非配对设计二值资料多水平多重Logistic回归分析

刘红伟[1]，张甜甜[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：李长平，E-mail：1067181059@qq.com）

【摘要】 本文目的是介绍非配对设计二值资料多水平多重logistic回归模型的构建与求解方法。首先介绍模型的有关概念及模型的构建原理，基于实例使用SAS软件对列联表资料进行分析，以proc glimmix和proc nlmixed过程构建和求解模型，并对相关结果进行解释和比较。

【关键词】 二值资料；多水平；SAS；多重logistic回归分析
中图分类号：R195.1 文献标识码：A doi：10.11886/j.issn.1007-3256.2019.05.002

Multi-level multiple Logistic regression analysis with the dichotomous choice data collected from the unpaired design

Liu Hongwei[1], *Zhang Tiantian*[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: Li Changping, E-mail: 1067181059@qq.com*)

【**Abstract**】 The purpose of this paper was to introduce the construction and solution of multi-level multiple logistic regression models for unpaired design binary data. Firstly, the related concepts of the model and the principle and construction of the model were introduced. The SAS software was used to analyze the contingency table data of the example. The model was constructed and solved by proc glimmix and proc nlmixed procedures, and the related results were explained and compared.

【**Keywords**】 Binary data; Multi-level; SAS software; Multiple logistic regression analysis

1 基本概念

1.1 二值资料

结局变量只有两个取值的资料称为“二值资料”，例如，在表1多中心临床试验数据中，治疗结局取值为“成功”或“失败”。

1.2 多水平数据

多水平数据或具有多水平层次结构的数据是多水平统计模型发展和应用的基础。多水平数据也就是自然形成的层次数据，例如，在多中心临床试验中，每个中心是水平2单位，受试者是水平1单位；在动物试验中，小鼠是水平1单位，窝别是水平2单位。多水平数据具有非独立性，故无法采用广义线性模型进行分析，因此提出了能够处理多水平数据的多水平模型[1-2]。

与广义线性模型相比，多水平模型稍显复杂，因为它同时包含了多个水平的数据，从而在多个水平上都存在残差。总体来说，其建模的思想就是把高水平上的差异估计出来（传统的线性模型不考虑这一差异，将其放到了一个统一的残差中），就使得残差变小，估计的结果更可靠[3-4]。

虽然理论上多水平模型可以有多个层次，但实际中最常用的是二水平模型，因此这里主要通过一份二水平数据简要介绍多水平模型构建与求解的思路。

2 数据结构

【例1】某地区开展多中心临床试验，拟比较两种药物治疗某疾病的效果。数据见表1。

项目基金：国家高技术研究发展计划课题资助（2015AA020102）

表 1 多中心临床试验数据

中心编号	药物种类	例数(人)			中心编号	药物种类	例数(人)		
		疗效:	成功	失败			疗效:	成功	失败
1	试验药		30	41	9	试验药		26	17
	对照药		16	32		对照药		18	15
2	试验药		25	8	10	试验药		34	14
	对照药		24	12		对照药		28	22
3	试验药		22	13	11	试验药		12	16
	对照药		12	18		对照药		6	16
4	试验药		14	23	12	试验药		4	15
	对照药		4	14		对照药		3	15
5	试验药		6	17	13	试验药		19	18
	对照药		3	19		对照药		15	23
6	试验药		3	7	14	试验药		14	24
	对照药		5	12		对照药		10	25
7	试验药		4	18	15	试验药		19	25
	对照药		2	9		对照药		13	21
8	试验药		13	4					
	对照药		22	4					

注:本例为假设资料

3 非配对设计二值资料多水平多重 logistic 回归模型的构建原理

3.1 回归模型的表达式

对于响应变量为二值变量的非层级结构数据,一般采用普通 logistic 回归模型分析,又称为固定效应 logistic 回归模型分析。设 $P(y=1|X)$(简记为 P)表示暴露因素为 X 时个体发生阳性事件(以 $y=1$ 表示发生阳性事件)的概率,而阳性事件发生的概率 P 与阴性事件发生的概率(1−P)之比称为优势比。对优势比进行自然对数变换即为对 P 的 logit 变换,得:

$$\text{logit}(P)=\ln\left(\frac{P}{1-P}\right)=\beta_0+\sum_{k=1}^{K}\beta_k x_{ki}=X\beta \quad (1)$$

