【统计理论与方法】

判别分析与 Logistic 回归的模拟比较

张初兵[1,2a]，高 康[2b]，杨贵军[2b]
（1. 天津大学 管理学院，天津 300072；2. 天津财经大学 a. 商学院，
b. 中国经济统计研究中心，天津 300222）

摘要：利用随机模拟方法，研究判别分析和 Logistic 回归分类的回判正确率。模拟结果显示，Logistic 回归的回判正确率优于判别分析。随着随机误差的增大，Logistic 回归与判别分析的回判正确率差异逐渐减小。随机误差超过一定界限，Logistic 回归的回判正确率低于判别分析。在随机模拟的基础上，引入修正 Logistic 回归分类，模拟结果显示，修正 Logistic 回归分类略优于 Logistic 回归。

关键词：判别分析；Logistic 回归；误判率；回判正确率
中图分类号：F224.7；O212.1 文献标志码：A 文章编号：1007－3116(2010)01－0019－07

一、引 言

在研究实际问题时，经常遇到分类问题。在很多情况下，为了研究目的，将研究对象分为两类，称为二分类问题。例如，为了研究保险公司破产原因，将所有保险公司分为两类，破产保险公司作为一类，正常运营保险公司作为另一类。通过这两类保险公司数据的对比，可以归纳出保险公司破产的影响因素以及分类方法，对当前运营保险公司破产可能性进行分析。鉴于二分类问题的代表性和广泛性，本文主要讨论二分类问题。本文的研究方法可以推广到多分类问题。

在统计学中，常用的分类方法是判别分析和 Logistic 回归。这两种方法简单实用，很多统计软件可以完成有关的计算，它们应用很广，特别是医学生物学领域和经济管理等研究领域。在医学生物学领域中，Biometrics，Biometrical Journal 等学术刊物每年都刊登很多判别分析或 Logistic 回归应用的论文。在国内学术刊物中，这两种方法的应用也很多。白玉峰等借助于判别分析对心血管功能进行定量的判别与预测[1]。张晓东、申洪利用判别分析探讨肺癌细胞核的有关体视学参数在肺癌诊断分型方面的意义[2]。易尚辉等对因大肠癌而住院的病历按治愈和未愈分两组进行非条件多因素 Logistic 回归分析[3]。王浩等通过 Logistic 回归探讨进展期胃癌淋巴结的转移规律[4]。

在经济管理领域，Zavgren 利用判别分析和 Logistic 回归对保险公司破产原因进行分析，量化保险公司倒闭前 5 年的公司金融问题信号，作为金融风险概率显著性的评价方法[5]。Lee、Hyun 和 Urrutia 利用 Logistic 模型预测非寿险公司偿付能力，并检测显著影响非寿险公司偿付能力的因素[6]。Dimitras 等和 Ganderton 等对判别分析和 Logistic 预测商务失败的效果进行了评价[7－8]。张玲等利用多元判别分析和神经网络对我国上市公司财务困境进行预警分析[9]。张立军等以我国沪市 A 股上市公司为研究对象，利用判别分析研究上市公司财务危机预警[10]。张成虎等基于个人消费信贷数据，建立

收稿日期：2009－10－15；修复日期：2009－11－25
基金项目：国家社会科学基金项目《基于数据挖掘技术的中国保险公司偿付能力研究》(07CTJ002)；教育部新世纪优秀人才支持计划《我国保险公司风险的监管量化技术及监管机制研究》(NCET－08－0909)
作者简介：张初兵(1984－)，男，安徽六安人，博士生，研究方向：营销管理、金融工程；
高 康(1987－)，男，河北鹿泉人，硕士生，研究方向：应用统计；
杨贵军(1970－)，男，黑龙江哈尔滨人，统计学博士，教授，研究方向：应用统计。

个人信用评分的线性判别模型[11]。熊熊等利用判别分析对商业银行监管和监控指标进行了研究[12]。蔡秋萍建立了北京、上海和江苏三省市上市公司分区域、分行业的 Logistic 财务预警模型[13]。

在上述应用中，判别分析和 Logistic 回归为解决实际问题提供了有价值信息。两种方法的统计理论基础并不完全相同。判别分析基于观测值与两个不同类别之间距离差异进行分类，距离包括欧氏距离和马氏距离等。Logistic 回归采用极大似然估计方法估计模型参数，依据回归函数值对观测数据进行分类。Logistic 回归不仅给出具体的分类算法，还能描述影响分类结果的影响因素。

