基于 STATA 的 FA-DEA 与 PCA-DEA①

彭江平

(湖南大学 工商管理学院, 长沙 410082)

摘 要: 针对应用 FA-DEA 与 PCA-DEA 模型时, 一般都需要同时使用统计分析软件与数据包络分析软件, 提出了在 STATA 中的实现过程与方法, 并进行了案例分析. 方便了在实际应用过程中使用 FA-DEA 与 PCA-DEA 的过程, 并为在统一 STATA 的环境下设计与应用新的数据包络法提供解决思路.

关键词: DEA; FA-DEA; PCA-DEA; STATA

Implementing FA-DEA and PCA-DEA in STATA

PENG Jiang-Ping

(Business School, Hunan University, Changsha 410079, China)

Abstract: In general, the dimension reduction methods, such as FA and PCA, are implemented in statistical analysis software, but the data envelopment analysis is doing by DEA software. Simultaneously implementing FA-DEA and PCA-DEA all in STATA is puts forward, and an example is also listed. This approach will facilitate the process using FA-DEA and PCA-DEA in the practical application, and will provided a new solution in the STATA to design and application of data envelopment analysis.

Key words: DEA; FA-DEA; PCA-DEA; STATA

1 引言

数据包络分析(data envelopment analysis, 简称 DEA)自 1978 年由 A.Charnes、W.W.Cooper 与 Rhodes 创立已来, 已成为重要的比较多个具有相同投入要素与产出要素的绩效的重要方法. 本文将在分析实际应用过程中综合应用因子分析与主成分分析方法与标准数据包络法必要性基础上, 提出在 STATA 的统一数据分析环境下, 结合这两种数据分析方法的 FA-DEA 与 PCA-DEA 的综合数据包络法的实现与应用.

2 数据包络、数据降维法与STATA

2.1 数据包络法

Koopmans(1951)首先提出技术效率的概念: 如果在不减少其他产出(或增加其他投入) 的情况下, 技术上不可能增加任何产出(或减少任何投入) , 则该投入产出向量是技术有效的, 技术有效的所有投入产出向量的集合构成生产前沿面[1]. Farrell(1957)首次提出了技术效率的前沿测定方法, 并得到了理论界的广泛认同, 成为了效率测度的基础[2]. 在实际应用中, 前沿面的确定方法主要有两种: “参数方法”通过计量模型对前沿生产函数的参数进行统计估计, 对技术效率进行测定; “非参数方法”通过求解数学中的线性规划来确定生产前沿面, 并进行技术效率的测定. 数据包络分析(data envelopment analysis, 简称 DEA),是最常用的一种非参数前沿效率分析法, 由 1978 年 A.Charnes、W.W.Cooper 与 Rhodes[3](简称为 CCR)创建, 随后有许多学者对 CCR 模型进行了扩充与完善, 如 BCC 模型、FG 模型、ST 模型等经典模型、CCGSS 加法模型、具有无穷多个 DMU 的 CCW 模型[4,5].

2.2 数据降维法

数据降维通过简化数据结构达到降低维度, 把多个存在相关关系的指标化成少数几个互不相关的新的

① 基金项目:国家自然科学基金(71171076);湖南省社科基金(13YBA074)
收稿时间:2015-01-11;收到修改稿时间:2015-03-18

综合性指标. 新产生的主成分和因子变量最大程度上反映了原始指标的信息(涵盖量达到 80%以上), 之间互不相关, 去除了重叠信息, 个数又较少, 而且层次较高, 综合性较强, 使形成的新指标体系达到最优. 这类方法主要包括主成分分析法和因子分析法[6].

2.3 STATA

STATA 不仅可以实现诸多的统计分析方法, 如单元统计、多元统计等内容; 还包括了许多经典和前沿的计量模型, 如单方程回归模型、离散选择模型、分位数回归、时间序列分析、面板数据分析、蒙特卡洛模拟和自举法等. 作为本文主题的数据降维, 有相应的因子分析法 Factor 与主成分分析法 PCA.

