流行病学常用的统计方法

Ⅳ.Logistic回归模型及其在流行病学中的应用

辽宁省卫生防疫站 章扬熙

分析性流行病学研究的目的在于分析暴露因素（exposure factor）和疾病（disease）之间的关系。但是，客观上往往存在着混杂因素（confounding factor），若不对其加以控制，会使得暴露因素与疾病之间呈现出假阳性或假阴性的情况，从而导致错误的判断和结论。为了控制混杂因素，可以在设计阶段采取措施，也可以在分析阶段采取措施，或二者兼而有之。在设计阶段，可通过选择对象、配对等办法来消除或均衡混杂因素对疾病的影响，甚至可把多因素问题化成单因素问题，单用这种方法有时会产生把相互联系的多因素问题化成单因、孤立的问题之弊。故也可在分析阶段采用分层分析或多因素分析等方法，把暴露因素的效应与混杂因素的效应分离开来，还可同时考察因素间的交互影响。在流行病学中，常用的分层分析的方法为1959年Mantel-Haenszel的方法（以下简称M-H法），这种方法仅适用于混杂因素较少、样本例数较多的情况，而且当暴露因素是连续性变量时，只能用等级分层方法，从而损失不少信息。有人曾用线性回归模型来弥补M-H法的不足，但估计的发病率可能出现小于0或大于1的不合理的情况，二十世纪六十年代，在流行病学领域开始应用了Logistic回归模型，它弥补了M-H法和线性回归的不足，经过近年的发展，现已成为分析性流行病学研究中的重要方法。本文将对此模型作一简要介绍。

单因素的Logistic回归模型

一、Logistic回归模型的原理：设暴露因素的量为X，发病率为P，可列出其微分方程为：

$$\frac{dP}{dX}=\beta_1 P(1-P) \quad \cdots\cdots(1)$$

分离变量，得

$$\frac{dP}{P(1-P)}=\beta_1 dX$$

两边积分，得

$$\int\left(\frac{1}{P}+\frac{1}{1-P}\right)dP=\int\beta_1 dX$$

$$\ln P-\ln(1-p)=\beta_1 X+\beta_0$$

$$\ln\frac{P}{1-P}=\beta_1 X+\beta_0 \quad \cdots\cdots(2)$$

$$\frac{P}{1-P}=e^{\beta_1 X+\beta_0},\quad \frac{1}{P}=1+e^{-(\beta_0+\beta_1 X)}$$

$$P=\frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \quad \cdots\cdots(3)$$

这就是有名的Logistic回归模型。

如果把（3）式右侧分子、分母用$e^{\beta_0+\beta_1 X}$来乘，得：

$$P=\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \quad \cdots\cdots(4)$$

这是Logistic回归模型的另一个表达式。

设Q为未发病率，则Q=1−P，P用（4）式代入，得

$$Q=1-\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

$$=\frac{1}{1+e^{\beta_0+\beta_1 X}} \quad \cdots\cdots(5)$$

可见，公式（2）、（3）、（4）、（5）都是Logistic回归模型的表达形式。

二、Logistic回归模型的基本性质：设$\beta_0=0$，$\beta_1=1$我们得到最简单的Logistic模型为：

$$P=\frac{1}{1+e^{-X}} \quad \cdots\cdots(6)$$

从上式显然可以看出，当$X=-\infty$时，$P=0$，当$X=+\infty$时，$P=1$。它是一条单调上升的S型曲线，如图1。

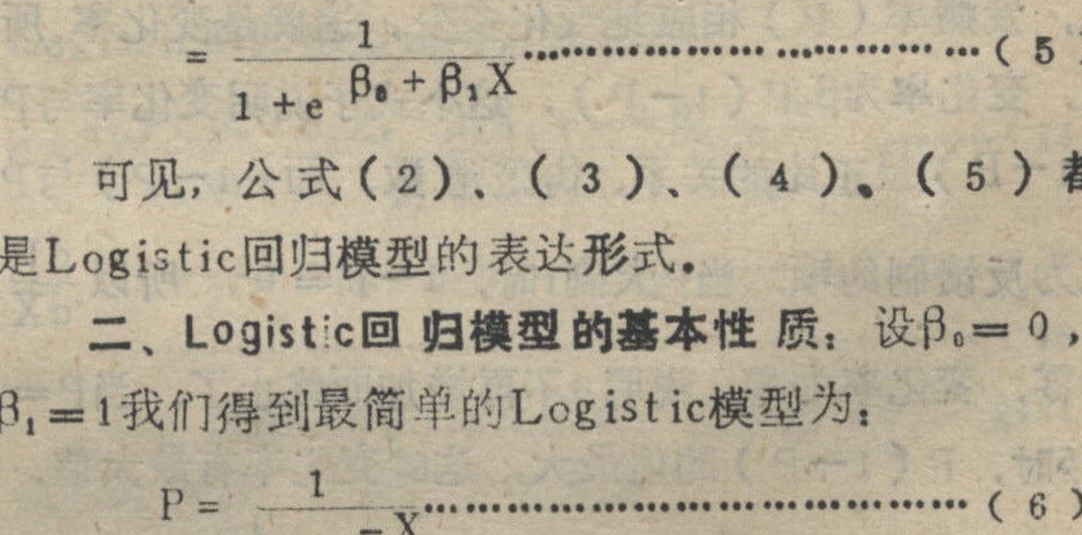

图1 $P=1/(1+e^{-X})$的图形

这条曲线有以下特征：

1.以（0，0.5）为中心的S型曲线，X越大，P越大。

2.当X→－∞时，P→1，所以X＝1是这条曲线的渐近线。

3.当X→∞时，P→0，所以X＝0是这条曲线的另一条渐近线。

我们知道，无论暴露因素的量（X）多小或多大，P不应小于0或大于1。显然，Logistic模型满足这个条件。

下面，我们再研究一下（3）式的情况，看看β_0与β_1对曲线形状的影响。β_1实际上是（2）式的回归系数（若把$\ln\frac{P}{1-P}$看成因变量，X为自变量），所以，当β_1越大，曲线越陡直，当β_1越小，曲线越平坦。β_0实际上是（2）式的截距，所以它影响曲线的左右移动，β_0越小，曲线越右移，β_0越大，曲线越左移。如果β_1为负值，画出的曲线为当X增加时P反而下降，这反映暴露因素是保护因素而不是危险因素。β_1为正值时，暴露因素为危险因素。

