Logistic回归结果的回归系数和OR值解读

作者：王江源 来源：爱科学

本文原地址：https://www.iikx.com/news/statistics/6062.html

Logistic回归结果的回归系数和OR值解读。Logistic回归虽然名字叫”回归”，但却是一种分类学习方法。使用场景大概有两个：第一用来预测，第二寻找因变量的影响因素。

一 从线性回归到Logistic回归

线性回归和Logistic回归都是广义线性模型的特例。

假设有一个因变量y和一组自变量$x_1, x_2, x_3, \dots, x_n$，其中y为连续变量，我们可以拟合一个线性方程：

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \dots + \beta_n * x_n$$

并通过最小二乘法估计各个β系数的值。

如果y为二分类变量，只能取值0或1，那么线性回归方程就会遇到困难: 方程右侧是一个连续的值，取值为负无穷到正无穷，而左侧只能取值[0,1]，无法对应。为了继续使用线性回归的思想，统计学家想到了一个变换方法，就是将方程右边的取值变换为[0,1]。最后选中了Logistic函数：

$$y = 1 / (1+e^{-x})$$

这是一个S型函数，值域为(0,1)，能将任何数值映射到(0,1)，且具有无限阶可导等优良数学性质。

我们将线性回归方程改写为：

$$y = 1 / (1+e^{-z}),$$

其中，$z = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \dots + \beta_n * x_n$

此时方程两边的取值都在0和1之间。

进一步数学变换，可以写为：

$$Ln(y/(1-y)) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \dots + \beta_n * x_n$$

Ln(y/(1-y))称为Logit变换。我们再将y视为y取值为1的概率p(y=1)，因此，1-y就是y取值为0的概率p(y=0)，所以上式改写为：

$p(y=1) = e^z/(1+e^z)$,

$p(y=0) = 1/(1+e^z)$,

其中，$z = \beta_0 + \beta_1*x_1 + \beta_2*x_2 + \beta_3*x_3 +...+ \beta_n*x_n$.

接下来就可以使用”最大似然法”估计出各个系数β。

二 odds与OR复习

odds: 称为几率、比值、比数，是指某事件发生的可能性(概率)与不发生的可能性（概率）之比。用p表示事件发生的概率，则：odds = p/(1-p)。

OR：比值比，为实验组的事件发生几率(odds1)/对照组的事件发生几率(odds2)。

三 Logistic回归结果的解读

我们用一个例子来说明，这个例子中包含200名学生数据，包括1个自变量和4个自变量：

因变量: hon，表示学生是否在荣誉班(honors class)，1表示是，0表示否；

自变量：

female ：性别，分类变量，1=女，0=男

read: 阅读成绩，为连续变量

write: 写作成绩，为连续变量

math：数学成绩，为连续变量

1 、不包含任何变量的Logistic回归

首先拟合一个不包含任何变量的Logistic回归，

模型为 $\ln(p/(1-p)) = \beta_0$

回归结果如下（结果经过编辑）：

hon	系数 β	标准误	P
截距	-1.12546	0.164	0.000

这里的系数 β 就是模型中的 β_0 = -1.12546,

我们用p表示学生在荣誉班的概率，所以有ln(p/(1-p) = β_0 = -1.12546,

解方程得：p = 0.245。

odds = p/1-p = 0.3245

这里的p是什么意思呢？p就是所有数据中hon=1的概率。

我们来统计一下整个hon的数据:

hon	例数	百分比
0	151	75.5%
1	49	24.5%

hon取值为1的概率p为49/(151+49) = 24.5% = 0.245，我们可以手动计算出ln(p/(1-p) = -1.12546，等于系数 β_0。可以得出关系：

β_0=ln(odds)。

2、包含一个二分类因变量的模型

拟合一个包含二分类因变量female的Logistic回归，

模型为 ln(p/(1-p) = β_0 + β_1* female.

