中国卫生统计，2005，22（5）：293-296

比例优势模型实现ROC分析的方法及其应用前景分析*

华中科技大学同济医学院公共卫生学院流行病与卫生统计学系　宇传华 余松林
第四军医大学预防医学系卫生统计学教研室　徐勇勇

比例优势模型（proportional odds model）是有序回归模型中的一种，有时也叫累积logit模型（cumulative logit model）[1]。李康等采用这一模型考虑了几种ROC分析的实际情况[2]，但未见有文献对这一模型实现ROC分析的实际应用价值作出评价。本研究试图将这种方法与经典公认的ROC分析方法作出比较，探讨这一模型在ROC分析中的实际应用前景。

原理与方法

1．ROC分析的比例优势模型

进行ROC分析时，数据至少需要有两个变量，一个是按“金标准”分类的疾病状态变量（state variable），一般为二分类，记为D，D＝1表示“异常组”，相应的概率记为π，D＝0表示“正常组”，相应的概率记为1－π；另一个是试验结果变量（test variable），它可以是有序分类变量（如放射医生将患者诊断分类为肯定无癌症、可能无癌症、疑似癌症、可能癌症、肯定癌症，分别数量化为1、2、3、4、5），也可以是连续型变量（如红细胞平均容积），记为T。事先假定T值越大，患者被诊断为阳性的可能性越大，且令T为反应变量，D为解释变量，则ROC分析所采用的比例优势模型可表示为：

$$\text{logit}[\pi(T \ge c_j)] = \ln\frac{\pi(T \ge c_j)}{1-\pi(T \ge c_j)} = \alpha_j + \beta D, \quad j = 1, \cdots\cdots, J-1 \qquad (1)$$

其中c_j为第j个诊断界值（cutpoint value），如果两组试验结果变量T有J个不重叠的数值，则通常将J个值按大到小顺序排列，最大值记为c_1，最小值记为c_J，以较大的J－1个不重叠T值（c_1~c_{J-1}）作为诊断界值，每一诊断界值将试验结果分成两部分，$T \ge c_j$则试验结果阳性，$T < c_j$则试验结果阴性。$\pi(T \ge c_j)$表示阳性试验结果的概率。ln为自然对数符号。α_j和β为模型的两个待估参数；对于每一个诊断界值c_j，公式（1）中的变量D均有相同的系数β；自然对数为底的β次幂，即$\exp(\beta)$称为优势比（odds ratio，OR），这也是为什么该模型被称为比例优势模型的原因。

将公式（1）进行简单的代数变换，可得到阳性试验结果的概率为：

$$\pi(T \ge c_j) = \frac{\exp(\alpha_j + \beta D)}{1+\exp(\alpha_j + \beta D)} = \frac{1}{1+\exp[-(\alpha_j + \beta D)]}, \quad j = 1, \cdots\cdots, J-1 \qquad (2)$$

假阳性率（FPR），即1-特异度，是“正常组”中试验结果为阳性的概率，根据公式（2）有：

$$\text{FPR}_j = \pi(T \ge c_j \,|\, D = 0) = \frac{\exp(\alpha_j + \beta \times 0)}{1+\exp(\alpha_j + \beta \times 0)} = \frac{1}{1+\exp(-\alpha_j)} \qquad (3)$$

真阳性率（TPR），即灵敏度，是“异常组”中试验结果为阳性的概率，根据公式（2）有：

* 国家自然科学基金资助项目（30371254）

$$\mathrm{TPR}_j = \pi(T \geq c_j \mid D=1) = \frac{\exp(\alpha_j + \beta \times 1)}{1+\exp(\alpha_j + \beta \times 1)} = \frac{1}{1+\exp(-\alpha_j - \beta)} \qquad (4)$$

