面板数据的聚类分析及其应用*

朱建平　陈民恳

内容提要:不同于传统的计量建模分析,本文探讨了多元统计方法在面板数据分析上的运用。文中介绍了面板数据的统计描述方法,构造了面板数据之间相似性的统计指标,并在此基础上提出了面板数据聚类分析的有效方法,通过实际应用取得了良好的效果。

关键词:面板数据;聚类分析;计量经济;多元统计
中图分类号:C812　　**文献标识码**:A　　**文章编号**:1002 - 4565(2007)04 - 0011 - 04

The Cluster Analysis of Panel Data and Its Application

Zhu Jianping & Chen Minken

Abstract: Unlike the traditional econometric modeling analysis, this paper discusses the application of multivariate statistical methods for panel data. It introduces the statistical description of panel data and constructs the statistical indicators for the similarity of the data, and thereby the method of clustering panel data is proposed. Finally, the method is proved to be effective through the practical application.

Key words: Panel data; cluster analysis; econometrics; multivariate statistics

一、引言

面板数据(Panel Data)是截面数据和时间序列数据的组合,是现实生活中很常见的数据形式,例如我国 31 个省级地区的 20 年的国内生产总值数据,就是一组"Panel Data"。从 1970 年代末以来,Panel Data 模型的理论方法已日渐成熟,涌现了大量有关的理论和经验分析文章,形成了现代计量经济学的一个相对独立的分支[1]。绝大多数有关 Panel Data 的理论[2],都是从计量建模的角度着手,从单方程模型到联立方程模型,从变截距模型到变系数模型,从线性模型到非线性模型等等,另一方面,都是着重于模型参数估计方法的研究。Bonzo D. C. 和 Hermosilla A. Y. 等统计学家则另辟蹊径,将多元统计方法引入到 Panel Data 的分析中来[3]。Bonzo D. C. 运用概率连接函数(probability link function)改进聚类分析的算法,从而将聚类分析用于面板数据的分析。然而,对面板数据的统计描述,以及刻画面板数据之间的相似性研究的不多,本文将针对此问题进行讨论,构造面板数据的相似指标,并在此基础上提出面板数据聚类分析的有效方法。

*本文获国家教育部"新世纪优秀人才支持计划"(NCET-04-0608)资助;国家教育部社科研究规划项目(06JA910003)资助。

参考文献

[1]Marc Boulle. Khiops: A Statistical Discretization Method of Continuous Attributes [J]. Machine Learning, 2004(55): 53—69.

[2]李刚,李霁伦,童[illegible]. WILD:基于加权信息损耗的离散化算法[J]. 南京大学学报(自然科学),2001(3):148—152.

[3]王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社,2003. 51.

[4]李立萍,张明友. 信息论导引[M]. 成都:电子科技大学出版社,2005. 33.

作者简介

刘云霞,山西省,1978 年生,女,厦门大学计划统计系 04 级博士研究生,研究方向为数据分析,厦门大学经济学院计划统计系 04 级博士。

曾五一,福建省,1953 年生,男,现为厦门大学计划统计系教授,博士生导师,中国统计学会副会长、教育部统计学教学指导分委员会副主任委员。

(责任编辑:竹　影)

二、面板数据的统计描述

对面板数据的研究已经形成了较为成熟的理论,但是对于面板数据的预处理往往被人们所忽视。一般人们根据实际情况,通常对面板数据就是通过计量经济模型进行分析,这样具有一定的盲目性,很难直接建立能反映实际问题的模型。面板数据实际上是一种复杂的数据结构形式,在对其进行深入的分析之前,特别是建立计量经济模型时,需要对面板数据有一个初步的了解,这样会从原始的面板数据中获得必要的信息。在此,引进面板数据的统计描述方法[4],不仅为面板数据的预处理提供了思路,而且为面板数据的深入分析奠定了理论基础。

设单指标的面板数据为 $x_i(t)$, $i=1,2,\cdots,N$, $0\leq t\leq T$,那么称

$$\bar{x}(t)=\frac{1}{N}\sum_{i=1}^{N}x_i(t),0\leq t\leq T \tag{1}$$

为 $x_i(t)$ 的均值函数(the mean function),$\bar{x}(t)$ 表示一种动态平均水平。称

$$var_x(t)=\frac{1}{N-1}\sum_{i=1}^{N}[x_i(t)-\bar{x}(t)]^2,0\leq t\leq T \tag{2}$$

为 $x_i(t)$ 的方差函数(the variance function),其平方根为 $x_i(t)$ 的标准差函数。

对于不同的时点 $0\leq t_1<t_2\leq T$,称

$$cov_x(t_1,t_2)=\frac{1}{N-1}\sum_{i=1}^{N}[x_i(t_1)-\bar{x}(t_1)][x_i(t_2)-\bar{x}(t_2)] \tag{3}$$

