正準相関分析と包絡分析法に関する研究

M2013SS009 信田真佑
指導教員：松田眞一

1　はじめに

包絡分析法 (DEA = Data Envelopment Analysis) とは，効率性を分析する方法の一つである．複数の項目を一度に扱うことができ，かつ，その評価方法から個性的な対象も評価される．さらに，定量的に項目を扱うので，具体的な目標値との差を示すことが可能である．正準相関分析は変数群間の相互関係を分析するために用いられる分析法である．包絡分析法にて変数選択を行う際、正準相関分析を用いた論文（上田 [2]）がすでにあり，その妥当性について検証する．統計学の視点から，包絡分析法より得られた結果と正準相関分析で分析を行った結果とを比較し，二つの分析法の間にどのような関係があるか調べる．

2　包絡分析法

包絡分析法とは，公共機関から民間企業などさまざまな事業体の効率性の評価のために適用される方法である．調査集団の中の優れた効率性をもつ事業体の集団が明示され，その事業体を基準として，非効率的な事業体の改善案を具体的に示すことができる．具体的な計算方法として，多入力，多出力系のシステムの効率性を相対的に評価するために線形計画法を用いる．想定される投入対産出の関係を記述する生産関数の形に応じて，いくつかモデルがある．本研究にて扱うモデルについて説明する（刀根 [1] 参照）．

2.1　CCR モデルとその双対問題について

本研究に用いる包絡分析法のモデルについて説明する．分析対象である事業体を一般に DMU (Decision Making Unit) といい，それぞれカテゴリー毎に似たような機能をもって活動している．ただし，ある程度独立した経営上の権限をもっているものとする．

n 個の活動 DMU_1, DMU_2, ...,DMU_n がある．入力は小さいほど好ましく，出力は大きいものほど好ましい．DMU_j の入力を $x_{ij}(1 \leq i \leq m)$, 出力データを $y_{rj}(1 \leq r \leq m)$ とするとき，対象の活動 DMU_o に対する次の線形計画法を考える．

$< CCR_o >$

$$\text{目的関数} \quad \max \ \theta = u_1 y_{1o} + \ldots + u_s y_{so} \tag{1}$$

$$\text{制約式} \quad v_1 x_{1o} + \ldots + v_m x_{mo} = 1 \tag{2}$$

$$u_1 y_{1j} + \ldots + u_s y_{sj} \leq v_1 x_{1j} + \ldots + v_m x_{mj} \quad (j = 1, \ldots, n) \tag{3}$$

$$v_1, v_2, \ldots, v_m \geq 0 \tag{4}$$

$$u_1, u_2, \ldots, u_s \geq 0. \tag{5}$$

$< CCR_o >$ の最適解を $(\boldsymbol{v}_o^*, \boldsymbol{u}_o^*)$ とし，目的関数値を θ_o^* とするとき，$\theta_o^* = 1$ なら DMU_o は D 効率的であり，$\theta_o^* < 1$ なら D 非効率であるという．この $(\boldsymbol{v}_o^*, \boldsymbol{u}_o^*)$ は DMU_o にとってもっとも好意的なウェイト付けの値であり，入出力の項目評価が分かる．ただし，$\theta_o^* = 1$ でも入力の余剰や出力の不足が生じている場合がある．

$\theta_o^* < 1$ のとき，制約式 (3) の中には等式が成立している活動があり，それを DMU_o に対する優位集合と呼ぶ．また，優位集合の和集合を効率的フロンティアと呼ぶ．一般的にはこの線形問題を直接解かず，その双対問題を解いて最適値を求める．

3　正準相関分析

多変量データにおいて，多数の変量が 2 つの変量群を構成するときに，変量群間の相互関係を分析するために用いられる．変量群の相関をもっともよく表わすような新しい変量に変換し，その変量によって 2 つの変量群間の相関について考えようとする

