非配对设计二值资料一水平多重Logistic回归分析

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】 本文的目的是介绍非配对设计二值资料一水平多重Logistic回归模型的构建与求解方法。基于SAS软件分别对以列联表和数据库形式呈现的定性资料进行全面分析，并得出了4个对提高模型拟合优度很有价值的结论：第一，若资料以列联表形式呈现，应拟合“加权”Logistic回归模型；第二，若资料中包含定量自变量，不适合将其定性化；第三，若资料中包含定量自变量，应依据定量自变量和二值自变量产生出派生自变量；第四，若资料中有定性自变量时，必须将多值名义或有序自变量进行哑变量变换，不需要依据二值自变量产生出派生自变量。

【关键词】 二值资料；一水平；派生变量；加权回归；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.1007-3256.2019.04.002

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

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 MedicineSocieties*, *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 model with binary data collected from the unpaired design. Based on SAS software package, the qualitative data presented in the form of contingency tables and databases were analyzed comprehensively and thoroughly, and four valuable conclusions were obtained to improve the goodness of fit. First, if the data were presented as contingency tables, the weighted Logistic regression model should be fitted. Second, if the data contained quantitative independent variables, which were not suitable for transforming into qualitative variables. Third, if the data contained quantitative independent variables, derived independent variables should be generated according to quantitative independent variables and binary independent variables. Fourth, if there were qualitative independent variables in the data, the multi-valued nominal or ordered independent variables should be transformed into dummy variables, and the derived independent variables did not need to be generated according to the binary independent variables.

【**Keywords**】 Binary data; One-level; Derived variables; Weighted regression; Multiple Logistic regression analysis

1 本文题目中有关名词概念

1.1 非配对设计

所谓“非配对设计”，就是未采用“配对设计”。实际上就是“单组设计”，即认为所有受试对象来自一个总体。

1.2 二值资料

这里的“二值资料”特指结果变量只有两种不同取值的资料，例如治疗结局中的“存活”与“死亡”。

项目基金：国家高技术研究发展计划课题资助（2015AA020102）

1.3 一水平

“一水平”即“一个层次”。医学资料常涉及“性别”“年龄”“接受治疗的方法”“正常人与患者”和“死亡与存活”等，受这些“原因与结果”影响的受试者似乎应该属于“多层次”的，为何大多数人都采用“一水平统计分析方法（注意：人们常用的统计分析方法的名称中并没有明确写出‘一水平’的字样，其实它是隐含的）”来处理资料呢？其实，“一水平”与“多水平”是指：所构建的“变量之间的依赖关系（例如：$\hat{y} = 3 + 0.5x$）”是否适合样本中的“所有个体”。若适合，可认为所有个体来自“一个总体”或“一个

层次”，这就是一个“一水平”的资料或问题；反之，就认为所有个体来自“两个”或“多个”“总体或层次”，这就需要构建“多水平回归模型”了[1-4]。

1.4 多重Logistic回归分析

“多重”是指原因变量或自变量的个数≥2；而Logistic回归分析是指专门构建定性结果变量（包括“二值”“多值有序”和“多值名义”结果变量）依赖一个或多个原因变量变化而变化的依赖关系的一类回归分析方法，它与“多重线性回归分析”的本质区别在于：前者属于“广义线性回归模型（实际就是非线性的）”，需要采用“最大似然法”和“迭代计算”等复杂方法来估计回归模型中“参数（截距和斜率）”的数值；而后者属于“一般线性回归模型”，只需要采用“普通最小二乘法”进行求解。

1.5 本文题目的简化表述

文献中常见的Logistic回归分析[5-7]基本上都属于“非配对设计二值资料一水平多重Logistic回归分析”，因为没有采取“配对设计”、结果变量为“二值变量”且隐含的假定为“样本中的全部受试对象来自同一个总体”（即“一水平”），自变量的个数通常都≥2（即“多重”）。于是，本文题目简化表述为“一般多重Logistic回归分析”或“一般多因素Logistic回归分析”。前面表述中的“一般”就是指“二值资料”，因为它出现的频率很高，而且，它又是“多值有序资料”和“多值名义资料”的重要基础。由此可知，本文题目甚至可以简化为“多重Logistic回归分析”或“多因素Logistic回归分析”。为了区别于“配对设计二值资料Logistic回归分析”，本文中所介绍的方法也被称为“非条件Logistic回归分析”。