三、Logistic回归模型的变化率：如前所述，单因子Logistic回归模型是由（1）式微分方程推导出来的，$\frac{dP}{dX}$表示暴露因子的量（X）每有一个微小变化，发病率（P）相应地变化多少，这就是变化率。所以，变化率为$\beta_1P(1-P)$，这个式子说明变化率与P（1－P）呈正比的关系，β_1是常数，而（1－P）与P互为反馈制约项，当P大到1时，1－P＝0，所以$\frac{dP}{dX}$为零，变化率为零，说明P不再增加而终止了。当P＝0.5时，P（1－P）的值最大，这时变化率有最大值，这一点由图1可以显见，当P小到零时，$\frac{dP}{dX}$也为零，说明P不再减少而终止了。所以，P值越接近于0或1，P（1－P）越小，变化率越低。

四、比数（Odds）、logit（ln Odds）与比值比（Odds Ratio）：以P表示某事件发生的概率，1－P则为不发生某事件的概率，两者的比值$\frac{P}{1-P}$叫作比数，也叫优势，比数若大于1说明发生的可能性大于不发生的可能性，也就是发生占优势；反之，比值小于1说明不发生占优势。比数的自然对数值$\ln\frac{P}{1-P}$，叫作Logit，即Logit $P=\ln\frac{P}{1-P}$，所以（2）式可以写成

$$\text{Logit } P=\beta_1X+\beta_0 \quad \cdots\cdots(7)$$

也可以用Odds来表示

$$\text{Odd}_s=\frac{P}{1-P}=e^{\beta_1X+\beta_0} \quad \cdots\cdots(8)$$

〔例1〕某单位调查吸烟者205人，10年内有43人发生慢性气管炎，不吸烟者205人，10年内有20人发生慢性气管炎，求各组的比数（Odds）和Logit。

吸烟组的发病率$P_1=43/205=0.2098$

比数Odds为$\frac{P_1}{1-P_1}=\frac{0.2098}{1-0.2098}=0.26543$

Logit $P_1=\ln0.26543=-1.32640$

同理，不吸烟组的发病率$P_0=20/205=0.09756$

比数$\frac{P_0}{1-P_0}=\frac{0.09756}{1-0.09756}=0.10811$

Logit $P_0=\ln0.10811=-2.22461$

在流行病学的研究中，往往有两个组，比如暴露组与未暴露组，这两个组的比数Odds的比值，叫作比值比（Odds Ratio，简称为OR）

$$OR=\frac{P_1/(1-P_1)}{P_0/(1-P_0)} \quad \cdots\cdots(9)$$

〔例2〕求例1的比值比（OR）。

从例1的解中知道，$P_1/(1-P_1)=0.26543$，$P_0/(1-P_0)=0.10811$，代入（9）式，得

$$OR=\frac{0.26543}{0.10811}=2.46$$

从（9）式可以看出，当P_1与P_0较小时，$1-P_1\to1$、$1-P_0\to1$，比值比OR与相对危险度$RR=P_1/P_0$相接近。

五、比值比OR与β_1的关系：如果设暴露因子的量为X，以X_0作为对照的量，两者的发病率分别为P和P_0，则有

Logit P－Logit $P_0=\beta_0+\beta_1X-(\beta_0+\beta_1X_0)=\beta_1(X-X_0)$而Logit P－Logit $P_0=\ln\frac{P}{1-P}-\ln\frac{P_0}{1-P_0}=\ln\frac{P/(1-P)}{P_0/(1-P_0)}=\ln OR$。可见$OR=e^{\beta_1(X-X_0)}$ ……（10）

如果有暴露因素X＝1，无暴露因素$X_0=0$，则（10）式变成为

$$OR=e^{\beta_1} \text{ 或 Logit } P-\text{Logit } P_0=\beta_1 \quad \cdots\cdots(11)$$

如前所述。当P与P_0较小时，OR与RR相接近，所以可以把OR作为RR的近似值。从（11）式中可以看出，当得到某暴露因素的Logistic回归系数β_1值之后，就可以求出OR这个RR的近似值。又从（10）式可以看出，β_1表示暴露因素改变一个单位，OR改变e^{β_1}个

单位.

〔例3〕用$Y=\beta_0+\beta_1X$(式中$Y=\text{Logit } P=\ln\frac{P}{1-P}$)Logistic模型，来分析例1、例2中的$\beta_1$与OR关系.

把例1列成四格表如下

表 1　　吸烟与患慢性气管炎的关系

		吸烟因子X	
		1(吸烟)	0(不吸烟)
反应情况D	1(发病)	43(0.2098)	20(0.09756)
	0(未发病)	162(0.7902)	185(0.90244)
	合计	205(1)	205(1)

注：括号内数字为概率值

当X=1时，

$$Y_1=\ln\left(\frac{0.2098}{0.7902}\right)=\ln\left(\frac{43}{162}\right)=-1.326396$$