本文主要研究两种方法的回判正确率。分类方法对全部观测值进行分类，其中分类结果正确的观测点所占比例为回判正确率。对于相同样本，判别分析和 Logistic 回归的回判正确率并不总是一样，有时差异很大。很多应用例子显示，Logistic 回归比判别分析稳健，回判正确率也优于判别分析，但并没有相关理论结果的证明。本文给出的模拟结果显示，在有些情况下，判别分析的回判正确率优于 Logistic回归。BarNiv 和 Hershbarger 的研究显示，当样本渐进正态分布，判别分析优于 Logistic 回归[14]。这增加了方法选择的难度。有些文献通过比较两种方法的回判正确率，最终选择误判率低的方法。在很多应用中，Logistic回归的回判正确率低于判别分析，直接采用 Logistic 回归进行分类。很少有文献对两种方法的误判率进行模拟比较，以及为降低回判正确率而对方法进行改进。

本文针对二分类问题，利用统计模拟方法对 Logistic 回归和判别分析的误判率进行比较。在此基础上，利用判别分析改进 Logistic 回归的误判率。模拟结果显示，新方法的误判率优于 Logistic 回归和判别分析。

二、判别分析与 Logistic 回归的符号和表示

用 R_1 和 R_2 表示两个互不相交的总体，$R_1 \cap R_2 = \varnothing$。记总体 R_1 和 R_2 的均值向量为 μ_1 和 μ_2，协方差阵为 $\sum_1$和 $\sum_2$。$\Omega = R_1 \cup R_2$ 表示两个总体的并集，包含全部的研究对象。二分类问题就是研究如何将样本数据划分为 R_1 和 R_2。

判别分析是最简单最直观的分类方法。利用样本构造判别函数，根据观测点与总体 R_1 和 R_2 的距离，将其归属于距离“最短”的总体。对于判别分析，距离的定义非常重要。在统计理论中，常用的是马氏距离。令 $x_1, x_2 \in R_1$，则 x_1 和 x_2 的马氏距离定义为 $d(x_1, x_2) = \sqrt{(x_1 - x_2)^T \sum_1^{-1} (x_1 - x_2)}$。定义 $x_1 \in \Omega$ 与总体 R_1 的马氏距离为 $d(x_1, R_1) = \sqrt{(x_1 - \mu_1)^T \sum_1^{-1} (x_1 - \mu_1)}$。本文讨论判别分析的距离是指马氏距离。

当 $\sum_1 = \sum_2 = \sum$ 时，对于任意 $x \in \Omega$，根据判别分析的统计思想，有 $R_1 = \{x \mid d(x, R_1) \leqslant d(x, R_2)\}$ 和 $R_2 = \{x \mid d(x, R_1) > d(x, R_2)\}$。于是，判别函数定义为 $w(x) = d^2(x, R_2) - d^2(x, R_1) = 2(x - \mu)^T \sum^{-1} (\mu_1 - \mu_2)$，其中 $\mu = (\mu_1 + \mu_2)/2$。判别准则为

$$R_1 = \{x \mid w(x) \geqslant 0\}$$
$$R_2 = \{x \mid w(x) < 0\} \qquad (1)$$

当 $\sum_1 \neq \sum_2$ 时，式(1) 中的判别函数为

$$w(x) = (x - \mu_2)^T \sum_2^{-1} (x - \mu_2) - (x - \mu_1)^T \sum_1^{-1} (x - \mu_1)。$$

在实际应用中，总体均值和总体协方差阵一般是未知的，需要用样本均值与样本协方差阵来代替。有关判别分析的细节请参考张润楚、王学仁、王松桂、张尧庭和方开泰的相关研究成果[15-17]。

Logistic 回归模型的响应变量 Y 是二分类变量，取值为 1 和 0。常用的 Logistic 回归模型为

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

其中 $X_1, \cdots, X_p$ 为回归模型的解释变量。误差项 ε 的分布与 Y 的分布有关。本文假定 Y 和 ε 都服从伯努利分布。Logistic 回归参数主要使用最大似然估计。令容量为 n 的样本 $Y_1, \cdots, Y_n$，则似然函数为

$$L(Y_1, \cdots, Y_n) = \prod_{i=1}^{n} P(Y_i = 1)^{Y_i} (1 - P(Y_i = 1))^{1-Y_i}$$

对数似然函数为

$$\ln(L(Y_1, \cdots, Y_n)) = \sum_{i=1}^{n} [Y_i \ln(P(Y_i = 1)) + (1 - Y_i)\ln(1 - P(Y_i = 1))] = \sum_{i=1}^{n} (Y_i(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip})) - \sum_{i=1}^{n} \ln(1 + e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}})$$

