基于 Gumbel Copula函数的多维 Logit模型

李华民 黄海军 王惠文

(北京航空航天大学 经济管理学院, 北京 100191)

摘 要: 针对多维 Logit模型中的独立同分布 (IID, Identically & Independently Distributed)条件假设,提出了一种基于 Copula函数的离散选择模型. 利用 Copula函数获得多元随机变量的联合分布函数以及 Gumbel Copula函数的特性,得到了任意 2个随机项之差的联合分布,它依然服从 Logistic分布,形式上只比现有的分布函数多了一个倍参数. 进一步将此结果推广至多维选择问题,得到了无需 IID条件下一个方案被选中的概率,从而克服了多维 Logit模型的应用障碍.

关 键 词: Logit模型; Gumbel Copula函数; Logistic分布

中图分类号: U 491

文献标识码: A **文 章 编 号**: 1001-5965(2009)12-1443-03

Multinomial Logit model based on Gumbel Copula

Li Huamin Huang Haijun Wang Huiwen

(School of Economics and Management, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: For the IID (identically & independently distributed) condition of the multinomial Logit model, which requires all random terms existing in the utilities of alternatives to be independent each other, a discrete choice model was proposed, based on the Copula function which can be used to derive the joint probability distribution of multi-random variables. The IID condition is weakened and the distribution of the difference between every two random terms is obtained using the Gumbel Copula function's property. It is found that this distribution still follows the Logistic type but the difference must be multiplied by a parameter which can be estimated by the likelihood method from survey data. This result is then extended and employed in the discrete choice problem with more than two alternatives. The probability of choosing a specific alternative is rigorously formulated. The work surmounts the difficulty of applying the multinomial Logit model.

Key words: Logit model; Gumbel Copula function; Logistic distribution

诺贝尔经济学奖获得者 McFadden教授提出了多维 Logit模型[1],该模型在表达上具有显式性特点,求解算法又比较简单,因而在经济和交通行为分析等领域得到了广泛应用[2]. 多维 Logit模型对效用函数的随机项有一个很重要的假设,即"非相关选择项的独立性"条件,简称 IIA (Independent from Irrelevant Alternative)特性[1],因为有了这个特性,大大简化和方便了多维 Logit模型的应用与分析,但这个条件过于苛刻,与实际情景有较大的差距. 为此,人们发展了许多新的多维 Logit模型和求解方法,如分层 Logit模型和混合 Logit模型,试图放宽或去掉这个条件.

分层 Logit模型在各层内依然要求 IIA特性成立. 混合 Logit模型克服了传统 Logit模型的 2个重要缺陷 (随机喜好性限制和 IIA特性),具有高度的适应性. McFadden证明,混合 Logit模型覆盖了所有随机效用模型 (如 Logit, Probit和分层 Logit模型)[3],即在适当的混合分布函数下,可以无限趋近任何一种随机效用模型,但这个适当的分布函数很难找到. 近年来关于混合 Logit模型的

收稿日期: 2008-11-21

基金项目: 国家自然科学基金资助项目 (70521001); 国家 973基础研究计划资助项目 (2006CB705503)

作者简介: 李华民 (1979 -),男,河南平顶山人,博士生, li_lhm@sohu.com.

研究主要在应用方面[4-5],而理论研究几乎停滞不前.事实上,混合 Logit模型在数学形式上只是多维 Logit模型的积分,当问题规模较大时只能用仿真方法求解.

基于上述考虑,本文应用 Copula函数放宽多维 Logit模型的 IIA条件,得到了二维选择问题的精确联合分布函数和多维选择问题的概率.

1 基本模型

在描述行为选择问题时,效用函数包括 2个可加部分,一个是可观测项,另一个是随机项,即

$$U_{iq} = V_{iq} + \varepsilon_{iq} \tag{1}$$

式中,i为决策者;q为选择集中的一个选择项;ε_{iq}为效用函数的随机项;V_{iq}为可观测到的效用,通常表示为 $V_{iq} = \beta^{\mathrm{T}} X_{iq}$,此处 X_{iq}为被观测到的参数值,包含选择主体 i和选择客体 q的特征,向量 β是将要被估计的参数.所有离散选择模型的研究和发展都是围绕随机变量的假设和处理进行的[6].

