M∶N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m∶n配对设计二值资料一水平多重Logistic回归分析方法。首先,介绍了需要了解的基本概念;其次,介绍了构建此类回归模型的基本原理;最后,通过一个实例介绍了使用SAS实现计算的全过程。在此过程中,获得了如下四点启示:其一,有必要确保所获得的科研资料是值得分析的;其二,有必要基于定量自变量产生派生变量;其三,有必要同时采用"逐步法""前进法"和"后退法"筛选自变量;其四,有必要采用多种方法评价不同回归模型对资料的拟合优度。

【关键词】 M∶N配对设计;派生变量;条件概率;似然函数;迭代计算
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2019.04.004

One-level multiple Logistic regression analysis of the dichotomous choice data collected from the m∶n paried 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 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 approach of multiple Logistic regression analysis for the binary data collected from the m∶n paired design. Firstly, the basic concepts that need to be understood were introduced. Secondly, the basic principle of constructing such regression model was presented. Finally, the whole process of computing with SAS was shown by an example. In this process, the following four useful enlightenments had been obtained: ① it was necessary to ensure that the scientific research data obtained were worth analyzing; ② it was necessary to generate the derived variables based on quantitative independent variables; ③ it was necessary to adopt the "stepwise method" "forward method" and "backward method" to screen the independent variables at the same time; ④ it was necessary to use several approaches to evaluate the goodness of fit of the different regression models.

【Keywords】 M∶N paired design; Derived variable; Conditional probability; Likelihood function; Iterative computation

1 基本概念

1.1 引言

在涉及"病例对照"的配对设计中,若病例数目相对较少,而对照数目比较多,研究设计者倾向于采取"1∶2"或"1∶r"的配对形式。然而,当病例和对照的数目都较多时,为了增大样本含量,提高研究的效率,研究设计者可能倾向于选择"m∶n"的配对形式。这里的"m"和"n"都是大于等于1的正整数,而且,在各个"匹配组"中,"m"和"n"最好取相同的数值,但也可取不同的数值。显然,后者更符合实际工作的需要。

1.2 合理确定配对条件

本期科研方法专题的第三篇文章提出了提高配对设计质量的六个要领,研究设计者在进行科研设计时,务必高度重视"合理确定配对条件";除此之外,还应有足够大的样本含量以及较好的样本代表性。

1.3 重视产生派生变量

当自变量中有定量变量时,应重视产生派生变量(即取对数、平方根、平方、立方和交叉乘积项等),将原有自变量和派生自变量(还包括基于多值名义或有序自变量产生的哑变量)全部纳入"自变

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

量筛选”；筛选自变量时，应尽可能多采取一些筛选策略，至少应包括“前进法”“后退法”和“逐步法”[1-7]。

1.4 关于本文的题目

本文的题目可能稍显“冗长”，其最精简的表述为“条件 Logistic 回归分析”。之所以采用现在这个题目，是希望把“设计类型（指‘M∶N 配对设计’）”“资料类型（指‘二值资料’）”“水平数（指‘一水平’）”“自变量数目（指‘多重’）”和“回归分析类型（指‘Logistic 回归分析’）”都一目了然地呈现出来。

2 回归模型的构建

M∶N 配对设计二值资料一水平多重 Logistic 回归分析的基本原理与 1∶1 及 1∶r 配对设计相同，只是由于每个匹配组中的病例数和对照数都有可能大于 1，并且不同匹配组中的受试对象数可能不相等，条件概率及条件似然函数的形式会比前两者略微复杂[8-9]，现简述如下。

设共有 h 个匹配组，编号为 1~h；第 i 个匹配组中有 m_i 例病例和 n_i 例对照，共（m_i+n_i）例受试对象，编号为 1，2，…，m_i，m_i+1，m_i+2，…，m_i+n_i，其中前 m_i 例代表病例，后 n_i 例代表对照；因变量取值为 1 代表病例，取值为 0 代表对照；自变量的个数为 p，编号为 1~p；自变量的观测值用 x_{111} 表示，下标中的第一个数字代表匹配组号，第二个数字代表受试对象的组内编号，第三个数字代表自变量的编号；向量 $\mathbf{X}_{it}=(x_{it1},x_{it2},\cdots,x_{itp})$ 表示第 i 个匹配组中第 t 个受试对象自变量的观察值。

在第 i 个匹配组内的（m_i+n_i）个受试对象中有 m_i 名病例的条件下，恰好前 m_i 个受试对象属于病例组的条件概率见式（1）：

$$P_i=\frac{\prod_{t=1}^{m_i}P(\mathbf{X}_{it}|Y=1)\prod_{t=m_i+1}^{m_i+n_i}P(\mathbf{X}_{it}|Y=0)}{\sum_{j}[\prod_{t_j=1}^{m_i}P(\mathbf{X}_{it_j}^{(j)}|Y=1)\prod_{t_j=m_i+1}^{m_i+n_i}P(\mathbf{X}_{it_j}^{(j)}|Y=0)]} \quad (1)$$