当X=0时，

$$Y_0=\ln\left(\frac{0.09756}{0.90244}\right)=\ln\left(\frac{20}{185}\right)=-2.224624=\beta_0$$

$$\beta_1=Y_1-Y_0=-1.326396-(-2.224624)=0.898228$$

将其作图如下

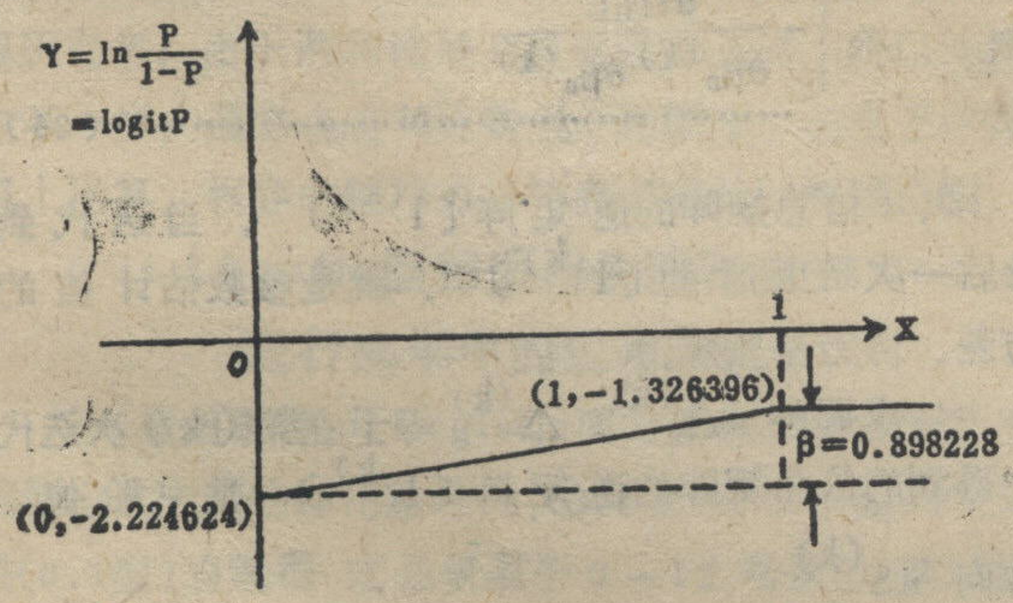

图 2　$Y=-2.224624+0.898228X$ 的图形

$$OR=e^{\beta_1}=e^{0.898228}=2.46$$

〔例4〕对日本广岛原子弹爆炸幸存者中白血病发病率进行了16年前瞻性观察，结果如表2，试求其Logistic回归方程式。

以LogitP为因变量，X为自变量，按最小二乘法求直线回归，得$\beta_1=0.00924175$，$\beta_0=-9.61578989$，Logistic回归方程式为$\text{LogitP}=-9.61578989+0.00924175X$，或

$$P=\frac{1}{1+e^{-(-9.61578989+0.00924175X)}}$$

关于参数β_1、β_0的估计，在Logistic回归模型中还有其他的方法，待以后述及。

表 2　　不同T_{DS}照射剂量的幸存者中白血病的发病率(/10万人·年)

T_{DS}总剂量范围(拉德)	组中值(X)	发病率(P)	$\text{Logit } P=\ln\frac{P}{1-P}$
<5	2.5	3.0	-10.4143
5～	12.5	5.1	- 9.8836
20～	35.0	20.9	- 8.4730
50～	75.0	18.3	- 8.6058
100～	150.0	41.5	- 7.7868
200～	250.0	55.6	- 7.4942
≧300	350.0	140.5	- 6.5663

多因素的Logistic回归模型概述

一、多因素的Logistic回归模型：当研究因素有p个，即X_1、$X_2\cdots X_p$，发病率为P，则与单因子相类似地有

$$P=\frac{1}{1+e^{-(\beta_0+\beta_1X+\beta_2X_2+\cdots+\beta_pX_p)}} \quad \cdots\cdots(12)$$

上式可以写成

$$P=\left\{1+\text{eXP}\left[-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p)\right]\right\}^{-1} \quad \cdots\cdots(13)$$

与单因素的Logistic回归模型相类似，多因素Logistic回归模型还可以表达为

$$P=\frac{\text{eXP}\left[\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p\right]}{1+\text{eXP}\left[\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p\right]} \quad \cdots\cdots(14)$$

$$Q=\left\{1+\text{eXP}\left[\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p\right]\right\}^{-1} \quad \cdots\cdots(15)$$

二、比数、Logit与比值比：在多因素的Logistic回归模型中，与单因素相类似

$$\text{Logit } P=\ln\frac{P}{1-P}=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p \quad \cdots\cdots(16)$$

$$OR=\text{eXP}\left[\beta_1(X_1-X_{1_0})+\beta_2(X_2-X_{2_0})+\cdots+\beta_p(X_p-X_{p_0})\right] \quad \cdots\cdots(17)$$

关于公式中β_0、β_1、$\beta_2\cdots\beta_p$各参数的估计，通常采用最大似然法。

三、最大似然法：其主要思想是这样的，比如，一个袋子中有100个球，其中红球可能为10个(p=10%)，也可能为90个(p=90%)，其余的均为白球，我们从袋子中摸出一个球，结果为红色，则可认为红球在袋子中有90个(最大似然是如此)，因为这个估计最有利于我们实际观测到的结果的出现。用最大似

然法来估计Logistic模型中的各参数β_k（k＝0，1，…，p），具体步骤为：

1.对资料构造出一个似然函数L，为了简化计算，再取其自然对数值lnL；

2.对这个对数似然函数lnL（即目标函数）求其各参数β_k的一阶偏导数：

$$\frac{\partial \ln L}{\partial \beta_k} \quad (k=0,\ 1,\ \cdots,\ p) \cdots\cdots\cdots\cdots (18)$$

3.通常是应用Newton－Raphson迭代法，解下列非线性方程组，求出满足下列方程组的各$\hat{\beta}_k$，即各参数β_k的最大似然估计值：

$$\left[\frac{\partial \ln L}{\partial \beta_k}\right]_{\beta=\hat{\beta}} = 0 \cdots\cdots\cdots\cdots (19)$$

