

重庆社会发展状况的因子分析

谭 婕¹ 王 蕾^{1,2}

(1. 重庆科技学院, 重庆 400050; 2. 重庆师范大学, 重庆 400047)

【摘 要】:利用多元统计分析中的因子分析方法,使用统计软件 SPSS 进行计算,对重庆下属的 40 个区县的社会发展状况进行了综合考察。

【关键词】:多元统计分析;因子分析;数学模型

【中图分类号】:F224:0212 **【文献标识码】**:A **【文章编号】**:1673-1980(2007)01-0126-03

多元统计分析是统计学的一个重要分支,随着电子计算机的普及和发展,其作用越来越大,在国民经济许多领域中有着广泛的应用,并已取得了许多成果。实践证明,多元统计分析方法是处理数据不可缺少的重要工具,并日益显示出无比的魅力。

1 基本思想及数学模型

因子分析法是将具有错综复杂关系的变量(或样品)综合为数量较少的几个因子,以再现原始变量与因子之间的相互关系,同时根据不同因子对变量进行分类,是属于多元分析中处理降维的一种统计方法。

因子分析的数学模型为:

$$\begin{cases} x_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1n}F_n + \varepsilon_1 \\ x_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2n}F_n + \varepsilon_2 \\ \dots \dots \dots \dots \dots \dots \\ x_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pn}F_n + \varepsilon_p \end{cases}$$

其中 $x_1, x_2, \dots, x_p$ 是 p 个变量; $F_1, F_2, \dots, F_n$ 是 n 个公因子,它们是对各变量都有影响的公因子,是潜在的因素,不容易直接观察或者不容易直接测量。 $\varepsilon_i (i=1, 2, \dots, p)$ 为特殊因子,是指对特定变量 $x_i (i=1, 2, \dots, p)$ 起作用的因子,在实际模型中往往忽略不计。 a_{ij} 称为因子载荷,是第 i 个变量在第 j 个因子上的载荷,某一变量在某一因子上的载荷愈大,则它与该因子的相关程度愈高。

令 $a_j^2 = \sum_{i=1}^p a_{ij}^2 (j=1, 2, \dots, p), \frac{g_j}{p}$ 是第 j 个公因子 F_j 的方差贡献率,表示第 j 个公因子 F_j 所反映的信息量占总体信息量

¹收稿日期:2006-12-10

作者简介:谭婕(1980-),女,重庆人,重庆科技学院数理系教师。

的比例。

$$\sum_{i=1}^m \frac{g_i^2}{p}$$

是前 $m(m \leq n)$ 个公因子 $F_1, F_2, \dots, F_m$ 的累积方差贡献率, 表示前 m 个公因子所反映的信息量占总体信息量的比例。

根据变量的相关关系选出第一主因子 F_1 , 它在各变量的公因子方差中所占的方差贡献率最大。然后消去这个因子的影响, 从剩余的相关关系中, 选出与 F_1 不相关的因子 F_2 , 使其在各变量的剩余因子方差贡献中最大。

用此方法分解完各个变量的公因子方差, 使 F_1 在总方差中所占的比重最大, F_2 次之, 其余因子 $F_3, \dots, F_n$ 的方差贡献率依此递减。

在实际研究中, 根据累积方差贡献率大于等于 80% 的原则, 选取前几个方差贡献大的因子, 舍弃那些方差贡献较小的因子。这样就减少了指标的数目, 抓住了主要矛盾。

对于所得到的各因子, 首先观察它们在哪些变量上的载荷较大, 在哪些变量上载荷较小。再根据载荷大的变量本身的内容来说明因子的含义。

本文用主成分估计参数法求初始因子载荷阵, 并对载荷阵作最大方差正交旋转, 载荷值越大, 表明变量能为该因子的解释提供的信息量越多。

2 数据处理及因子分析

下面我们根据《中国 2005 年统计年鉴》所得到的数据, 用多元统计方法中的因子分析方法就重庆下属的 40 个区县的社会发

展状况进行综合考察。对原始数据矩阵 X 进行标准化, 即 $\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$, 其中 x_j 和 s_j 分别是第 j 个指标的样本均值和样本标准差,

