

SAS 软件在对农村居民家庭消费评析中的应用

刘 闯 千晓蓉*

(昆明理工大学理学院,昆明 650093)

摘 要 在现代各种统计数据的分析中,各种软件程序扮演了越来越重要的角色,而 SAS 是其中最重要的软件之一。利用 SAS 软件,提出用主成分分析的方法来对各地区农村综合消费水平排序,从而在一定程度上反映了各地区农村经济发展,综合实力状况。

关键词 SAS 主成份分析 农村居民家庭的综合消费水平 聚类分析

中图法分类号 TP391.3; **文献标志码** A

消费是社会再生产的重要组成部分,离开了消费,社会再生产便无法继续进行。消费既是社会再生产的起点,也是社会再生产的终点。消费水平作为消费的重要内容之一,是指国民在一年内平均消费的商品和劳务价值额,同时也可以用来指称国民的消费总规模,即社会总消费。研究消费水平,对确定社会生产中的积累与消费的比例和确定社会经济发展的战略具有重要意义。

自从我国改革开放以来,各地区经济蓬勃发展,展现出一派欣欣向荣的景象。各地区农村居民的生活水平已得到明显的提高,为更好的贯彻党中央的方针,建立一个更加和谐的社会,研究农村居民消费水平,确立发展农村的相关战略已是当务之急。在我国由于广大农村自然条件不同,生产力布局不同以及对某些地区采取“倾斜”政策和劳动差别和非劳动因素造成经济发展水平不同,劳动报酬不同,从而形成了消费水平的差异。因此有必要对各农村居民的综合消费水平做个评价,以其为今后的经济发展提供参考。

各个地区农村居民的消费指标主要是食品支

出、衣着支出、居住支出、家庭设备及服务支出、交通和通讯支出、文教娱乐用品及服务支出、医疗保健支出、其它商品及服务支出。本文利用 SAS 软件采用主成份分析对我国 31 个省,自治区,直辖市农村消费指标进行了分析,提出了使用第一成份来评价各个地区的综合消费水平的方法。

1 分析方法的理论介绍

主成份分析概念首先由 karl pason 在 1901 年提出,在 1933 年 Hotelling 将这个概念进行了推广。

主成份分析^[1]是指将多指标化为少数几个综合指标的一种统计分析方法。在实际问题中,研究多变量(多指标)问题是我们经常遇到的问题。因为变量个数太多,并且彼此之间存在着一定的相关性,因而使得所观测到的数据在一定程度上反映的信息有所重叠。而且当变量较多时,在高维空间中研究样本的分布规律比较复杂,势必增加分析问题的复杂性。因此人们希望用较少的综合变量来代替原来较多的变量,同时这几个综合变量又能够尽可能多地反映原来变量的信息,并且彼此之间不相关。主成份分析就是利用这种降维的思想来解决

问题,其关键在于消除评价体系中各个指标间的相关性。

主成份分析法的相关步骤^[2]如下:

(1) 原始统计数据的标准化处理:

2010 年 6 月 4 日收到

第一作者简介:刘 闯,男,湖南怀化人,硕士,研究方向:统计模式识别。

* 通讯作者简介:千晓蓉,女,四川省成都市人,教授,研究方向:统计模式识别。

$$x_{ij}^* = \frac{x_{ij} - E(x_j)}{\sqrt{\text{var}(x_j)}} (i = 1, 2, \dots, n; j = 1, 2, \dots, m)。$$

其中: x_{ij} 为第*i*个地区第*j*个指标的观测值; $E(x_j)$ 为第*j*个指标的样本均值; $\sqrt{\text{var}(x_j)}$ 为第个*j*指标的标准差。 x_{ij}^* 为标准化以后的指标值。记 $A = (x_{ij}^*)_{n \times m}。$