2 两种形式的数据结构

2.1 基于列联表形式呈现的数据结构

【例1】某医学研究机构收集到来自三个医疗中心（东京、波士顿和格拉摩根）的乳腺癌患者的有关资料[8]。见表1。表1前4列为4个影响因素，分别有3、3、2、2个水平；最后2列为“二值结果变量（死、活）及其频数”。

2.2 基于数据库形式呈现的数据结构

【例2】某临床医生收集到前来就诊患者的有关资料。见表2。

表1 四个因素影响下乳腺癌患者三年存活情况的调查结果（n=764）

医疗中心	患者年龄（岁）	慢性炎症反应程度	核的量级	例数	
				三年结局：	死	活
东京	<50	轻	恶性		9	26
			良性		7	68
		重	恶性		4	25
			良性		3	9
	50~	轻	恶性		9	20
			良性		9	46
		重	恶性		11	18
			良性		2	5
	70~	轻	恶性		2	1
			良性		3	6
		重	恶性		1	5
			良性		0	1
波士顿	<50	轻	恶性		6	11
			良性		7	24
		重	恶性		6	4
			良性		0	0
	50~	轻	恶性		8	18
			良性		20	58
		重	恶性		3	10
			良性		2	3
	70~	轻	恶性		9	15
			良性		18	26
		重	恶性		3	1
			良性		0	1
格拉摩根	<50	轻	恶性		16	16
			良性		7	20
		重	恶性		3	8
			良性		0	1
	50~	轻	恶性		14	27
			良性		12	39
		重	恶性		3	10
			良性		0	4
	70~	轻	恶性		3	12
			良性		7	11
		重	恶性		3	4
			良性		0	1

3 一水平二值资料多重Logistic回归模型

3.1 回归模型的构建

因变量Y取值0和1分别表示阴性与阳性结果，X_1、X_2、…、X_m代表m个自变量。设$P(Y=1|X_1, X_2, \cdots, X_m)$表示在自变量$X_1$、$X_2$、…、$X_m$存在的条件下出现阳性结果的概率（$P$）。多重Logistic回归模型表示为：

$$P=\frac{\exp(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m)}{1+\exp(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m)} \quad (1)$$

式中β_0为常数项，β_1、β_2、…、β_m分别为各个自变

量所对应的回归系数。与之等价的模型为：

$$P=\frac{1}{1+\exp[-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m)]} \quad (2)$$

阳性结果发生的概率为P，则阴性结果发生的概率为1−P。P与1−P之比叫做优势，记作“od”，对od值取自然对数，称为对P作logit变换，用logit(P)表示：

$$\text{logit}(P)=\ln\frac{P}{1-P} \quad (3)$$

此Logistic模型又可以表示为如下形式：

$$\text{logit}(P)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m \quad (4)$$

概率P与自变量X_1、X_2、…、X_m之间的关系是非线性的，但logit(P)与自变量之间呈线性关系。

Logistic回归模型中各参数都有明确的实际意义，回归系数β_i表示在其他影响因素（即自变量）不变的情况下，自变量X_i每变化一个单位时所引起的logit(P)的改变量，X_i分别在其两个不同水平对应的优势od_1与od_2之比记为OR_i，其计算公式为：

$$OR_i=\exp\beta_i \quad (5)$$

当某疾病的发病率或死亡率很低时，可用优势比OR近似估计相对危险度RR。应注意：由一个多值名义或有序变量产生的K−1个“哑变量”的“OR值”不能用公式(5)计算，因为K−1个“哑变量”都以另一个“未出现的水平”为“基准”，原变量的“K个水平”的回归系数之和必须为0。所以，求每个哑变量“OR值”时，实际上是所求水平与基准水平的“OR值”之比。