其中 $X_{i1}, \cdots, X_{ip}$ 是与 Y_i 相对应解释变量的观测值。将对数似然函数分别对 $\beta_0, \beta_1, \cdots, \beta_p$ 求偏导数，并令偏导数为 0，得到似然方程。似然方程的解 $\hat{\beta}_0, \hat{\beta}_1,$

…, $\hat{\beta}_p$ 为回归参数 β_0, β_1, …, β_p 的估计值。通常似然方程是非线性的, 很难得到方程的精确解, 一般采用 Newton－Raphson 迭代法求得近似解。

对于任意 $x=(x_1, \cdots, x_p)\in\Omega$, 计算发生概率 $P(y=1\mid x)=[\exp(\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p)]/[1+\exp(\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p)]$, 选择0.5作为分割点, 则

$$R_1=\{x\mid P(y=1\mid x)\geqslant 0.5\}$$
$$R_2=\{x\mid P(y=1\mid x)<0.5\} \qquad (2)$$

有关 Logistic 回归的细节参见 Agrest 的研究成果[18]。

三、判别分析和 Logistic 回归的模拟比较

为了挑选分类方法, 通常将样本分为两个部分, 训练样本和测试样本。训练样本用于构造分类模型, 测试样本用于检验分类模型的误判率。一般情况下, 分类模型对训练样本的回判正确率低, 对测试样本分类的误判率也低。实际应用中, 有时也将全部样本作为训练样本, 构建模型, 计算模型对训练样本分类的回判正确率, 回判正确率最低的模型是最优的。目前, 统计模拟方法已经被用于很多统计问题的研究中, Yang 和 Yang、Liu 与 Zhang 的研究成果就是其中的一部分[19－20]。本文利用统计模拟方法对判别分析和 Logistic 回归的回判正确率进行评价。

(一) 判别分析和 Logistic 回归比较的模拟方法

为了生成随机数, 定义函数 $f(X_1, \cdots, X_p)=b_0+b_1X_1+\cdots+b_pX_p+\varepsilon$, 其中 b_0, b_1, …, b_p 是函数系数, 已知的, 由标准正态分布随机数代替。ε 服从均值为 0、方差为 σ^2 的正态分布, 记为 $\varepsilon\sim N(0, \sigma^2)$。$X_1$, …, X_p 是标准正态分布所生成的随机数。容量为 n 的样本生成过程如下。首先, 生成 $p+1$ 个标准正态分布的随机数, 依次令 b_0, b_1, …, b_p 等于这 $p+1$ 个数。其次, 生成 n 组正态分布随机数 x_{i1}, …, x_{ip} ($i=1, \cdots, n$), n 个服从 $N(0, \sigma^2)$ 的随机数赋给 ε_1, …, ε_n。计算函数值 $f_i(x_{i1}, \cdots, x_{ip})=b_0+b_1x_{i1}+\cdots+b_px_{ip}+\varepsilon_i$, 并计算函数值的中位数, 记为 Median($f$)。最后, 对于 $i=1, \cdots, n$, 令

$$y_i=\begin{cases}1, & f_i(x_{i1}, \cdots, x_{ip})\geqslant \text{Median}(f)\\ 0, & f_i(x_{i1}, \cdots, x_{ip})< \text{Median}(f)\end{cases}$$

则 y_i, x_{i1}, …, x_{ip} ($i=1, \cdots, n$) 即为所分析的样本数据。当有多个函数值等于 Median(f) 时, 需要继续生成一些随机数, 以保证类别 1 和类别 0 的容量都等于 $n/2$。对于这个样本, 利用判别分析和 Logistic 回归进行分类, 将分类结果与原始数据相比较, 计算回判正确率。对于给定的 p、n、σ^2, 将上述过程重复 200 次, 比较两种方法分类结果, 并分析 p、n、σ^2 的变化对两种方法回判正确率的影响。

(二) 判别分析和 Logistic 回归比较的模拟结果

1. 对于 $\sigma^2=1$, $p=2$, $n=100$ 的模拟结果

当 $\sigma^2=1$, $p=2$, $n=100$ 时, 某个模拟数据见图 1。图 1 的横坐标和纵坐标分别代表变量 X_1 和 X_2。图 1 的实心点属于 $Y=1$ 类, 空心点属于 $Y=0$ 类。图 1 显示随机生成的两类数据的交错很复杂, 没有明显界限。对于 $\sigma^2=1$, 分别增加 p 和 n 的取值, 模拟结果显示两种方法的回判正确率与 p 相关性很高。在固定 p 的条件下, 两种方法的回判正确率与样本量的相关性不大。对于 $\sigma^2=1$ 和 $n=100$, 判别分析与 Logistic 回归的回判正确率平均值比较见图 2。图中横坐标代表 p, 纵坐标表示回判正确率。图 2 显示两种方法的回判正确率随着变量个数 p 的增加而增高, 判别分析的分类结果略优于 Logistic 回归。