4.各参数估计值$\hat{\beta}_k$的方差、协方差矩阵为迭代终止时得到的信息矩阵的逆矩阵，即目标函数二阶偏导数的负值组成的矩阵的逆矩阵：

$$[I]^{-1} = \left[-\frac{\partial^2 \ln L}{\partial \beta_j \partial \beta_k}\right]^{-1} \quad (j,\ k=0,\ 1,\ \cdots,\ p)$$

$$\cdots\cdots\cdots\cdots (20)$$

应用这些信息，可对$\hat{\beta}_k$进行假设检验和参数的区间估计。

5.通过对比两个包含不同参数个数的lnL值，可以对模型中的参数贡献做假设检验。所用统计量是似然比统计量G，用下式计算：

$$G_f = 2\ln\left[\frac{L_{t+f}}{L_f}\right] = 2(\ln L_{t+f} - \ln L_t)$$

$$\cdots\cdots\cdots\cdots (21)$$

上式中t是原包含在模型中的参数的个数，f为所加入的参数的个数，统计量G_f服从自由度为f的χ^2分布。通过χ^2检验来考察f个参数贡献有没有显著意义。

四、Newton-Raphson迭代法：其计算公式为

$$\beta^{(k+1)} = \beta^{(k)} + [I^{(k)}]^{-1} \cdot S^{(k)}$$

$$\cdots\cdots\cdots\cdots (22)$$

上式中右上标（k）及（k+1）表示迭代次数，$\beta^{(k)}$表示在第k次迭代时的试验值，它是（P+1）$\times_1$维参数向量

$$\beta^{(k)} = (\beta_0^{(k)},\ \beta_1^{(k)},\ \cdots,\ \beta_p^{(k)})'$$

$$\cdots\cdots\cdots\cdots (23)$$

上式中“′”表示转置。

在计算中所用的目标函数为lnL，具体的迭代步骤是：对各参数给定初始值$\beta_k^{(o)}$，通常赋值为零，规定控制误差ε的大小，然后开始迭代。各次迭代的过程是一样的，为了说明一般情况，假设迭代已进行到了第（k）次。

1.对目标函数lnL求各参数试验值$\beta_k^{(k)}$的一阶偏导数，并由其组成向量$S^{(k)}$为

$$S^{(k)} = \left(\frac{\partial \ln L}{\partial \beta_0^{(k)}},\ \frac{\partial \ln L}{\partial \beta_1^{(k)}},\ \cdots,\ \frac{\partial \ln L}{\partial \beta_p^{(k)}}\right)'$$

$$\cdots\cdots\cdots\cdots (24)$$

2.对目标函数lnL求各参数试验值$\beta_k^{(k)}$的二阶偏导数

$$\frac{\partial^2 \ln L}{\partial \beta_j^{(k)} \partial \beta_k^{(k)}} \quad (j,\ k=0,\ 1\cdots,\ p) \cdots\cdots (25)$$

由各二阶偏导数的负值组成信息矩阵为

$$[I^{(k)}] = \begin{bmatrix} -\frac{\partial^2 \ln L}{\partial \beta_0^{(k)} \partial \beta_0^{(n)}}, & -\frac{\partial^2 \ln L}{\partial \beta_0^{(k)} \partial \beta_1^{(k)}}, & \cdots, & \frac{-\partial^2 \ln L}{\partial \beta_0^{(k)} \partial \beta_p^{(k)}} \\ -\frac{\partial^2 \ln L}{\partial \beta_1^{(k)} \beta_0^{(k)}}, & -\frac{\partial^2 \ln L}{\partial \beta_1^{(k)} \partial \beta_1^{(k)}}, & \cdots, & \frac{-\partial^2 \ln L}{\partial \beta_1^{(k)} \partial \beta_p^{(k)}} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ -\frac{\partial^2 \ln L}{\partial \beta_p^{(k)} \partial \beta_0^{(n)}}, & -\frac{\partial^2 \ln L}{\partial \beta_p^{(k)} \partial \beta_1^{(k)}}, & \cdots, & -\frac{\partial^2 \ln L}{\partial \beta_p^{(k)} \partial \beta_p^{(k)}} \end{bmatrix}$$

$$\cdots\cdots\cdots\cdots (26)$$

3.求信息矩阵的逆矩阵$[I^{(k)}]^{-1}$。当迭代到最后一次结束时的$[I^{(k)}]^{-1}$，就是参数估计值的方差、协方差矩阵。

4.求调整量$\Delta^{(k)}$。$\Delta^{(k)}$等于在第（k）次迭代中得到的信息矩阵的逆矩阵$[I^{(k)}]^{-1}$乘一阶偏导数向量$S^{(k)}$

$$\Delta^{(k)} = [I^{(k)}]^{-1} \cdot S^{(k)} \cdots\cdots\cdots\cdots (27)$$

若$|\Delta^{(k)}| \leqslant \varepsilon$，则迭代结束，就把$\beta^{(k)}$作为参数$\beta$的最优估计值$\beta$，否则继续迭代。

5.下一次迭代时的各参数试验值等于本次试验值$\beta^{(k)}$加上调整量$\Delta^{(k)}$，即

$$\beta^{k+1} = \beta^{(k)} + \Delta^{(k)} \cdots\cdots\cdots\cdots (28)$$

用向量表示为

$$\begin{pmatrix} \beta_0^{(k+1)} \\ \beta_1^{(k+1)} \\ \vdots \\ \beta_p^{(k+1)} \end{pmatrix} = \begin{pmatrix} \beta_0^{(k)} \\ \beta_1^{(k)} \\ \vdots \\ \beta_p^{(k)} \end{pmatrix} + \begin{pmatrix} \Delta_0^{(k)} \\ \Delta_1^{(k)} \\ \vdots \\ \Delta_p^{(k)} \end{pmatrix} \cdots\cdots (29)$$

然后进行第（k+1）次的迭代，过程与第k次迭代相同。

五、应用Logistic回归模型进行多因素分析：其目的在于把暴露因素效应有显著意义的变量纳入模型之中，通常采用阶梯式配合技术，即由最简单的模型开始配合起，一直配合到把所有贡献有显著意义的因素变量全部纳入模型之中。最后把有显著意义的交互作用项也纳入模型之中。各模型对成组的资料拟合的好坏，可用下式进行拟合优度检验：

$$\chi^2=\sum_{g=1}^{m}\frac{(r_g-n_g\hat{p}_g)^2}{n_g\hat{p}_g(1-\hat{p}_g)}\cdots\cdots\cdots\cdots\cdots\cdots(30)$$