文献[7]总结了经典的多维 Logit模型,假设所有的随机项是相互独立的,且都服从极值分布(又称为 Gumbel分布).根据效用最大化理论,任何个人都选择自身效用最大的选择项,即个人 i选择 q的概率等于选择项 q带给选择者的效用比其他任何选择项带给选择者的效用都大的概率,可以表示为

$$P_{iq} = P(U_{iq} \geq U_{ir},\ r \in C_i) \tag{2}$$

式中 C_i表示个人 i的所有选择项的集合.将式(1)代入式(2),可以得到

$$P_{iq} = P(V_{iq} + \varepsilon_{iq} \geq V_{ir} + \varepsilon_{ir},\ r \in C_i) = P(\varepsilon_{ir} - \varepsilon_{iq} \leq V_{iq} - V_{ir},\ r \in C_i) \tag{3}$$

2个服从 IID(Identically & Independently Distributed)-Gumbel分布的随机变量的差是 Logistic分布的随机变量,所以 $\varepsilon_{irq} = \varepsilon_{ir} - \varepsilon_{iq}$的分布函数为

$$F(\varepsilon_{irq}) = \frac{\exp \varepsilon_{irq}}{1 + \exp \varepsilon_{irq}} \tag{4}$$

从而可以得到个人 i选择 q的概率为

$$P_{iq} = \frac{1}{\sum_{r \in C_i} \exp(V_{ir} - V_{iq})} = \frac{\exp V_{iq}}{\sum_{r \in C_i} \exp V_{ir}} \tag{5}$$

2 Copula函数

Copula函数是一个把多个随机变量的联合分布函数与它们各自的边缘分布相连接的函数,其基本概念可参考文献[8],这里不再赘述.

如果 C是一个 n维的 Copula函数,且 $F_1, F_2, \cdots, F_n$都是分布函数,则由下式

$$F(x_1, x_2, \cdots, x_n) = C[F_1(x_1), F_2(x_2), \cdots, F_n(x_n)] \tag{6}$$

确定的函数 F是一个 n维联合分布函数,它的边缘分布是 $F_1, F_2, \cdots, F_n$.在很一般的条件下,式(6)中的 $C(u_1, u_2, \cdots, u_n)$是存在的.

迄今已发展了多种 Copula函数,其中 Gumbel Copula函数的统计特性比较突出,其一般形式为[9-10]

$$C(u_1, u_2, \cdots, u_n) = \exp\{-[(-\ln u_1)^{1/\beta} + (-\ln u_2)^{1/\beta} + \cdots + (-\ln u_n)^{1/\beta}]^{\beta}\} \tag{7}$$

式中,$u_i\ (i = 1, 2, \cdots, n)$为边缘分布函数;$0 < \beta \leq 1$,$\beta = 1$表示 $u_i\ (i = 1, 2, \cdots, n)$是相互独立的.Copula函数中的参数估计可以由 IFM(Inference For the Margins)算法完成[11],先用极大似然法估计单变量边缘分布的参数,再用同样的极大似然法估计 Copula函数中的参数 β.

3 基于 *Gumbel Copula* 函数的 *Logit* 模型

本文仍然假设 ε_{iq}服从 Gumbel分布,但不要求它们具有完全 ID特性,即随机项之间可以是不相互独立的.不失一般性,假定 ε_{iq}的分布函数为

$$F(x_{iq}) = \exp(-e^{-x_{iq}}) \tag{8}$$

式中 $x_{iq} = \varepsilon_{iq}$是服从标准 Gumbel分布的随机变量.利用 Gumbel Copula函数的性质式(6)和式(7),可得 ε_{ir}和 ε_{iq}的联合分布函数为

$$F_{rq}(x_{ir}, x_{iq}) = \exp\{-[(e^{-x_{ir}})^{1/\beta'} + (e^{-x_{iq}})^{1/\beta'}]^{\beta'}\} \quad 0 < \beta' \leq 1 \tag{9}$$

那么,$y = \varepsilon_{irq} = \varepsilon_{ir} - \varepsilon_{iq}$的分布函数如下:

$$F_{irq}(y) = \frac{\exp(y/\beta)}{1 + \exp(y/\beta)} \tag{10}$$

显然,式(10)表明 $\varepsilon_{irq} = \varepsilon_{ir} - \varepsilon_{iq}$是一个位置参数为 0、尺度参数为 β的 Logistic分布.将式(10)与式(4)比较,容易发现,差别仅在于式(10)中的随机变量被系数 $1/\beta$所倍乘.式(4)假定 ε_{ir}和 ε_{iq}是 ID-Gumbel分布,而式(10)只要求 ε_{ir}和 ε_{iq}分别服从 Gumbel分布,并不要求它们之间相互独立.所以,系数 β在某种程度上反映了 ε_{ir}和 ε_{iq}之间的相关性.这里的 β类似于分层 Logit模型中的参数,它衡量了随机项 ε_{ir}和 ε_{iq}的独立程度.β的值越大,ε_{ir}和 ε_{iq}的独立程度越高,也就是意味着二者的相关性越低.特别地,当 $\beta' = 1$时,ε_{ir}和 ε_{iq}相