今，p 個の変量群 $X = (X_1, \ldots, X_p)$ と q 個の変量群 $Y = (X_1, \ldots, Y_q)$ がある．まず，任意の係数 $\boldsymbol{s} = (s_{11}, \ldots, s_{p1}), \boldsymbol{t} = (t_{11}, \ldots, t_{q1})$ を用いて，各変量群内で重みつきの合計を考えると，

$$u_1 = s_{11} X_1 + \ldots + s_{p1} X_p \tag{6}$$

$$v_1 = t_{11} Y_1 + \ldots + t_{q1} Y_p \tag{7}$$

となる．重みつきの合計の値は，重み $\boldsymbol{s}, \boldsymbol{t}$ の選び方によって変化する．u_1 と v_1 の間の相関が最大になるように重み係数を選ぶ．求められた u_1, v_1 と無相関で，かつ，相関が最大となるよう u_2, v_2 を求める．この手順を繰り返し，得られた結果を $\boldsymbol{u} = (u_1, \ldots, u_i), \boldsymbol{v} = (v_1, \ldots, v_i)(i = \min(p, q))$ とする．$\boldsymbol{u}$ と $\boldsymbol{v}$ の間の相関係数を正準相関係数と呼び，この相関係数が最大となるように重みの値を決定する方法を正準相関分析という（小笠原 [3] 参照）．

3.1　包絡分析法との違い

包絡分析法は入出力の項目に重みを付け，仮想的な入出力に対して比率尺度が最大となるような係数を考える．正準相関分析は，変量群の重み付き合計の間の相関が最大になるように，重み係数を選択する．2 つの分析法の重みのつけ方には違いがある．包絡分析法は事業体別に項目ごとに好ましいように重みをつけることができるが，正準相関分析では全体として相関が最大となるように重み係数をつけている．

4　プログラム

R の拡張パッケージの一つである lpSolve パッケージを導入し，R 上でも線形計画法を行えるようにした上で，

CCR モデルを入力指向型の双対問題で解くプログラムを自作した（金 [5]，及び逆瀬川 [6] 参照）．計算結果として，D 効率値，双対問題の最適解である $\boldsymbol{\lambda}, \boldsymbol{s}_x, \boldsymbol{s}_y$（$\boldsymbol{s}_x, \boldsymbol{s}_y$ は仮想的な入出力の余過剰でもある），入出力の重み係数である $\boldsymbol{v}_j, \boldsymbol{u}_j$，各事業体に対する優位集合，そして効率的フロンティを得る仕組みになっている．出力の重み $\boldsymbol{u}$ と入力の重み $\boldsymbol{v}$ については，対象の事業体と算出した優位集合のデータのみで再度 CCR モデルによる線形計画法を解き，計算する形をとった．

プログラムの流れは下記の手順になる．

1. 双対問題を解き，最適解 $\boldsymbol{\lambda}, \boldsymbol{s}_x, \boldsymbol{s}_y$ と各事業体の優位集合を得る．
2. 求められた優位集合から対象の事業体ごとに関わる事業体のみにデータを絞りこむ．
3. 入出力の重み係数 $\boldsymbol{v}_j, \boldsymbol{u}_j$ を再度 CCR モデルの線形計画法を行うことによって求める．
4. 効率値が 1 となる事業体については効率的フロンティアから入出力の重み係数を求める．

5　プログラムの適用例

各都道府県で 2010 年の 1 年間で起きた事故に関するデータを利用する．それぞれ入力は，各都道府県の乗用車保有台数を入力 1, 貨物車保有台数を入力 2, 人口を入力 3, 面積を入力 4, シートベルトの着用率を入力 5 とし，出力は交通事故死亡者数を出力 1, 事故発生件数を出力 2 とした．また，出力に当たる 2 つの変数はともに少ない方がよいので逆数にし，入力に当たる 5 つの変数は事故を増やす要因なら同じく逆数にした．ここでは，乗用車保有台数，貨物車保有台数，人口の 3 つのみ逆数にした．このデータを用いて，上田 [4] にて行われた変数選択の手法を検討する．

5.1　正準相関分析によるデータの変数選択

まずデータ間の影響をみるため偏相関係数を求めた．入力の各変数は出力と正の偏相関を持つのがよいが，出力の 2 つの変数とともに正の相関をもつのは乗用車保有台数とシートベルト着用率のみとなっている．貨物車保有台数，人口，面積は正と負に分かれるが，絶対値の大きい方が正となっているのでこれで良いものとする．この結果を踏まえ，データを R の内部関数で標準化を行い，正準相関分析で分析を行った．

