广东省社会综合发展水平研究

——基于 2017 年统计年鉴数据的因子分析

LPPL 小组：张昀 康国钰 唐晔

指导老师：蒋学军

[摘要]：广东省位于南岭以南，南海之滨，是中国经济最发达的省份。省会广州，辖 21 个省辖市，其中副省级城市 2 个（广州、深圳），地级市 19 个。广东珠三角 9 市将联手港澳打造粤港澳大湾区，成为世界四大湾之一。本文采用因子分析方法对广东省 2017 年 21 个市的 11 项指标进行社会经济发展水平分析，结果表明：广州和深圳是广东省经济社会文化的中心，东莞，珠海，佛山，中山等城市的综合实力也不容小觑,此外虽然揭阳，茂名，湛江等的经济水平相对落后，然而教育文化水平排名却不低。

[关键词]：广东省，社会综合发展，因子分析

一、研究方法

1、因子分析

衡量一个地区的社会经济发展水平，可以对促进社会发展的多个方面进行综合分析，得到区域（市）经济发展的多重影响因素，如科教、文化娱乐、医疗卫生、经济发展等方面。多个变量存在中等程度的相关性，可以通过降维的思路，减少用于分析的变量数目，并归纳为潜在因子来描述原始变量间的相关性。因子分析能够有效处理相关性强的多变量数据，归纳出简洁便于理解的因子，揭露事物间最本质的联系。

1.1 因子分析原理：

$\mathrm{X} = \mu + \mathrm{Af} + \varepsilon$，其中，X 的均值$\mu = (\mu_1, \ldots, \mu_p)^T$；公共因子为$\mathrm{f} = (\mathrm{f}_1, \ldots, \mathrm{f}_m)^T$ ，

a_{ij}为原始变量x_i在因子f_j上的载荷，反映因子f_j与变量x_i的贡献；特殊因子ε_i称为误差或特殊因子，是公共因子无法解释的部分。p 个原始变量由 m 个公共因子线性表出。

1.2 模型假设：

（1）公共因子互不相关，线性意义上信息互不重复；

（2）特殊因子无法被公共因子 f 解释，故二者不相关，

（3）特殊因子互不相关，因此公共因子可以解释原始变量的相关性，

用数学语言表示上述假设：

$$\begin{cases} E(f)=\mathbf{0} \\ V(f)=\mathbf{I} \\ E(\boldsymbol{\varepsilon})=\mathbf{0} \\ V(\boldsymbol{\varepsilon})=\boldsymbol{D}=\operatorname{diag}\left(\sigma_1^2,\sigma_2^2,\cdots,\sigma_p^2\right) \\ \operatorname{Cov}(\boldsymbol{f},\boldsymbol{\varepsilon})=E\left(\boldsymbol{f}\boldsymbol{\varepsilon}'\right)=\mathbf{0} \end{cases}$$

其中，对数据进行标准化后，$E(X)=\mu=0$; $V(X)=I$; $D=I$，此时 X 的协方差矩阵就是相关矩阵 R，本文中使用 python 实现的因子方差均是基于标准化后的数据。

$$a_{ij}=Cov(x_i,f_j)=\frac{Cov(x_i,f_j)}{\sqrt{V(x_i)}\sqrt{V(,f_j)}}=\rho(x_i,f_j),\mathrm{i}=1,2,\ldots,\mathrm{p},\mathrm{j}=1,2,\ldots,\mathrm{m}$$

1.3 因子分析模型:

$$\begin{cases} x_1=\mu_1+a_{11}f_1+a_{12}f_2+\cdots+a_{1m}f_m+\varepsilon_1 \\ x_2=\mu_2+a_{21}f_1+a_{22}f_2+\cdots+a_{2m}f_m+\varepsilon_2 \\ \cdots \\ x_p=\mu_p+a_{p1}f_1+a_{p2}f_2+\cdots+a_mf_m+\varepsilon_p \end{cases}$$

用矩阵表示简记为：$\mathbf{X}=\mathbf{AF}+\boldsymbol{\varepsilon}$，再利用回归$\boldsymbol{\varepsilon}$，再利用回归估计的方法将各因子用 X 线性展开，可进一步计算出因子得分，公式表示为：

$$f_j=b_{j1}X_1+b_{j2}X_2+\cdots+b_{jp}X_p,\mathrm{j}=1,2,\ldots,\mathrm{m}$$

2、构建指标体系--原始变量

本文使用《2018 广东统计年鉴》中调查的 2017 年数据，涉及对外经济、住

宿餐饮业和旅游、教育和科技、文化和体育、卫生社会福利及社会保障和其他、县（市、区）主要经济指标六个领域，选取 11 个指标：X1 地区生产总值（万元）；X2 地方一般公共预算收入；X3 社会消费品零售总额；X4 出口总额；X5 各式国际旅行外汇；X6 社会研究与实验发展经费；X7 图书馆数目；X8 中学数目；X9 医院床位数；X10 卫生技术人数；X11 城乡基本医疗保险参保人数。

