逐步回归与判别分析的应用研究
——在乳腺疾病建模中的应用

吉国力[1]，陈舒婷[1]，张延坤[1]，李 健[2]，陈承祺[3]
(1. 厦门大学自动化系，福建 厦门 361005; 2. 厦门市妇幼保健医院，福建 厦门 361005;
3. 厦门市第一医院肿瘤科，福建 厦门 361005)

[摘 要] 体内激素紊乱是导致乳腺疾病发生的重要原因. 将逐步回归法与 Fisher判别法结合应用于乳腺疾病建模中，对乳腺疾病与 6种内分泌激素及患者年龄的关系进行了多元线性判定分析，建立了变量与相关疾病的逻辑回归模型与判别模型. 该模型通过实际应用，判别效果较好，有助于乳腺疾病的诊断.

[关键词] 乳腺增生; 乳腺癌; 激素; 模型; 判别分析
[中图分类号] O 212; R 655.8 [文献标识码] A [文章编号] 1008 - 3804 (2006) 02 - 0022 - 05

乳腺癌是妇女中十分常见的恶性肿瘤，我国虽然为乳腺癌的低发区，但近年来发病呈上升趋势，已成为影响妇女健康的主要因素，防治乳腺癌的重要性已经日渐突出. 乳腺癌的病因在医学界还不十分清楚，激素、遗传、免疫、生理与环境因素相互作用，可能共同参与乳腺组织的癌变和演变过程[1].

Fisher判别法是多变量分析方法中较成熟的一类方法，近年来在医学领域获得了广泛的应用. 但是，在判别模型建立过程中，从可获得的实际数据出发来选择模型变量，以此建立的判别函数其判别精度高低受变量之间的相关性影响很大，有些情形会提高判别效果，有些情形相反. 因此恰当选择模型变量是一个重要问题[2].

逐步回归法的基本思想是变量一一引入，即偏回归平方和 F检验显著的变量才作为引入变量，新变量引入后还要重新对所有已引入的变量进行检验，不显著者从方程中剔除，直到没有变量可剔除也没有变量可引入时为止，最后对所选定变量建立线性回归方程[3].

本文将逐步回归法与 Fisher判别法结合应用于乳腺疾病建模中，并借助秩和检验与 Spearman秩相关系数法进行辅助分析. 用 124例训练样本建模得到的判别模型可以根据激素水平对乳腺疾病进行分类，通过 29例新样本检验表明：该模型具有较高的判别率.

1 理论模型

1.1 逐步回归分析的实现原理

设共有 n个可供选择的变量 $x_1, x_2, \cdots, x_n$，则可以用逐步回归法构建下述回归模型：

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

逐步回归是将变量一一引入. 首先，选择第一个变量. 记各自变量的偏回归平方和为：

$$V_j^{(1)} = (r_{jy}^{(0)})^2 / r_{jj}^{(0)}, \ j = 1, 2, \cdots, n$$

设 $k1$使得 $V_{k1}^{(1)} = \max_j \{V_j^{(1)}\}$，则 x_{k1}入选. 作入选检验，在不相关的假设下，取统计量：

$$F_{进}(k1) = \frac{(n - 1 - 1)V_{k1}^{(1)}}{r_{yy}^{(0)} - V_{k1}^{(1)}} \sim F(1, n - 2)$$

[收稿日期] 2006 - 02 - 25 [修回日期] 2006 - 04 - 10
[基金项目] 厦门市社会发展计划项目 (3502Z20044003)
[作者简介] 吉国力 (1960 -)，男，山西侯马人，教授，硕士，从事系统工程理论方法与应用的研究.

若 $F_{进}(k1) > F_{\alpha}^{*}$，则 x_{k1}进入回归方程．选第二个变量．计算偏回归平方和：$V_j^{(2)} = (r_{jy}^{(1)})^2 / r_{jj}^{(1)}$，$j \neq k1$，设 $k2$使得 $V_{k2}^{(2)} = \max_j \{V_j^{(2)}\}$，则 x_{k2}入选．作入选检验，在不相关的假设下，取统计量：$F_{进}(k2) = \dfrac{(n-2-1)V_{k2}^{(2)}}{r_{yy}^{(1)} - V_{k2}^{(2)}} \sim F(1, n-3)$，若 $F_{进}(k2) > F_{\alpha}^{*}$，则 x_{k2}进入回归方程[4]．由于入选 x_{k2}，需对 x_{k1}重新作显著性检验．一般讲，设前面已选入了变量 $x_{k1}, x_{k2}, \cdots, x_{kl}$，由于引入了新变量，需对原有变量的显著性逐一作检验，考察是否有变量需要剔除．假如无变量可剔除，也无变量可引入，则可写出关于 $x_{k1}, \cdots, x_{kl}$的回归方程：