表 1　事故データの相関と入出力に対する推定係数

		1	2
正準相関係数		0.944	0.242
入力に対する推定係数	乗用車保有台数	0.088	−0.080
	貨物車保有台数	0.040	−0.231
	人口	0.013	0.335
	面積	0.004	0.113
	シートベルト着用率	0.024	−0.006
出力に対する推定係数	交通事故死亡者数	0.129	−0.227
	事故発生件数	0.022	0.260

表 1 より，第 1 正準相関係数が 0.944 と出ている．

上田 [4] にて，乗数，すなわち推定係数に非負制約を課したとあるが，この場合は自然とすべてが正の方向に揃った．都道府県の事故に関する総合的な値と見て取れるので，正の方向にいくほど，事故との関連が高いことになると言える．そのため，各係数には強弱があるものの上田 [4] のように係数を変数選択の基準にすることはできない．

次に第 2 正準相関係数は 0.242 である．正の推定係数がでているのは，各都道府県の人口，面積，事故発生件数の 3 つのデータである．これより，入力を人口，面積，出力を事故発生件数とする包絡分析法を行う場合を考えることができる．また，正と負に分かれることから，それを特色ととらえ，同じ正と負を示すものの中から変数選択で減らせるものがあるともいえる．

正の方向について考えてみると，人口，面積，事故発生件数である．対象となる都道府県がどの程度事故の起きやすい環境であるかと考えることが可能である．しかし，面積の推定係数が他の係数と比べ小さいので，人の多い場所で起きた事故件数の多さと考えてもよいだろう．

負の方向については，乗用車保有台数，貨物車保有台数，シートベルトの着用率，交通事故死亡者数である．シートベルトの着用率と乗用車保有台数の推定係数は小さいので除いて考えれば，貨物車の関係した交通死亡事故の多さと捉えることができる．以上を踏まえた上で，入力を人口，貨物車保有台数とし，出力を交通事故死亡者数，事故発生件数とする場合も考えることが可能である．

5.2　分析結果の比較

すべてのデータを使い包絡分析法を行った場合と正準相関分析にて変数選択を行った後に包絡分析法を行った場合を D 効率値にて比較した．表 2 にて結果を示す．ただし，入力を人口，面積，出力を事故発生件数とした場合を変数選択後 1 とした．また，入力を貨物車保有台数，人口とし，出力を交通事故死亡者数，事故発生件数とする場合を変数選択後 2 とする．

変数選択前に効率値が 1 であった都道府県は，山形県，群馬県，東京都，長野県，奈良県，鳥取県，島根県，長崎県，沖縄県である．

変数選択後 1 の場合，効率値が 1 なのは鳥取県だけである．選択後 1 の場合の入出力のデータを考えると，この効率値県の面積の小ささや人口の大きさから考えた事故件数の効率がよいことを意味する．県の規模が大きく人口が少ないほど効率値は低くなり，かつ事故発生件数が少ないほど効率値が 1 に近づくことになる．鳥取県の実際の面積と人口のデータを見ると，事故発生件数は他の都道府県よりも少ないが，それは人口の少なさや面積の大きさから考えたものよりよいため効率値が 1 となった．実際人口一人あたり事故発生件数と　面積×事故発生件数をプロットすると図 1 となり，鳥取県が左下の方に位置して効率がいいことが分かる（図は見やすいように対数変換している）．

