Logistic 回归原理及计算机实现①

杨俊仙[1]，刘维奇[2]

（1. 山西大学　经济与管理学院，太原　030006；2. 山西大学　管理与决策研究所，太原　030006）

摘　要：Logistic 回归分析是用来处理分类因变量的最常用的统计方法。本文在对 Logit 模型、Probit 模型和 Tobit 模型区别与联系认识的基础上，通过统计分析，从应用领域、研究层次、研究成果时间序列特征等方面梳理归纳了三种模型的文献分布状况及发展。进一步详细介绍二分类 logistic 回归模型的原理、模型参数估计方法和假设检验方法，并以两个案例给出了其相应的 SPSS 和 SAS 计算机实现。

关键词：Logit 模型；SPSS；SAS

中图分类号：F224.0　**文献标识码**：A　**文章编号**：（2016）01-0081-13

0　引言

线性回归模型，是经济统计预测中常用的一种方法，但也存在局限性。在许多实际问题中，经常出现因变量是分类变量而不是连续变量的情形，这时线性回归模型就不再适用。比如，是否购买或销售某种产品，是否进行企业并购，风险偏好属于风险厌恶还是风险中性，企业是否破产等等。Logistic 回归分析是用来处理分类因变量的最常用的统计方法。1838 年，比利时学者 P. F. Verhuist 利用 Logistic 概率函数的 S 形增长曲线研究了人口问题，首次将 Logistic 回归分析这一函数引入社会科学研究领域。随后，国内外学者产出了许多应用研究与理论研究文献。一方面，Logistic 分布与极值分布的多种关系、Logistic 分布与指数分布之间的关系等的理论研究取得了一系列重要的成果，极大地推动了 Logistic 分布的参数估计及其分布的拟合优度检验的理论研究。而且，进一步研究发现了 Logistic 回归模型的扩展模型、替代模型：Probit 模型和 Tobit 模型等。Logistic 模型（也称 Logit 模型）使用的是累计逻辑函数，这并非唯一可使用的累积分布函数（CDF），研究发现正态 CDF 也是可行的，来自正态 CDF 的估计模型通常称为 Probit 模型（有时也称 Normit 模型）。在大多数应用中，两个模型十分类似，主要区别在于 Logit 的条件概率比 Probit 的以更慢的速度趋近于 0 或 1，Logistic 分布有稍微平坦的尾部。而 Probit 模型的一个扩展也就是 Tobit 模型，最先是 1958 年由 James Tobin 提出的，在这个模型中因变量仅仅在某些条件满足时才可以观测到，因此也叫限值因变量模型或截取回归模型。另一方面，许多学者已经开始并继续致力于 Logistic 回归应用研究的拓展，直至今日，Logistic 回归分析作为一种有效的数据处理方法在生物医学、生态工程、健康学、语言学、生物学、管理学和经济学等很多领域都有广泛的应用。下面对 Logit 模型、Probit 模型和 Tobit 模型三种定性因变量回归模型的文献，从研究层次、应用领域等方面进行统计分析，以期了解三种模型的文献分布状况及发展。

1　文献统计分析

1.1　时间序列特征

近 10 年，Logit 模型、Probit 模型和 Tobit 模型的研究论文数量增长快速。具体以中国知网数据库的

① 作者简介：杨俊仙（1972—），女，山西太谷人，山西大学经济与管理学院，副教授，研究方向：资产定价，金融计量经济学，Email：jxyang88@sxu.edu.cn；刘维奇（1963—），男，山西忻州人，博士，山西大学管理与决策研究所，教授，博士生导师，研究方向：金融工程与风险管理、时间序列分析，Email：liuwq@sxu.edu.cn。

期刊和硕博论文领域为平台搜索并对数据进行统计分析。关于 Logit 模型的文献方面，以“Logit”为关键词，选择全文选项搜索到 4595 篇公开发表于期刊的论文，从历年的中文期刊论文发表数量图可看出（图 1），1999 年之前 19 年里 Logit 模型的论文仅有 142 篇，且增速缓慢，之后论文数量出现快速增长，转折点是 2005 年，之后每年发文数量比上一年平均净增长 70 篇左右；而从 2000 年开始到现在，Logit 模型的硕博士学位论文共 5158 篇，2004 年后快速增长，2012 年的论文数量比上年增长 51%。