互独立;当 $\beta \to 0$时,ε_{ir}和 ε_{iq}趋向于完全相关.统计量 $1-\beta$则是 ε_{ir}和 ε_{iq}相关性的度量,其值越大,二者的相关性越高.但需要注意的是,$1-\beta$并不完全是二者之间的相关系数,因为 β是在确定 ε_{ir}和 ε_{iq}的联合分布函数过程中出现的.利用文献[12]中的方法,可以得到 β基于截尾样本的近似极大似然估计值 $\hat{\beta}'$,具体结果如下.

设 $y_1, y_2, \cdots, y_n$ 是抽自 Logistic 总体 $\frac{\exp(y/\beta)}{1+\exp(y/\beta)}$的一个简单样本,$y_{(1)}, y_{(2)}, \cdots, y_{(n)}$为相应的次序样本.若 R_1 与 R_2 分别为从次序样本中剔除较小和较大观测值的个数,称 $y_{(R_1+1)}, \cdots, y_{(n-R_2)}$为 Logstic总体截尾样本.记

$$\gamma_i = \frac{i}{n+1} - \frac{i}{n+1}\left(1-\frac{i}{n+1}\right)\ln\frac{i}{n+1-i}$$

$$\delta_i = \frac{i}{n+1}\left(1-\frac{i}{n+1}\right)$$

$$i = R_1+1, \cdots, n-R_2$$

$$A = n - R_1 - R_2$$

$$m = R_1\delta_{R_1+1} + R_2\delta_{n-R_2} + 2\sum_{i=R_1+1}^{n-R_2}\delta_i$$

$$B = \frac{1}{m}\left(R_1\delta_{R_1+1}y_{(R_1+1)} + R_2\delta_{n-R_2}y_{(n-R_2)} + 2\sum_{i=R_1+1}^{n-R_2}\delta_i y_{(i)}\right)$$

$$C = \frac{1}{m}\left(R_1\gamma_{R_1+1} + R_2\gamma_{n-R_2} + 2\sum_{i=R_1+1}^{n-R_2}\gamma_i - (n-R_2)\right)$$

$$D = \sum_{i=R_1+1}^{n-R_2}(2\gamma_i - 1)y_{(i)} - R_1(1-\gamma_{R_1+1})y_{(R_1+1)} + R_2\gamma_{n-R_2}y_{(n-R_2)} - mBC$$

$$E = R_1\delta_{R_1+1}y_{(R_1+1)}^2 + R_2\delta_{n-R_2}y_{(n-R_2)}^2 + 2\sum_{i=R_1+1}^{n-R_2}\delta_i y_{(i)}^2 - mB^2$$

则 β的近似极大似然估计 $\hat{\beta}'$为

$$\hat{\beta}' = \frac{D+mBC+\sqrt{(D+mBC)^2+4A(E+mB^2)}}{2A}$$

$y=\varepsilon_{irq}=\varepsilon_{ir}-\varepsilon_{iq}$的分布函数为

$$F_{irq}(y) = \frac{\exp(y/\beta')}{1+\exp(y/\beta')} \tag{11}$$

现在考虑多维情形,若个人 i面对 M 个选择项(如交通方式),则 $M-1$个 $\varepsilon_{irq}=\varepsilon_{ir}-\varepsilon_{iq}$ $(r=r_1, \cdots, r_{q-1}, r_{q+1}, \cdots, r_M)$的分布函数都可以得到.记 $\varepsilon_{ir_1}-\varepsilon_{iq}, \cdots, \varepsilon_{ir_{q-1}}-\varepsilon_{iq}, \varepsilon_{ir_{q+1}}-\varepsilon_{iq}, \cdots, \varepsilon_{ir_M}-\varepsilon_{iq}$所对应的分布函数分别是 $F_{ir_1q}(y_{r_1})$, $\cdots$, $F_{ir_{q-1}q}(y_{r_{q-1}})$, $F_{ir_{q+1}q}(y_{r_{q+1}})$, $\cdots$, $F_{ir_Mq}(y_{r_M})$,有

$$F_{ir_1q}(y_{r_1}) = \frac{\exp(y_{r_1}/\hat{\beta}_1')}{1+\exp(y_{r_1}/\hat{\beta}_1')}$$