令 $z=\beta_0+\sum_{k=1}^{K}\beta_k x_{ki}$,则有:

$$P=\frac{\exp(z)}{1+\exp(z)} \quad (2)$$

对于响应变量为二值变量的层级数据结构,可采用多水平 logistic 回归模型分析,其模型可表达为:

$$\ln\left(\frac{P}{1-P}\right)=X\beta+ZU \quad (3)$$

其中,U 是随机回归系数向量,服从 $N(0,G)$,G 为协方差矩阵,β 是水平 1 固定回归系数向量,X 和 Z 分别是固定效应和随机效应的解释变量设计矩阵。

3.2 回归模型的构建

以例 1 为例,以变量 zhongxin 表示“中心”,以变量 drug 表示“药物种类”,以变量 y 表示疗效($y=0$ 表示治疗成功,$y=1$ 表示治疗失败),以 P_{ij} 表示个体 $y=0$ 发生的概率。建模过程如下:

第一步,建立空模型,计算组内相关系数 ICC 的值。空模型中仅有一个随机截距而不包含任何解释变量,其模型为:$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right)=\beta_{0j}$ 和 $\beta_{0j}=\beta_0+\mu_{0j}$,上述模型可合并为:

$$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right)=\beta_0+\mu_{0j} \quad (4)$$

其中,β_0 为 $y=0$ 的总平均 logit 值,μ_{0j} 为组水平(本资料为中心)的平均 logit 值的变异量,表示第 j 个组的平均 logit 值与总平均 logit 值之间的差异,且 $\mu_{0j}\sim N(0,\sigma^2_{\mu_0})$。

多水平 logistic 回归模型的组间变异也可用组内相关系数进行评估,因 logistic 回归模型的残差方差为 $\pi^2/3$,所以:

$$\text{ICC}=\frac{\sigma^2_{\mu_0}}{\sigma^2_{\mu_0}+\pi^2/3}=\frac{\sigma^2_{\mu_0}}{\sigma^2_{\mu_0}+3.289868134} \quad (5)$$

第二步,建立包含解释变量的随机截距模型,即在随机截距的基础上再考察变量 drug 的固定效应。模型如下:$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right)=\beta_{0j}+\beta_1 drug_{ij}$ 和 $\beta_{0j}=\beta_0+\mu_{0j}$,上述模型可合并为:

$$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right)=\beta_0+\beta_1 drug_{ij}+\mu_{0j} \quad (6)$$

其中，$\beta_{0j}+\beta_1 drug_{ij}$为固定效应，$\mu_{0j}$为随机效应，且$\mu_{0j}\sim N(0,\sigma^2_{\mu_0})$。

第三步，建立包含解释变量的随机截距-斜率模型，即截距项和解释变量drug的系数均为随机系数，模型如下：$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right)=\beta_{0j}+\beta_{1j}drug_{ij}$，$\beta_{0j}=\beta_0+\mu_{0j}$和$\beta_{1j}=\beta_1+\mu_{1j}$，上述模型可合并为：

$$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right)=\left(\beta_0+\beta_1 drug_{ij}\right)+\left(\mu_{0j}+\mu_{1j}drug_{ij}\right) \quad (7)$$

其中，$\beta_0+\beta_1 drug_{ij}$为固定效应，$\mu_{0j}+\mu_{1j}drug_{ij}$为随机效应，且$\mu_{0j}\sim N(0,\sigma^2_{\mu_0})$，$\mu_{1j}\sim N(0,\sigma^2_{\mu_1})$，$\mu_{0j}$与$\mu_{1j}$之间的协方差可能有统计学意义；若无统计学意义，则将它们之间的协方差设为0。

例1较为特殊，其最终模型是包含解释变量drug的随机截距模型，但又与第二步所建模型略有不同，区别在于本资料截距项为0，即β_0=0。例1的最终模型中包含一个固定效应和一个随机效应，模型如下：

$$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right)=\beta_1 drug_{ij}+\mu_{0j} \quad (8)$$

4 基于SAS分析实例

4.1 分析与解答

在例1中，研究者欲考察试验药与对照药治疗某种感染的效果。资料中涉及两个原因变量——中心和药物种类，响应变量为二值变量，由于不同医院对同一疾病的治疗效果可能有差异，而在同一医院中，相同疾病的治疗效果也并不完全独立。故可考虑采用多水平logistic回归模型分析，SAS程序如下：

```
                                        /*1*/
data aa;
do zhongxin=1 to 15;
do drug=0 to 1;
do y=0 to 1;
input f @@;
do i=1 to f
output;
end;end;end;end;
datalines;
```