式中g为分组号，共m组数据，r_g为g组的病例数，n_g为g组的总人数，$\hat{P}_g$为g组的估计发病率。另外，为了考察各β_k的贡献，可进行G检验（用公式21）。通过对各模型的比较和检验，确定最后所采用的模型。然后再列出该模型各暴露因素的参数估计值$\hat{\beta}_k$、估计方差$\hat{V}(\hat{\beta}_k)$、估计标准误$\sqrt{\hat{V}(\hat{\beta}_k)}$、标准化参数估计值$\hat{\beta}_k/\sqrt{\hat{V}(\hat{\beta}_k)}$及比值比eXP$(\hat{\beta}_k)$。

六、两种Logistic回归模型：即非条件Logistic回归模型与条件Logistic回归模型，前者可用于成组的病例对照研究和定群研究，后者用于配比的病例对照研究等。由于两种模型的似然函数不同，所以计算也不一样，通常各有其计算机专用程序，在电子计算机上运算。程序的使用方法可参考有关的使用说明。

Logistic回归适用条件和在流行病学中的应用范围

一、适用条件：从Logistic回归模型的基本性质不难看出其应用的适用条件为：①因变量P必须是两项分类(0,1型)的数据，或必须限于0～1之间数值的数据。②自变量X与因变量P之间的关系呈或基本上呈S形曲线关系，或者说自变量X与LogitP之间呈线性关系。

二、流行病学中应用的范围：分析性流行病学调查研究方法有多种，比如队列研究、病例对照研究（包括成组和配比）等，只要所得资料符合上述条件，均可用Logistic回归模型来进行分析。应当指出，有些流行病学资料并不满足上述的适用条件，比如年龄因子与死亡率的关系呈✓型曲线而非S型，则不宜用Logistic回归模型来进行分析。

非条件Logistic回归模型

一、定群研究与病例对照研究：这两项成组资料都可以用非条件Logistic回归模型公式（16）进行研究。不容置疑，公式（16）是适用于定群研究的。定群研究是从因到果的研究，病例对照研究则相反，它是在发病总体中按比例π_1抽取病例，在不发病总体中按比例π_2抽取对照。对于一组因素X_k（k=1，2，…，p）取值，由于其发病概率为P，不发病概率为1−P，所以对于该组因素取值时，根据概率乘法定理，是病人被抽到的概率为$P\pi_1$；是非病人（对照）被抽到的概率为$(1-P)\pi_2$。这样，是病人而且具有该组因素取值的概率为

$$P'=\frac{P\pi_1}{P\pi_1+(1-P)\pi_2}$$

非病人而且具有该组因素取值的概率为

$$1-P'=\frac{(1-P)\pi_2}{P\pi_1+(1-P)\pi_2}$$

故 $$\frac{P'}{1-P'}=\frac{P\pi_1}{(1-P)\pi_2}$$

$$\text{Logit}\,P'=\ln\frac{P'}{1-P'}=\ln\frac{\pi_1}{\pi_2}+\ln\frac{P}{1-P}=\beta_0'+\sum_{k=1}^{m}\beta_kX_k$$

由于 $\ln\frac{P}{1-P}=\beta_0+\sum_{k=1}^{m}\beta_kX_k$，故

$$\beta_0'=\beta_0+\ln\frac{\pi_1}{\pi_2}\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(31)$$

这就证明了：定群研究与病例对照研究所建立的Logistic回归方程，除常数项相差一个$\ln\frac{\pi_1}{\pi_2}$外，其他各偏回归系数相同。

二、举例：基于上述特点，对于定群研究与病例对照研究的成组资料进行Logistic回归模型配合是类似的。下面以病例对照研究的成组资料来说明其具体步骤，读者可触类旁通。

〔**例5**〕应用病例对照研究来探索饮酒与食管癌发病的关系。为了排除年龄因素的干扰，把年龄划分为6个组（X_1～X_6），每组又按饮酒水平分两组（每次饮酒量少于80毫升者记为非接触$X_7=0$，等于或多于80毫升者记为接触$X_7=1$），结果如表3，试用Logistic回归模型进行配合。

对Logistic模型中的参数估计采用最大似然法，首先对资料根据二项分布原理构造出似然函数为

$$L=\prod_{g=1}^{m}\binom{r_g}{n_g}P_g^{r_g}Q_g^{n_g-r_g}\quad(\text{m为层数})\cdots\cdots(32)$$

相应的对数似然函数为

$$\ln L=\sum_{g=1}^{m}\ln\binom{r_g}{n_g}+\sum_{g=1}^{m}r_g\ln P_g+\sum_{g=1}^{m}(n_g-r_g)\ln Q_g\cdots\cdots(33)$$

表 3　　饮酒与食管癌发病关系的病例对照研究资料

分层	年龄组	年龄变量						饮酒变量	病例数	合计数
(g)	(i)	X_1	X_2	X_3	X_4	X_5	X_6	X_7	r_g	n_g
1	1	1	0	0	0	0	0	1	1	10
2	1	1	0	0	0	0	0	0	0	106
3	2	0	1	0	0	0	0	1	4	30
4	2	0	1	0	0	0	0	0	5	169
5	3	0	0	1	0	0	0	1	25	54
6	3	0	0	1	0	0	0	0	21	159
7	4	0	0	0	1	0	0	1	42	69
8	4	0	0	0	1	0	0	0	34	173
9	5	0	0	0	0	1	0	1	19	37
10	5	0	0	0	0	1	0	0	36	124
11	6	0	0	0	0	0	1	1	5	5
12	6	0	0	0	0	0	1	0	8	39

注：$X_1 \sim X_6$为年龄变量；25～34岁 $X_1=1$，其余为0；35～44岁 $X_2=1$；45～54岁 $X_3=1$；55～64岁 $X_4=1$；65～74岁 $X_5=1$；75岁以上 $X_6=1$；余为0；

