广西科学院学报
Journal of Guangxi Academy of Sciences

2011,27(3):190~192
Vol. 27, No. 3 August 2011

基于二元选择模型的蠓虫分类方法
The Classification Method of Midges based on Binary Choice Model

冯 烽[1,3],黄 晗[1],韦 范[2],缪剑华[2]
FENG Feng[1,3], HUANG Han[1], WEI Fan[2], Miao Jian-hua[2]

(1.广西财经学院数学与统计系,广西南宁 530003;2.广西壮族自治区药用植物园,广西南宁 530023;3.福州大学管理学院,福建福州 350002)
(1. Department of Mathematics & Statistics, Guangxi University of Finance & Economics, Nanning, Guangxi, 530003, China; 2. Guangxi Botanical Garden of Medicinal Plant, Nanning, Guangxi, 530023, China; 3. School of Management, Fuzhou University, Fuzhou, Fujian, 350002, China)

摘要:利用 Probit、Logit 和 Extreme 三种二元选择模型研究蠓虫分类问题,并用极大似然法进行参数估计,所得方法具有较高的拟合优度和较好的预测效果,可以用于有效鉴别两类蠓虫。
关键词:模式识别 蠓虫分类 二元选择模型
中图法分类号:O29 **文献标识码**:A **文章编号**:1002-7378(2011)03-0190-03
Abstract: The classification problem of midges is studied by using three kinds of binary choice models: Probit, Logit and Extreme Model, and the Method of Maximum Likelihood is used for the parameters estimation. The results show that the binary choice models have high goodness of fitting and good prediction accuracy, two kinds of midges can be classified by means the binary choice models.
Key words: pattern recognition, midges classification, binary choice model

生物学家 Grogan 和 Wirth 曾试图就两种蠓虫 Af 和 Apf 的鉴别问题进行研究,Af 是宝贵的传粉益虫,Apf 则是某种疾病的载体毒蠓,希望建立一种正确区分两种蠓虫的模型[1]。尽管这只是一道美国大学生数学建模竞赛赛题,然而这个蠓虫分类问题对于正确识别蠓虫的种类,进而为最大限度的消灭毒蠓 Apf 以及保护益蠓 Af 颇具实际意义。现在有对蠓虫分类问题的研究并不多何水明[2]为克服标准 BP 算法中存在的网络学习收敛速度慢、容易陷入局部极小等问题,引入同伦算法,提出了一种将同伦与快速 BP 算法结合的改进算法来进行蠓虫分类;冯增哲等[3]使用支持向量机算法建模了蠓虫分类问题,即利用极大化"间隔"的思想,将分类问题转化为一个二次规划及其对偶规划问题进行求解;王琪[4]使用模糊聚类获得模糊模式,进而通过贴近度模式识别判断待识别对象的归类情况。

对于蠓虫分类的问题的现有文献基本上是运用数学模型的方法进行研究,这些方法主要有神经网络分析方法、支持向量机方法、模糊模式识别方法等,这些方法的优点是通常可以获得较高的精度要求,然而上述分类方法都存在着需要编程实现及运算时间较长的不便,特别是无法获得个体的指标及其所属类别之间的解析关系式和相应的判定概率。由于二元选择模型(Binary choice model)作为一类重要的计量经济模型具有可刻画变量间的解析关系且易于求解的优点[5,6]。为此,本文拟使用二元选择模型对蠓虫分类问题进行研究。

1 蠓虫分类的二元选择模型

蠓虫分类的二元选择模型的设定如下:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i = x_i^T \beta + u_i \quad (1)$$

收稿日期:2011-05-04
修回日期:2011-05-28
作者简介:冯 烽(1980-),男,博士研究生,讲师,主要从事建模分析、金融计量的研究。

其中，$\beta=(\beta_0,\beta_1,\beta_2)^T$ 为参数向量；$x_i=(1,x_{1i},x_{2i})^T$，x_{1i} 为蠓虫 i 的触角长度；x_{2i} 为蠓虫 i 的翅膀长度；u_i 是均值为 0 且相互独立的随机扰动项；被解释变量 y_i 的取值表示蠓虫 i 所属的类型，$y_i=\begin{cases}1,\text{蠓虫 } i \text{ 为 Af}\\0,\text{蠓虫 } i \text{ 为 Apf}\end{cases}$。可以证明式(1)所定义的二元选择模型中的残差项一定是异方差的[5]。因此，不能采用普通最小二乘法对二元选择模型进行参数估计。