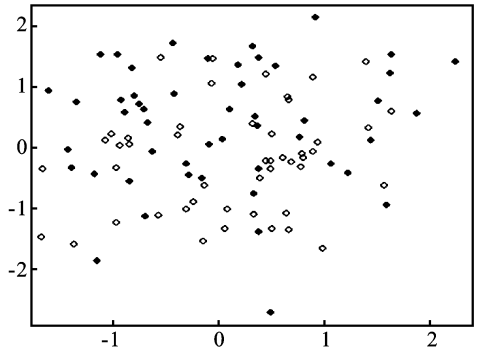

图 1　$\sigma^2=1$、$p=2$、$n=100$ 的散点图

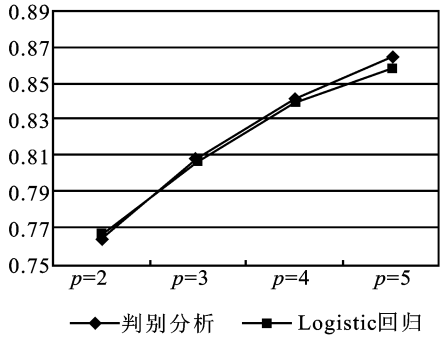

图 2　$\sigma^2=1$、$n=100$ 的 Logistic 回归与判别分析的回判正确率比较图

2. 对于 $\sigma^2=0.25$、$p=2$、$n=100$ 的模拟结果

当 $\sigma^2=0.25$、$p=2$、$n=100$ 时, 某个模拟数据

见图 3。图 3 的横坐标和纵坐标分别代表变量 X_1 和 X_2，实心点属于 $Y=1$ 类，空心点属于 $Y=0$ 类。图 3 好于图 1，随机生成的两类数据的交错复杂程度略低。对于 $\sigma^2=0.25$、$n=100$，判别分析与 Logistic 回归的回判正确率平均值比较见图 4。图 4 的横坐标代表变量个数 p，纵坐标表示回判正确率。图 4 显示两种方法的回判正确率随着 p 的增加而增高，Logistic 回归的分类结果略优于判别分析。对于 $\sigma^2=0.25$，对于 p、n 的不同取值，进行 200 次模拟，计算两种方法的回判正确率，结果见表 1。

表 1　判别分析与 Logistic 回归的 200 次模拟回判正确率比较（$\sigma^2=0.25$）

模拟次数	$p=2$	$p=3$	$p=4$	$p=5$
$n=20$	(45, 79)	(43, 76)		
$n=40$	(45, 76)	(53, 89)	(43, 111)	(40, 121)
$n=60$	(51, 78)	(57, 94)	(58, 95)	(53, 100)
$n=80$	(65, 71)	(65, 89)	(66, 95)	(57, 109)
$n=100$	(62, 87)	(71, 95)	(67, 105)	(60, 101)

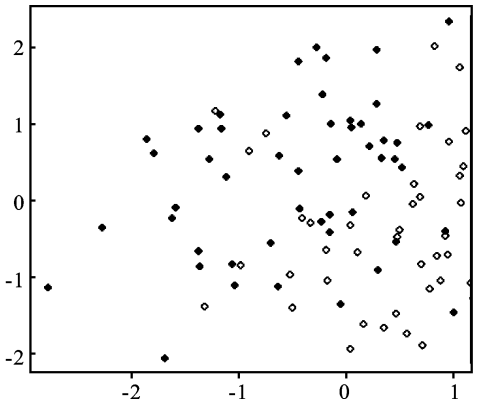

图 3　$\sigma^2=0.25$、$p=2$、$n=100$ 的散点图

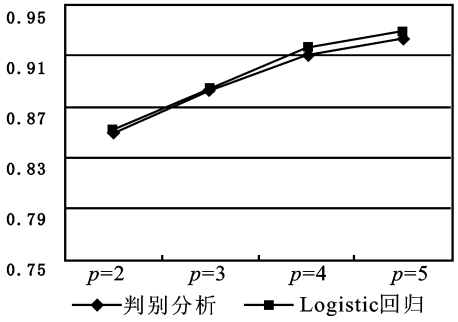

图 4　$\sigma^2=0.25$、$n=100$ 的 Logistic 回归与判别分析回判正确率比较图

在表 1 中，每个二元数组的两个数依次表示在 200 模拟结果中，判别分析优于 Logistic 回归的次数和 Logistic 回归优于判别分析的次数。根据表 1，200 次模拟结果显示 Logistic 回归优于判别分析的次数更多。当变量个数 p 固定时，随着样本量 n 增加，两种方法的回判正确率差异变化不大。对于固定的样本容量 n，随着变量个数 p 增加，两种方法的回判正确率差异变化大，Logistic 回归分类结果明显优于判别分析。