对照数 = 合计数 − 病例数

式中

$$P_g=\frac{e^{XP\left(\sum_{k=1}^{P}\beta_k X_k\right)}}{1+e^{XP\left(\sum_{k=1}^{P}\beta_k X_k\right)}}\quad(P为自变量数)\cdots\cdots(31)$$

$$Q_g=\frac{1}{1+e^{XP\left(\sum_{k=1}^{P}\beta_k X_k\right)}}\cdots\cdots(35)$$

根据Newton-Raphson迭代法的要求，需对lnL求出各βi的一阶及二阶偏导数，可是βi是包含在P_g及Q_g中，为了计算的方便，可先对P_g、Q_g、$\ln P_g$、$\ln Q_g$求出各βk的一阶偏导数为

$$\frac{\partial P_g}{\partial \beta_k}=X_{gk}P_gQ_g\qquad\frac{\partial Q_g}{\partial \beta_k}=-X_{gk}P_gQ_g$$

$$\frac{\partial \ln P_g}{\partial \beta_k}=X_{gk}Q_g\qquad\frac{\partial Q_g}{\partial \beta_k}=-X_{gk}P_g$$

因此有

$$\frac{\partial \ln L}{\partial \beta_k}=\sum_{g=1}^{m}r_gX_{gk}Q_g-\sum_{g=1}^{m}(n_g-r_g)X_{gk}P_g$$

$$=\sum_{g=1}^{m}(r_g-n_gP_g)X_{gk}\cdots\cdots(36)$$

$$\frac{\partial^2 \ln L}{\partial \beta_k\,\beta_j}=-\sum_{g=1}^{m}r_gX_{gi}X_{gk}P_gQ_g-\sum_{g=1}^{m}(n_g-r_g)$$

$$X_{gi}X_{gk}P_gQ_g=-\sum_{g=1}^{m}n_gX_{gi}X_{gk}P_gQ_g\cdots\cdots(37)$$

采用Newton-Raphson迭代法求出模型中各参数估计值。再利用公式(17)求比值比

$$OR=e^{XP\left[\sum_{k=1}^{P}\beta_k(X_k-X_{ko})\right]}=\prod_{k=1}^{P}e^{XP\left[\beta_k(X_k-X_{ko})\right]}$$

从上式可以看出，多因素的联合比值比为各个单因素比值比的乘积，这个相乘作用也是Logistic模型的一个适用条件，使用此模型时应予注意。

本例可采用以下三种从简单到复杂的Logistic回归模型来进行配合：

①$LogitP=\beta_1X_1+\beta_2X_2+\cdots+\beta_6X_6$

②$LogitP=\beta_1X_1+\beta_2X_2+\cdots+\beta_6X_6+\beta_7X_7$

③$LogitP=\beta_1X_1+\beta_2X_2+\cdots+\beta_6X_6+\beta_7X_7+\beta_8X_8$

模型①为发病仅与年龄有关，与饮酒无关，模型②为发病与年龄及饮酒都有联系，模型③为在模型②的基础上，增加了X_8（年龄与饮酒的交互影响项，$X_8=(i-\bar{i})X_7$，$\bar{i}=(1+2+3+4+5+6)/6=3.5$，当$i=1$，$X_7=1$时，$X_8=(1-3.5)\times1=-2.5$，余类推。结果各层的$X_8$依次为−2.5，0，−1.5，0，−0.5，0，0.5，0，1.5，0，2.5，0。

首先，配合模型①，由于模型①不考虑饮酒的作用，所以把每两饮酒的水平合并，这样12个层就合并成6层，如表4的格式。

根据最大似然估计法，用Newton-Raphson法迭代求出各参数的估计值$\hat{\beta}k$（见表5的模型①部分）。lnL为对数似然函数值，得−434.22。G为似然比的统计量，用下式求得

$$G=2\sum_{g=1}^{m}\left[r_g\ln(r_g/(n_g\hat{P})+(n_g-r_g)\ln\{(n_g-r_g)/\right.$$

$\{n_g(1-\hat{P})\}$ ……(38)

表 4　配合模型①的数据格式

分层 (g)	年龄组 (i)	年龄变量 X_1	X_2	X_3	X_4	X_5	X_6	病例数 r_g	合计数 n_g
1	1	1	0	0	0	0	0	1	116
2	2	0	1	0	0	0	0	9	199
3	3	0	0	1	0	0	0	46	213
4	4	0	0	0	1	0	0	76	242
5	5	0	0	0	0	1	0	55	161
6	6	0	0	0	0	0	1	13	44

表 5　例5配合Logistic模型的结果

	模型①	模型②	模型③
变量数	6	7	8
lnL	-434.222	-394.461	-394.245
G	90.560	11.040	10.610
χ^2	101.771	9.319	8.467
df	6	5	4
β_1	-4.745	-5.054	-5.182
β_2	-3.050	-3.512	-3.616
β_3	-1.289	-1.856	-1.901
β_4	-0.781	-1.341	-1.334
β_5	-0.656	-1.088	-1.049
β_6	-0.869	-1.092	-1.055
β_7		1.670	1.714
$SE(\beta_7)$		0.190	0.201
β_8			-0.125
$SE(\beta_8)$			0.189

本例m=12，r_g及n_g皆为表 3 中的数值，$\hat{P}$用表 4 中的数值求模型①的G值，比如表 3 中的1～2层的$\hat{P}$用表 4 的1层求$\hat{P}=\frac{1}{116}=0.0086$，表 3 中的 3～4 层的$\hat{P}$用表 4 的2层求$\hat{P}=\frac{9}{199}=0.0454$，余类推，应当指出，用公式(38)与用公式（21）求G的结果一样。χ^2值的计算，应用公式（30）。模型①的χ^2值中$\hat{P}$由表 4 求得，r_g及n_g为表 3 中的值，这与求G的情况相同。模型①的$\chi^2=101.77$，自由度＝组数－变量数，本例自由度df＝12－6＝6，$P<0.005$，资料配合模型①不理想。于是，用表 3 的数据配合模型②，$\chi^2=9.319$，自由度df＝12－7＝5，$P>0.05$，配合较好。G＝11.04，两个模型G的差值近似χ^2分布，其自由度为两个模型变量数的差。把模型①与模型②比较，$G_1-G_2=90.56-11.04=79.52$，自由度df＝7－6＝1，$P<0.005$，说明资料用模型②配合与用模型①配合有了极大的改善。再利用各层$X_1\sim X_8$的信息用模型③配合，得$G_3=10.61$，$G_2-G_3=11.04-10.61=0.43$自由度df＝8－7＝1，$P>0.5$，未见显著改善，所以不能认为有年龄与饮酒交互作用的存在，在模型中无理由加入X_8，结果在 3 个模型中选用模型②。详细的模型②结果见表 6 。