为 $x_i(t)$ 的协方差函数(the covariance function)。称

$$corr_x(t_1,t_2)=\frac{cov_x(t_1,t_2)}{\sqrt{var_x(t_1)\,var_x(t_2)}},0\leq t_1<t_2\leq T \tag{4}$$

为 $x_i(t)$ 的联合相关函数(the associated correlation function)。

设有一对面板数据$(x_i(t),y_i(t))$, $i=1,2,\cdots,N$,$0\leq t\leq T$,称

$$cov_{x,y}(t_1,t_2)=\frac{1}{N-1}\sum_{i=1}^{N}[x_i(t_1)-\bar{x}(t_1)][y_i(t_2)-\bar{y}(t_2)],0\leq t_1<t_2\leq T \tag{5}$$

为$(x_i(t),y_i(t))$的交叉协方差函数(the cross-covariance function)。称

$$corr_{x,y}(t_1,t_2)=\frac{cov_{x,y}(t_1,t_2)}{\sqrt{var_x(t_1)\,var_y(t_2)}},0\leq t_1<t_2\leq T \tag{6}$$

为$(x_i(t),y_i(t))$的交叉相关函数(cross-correlation function)。

有了对面板数据的统计描述后,可以利用所获得的信息,根据要解决的实际问题构建面板数据的计量经济模型、面板数据主成分分析、时间函数的光顺性分析、面板数据的聚类分析等等。

三、面板数据的聚类分析

1. 面板数据的相似指标

由于大型数据库中面板数据的出现,聚类分析的研究工作自然涉及到面板数据的有效聚类分析上。那么,面板数据的聚类分析所针对的数据类型如何呢?这一问题的明确,将会为面板数据的聚类分析方法的研究澄清思路。

对于面板数据 $x_i(t)$, $i=1,2,\cdots,N$,$0\leq t\leq T$,考虑 N 个面板数据之间的近似性用面板之间的距离表示,其表现形式是一个 $N\times N$ 的对称阵,即

$$\begin{bmatrix}0&\delta_{1,2}&\delta_{1,3}&\cdots&\delta_{1,N}\\&0&\delta_{12}&\cdots&\delta_{2,N}\\&&\ddots&\ddots&\cdots\\&&&\ddots&\delta_{N-1,N}\\&&&&0\end{bmatrix} \tag{7}$$

其中 $\delta_{i,j}$ 是第 i 个面板数据与第 j 个面板数据之间的相异程度的量化表示,当第 i 个与第 j 个面板数据相似或"接近",其值越接近于 0。

对于设定的面板数据 $x_i(t)$, $i=1,2,\cdots,N$,$0\leq t\leq T$,那么,面板数据之间的相似指标可用:

(1)差异的上确界。

$$\delta_{ij}^{(1)}=\sup\{|x_i(t)-x_j(t)|,0\leq t\leq T\} \tag{8}$$

(2)一致差异。

$$\delta_{ij}^{(2)}=\int_0^T|x_i(t)-x_j(t)|\,dt \tag{9}$$

如果针对间断型的面板数据 $x_i(t_k)$, $i=1,2,\cdots,N$,$0\leq t_1<t_2<\cdots<t_m\leq T$,面板数据之间的相似指标可用:

(3)差异的最大值:

$$\delta_{ij}^{(3)}=\max_{1\leq k\leq m}|x_i(t_k)-x_j(t_k)| \tag{10}$$

(4)差异的绝对和:

$$\delta_{ij}^{(4)}=\sum_{k=1}^{m}|x_i(t_k)-x_j(t_k)| \tag{11}$$

(5)差异的欧氏距离:

$$\delta_{ij}^{(5)}=\sum_{k=1}^{m}[x_i(t_k)-x_j(t_k)]^2 \tag{12}$$

2. 面板数据的聚类分析

聚类分析的关键是,对所研究的问题构造数据

之间的相似指标。针对复杂的面板数据,根据连续和间断的情形,从不同的角度,提出了描述面板数据之间相似程度的指标,由此生成了相应的相似矩阵式(7)。

在此基础上,利用系统聚类分析就可以得到分析结果,其聚类的基本过程是:假设面板数据 $x_i(t)$, $i=1,2,\cdots,N$, $0\leq t\leq T$,第一步将每个数据 $x_i(t)$, $i=1,2,\cdots,N$,独自聚成一类,共有 N 类;第二步根据所确定的面板数据的相似指标把"距离"较近的两个面板数据聚合为一类,其它的面板数据仍各自聚为一类,共聚成 $N-1$ 类;第三步将"距离"最近的两个类进一步聚成一类,共聚成 $N-2$ 类;……,以上步骤一直进行下去,最后将所研究的面板数据全聚成一类。为了直观地反映以上的系统聚类过程,将对实际问题进行研究。