设某多值名义或有序变量W有4个水平，假定以其第一水平为“基准或参照水平”，产生3个“哑变量”：W_2、W_3和W_4，假定它们的回归系数分别为b_2、b_3和b_4，于是，W的第一水平的回归系数（不显示出来）$b_1=-(b_2+b_3+b_4)$。此时，若希望求出三个哑变量对应的“OR值”，就需要按如下公式计算：

$$OR_{w_2}=\exp(b_2)/\exp(b_1) \quad (6)$$

$$OR_{w_3}=\exp(b_3)/\exp(b_1) \quad (7)$$

$$OR_{w_4}=\exp(b_4)/\exp(b_1) \quad (8)$$

表2 五种疾病的患者是否出现晕厥对结局（死与活）的影响（n=2789）

编号	年龄(age)	性别(sex)	是否晕厥(yj)	疾病种类(jb)	结局(sw)
1	80	女	否	1	存活
2	58	男	否	1	存活
3	72	男	否	1	存活
…	…	…	…	…	…
2787	76	女	晕厥	5	死亡
2788	73	女	晕厥	5	存活
2789	58	男	否	5	死亡

注：将“age”定性化后改名为“ag”，ag=1（<30岁）、ag=2（30~60岁）、ag=3（>60岁）；将“sex”改名为“sb”，sb=0（女）、sb=1（男）；将“yj”改名为“hj”，hj=0（否）、hj=1（晕厥）；将“sw”改名为“y”，y=0（死亡）、y=1（存活）

3.2 回归模型的求解

3.2.1 Logistic回归模型的参数估计和假设检验

在Logistic回归分析中，参数的估计通常采用最大似然法。首先建立样本的似然函数，取似然函数的自然对数，得到对数似然函数，将其作为目标函数。用对数似然函数代替似然函数的目的在于：求似然函数极大值的过程比较困难，使用对数似然函数可以简化计算。然后求对数似然函数关于各个参数的一阶偏导数并使之为0，便得到了似然方程组。最后再解似然方程组，就可以得到参数的估计值，由于似然方程组是非线性的，需要使用迭代方法对其求解，通常使用的是*Newton-Raphson*方法。

在估计出回归系数以后，要对其进行假设检验，包括对全部回归系数是否均为0进行检验和对单个回归系数进行检验，常用的检验方法有似然比检验、计分检验和*Wald*检验。

3.2.2 多重Logistic回归模型中自变量的筛选方法

与多重线性回归分析一样，多重Logistic回归分析中也需对自变量进行筛选，只保留对因变量具有统计学意义的自变量。筛选自变量的方法主要有前进法、后退法、逐步法和最优子集法。与多重线性回归不同的是，筛选自变量时所用的检验统计量不再是F统计量，而是似然比统计量、计分统计量和*Wald*统计量之一。

3.2.3 回归模型的拟合优度检验

建立回归模型后，往往需要对模型做出评价，考察模型与实际数据的符合情况，称为拟合优度检验。在Logistic回归分析中，用于拟合优度评价的统计量主要包括Pearsonχ^2、偏差、Hosmer-Lemeshow统计量和一些信息测量指标。

Pearsonχ^2通过比较模型预测的和实际观察到的事件发生与不发生的频数检验模型成立的假设。当该统计量很小时，对应的P值大于规定的显著性水平，显示预测值和观测值之间差异无统计学意义，说明模型较好地拟合了数据；如果该统计量很大，P值小于显著性水平，则说明拟合效果不佳。