3. 对于 $\sigma^2=0.04$、$p=2$、$n=100$ 的模拟结果

当 $\sigma^2=0.04$、$p=2$、$n=100$ 时，一组模拟数据见图 5。

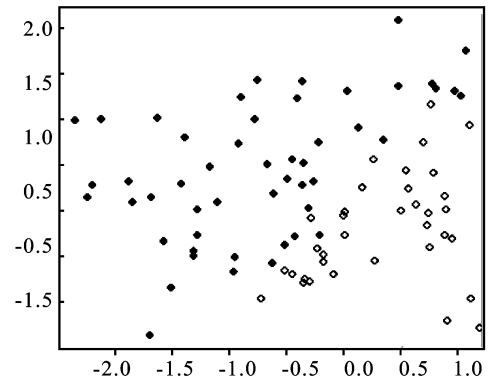

图 5　$\sigma^2=0.04$、$p=2$、$n=100$ 的散点图

图 5 的横坐标和纵坐标分别代表变量 X_1 和 X_2，实心点属于 $Y=1$ 类，空心点属于 $Y=0$ 类。图 5 仅有少数点交错，类别界限较显。对于 $\sigma^2=0.04$、$n=100$，判别分析与 Logistic 回归的回判正确率平均值比较见图 6。图 6 的横坐标代表变量个数 p，纵坐标表示回判正确率。图 6 显示两种方法的回判正确率随 p 增加而增高，但增加量较小，并且 Logistic 回归的分类结果优于判别分析。对于 $\sigma^2=0.04$，对于 p、n 的不同取值，进行 200 次模拟，计算两种方法的回判正确率，结果见表 2。

表 2　判别分析与 Logistic 回归的 200 次模拟回判正确率比较（$\sigma^2=0.04$）

模拟次数	$p=2$	$p=3$	$p=4$	$p=5$
$n=20$	(23, 79)	(9, 95)		
$n=40$	(40, 79)	(26, 110)	(15, 133)	(5, 149)
$n=60$	(42, 70)	(30, 124)	(32, 115)	(8, 161)
$n=80$	(45, 96)	(44, 113)	(35, 127)	(20, 150)
$n=100$	(42, 92)	(29, 129)	(28, 145)	(21, 159)

在表 2 中，每个二元数组的两个数依次表示在 200 模拟结果中，判别分析优于 Logistic 回归的次数和 Logistic 回归优于判别分析的次数。根据表 2，200 次模拟结果显示 Logistic 回归优于判别分析的次数显著增多。当变量个数 p 固定时，随着样本量 n 增加，两种方法的回判正确率差异变化不大。对于固定的样本容量 n，随着变量个数 p 增加，两种方法的回

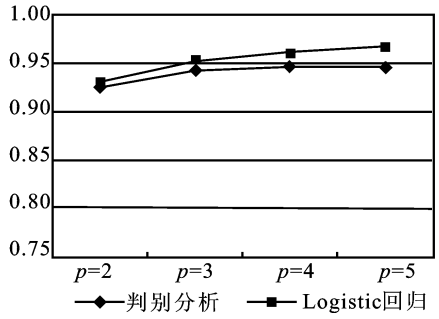

图 6　$\sigma^2 = 0.04$、$n = 100$ 的 Logistic 回归与判别分析回判正确率比较

判正确率差异变化大, Logistic 回归分类结果显著优于判别分析。与表 1 相比, 判别分析的回判正确率优于 Logistic 回归的次数明显减少, 而 Logistic 回归的回判正确率优于判别分析的次数明显增加。