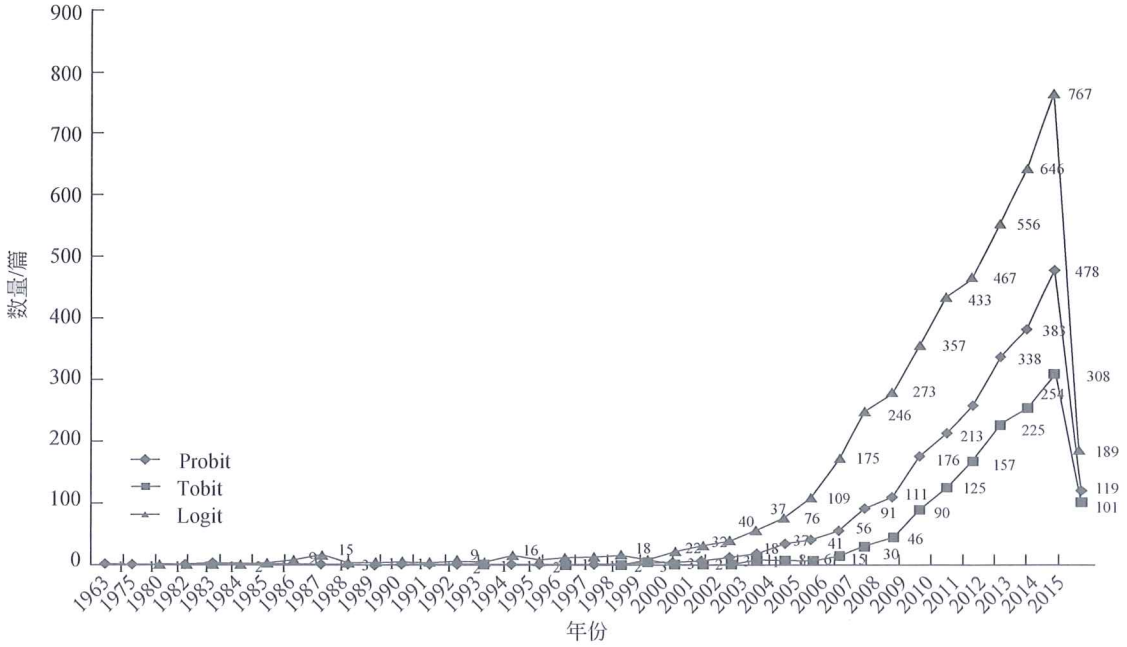

图 1　Logit、Probit 和 Tobit 相关中文期刊论文发表数量

Fig. 1　The number of published papers in Chinese about Logit, Probit and Tobit

关于 Probit 模型的文献方面，以“Probit”为关键词，选择全文选项搜索到 4519 条结果，其中发表在期刊上的有 2377 篇，硕博士学位论文有 1959 篇。从图 1 可以看出，从发表论文的数量来看，Probit 方法相关论文数量加速增长，2001 年之前 19 年里与 Probit 方法相关的论文仅有 47 篇，2002 年之后相关论文数量快速增长，每年发文数量比上一年净增长 25% 左右。从 2003 年开始与 Probit 方法相关的硕博士学位论文数量迅速增长，其中 2011～2012 年增长 127 篇，2013 年全年与 Probit 方法相关的硕博士学位论文数量达到 331 篇。

关于 Tobit 模型的文献方面，以“Tobit”为关键词，选择全文选项搜索到 2503 条结果，其中发表在期刊上的有 1398 篇，2005 年之前与 Tobit 方法相关的论文仅有 37 篇，2005 年之后相关论文数量快速增长，每年发文数量比上一年净增长 40 篇左右。此外，与 Tobit 模型相关的硕博士学位论文共 1001 篇，从 2007 年开始与 Logit 方法相关的硕博士学位论文数量迅速增长，2012 年全年的论文数量达到 226 篇。

1.2　学科分布特点

将中国知网分别以“Logit”“Probit”和“Tobit”为关键词检索文献的学科分布情况为：Logit 模型主要集中于经济、金融、数学和医药学方面的应用，涉及宏观经济管理与可持续发展、企业经济、金融、农业经济、公路与水路运输、投资、数学、证券、医药学（包括临床医学、中医学、畜牧与动物医学和预防医学与卫生学）和贸易经济等学科，其中位居前四位的宏观经济管理与可持续发展、企业经济、金融、农业经济占到 50%；Probit 模型、Tobit 模型的学科分布特点与 Logit 模型差别不大。另外，医药学三种方法的总和是 374 篇。

1.3　研究层次情况

分别以“Logit”“Probit”和“Tobit”为关键词进行搜索，选择研究层次选项检索结果：Logit 模型相

关的论文60.9%发表于核心期刊，36.9%来源于CSSCI，1.5%来源于EI，0.7%来源于SCI，而Tobit模型依次为64.7%，30.6%，3.7%，1.0%。与Probit模型相关的论文60.9%来源于核心期刊，来源于CSSCI的为38.8%，来源于SCI的为0.3%（表1）。

表1 Logit、Probit和Tobit相关的中文论文研究层次情况

Table 1 The research levels of papers in Chinese about Logit, Probit and Tobit

学科门类	研究层次名称	Logit		Probit		Tobit	
		数量	占比（%）	数量	占比（%）	数量	占比（%）
社会科学	基础研究	2478	71.62	1606	87.19	1003	82.28
	行业指导	370	10.69	128	6.95	104	8.53
	政策研究	141	4.08	88	4.78	90	7.38
	职业指导	100	2.89	20	1.09	22	1.80
	高级科普	1	0.03	—	—	—	—
	行业技术指导	370	10.69	—	—	—	—
	合计	3460	100.00	1842	100.00	1219	100.00
自然科学	工程技术	667	45.91	112	22.40	7	4.38
	政策研究	22	1.51	11	2.20	15	9.38
	高级科普	—	—	1	0.20	—	—
	专业实用技术	3	0.21	—	—	3	1.88
	行业技术指导	72	4.96	27	5.40	12	7.50
	标准与质量控制	14	0.96	2	0.40	—	—
	基础与应用基础研究	675	46.46	347	69.40	123	76.88
	合计	1453	100.00	500	100.00	160	100.00
其他	高等教育	17	65.38	8	72.73	—	—
	经济信息	3	11.54	2	18.18	—	—
	基础教育与中等职业教育	6	23.08	1	9.09	—	—
	合计	26	100.00	11	100.00	—	—

从学科分布来看，研究层次呈现以下特点。第一，社会科学方面的研究最多，Logit、Probit和Tobit方法在社会科学方面的研究分别占总体的70%、78%和88%，自然科学方面的研究次之，还有一小部分是关于高等教育、经济信息和基础教育与中等职业教育的研究。第二，在社会科学和自然科学的研究中基础研究均占最大的比重，基本都超过了50%，除了Logit在自然科学基础与应用基础研究方面占比为46.46%。在社会科学研究中占比排第二的为行业指导，在自然科学中工程技术方面的研究占比为第二。在自然科学的研究中主要为基础与应用基础研究和工程技术研究，二者所占比例超过了80%。

1.4 经济管理和统计类论文发表期刊分布情况

图2（a）、图2（b）分别是Logit模型、Probit模型经济管理和统计学学科中文论文发表期刊分布图，可以看出，Logit模型、Probit模型在《经济研究》《管理世界》和《中国农村经济》的占比较高。《统计研究》《数量经济技术经济研究》《金融研究》《财经研究》等期刊也发文不少。Tobit模型的分布状况趋同于其他两个模型。