$$\cdots$$

$$F_{ir_{q-1}q}(y_{r_{q-1}}) = \frac{\exp(y_{r_{q-1}}/\hat{\beta}_{q-1}')}{1+\exp(y_{r_{q-1}}/\hat{\beta}_{q-1}')}$$

$$F_{ir_{q+1}q}(y_{r_{q+1}}) = \frac{\exp(y_{r_{q+1}}/\hat{\beta}_{q+1}')}{1+\exp(y_{r_{q+1}}/\hat{\beta}_{q+1}')}$$

$$\cdots$$

$$F_{ir_Mq}(y_{r_M}) = \frac{\exp(y_{r_M}/\hat{\beta}_M')}{1+\exp(y_{r_M}/\hat{\beta}_M')}$$

其中,$\hat{\beta}_1', \cdots, \hat{\beta}_{q-1}', \hat{\beta}_{q+1}', \cdots, \hat{\beta}_M'$为相应的 $M-1$个参数的估计值,$0<\hat{\beta}_j\le 1$, $j=1, \cdots, q-1, q+1, \cdots, M$.然后,再利用 Gumbel Copula函数的性质,得到 $\varepsilon_{ir_1}-\varepsilon_{iq}, \cdots, \varepsilon_{ir_{q-1}}-\varepsilon_{iq}, \varepsilon_{ir_{q+1}}-\varepsilon_{iq}, \cdots, \varepsilon_{ir_M}-\varepsilon_{iq}$的联合分布函数如下:

$$\exp\{-[(-\ln F_{ir_1q}(y_{r_1}))^{1/\beta} + \cdots + (-\ln F_{ir_{q-1}q}(y_{r_{q-1}}))^{1/\beta} + (-\ln F_{ir_{q+1}q}(y_{r_{q+1}}))^{1/\beta} + \cdots + (-\ln F_{ir_Mq}(y_{r_M}))^{1/\beta}]^{\beta}\}$$

其中参数 β仍然采用极大似然法估计,但其显式很难得到,记估计值为 $\hat{\beta}$

最后,得到个人 i选择交通方式 q的概率如下:

$$P_{iq} = P(V_{iq}+\varepsilon_{iq} > V_{ir}+\varepsilon_{iq}, r\ne q) = P(\varepsilon_{ir}-\varepsilon_{iq} < V_{iq}-V_{ir}, r\ne q) =$$

$$\exp\left\{-\left[\left(-\ln\frac{e^{(V_{iq}-V_{ir_1})/\hat{\beta}_1}}{1+e^{(V_{iq}-V_{ir_1})/\hat{\beta}_1}}\right)^{1/\hat{\beta}} + \cdots + \left(-\ln\frac{e^{(V_{iq}-V_{ir_{q-1}})/\hat{\beta}_{q-1}}}{1+e^{(V_{iq}-V_{ir_{q-1}})/\hat{\beta}_{q-1}}}\right)^{1/\hat{\beta}} + \left(-\ln\frac{e^{(V_{iq}-V_{ir_{q+1}})/\hat{\beta}_{q+1}}}{1+e^{(V_{iq}-V_{ir_{q+1}})/\hat{\beta}_{q+1}}}\right)^{1/\hat{\beta}} + \cdots + \left(-\ln\frac{e^{(V_{iq}-V_{ir_M})/\hat{\beta}_M}}{1+e^{(V_{iq}-V_{ir_M})/\hat{\beta}_M}}\right)^{1/\hat{\beta}}\right]^{\hat{\beta}}\right\} \tag{12}$$

式(12)在形式上比较复杂,而且需要对 M 个参数进行估计(先计算 $\hat{\beta}_i$, $i=1, \cdots, q-1, q+1, \cdots, M$,再计算 $\hat{\beta}$).

(下转第 1463页)

参考文献 (References)

[1] 陈叶菁. 装备维修保障设计方案评估方法研究 [D]. 长沙:国防科技大学信息系统与管理学院, 2006
Chen Yejing Study on evaluation method of materiel maintenance support plan[D]. Changsha: School of Information System & Mangement, National University of Defense Technology, 2006 (in Chinese)

[2] 徐东. 装备综合保障关键技术研究 [D]. 长沙:国防科技大学机电工程与自动化学院, 2006
Xu Dong The research on key technologies of integrated logistics support[D]. Changsha: College of Mechatronics Engneering and Automation, National University of Defense Technology, 2006 (in Chinese)