将公式（3）整理为$\exp(-\alpha_j) = (\mathrm{FRP}_j^{-1} - 1)$后，代入公式（4）得ROC曲线方程为：

$$\mathrm{TPR} = \frac{1}{1+(\mathrm{FPR}^{-1}-1)\exp(-|\beta|)} \qquad (5)$$

因为消除了与*j*有关的项α_j，所以公式（5）去除了TPR与FPR间的下标。之所以在公式（5）中对β取绝对值，是为了保证曲线下面积大于0.5。通过积分计算，可求得到ROC曲线下面积*A*为

$$A = \int_0^1 \mathrm{TPR}d(\mathrm{FPR}) = 1 + \frac{1}{OR-1} - \frac{OR\ln OR}{(OR-1)^2} \qquad (6)$$

其中，$OR = \exp(|\beta|)$，称为优势比。当$|\beta| = 0$或$OR = 1$，有面积$A = 0.5$，表示试验无诊断能力；当$|\beta|$很大，$OR = \infty$，有$A = 1$。利用泰勒展开式，A的近似标准误为：

$$SE(A) = \frac{\ln OR(OR+1) - 2(OR-1)}{(OR-1)^3} OR \cdot SE(\beta) \qquad (7)$$

在大样本情况下，按近似正态分布原理，可通过$A \pm 1.96SE(A)$计算ROC曲线下面积的95%置信区间。

2. 模型的评价指标

为了将比例优势模型与经典ROC分析参数方法——双正态模型（binormal model）进行比较，分别计算每一所需比较模型的灵敏度残差平方和与决定系数两个指标。灵敏度的残差平方和计算公式为$\sum(\hat{\mathrm{TPR}} - \mathrm{TPR})^2$，其值越小模型越好；灵敏度决定系数是灵敏度估计值$\hat{\mathrm{TPR}}$与实际值$\mathrm{TPR}$之间的简单相关系数之平方，记为$R^2$，其值越大模型越好。其中的$\mathrm{TPR}$是由真实数据计算得到的实际散点灵敏度值（相当于直线回归散点的纵坐标值），以此值作为真值，将比例优势模型、双正态模型等的灵敏度拟合值（$\hat{\mathrm{TPR}}$）与之相结合，分别计算上述两个模型评价指标。

实例分析

ROC分析的资料类型一般分为有序分类与连续型两种。下面每种类型各举一例。

例1 表1为某放射医生对234张影像片按正常、可能正常、疑似异常、可能异常和异常（分别计分为1、2、3、4、5）进行分类的结果，问该放射医生的诊断准确度如何？

表1　有序分类资料

分类	1	2	3	4	5	合计
正常组	35	68	49	29	12	193
异常组	2	3	8	16	12	41

对于表1数据，拟合公式（1）的比例优势模型，得到疾病状态变量*D*的回归系数β=1.9894，优势比*OR* =exp(1.9894)=7.3111，比例优势模型的ROC曲线方程为：

$$\hat{TPR} = \frac{1}{1+(FPR^{-1}-1)\exp(-1.9894)} \qquad (8)$$

由公式（5）和（6）得到的ROC曲线下面积及其标准误见表2的第2、3列，表2第4列为按正态分布原理计算得到的95％ROC面积的置信区间。为了与公认的几种方法比较，表2也列出了针对表1数据的双正态模型[3]、Hanley-McNeil非参数法[4]、Delong-Delong非参数法[5]的计算结果。由表2可见，这些方法获得的结果很相似。

表2　四种ROC分析方法对表1数据的计算结果

方法	ROC面积	标准误	95%置信区间
比例优势模型	0.7933	0.0381	(0.7185,0.8680)
双正态模型	0.7855	0.0397	(0.6999,0.8549)
Hanley-McNeil非参数法	0.7797	0.0403	(0.7006,0.8588)
Delong- Delong非参数法	0.7797	0.0396	(0.7021,0.8574)