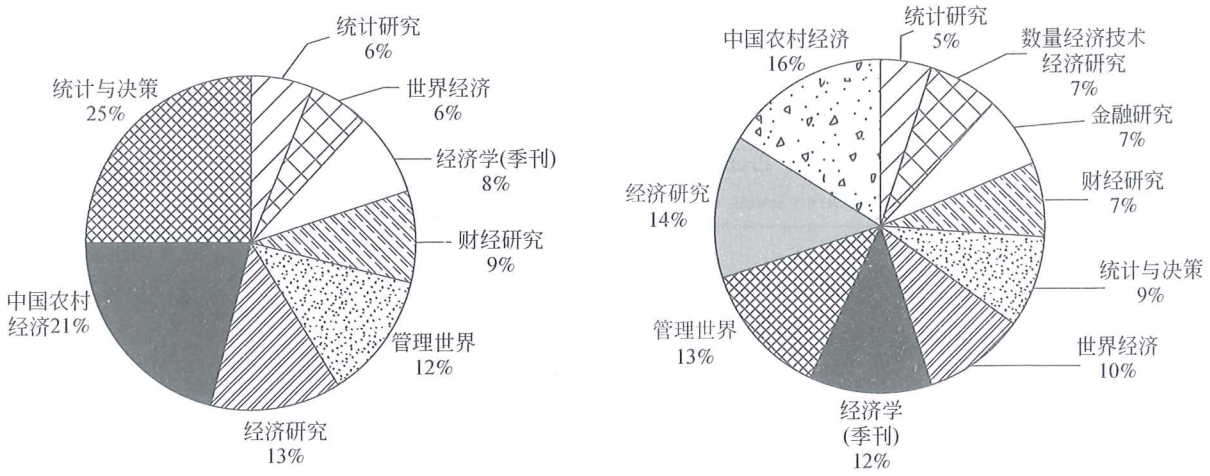

图 2　Logit 和 Probit 相关经济管理学科中文论文发表期刊分布

Fig. 2　The outline of published papers in Chinese about Logit, Probit and Tobit in Economic Management

1.5　高被引次数的论文情况

Logit 模型、Probit 模型和 Tobit 模型研究论文数量颇丰，将每个模型被引次数位居前 10 位的论文、作者、刊物情况进行统计，（前 5 位列于表 2），三种方法被引次数比较高的论文发表于 2004 ~ 2010 年的有 23 篇，占到 77%。

表 2　Logit、Probit 和 Tobit 被引次数最高的中文论文

Table 2　The highly cited papers in Chinese about Logit, Probit and Tobit

模型	论文	作者	刊名	年/期	被引（次）
Logit	1. 中国上市公司股权融资偏好解析——偏好股权融资就是缘于融资成本低吗?	陆正飞，叶康涛	经济研究	2004/04	922
	2. 商业银行信用风险评估及其实证研究	王春峰，万海晖 张维	管理科学学报	1998/01	510
	3. 公司绩效与高层更换	龚玉池	经济研究	2001/10	479
	4. 企业失败判别模型实证研究	高培业，张道奎	统计研究	2000/10	388
	5. 公司治理与财务舞弊关系的经验分析	蔡宁，梁丽珍	财经理论与实践	2003/06	302
Probit	1. 中国上市公司资本结构的影响因素和股权融资偏好	肖泽忠，邹宏	经济研究	2008/06	278
	2. 农村工业化以及人力资本在农村劳动力市场中的角色	陈玉宇，邢春冰	经济研究	2004/08	209
	3. 迁移的双重动因及其政策含义——检验相对贫困假说	蔡昉，都阳	中国人口科学	2002/04	208
	4. 影响农户购买农业保险决策因素的实证分析——以新疆玛纳斯河流域为例	宁满秀，邢郦 钟甫宁	农业经济问题	2005/06	189
	5. 贫困山区农业技术采用的决定因素分析	朱希刚，赵绪福	农业技术经济	1995/05	175
Tobit	1. 关于我国国有商业银行效率的实证分析与改革策略	朱南，卓贤 董屹	管理世界	2004/02	563
	2. 中国地方政府财政支出效率研究：1978—2005	陈诗一，张军	中国社会科学	2008/04	303
	3. 多元化经营与企业价值：我国上市公司多元化溢价的实证分析	苏冬蔚	经济学（季刊）	2005/11	296
	4. 影响中国农户借贷需求的因素分析	周小斌，耿洁 李秉龙	中国农村经济	2004/08	230
	5. 中国出口增长的二元边际及其因素决定	钱学锋，熊平	经济研究	2010/01	228

被引论文所发表的期刊级别较高，有《经济研究》《中国工业经济》和《管理科学学报》等，其中《经济研究》发表的相关论文引用最多。Logit 方法相关论文被引前十名中有五篇发表于《经济研究》，其中陆正飞与叶康涛发表于《经济研究》2004 年 4 期的《中国上市公司股权融资偏好解析——偏好股权融资就是缘于融资成本低吗?》一文引用最多，被引次数高达 922 次；龚玉池发表于《经济研究》的《公司绩效与高层更换》一文被引 479 次，位居第三。Probit 模型的论文被引次数位居一、二的均发表于《经济研究》，肖泽忠和邹宏的《中国上市公司资本结构的影响因素和股权融资偏好》被引用高达 278 次，陈玉宇和邢春冰的《农村工业化以及人力资本在农村劳动力市场中的角色》，被引 209 次。Tobit 方法中被引次数前十名中有两篇论文发表在《经济研究》，分别为钱学锋和熊平的《中国出口增长的二元边际及其因素决定》，被引 228 次，张宁等的“应用 DEA 方法评测中国各地区健康生产效率”，被引 176 次。

被引次数比较高的论文中研究方向既集中又分散。Logit 被引次数最高的前 10 篇文章有 7 篇是针对公司层面一些问题的研究，Probit 方法的 10 篇论文中有 6 篇是针对农村城镇化和农户借贷等农村经济问题的研究，Tobit 方法被引次数最高的前 10 篇论文中有 5 篇是针对效率的研究，包括能源效率、健康生产效率和商业银行效率等。