[3] 姜晨,赵喜春,殷志祥. 基于 eEPC的保障方案权衡建模仿真 [J]. 弹箭与制导学报, 2007, 27(2): 320 - 324
Jiang Chen, Zhao Xichun, Yin Zhixiang eEPC-based modeling and simulation in support concept trade off[J]. Journal of Projectiles, Rockets, Missiles and Guidance, 2007, 27(2): 320 - 324 (in Chinese)

[4] Bevilacqua M, Braglia M. The analytic hierarchy process applied to maintenance strategy selection [J]. Reliability Engineering and System Safety, 2000, 70(1): 71 - 83

[5] Michael A Bourcier. Logistic test and evaluation in flight test [R]. RTO-AG-300-V20, 2001

[6] Dinesh Kumar U. Nowicki David, Jose E Ramirez-Marquez, et al A goal programming model for optimizing reliability, maintainability and supportability under performance based logistics [J]. International Journal of Reliability, Quality and Safety Engineering, 2007, 14(3): 251 - 261

[7] Headquarters Department of the Army. Integrated logistics support [R]. AR700-127, 2005

[8] Charnes A, Cooper W W, Rhode E. Measuring the efficiency of decision making units [J]. European Journal of Operational Research, 1978, 12: 429 - 444

[9] 魏权龄. 数据包络分析 [M]. 北京:科学出版社, 2004
Wei Quanling Data envelopment analysis [M]. Beijing: Science Press, 2004 (in Chinese)

[10] 黄朝峰,沈永平,陈英武. 基于生产前沿面的 DEA有效单元再评价研究 [J]. 运筹与管理, 2008, 17(2): 48 - 52
Huang Chaofeng, Shen Yongping, Chen Yingwu. Research on re-evaluating DEA efficient DMU based on production frontier [J]. Operations Research and Management Science, 2008, 17(2): 48 - 52 (in Chinese)

(编　辑:娄　嘉)

(上接第 1445页)

4 结束语

本文所得到的选择概率计算量比较大,这是放宽 ID条件所必需付出的代价. 重要的是,虽然在推导过程中假设随机项服从 Gumbel分布,但该过程可以用于任何其他分布情形,这为解决实际问题提供了普适性途径,不必再局限于 ID条件和 Gumbel分布.

参考文献 (References)

[1] McFadden D. Modeling the choice of residential location [C]// Karlqvist A, Snickars F, Weibull J. Spatial Interaction Theory and Planning Models Amsterdam: North-Holland, 1978: 105 - 142

[2] 关宏志. 非集计模型:交通行为分析的工具 [M]. 北京:人民交通出版社, 2004
Guan Hongzhi Disaggregate mode: a tool of traffic behavior analysis[M]. Beijing: China Communications Press, 2004 (in Chinese)

[3] McFadden D, Train K E. Mixed MNL models of discrete response [J]. Journal of Applied Econometrics, 2000, 15(5): 447 - 470

[4] Hess S, Bierlaire M. Estimation of value of travel-time savings using mixed logit models [J]. Transportation Research Part A, 2005, 39(3): 221 - 236

[5] Greene W H, Hensher D A. Accounting for heterogeneity in the variance of unobserved effects in mixed logit models [J]. Transportation Research Part B, 2006, 40(1): 75 - 92

[6] 陆化普,黄海军. 交通规划理论研究前沿 [M]. 北京:清华大学出版社, 2007: 77 - 91
Lu Huapu, Huang Haijun. Theory research progress of transportation planning [M]. Beijing: Tsinghua University Press, 2007: 77 - 91 (in Chinese)

[7] Train K E. Discrete choice methods with simulation [M]. UK: Cambridge University Press, 2003

[8] Nelsen R B. An introduction to Copulas[M]. Second Edition. New York: Springer Verlag, 2006

[9] Genest C, Rivest L P. A characterization of Gumbel's family of extreme value distributions [J]. Statist Probab Lett, 1989, 8(3): 207 - 211

[10] Hutchinson T P, Lai C D. Continuous bivariate distributions, emphasising applications [M]. Adelaide: Rumsby Scientific Publishing, 1990

[11] Joe H, Xu J J. The estimation method of inference functions for margins for multivariate models [R]. Department of Statistics, University of British Columbia, Technical Report, 166, 1996

[12] 杨振海,程维虎. 基于 Logistic总体 II型截尾样本分布参数的近似极大似然估计 [J]. 北京工业大学学报, 2004, 30(2): 235 - 240
Yang Zhenhai, Cheng Weihu Approximate maximum likelihood estimation based on Logistic distribution parameters of type II censored samples [J]. Journal of Beijing Polytechnic University, 2004, 30(2): 235 - 240 (in Chinese)

(编　辑:赵海容)