公式（8）（比例优势模型的ROC曲线方程）对应的ROC曲线见图1。为了比较，图1中也绘出了实际散点（FPR, TPR），以及双正态模型的拟合曲线（参数a =1.1172，b =0.9973，由ROCKIT软件计算）（因为非参数法不能绘制光滑ROC曲线，所以这里只比较参数模型的曲线）。从图1可见，实际散点分布在比例优势模型与双正态模型所拟合曲线的周围，两模型拟合均较好。比例优势模型与双正态模型所得的灵敏度残差平方和分别为0.0021与0.0070，决定系数分别为0.9938与0.9747。由此可见，两模型的这两个指标相当接近，其中比例优势模型获得的灵敏度残差平方和相对较小，决定系数较大，比例优势模型的拟合似乎略好于双正态模型，从图1也可以直观看出这一点。

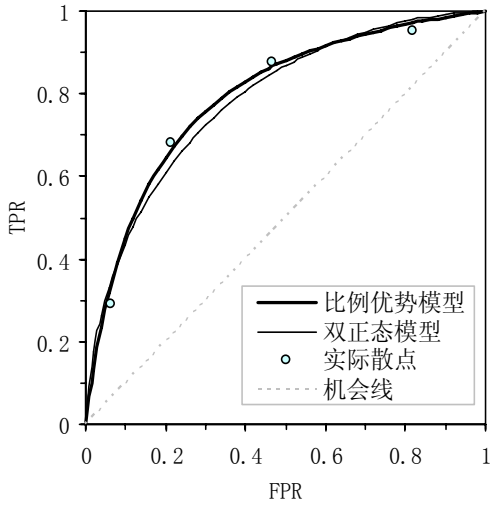

图1　例1数据的ROC曲线

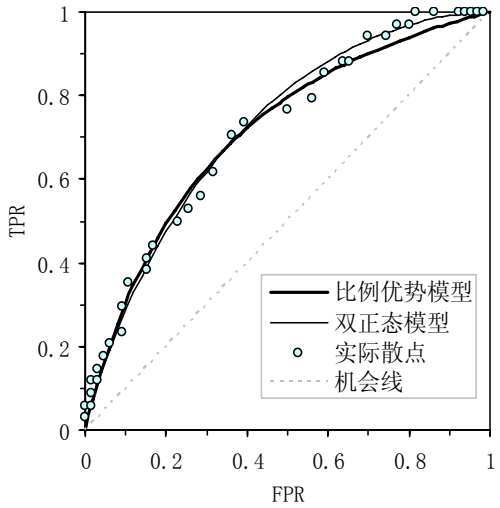

图2　例2数据的ROC曲线

表 3　连续型资料

骨髓诊断	红细胞平均容积 MCV 结果																
正常组	60	66	68	69	71	71	73	74	74	74	76	77	77	77	77	78	78
	79	79	80	80	81	81	81	82	82	83	83	83	83	83	83	83	84
	84	84	84	85	85	86	86	86	87	88	88	88	89	89	89	90	90
	91	91	92	93	93	93	94	94	94	94	96	97	98	100	103		
异常组	52	58	62	65	67	68	69	71	72	72	73	73	74	75	76	77	77
	78	79	80	80	81	81	81	82	83	84	85	85	86	88	88	90	92

例 2 采用骨髓诊断作为金标准，将 100 例患者中的 34 例确诊为缺铁性贫血（异常组），其余 66 例确诊为非缺铁性贫血（正常组），事先测得每个患者的红细胞平均容积（MCV）见表 3，试采用 ROC 分析评价 MCV 诊断缺铁性贫血的价值。

对表3数据拟合公式（1）的比例优势模型，得到变量D的回归系数β=−1.3614（本例是MCV结果较小时诊断为阳性，所以得到负的β值），比例优势模型的ROC曲线方程为：

$$\hat{\mathrm{TPR}} = \frac{1}{1+(\mathrm{FPR}^{-1}-1)\exp(-1.3614)} \qquad (9)$$

本例优势比$OR=\exp\left(\left|-1.3614\right|\right)$=3.9017，四种方法计算得到的ROC曲线下面积、标准误及其95％置信区间见表4。同样，这四种方法获得的结果十分相似。该数据的比例优势模型和双正态模型对应的ROC曲线见图2。比例优势模型所得的灵敏度残差平方和为0.0355，决定系数为0.9929；双正态模型所得的灵敏度残差平方和为0.0307，决定系数为0.9948。由此可见，两模型的这两个指标相当接近，其中以比例优势模型拟合ROC曲线的效果略差于双正态模型。从图2也可以直观看出这一点。