(2) 计算矩阵 $(x_{ij}^*)_{n \times m}$ 的相关系数矩阵 $R, R = \frac{1}{n-1}A \cdot A$,并计算 R 的特征值: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq 0$,以及对应的正交化特征向量 $u_1, u_2, \dots, u_m$, 其中向量 $u_j = (u_{j1}, u_{j2}, \dots, u_{jm}), j = 1, 2, \dots, m。$

(3) 计算特征值的累计贡献率: $Q = \frac{\sum_{k=1}^p \lambda_k}{\sum_{k=1}^m \lambda_k}$, 取

当 $Q \geq 0.85$ 成立时的最小的整数 P 的值作为主成分的个数。

(4) 提出 P 个主成份: $z_j = \sum_{k=1}^m u_{jk}x_k, j = 1, 2, \dots, p。$

2 具体计算以及结果分析

全国 31 个省,自治区,直辖市附近的农村居民各消费支出作为样本,设 X_1 为食品支出, X_2 为衣着指出, X_3 为居住支出, X_4 为家庭设备及服务支出, X_5 为交通和通讯支出, X_6 为文教娱乐用品及服务支出, X_7 为医疗保健支出, X_8 为其它商品及服务支出,这样就得到了 31 行 8 列的原始数据集 a ,相关数据来源于中国统计年鉴(2008)。

(1) 利用 SAS 软件先将数据集 a 标准化,再求出相关系数阵,其中各元素的值如表 1 所示。

表 1 标准化的数据集								
	X1	X2	X3	X4	X5	X6	X7	X8
X1	1							
X2	0.642 9	1						
X3	0.855 5	0.770 5	1					
X4	0.903 1	0.830 9	0.912 2	1				
X5	0.875 2	0.869 5	0.900 2	0.935 7	1			
X6	0.786 9	0.853 8	0.809 0	0.890 9	0.942 5	1		
X7	0.686 9	0.889 0	0.782 5	0.822 8	0.895 1	0.887 2	1	
X8	0.905 1	0.715 3	0.860 1	0.864 5	0.858 9	0.787 6	0.689 3	1

(2) 计算相关系数矩阵的特征值及贡献率,如表 2。

表 2 系数矩阵特征值及贡献率				
	特征值	主成分分量的方差	贡献率	累计贡献率
1	6.864 542 33	6.289 429 97	0.848 1	0.848 1
2	0.575 112 36	0.406 194 44	0.081 9	0.930 0
3	0.168 917 92	0.023 951 36	0.021 1	0.951 1
4	0.144 966 57	0.046 091 55	0.018 1	0.969 2
5	0.098 875 02	0.015 057 50	0.012 4	0.981 6
6	0.083 817 53	0.040 924 54	0.010 5	0.992 0
7	0.042 892 99	0.022 017 70	0.005 4	0.997 4
8	0.020 875 28		0.002 6	1.000 0

(3) 取累计贡献率 $Q = 0.93$, 即 $P = 2$, 也就是说可以 2 个主成分 z_1, z_2 来代替 8 个变量。而对于这两个主成份对应的特征向量如表 3。

表 3 主成份的特征向量								
变量	X1	X2	X3	X4	X5	X6	X7	X8
z_1	0.343	0.338	0.355	0.369	0.375	0.359	0.343	0.344
z_2	-0.504	0.487	-0.197	-0.109	0.055	0.221	0.478	-0.423

(4) 下面提出两个主成份,并计算第一主成分的值。
第一主成份:

$$z_1 = 0.343X_1 + 0.338X_2 + 0.355X_3 + 0.369X_4 + 0.375X_5 + 0.359X_6 + 0.343X_7 + 0.344X_8。$$

第二主成份:
$$z_2 = (-0.504)X_1 + 0.487X_2 + (-0.197)X_3 + (-0.109)X_4 + 0.055X_5 + 0.221X_6 + 0.478X_7 + (-0.423)X_8。$$