30 41 16 32 25 8 24 12 22 13 12 18 14 23 4 14 6 17 3 19 3 7 5 12 4 18 2 9 13 4 22 4 26 17 18 15 34 14 28 22 12 16 6 16 4 15 3 15 19 18 15 23 14 24 10 25 19 25 13 21

;

```
run;
ods html ;                              /*2*/
proc glimmix method=rspl;
class zhongxin;
model y(event='0')=/s dist=binary link=logit ddfm=bw;
random int/sub=zhongxin;
run;
proc nlmixed;                           /*3*/
parms b0=-0.3018 v_u0=0.5988;
z=b0+u0j;
if (y=0) then p=exp(z)/(1+exp(z)); else p=1-(exp(z)/(1+exp(z)));
ll=log(p);
model y~general(ll);
random u0j~normal(0,v_u0) sub=zhongxin;
estimate'ICC' v_u0/(v_u0+3.289868134);
run;
proc glimmix method=rspl;               /*4*/
class zhongxin;
model y(event='0')=drug/s dist=binary link=logit ddfm=bw noint;
random int/sub=zhongxin;
run;
proc nlmixed;                           /*5*/
parms b1=-0.4343 v_u0=0.5984; z=b1*drug+u0j;
if (y=0) then p=exp(z)/(1+exp(z)); else p=1-(exp(z)/(1+exp(z)));
ll=log(p);
model y~general(ll);
random u0j~normal(0,v_u0) sub=zhongxin;
run;
ods html close;
```

【说明】程序共5步，包括1个数据步和4个过程步。程序第2步、第3步是建立不包含任何解释变量的空模型，以计算ICC值。程序第4、5步均是建立包含解释变量drug的随机截距模型。GLIMMIX过程的计算结果与NLMIXED过程的结果会略有差异，前者运算速度快、用法简单，但在评估模型拟合效果时使用虚拟的对数似然值，而非真实值，不能用于模型的比较，且SAS 9.4的中GLIMMIX过程没有提供随机效应的假设检验，其结果虽有随机系数方差的参数估计值及标准误，但两者的比值只能作为参考，不能采用t检验计算相应的P值。NLMIXED

过程可以提供真实的对数似然值，并为随机效应提供假设检验的结果，也可以通过似然比检验对嵌套模型的拟合效果进行比较，但其用法较复杂，需设置模型和参数的初始值，不便于使用。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。对于相对简单的模型而言，NLMIXED过程对参数初始值并不敏感，此时采用其默认的初始值1即可。调用NLMIXED过程运行包含解释变量的随机截距-斜率模型，所用程序与本节程序第5步有较大修改。参考程序如下：

```
proc nlmixed;
parms b0= b1= v_u0= cov_u01= v_u1= ; z=b0+b1*drug+u0j+u1j;
if (y=0) then p=exp(z)/(1+exp(z)); else p=1-(exp(z)/(1+exp(z)));
ll=log(p);
model y~general(ll);
random u0j u1j~normal([0, 0],[v_u0, cov_u01, v_u1]) sub=zhongxin;
run;
```

其中，“b0”“b1”“v_u0”“cov_u01”“v_u1”分别相当于公式(7)中β_0、β_1、μ_{0j}的方差，μ_{0j}与μ_{1j}之间的协方差，μ_{1j}的方差。

【主要输出结果及解释】以下是第一个过程步输出结果，即调用GLIMMIX过程运行空模型。模型构建是以”y=0”为基础的，即计算”y=0”发生的概率模型。

Fit Statistics

-2 Res Log Pseudo-Likelihood	4100.75
Generalized Chi-Square	926.05
Gener Chi-Square/DF	0.98

以上是模型拟合的有关信息，第一行即为-2倍的限制性/残差虚拟对数似然值，此统计量不能用于不同模型的比较。

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
Intercept	zhongxin	0.5988	0.2686

以上是协方差参数估计的结果，给出了随机效应方差的估计值及相关假设检验的结果。可见随机截距方差(即$\sigma^2_{\mu_0}$)的估计值为0.5988，标准误为0.2686。但此处未给出随机截距方差是否为0的假设检验结果。故没有客观依据判定$\sigma^2_{\mu_0}$与0之间的差异是否有统计学意义。

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t value	Pr>|t|
Intercept	-0.3018	0.2138	14	-1.41	0.1798