四、实证分析

改革开放近 30 年来,我国经济持续高速增长,人民的收入也不断提高。但是,长期以来,粗放型的经济增长方式没有得到根本改变,依靠高投入而带来高产出的做法越来越难以维系。在市场经济条件下,有效需求才是维持经济稳定增长的持久动力源。有效需求为有支付能力的需求,表现为居民的实际支出,它主要取决于居民的收入状况。近年来,我国实行了一些旨在扩大内需的宏观经济政策,但效果不甚明显,消费需求对经济增长的拉动作用没有充分发挥出来,使经济增长需求结构的偏斜更加突出,不利于我国经济长期健康稳定的发展。

为了深入了解近年来我国城镇居民的收入和支出情况,找出不同区域的城镇居民在收入状况和支出行为上的差异,以便因地制宜地引导居民消费,笔者将对 1995 年至 2004 年全国 31 个省市城镇居民的人均年收入和消费性支出的面板数据进行聚类分析。文中的数据来自 1996 年至 2005 年的《中国统计年鉴》,由于篇幅有限,原始数据不再列出。

通过对 1995 年至 2004 年全国 31 个省市的城镇居民收入状况面板数据的直观分析,只能大致看出 10 年来,我国城镇居民的收入水平在不断提高,却无法对各个地区的收入状况作出准确判断和区分。因而,有必要对其作聚类分析。

根据上面介绍的方法,选择式(12),欧氏距离作为该面板数据的相似指标,根据式(2)所定义的方差函数,选用离差平方和法进行聚类分析得到全国 31 个省市的城镇居民收入面板数据的聚类树形图,见图 1。

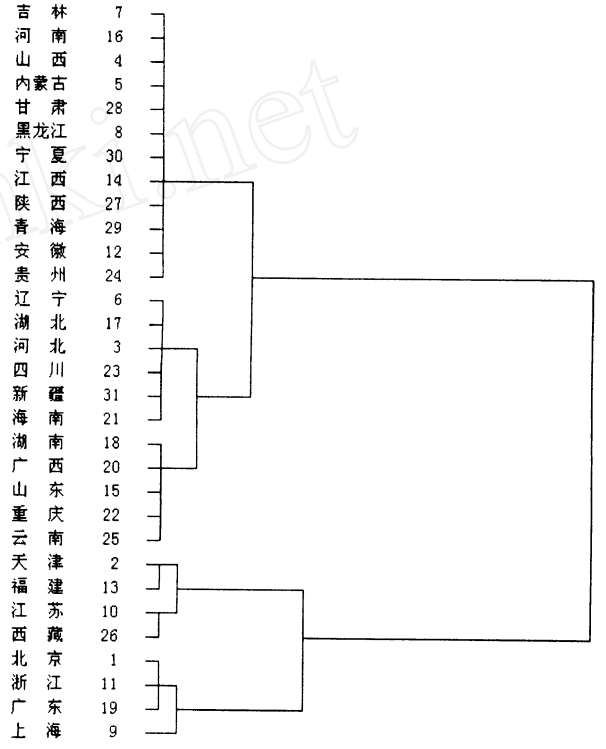

图 1　1995—2004 年全国 31 个省市城镇居民收入的聚类树形图

从图 1 中,可以清楚地得知,近 10 年期间,北京、浙江、广州和上海的城镇居民的人均年收入位居全国首列;天津、江苏、福建和西藏的城镇居民的人均年收入则属于第二个层次;山西、内蒙古、宁夏、贵州等广大中西部地区的城镇居民的人均收入水平则居全国末位。如果将最高收入省市的城镇居民的人均年收入与最低收入省市的城镇居民的人均年收入作比较,可以发现,十年来它们之间的差距维持在 2.03 倍左右,但是略有扩大的趋势,见图 2。

有效需求表现为居民的实际支出,并取决于居民的收入状况,那么具有高收入的省市的城镇居民是否会表现为高支出,各地区城镇居民的消费行为是否会与收入水平相对应呢?为此,对 1995 年至 2004 年全国 31 个省市的城镇居民的人均消费性支出状况面板数据也作了聚类分析,并得到聚类树形图,见图 3。