对各主成份的系数进行分析。因为第一主成份的系数都是正的,且数值相差不大,所以认为第一主成分可以代表我国农村居民的综合消费水平。比较第一主成分的各系数, X_4, X_5 的系数相对较大一些,说明家庭设备及服务支出与交通和通讯支出相对其它六种支出而言更能影响农村的综合消费水平。按第一主成份的值进行排序,见表 4,我们能够知道,农村居民家庭综合消费水平最高的属上海, z_1 的值为 9.245 24,北京居于第二位,浙江次之;但贵州,甘肃,西藏的综合消费水平排名最后。

从第二主成份的公式可以看出,它在 X_1 和 X_8 上有较大的负值,在 X_2 和 X_7 上有较大的正值。它意味着较小的食品支出和商品及服务支出,较多的衣着支出和医疗保健支出将获得较大的数值。通过 z_2 可粗略地认为我国农村居民的温饱状况已得到明显的改善,但需加大经济投入建立更加健全的医疗保健体系。

表 4 z_1 的排序		
按第一主成分之值排名	地区名	z_1
1	上海	9.245 24
2	北京	5.954 06
3	浙江	5.795 05
4	江苏	2.479 74
5	福建	0.931 76
6	广东	0.848 36
7	山东	0.543 90
8	辽宁	0.448 68
9	天津	0.330 82
10	内蒙古	0.011 14
11	吉林	-0.121 31
12	黑龙江	-0.235 72
13	湖南	-0.269 78
14	湖北	-0.431 92
15	山西	-0.822 21
16	河北	-0.892 28
17	江西	-1.037 38
18	河南	-1.074 10
19	安徽	-1.091 79
20	宁夏	-1.187 29
21	陕西	-1.235 04
22	四川	-1.312 96
23	青海	-1.399 95
24	新疆	-1.585 03
25	广西	-1.645 64
26	重庆	-1.698 58
27	云南	-1.721 74
28	海南	-1.788 83
29	西藏	-1.838 57
30	甘肃	-2.400 50
31	贵州	-2.798 14

定义样本间的距离为欧式距离,利用 SAS 软件采用聚类分析中离差平方和法,对我国各地区进行聚类分析^[3],能将全国 31 个省,自治区,直辖市大致分成三类,结果如表 5。

表 5 聚类分析			
类别	第一类	第二类	第三类
该类所包含的地区	河北,河南,广西,云南,重庆,四川,青海,宁夏,安徽,江西,新疆,甘肃,贵州,山西,海南,陕西,西藏	湖北,湖南,内蒙古,吉林,天津,黑龙江,辽宁,福建,广东,山东,江苏	北京,浙江,上海
食品平均消费	1 152. 96	1 522. 77	2 607. 53
衣着平均消费	161. 23	225. 61	464. 76
居住平均消费	439. 88	595. 80	1 539. 64
家庭设备及服务平均消费	116. 60	154. 89	376. 78
交通和通讯平均消费	241. 02	385. 92	815. 07
文教娱乐用品及服务平均消费	207. 85	364. 30	826. 09
医疗保健平均消费	165. 51	245. 76	551. 02
其它商品及服务平均消费	52. 79	93. 72	167. 68

从分类结果可以看出,第一类的地区主要在西部,因此可以称为西部地区,发展稍慢,农村居民的家庭消费水平较低;第二类地区主要分布在中部,东部,我们可以称其为中东部地区,经济发展中等,农村居民的家庭消费水平中等偏上;第三类地区主要在沿海,因此称为沿海地区,发展最快,农村居民的消费水平最高,他们已基本进入小康社会。因此国家在刺激中东部地区,沿海地区农村经济发展的同时,更要加大对西部偏远农村的经济扶持力度,继续采取一定的“倾斜”政策,这样,我们将会建立一个更加和谐美好的社会。分类结果与主成份分