式（1）中分子为当前样本出现的概率，分母为各种可能组合情况下的概率之和，所谓各种可能组合，是指（m_i+n_i）个受试对象中有 m_i 名病例的所有可能组合共有 $C_{m_i+n_i}^{m_i}$ 种。t 表示受试对象的实际编号，t_j 表示第 j 种可能组合下受试对象的新编号，$\mathbf{X}_{it_j}^{(j)}$ 为相应编号为 t_j 的自变量观察值向量。

根据 Bayes 原理对式（1）进行转换，将式（2）中的 Logistic 回归模型表达式代入式（1）后化简得到式（3）：

$$P(Y=1|\mathbf{X}_{it})=P_i(Y=1|\mathbf{X}_{it})=\frac{1}{1+\exp[-(\beta_{0i}+\sum_{k=1}^{p}\beta_k x_{itk})]} \quad (2)$$

$$P_i=\frac{\prod_{t=1}^{m_i}\exp(\sum_{k=1}^{p}\beta_k x_{itk})}{\sum_{j}\prod_{t_j=1}^{m_i}\exp(\sum_{k=1}^{p}\beta_k x_{it_jk}^{(j)})} \quad (3)$$

综合 h 个匹配组，得到总的条件似然函数见式（4）：

$$L=\prod_{i=1}^{h}\frac{\prod_{t=1}^{m_i}\exp(\sum_{k=1}^{p}\beta_k x_{itk})}{\sum_{j}\prod_{t_j=1}^{m_i}\exp(\sum_{k=1}^{p}\beta_k x_{it_jk}^{(j)})} \quad (4)$$

有了条件似然函数，就可对其取对数变换，然后对其中的未知参数求偏导数，可获得联立方程组。进而采取合适的非线性迭代算法，可求得参数的估计值。

【说明】配对设计二值资料一水平多重 Logistic 回归模型的求解（包括参数估计、假设检验、拟合优度评价等）非常复杂，一般需要借助统计软件来实现，因篇幅所限，此处从略。

3 实例分析

3.1 问题与数据

【例 1】为探索分娩体重小于胎龄儿的原因，收集 226 例产妇的资料[8]。其中 47 例产妇分娩体重小于胎龄儿，179 例产妇分娩体重正常儿。对产妇的年龄（A）进行配对，分析初潮年龄（X_1）、身高（X_2）、孕期是否补铁（X_3）、孕期是否补钙（X_4）和孕期是否补锌（X_5）与体重小于胎龄儿的关系。令体重小于胎龄儿为 1（Y=1）、正常体重儿为 0（Y=0）；服用过相应营养素补充剂赋值为 1（X_3=1，X_4=1，X_5=1）、未服用过相应营养素补充剂赋值为 0（X_3=0，X_4=0，X_5=0）。见表 1。

表 1 分娩体重小于胎龄儿影响因素的 m∶n 配对病例对照研究资料

id	Y	A	X_1	X_2	X_3	X_4	X_5
1	1	20	14	162	1	1	1
2	1	20	14	158	0	0	0
3	1	20	14	165	0	0	0
…	…	…	…	…	…	…	…
224	0	30	12	168	0	0	0
225	0	30	12	152	0	0	0
226	0	30	13	150	0	0	0

注：因篇幅所限，详细数据见文献[8]

3.2 分析策略

3.2.1 仅依据题中的自变量并采取三种方法筛选自变量

3.2.1.1 变量说明

设 Y 为因变量(体重是否小于胎龄儿),“Y=1”代表体重小于胎龄儿、“Y=0”代表正常体重胎龄儿。再设 age=A,即把“age(年龄)”视为一个计量自变量,同时,A(年龄)又是一个“配对标记变量”或“分层变量”;X_1 和 X_2 是另外两个计量变量;X_3、X_4 和 X_5 都是“二值变量”。

3.2.1.2 SAS数据步程序

利用以下SAS数据步程序,可创建SAS数据集a:

```
data a;
infile'D:\sastjfx\data10_4.txt';
input id Y A X1-X5 @@;
age=A;
run;
```

【说明】先将题中的全部数据按226行8列输入计算机,并以文件名“data10_4.txt”存储在D盘文件夹名为“sastjfx”之内。8列变量名依次为:“id Y A X1-X5”,其中,“id”为“编号”,即受试对象的编号,id=1、2、3、…、226;其他变量的含义如题中所述。数据文件的格式为“文本文件”,其内的第1行就是第1位受试对象的数据,绝对不能输入“变量名”。“age=A;”是一个赋值语句,即产生一个新变量age,其各行的数值与变量A相同。

