Pure Mathematics 理论数学, 2018, 8(6), 604-612
Published Online November 2018 in Hans. http://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2018.86081

Hans汉斯

The Discriminant Analysis and Logistic Regression Analysis of SMEs Bankruptcy Model

Yalei Pei

Department of Mathematics, Taiyuan Normal University, Taiyuan Shanxi
Email: peileelee@163.com

Received: Oct. 15th, 2018; accepted: Oct. 27th, 2018; published: Nov. 8th, 2018

Abstract

In multivariate statistical analysis, discriminant analysis and Logistic regression analysis are both used to predict and interpret the classification. Regression models are used to predict and interpret metric variables, while discriminant analysis and Logistic regression analysis are used to solve situations where explanatory variables are non-metric variables. When the explanatory variable contains two types, both discriminant analysis and Logistic regression analysis are applicable; when the explanatory variable contains more than two types, only discriminant analysis is applicable. However, discriminant analysis is only applicable when the explanatory variables satisfy the multivariate normality and the equivalent covariance matrix hypothesis. Logistic regression does not require a series of assumptions about explanatory variables, and good results can still be obtained. In this paper, the bankruptcy model of SMEs (Small and Medium Enterprises) was analyzed by discriminant analysis and Logistic regression analysis respectively, and the differences and similarities between the two classification methods were compared.

Keywords

Discriminant Analysis, Fisher Discriminant, Logistic Regression, Maximum Likelihood Estimation

中小企业破产模型的判别分析与Logistic回归分析

裴亚蕾

太原师范学院数学系，山西 太原
Email: peileelee@163.com

文章引用: 裴亚蕾. 中小企业破产模型的判别分析与 Logistic 回归分析[J]. 理论数学, 2018, 8(6): 604-612.
DOI: 10.12677/pm.2018.86081

收稿日期：2018年10月15日；录用日期：2018年10月27日；发布日期：2018年11月8日

摘　要

多元统计分析中，判别分析和Logistic回归分析都是用来预测和解释一个对象所属类别的分类方法。回归模型用于预测和解释度量变量，而判别分析和Logistic回归分析用来解决被解释变量是非度量变量的情况。被解释变量包含两类时，判别分析和Logistic回归分析都适用；而被解释变量包含两类以上时，只有判别分析适用。但是，只有解释变量满足多元正态性和相等协方差阵假设时，判别分析才适用。而Logistic回归不需要解释变量的一系列的假设，仍可以得到良好的结果。本文分别用判别分析和Logistic回归分析对中小企业的破产模型进行分析，并对比两种分类方法的异同。

关键词

判别分析，Fisher判别，Logistic回归，极大似然估计

1. 引言

在现实生活中，人们可能会面临这样一类问题，判断某一对象属于哪个类别。比如一个公司是不是潜在客户，一个广告方案能否成功。这就需要选择对象所包含的变量作为解释变量，根据一定的判别准则，定义对象与类别之间的“距离”，所观察对象与哪个类别“距离”近，就说明该对象属于哪个类别。常用的判别方法有判别分析和 Logistic 回归分析，两者既有区别又有联系，本文通过一个实例来比较两种判别方法之间的异同。

2. 判别分析的基本思想

回归模型用来预测和解释度量变量，在回归模型中，解释变量和被解释变量都是度量变量，也就是定量变量。而实际生活中，往往面对被解释变量是非度量变量，也就是定性变量，在这种情况下，传统的回归分析是失效的。而判别分析正是用来解决被解释变量是非度量变量的情形。

判别分析的步骤：

1) 选择恰当的解释变量，解释变量不是越多越好，要“越有代表性越好”，解释变量多了会出现多重共线性的结果，影响判别分析方法的使用。

2) 判断解释变量是否满足判别分析的假设条件：

i. 每个解释变量不能是其他解释变量的线性组合；

ii. 各组变量的协方差阵相等；

iii. 各个解释变量间具有多元正态分布[1]。

3) 选择合适的准则判别对象和类别之间的距离。常用的判别方法有：

i. 距离判别；

ii. Bayes 判别；

iii. Fisher 判别；

iv. 逐步判别。

4) 对模型的拟合效果进行显著性检验。

5) 对模型的结果进行分析和解释。

3. Logistic 回归的基本思想

当被解释变量只有两组时，Logistic 回归也可以用于预测和分类[1]。而且不需要解释变量满足多元正态性和相等协方差阵假设，Logistic 回归就可以得到良好的结果。