以上是固定效应的解。因为此过程步运行的是空模型，所以这里只有一个固定效应，即截距，其值为-0.3018，表示y=0的总平均logit值为-0.3018。

以下是模型拟合的有关信息，包括三种信息标准的估计值和-2倍的对数似然值。这些统计量本身不能说明模型拟合的优劣，但可用于含相同自变量数目的不同模型的比较。

Fit Statistics

-2 Log Likelihood	1233.8
AIC(smaller is better)	1237.8
AICC(smaller is better)	1237.9
BIC(smaller is better)	1239.3

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr>|t|	95% Confidence Limits		Gradient
b0	-0.3079	0.2105	14	-1.46	0.1656	-0.7593	0.1435	4.202E-7
v_u0	0.5747	0.2538	14	2.26	0.0399	0.0304	1.1190	1.601E-6

以上是模型中参数的估计结果，包括固定效应和随机效应的参数估计值及相应的假设检验结果。注意，随机效应假设检验给出的是双侧检验的结果，而实际上检验方差是否为0应采用单侧检验，故此处所得的P值应除以2才是正确的P值，后同。“v_u0”对应的P值为0.0399/2<0.05，说明$\sigma^2_{\mu_0}$与0之间差异有统计学意义，分析时应采用多水平logistic回归模型分析。

Additional Estimates

Label	Estimate	Standard Error	DF	t Value	Pr>|t|	Alpha	Lower	Upper
ICC	0.1487	0.05590	14	2.66	0.0187	0.05	0.02881	0.2686

以上是ICC的计算结果，其值为0.1487，对应的P值为0.0187<0.05，说明数据存在一定的组内同质性，需采用多水平logistic模型分析该资料。

以下是第三个过程步的输出结果，即调用GLIMMIX过程运行含解释变量drug的随机截距模型的结果，其截距项设定为0。可见，$\sigma^2_{\mu_0}$值为0.5984，解释变量系数为-0.4343。

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
Intercept	zhongxin	0.5984	0.20609

Solutions for Fixed Effects

Effect	Estimate	Standard Erro	DF	t Value	Pr>|t|
drug	-0.4343	0.1341	927	-3.24	0.0012

以下是第四个过程步的输出结果，即调用NLMIXED过程运行含解释变量drug的随机截距模型的结果，其截距项设定为0。可见，解释变量系数为-0.4409（P=0.0057<0.05），$\sigma^2_{\mu_0}$值为0.6221（P=0.0390<0.05），二者与0的差异均有统计学意义。另外，由“Fit Statistics”部分结果可知，-2倍对数似然值为1225.1，略大于第四步构造的模型；但由“Parameter Estimates”部分结果可知，模型中共包含两个参数，参数个数较之前少了一个，且两个模型的拟合效果并无统计学差异（χ^2=1225.1-1224.9=0.2，P>0.05）。所以，使用此模型更合适。

Fit Statistics

-2 Log Likelihood	1225.1
AIC (smaller is better)	1229.1
AICC (smaller is better)	1229.1
BIC (smaller is better)	1230.6

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr>|t|	95% Confidence Limits		Gradient
b1	-0.4409	0.1352	14	-3.26	0.0057	-0.7308	-0.1509	0.000039
v_u0	0.6221	0.2732	14	2.28	0.0390	0.0362	1.2081	0.000237

解释变量drug的系数为-0.4409，且与0的差异有统计学意义（P=0.0057），说明试验药组与对照药组的疗效之差具有统计学意义。因exp(-0.4409)=0.6435，所以对照药组治疗成功率是试验药组治疗成功率的0.6435倍。随机截距的方差“v_u0”估计值为0.6221，与0的差异有统计学意义（P=0.0390/2<0.05），说明水平1截距跨中心变异显著，即不同中心μ_{0j}值存在差异。

5 讨 论

一般而言，响应变量为二值变量的高维列联表资料采用一般logistic回归分析，但此法要求所有观测结局相互独立。对于研究个体存在聚集性特征时，应采用多水平模型。这样可将传统模型中的随机误差分解到数据层级结构相应的水平上，使得个体的随机误差更纯[5]。

采用PROC GLIMMIX和PROC NLMIXED过程来构建模型：首先建立不包含任何解释变量的空模型，以计算ICC值。若存在组内相关，则构建截距项不为0的模型。若经检验得到此截距项与0的差异没有统计学意义，则构建截距项为0的模型。

参考文献

[1] 冯国双. 白话统计[M]. 北京：电子工业出版社，2018：112-119.

[2] 杨珉，李晓松. 医学和公共卫生研究常用多水平统计模型[M]. 北京：北京大学医学出版社，2007：69-91.

[3] 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，2008：128-149.

[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.

[5] 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.

（收稿日期：2019-09-27）

（本文编辑：陈 霞）