3 基于STATA的FA-DEA与PCA-DEA

3.1 综合数据降维法与数据包络法

Ueda and Hoshiai (1997)首先将主成分分析(PCA)与数据包络法(DEA)整合起来, 克服相当 DMU 数而言有大量输入与输出因素时, DEA 标准 CCR 模型识别效率低下的困难, 同时也能在一定程度上消除随机波动对 DEA 分析的影响. 此后, 出现了大量文献通过用因子分析法或主成分分析法同时对投入指标与产出指标进行筛选和综合, 再采用 DEA 方法进行评价, 解决了 DEA 方法对指标数量限制的问题[7], 形成相应的 PCA-DEA、FA-DEA 系列模型.

DEA 事实上是一系列线性规划的求解, 而数据降维, 包括因子分析法与主成分分析法, 属于统计分析方法, 基于 FA-DEA 与 PCA-DEA 的实施过程中, 数据降维一般由标准的统计分析软件如 SAS、SPSS 与 STATA 等的相应算法完成, 而 DEA 部分由专业的数据包络分析软件包如 DEA Solver Pro、Warwick DEA、DEA Excel Solver 与 EMS 等. 需要在多个软件之间重复地输入与输出数据、互用中间结果信息. 既增加了学习的难度, 也难以避免转换过程中可能产生的问题.

3.2 基于 STATA 的统一实现

在标准的 STATA 软件包中, 没有相应的数据包络分析, 但自 V9.0 后引入了强大的 MATA 矩阵运算, 这为在 STATA 中应用 DEA 方法奠定了基础, 更方便的是, Lee, C., J. Lee(1997)已设计了基于 STAT 的数据包络软件包, 并在 SSC(Statistical Software Components)上提供了相应的安装包 st0193[8,9]. 基于 STATA 统一实现 FA-DEA 与 PCA-DEA, 主要过程包括步骤:

第一步, 分别对输入变量组与输出变量组进行数据降维处理. 数据降维的具体实现可以依据实际需要, 采用因子分析法或者主成分分析法.

第二步, 主成分分析或因子分析的产出数据的非负转换. DEA 方法要求投入产出数据均为正值, 而采用因子分析或主成分分析方法进行产出因子得分所得到的数据会有一部分为负值, 所以, 所得到的数据需要进行一定的处理再进行 DEA 分析. 具体计算方法如下:

$$X'_{ij}=0.1+\frac{X_{ij}-b_j}{a_j-b_j}\times 0.9 \quad 0.1\le X'_{ij}\le 1 \qquad (1)$$

式中, 为第 j 个指标的最大值, 为第 j 个指标的最小值, 是初始数据, 是通过变换所得到的数据. 这种变换可以将所有的数据变换为[0.1, 1]区间上的数据, 并不影响评价结果[4].

第三步, 基于数据降维及转换处理后的新变量, 实行标准的数据包络分析.

3.3 Stata 中 FA-DEA 与 PCA-DEA 实现

为方便读者在数据包络分析应用研究过程中使用, 简要给出在 STATA 中的实现.

第一步, 使用 STATA 中提供的标准命令 Factor(因子分析法)与 PCA(主成分分析法). 以主成分分析为例, 说明数据降维过程如下:

```
pca   in_1 in2 in_3 in_4 in_5
predict pin_p1 pin_p2
//依据特征根标准选择生成两个主成分变量
```

第二步, 因子或主成分非负转换的实现. 假定 pin_p1 是主成分析分析法得到的某个主成分变量, 则依据上面的公式(1), 转换实现如下:

```
sum(pin_p1)
scalar in_p1_max=r(max)
scalar in_p1_min=r(min)
gens_pin_p1=0.1+0.9*(pin_p1-in_p1_min)/(in_p1_
max-in_p1_min)
```

第三步, 数据包络分析. 安装数据包络分析的软件包“st0193”, 然后调用其中的 dea 方法. 大致实现过程如下:

```
net search st0193 //查询相应的软件包并安装
dea   in_1 in2 in_3 in_4 in_5 =   ou_1 ou_2 ou3
 //数据包络分析
```

4 应用案例分析

4.1 数据来源

案例数据集选择自 Hokkanen J and Salminen P (1997)及Sarkis J (2000)关于芬兰Oulu地区垃圾管理系统绩效评估管理的数据集, 包括五个输入变量、三个输出变量及 22 个 DMU[10,11]. 原始数据集由于篇幅原因略去, 读者可自行从文献中获取.

4.2 实现过程

为对比分析设计了三个模型.

模型 A: 使用标准的 DEA 的 CCR 方法.

模型 B: 输入变量组的两个主成分变量作为输入变量, 而输出变量不变.