当研究某一随机现象发生的概率 p 的大小，比如一个公司是不是潜在客户，一个广告方案能否成功，以及讨论 p 与哪些因素有关。但是因为概率 p 的取值是 0 到 1 之间的实数，变化范围非常小，这就意味着，当 p 作为被解释变量的时候，可能对解释变量的变化不够敏感，也就造成了判别方法的失效，所以直接对概率 p 进行数学上的处理有一定的难度。为了数学上处理的方便，我们构造 p 的一个严格单调函数 $Q=Q(p)$。$Q(p)$ 与 p 同增同减，$Q(p)$ 在 $p=0$ 或者 $p=1$ 的附近微小变化要很敏感，因 $p=0$ 或者 $p=1$ 的极端情形，往往正是研究者所关心的问题。也就要求 $\frac{\mathrm{d}Q}{\mathrm{d}p}$ 应与 $\frac{1}{p(1-p)}$ 成比例，于是令

$$Q=\ln\frac{p}{1-p}$$

上述变换称为 Logit 变换[3]。

然后，我们可以将 Q 看作新的被解释变量，然后构造 Q 和解释变量的函数关系，并从中解出 p 值。比如 $Q=bx$，则 $p=\frac{\mathrm{e}^{bx}}{1+\mathrm{e}^{bx}}$。

当比例只取 0 和 1 两个值时，被解释变量 y 取 1 的概率 $p(y=1)$ 就是要研究的对象。将影响被解释变量 y 的解释变量，记为 $x_1,x_2,\cdots,x_n$，这些 $x_i\,(i=1,2,\cdots,n)$ 中既可以包含定性变量，也可以包含定量变量。因为下式成立

$$\ln\frac{p}{1-p}=b_0+b_1x_1+\cdots+b_nx_n$$

所以 $\ln\frac{Ey}{1-Ey}$ 是 $x_1,x_2,\cdots,x_n$ 的线性函数，满足上面条件的称为 Logistic 线性回归[2]。

Logistic 回归的步骤：

1) 选择恰当的解释变量和被解释变量[2]。

2) 令 $Q=\ln\frac{p}{1-p}=b_0+b_1x_1+\cdots+b_nx_n$。

3) 使用极大似然估计 $b_0,b_1,\cdots,b_n$。

4) 解出 p 值。

5) 对模型的拟合效果进行显著性检验。

6) 对模型的结果进行分析和解释。

4. 数据背景

为了比较判别分析与 Logistic 回归的异同，以中小企业的破产模型为例，收集 21 个破产企业和 25 个财务良好的企业破产前两年的年度财务数据。将财务数据作为解释变量，检验这些解释变量对企业是否破产有怎样的影响。

财务数据涉及四个解释变量：

x_1 = 现金流量/总债务；

x_2 = 净收入/总资产；

x_3 = 流动资产/流动债务；

x_4 = 流动资产/净销售额[2]。

5. 判别分析

Table 1. Ogarithmic determinant
表 1. 对数行列式

分类	秩	对数行列式
1.00	2	−5.888
2.00	2	−6.027
汇聚的组内	2	−5.001

Table 2. Test results
表 2. 检验结果

Box’M		42.330
F	近似	13.407
	df1	3
	df2	6935351.351
	Sig.	0.000

上面两张表(表 1，表 2)是关于解释变量的协方差是否相等的 Box’M 检验。根据进行判别分析所需的假设条件，只有解释变量协方差相等，判别分析才是适用的，判别分析的结果才是可靠的。上表显示解释变量通过检验。

Wilks’s Lambda 准则用来评估判别函数的判别效力的显著性。Spss 默认引入变量的临界值为 3.87，剔除变量的临界值为 2.71。

第一步：表 3 步骤 0 中表明 x_3 (流动资产/流动债务)的 F 值最大，为 26.610，大于引入变量的临界值 3.87，Wilks’s Lambda 最小，为 0.632，x_3 (流动资产/流动债务)第一个进入模型，这在表 4 中反映出来[1]。

第二步：表 3 步骤 1 中，在 x_3 (流动资产/流动债务)进入模型后，模型外的三个变量中 x_2 (净收入/总资产)的 F 值最大，为 7.446，大于 3.87，Wilks’s Lambda = 0.531 最小，因此第二个进入模型的是 x_2 (净收入/总资产) [2]。表 4 步骤 2 中，x_2 (净收入/总资产)，x_3 (流动资产/流动债务)的 F 值都大于 2.71，因此，x_2(净收入/总资产)，x_3 (流动资产/流动债务)都进入模型。

第三步：表 3 步骤 2 中，x_1 (现金流量/总债务)，x_4 (流动资产/净销售额)的 F 值都小于 3.87，分别为 0.403，1.163，不能进入模型。