$$Y = \beta_{k1}^{(l)} x_{k1} + \cdots + \beta_{kl}^{(l)} x_{kl}$$

1.2　Fisher判别模型

设 $u(X)$是要寻找的最佳判别函数，其中：

$$X = (x_1, x_2, \cdots, x_n)^T$$

为待判样本．Fisher准则的思想是：要使得母体 π_1与 π_2间的距离 $(\bar{u}_1 - \bar{u}_2)^2$尽可能地大，而同时使各母体内方差 σ_1^2和 σ_2^2尽可能地小，即在：

$$\lambda = (\bar{u}_1 - \bar{u}_2)^2 / (q_1\sigma_1^2 + q_2\sigma_2^2)$$

为极大的条件下，$u(X)$作为判别函数．其中 $\bar{u}_i$表示均值，q_i为 π_i的某种先验率或者权数[3]．

如记 $f_i(X)$为母体 π_i中 X概率密度，则：

$$\bar{u}_i = \int u(X) f_i(X)\,\mathrm{d}X, \quad \sigma_i^2 = \int (u - \bar{u}_i)^2 f_i(X)\,\mathrm{d}X$$

在 Fisher准则下的线性判别函数为：

$$u(X) = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n \triangleq C^T X,$$

$$C = (c_1, c_2, \cdots c_n)^T$$

此时：

$$\bar{u}_i = C^T \mu^{(i)}, \quad i = 1, 2; \qquad \sigma_i^2 = C^T \sum^{(i)} C, \quad i = 1, 2$$

其中 $\mu^{(i)}$，$\sum^{(i)}$为 π_i中 X的均值及协方差阵．则：

$$C = (q_1 \sum^{(1)} + q_2 \sum^{(2)})^{-1} (\mu^{(1)} - \mu^{(2)})$$

由此，根据两个已知不同母体的抽样结果，获得一系列的可观测变量，建立数学模型，从而可用来判别任一新观测的样品应归属哪一个母体．

2　实例应用

模型所采用激素数据由厦门市第一医院核医学科检测，由美国康仁公司生产的 ACS/180SE化学发光仪提供配套检测试剂，检测项目包括垂体激素 3项：促催乳激素 (PRL μg/L)、促卵泡激素 (FSH IU/L)、促黄体激素 (LH IU/L)，性类固醇激素 3项：雌二醇 (E2μg/L)、孕酮 (P μg/L)、睾酮 (T μg/L)．从第一医院 2000～2005年共 7 781条记录中抽取以下两组：

确诊乳腺癌 62例．患者年龄 22～49岁，平均年龄 38.7岁，排除其他伴随疾病史 (垂体、肾上腺亚临床微腺瘤、肝胰胆疾病、子宫卵巢疾病史)，于治疗前，滤泡期抽血检测激素水平．

确诊乳腺增生 62例．患者年龄 19～46岁，平均年龄 33.1岁，排除其他伴随疾病史 (垂体、肾上腺亚临床微腺瘤、肝胰胆疾病、子宫卵巢疾病史)．患者月经规律正常或失调，经前乳腺疼痛明显，时间长达 1～2年，临床可触及不同导管分布区片状增厚组织，呈结节或沙粒状，有触痛．于治疗前，滤泡期抽血检测激素水平．

血浆 E2、FSH、LH、P、T、PRL的测定结果，以中位数 Median $(P_{25} - P_{75})$表示；秩和检验进行组间比较；用 Spearman秩相关系数法对各激素变量以及年龄进行相关分析，判断变量之间的密切程度．

由于随机变量 Y为患者是患乳腺癌还是乳腺增生，服从两点分布，因此，将 Y定义为：

$$Y=\begin{cases}1 & \text{乳腺癌}\\ 0 & \text{乳腺增生}\end{cases}$$

现把乳腺癌发生的概率假定为 p，乳腺增生发生的概率为 $1-p$，建立逻辑回归方程：

$$\ln\left(\frac{p}{1-p}\right)\beta_0+\beta_1 E2+\beta_2 FSH+\beta_3 LH+\beta_4 P+\beta_5 T+\beta_6 PRL+\beta_7 Age$$

其中，E2、FSH、LH、P、T、PRL为检查者测量出的6个激素变量；Age表示患者的年龄；$(\beta_0,\beta_1,\cdots,\beta_7)$为回归系数.

回归过程使用逐步回归分析，从E2、FSH、LH、P、T、PRL和Age 7个元素中挑选出适当的自变量组合，确定入选逻辑回归模型的元素，以建立对这批观测数据来说是最优的回归方程[5].