可见，logistic 回归是一个被广泛研究和使用的方法。下面我们就二分类 logistic 回归模型的原理、模型参数估计、假设检验及计算机实现来做介绍。

2 二分类 Logit 模型

2.1 Logit 模型原理

通常我们需要研究公司成功与失败等现象发生的概率 p 的大小，以及讨论 p 的大小与哪些因素有关，即

$$p=\frac{1}{1+e^{-Z}}=\frac{e^{Z}}{1+e^{Z}} \tag{1}$$

其中 Z 是影响因素 x_1，x_2，…，x_k 的线性函数。该方程是基于累积逻辑分布函数（cumulative logistic distribution function）的模型。

但是直接处理数值 p 存在两个困难：①由于 $0\leqslant p\leqslant 1$，以致 p 与自变量的关系难以用线性模型来描述，意味着 OLS 估计模型参数失效；②当 p 接近于 0 或 1 时，p 值的微小变化用普通的方法难以发现，不能很好地处理，因此需要转换为处理 p 的函数 Q（p）。要求构造的 Q（p）是 p 的严格单调函数，并且 Q（p）对 $p=0$ 或 $p=1$ 的附近的微小变化很敏感，即

$$\frac{dQ}{dp}\propto\frac{1}{p\ (1-p)} \tag{2}$$

所以，假设 Q（p）遵循下式时：

$$Q\ (p)=\ln\frac{1}{p\ (1-p)} \tag{3}$$

变换（3）称为 Logit 变换。当 p 从 0 变化到 1 时，Q（p）从 $-\infty$ 变化到 $+\infty$，这一变换在数据处理上带来很多方便。

基于上面的思想，当因变量是一个二元变量，只取 0 与 1 两个数值时，因变量取 1 的概率 p（$y=1$）就是要研究的对象。如果有很多因素影响 y 的取值，这些因素就是自变量，记为 x_1，x_2，…，x_k，这些 x_i 中既有定性变量，也有定量变量。模型如下：

$$L=\ln\frac{p}{1-p}=b_0+b_1x_1+\cdots+b_kx_k \tag{4}$$

就概念而言，公式（1）是 Logistics 分布函数表达的“Logistic 回归”，而公式（4）采用 Logit 形式表达的是“Logit 模型”。有些研究者对 Logistics 回归和 Logit 模型是根据所用自变量是否为连续变量来划分，将以分类自变量（categorical independent variables）构成的模型称为 Logit 模型，而将既有分类自变量又有连续自变量（continuous independent variables）的模型称为 Logistic 回归模型，有时为了方便，不管自变量

是哪种类型，人们将Logistic回归模型统称为Logit模型，甚至于将Logistic回归、Logistic模型、Logistic回归模型、Logit模型称谓通用，在此文中，我们称为Logit模型。

2.2 Probit模型

在二分类因变量的统计模型中，Probit模型提供了对Logit模型的一种替代选择，Logit模型是通过事件发生概率 p 进行Logit转换之后得到的。同样，这里事件发生概率 p 的非线性函数也可以通过Probit转换，得到一个关于 p 的单调函数，且该函数与自变量呈线性关系。

以 p_i 表示第 i 个观测案例发生某一事件的概率，它由以下标准累积正态分布函数给出：

$$p_i = \int_{-\infty}^{\eta_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\mu_i^2\right) d\mu_i \tag{5}$$

其中 Φ 表示标准正态分布的累积分布函数，式（5）常记作 $p_i=\Phi(\eta_i)$，$\eta_i \in (-\infty, +\infty)$。通过求标准累积正态分布函数的反函数，得到Probit转换（也称Normit转换），即

$$\eta_i = \Phi^{-1}(p_i) = \text{probit}(p_i) \tag{6}$$

Probit模型一般被写作：

$$\Phi^{-1}(p_i) = \eta_i = \sum_{k=0}^{K} \beta_k x_{ik} \tag{7}$$

或

$$p_i = \Phi\left(\sum_{k=0}^{K} \beta_k x_{ik}\right)。 \tag{8}$$

其中，k 是影响因素的个数，β_k 是回归系数，x_{ik} 表示第 i 个观察案例的第 k 个影响因素。Probit转换也克服了概率取值超出区间［0，1］的问题。

2.3 Logit模型与Probit模型的关系

通过不同的转换，Logit模型和Probit模型都能避免线性模型在处理二分类变量时存在的最大问题，即预测值取值范围的荒谬性。两者均是对事件发生概率 p 的一种非线性单调转换，只不过它们在转换过程中采用了不同的函数而已，两种转换所得到的结果非常相似，通常Logit估计值约是Probit估计值的1.8倍。这是因为对于logit模型其残差的标准差为 $\pi/\sqrt{3}=1.8138$，而对于probit模型，其残差的标准差为1。当 p 处于0.2到0.8之间时，这两种转换基本上都属于线性的，当 p 处于其他情形时，两者呈现出高度的非线性特征，非线性意味着如果将 p 作为自变量 x 的函数来进行建模的话，那么 x 对 p 的作用不是固定不变的，而是随着 x 取值的变化而变化。这一点与线性回归的情况极为不同。

在实际研究中，研究者并不知道Logit模型和Probit模型中哪一个是适合的模型。不过，出于概率分布函数的简洁性的考虑，同时考虑到Logit模型中对数发生比率比在解释形式上的便利性，许多学者选择Logit模型。相比之下，正态分布没有简洁的封闭表达。

3 Logit模型参数估计

通常用以估计Logit模型的数据有两种形式：群组或重复观测数据，或是个体水平上的数据。群组数据也称宏观数据，来自于汇总；个体水平上的数据也称微观数据，一般从抽样调查中取得。下面以案例的形式分别介绍基于分组和未分组两种数据的模型参数估计。

3.1 基于分组数据的参数估计

案例1[2]：在一次住房展销会上，与房地产商签订初步购房意向书的共有 $n=325$ 人，在随后的3个月的时间内，只有一部分顾客确实购买了房屋。购买了房屋的顾客记为1，没有购买房屋的顾客记为0，以顾客的年家庭收入（万元）为自变量 x，对表3中的数据，建立Logit模型。