析法得出的大体一致。

3 结论

对各地区农村居民家庭消费评析是一种典型的多指标的问题,使用主成份分析法,利用较少的变量来代替原来的众多指标。以第一主成份的数值来作为农村综合消费水平的度量,能够反映原始变量的主要信息,结果可靠。加上第二主成分,可以对各个地区的农村居民综合消费水平有了比较清楚的了解,从而进一步了解到各个地区农村的经济发展,综合实力状况,其分析结果与聚类分析的结果大体一致。通过观察数据的分析过程,我们也很容易发现 SAS 的计算速度快,运算结果系统全面等优点,这样就能大大节省了我们分析多指标问题的时间,提高了工作的效率。

附录

主成份分析实现的 SAS 程序如下:

(1) 建立数据集^[4]

```
Data a;  
Input name MYM x1-x8;  
Cards;  
地名 数据  
Run;
```

(2) 进行主成份分析,产生相关系数矩阵,特征值及特征向量。

```
Proc princomp data=a out=acomp;  
Var x1-x8;  
Run;
```

(3) 对产生的第一主份值进行排序,如下表 D

```
Proc sort data=acomp;  
By prin1;  
Run;  
Proc print ;  
Id name ;
```

```
Var prin1;  
Run;
```

聚类分析的 SAS 程序^[5]:

(4) 对原始数据集 *a* 做标准化变换,采用离差平方和法进行聚类分析:

```
proc cluster data=a outtree=olunwen method=ward  
std pseudo ccc;  
id group;  
var x1-x8;  
run;
```

(5) 对(4)过程生成的数据集 olunwen 绘制水平方向的树状输出结构图

```
Proc tree data=olunwen horizontal graphics n=3  
out=hlunwen;  
copy group x1-x8;  
run;
```

(6) 对(5)生成的数据集 hlunwen 按照类、地区、X1 到 X8 的各指标的输出生成分类结果

```
proc sort data=hlunwen;  
by cluster;  
run;  
proc print data=hlunwen;  
var cluster group x1-x8;  
run;
```

(7) 将我国各地区分为三类,按分类顺序计算出各地区农村居民家庭的平均消费水平。

```
proc means data=hlunwen;  
by cluster;  
var x1-x8;  
run;
```

参 考 文 献

1 张尧庭,方开泰.多元统计分析引论.北京:科学出版社,1982

2 于秀林,任雪松.多元统计分析.北京:中国统计出版社,1999

3 高惠璇.应用多元统计分析.北京:北京大学出版社,2005

4 汪嘉冈.SAS V8 基础教程.北京:中国统计出版社,2003

5 邓祖新.数据分析方法与 SAS 系统.上海:上海财经大学出版社,2006

The Application of SAS Software on the Rural Regional Household Consumption Levels

LIU Chuang, GAN Xiao-rong *

(Faculty of Science, Kunming University of Science and Technology, Kunming 650093, P. R. China)

[**Abstract**] Now ,the variety of software programs play an important role in the analysis of the different kinds of modern statistical data. While the SAS is one of the most important software. The SAS software is used and proposed sorting the comprehensive consumption levels of the various rural regions by using the method of the principal component analysis. To the extend ,it can reflect the rural economic development and comprehensive strength conditions of the various regions.

[**Key words**] SAS principal component analysis comprehensive household consumption levels of the rural region cluster analysis

(上接第 6292 页)

- 2 刘少辉. Rough 集高效算法的研究. 计算机学报, 2003;26(5): 524—529
- 3 苗夺谦, 胡桂荣. 知识约简的一种启发式算法. 计算研究与发展, 1999; 36(6): 681—684

Attribute Reduction Algorithm Based on SQL

LIU Jing-lian

(Computer Department, Suihua Institute, Suihua 152061, P. R. China)

[**Abstract**] A theorem is given about whether or not an attribute is necessary for a certain attribute sets, and the corresponding algorithm is given based on SQL. Based on these research results, an improved attribute reduction algorithm based on rough set is presented.

[**Key words**] rough set attribute reduction algorithm SQL