3.2.1.3 SAS过程步程序

需要调用LOGISTIC过程,采用逐步法、前进法和后退法筛选自变量的结果相同,故仅呈现后退法对应的程序如下:

```
proc logistic data=a descending;
model Y=age X1-X5/selection=backward sls=0.05;
strata A;
run;
quit;
```

【说明】“descending”选项是要求给出“Y=1”(即体重小于胎龄儿)发生概率的计算结果,否则,给出“Y=0”(即体重正常)发生概率的计算结果;“strata A;”语句表示以“年龄(A)”为“配对标记”或“分层变量”,表明将创建配对设计二值资料多重Logistic回归模型。

3.2.1.4 主要输出结果

主要输出结果见式(5):

$$P(Y=1)=\frac{e^{-0.1563X_2}}{1+e^{-0.1563X_2}} \quad (5)$$

反映模型对资料的拟合优度的一个指标“-2L=174.159”,df=1。“-2L”是“-2倍的对数似然函数的值”,自由度相同时,此值越小,表明所对应的模型对资料的拟合优度越高。

3.2.2 增加由计量自变量产生派生自变量并采取三种方法筛选自变量

3.2.2.1 增加由计量自变量产生派生自变量

```
data b;
    set a;
w1=age*age; w2=age*X1;
w3=age*X2; w4=X1*X1;
w5=X1*X2; w6=X2*X2;
w7=age*age*age; w8=X1*X1*X1;
w9=X2*X2*X2; w10=log10(age);
w11=log10(X1); w12=log10(X2);
w13=exp(age); w14=exp(X1);
w15=exp(X2); w16=w13/(1+w13);
w17=w14/(1+w14); w18=w15/(1+w15);
w19=1/age;w20=1/X1; w21=1/X2;
run;
```

【说明】在数据集a的基础上,增加21个派生自变量w1-w21,创建数据集b。其中,有些是平方变换(如w1、w4、w6),立方变换(如w7、w8、w9),交叉乘积变换(如w2、w3、w5),对数变换(如w10-w12),指数变换(如w13-w15),Logistic变换(如w16-w18),倒数变换(如w19-w21)。

3.2.2.2 SAS过程步程序

采用逐步法和前进法筛选自变量的结果相同,模型中仅保留w3,其“-2L=176.585”,df=1,略差于前面的拟合效果,程序从略。而采用后退法,结果有了一定的改善,其程序如下:

```
proc logistic data=b descending;
model Y=age X1-X5 w1-w21/selection=backward sls
=0.05;
strata A;
run;
quit;
```

3.2.2.3　主要输出结果

主要输出结果见式(6)：

$$P(Y=1)=\frac{e^{Z}}{1+e^{Z}} \quad (6)$$

在式(6)中，Z=−1.2264X_2+0.1377w_2−0.1186w_4+0.00334w_6

反映模型对资料的拟合优度的一个指标“−2L=167.015”，df=4。

若将前面过程步的“model”语句中“w1−w21”修改成“w1−w9”，其他保持不变，则主要输出结果见式(7)：

$$P(Y=1)=\frac{e^{Z}}{1+e^{Z}} \quad (7)$$

上式中，Z=−0.9729X_2+0.1083w_2−0.00438w_8+0.000011w_9。

反映模型对资料的拟合优度的一个指标“−2L=166.860”，df=4。

比较模型(6)与(7)可知：后者稍微优于前者。

3.3　拟合优度比较

3.3.1　依据“−2L”进行比较

模型(5)与(7)中所含参数数目不等，两者中何者为优？可采用以下方法进行假设检验：

含参数少的模型的“−2L”与含参数多的模型的“−2L”之差，服从自由度为df的χ^2分布。这里，$df=df_{多}-df_{少}$，本例中，df=4−1=3。

χ^2 = 174.159 − 166.860 = 7.299，df=3

查χ^2值表，得：$\chi^2_{0.05(3)}$ = 7.815，因7.299<7.815，说明P>0.05。就本例而言，含4个自变量的模型(7)并未明显优于仅含1个自变量的模型(5)。

3.3.2　依据kappa值比较

可以分别基于模型(5)和(7)计算出“预测概率$\hat{P}$”，它是将各受试对象的自变量的取值代入模型计算出来的，它代表“Y=1”发生的概率。于是，当“$\hat{P}$”大于某个“界值P_c”时，可判断该个体属于“Y=1”，否则，判断该个体属于“Y=0”。这样就可以获得一个预测的二值结果变量Y_{hat}，它的取值为0或1。接下去，就可以做出一个关于“Y”与“Y_{hat}”的交叉频数表。见表2。