4. 对于 $\sigma^2 = 0.01$、$p = 2$、$n = 100$ 的模拟结果

当 $\sigma^2 = 0.01$、$p = 2$、$n = 100$ 时, 某个模拟数据见图 7。

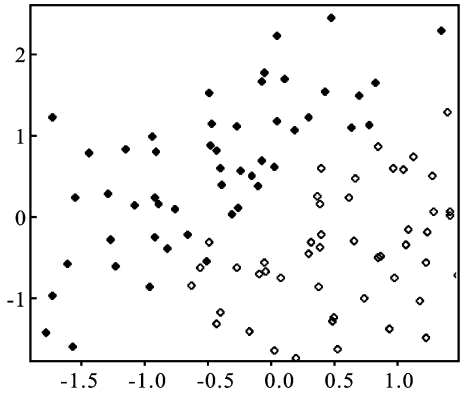

图 7　$\sigma^2 = 0.01$、$p = 2$、$n = 100$ 的散点图

图 7 的横坐标和纵坐标分别代表变量 X_1 和 X_2, 实心点属于 $Y = 1$ 类, 空心点属于 $Y = 0$ 类。图 7 显示, 随机生成的数据点几乎没有交错, 类别界限非常明显。图 8 的横坐标代表变量个数 p, 纵坐标表示回判正确率。图 8 显示两种方法的回判正确率随 p 增加而增高, 但增加量较小, 并且 Logistic 回归的分类结果优于判别分析。对于 $\sigma^2 = 0.01$, 表 3 给出了对于 p、n 的不同取值, 进行 200 次模拟, 两种方法的回判正确率。表 3 的每个二元数组的两个数依次表示在 200 次模拟结果中, 判别分析优于 Logistic 回归的次数和 Logistic 回归优于判别分析的次数。根据表 3, 200 次模拟结果显示 Logistic 回归优于判别分析的次数显著增多。当变量个数 p 固定时, 随着样本量 n 增加, 两种方法的回判正确率差异变化不大。对于固定的样本容量 n, 随着变量个数 p 增加, 两种方法的回判正确率差异变化大, Logistic 回归分类结果显著优于判别分析。

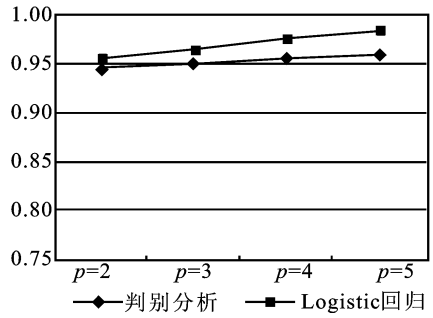

图 8　$\sigma^2 = 0.01$、$n = 100$ 的 Logistic 回归与判别分析回判正确率比较图

综上所述, 对于二分类问题, 判别分析与 Logistic 回归分类结果的差异主要与误差项 σ^2 和变量个数 p 关系密切。当 $\sigma^2 = 1$ 且 $p > 3$ 时, 判别分析优于 Logistic 回归。当 $\sigma^2 = 0.01$, Logistic 回归优于判别分析。当 $\sigma^2 = 0.04$ 或 $\sigma^2 = 0.25$, Logistic 回归略优于判别分析。

表 3　判别分析与 Logistic 回归的 200 次模拟回判正确率比较 ($\sigma^2 = 0.01$)

模拟次数	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$n = 20$	(8, 93)	(2, 110)		
$n = 40$	(28, 98)	(12, 130)	(2, 148)	(1, 147)
$n = 60$	(26, 109)	(17, 139)	(5, 160)	(1, 168)
$n = 80$	(43, 105)	(28, 127)	(13, 165)	(4, 178)
$n = 100$	(37, 133)	(30, 144)	(18, 152)	(8, 179)

四、利用判别分析修正 Logistic 回归的分类

在前面的模拟比较中, 判别分析和 Logistic 回归对某些点分类结果不一致, 有时判别分析能够正确分类而 Logistic 回归却给出错误分类。另外, 很多情况下, Logistic 回归的回判正确率高于判别分析。这里试图利用判别分析修正 Logistic 回归的分类结果, 以进一步提高回判正确率。

根据式(1) 和式(2), 判别分析是依据判别函数 $w(x)$ 进行分类。直观上, 判别函数 $w(x)$ 的绝对值越大, 越容易决定属于哪个分类。Logistic 回归函数是依据概率 $P(y = 1 \mid x)$ 与 0.5 差值的大小进行分类。这个差值绝对值越大, 越可能做出正确分类。基于这样的思想, 利用判别分析修正 Logistic 回归分类结果的过程如下。

1. 记 $\max_w = \max(\mid w(x_1) \mid, \cdots, \mid w(x_n) \mid)$, $\max_p = \max(\mid P(x_1) - 0.5 \mid, \cdots, \mid P(x_n) - 0.5 \mid)$,