表 3 房地产商签订购房意向情况表

Table 3 The condition of purchase intention signed by estate agents

组号	年家庭收入（万元）x	签订意向书人数（人）n_i	实际购房人数（人）m_i	实际购房比例 $\hat{p}_i=m_i/n_i$
1	1.5	25	8	0.320 000
2	2.5	32	13	0.406 250
3	3.5	58	26	0.448 276
4	4.5	52	22	0.423 077
5	5.5	43	20	0.465 116
6	6.5	39	22	0.564 103
7	7.5	28	16	0.571 429
8	8.5	21	12	0.571 429
9	9.5	15	10	0.666 667

这是一个分类自变量的问题。此案例的 Logit 模型为

$$p_i=\frac{\exp\ (b_0+b_1x_i)}{1+\exp\ (b_0+b_1x_i)}\ (i=1,\ 2,\ \cdots c) \tag{9}$$

其中，p_i 是组 i 的概率，b_0，b_1 为回归系数，x_i 是第 i 影响因素。经 logit 变换，得到

$$L_i=\ln\left(\frac{p_i}{1-p_i}\right)=b_0+b_1x_i+\mu_i \tag{10}$$

其中，L_i 是 p_i 的 Logit 变换，μ_i 是误差项，c 为公式中 i 的取值部分的分组数据组数，此例中 $c=9$。如果样本相当大，$\hat{p}_i$ 是 p_i 的良好估计值，利用 $\hat{p}_i$，令 $\hat{L}_i=\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)$ 得到一个相当好的估计 Logit 为

$$\hat{L}_i=\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)=\hat{b}_0+\hat{b}_1x_i \tag{11}$$

由于 $\mu_i\sim N\left(0,\ \frac{1}{N_iP_i\ (1-P_i)}\right)$ 是异方差，因此最小二乘法 OLS 估计失效，需要用加权最小二乘法 WLS 进行参数估计。

具体来讲，只有一个自变量的估计 Logit 模型的步骤如下：

步骤一：对每一收入水平 X_i，计算购买房屋的估计概率 $\hat{p}_i=m_i/n_i$；

步骤二：对每一 X_i 求 logit：$\hat{L}_i=\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)$；

步骤三：为解决异方差性的问题，将公式（10）变换为

$$\sqrt{\omega_i}L_i=b_0\sqrt{\omega_i}+b_1\sqrt{\omega_i}x_i+\sqrt{\omega_i}\mu_i \tag{12}$$

记为：$L^*{}_i=b_0\sqrt{\omega_i}+b_1x_i^*+\nu_i$，可证明变换后的误差项 ν_i 是同方差的。当 n_i 较大时，可以选取估计的方差的倒数作为权重（但这不是唯一选择），即 $\omega_i=n_i\hat{p}_i\ (1-\hat{p}_i)$。

步骤四：用 OLS 估计（3），不含截距项；

步骤五：依据 OLS 方式建立置信区间和检验假设。

分组数据的 Logit 模型可以很方便地推广到多个变量的情况，在此不作具体推广。这种方法只适用于大样本的情形。如果是小样本的未分组的数据则可以使用极大似然估计的方法进行模型估计。

3.2 基于个体水平上的（未分组的）数据的参数估计

案例 2[2]：在一次关于公共交通的社会调查中，一个调查项目为“是乘坐公交车上下班，还是骑自行车上下班”。因变量 $y=1$ 表示主要乘坐公交车上下班，$y=0$ 表示主要骑自行车上下班。自变量 x_1 是年龄（岁），作为连续型变量；x_2 是月收入（元）；x_3 是性别，$x_3=1$ 表示男性，$x_3=0$ 表示女性。调查对象是工

薪族群体，数据见表 4，试建立 y 与自变量 x_1，x_2，x_3 间的 Logit 模型。

表 4　公共交通社会调查表

Table 4　Public transportation social questionnaire

序号	性别	年龄	月收入	出行方式	序号	性别	年龄	月收入	出行方式
1	0	18	850	0	15	1	20	1000	0
2	0	21	1200	0	16	1	25	1200	0
3	0	23	850	1	17	1	27	1300	0
4	0	23	950	1	18	1	28	1500	0
5	0	28	1200	1	19	1	30	950	1
6	0	31	850	0	20	1	32	1000	0
7	0	36	1500	1	21	1	33	1800	0
8	0	42	1000	1	22	1	33	1000	0
9	0	46	950	1	23	1	38	1200	0
10	0	48	1200	0	24	1	41	1500	0
11	0	55	1800	1	25	1	45	1800	1
12	0	56	2100	1	26	1	48	1000	0
13	0	58	1800	1	27	1	52	1500	1
14	1	18	850	0	28	1	56	1800	1

Logit 的非线性特征使得在估计模型时采用最大似然估计的迭代方法，找到系数“最可能”的估计。在计算整个模型拟合度的时候，采用似然值而不是最小二乘估计所用的离差平方和。案例 2 是三因变量的情形，具体求解在计算机实现部分。下面，我们给出多因变量未分组数据的极大似然估计的思想及原理：

设 y 是 0-1 型变量，x_1，x_2，…，x_p 是与 y 相关的确定性变量，n 组观测数据为（x_{i1}，x_{i2}，…，x_{ip}；y_i），$i=1$，2，…，n，其中 y_1，y_2，…，y_n 是取值 0 或 1 的随机变量，y_i 与 x_{i1}，x_{i2}，…，x_{ip} 的关系式为

$$E\ (y_i)\ =p_i=f\ (b_0+b_1x_{i1}+b_2x_{i2}+\cdots+b_px_{ip}) \tag{13}$$

其中函数 $f\ (x)$ 是值域在［0，1］内的单调函数，b_0，b_1，…，b_p 是回归系数，对于 Logit 模型，有

$$f\ (x)\ =\frac{\exp(x)}{1+\exp(x)} \tag{14}$$

于是 y_i 是均值为 $E\ (y_i)\ =p_i=f\ (b_0+b_1x_{i1}+b_2x_{i2}+\cdots+b_px_{ip})$ 的 0-1 分布，概率函数为

$$\begin{cases}P\ (y_i=1)\ =p_i\\P\ (y_i=0)\ =1-p_i\end{cases}$$

即

$$P\ (y_i)\ =p_i^{y_i}\ (1-p_i)^{1-y_i}\ \ (y_i=0;\ i=1,\ 2,\ \cdots,\ n) \tag{15}$$

