过程方面实则不同，具体的变量筛选原理可参阅文献[5]。从多种筛选自变量方法产生的回归方程中选择最优的回归方程，可参考的标准主要有以下几条：第一，整个回归方程以及筛选出的自变量具有统计学意义，并在专业上有合理的解释；第二，若回归方程中所含自变量的个数相同，取赤池信息标准值(Akaike information criteria，AIC)较小者，其次模型的结果以简单为主。本案例中三种变量筛选方法的AIC值均为409.338，且纳入的自变量相同，故最后结果相同。

除此之外，累积logit回归分析多值有序数据时依然还需要注意一些问题：(1)平行线假设：在拟合有序logistic回归时，需要对拟合的$J-1$个方程对应的累积概率曲线的平行性进行检验。当平行线假设未满足时，说明资料不适合有序logistic回归模型，应采用多值名义的logistic回归模型；(2)个体独立性：拟合多值有序logistic回归模型时，要求研究个体之间是相互独立的，即不存在组内个体同质、组间个体异质的现象，若资料不满足该情况则可以

采用多值有序多水平的logistic回归分析；(3)在建模时，还可以引入一些派生自变量(如连续变量的平方项、交叉乘积项等)参与自变量的筛选，有时可能获得拟合优度更高的回归模型。因篇幅所限，此处暂不赘述，可参阅文献[6-9]。

参考文献

[1] 陈佩珍，陈峰. 累积比数logistic回归在医学研究中的应用[J]. 南通医学院学报，2001，21(2)：140-142.

[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，2008：153-154.

[3] 鲍晓蕾，王小利，胡良平. 如何用SAS软件正确分析生物医学科研资料XXIV.结果变量为多值有序变量的高维列联表资料的统计分析与SAS软件实现(二)[J]. 中国医药生物技术，2013，8(4)：311-314.

[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，2012：518-526.

[5] 许汝福. 回归变量筛选及回归方法选择实例分析[J]. 中国循证医学杂志，2016，16(11)：1360-1364.

[6] 谷恒明，胡良平. 基于经典统计思想实现多重线性回归分析[J]. 四川精神卫生，2018，31(1)：7-11.

[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.

[8] 胡良平. 提高回归模型拟合优度的策略(Ⅱ)——算术均值变换与其他变量变换[J]. 四川精神卫生，2019，32(1)：9-15.

[9] 胡良平. 提高回归模型拟合优度的策略(Ⅲ)——校正均值变换与其他变量变换[J]. 四川精神卫生，2019，32(1)：16-20.

(收稿日期：2019-09-27)

(本文编辑：陈 霞)