其中 $x_i = (x_{i1}, \cdots, x_{ip})$。

2. 确定 Logistic 回归在回判分析中误判的点，共计 l 个，分别记为 $x'_1, \cdots, x'_l$。令 Median(P) 为 $|P(x'_1)-0.5|, \cdots, |P(x'_l)-0.5|$ 的中位数。

3. 对于 $i = 1, \cdots, l$，判断 $|P(x'_i)-0.5| < \text{Median}(P)$ 与 $|w(x'_i)-0| > \delta * \text{Median}(P) * \max_w / \max_P$ 是否同时成立。如果成立，则利用判别分析分类，否则采用 Logistic回归分类。其中 $\delta \in (0, \max_p / \text{Median}(P))$，称为修正权数。$\delta$越大表示判别分析对 Logistic 回归分类结果改进小。

为了比较 Logistic回归与修正 Logistic回归分类的回判正确率，继续使用上述随机模拟过程。每组参数的模拟次数为100次。对于 $\sigma^2 = 1$，由于样本量增大对回判正确率的影响小，取 $n = 60$。根据经验，$\delta = 0.8$ 较好。在 $\sigma^2 = 1$、$n = 60$ 和 $\delta = 0.8$ 的情况下，对于不同变量个数，Logistic 回归与修正 Logistic 回归分类的模拟比较结果见图 9。

图 9 的横坐标为变量个数，纵坐标为回判正确率。在平均意义上，修正 Logistic 回归分类比 Logistic回归略优。对于每次模拟数据，计算两种方法的回判正确率，Logistic 回归优于修正 Logistic 回归的次数和修正 Logistic 回归优于 Logistic 回归的次数，作为二元数组在表 4 中给出。

表 4　Logistic 回归与
修正 Logistic 回归分类的比较表

变量个数	$p=2$	$p=3$	$p=4$	$p=5$
次数	(0, 6)	(1, 5)	(0, 4)	(1, 5)

由表 4，在 100 次随机模拟中，修正 Logistic 回归分类优于 Logistic回归的次数多。

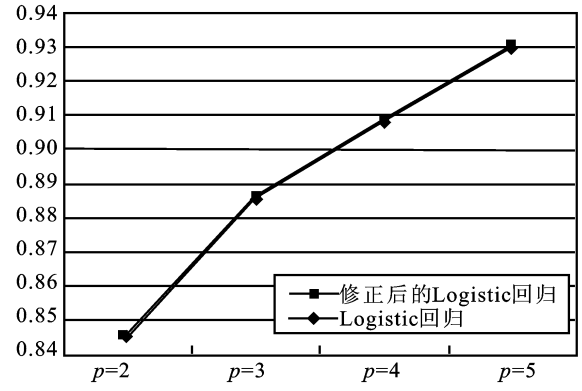

图 9　Logistic 回归与修正 Logistic回归分类的回判正确率比较图

五、结　论

针对二分类问题，本文利用随机模拟方法研究判别分析和 Logistic 回归分类的回判正确率。模拟结果显示，判别分析和 Logistic 回归分类的回判正确率受随机误差大小和变量个数的影响大。一般情况下，Logistic 回归的回判正确率优于判别分析。随着随机误差的增大，Logistic回归与判别分析的回判正确率差异逐渐减小。随机误差超过一定界限，Logistic 回归的回判正确率低于判别分析。在随机模拟的基础上，本文引入了修正 Logistic 回归分类。模拟结果显示修正 Logistic 回归分类略优于 Logistic 回归。

参考文献：

[1] 白玉峰，耿美英，连江宏，罗志昌，张松，杨文鸣. 逐步 Bayes 判别分析在心血管功能评定中的应用[J]. 北京工业大学学报，1994(3)：54－60.

[2] 张晓东，申洪. 肺癌细胞核三维结构的体视学定量研究(Ⅱ)－肺癌诊断分型的判别分析[J]. 中国体视学与图像分析，2006(3)：177－182.

[3] 易尚辉，易银沙，刘桃成，吕媛. 大肠癌预后 logistic回归分析[J]. 中国现代医学杂志，2008(7)：969－970.

[4] 王浩，周岩冰，陈士远. 进展期胃癌淋巴结转移规律的 Logistic 回归分析[J]. 青岛大学医学院学报，2008(10)：414－418.

[5] Christine V Zavgren. Assessing the vulnerability to failure of American industrial firms：a logistic analysis[J]. Journal of Business Finance and Accounting，1985(3)：19－45.

[6] Suk Hun Lee，Hyun Mo Sung，Jorge L Urrutia. The impact of the persian gulf crisis on the prices of LDCs' loans[J]. Journal of Financial Services Research，1996(10)：143－162.