判别分析的自变量选择结束，x_1 (现金流量/总债务)，x_4 (流动资产/净销售额)对判别函数的贡献不显著，其他两个自变量进入判别方程。

Table 3. Variables not in the analysis
表 3. 不在分析中的变量

步骤		容差	最小容差	要输入的 F	Wilks'Lambda
0	x_2	1.000	1.000	19.765	0.690
	x_1	1.000	1.000	23.106	0.656
	x_4	1.000	1.000	0.039	0.999
	x_3	1.000	1.000	26.610	0.623
1	x_2	0.961	0.961	7.446	0.531
	x_1	0.892	0.892	6.602	0.540
	x_4	0.952	0.952	1.121	0.607
2	x_1	0.345	0.345	0.403	0.526
	x_4	0.950	0.920	1.163	0.517

Table 4. Variables in the analysis
表 4. 分析中的变量

步骤		容差	要删除的 F	Wilks'Lambda
1	x_3	1.000	26.610	
2	x_3	0.961	12.861	0.690
	x_2	0.961	7.446	0.623

Table 5. Classification function coefficients
表 5. 分类函数系数

	分类	
	1.00	2.00
x_2	−11.168	−1.028
x_3	2.402	4.043
(常量)	−2.789	−5.908

Table 6. Canonical discriminant function coefficient
表 6. 典型判别式函数系数

	Fuction
	1
x_2	5.497
x_3	0.890
(常量)	−1.771

a) 由表 5 可以看出两类的 Fisher 判别函数分别是

$$f_1 = -2.789 + 2.402x_2 - 11.169x_3$$

$$f_2 = -5.908 + 4.043x_2 - 1.028x_3$$

b) 由表 6 可以看出非标准化的判别函数为

$$f(x) = -1.771 + 5.497x_2 + 0.89x_3$$

c) 根据 Fisher 线性判别函数对原始数据进行回判，根据非标准的线性判别函数计算每个观测的 Z 得分。由表 7 判别函数在 $y = 1$ 的重心为−1.003，而在 $y = 2$ 的重心为 0.842。计算分割点为 0，可以根据待判样品的每个观测的 Z 得分进行分类[1]。

Table 7. Functions at group centroids
表 7. 组质心处的函数

分类	fuction
	1
1.00	−1.003
2.00	0.842

Table 8. Wilks' Lambda
表 8. Wilks 的 Lambda 检验

步骤	输入的	Wilks'Lambda							
		统计量	df1	df2	df3	精确F			
						统计量	df1	df2	Sig.
1	x_3	0.623	1	1	44.000	26.610	1	44.000	0.000
2	x_2	0.531	2	1	44.000	18.977	2	43.000	0.000

表 8 是对两个判别函数的 Wilks' Lambda 检验，说明判别函数在 0.05 的显著性水平上是显著的，模型拟合比较好[1]。

通过逐步分析法，x_2 (净收入/总资产)，x_3 (流动资产/流动债务)贡献比较大的保留下来。另外两个变量 x_1 (现金流量/总债务)、x_4 (流动资产/净销售额)对因变量影响较小而被剔除。表 9 和表 10 判别载荷和标准判别函数证实了这一点[2]。

Table 9. Structure matrix
表 9. 结构矩阵

	函数
	1
x_3	0.828
x_2	0.713
x_1[a]	0.687
x_4[a]	0.205

Table 10. Canonical discriminant function coefficient
表 10. 标准化的典型判别式函数系数

	函数
	1
x_2	0.572
x_3	0.715

6. Logistic 回归分析

表 11、表 12 是对整个模型的拟合效果的检验，表中的结果表明模型是非常显著的，拟合效果良好，可以用来做解释和预测。

Table 11. Omnibus test of model coefficients
表 11. 模型系数的综合检验

		卡方	df	Sig.
步骤 1	步骤	35.978	4	0.000
	块	35.978	4	0.000
	模型	35.978	4	0.000

Table 12. Model summary
表 12. 模型汇总

步骤	−2 对数似然值	Cox & Snell R 方	Nagelkerke R 方
1	27.443[a]	0.543	0.725

Table 13. Hosmer and Lemeshow test
表 13. Hosmer 和 Lemeshow 检验

步骤	卡方	df	Sig.
1	7.103	7	0.418

表 13 是 Hosmer-Lemeshow 检验，检验因变量实际值与预测值的分布是否有显著的差异，结果表明不显著，也就是说因变量实际值与预测值的分布没有显著差异，模型拟合较好[1]。

Table 14. Variables in the function
表 14. 方程中的变量

		B	S.E,	Wald	Df	Sig.	Exp (B)	EXP(B)的95% C.I.	
								下限	上限
步骤 1[a]	x_1	7.138	6.002	1.414	1	0.234	1258.662	0.010	1.617E8
	x_2	−3.703	13.670	0.073	1	0.786	0.025	0.000	1.066E10
	x_3	3.415	1.204	8.049	1	0.005	30.412	2.874	321.822
	x_4	−2.968	3.065	0.938	1	0.333	0.051	0.000	20.897
	常量	−5.320	2.366	5.053	1	0.025	0.005		