得到标准化后的矩阵 $\tilde{X}$ 。标准化的主要目的是消去量纲和数量级的影响, 建立 $\tilde{X}$ 的相关系数阵 R 见表 1。

表 1 相关系数矩阵

	人均 GDP	全社会固定资产投资	职工平均工资	商品零售价格指数	工业总产值	年末就业人员数
人均 GDP	1.000	0.540	0.856	0.121	0.742	-0.201
全社会固定资产投资	0.540	1.000	0.567	0.755	0.780	0.221
职工平均工资	0.856	0.567	1.000	0.246	0.728	-0.318
商品零售价格指数	0.121	0.755	0.246	1.000	0.458	0.167
工业总产值	0.742	0.780	0.728	0.458	1.000	-0.029
年末全部就业人员数	-0.201	0.221	-0.318	0.167	-0.029	1.000

应用 KMO and Bartlett's Test 得到 $KMO=0.686>0.5$ 且 Bartlett 球型检验 $P<0.05$, 可以进行因子分析。同时可从表 2 中看出各个变量的信息都提取得比较充分。

表 2 公因子方差率表

	初值	方差率
人均 GDP	1.000	0.843
全社会固定资产投资	1.000	0.935
职工平均工资	1.000	0.895
商品零售价格指数	1.000	0.713
工业总产值	1.000	0.848
年末全部就业人员数	1.000	0.644

下面计算各因子的特征值和累计贡献率：

表 3 主成分公因子方差分析

	初始特征值			提取的平方和负载量		
	总数	方差率	累计贡献率	总数	方差率	累计贡献率
人均 GDP	3.378	56.301	56.301	3.378	56.301	56.301
全社会固定资产投资	1.499	24.990	81.292	1.499	24.990	81.292
职工平均工资	0.700	11.669	92.960			
商品零售价格指数	0.213	3.548	96.508			
工业总产值	0.116	1.933	98.441			
年末全部就业人员数	0.094	1.559	100.000			

通过表 3 可见前两个特征值累计贡献率已达到 81.292%。同时,我们从图 1(就是按特征根大小排列的主成分散点图)也可以看出,从第三个主成分开始特征根都非常低,该图从另一个侧面说明了只需要提取两个主成分即可。

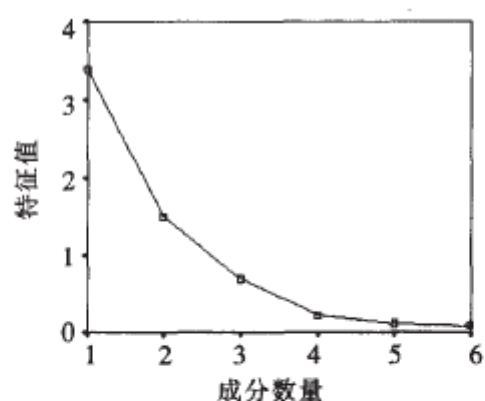


图 1 主成分散点图

表 4 成分组成矩阵

	成分组成	
	1	2
人均 GDP	0.830	-0.392
全社会固定资产投资	0.873	0.417
职工平均工资	0.859	-0.396
商品零售价格指数	0.580	0.613
工业总产值	0.921	0.007
年末全部就业人员数	-0.074	0.799

故取前两个特征值建立因子载荷阵, 如表 4 所示: 从表 4 中发现各个因子的典型代表指标的经济意义不是很突出, 难以得到合理解释, 故需对因子进行旋转。

这样因子解的实际意义更容易理解。因子经过 25 轮旋转后, 其载荷阵参见表 5, 旋转后两个主因子的累计方差贡献率仍为 81.292% 因子旋转不改变模型对数据的拟合。

表 5 旋转后成分组成矩阵

	成分组成	
	1	2
人均 GDP	0.918	0.017
全社会固定资产投资	0.597	0.761
职工平均工资	0.945	0.026
商品零售价格指数	0.248	0.807
工业总产值	0.822	0.415
年末全部就业人员数	-0.422	0.683

由表 5 可见, 每个因子只对应少数几个指标的因子载荷较大, 因此可根据上表对指标进行分类。将 6 项指标按高载荷分成两类: 第一类命名为实力因子, 从人均 GDP、职工平均工资、工业总产值 3 个方面反映人们经济状况, 这类因子基本反映了各地区经