于是 y_1，y_2，…，y_n 的似然函数为

$$L=\prod_{i=1}^{n}P(y_i)=\prod_{i=1}^{n}p_i^{y_i}\ (1-p_i)^{1-y_i} \tag{16}$$

则 y_1，y_2，…，y_n 的对数似然函数为

$$\begin{aligned}\ln L&=\sum_{i=1}^{n}[y_i\ln p_i+(1-y_i)\ln(1-p_i)]\\&=\sum_{i=1}^{n}\left[y_i\ln\frac{p_i}{1-p_i}+\ln(1-p_i)\right]\\&=\sum_{i=1}^{n}\{y_i(b_0+b_1x_{i1}+\cdots+b_px_{ip})-\ln[1+\exp(b_0+b_1x_{i1}+\cdots+b_px_{ip})]\}\end{aligned} \tag{17}$$

然后求 b_0，b_1，b_2，…，b_p 满足对数似然函数最大的极大似然估计值 $\hat{b}_0$，$\hat{b}_1$，$\hat{b}_2$，…，$\hat{b}_p$。由于一阶

条件得到的似然方程是待估参数的非线性方程组，求解十分困难，需借助现代计算技术，进行数值迭代完成计算。

4 Logit 模型的假设检验

4.1 拟合优度检验

当用最大似然估计来拟合 logit 模型时，偏差（Deviance）可以较好地测量拟合优度。$\hat{L}_s$ 为设定模型所估计的最大似然值，它概括了样本数据由这一模型所拟合的程度。$\hat{L}_f$ 是饱和模型的最大似然估计值，$\hat{L}_s/\hat{L}_f$ 称为似然比。当样本足够大时，$-2\ln(\hat{L}_s/\hat{L}_f)$ 服从 χ^2 分布，自由度为所设模型中协变类型个数与系数个数之差，称作偏差，通常用 D 来表示，简记为 $-2LL$，即

$$D=-2\ln(\hat{L}_s/\hat{L}_f)=-2\ (\ln\hat{L}_s-\ln\hat{L}_f) \tag{18}$$

如果模型完全拟合，则似然比值为 1，这时 $D=-2\ln(\hat{L}_s/\hat{L}_f)$ 应当达到最小值 0，一个好的模型应该有较小的 D 值，换句话说，$\hat{L}_s$ 值相对于 $\hat{L}_f$ 值较小时，就会有较大的 D 值，此时所设模型拟合较差。

4.2 回归系数的统计显著性检验

Logit 模型的回归系数的最大似然估计是总体参数的渐进无偏的、有效的点估计。对系数的估计标准误提供了在换用不同样本时估计系数的可能变化范围。logit 模型回归系数的 MLE 估计近似服从正态分布，我们可以直接对回归系数进行显著性统计检验。假设零假设为

$$H_0:\ \beta_k=0 \tag{19}$$

β_k 为回归系数如果接受零假设，说明自变量 x_k 对事件发生可能性没有影响；如果拒绝原假设，说明事件发生可能性依赖于自变量 x_k 的变化。

对 logit 模型回归系数进行显著性检验，通常使用 Wald 检验。统计量为

$$W=(\hat{\beta}_k/\mathrm{SE}_{\hat{\beta}_k})^2 \tag{20}$$

其中，$\mathrm{SE}_{\hat{\beta}_k}$ 是 $\hat{\beta}_k$ 的标准误。W 服从自由度为自变量个数的近似 χ^2 分布，此例是自由度是 3，其在检验水平 α 下的常用的临界值为：$\chi^2_{0.05}$（3），$\chi^2_{0.01}$（3）或 $\chi^2_{0.001}$（3）。根据正态分布理论，Wald 统计量很容易计算，但是当回归系数的绝对值很大时，这一系数的估计标准误会膨胀，导致 Wald 统计量变得很小，以致犯第二类错误的概率增加。

5 Logit 模型的计算机实现

5.1 SPSS 的窗口实现

5.1.1 分组数据的情形——案例 1 的 WLS

首先，在 SPSS 软件操作中，点选 Transform→Compute Variable，计算 $\hat{p}_i=m_i/n_i$，$\ln\left(\dfrac{\hat{p}_i}{1-\hat{p}_i}\right)$ 和权重 $\omega_i=n_i\hat{p}_i\ (1-\hat{p}_i)$。

其次，在 SPSS 软件操作中，点选 Analyze→Regression→Liner Regression，Dependent：logit 变换 $\ln\left(\dfrac{\hat{p}_i}{1-\hat{p}_i}\right)$，Independent：年家庭收入 x，WLS Weight：权重 ω_i，见图 3。输出结果见图 4。