表 14 中输出了全部自变量的系数和各变量的相关统计量，Sig 是 Wald 检验的显著性概率。可以看到因素 x_2 (净收入/总资产)的系数的 Wald 检验在显著性水平 0.05 上仍然不显著，将其剔除。用 y 对 x_1 (现金流量/总债务)、x_3 (流动资产/流动债务)、x_4 (流动资产/净销售额)三个自变量做回归，输出结果见表 15。

Table 15. Variables in the function
表 15. 方程中的变量

		B	S.E,	Wald	Df	Sig.	Exp (B)	EXP(B)的95% C.I.	
								下限	上限
步骤 1[a]	x_1	5.772	3.005	3.690	1	0.055	321.324	0.889	116121.646
	x_3	3.289	1.085	9.183	1	0.002	26.810	3.195	224.952
	x_4	−2.979	3.025	0.970	1	0.325	0.051	0.000	19.089
	常量	−5.038	2.060	5.983	1	0.014	0.006		

从表 15 中得到结论，自变量 x_4 (流动资产/净销售额)的系数的 Wald 检验在显著性水平 0.05 上仍然不显著，将其剔除，再用 y 对 x_1 (现金流量/总债务)和 x_3 (流动资产/流动债务)做回归。

Table 16. Variables in the function
表 16. 方程中的变量

		B	S.E,	Wald	Df	Sig.	Exp (B)	EXP(B)的95% C.I.	
								下限	上限
步骤 1[a]	x_1	6.556	2.905	5.092	1	0.024	703.744	2.367	209200.474
	x_3	3.019	1.002	9.077	1	0.003	20.473	2.872	145.937
	常量	−5.940	1.986	8.950	1	0.003	0.003		

从表 16 中可以得到下面模型：

$$\frac{\hat{p}}{1-\hat{p}} = \mathrm{e}^{-5.940+6.556x_1+3.016x_3}$$

Table 17. Classification table
表 17. 分类表 [a]

已观测			已预测		
			分类		百分比校正
			1.00	2.00	
步骤1	分类	1.00	18	3	85.7
		2.00	1	24	96.0
	总计百分比				91.3

a.切割值为 0.500。

由表 17 可以看出，组 1 的正确判断率为 85.7%，组 2 的正确判断率为 96%，总的正确判断率为 91.3%。Logistic 回归方程判别效果良好。

7. 判别分析与 Logistic 回归分析的比较

本例中， Logistic 回归的判别效果比判别分析好。

从解释变量的贡献程度来看，Logistic 回归分析的结论是 x_1 (现金流量/总债务)和 x_3 (流动资产/流动债务)贡献较大，而 x_3 (流动资产/流动债务)的贡献最大(wald 值最大)。判别分析的结论是 x_2 (净收入/总资产)

和 x_3 (流动资产/流动债务)贡献较大，而 x_3 (流动资产/流动债务)贡献最大(载荷因子最大)。两种分类方法中，解释变量 x_3 (流动资产/流动债务)贡献最大，解释变量 x_4 (流动资产/净销售额)都被剔出了。两种方法有一致性。

8. 结论

在所有参加 Logistic 回归分析的 4 个因素中，x_2 (净收入/总资产)首先被剔除，其次 x_4 (流动资产/净销售额)被剔除，说明它们对中小企业是否破产影响不大。我们把两个重要指标 x_1 (现金流量/总债务)和 x_3 (流动资产/流动债务)引入模型，事实上 x_1 (现金流量/总债务)和 x_3 (流动资产/流动债务)可能存在共线性，其中某一个因素的引入可能会影响另一个因素进入方程，在判别分析中 x_1 (现金流量/总债务)就被剔除。

在参与判别分析的四个因素中，x_2 (净收入/总资产)和 x_3 (流动资产/流动债务)被保留下来，而且由表 9 和表 10 可知，x_3 (流动资产/流动债务)的影响强于 x_2 (净收入/总资产)，也与 Logistic 回归的结论一致性。

参考文献

[1] 郭蕾. 2 型糖尿病的判别分析和 Logistic 回归分析[D]: [硕士学位论文]. 长沙: 中南大学, 2007.

[2] 何晓群. 多元统计分析[M]. 第 4 版. 北京: 中国人民大学出版社, 2014: 105-305.

[3] 郭志刚. 社会统计分析方法: SPSS 软件应用[M]. 第 2 版. 北京: 中国人民大学出版社, 2015: 177-306.