由表2可以计算出两种情况下的“kappa值”(即反映模型与资料吻合程度或一致性高低的统计指标)，此值越大，说明模型与资料越吻合。

按常理，应取P_c=0.5，但此时kappa值并不够高。

表2　基于模型(5)和(7)计算得到的交叉频数表

Y	例数					
	模型(5)Y_{hat}：	0	1	模型(7)Y_{hat}：	0	1
0		154	25		160	19
1		24	23		24	23

对于模型(5)而言，P_c = 0.32时，kappa=0.3470为最大值，P<0.0001(说明总体kappa值不为0)；而对于模型(7)而言，经尝试，P_c = 0.35时，kappa=0.3989为最大值，P<0.0001(说明总体kappa值不为0)。

由此可知，模型(7)稍优于模型(5)(注意：此处不便对两个kappa值进行假设检验)。

基于SAS计算出表2中与模型(5)对应结果所需要的SAS程序如下：

```
proc logistic data=a descending;
  model Y=X2;strata A;
  output out=aaa1 p=prob;
run;quit;
data bbb1;
  set aaa1;
  if prob>0.32 then yhat=1;
  else if prob<=0.32 then yhat=0;
proc freq data=bbb1;
  tables y*yhat/agree;test kappa;
run;
```

基于SAS计算出表2中与模型(7)对应结果所需要的SAS程序如下：

```
proc logistic data=b descending;
model Y=X2 W2 W8 W9;strata A;
output out=aaa2 p=prob; run; quit;
data bbb2;
  set aaa2;
  if prob>0.35 then yhat=1;
  else if prob<=0.35 then yhat=0;
proc freq data=bbb2;
  tables y*yhat/agree; test kappa;
run;
```

4　讨论与小结

4.1　讨论

对配对设计二值资料构建一水平多重Logistic回归模型，不能计算ROC曲线下的面积AUC，故不能基于此比较不同模型对资料的拟合优度。基于“−2L”数值比较不同模型对资料的拟合优度时，不

够直观，灵敏度也不够高。由模型导出的“kappa值”来比较不同模型对资料的拟合优度，仅是一种尝试，且不便进行不同模型对应的kappa值的假设检验。

当资料中含有较多计量自变量时，产生派生自变量在一定程度上有利于构建出具有较高拟合优度的回归模型；当自变量中有“多值名义或有序变量”时，应产生“哑变量”，且应将来自每个原变量的“多个哑变量”视为一个“整体”（其所有“水平”的回归系数之和等于0），实现的方法是在LOGISTIC过程步中使用“class语句”（参见本期科研方法专题第二篇文章）；对自变量中的“二值变量”不必产生派生变量，因为那些派生自变量没有实用价值。

筛选自变量时，应尽可能采用多种策略，如“逐步法”“前进法”和“后退法”。当全部自变量数目较多时，往往有可能获得拟合效果不同的回归模型[1]。

4.2 小结

本文介绍了m∶n配对设计二值资料一水平多重Logistic回归分析的基本原理，并通过一个实例，介绍了使用SAS软件实现统计分析的全过程。获得了如下四点启示：其一，有必要确保所获得的科研资料是值得分析的；其二，有必要基于定量自变量产生派生变量；其三，有必要同时采用“逐步法”“前进法”和“后退法”筛选自变量；其四，有必要采用多种方法评价不同回归模型对资料的拟合优度。

由基本常识可知，能否获得非常理想的统计分析结果，仅在较小的程度上取决于所采用的统计分析方法和实现分析的策略。真正起决定性作用的是研究设计是否科学、严谨和完善；研究过程是否依据标准操作规程、是否有严格的质量控制措施；数据是否精准、可靠，是否具有重现性，它们才是决定所获得的数据是否值得分析，以及分析结果是否真实揭示了研究问题内在客观规律的重要基础。

参考文献

[1] 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.
[2] 谷恒明，胡良平. 基于贝叶斯统计思想实现多重线性回归分析[J]. 四川精神卫生，2018，31(1)：12-14.
[3] 谷恒明，胡良平. 基于机器学习统计思想实现多重线性回归分析[J]. 四川精神卫生，2018，31(1)：15-18.
[4] 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.
[5] 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.
[6] 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.
[7] 胡良平. 提高回归模型拟合优度的策略(IV)——优化计分变换与其他变量变换[J]. 四川精神卫生，2019，32(1)：21-28.
[8] 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，2012：451-462.
[9] 柳青. 中国医学统计百科全书-多元统计分册[M]. 北京：人民卫生出版社，2004：212-217.

（收稿日期：2019-08-01）
（本文编辑：吴俊林）