引入潜在变量 y_i^*，它与解释变量 x_i 之间具有线性关系：

$$y_i^*=x_i^T\beta+u_i^*, \tag{2}$$

其中，u_i^* 为随机扰动项，其密度和分布函数分别记为 $f(u)$ 和 $F(u)$。y_i 和 y_i^* 关系如下：

$$y_i=\begin{cases}1,y_i^*\geqslant 0,\\0,y_i^*<0。\end{cases} \tag{3}$$

式(3)中把临界值选为 0，但事实上只要式(2)中包含有常数项，则临界值的选取不影响模型的结果，故不妨设为 0。这样，可得 y_i 条件概率和条件均值：

$$P(y_i=1\mid x_i)=P(y_i^*\geqslant 0)=P(u_i^*\geqslant -x_i^T\beta)=1-F(-x_i^T\beta),$$

$$P(y_i=0\mid x_i)=P(y_i^*<0)=P(u_i^*<-x_i^T\beta)=F(-x_i^T\beta),$$

$$E(y_i\mid x_i)=1-F(-x_i^T\beta)。$$

于是，可以得到 y_i 关于它的条件均值的回归模型：

$$y_i=1-F(-x_i^T\beta)+u_i。 \tag{4}$$

式(2)中 u_i^* 的分布函数类型决定了二元选择模型的类型。常用的三种二元选择模型如表 1 所示。

表 1　二元选择模型的类型

u_i^* 的分布	分布函数 F	对应的二元选择模型
标准正态分布	$\Phi(x)$	Probit 模型
逻辑分布	$e^x/(1+e^x)$	Logit 模型
极值分布	$1-\exp(-e^x)$	Extreme 模型

2　模型的参数估计

采用极大似然法进行模型的参数估计。以样本出现的联合概率作为似然函数：

$$L=\prod_{y_i=0}P(y_i=0)\prod_{y_i=1}P(y_i=1)=\prod_{i=1}^{N}[F(x_i^T\beta)]^{y_i}[1-F(x_i^T\beta)]^{1-y_i}。 \tag{5}$$

对式(5)取对数获得对数似然函数式(6)，并使其最大化可得参数的极大似然估计值。

$$\ln L=\sum_{i=1}^{N}[y_i\ln F(x_i^T\beta)+(1-y_i)\ln(1-F(x_i^T\beta))]。 \tag{6}$$

由于 y_i 的取值只能是离散的两个值，因此与线性回归模型所不同的是，二元选择模型中的系数不能看作对被解释变量的边际影响。此时，可以把边际效应考虑为：

$$\frac{\partial E(y_i\mid x_i)}{\partial x_{ij}}=\frac{\partial[1-F(-x_i^T\beta)]}{\partial x_{ij}}=f(-x_i^T\beta)\beta_j。 \tag{7}$$

式(7)表明，尽管系数的大小不再是边际影响，但是我们还是可以利用系数的符号进行一种定性的分析，即如果系数为正，则被解释变量 y_i 取 1 的概率越大；反之，如果系数为负，则相应的概率越小。

3　模型的应用

设样本中 9 只 Af 对应于模型中 y_i 取值为 1，$i=1,\cdots,9$；6 只 Apf 对应于模型中 y_i 取值为 0，$i=10,\cdots,15$。将样本数据(表 2)代入模型中，并用 Eviews6.0 进行求解的结果如表 3 所示。

表 2　蠓虫样本集

样本	触角长度	翅膀长度	分类	样本	触角长度	翅膀长度	分类
1	1.24	1.27	Af	10	1.14	1.78	Apf
2	1.36	1.74	Af	11	1.18	1.96	Apf
3	1.38	1.64	Af	12	1.2	1.86	Apf
4	1.38	1.82	Af	13	1.26	2	Apf
5	1.38	1.9	Af	14	1.28	2	Apf
6	1.40	1.7	Af	15	1.3	1.96	Apf
7	1.48	1.82	Af	16	1.24	1.8	待定
8	1.54	1.82	Af	17	1.28	1.84	待定
9	1.56	2.08	Af	18	1.4	2.04	待定

表 3　蠓虫分类的二元选择模型估计结果

模型	β_0	β_1	β_2	McFadden R-squared	Prob (LR statistic)
Probit	−200.423	479.4239	−228.788	1.000000	0.000041
Logit	−389.241	872.8426	−404.485	1.000000	0.000041
Extreme	−250.045	620.482	−293.048	1.000000	0.000041

表 3 的估计结果显示，三种模型下的 McFadden R-squared 值均为 1 表明了这三种模型样本内的拟合效果非常好，因为解释变量 可完全解释被解释变量 的变化。同时，三种模型估计得到的系数符号是完全一致的，均表明触角越长的蠓虫越可能是 Af，翅膀越长的蠓虫是 Af 的可能性越小，这与从样本数据(表 2)中获得的直观认识是一致的。

从表 4 三种模型下的蠓虫样本内的拟合值和样本外的预测结果可以看出，三种方法在样本内(即样本 1～15)的拟合结果与实际观察值完全一致，均体现了很好的拟合优度。但是这三种模型在进行样本外(即样本 16～18)的预测时的结果并不完全一致。

Probit模型及Logit模型均把蠓虫个体16、17判定为Apf，而把蠓虫个体18判定为Af；Extreme模型则是把蠓虫个体16判定为Apf，而把蠓虫个体17、18判定为Af。也就是说，对于蠓虫个体17的判定三种模型并不完全一致，此时，如果认为将Apf错判为Af的危害要比把Af误判为Apf的大，则建议采用Probit模型及Logit模型且判定为Apf；否则，采用Extreme模型而且判定为Af。