[7] Dimitras，Slowinski，Susmaga，Zopounidis. Business failure prediction using rough sets[J]. European Journal of Operational Research，1999(2)：263－280.

[8] Philip T Ganderton，David S Brookshire，Michael McKee，Steve Stewart，Hale Thurston. Buying insurance for disaster－type risks：experimental evidence[J]. Journal of Risk and Uncertainty，2000(3)：271－289.

[9] 张玲, 陈收, 张昕. 基于多元判别分析和神经网络技术的公司财务困境预警[J]. 系统工程, 2005(11): 50－56.
[10] 张立军, 王瑛, 刘菊红. 基于贝叶斯判别分析的上市公司财务危机预警模型研究[J]. 商业研究, 2009(4): 112－114.
[11] 张成虎, 李育林, 吴鸣. 基于判别分析的个人信用评分模型研究与实证分析[J]. 大连理工大学学报: 社会科学版, 2009(3): 6－9.
[12] 熊熊, 张维, 寇悦. 基于逐步判别分析方法的商业银行监管、监控指标分析[J]. 天津大学学报: 社会科学版, 2007(1): 51－54.
[13] 蔡秋萍. 基于 Logistic 分析的我国上市公司财务预警区域研究[J]. 华东经济管理, 2006(10): 101－104.
[14] BarNiv, Hershbarger. Classifying financial distress in the life insurance industry[J]. Journal of Risk & Insurance, 1990(1): 110－136.
[15] 张润楚. 多元统计分析[M]. 北京: 科学出版社, 2006: 415－520.
[16] 王学仁, 王松桂. 实用多元统计分析[M]. 上海: 上海科学技术出版社, 1990: 117－265.
[17] 张尧庭, 方开泰. 多元统计分析引论[M]. 北京: 科学出版社, 1983: 128－219.
[18] Agresti A. Categorial Data Analysis[M]. 第 2 版. New York: John Wiley & Sons, 2002: 271－376.
[19] Yang Guijun. Genaralized quasi－regression[J]. Communications in statistics, Simulation and Computation, 2008(4): 731－745.
[20] Yang Guijun, Lin Lu, Zhang Runchu. Unbiased quasi－regression[J]. Chinese Annals of Mathematics Series B, 2007(2): 177－186.

An Analogue Comparison of Discriminant Analysis and Logistic Regression

ZHANG Chu-bing[1,2a], GAO Kang[2b], YANG Gui-jun[2b]

(1. School of Management, Tianjin University, Tianjin 300072, China; 2. Tianjin University of Finance and Economics, a. School of Business; b. Center of China Economics and Statistics Research, Tianjin 300222, China)

Abstract: By stochastic simulation, the discriminant accuracy rate of Discriminant analysis and Logistic regression is studied. The result shows that the discriminant accuracy rate of Logistic regression is better than Discriminant analysis. As random error becomes bigger, the differences between the discriminant accuracy rate of two methods gradually reduced. When random error exceeded a certain limit, the discriminant accuracy rate of Logistic regression was worse than Discriminant analysis. Furthermore, the amended Logistic regression is introduced, which showed the amended Logistic regression was slightly superior to Logistic regression.

Key words: discriminant analysis; logistic regression; error rate; discriminant accuracy rate

(责任编辑: 张治国)

《统计与信息论坛》再次入选 CSSCI 来源期刊

经过中文社会科学引文索引指导委员会第八次会议审议通过, 中文社会科学引文索引(CSSCI)来源期刊 2010－2011 年共选出 527 种, 扩展版 173 种,《统计与信息论坛》再次入选来源期刊。在此, 我们对长期以来关心、关爱并提供各种帮助的各界人士、本刊编委、广大作者表示衷心感谢！我们将继续秉承长期以来形成的办刊方针和办刊理念, 坚持学术性、专业性、应用性和可读性的有机统一, 不断扩大作者群和读者群, 把更多的统计学者和相关领域专家吸纳到自己周围, 在统计理论与方法、统计应用研究、统计调查、统计教育等领域, 以统计学科的视角和方法, 紧密结合我国国情, 关注和研究关乎国计民生与社会转型时期重大的经济、社会问题。我们欢迎新老作者积极为本刊赐稿, 同时也乐意倾听提高期刊质量的建议和意见。

在新年到来之际, 特向关爱本刊的各界人士送上新春的问候, 愿老朋友情谊更深, 祝新朋友志趣相投, 祈《统计与信息论坛》这块平台为我国的统计事业和学术发展留下更多的学术佳作和学术佳话。

本刊编辑部