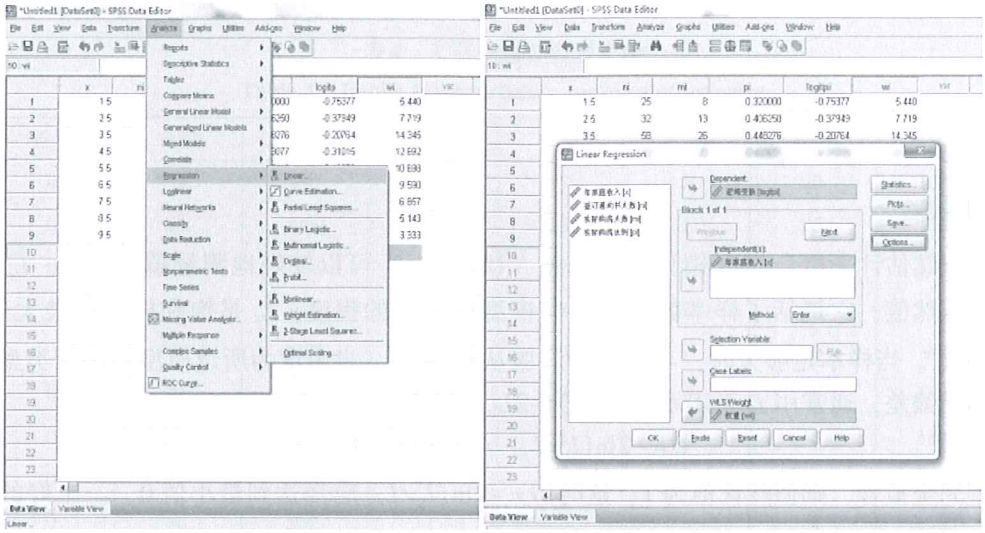

图 3　WLS 回归的操作界面

Fig. 3　The operation interface of WLS regression

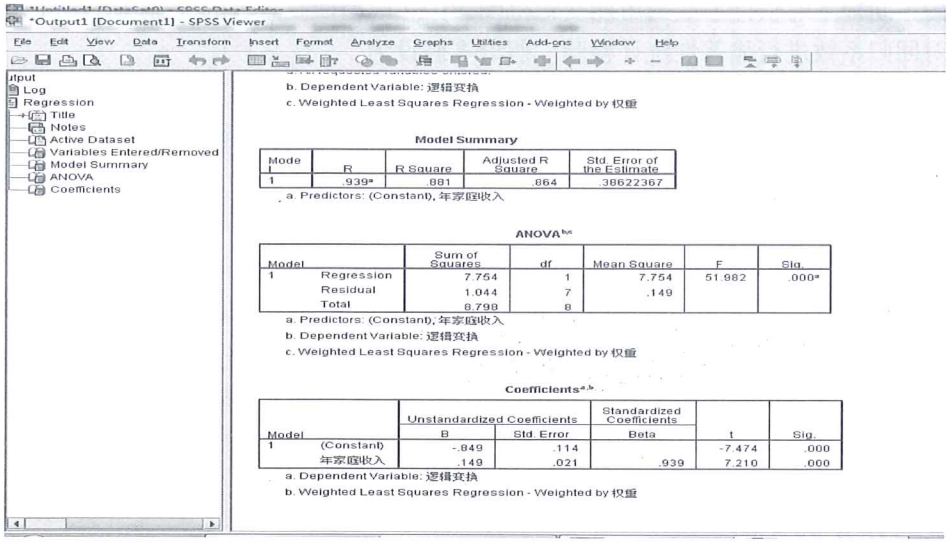

图 4　输出结果界面

Fig. 4　The output interface of results

从图 4 输出结果中可以看到：ANOVA 表中 $F=51.992$，显著性水平 Sig. $=0.000$，表明模型拟合很好。Coefficients 表中依据 t 值或显著性水平 *Sig.* 得到自变量年家庭收入的系数是显著的。因此，用加权最小二乘法得到的 Logit 模型为

$$\hat{p}_i=\frac{\exp(-0.849+0.149x)}{1+\exp(-0.849+0.149x)} \tag{21}$$

5.1.2　未分组数据的情形——案例 2 的 MLE

在 SPSS 软件操作中，点选 Analyze→Regression→Binary Logistic，进入 Logistic 回归对话框，如图 5，Dependent：出行交通方式 y，Covariate：性别、年龄、月收入，点击 OK，输出结果见图 6。

在输出结果中，Omnibus Test of Model Coefficients 表中 Chi-square 值和 Sig. 与 Model Summary 表中 -2Log likelihood 值，可以对整个模型检验，此例的图 6 结果表明模型是非常显著的。Variables in the Equation 表是模型的系数及检验，月收入 x_2 的 Wald 值 $=0.661$，*Sig.* $=0.416$，表明在 0.05 水平上，x_2 影响不显著。于是，只选取表现显著的自变量性别和年龄重新进行极大似然估计，结果见图 6，在 0.05 水平上，经拟合优度检验和模型参数显著性检验，得到此案例 2 的 Logit 模型为