从年龄变量$X_1\sim X_6$可以看出，年龄越大，$\hat{\beta}k$越大（负值越小），发病的概率越高，模型中应纳入此变量。饮酒的变量X_7的OR＝5.3111，说明饮酒是患食管癌的危险因素，总体OR的95%可信限为

$$e^{\hat{\beta}k\pm1.96\hat{\delta}(\hat{\beta}k)}=e^{1.6698\pm1.96\times0.1895}$$
$$=3.6634\sim7.7000$$

表 6 中STD$\hat{\beta}k$系用下式求得

$$STD\hat{\beta}k=\hat{\beta}k/\hat{\delta}(\hat{\beta}k) \quad \text{……(39)}$$

在大样本中，标准化$\hat{\beta}k$可视为U值，对$\hat{\beta}k$进行假设检验，表6各STD$\hat{\beta}k$的绝对值都大于 3，$P<0.01$，说明各$\hat{\beta}k$均有作用，有理由留在模型之中，另外，$\hat{\beta}k$的方差系从信息矩阵的逆矩阵中得来，$\hat{\beta}k$的标准误为其平方根。

表 6　例5配合模型②所得的各参数估计值

变量 X_k	偏回归系数 $\hat{\beta}_k$	$\hat{\beta}_k$ 的方差 $\hat{V}(\hat{\beta}_k)$	$\hat{\beta}_k$ 的标准误 $\hat{\sigma}(\hat{\beta}_k)$	标准化的$\hat{\beta}_k$ $STD\hat{\beta}_k$	比值比OR $\exp\hat{\beta}_k$
X_1	-5.0543	1.0180	1.0090	-5.0094	0.0064
X_2	-3.5120	0.1269	0.3562	-9.8588	0.0298
X_3	-1.8555	0.0375	0.1936	-9.5818	0.1564
X_4	-1.3408	0.0270	0.1643	-8.1598	0.2616
X_5	-1.0874	0.0339	0.1841	-5.9059	0.3391
X_6	-1.0921	0.1184	0.3441	-3.1739	0.3355
X_7	1.6698	0.0359	0.1895	8.8129	5.3111

条件Logistic回归模型

一、1：r配比的病例对照研究，设有几个配比组

（i=1，…，n），每组包括一个病例和r个对照（j=0，1，…，r_i），所研究的暴露因素共P个（k=1，…，P）。在第i个配比组内，有一个病例和r个对照；与这个配比组内第j个个体相联系的有一个危险因素向量$X_{ij}=(X_{ij1},\cdots,X_{ijP})'$，利用条件概率表示法，记病例接触暴露因素的概率为$P(X_{io}|D=1)$，X下标i表示第i组。O表示病例，D=1表示在发病的条件下；对照接触暴露因素的概率为$P(x_{ij}|D=0)$，X下标j表示为第j个对照（$j=1,\cdots,r_i$）。我们构造的条件概率为观察到的第一个暴露因素向量属于病例而其他暴露因素向量属于对照的概率与任何一个向量属病例、其余向量属对照的概率和的比值，即

$$\frac{P(X_{io}|D=1)\cdot\prod_{j=1}^{r_i}P(X_{ij}|D=0)}{\sum_{j=0}^{r_i}\left[P(X_{ij}|D=1)\cdot\prod_{\substack{j'\neq j\\ j'=0}}^{r_i}P(X_{ij'},|D=0)\right]}\quad\cdots\cdots(40)$$

再利用以下条件概率公式

$$P(X_{ij}|D=1)=[P(D=1|X_{ij})\cdot P(X_{ij})]/P(D=1)$$

$$P(X_{ij}|D=0)=[P(D=0|X_{ij})\cdot P(X_{ij})]/P(D=0)$$

公式（40）可写成

$$\frac{P(D=1|X_{io})\cdot\prod_{j=1}^{r_i}P(D=0|X_{ij})}{\sum_{j=0}^{r_i}\left[P(D=1|X_{ij})\cdot\prod_{\substack{j'\neq j\\ j'=0}}^{r_i}P(D=0|X_{ij'})\right]}\quad\cdots\cdots(41)$$

公式（40）与公式（41）等价，前者为回顾性研究，后者为前瞻性研究；从Logistic函数的定义有

$$P(D=1|X_{ij})=\frac{\exp(\alpha+\sum_{k=1}^{P}\beta_kX_{ijk})}{1+\exp(\alpha+\sum_{k=1}^{P}\beta_kX_{ijk})}$$

$$P(D=0|X_{ij})=\frac{1}{1+\exp(\alpha+\sum_{k=1}^{P}\beta_kX_{ijk})}$$

将上式代入（41）式并简化得（42）式

$$\frac{\exp(\sum_{k=1}^{P}\beta_kX_{iok})}{\sum_{j=0}^{r_i}\exp(\sum_{k=1}^{P}\beta_kX_{ijk})}=\frac{1}{1+\sum_{j=1}^{r_i}\exp[\sum_{k=1}^{P}\beta_k(X_{ijk}-X_{iok})]}\quad\cdots\cdots(42)$$

上式仅是对第i组而言，整个资料共有n个组，根据概率乘法定理，每一组内都是第一个暴露因素向量与病例相联系，其他向量与对照联系的概率为各配比率概率的连乘积，即条件似然函数为公式（43），亦即条件Logistic回归模型的计算公式。

$$L^*=\prod_{i=1}^{n}\{1/[1+\sum_{j=1}^{r_i}\exp(\sum_{k=1}^{P}\beta_kd_{ijk})]\}\quad\cdots\cdots(43)$$