偏差统计量在样本含量较大时服从χ^2分布。与Pearsonχ^2相似，当偏差统计量较小，若$P>0.05$，说明拟合效果较好；反之则提示拟合效果较差。

当自变量数量增加时，尤其是连续自变量纳入模型之后，自变量组合方式的数量便会很大，于是许多组合方式下只有很少的观测例数，在这种情况

下，Pearsonχ^2和偏差不再适用于评价拟合优度。此时可以采用Hosmer-Lemeshow统计量来度量模型的拟合优度。Hosmer-Lemeshow统计量是一种类似于Pearsonχ^2统计量的指标，对应的P值大于规定的显著性水平，说明拟合较好；反之，则拟合不好。

信息测量指标包括*Akaike*信息准则（*AIC*）、贝叶斯信息准则（*BIC*）、-2倍对数似然函数值L（简记为“-2L”）和ROC曲线下的面积AUC，在这四个指标中，前三个取值越小，说明模型对资料的拟合度越高；而AUC取值越大，说明模型对资料的拟合度越高。

4 基于SAS分析实例

4.1 基于SAS分析例1

4.1.1 分析策略

策略一：基于原始数据集a1（指仅采用表1中的4个原因变量和1个二值结果变量以及各网格上的频数），不考虑“权重”，并采用前进法、后退法和逐步法筛选自变量。

策略二：基于原始数据集a1，考虑“权重”（即在LOGISTIC过程步中，增加“weight f;”语句，其中，“f”代表各网格上的“频数”），并采用前进法、后退法和逐步法筛选自变量。

策略三：基于原始数据集a1，引入5个派生自变量，形成含11个自变量（即4个哑变量、2个二值变量、5个派生变量）的数据集a2，基于此数据集，不考虑“权重”，并采用前进法、后退法和逐步法筛选自变量。

策略四：基于原始数据集a1，引入5个派生自变量，形成含11个自变量的数据集a2，基于此数据集，考虑“权重”（即在LOGISTIC过程步中增加“weight f;”语句，其中，“f”代表各网格上的“频数”），并采用前进法、后退法和逐步法筛选自变量。

【说明】产生派生变量的方法：由2个二值原因变量产生它们的平方项（2个）、交互项（1个）、立方项（2个），共5个派生变量。

4.1.2 以摘要形式呈现的分析结果

SAS输出结果很多，为节省篇幅，下面仅以“ROC曲线下面积AUC”作为模型对资料的拟合效果的评价指标，基于某个多重Logistic回归模型算得的AUC越大，表明此回归模型对资料的拟合优度越高，计算结果见表3。

表3 基于4种分析策略对例1进行多重Logistic回归分析的部分结果

分析策略	参入自变量数	保留自变量数	AUC值	AUC 95% *CI*
策略一	6	3	0.6014	0.5582~0.6446
策略二	6	5	0.6088	0.5656~0.6520
策略三	11	3	0.6014	0.5582~0.6446
策略四	11	5	0.6068	0.5656~0.6520

【说明】以下呈现出四个策略所对应的模型的“-2倍对数似然函数值（-2L）”，模型中含相同个数的参数时，此值越小，代表模型对资料的拟合优度越高。

策略一：-2L=877.564，$df_{模型}$=4（包括截距项）；

策略二：-2L=12401.619，$df_{模型}$=6（包括截距项）；

策略三：-2L=877.564，$df_{模型}$=4（包括截距项）；

策略四：-2L=12391.808，$df_{模型}$=7（包括截距项）。

策略二对应的AUC值最大，此时，二值资料一水平多重Logistic回归模型的参数估计结果如下：

最大似然估计值分析

参数	自由度	估计值	标准误差	Wald卡方	Pr>卡方
Intercept	1	-1.1876	0.1958	36.8029	<0.0001
cent 2	1	0.4225	0.1846	5.2387	0.0221
cent 3	1	0.3725	0.1829	4.1476	0.0417
age 2	1	-0.2911	0.1656	3.0884	0.0789
age 3	1	0.6034	0.2248	7.2035	0.0073
he	1	-1.0846	0.2570	17.8094	<0.0001

由此，多重Logistic回归模型表达式为：

$$P(y=0)=\frac{e^{Z}}{1+e^{Z}}$$

上式中，y=0代表“死亡”，$Z=-1.1876+0.4225cent_2+0.3725cent_3-0.2911age_2+0.6034age_3-1.0846he$