表4　四种ROC分析方法对表3数据的计算结果

方法	ROC面积	标准误	95%置信区间
比例优势模型	0.7138	0.0537	(0.6085,0.8190)
Metz双正态模型	0.7233	0.0516	(0.6144,0.8145)
Hanley-McNeil非参数法	0.7170	0.0526	(0.6139,0.8201)
Delong- Delong非参数法	0.7170	0.0529	(0.6134,0.8206)

讨论

无论ROC分析资料类型属于有序分类数据，还是属于连续型数据，比例优势模型对实际散点（FPR,TPR）的拟合均较好。由比例优势模型得到的ROC曲线下面积、标准误及其95％置信区间的计算结果与公认的双正态参数模型、以及其他两种非参数方法的计算结果十分接近。

与双正态模型相比，比例优势模型有如下优点：

1. 可借助于现有软件实现

为了获得比例优势模型的ROC曲线方程中变量D 的回归系数β值，对于连续型资料（如例2），其主要SAS程序为“**PROC LOGISTIC** DESCENDING;MODEL T=D;”，其中T为试验结果变量，D为疾病状态变量；对于有序分类资料（如例1），如果频数变量名记为freq，则在上述SAS程序的MODEL语句前需加上“FREQ freq;”。

为了获得各诊断界值的实际散点（FPR, TPR），对于有序分类资料，其主要SAS程序为“**PROC LOGISTIC** DESCENDING; FREQ freq; MODEL D=T/OUTROC=roc; **PROC PRINT**; **RUN**;”，连续型资料省略“FREQ freq;”语句即可。对于双正态模型，目前难以采用标准软件（如SAS）实现，通常需采用Metz等开发的专有软件[3]，如ROCKIT软件。

2. 可简单考虑协变量的影响

如果假定正常组与异常组的协变量X（如性别、年龄等）相同，将这些协变量放入模型公式（1）中，且不考虑协变量与变量D间的交互作用，按上述方法进行推导，得到的ROC曲线方程、曲线下面积及其标准误的计算公式，分别与公式（5）、（6）、（7）相同。也就是说，与α_j类似，ROC曲线方程中不含与协变量有关的项（$\beta_X X$）。这一结果表明，为了控制协变量对ROC曲线的影响，只需将协变量放入比例优势模型中，将获得的疾病状态变量D

的回归系数（或优势比）及其标准误代入公式（6）、（7）中，便可简单计算ROC曲线下面积及其标准误。Metz等软件的双正态模型目前难以考虑协变量的影响。

由于比例优势模型假定每一诊断界值对应的变量*D*的优势比相等，所以比例优势模型获得的ROC曲线形状单一，这是该模型的致命缺陷。令参数 β 分别为0.5、1、1.5、2、2.5，将此参数代入公式（5），得到的5条ROC曲线见图3所示，这5条曲线明显表现出了与主对角线间的对称关系。如果实际数据散点分布偏离比例优势模型的假定，则比例优势模型的ROC曲线拟合效果明显差于双正态模型（见图4）。因此，比例优势模型在ROC分析中的应用尚有待改进。解决此问题的一种方法是：在模型（1）的基础上增加尺度参数项[6]。