回归结果如下（结果经过编辑）：

hon	系数 β	标准误	P
female	0.593	.3414294	0.083
截距	-1.47	.2689555	0.000

在解读这个结果之前，先看一下hon和female的交叉表：

hon	female		Total
	Male	Female	
0	74	77	151
1	17	32	49

Total	91	109	

根据这个交叉表，对于男性（Male），其处在荣誉班级的概率为17/91，处在非荣誉班级的概率为74/91，所以其处在荣誉班级的几率odds1=(17/91)/(74/91) = 17/74 = 0.23；相应的，女性处于荣誉班级的几率odds2 = (32/109)/(77/109)=32/77 = 0.42。女性对男性的几率之比OR = odds2/odds1 = 0.42/0.23 = 1.809。我们可以说，女性比男性在荣誉班的几率高80.9%。

回到Logistic回归结果。截距的系数-1.47是男性odds的对数（因为男性用female=0表示，是对照组），ln(0.23) = -1.47。变量female的系数为0.593，是女性对男性的OR值的对数，ln(1.809) = 0.593。所以我们可以得出关系: OR = exp(β)，或者 β = ln(OR)（exp(x)函数为指数函数，代表e的x次方）。

3、包含一个连续变量的模型

拟合一个包含连续变量math的Logistic回归，

模型为 $\ln(p/(1-p)) = \beta_0 + \beta_1 * math.$

回归结果如下（结果经过编辑）：

hon	系数 β	标准误	P
math	.1563404	.0256095	0.000
截距	-9.793942	1.481745	0.000

这里截距系数的含义是在荣誉班中math成绩为0的odds的对数。我们计算出odds = exp(-9.793942) = .00005579，是非常小的。因为在我们的数据中，没有math成绩为0的学生，所以这是一个外推出来的假想值。

怎么解释math的系数呢？根据拟合的模型，有：

ln(p/(1-p)) = - 9.793942 + .1563404*math

我们先假设math=54，有：

ln(p/(1-p))(math=54) = - 9.793942 + .1563404 *54

然后我们把math提高提高一个单位，令math=55，有：

ln(p/(1-p))(math=55) = - 9.793942 + .1563404 *55

两者之差：

ln(p/(1-p))(math=55) - ln(p/1-p))(math = 54) = 0.1563404.

正好是变量math的系数。

由此我们可以说，math每提高1个单位，odds（即p/(1-p)，也即处于荣誉班的几率）的对数增加0.1563404。

那么odds增加多少呢？根据对数公式：

ln(p/(1-p))(math=55) - ln(p/1-p))(math = 54) = ln((p/(1-p)(math=55)/ (p/(1-p)(math=54))) = ln(odds(math=55)/ odds(math=54)) = 0.1563404.

所以：

odds(math=55)/ odds(math=54) = exp(0.1563404) = 1.169.

因此我们可以说，math每升高一个单位，odds增加16.9%。且与math的所处的绝对值无关。

聪明的读者肯定发现，odds(math=55)/ odds(math=54)不就是OR嘛！

4 、包含多个变量的模型（无交互效应）

拟合一个包含female、math、read的Logistic回归，

模型为 $\ln(p/(1-p)) = \beta_0 + \beta_1 * \text{math} + \beta_2 * \text{female} + \beta_3 * \text{read}$.

回归结果如下（结果经过编辑）：

hon	系数β	标准误	P
math	.1229589	略	0.000
female	0.979948	略	0.020
read	.0590632	略	0.026
截距	-11.77025	略	0.000

该结果说明:

（1） 性别：在math和read成绩都相同的条件下，女性（female=1）进入荣誉班的几率（odds）是男性（female=0）的exp(0.979948) = 2.66倍，或者说，女性的几率比男性高166%。

（2） math成绩：在female和read都相同的条件下，math成绩每提高1，进入荣誉班的几率提高13%（因为exp(0.1229589) = 1.13）。

（3）read的解读类似math。

5 、包含交互相应的模型

拟合一个包含female、math和两者交互相应的Logistic回归，

模型为 $\ln(p/(1-p)) = \beta_0 + \beta_1 * female + \beta_2 * math + \beta_3 * female * math.$

所谓交互效应，是指一个变量对结果的影响因另一个变量取值的不同而不同。

回归结果如下（结果经过编辑）：

hon	系数 β	标准误	P
female	-2.899863	略	0.349
math	.1293781	略	0.000
female*math	.0669951	略	0.210
截距	-8.745841	略	0.000

注意：female*math项的P为0.21，可以认为没有交互相应。但这里我们为了讲解交互效应，暂时忽略P值，姑且认为他们是存在交互效应的。

由于交互效应的存在，我们就不能说在保持math和female*math不变的情况下，female的影响如何如何，因为math和female*math是不可能保持不变的!

对于这种简单的情况，我们可以分别拟合两个方程，

对于男性（female=0）：

$\log(p/(1-p)) = \beta_0 + \beta_2 * math.$

对于女性（female=1）：

$\log(p/(1-p)) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) * math.$

然后分别解释。