济发展的水平和综合实力;第二类命名为资金因子,从全社会固定资产投资,商品零售价格指数和年末全部就业人员数 3 方面反映经济发展状况。

3 结论

根据因子载荷矩阵可以计算得到每个样本(即每个地区)相对于两个主因子的得分情况,然后再根据各因子的方差贡献率之比,对得分进行加权汇总,从而得到每个地区的总因子得分,按总得分进行排序。

第一主因子得分:

$$\tilde{f}_1=0.918\tilde{x}_1+0.597\tilde{x}_2+0.945\tilde{x}_3+0.248\tilde{x}_4+0.822\tilde{x}_5-0.422\tilde{x}_6$$

第二主因子得分:

$$\tilde{f}_2=0.017\tilde{x}_1+0.761\tilde{x}_2+0.026\tilde{x}_3+0.807\tilde{x}_4+0.41\tilde{x}_5+0.683\tilde{x}_6$$

以各因子的贡献率占两个主因子的贡献率的比例作为权重,构造综合因子得分为:

$$\tilde{f}=\frac{56.301}{81.292}\tilde{f}_1+\frac{24.990}{81.292}\tilde{f}_2$$

对标准化数据的主因子得分 $\tilde{\mathbf{x}}=(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_6)$ 和综合得分进行排名,见表 6。

表 6 计算的各地市主因子得分和综合得分及得分排名

地区名称	按第一因子 得分排名	按第二因子 得分排名	综合得分	综合排名
渝中区	2	12	4.592 764 1	5
大渡口区	6	34	3.448 828 2	7
江北区	4	5	4.783 646 2	4
沙坪坝区	5	3	4.815 024 4	3
九龙坡区	1	2	5.737 167 1	2
南岸区	7	4	4.436 191 7	6
北碚区	9	16	1.150 245 7	9
渝北区	3	1	6.554 742 8	1
巴南区	10	7	1.538 018 3	8
万盛区	17	38	-1.212 827 3	22
双桥区	8	40	1.026 367 9	10
綦江县	19	15	-0.712 048 5	18
潼南县	25	18	-1.261 029 7	24
铜梁县	21	17	-1.023 116 4	20
大足县	22	21	-1.259 842 8	23
荣昌县	24	26	-1.498 562 2	26
璧山县	13	19	0.097 704 4	14
江津市	16	10	0.134 688 6	13

合川市	28	8	-0.746 952 2	19
永川市	12	13	0.350 229	12
南川市	20	24	-1.188 387 7	21
万州区	18	6	-0.075 869 2	15
涪陵区	11	9	0.745 814 4	11
黔江区	14	32	-0.317 675 4	17
长寿区	15	23	-0.189 762	16
梁平县	23	29	-1.569 889 8	27
城口县	31	39	-2.344 921 1	38
丰都县	33	27	-1.871 363 3	32
垫江县	26	30	-1.597 164 1	28
武隆县	30	35	-1.945 081 8	34
忠 县	34	20	-1.728 258 8	30
开 县	37	11	-1.751 690 8	31
云阳县	40	14	-2.108 975 5	35
奉节县	36	25	-2.211 699 1	36
巫山县	35	36	-2.319 504 5	37
巫溪县	39	37	-2.731 317 7	40
石柱县	32	31	-1.922 018 8	33
秀山县	29	33	-1.723 780 9	29
酉阳县	38	28	-2.396 528 7	39
彭水县	27	22	-1.469 468	25

从表 6 可以看出第一因子得分排名和综合因子排名差距较小,说明第一因子对衡量各区县经济发展水平起着主导作用。

同时,重庆市各区县经济发展很不平衡,不发达地区相对较多,都市经济圈实力较强;而三峡库区基础薄弱,资金投入不多,致使排名基本上靠后。这表明政府应加大库区建设,抓住西部大开发这样的历史性机遇,加速库区的发展。

参考文献

[1]于秀林,任雪松.多元统计分析[M]北京:中国统计出版社,1999.

[2]中华人民共和国统计局.中国统计年鉴-2005[M],北京:中国统计出版社,2005.

[3]张文彤.SPSS11 统计分析教程[M],北京:希望电子出版社,2002.