模型 C: 第一个输入变量加上输入变量组的两个主成分变量作为输入变量, 输出变量使用输出变量组的两个转换后主成分变量. 模型分析结果下表 1.

表 1 模型分析结果表

	模型 A		模型 C		模型 D	
DMU	CRS	VRS	CRS	VRS	CRS	VRS
dmu01	0.8374	1	0.4637	0.5719	0.6262	0.6649
dmu02	0.8711	1	0.4947	0.6712	0.7013	1
dmu03	1	1	0.3969	1	0.5003	1
dmu04	1	1	0.814	0.8772	0.7769	0.8318
dmu05	0.9912	1	0.8697	0.8884	0.8791	0.8884
dmu06	0.9859	1	0.8368	0.9725	0.8334	0.9232
dmu07	1	1	0.943	1	0.9054	0.9693
dmu08	1	1	0.9802	1	0.9895	1
dmu09	1	1	1	1	1	1
dmu10	1	1	1	1	1	1
dmu11	1	1	1	1	1	1
dmu12	1	1	1	1	1	1
dmu13	1	1	0.6441	1	0.5933	0.7034
dmu14	1	1	0.7857	1	0.8053	0.9854
dmu15	0.9778	1	0.3912	0.8465	0.3957	0.8206
dmu16	1	1	0.6272	1	0.5757	0.6826
dmu17	1	1	0.7446	0.9678	0.7542	0.9229
dmu18	0.9778	1	0.3891	0.8308	0.3936	0.8053
dmu19	1	1	0.534	0.534	0.5363	0.5773
dmu20	1	1	0.7594	1	0.8732	1
dmu21	1	1	0.4677	1	0.4267	1
dmu22	1	1	0.4595	1	0.4277	1

4.3 案例结论

不难看出, 无论是CRS组还是VRS组, 从模型A到模型 C, 模型识别能力不断提高, 从结果列数据中为非“1”值的个数就可看出. 如模型A的标准数据包络法的 VRS 方法中所有决策单元的效率都是“1”, 无法依据效率值对这些决策单元排序; 而对应的CRS方法中, 也有超过 70%的决策单元的效率值都是“1”, 无法对于这些决策单元进行排序.

另外, 因为采用数据降维的主成分分析或因子分析方法后, 变量个数减少了, 可以大大减少计算时间.

5 结论

分析展示了在STATA下实现FA-DEA与PCA-DEA的过程与方法, 验证了 FA-DEA 与 PCA-DEA 对提高数据包络法的识别能力与节约计算资源方面的优势. 为在实际应用过程中应用FA-DEA与PCA-DEA, 并为在统一 STATA 的环境下设计与应用数据包络法提供了相应的解决思路.

参考文献

1 Koopmans TC. An Analysis of Production as an Efficient Combination of Activities. Koopmans TC. Activity Analysis of Production and Allocation, Cowles Commission for Research in Economics. Monograph[Thesis], Wiley, New York, 1951.

2 Farrell MJ. The measurement of production efficiency. Journal of Royal Statistical Society, Series A, 1957, 120 (3): 253–281.

3 Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. European Journal of Operational Research, 1978, 2(6): 429–444.

4 魏权龄.评价相对有效性的数据包络分析模型—DEA 和网络 DEA.北京:中国人民大学出版社,2012.

5 李双杰,范超.随机前沿分析与数据包络分析方法的评析与比较.统计与决策,2009,(7):25–28.

6 胡博. Stata 统计分析与应用.北京:电子工业出版社,2013.

7 Ueda T, Hoshiai Y. Application of principal component analysis for parsimonious summarization of DEA inputs and/or outputs. J Op Res Soc of Japan, 1997, 40: 466–478.

8 Lee C, Lee J, Kim T. Innovation policy for defense acquisition and dynamics of productive efficiency: A DEA application to the Korean defense industry. Asian Journal of Technology Innovation, 2009, 17: 151–171.

9 Lee C, Ji Y. Data Envelopment Analysis in Stata. the Stata Journal, 2010, 10(2): 267–280.

10 Hokkanen J, Salminen P. Choosing a solid waste management system using multicriteria decision analysis. European J Opl Res. 1997, 90: 461–72.

11 Sarkis J. A comparative analysis of DEA as a discrete alternative multiple criteria decision tool. European J Opl Res., 2000, 123: 543–57.