式中$d_{ijk}=(X_{ijk}-X_{iok})$，为对照与病例对同一暴露因素接触水平的差数。

对数的条件似然函数为公式（44）

$$\ln L^*=-\sum_{i=1}^{n}\ln[1+\sum_{j=1}^{r_i}\exp(\sum_{k=1}^{P}\beta_kd_{ijk})]\quad\cdots\cdots(44)$$

用最大似然估计法及Newton-Raphson迭代来解出各$\hat{\beta}_k$值及其他参数的估计值。对于暴露因素的假设检验可用似然比统计量G，G值用公式(21)求得。

二、举例：以下用实例来说明条件Logistic回归模型配合配比的病例对照研究资料。

〔**例6**〕应用1：2配比的病例对照研究来研究肥胖（X_1）和口服避孕药雌激素（X_2）与子宫内膜癌发病的关系，共调查了20组，试配合条件Logistic回归模型（表7）。

表7　1:2配比病例对照研究的数据资料

配比组号 i	病例 0		对照 1		对照 2	
	X_{i01}	X_{i02}	X_{i11}	X_{i12}	X_{i21}	X_{i22}
1	1	1	0	0	0	0
2	0	1	0	1	1	0
3	1	1	1	1	0	1
4	0	1	0	1	0	1
5	1	1	0	1	1	1
6	1	1	0	0	1	1
7	0	1	0	0	0	1
8	1	1	0	0	1	0
9	0	0	1	0	0	1
10	1	1	0	0	0	1
11	1	1	0	0	1	0
12	0	1	0	1	0	1
13	1	1	0	1	1	1
14	0	1	0	0	1	0
15	1	1	1	0	1	1
16	1	1	1	0	0	1
17	1	0	1	1	1	1
18	1	0	0	1	0	1
19	0	1	0	1	0	0
20	1	1	0	1	0	0

模型可考虑单因素X_1、单因素X_2，双因素X_1+X_2，双因素及其交互影响（X_3），即$X_1+X_2+X_3$四种模型。

应用最大似然法进行计算，各模型的计算结果如表8。

表 8　例6资料配合不同模型的计算结果

	模型①	模型②	模型③	模型④
	X_1	X_2	X_1+X_2	$X_1+X_2+X_3$
lnL*	-19.263	-19.777	-16.653	-16.301
G	5.419	4.390	5.218**	0.704
β_1	1.4363		1.8239	0.8977
β_2		1.2833	1.5896	0.7398
β_3				1.4944

** 为比较只含X_1模型与含X_1及X_2模型的G值

计算得无有因素X的模型的$\ln L_0^*=-21.972$，模型①与无因素模型比较的统计量G为$G_1=2\times[(-19.263)-(-21.972)]=5.419$，自由度$df=1$，$P<0.05$，说明$X_1$（肥胖）对子宫内膜癌的发生有作用。同理，模型②的$G_2=4.390$，$df=1$，$P<0.05$，说明$X_2$（雌激素）对子宫内膜癌的发生也有作用。再把双因素模型③与单因素模型①比较，得$G_3=2[(-16.6534)-(-19.2626)]=5.218$，$df=1$，$P<0.05$，说明双因素比单因素（$X_1$或$X_2$）都优越（因为$\ln L_2^*$大于$\ln L_3^*$，若模型③与模型②比较所得的G值比$G_3$还大，$P<0.05$是必然的）。再考察模型④，$G=2[(-16.8013)-(-16.6534)]=0.7042$，$df=1$，$P>0.05$，说明$X_1$与$X_2$的交互影响$X_3$的作用不明显，不应纳入方程，最后选定用模型③来配合该资料，有关模型③的详细结果见表9。

这个结果说明肥胖与雌激素均为发生子宫内膜癌的危险因素，比值比分别为6.1961及4.9019，惟因本例的例数较少，变异较大，OR的95%可信限范围较宽。

表 9　例6配合模型③所得的各参数估计值

变量 X_k	偏回归系数 $\hat{\beta}k$	β_k 的标准误 $(\hat{\sigma}\beta k)$	标准化的 $\hat{\beta}_k$ $STD\hat{\beta}_k$	OR $\exp\hat{\beta}_k$	OR的95%可信限
X_1	1.8239	0.8495	2.1471	6.1961	1.17～32.75
X_2	1.5896	0.8090	1.9650	4.9019	1.00～23.93

四川省D型肝炎病毒感染的初步检测

四川省卫生防疫站　王道钦　刘丽华　王宏禧　殷大常　屠云人

血清流行病学研究表明，D型肝炎病毒（HDV）感染在世界上分布广泛，有散发和地方性感染两种形式。目前尚未弄清楚，为何在有高度HBV感染的东南亚地区和中国，HDV感染反而很低。四川一般人群的HBV携带率7.9%（RPHA法），而HDV感染的情况尚不清楚。本文简要报道四川地区HDV的初步检测结果。

1983年1月至1985年6月，在四川省内有代表性的地区及人群采集到HBsAg阳性患者血清515份，加0.1%叠氮钠（NaN_3）防腐，干冰条件下送加拿大IDRC。初测系作者在加拿大Alberta省公共卫生实验室用RIA法（Abbott药盒）进行。复测系由美国Georgetown大学Gerin实验室作进一步鉴定。

515份HBsAg阳性血清中有34份用（Abbott药盒）RIA法检测属可疑，经Gerin实验室证实其中4份为Anti-HD阳性，阳性率为0.8%（4/515）。其中急性B型肝炎的Anti-HD阳性率为2.0%（2/99），慢性B型肝炎0.9%（1/104），HBsAg携带者0.3%（1/300）；12份肝癌患者血清中，未检出Anti-HD。从地区分布看，成都市Anti-HD阳性率1.37%（3/219），永川4.55%（1/22），其余地区尚未发现。在71份藏族和37份彝族HBsAg携带者血清中均未查到Anti-HD。

本次初步调查结果证实我省确实有D型肝炎病毒感染存在。

（本项研究得到加拿大IDRC的资助及G.Y.Minuk、C.M.Anand、T.C.Stowe和K.A.Buchan的指导，谨致谢意）