二、因子分析过程

1、KMO 检验和 Bartlett 检验

KMO （Kaiser-Meyer-Olkin）统计量是取值在 0 和 1 之间。当所有变量间的简单相关系数平方和远远大于偏相关系数平方和时，KMO 值接近 1。KMO 值越接近于 1，意味着变量间的相关性越强，表明原始变量越适合做因子分析。在实践中，KMO 的值在 0.7 以上时适合做因子分析。Bartlett 检验 p-value<0.05，统计学意义下显著，可以进行因子分析。如下表 1 为本案例的检验结果，可以看出本 KMO=0.794>0.7，且 SIG 接近于 0，远小于 0.05。因此，该数据是适合做因子分析的。

取样足够度的 KMO 度量		0.794
Bartlett 的球形度检验	近似卡方	561.305
	显著性	2.346E-85

表 2.1 KMO 和 Bartlett 的检验

2、提取公共因子的结果

图 2. 1 碎石图表示原始数据相关系数矩阵的特征值，从图中可以清楚看出，只有前两个因子的特征值大于 1，但是从第四个特征值开始，特征值趋于 0。因此可以考虑选取前三个或者四个因子进行因子分析。

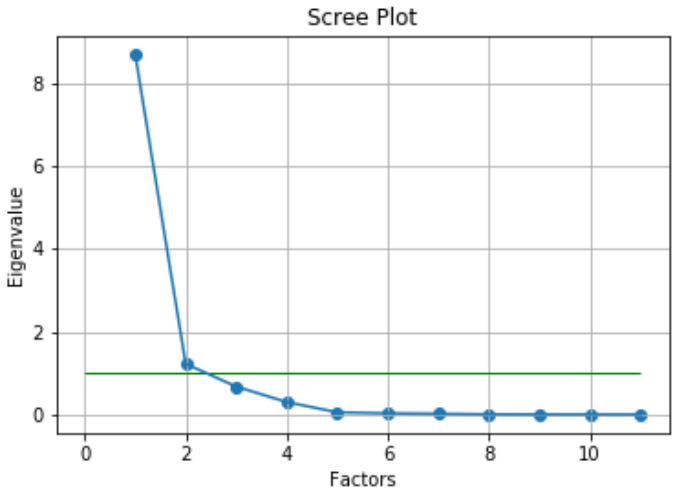

图 2.1

表 2.2 为未进行因子旋转的正交因子模型的因子载荷矩阵（原始变量与前 4 个公共因子）。其中载荷矩阵每一列的元素绝对值没有明显分别，如第一列，几乎所有载荷都＞0.75，然而第二列第三列没有载荷>0.5。可见，目前的因子载荷矩阵无法对因子进行合理解释。因此考虑进行因子旋转。

表 2.2 旋转前的正交因子模型载荷矩阵

旋转前 - 因子载荷(前4个)		F1	F2	F3	F4
X1	地区生产总值(万元)	0.994166	0.074719	-0.049846	-0.017524
X2	地方一般公共预算收入	0.901298	0.428215	0.027649	0.010581
X3	社会消费品零售总额	0.974514	-0.201368	-0.050787	0.026842
X4	出口总额	0.84078	0.485218	-0.029944	-0.159266
X5	各式国际旅行外汇	0.970776	-0.027275	-0.152986	0.121975
X6	社会研究与实验发展经费	0.931611	0.359741	-0.008415	0.006109
X7	图书馆	0.446014	-0.157455	0.471874	0.482199
X8	中学数目	0.774857	-0.312469	0.494879	0.062228
X9	医院床位数	0.902647	-0.421336	-0.010747	-0.053537
X10	卫生技术人数	0.956101	-0.284838	0.007919	0.002848
X11	城乡基本医疗保险参保人数(万人)	0.905927	0.060309	0.397423	-0.030485

表 2.3 表示进行因子旋转后，各个因子的方差贡献率和以及累积方差贡献率。累积贡献率用来衡量公共因子对原始变量的重要性，是公共因子解释总方差的比例。表中显示前四个公共因子的方差贡献率分别为 40.03%、36.26%，13.00%，

7.4%；前两个因子的累积贡献率为 76.3%，前三个因子的累积贡献率为 89.3%，前四个因子的累积贡献率为 96.7%。从方差贡献率角度可以选择三个公共因子。表 2.4 为根据旋转后的因子载荷矩阵，前三个因子可以较容易的给出解释，而第四个因子难以找到合理解释。因此，提取前三个因子为公共因子。