总之，比例优势模型实现 ROC 分析具有可控制协变量的影响，且参数的计算相对容易等特点；但由于该模型获得的 ROC 曲线形状单一，实际应用时应特别加以注意。

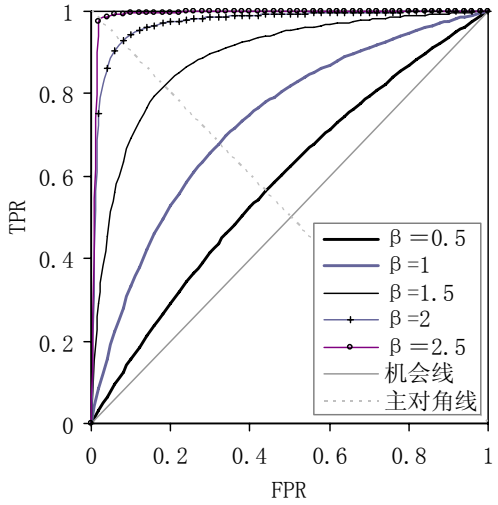

图3　5个 β 参数对应的5条ROC曲线

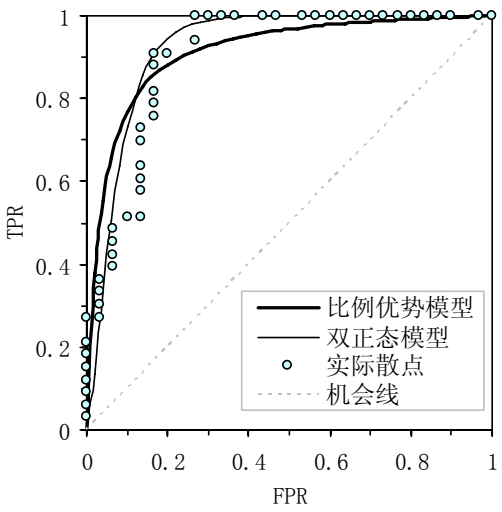

图4　实际散点分布较特殊的情况

参考文献

1. 余松林 主编. 医学统计学（7 年制规划教材）. 北京：人民卫生出版社, 2002, 224-228.
2. 李康, 马葆华, 赵亚双等. 具有协变量或干扰因素的诊断试验数据的 ROC 分析. 中国卫生统计, 2002, 19(2): 67-70.
3. Metz CE, Herman BA, Shen JH. Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continously-distributed data. Statistics in Medicine 1998; 17:1033-1053.
4. Hanley JA,McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.
5. DeLong ER, DeLong DM and Clarke-Pearson DL. Comparing the areas under two or more correlated receive operating characteristic curves: a nonparametric approach. Biometrics 1988; 44:837-845.
6. Cox, C. Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach. Statistics in Medicine 1995; 14: 1191-1203.

【提 要】 目的 探讨比例优势模型在 ROC 分析中的应用前景。**方法** 比较比例优势模型与双正态模型等经典方法所计算的 ROC 曲线下面积及其标准误；采用灵敏度残差平方和与决定系数两个指标评价参数模型的拟合优度。**结果** 在一般情况下，由比例优势模型所得到的 ROC 曲线指标结果与经典方法很接近；对于有序分类资料和连续型资料，该模型的拟合效果均较好；但由于该模型获得的 ROC 曲线形状单一，有些情况下该模型的拟合不理想。**结论** 与经典的方法相比，比例优势模型有其自身的特点，实际应用时应慎重做出选择。

The proportional odds model and its prospect applied to ROC analysis　Yu Chuanhua, Yu Songlin. Department of Epidemiology and Health Statistics, Tongji Medical College, Huazhong University of Science and Technology　Xu Yongyong. Department of Health Statistics, the Fourth Military Medical University

[**Abstract**]　**Objective** Explore the role of proportional odds model in ROC analysis. **Methods** Compare the area under ROC curve of proportional odds model and its standard errors with of common methods, such as binormal model; evaluate the goodness of fit for the parametric models using the sum of square of residuals of sensitivity and determination coefficient between actual sensitivity and estimated sensitivity. **Results** The values of the measures of ROC analysis from the proportional odds model are similar with of common methods; whether ordinal category data or continuous data, the effect of fit for the model are generally good. **Conclusion** Comparing with common methods of ROC analysis，the proportional odds model have itself characteristic. It is an alternative method for ROC analysis in practice.

关键词：　ROC 曲线　比例优势模型　模型评价 应用

Key Words:　ROC curve　Proportional odds model　Model evaluation　Application