用逐步回归分析得到的这个回归方程，包含了所有对 Y 显著的变量而不包含对 Y 不显著的变量. 这样，对这批观测数据而言，这组变量就是用来分类的最好一组变量. 由此，判别分析就根据逐步回归分析筛选出的变量对乳腺增生和乳腺癌这两种疾病类型建立判别模型.

设 Y_1 为诊断为乳腺增生的预测值；Y_2 为诊断为乳腺癌的预测值. 使用Fisher判别法，得到的判别函数为：

$$Y_1=c_0+c_1 E2+c_2 FSH+c_3 LH+c_4 P+c_5 T+c_6 PRL+c_7 Age$$

$$Y_2=c_0'+c_1' E2+c_2' FSH+c_3' LH+c_4' P+c_5' T+c_6' PRL+c_7' Age$$

其中，E2、FSH、LH、P、T、PRL为6个激素变量，Age表示患者的年龄；$(c_0,c_1,\cdots,c_7)$与$(c_0',c_1',\cdots,c_7')$为判别方程系数. 逐步回归筛选最终进入模型的变量，若有变量没有进入判别方程，则该变量对应的方程系数为0.

3　结果与分析

对乳腺癌和乳腺增生患者血浆中的E2、FSH、LH、P、T和PRL 6项激素的测定结果进行频数分析，分布水平采用中位数 $(P_{25}-P_{75})$ 表示，结果见表1:

采用Wilcoxon符号平均秩检验对这两组数据进行组间比较，结果见表2:

表1　中位数 $(P_{25}-P_{75})$

激素名称	乳腺疾病名称	
	乳腺癌	乳腺增生
E2	169.71(101.36-246.80)	77.06(68.29-88.64)
FSH	18.88(14.24-34.92)	7.44(4.95-9.59)
LH	23.24(10.54-43.46)	7.60(6.64-9.25)
P	0.56(0.32-1.22)	0.49(0.27-0.78)
T	48.19(33.56-65.22)	64.43(43.51-92.22)
PRL	11.59(7.81-17.15)	15.70(8.79-26.98)

表2　秩和检验

激素名称	检验 p 值	比较
E2	$p=0.000$	<0.05
FSH	$p=0.000$	<0.05
LH	$p=0.000$	<0.05
P	$p=0.285$	>0.05
T	$p=0.002$	<0.05
PRL	$p=0.109$	>0.05

Spearman秩相关系数法对6种激素和年龄进行相关分析，各激素数据和年龄评秩见表3:

表3　7个变量评秩

编号	1	2	3	……	122	123	124
E2	80.0	59.0	17.0	……	61.0	10.0	3.0
FSH	65.0	1.0	15.0	……	45.0	103.0	49.0
LH	75.0	51.0	50.0	……	2.0	66.0	5.0
P	101.0	20.0	75.0	……	52.5	102.0	45.5
T	114.0	21.0	1.0	……	105.0	68.0	11.0
PRL	64.0	51.0	74.0	……	95.0	122.0	17.0
Age	71.0	28.0	45.0	……	71.0	59.5	71.0

Spearman秩相关系数 $\gamma_{xy}^{s} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$，$d_i$表示评秩差值. 选择统计量 $t = \gamma_{xy}^{s}\sqrt{\frac{n-2}{1-\gamma_{xy}^{s2}}}$对相关系数显著性进行双侧检验.

得到的各相关系数见表 4:

表 4　7个变量的 Spearman秩相关系数矩阵

激素名称	E2	FSH	LH	P	T	PRL	Age(年龄)
	相关系数						
E2	1.000						
FSH	0.417*	1.000					
LH	0.574**	0.663**	1.000				
P	0.222*	0.123	0.117	1.000			
T	0.029	-0.181*	-0.019	0.285**	1.000		
PRL	-0.141	-0.156	-0.124	0.015	0.179*	1.000	
Age	0.211*	0.267**	0.068*	-0.035	-0.272**	-0.234**	1.000

*表示 <0.05, 相关系数在 0.05水平下显著; **表示 <0.01, 相关系数在 0.01水平下显著.

表 1与表 2结果显示乳腺癌组的 E2、FSH、LH水平高于乳腺增生组; 乳腺癌组 T的水平低于乳腺增生组, 表明人体血浆 E2、FSH、LH水平升高有增加女性乳腺癌危险性的作用. 乳腺增生组与乳腺癌组 P、PRL区别不显著, 无统计学意义.

表 4分析结果显示激素相互之间有密切的关系. 如 E2与 FSH、E2与 LH、FSH与 LH正线性相关, T与 FSH负线性相关、与 P正线性相关, 说明在 HPO轴系里, 由于存在前馈和反馈作用, 使得激素间存在着复杂的相关性. 又如 Age (年龄) 和除 P以外的激素间都具有显著的相关性, 表明年龄大小也影响激素水平的分布.