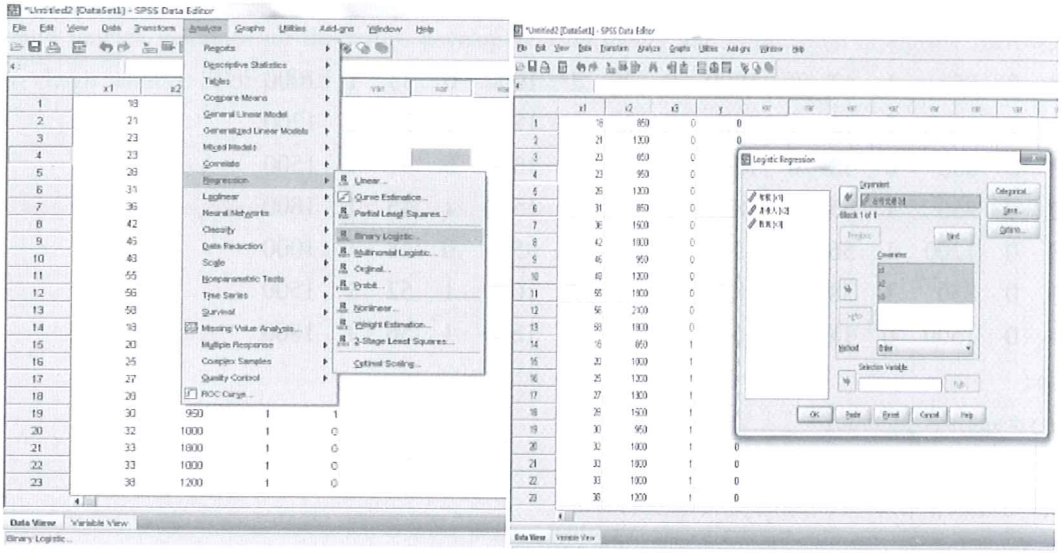

图 5　MLE 的操作界面

Fig. 5　The operation interface of MLE

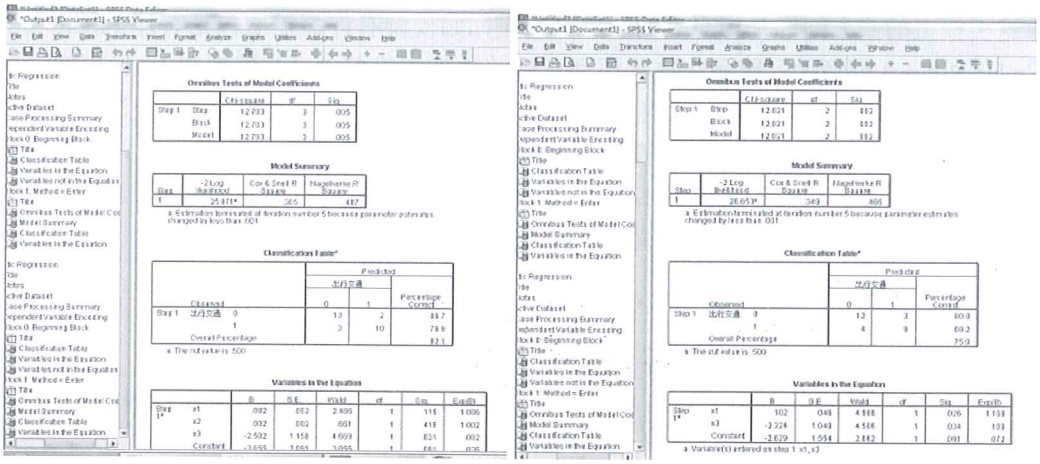

图 6　MLE 的输出界面

Fig. 6　The output interface of MLE

$$\frac{\hat{p}}{1-\hat{p}} = e^{-2.629+0.102x_1-2.224x_3} \tag{22}$$

5.2　SAS 程序

SAS 的 PROC LOGISTIC 过程可以处理 Logit 模型的极大似然估计。下面这些语句用于 LOGISTIC 过程：

/ * * 以下为必须的语句 * * /

PROC LOGISTIC <*option* >;

MODEL*response* = *independents* <*option*>;

/ * * 以下为可选的语句 * * /

BY*variables*;

FREQ*variable*;

OUTPUT <OUT = *SAS-data-set*>　　<*keword* = *name* . . . >;

WEIGHT*variable* <*option*>;

上述案例 2 的 SAS 程序如下：

```
data a;
```

```
    input y age sex income @@;
cards;
0  18  0  850   1  42  0  1000  0  20  1  1000  0  33  1  1000
0  21  0  1200  1  46  0  950   0  25  1  1200  0  38  1  1200
1  23  0  850   0  48  0  1200  0  27  1  1300  0  41  1  1500
1  23  0  950   1  55  0  1800  0  28  1  1500  1  45  1  1800
1  28  0  1200  1  56  0  2100  1  30  1  950   0  48  1  1000
0  31  0  850   1  58  0  1800  0  32  1  1000  1  52  1  1500
1  36  0  1500  0  18  1  850   0  33  1  1800  1  56  1  1800
proc logistic;
    model y=age sex income;
run;
```

其他语句可以根据研究问题的需要来选择添加。

6 结论与研究展望

无论是Logit模型还是Probit模型和Tobit模型，它们都属于定性响应回归模型，在离散资料的分析中应用广泛，这类模型的研究问题是相当多的。二分响应回归模型有许多扩展包括有序Probit和Logit模型，以及名义Probit和Logit模型等，隐藏在这些模型背后的哲理和简单Probit模型和Logit模型一样，但是其数学问题更加复杂，所以一直是学者们关注的研究内容。另外，持续时间模型已被关注，在该模型中，现象的持续时间取决于几个因素，关于这类模型，持续时间的长度现已成为研究者感兴趣的变量。Logistic回归模型目前研究集中在共线性、动态的Logit、模糊状态风险分析的广义logistic回归理论、样本量估计、统计功效估计等方面，相信该类模型将继续在理论研究和应用研究领域举足轻重。

参考文献：

[1] 张尧庭．定性资料的统计分析［M］．桂林：广西师范大学出版社，1991.
Zhang, Y. Statistical analysis of categorical data［M］. Guilin: *Guangxi Normal University Press*, 1991. (in Chinese)

[2] 何晓群．多元统计分析［M］．北京：中国人民大学出版社，2015.
He, X. Multivariate statistical Analysis［M］. Beijing: *China Renmin University Press*, 2015. (in Chinese)

[3] 谢宇．回归分析（修订版）［M］．北京：社会科学文献出版社，2013.
Xie, Y, . Regression analysis (Revision)［M］. Beijing: *Social sciences academic press*, 2013. (in Chinese)

[4] Lukas, M. , V. D. G. Sara, B. Peter. The group lasso for logistic regression［J］. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2008, 70 (1): 53-71.

[5] Drew, F. , S. Tomasz. Dynamic logit with choice aversion［J］. *Econometrica*, 2015, 83 (2): 651-691.

[6] Jan, B. , L. Kai. Corporate innovations and mergers and acquisitions［J］. *The Journal of Finance*, 2014, 69 (5): 1923-1960.

Logistic Regression: Principle and Computer Implementation

Yang Junxian[1], *Liu Weiqi*[2]

1. School of Economics and Management, Shanxi University , Taiyuan 030006, China;
2. Institute of Management and Decision, Shanxi University, Taiyuan 030006, China

Abstract: Logistic regression was the most commonly used statistical method used to handle categorical dependent variables. There were three common forms, logit model, probit model and tobit model. We counted and analysed the existing literatures of them, summed up the distribution and development of three models respectively. Moreover, we introduced the principle, parameter estimation method and hypothesis testing method of two-category

logistic regression. Further, based on two cases detailed Windows operating steps in software, such as SAS and SPSS. It would provide guidance for the future development on theory and application of logistic regression.

Key words: Logit model; SPSS; SAS