上述回归模型中，“$cent_2$”与“$cent_3$”是由“医疗中心cent”产生的两个哑变量，它们都是相对于第一个医疗中心（$cent_1$）而言的；同理，“age_2”与“age_3”是由“年龄组age”产生的两个哑变量，它们都是相对于第一个年龄组（age_1，即“<50岁”）而言的；最后

一个变量“he”代表“核的量级”，它本身就是一个“二值变量”，he=0（恶性）、he=1（良性）。上述回归模型中各变量的“优势比估计值”的计算结果如下：

优比估计值

效应	点估计值	95% CI Wald
cent 1-2	3.379	1.700~6.715
cent 1-3	3.214	1.626~6.355
age 1-2	1.021	0.562~1.856
age 1-3	2.498	1.135~5.501
he	0.338	0.204~0.559

以上呈现的是回归模型中各自变量的“优势比 *OR* 值的估计值”及其“95% *CI*（按 Wald 方法计算）”，其中，“cent 1-2”和“cent 1-3”分别代表“cent 的 2 水平与 1 水平比较”和“cent 的 3 水平与 1 水平比较”；同理，可理解“age 1-2”与“age 1-3”；而“he”本身是一个独立的“二值变量”，其 *OR* 值可直接按公式（5）计算出来：

$$OR_{he} = \exp(-1.0846) = 0.338$$

但是，“cent 1-2”和“cent 1-3”这两行的 *OR* 值就需要借助公式（6）和（7）来计算了。首先，需要求出“cent 取一水平”时的“回归系数 b_1”：

$$b_1 = -(0.4225 + 0.3725) = -0.7950$$

$$\text{cent } 1-2=\frac{\exp(0.4225)}{\exp(-0.7950)}=\exp(1.2175)=3.378$$

$$\text{cent } 1-3=\frac{\exp(0.3725)}{\exp(-0.7950)}=\exp(1.1675)=3.214$$

同理，读者可参照以上做法，计算出上面与“age 1-2”与“age 1-3”对应的“*OR* 值”，此处不再赘述。

4.1.3 表 3 中 4 种分析策略下计算结果优劣比较

比较“策略一”与“策略二”、“策略三”与“策略四”可知：增加“weight f;”语句，即采用“加权”的多重 Logistic 回归模型比采用“未加权”的多重 Logistic 回归模型计算，ROC 曲线下的面积 AUC 会有所增大。

比较“策略一”与“策略三”、“策略二”与“策略四”可知：在模型中引入由“二值自变量”产生的派生变量，对模型拟合优度的提高几乎没有任何作用，有时，甚至还可能有所降低。

4.1.4 分析例 1 所需要的 SAS 程序

4.1.4.1 策略一和策略二所需要的 SAS 数据步程序

```
data a1;
do cent=1 to 3;do age=1 to 3;
do yz=0 to 1;do he=0 to 1;
do y=0 to 1;
    input f @@; output;
end;end;end;end;end;
cards;
（此处输入表 1 中 36 行 2 列数据）
;
run;
```

4.1.4.2 策略一所需要的 SAS 过程步程序

```
proc logistic data=a1;
    class cent(ref="1") age(ref="1");
    model y=cent age yz he/selection=backward sls=0.05;
    freq f;
    roc;
run;
```

4.1.4.3 策略二所需要的 SAS 过程步程序

在前面的 SAS 过程步程序（见 4.1.4.2 小节）的“roc;”语句前增加一个“weight f/norm;”语句即可。其中，选项“/norm”是使“权重”正规化，即确保总权重之和等于总频数，下同。

4.1.4.4 策略三和策略四所需要的 SAS 数据步程序

在数据集 a1 的基础上，依据三个二值自变量产生出 5 个派生变量，形成数据集 a2：

```
data a2;
    set a1;
x1=yz*yz; x2=yz*he;
x3=he*he; x4=yz*yz*yz;
x5=he*he*he;
run;
```