从图 3 中,发现近 10 年期间,北京、上海、浙江和广州的城镇居民的人均消费性支出水平仍位居全

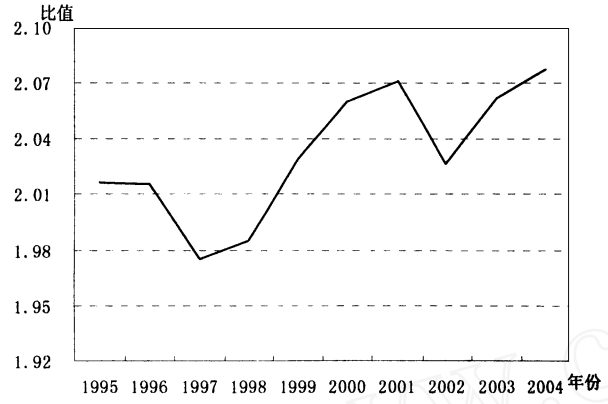

图 2 最高收入省市城镇居民的人均收入与最低收入省市的人均收入比值

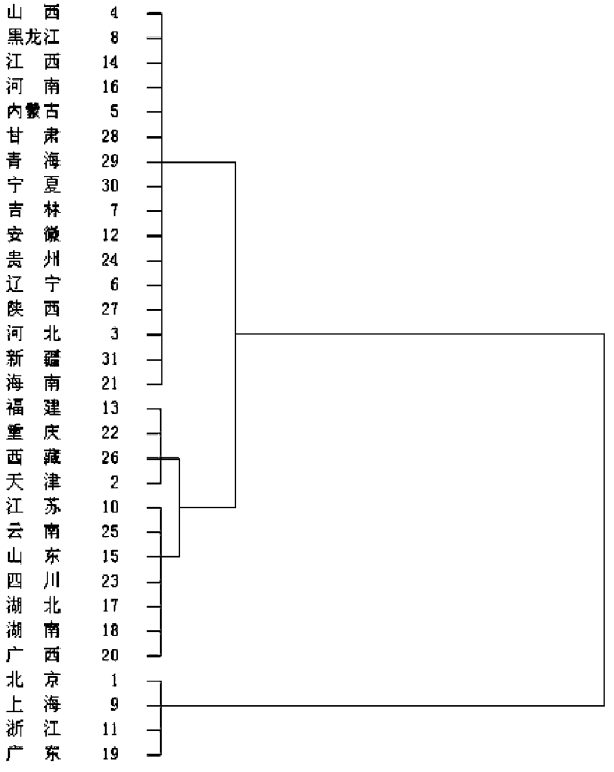

图 3 1995 —2004 年全国 31 个省市城镇居民支出的聚类树形图

国首列,而收入水平较高的福建、天津、西藏和江苏则与收入水平低它们一筹的云南、四川、山东、重庆和湖北等地的支出水平相近,收入最低的广大中西部省市的城镇居民的支出水平仍排在全国的末位。

通过分析发现,1995 年以来,我国城镇居民的收入水平不断提高,人民生活不断改善,但贫富差距不断扩大,地区发展不均衡,广大中西部地区仍大大落后于东部沿海地区。从动态角度看,各地区城镇居民的支出行为存在明显差异,部分省市的有效需求激发不足。北京、上海、浙江和广东四省市的城镇居民,由于现期收入很高,加上良好的经济发展势头使得他们有着乐观的未来预期,故而能充分激发出有效需求。福建、天津和江苏等地,收入水平较高,但还未能充分引导居民消费,有效需求有待进一步提升。

五、结束语

通过面板数据聚类分析方法的具体应用,发现原有的聚类分析方法只能解决静态问题,进行面板数据聚类分析的研究,不仅可以弥补聚类分析的理论,而且可以从动态的角度描述事物的类别,进一步对实际问题进行深入的研究。面板数据的聚类分析只是从统计学的角度研究面板数据的一小部分内容,而且聚类分析的方法和思路也不能局限于此,还有多指标面板数据的聚类分析,不同时间间隔情形的聚类分析问题,这也是正在研究的内容之一。

参考文献

[1] Hsiao C., Analysis of Panel Data [M]. 北京:北京大学出版社, 2005. 21 —92.

[2] Hsiao T. P. and Chih Y. Y. Comparison of Linear and Nonlinear Models for Panel Data Forecasting: Debt Policy in Taiwan [J]. Review of Pacific Basin Financial Markets and Policies, 2005(3): 525 —541.

[3] Bonzo D. C. and Hermosilla A. Y. Clustering Panel Data via Perturbed Adaptive Simulated Annealing and Genetic Algorithms [J]. Advances in Complex Systems, 2002(4): 339 —360.

[4] Ramsay J. O. and Silverman B. W. Functional Data Analysis [M]. New York: Springer-Verlag New York, Inc, 1997. 11 —83.

作者简介

朱建平,河南省,1962 年生,男,2003 年毕业于南开大学数学科学学院统计学系,获理学博士学位,现为厦门大学经济学院教授,博士生导师,计划统计系副主任,主要研究方向为数理统计、数据挖掘。

陈民恳,浙江省,1982 年生,男,厦门大学经济学院计划统计系硕士研究生,研究方向为多元统计与数据挖掘。

(责任编辑:李峻浩)