表4 蠓虫分类的二元选择模型的拟合结果和预测结果

模型样本	Probit模型		Logit模型		Extreme模型	
	拟合值	拟合概率	拟合值	拟合概率	拟合值	拟合概率
1	103.50	1.00	179.39	1.00	147.18	1.00
2	53.50	1.00	94.02	1.00	83.91	1.00
3	85.97	1.00	151.93	1.00	125.62	1.00
4	44.79	1.00	79.12	1.00	72.87	1.00
5	26.49	1.00	46.76	1.00	49.43	1.00
6	81.83	1.00	145.11	1.00	120.45	1.00
7	92.73	1.00	166.40	1.00	134.92	1.00
8	121.50	1.00	218.77	1.00	172.15	1.00
9	71.60	1.00	131.07	1.00	108.37	1.00
10	−61.12	0.00	−114.18	0.00	−64.32	0.00
11	−83.13	0.00	−152.08	0.00	−92.25	0.00
12	−50.66	0.00	−94.17	0.00	−50.53	0.00
13	−53.92	0.00	−98.43	0.00	−54.33	0.00
14	−44.34	0.00	−80.97	0.00	−41.92	0.00
15	−25.60	0.00	−47.34	0.00	−17.79	0.00
16	−17.75	0.00	−34.99	6.66×10^{-16}	−8.13	0.00
17	−7.73	5.44×10^{-15}	−16.25	8.72×10^{-8}	4.96	0.99
18	4.04	0.99	7.59	0.99	20.81	1

3 结束语

采用二元选择模型对蠓虫分类问题进行研究具有极高的拟合优度及较好的估计效果。这种分类方法具有直观的解析意义及便于软件实现的优点，可应用于疾病的诊断、投资的决策分析、项目的评估等模式识别问题的研究。

参考文献：

[1] Grogan W L, Wirth W W. A new American genus of predaceous midges related to palpomyia and bezzia (Diptera: Ceratopogonidae)[J]. Proceedings of the Biological Society of Washington, 1981, 94: 1279-1305.

[2] 何水明. 采用同伦BP算法进行蠓虫分类[J]. 湖北工业大学学报, 2007, 22(1): 65-66.

[3] 冯增哲，王清，王昌元，等. 一种基于支持向量机的蠓虫分类方法[J]. 中国科技信息, 2007(4): 207-268.

[4] 王琪. 基于模糊模式识别的蠓虫分类数学模型[J]. 咸宁学院学报, 2010, 30(12): 59-60.

[5] 李子奈，叶阿忠. 高等计量经济学[M]. 北京：清华大学出版社, 2000: 155-163.

[6] 高铁梅. 计量经济分析方法与建模——Eviews应用及实例[M]. 北京：清华大学出版社, 2009: 219-242.

（责任编辑：邓大玉）

（上接第183页）

量为零，从而系统(6)$|_{p=1,q=2}$在原点可积的充分必要条件是定理4中的3个条件之一成立. 由于变换是同胚的，即系统(2)$|_{p=1,q=2}$在原点可积的充分必要条件是定理4中的3个条件之一成立.

证明 如果条件(Ⅰ)成立，系统(2)$|_{p=1,q=2}$有积分因子 $J=u^{-4}v^{-3}\left(1+\frac{1}{2}v^2b_{20}\right)^{-\frac{1}{2}-\frac{a_{02}}{b_{20}}}$. 如果条件(Ⅱ)成立，系统(2)$|_{p=1,q=2}$有积分因子 $J=u^{-6}v^{-4}$. 如果条件(Ⅲ)成立，系统(2)$|_{p=1,q=2}$有积分因子 $J=u^{-\frac{2(4a_{20}-b_{02})}{2a_{20}-b_{02}}}v^{-\frac{2(3a_{20}-b_{02})}{2a_{20}-b_{02}}}$. 证明完毕.

参考文献：

[1] 肖萍. 复平面多项式共振微分系统的奇点量与可积性条件[D]. 长沙：中南大学博士学位论文, 2005.

[2] Jaume Giné, Valery G Romanovski. Integrability conditions for Lotka-Volterra planar complex quintic systems[J]. Nonlinear Analysis: Real World Applications, 2010, 11(3): 2100-2105.

[3] 刘一戎. 高次奇点与无穷远点的中心焦点理论[J]. 中国科学, 2001(A31): 37-48.

[4] Wu Yusen, Zhang Gui, Li Peiluan. Isochronicity problem of higher-order singular point for polynomial differential systems [J]. Acta Appl Math, 2010, 110: 1429-1448.

[5] 黄文韬. 微分自治系统的几类极限环分支与等时中心问题[D]. 长沙：中南大学博士学位论文, 2004.

[6] Zhang Qi, Gui Weihua, Liu Yirong. The generalized center problem of degenerate resonant singular point [J]. Applied Mathematics and Computation, 2009, 215: 1507-1512.

[7] Zhang Qi, Liu Yirong, Gui Weihua. Generalized singular point quantity and integrability of degenerate resonant singular point[J]. Bull Sci math, 2009, 133: 198-204.

[8] 刘一戎，李继彬. 论复自治系统的奇点量[J]. 中国科学, 1989(A3): 245-255.

（责任编辑：尹 闯）