【说明】最好依据“定量自变量”产生派生自变量，本例中无“定量自变量”，尝试依据“二值自变量”来产生派生变量，但不适合依据“多值名义自变量”和“多值有序自变量”来产生派生自变量。

4.1.4.5 策略三所需要的 SAS 过程步程序

```
proc logistic data=a2;
    class cent(ref="1") age(ref="1");
    model y=cent age yz he x1-x5/selection=backward sls=0.05;
    freq f;roc;
run;
```

4.1.4.6 策略四所需要的 SAS 过程步程序

在前面的 SAS 过程步程序（见 4.1.4.5 小节）的

“roc;”语句前增加一个“weight f/norm;”语句即可。

【说明】以上SAS模型语句中，采用的是后退法（backward），若换成前进法（forward）和逐步法（stepwise），计算结果完全相同。

4.2 基于SAS分析例2

4.2.1 分析策略

策略一：完全基于原始数据集a1（指仅采用表2中的4个原因变量和1个二值结果变量及其取值），即把“年龄（age）”视为“计量资料”；把“疾病种类（jb）”视为“5分类变量”，产生出4个“哑变量”。基于此数据集，分别采用前进法、后退法和逐步法筛选自变量。

策略二：在数据集a1的基础上，将“年龄（age）”定性化为ag，即结合专业知识，将年龄按“<30岁”“30岁≤年龄≤60岁”和“>60岁”划分为三档，记为ag=1、2、3；进而，将ag产生2个哑变量。由此，形成数据集b1。基于此数据集，分别采用前进法、后退法和逐步法筛选自变量。

策略三：基于数据集a1，引入9个派生自变量（由age、xb和hj生成平方项、交叉乘积项和立方项，x1-x9），形成含16个自变量（age、xb、hj、x1-x9和由jb产生的4个哑变量）的数据集a2。基于此数据集，分别采用前进法、后退法和逐步法筛选自变量。

策略四：基于数据集b1，引入5个派生自变量（由xb和hj生成平方项、交叉乘积项和立方项，w1-w5），形成含13个自变量（xb、hj、w1-x5、由jb产生的4个哑变量和由ag产生的2个哑变量）的数据集b2。基于此数据集，分别采用前进法、后退法和逐步法筛选自变量。

4.2.2 以摘要形式呈现的分析结果

计算结果见表4。

表4 基于4种分析策略对例2进行多重Logistic回归分析的结果

分析策略	参入自变量数	保留自变量数	AUC值	AUC 95% *CI*
策略一	7	6	0.8486	0.8193~0.8778
策略二	8	8	0.8298	0.8001~0.8596
策略三	14	9	0.8582	0.8305~0.8860
策略四	13	8	0.8298	0.8001~0.8596

【说明】4种分析策略对应的多重Logistic回归模型的拟合优度评价指标之一，即“-2倍对数似然函数L的数值”（简写成“-2L”）（含相同数目参数的条件下，取值越小越好）如下：

策略一：-2L=887.457，$df_{模型}$=7（包括截距项，其中jb的df=4）；

策略二：-2L=912.273，$df_{模型}$=8（包括截距项，其中jb的df=4、ag的df=2，ag代表三个年龄段）；

策略三：-2L=867.939，$df_{模型}$=10（包括截距项，其中jb的df=4）；

策略四：同“策略二”，此处从略。

4.2.3 表4中4种分析策略计算结果优劣比较

“策略一”与“策略二”可比，它们都基于原始自变量，区别仅在于是否将计量变量“年龄（age）”定性化。由表4的第一、二行以及表4后的“说明”中的“第（1）行与第（2）行”可知：“策略一”优于“策略二”（AUC值大好、-2L值小好），即不适合将“计量自变量”转换成“多值有序自变量”。