将乳腺增生和乳腺癌分别赋值为 0和 1, 进行逻辑逐步回归分析, 所得方程为:

$$\ln\left(\frac{p}{1-p}\right) = -9.949 + 0.021E2 + 0.336FSH - 0.048T + 0.056PRL + 0.151Age$$

其中 p为患乳腺癌的概率.

最后保留在方程中的元素为 E2, FSH, T, PRL和 Age (年龄). 由上面的秩和检验分析结果, 因为激素 P组间区别不显著, 而 LH虽然秩和检验分析显示区别显著, 但相关分析显示与 E2, FSH存在正线性相关关系, 即存在协同作用, 故这两个激素均没有入选回归方程.

逻辑回归方程的参数估计见表 5, 由表 5可以看出, 入选方程的变量显著水平均小于 0.05, 对因变量的影响显著.

表 5　参数估计

变量名称	似然估计	Wald	检验 p值
E2	0.008	6.397	0.011
FSH	0.093	13.044	0.000
T	0.016	8.788	0.003
PRL	0.022	6.348	0.012
Age	0.056	7.181	0.007

以回归方程入选的 E2, FSH, T, PRL和年龄 5个元素, 用 Fisher判别法进行判别分析. 设 Y_1为诊断为乳腺增生的预测值, Y_2为诊断为乳腺癌的预测值, 得到的判别方程为:

$$Y_1 = -17.545 + 0.015E2 + 0.105FSH + 0.076T + 0.118PRL + 0.740Age$$

$$Y_2 = -26.595 + 0.033E2 + 0.244FSH + 0.053T + 0.158PRL + 0.881Age$$

将 124例训练样本回代判别方程进行检验：对乳腺增生的判别准确率高达 100%，对乳腺癌的判别准确率为 79%，总体判别准确率为 89.5%，说明建立的判别方程还是比较灵敏可靠的，能够较为准确地对乳腺癌和乳腺增生加以判别.

用所得判别方程对新样本进行检验，对确诊乳腺增生 15例、乳腺癌 14例的测定激素进行预测：对乳腺增生的判别准确率为 93.3%，对乳腺癌的判别准确率为 92.9%，总体判别准确率为 93.1%，从小样本上证明判别方程对乳腺增生、乳腺癌的判别是有效的.

4 结语

本文采用逐步回归法与 Fisher判别法结合的方法对乳腺疾病与 6种内分泌激素的关系进行了多元线性判定分析，建立了变量与相关疾病的逻辑回归模型与判别模型，考察了究竟是哪些激素变量的改变影响着相关乳腺疾病的发生. 从而用来判别任一新观测到的激素样品应归属于乳腺增生或是乳腺癌，有助于乳腺疾病诊断分析.

目前存在的主要问题：1）样本量少，影响模型参数确定；2）在数据收集过程中，激素与疾病症状、患者体征等数据没有及时记录，影响模型变量确定；3）在判别分析方面，一些新的医学测定还没有结合考虑.

致谢：在研究过程中，从数据来源、分析和对研究结果的检验，得到了厦门市妇幼保健医院钼靶X光科林玉斌副主任、厦门市第一医院俞丹副主任和厦门市妇幼保健医院钱浩勇主任的大力协助，在此表示感谢.

[参考文献]

[1] 肖凤，孙朝越. 乳腺癌的危险因素及预后分析 [J]. 中国基层医药，2004，(11)：1258-1260.
[2] 孙尚拱，潘恩沛. 实用判别分析 [M]. 北京：科学出版社，1990.
[3] 周纪芗. 回归分析 [M]. 上海：华东师范大学出版社，1993.
[4] 骆振华. 概率统计简明教程 [M]. 厦门：厦门大学出版社，1990.
[5] 周文芳，李民. 逐步回归分析法的一点不足之处 [J]. 西北水电，2004，(4)：49-50.

Stepwise Regression and Discriminant Analysis in Breast Disease Modeling

JI Guo-li[1], CHEN Shu-ting[1], ZHANG Yan-kun[1], LI Jian[2], CHEN Cheng-qi[3]

(1. Department of Auto. Xiamen University, Xiamen 361005, China; 2. Xiamen Women and Children Health Care Hospital, Xiamen 361003, China; 3. First Hospital of Xiamen, Xiamen 361003, China)

Abstract: Hormone disorder is an important cause leading to breast disease. This article combined Stepwise Regression and Fisher Discriminant in building a model of breast disease. The relationship between six kinds of endocrine hormone, patient's age and breast disease is analyzed through Multiple Linear Discriminant Method. In the end, for element variable and correlative disease, this article gets a logistic regression model and a discriminant model. Through application, the model is testified to be effective for the diagnosis of breast disease.

Key words: hyperplasic of the breast; beast cancer; hormone; discriminant analysis