表 2.3 旋转后的因子方差贡献率

SS Loadings	4.403647	3.988767	1.430808	0.81631	0.049983	0.017777
Proportion Var	0.400332	0.362615	0.130073	0.07421	0.004544	0.001616
Cumulative Var	0.400332	0.762947	0.89302	0.96723	0.971774	0.97339

表 2.4 旋转后的因子载荷矩阵

旋转后 - 因子载荷(前3个)		F1	F2	F3
X1	地区生产总值(万元)	0.72102	0.645728	0.191867
X2	地方一般公共预算收入	0.913372	0.324044	0.211571
X3	社会消费品零售总额	0.49975	0.809883	0.226495
X4	出口总额	0.94107	0.275997	-0.013388
X5	各式国际旅行外汇	0.628711	0.721264	0.20568
X6	社会研究与实验发展经费	0.88536	0.401723	0.20196
X7	图书馆	0.108978	0.218567	0.833471
X8	中学数目	0.249349	0.557842	0.464685
X9	医院床位数	0.29632	0.894753	0.234264
X10	卫生技术人数	0.427719	0.831753	0.273535
X11	城乡基本医疗保险参保人数(万人)	0.62468	0.449173	0.417557

3、公共因子命名的结果

表 2.4 为使用正交旋转法旋转后的三个公共因子载荷值。一般而言，因子载荷值越大，表明原始变量x_i与公共因子f_j的相关程度越高，公共因子f_j能够解释的信息量越大。

第一个公共因子在 X1、X2、X4、X6 上有较高的载荷值，在 X3、X5 上有中等大小的载荷值，根据以上指标的性质，可以讲第一个公共因子命名为“经济发展

水平因子”（F1）。第二个公共因子在 X9、X10 上有较大载荷值，均在 0.8 以上，根据这两个指标的性质，将第二个公共因子归纳为“医疗卫生水平因子”（F2）。第三个公共因子在 X7、X8 上有较高的载荷值，因此可以讲第三个公共因子归纳为“科教水平因子”（F3）。

4、因子得分和综合排名结果

表 2.5 显示广东省 21 市分别在三项主因子的得分情况。计算结果见下表：

表 2.5 广东省 21 市的因子得分

	F1	排名	F2	排名	F3	排名
广州市	-0.073775	7	3.991805	1	1.378399	1
深圳市	4.238835	1	-0.565188	17	1.121744	2
珠海市	0.228229	3	0.080221	5	-1.522783	20
汕头市	-0.543008	19	-0.40273	13	-0.137634	15
佛山市	0.227957	4	1.116618	2	-0.844435	18
韶关市	-0.330259	14	-0.353507	12	0.832785	5
河源市	-0.326245	13	-0.635841	21	0.311549	9
梅州市	-0.453536	17	-0.615049	19	0.988312	3
惠州市	-0.02274	6	-0.328259	11	0.163111	11
汕尾市	-0.274944	12	-0.63099	20	-0.015582	13
东莞市	0.627205	2	0.821098	3	-2.981238	21
中山市	0.163668	5	0.144064	4	-1.170052	19
江门市	-0.223664	10	0.006056	6	0.407987	8
阳江市	-0.250178	11	-0.241418	9	-0.126986	14
湛江市	-0.74776	21	-0.054528	7	0.919029	4
茂名市	-0.615503	20	-0.123923	8	0.540334	7
肇庆市	-0.351417	15	-0.315215	10	0.599845	6
清远市	-0.365091	16	-0.405102	14	0.286378	10
潮州市	-0.204829	9	-0.499296	16	-0.594545	17
揭阳市	-0.505799	18	-0.572556	18	-0.00891	12
云浮市	-0.197146	8	-0.41626	15	-0.147307	16

三、因子分析结果评价

从经济发展水平指标的排名可以看出，广州，深圳，珠海，东莞，佛山，中山的经济发展水平较高，东莞市有发达的制造业，其 2017 年出口总额占 21 个城市中的第二名，导致它的第一个因子的得分很高；从医疗发展水平来看，广

州，佛山，东莞，中山的排名靠前；从教育发展水平指标的排名可以看出广州，深圳，梅州，湛江处于前四名。

通过以上数据指标与数据分析结果，可以看出，广州和深圳是广东省经济社会文化的中心．东莞，珠海，佛山，中山等城市的实力也不容小觑．医疗卫生水平是深圳经济社会发展的短板。从分析结果中也可以看出，广东省省内经济发展水平差别很大，有些城市如揭阳，茂名，湛江等的经济水平相对落后。从结果中也可以看出，某些经济不发达的城市的教育发展水平很高。

参考文献

[1] 吴磊. (2015). 珠江—西江经济带社会经济发展水平研究——基于2013年数据的因子分析. 商场现代化, (28), 140-141.

[2] Introduction to Factor Analysis in Python. (2019). Retrieved from https://www.datacamp.com/community/tutorials/introduction-factor-analysis