“策略三”与“策略四”可比，它们都在前面分析“策略一”与“策略二”的基础上，引入了“派生自变量（由计量变量和二值变量产生，不适合由多值名义或多值有序变量产生）”。由表4的第三、四行以及表4后面的“说明”中的“第（3）行与第（4）行”可知：“策略三”优于“策略四”（AUC值大好、-2L值小好），即不适合将“计量自变量”转换成“多值有序自变量”。

“策略一”与“策略三”可比，它们都没有将计量自变量定性化，区别仅在于是否引入派生自变量，由表4中第一行与第三行以及表4后面的“说明”中的“第（1）行”与“第（3）行”可知：“策略三”优于“策略一”（AUC值大好、-2L值小好），即应该尽可能引入派生自变量。

“策略二”与“策略四”可比，但它们的结果完全相同，即将计量自变量定性化后，仅基于某些“二值变量”产生派生自变量，对模型的拟合优度几乎没有改善。

因篇幅所限，关于例2的回归模型的详细计算结果此处从略。

4.2.4 分析例2所需要的SAS程序

所需要的SAS程序与“分析例1的SAS程序”基本相同，此处从略。

5 讨论与小结

5.1 对“定量自变量”是否应定性化

在进行回归建模时，对于“定量自变量”是否需要定性化，人们持有不同的观点。例2的分析结果表明：不将其定性化，有利于提高模型的拟合优度。

5.2 是否应当引入派生自变量

是否需要引入派生自变量参与自变量的筛选，人们的习惯做法是不引入派生自变量。例 1 和例 2 的分析过程及分析结果表明：当自变量中含有“定量自变量”时，引入派生变量参与自变量的筛选，有可能提高模型的拟合优度；而当自变量中没有“定量自变量”时，仅依据“二值自变量”产生派生自变量，对提高模型的拟合优度没有价值。

5.3 分析列联表资料时是否需要进行“加权回归分析”

在对高维列联表资料进行多重 Logistic 回归分析时，常规的做法是构建“未加权”的回归模型。但事实上，采取“加权回归分析”有利于提高模型的拟合优度。

一般来说，如果能找到“合适的权重系数”，基于加权回归分析的效果可能会优于未加权的回归分析的效果。无论是简单直线回归分析、多重线性回归分析还是曲线回归分析（因为 Logistic 回归分析就是曲线回归分析中的一种），都是如此[9]。

5.4 小结

在构建二值资料一水平多重 Logistic 回归模型的过程中，把握好以下四点有利于提高模型对资料的拟合优度：第一，若资料以列联表形式呈现，应拟合“加权”Logistic 回归模型；第二，若资料中包含定量自变量，不适合将其定性化；第三，若资料中包含定量自变量，应依据定量自变量和二值自变量产生出派生自变量；第四，若资料中有定性自变量时，必须对多值名义或有序自变量进行哑变量变换，不需要依据二值自变量产生出派生自变量。

参考文献

[1] 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.

[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，2008：128-149.

[3] 王济川，谢海义，费舍尔. 多层统计分析模型：SAS 与应用[M]. 北京：高等教育出版社，2009：116-136.

[4] 石磊. 多水平模型及其统计诊断[M]. 北京：科学出版社，2008：27-50.

[5] 陈邦定，彭东桃，阳波，等. 常德地区严重精神障碍患者暴力攻击行为研究[J]. 四川精神卫生，2019，32(1)：53-57.

[6] 弋可，曾强，吴俊林，等. 绵阳市 COPD 患者焦虑与抑郁检出情况及相关因素分析[J]. 四川精神卫生，2018，31(6)：526-530.

[7] 朱意平，李春阳，陈红红，等. 住院精神分裂症患者合并代谢综合征的影响因素[J]. 四川精神卫生，2018，31(6)：540-543.

[8] 胡良平. 统计学三型理论在统计表达与描述中的应用[M]. 北京：人民军医出版社，2008：160-162.

[9] 胡良平，胡纯严，鲍晓蕾. 应用数理统计[M]. 北京：电子工业出版社，2015：155-184.

（收稿日期：2019-08-01）

（本文编辑：陈 霞）