変数選択後 2 の場合，効率値が 1 なのは山形県，東京都，鳥取県，島根県である．入出力のデータを考えると，この場合の効率値は人口が小さく，走行する貨物車が少

表 2 事故データの変数選択前後の D 効率値の比較

	選択前	選択後 1	選択後 2
北海道	0.986	0.873	0.921
青森県	0.969	0.602	0.969
岩手県	0.946	0.946	0.946
宮城県	0.850	0.518	0.763
秋田県	0.941	0.938	0.941
山形県	1	0.359	1
福島県	0.906	0.493	0.864
茨城県	0.749	0.486	0.652
栃木県	0.841	0.587	0.662
群馬県	1	0.247	0.897
埼玉県	0.953	0.470	0.783
千葉県	0.870	0.630	0.756
東京都	1	0.681	1
神奈川県	0.991	0.581	0.791
山梨県	0.969	0.370	0.834
新潟県	0.849	0.683	0.769
富山県	0.862	0.511	0.654
石川県	0.763	0.548	0.587
長野県	0.889	0.475	0.889
福井県	0.656	0.606	0.606
岐阜県	0.670	0.485	0.558
静岡県	0.650	0.232	0.581
愛知県	0.988	0.329	0.780
三重県	0.732	0.411	0.650
滋賀県	0.668	0.391	0.543
京都府	0.991	0.503	0.793
大阪府	0.914	0.418	0.866
奈良県	1	0.600	0.769
和歌山県	0.827	0.459	0.762
兵庫県	0.680	0.371	0.613
鳥取県	1	1	1
島根県	1	0.947	1
岡山県	0.618	0.298	0.591
広島県	0.723	0.433	0.643
山口県	0.763	0.456	0.689
徳島県	0.657	0.356	0.590
香川県	0.850	0.237	0.579
愛媛県	0.700	0.465	0.696
高知県	0.724	0.562	0.716
福岡県	0.927	0.252	0.869
佐賀県	0.784	0.197	0.619
長崎県	1	0.433	0.876
熊本県	0.740	0.453	0.739
大分県	0.705	0.451	0.676
宮崎県	0.754	0.236	0.744
鹿児島県	0.757	0.403	0.757
沖縄県	1	0.454	0.826

ないほど効率値は低くなり，交通事故死亡者数と事故発生件数が減るほど効率値が 1 に近づいていく．表 3 で入

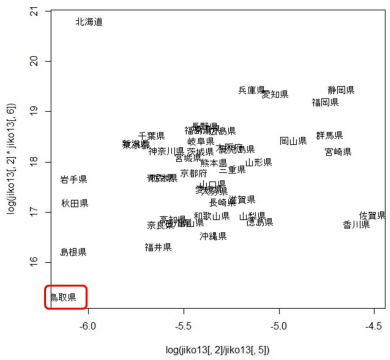

図 1 事故データの変数選択後 1 の人口一人あたり事故発生件数と面積×事故発生件数

出力係数を示す．鳥取県は人口の多さからすると事故発生件数が少ないことから効率がよいのは変数選択後 1 と同じであるが，山形県は貨物保有台数に対する死亡事故者数の少なさから効率がよいと考えられる．効率を視覚化すると，貨物車保有台数 1 台あたりの事故発生件数と人口 1 人あたりの事故発生件数をプロットしたのが図 2 となり，貨物車保有台数 1 台あたりの死亡事故者数と人口 1 人あたりの死亡事故者数をプロットしたのが図 3 となる．上記の説明の様子がプロット図からも読み取れる（これらの図も見やすいように対数変換している）．

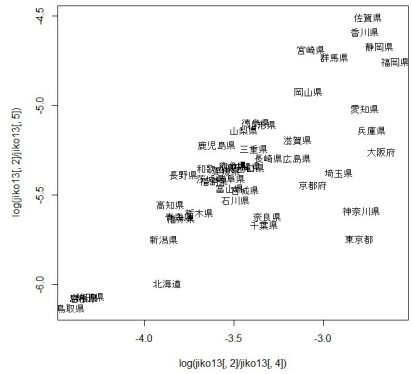

図 2 事故データの変数選択後 2 の貨物車保有台数 1 台あたりの事故発生件数と人口 1 人あたりの事故発生件数

6 考察

CCR モデルで解く問題点として，効率値 1 に関しては自由度が残るので数値の安定性にかけることが挙げられる．入出力の重み係数 $\boldsymbol{v}, \boldsymbol{u}$ は優位集合が効率値 1 の事業体自身で計算することになるので，値の結果に違いが出る．これは本研究でのプログラムに限らず他のプログラムでも本と一致しない点から一般的な問題点と思われる．また，正準相関分析で変数選択を行う場合，第 1 正準相関の係数がすべて有効となったら上田 [4] の方法では変数選択できない．よって，第 2 正準相関の推定係

表 3 事故データの変数選択後 2 の重み係数

	v		u	
	貨物車保有台数	人口	交通事故死亡者数	事故発生件数
北海道	0	5506419	28	10528
青森県	156817	407899	41	555
岩手県	0	1330147	0	2892
宮城県	228362	593996	60	809
秋田県	0	1085997	6	2076
山形県	199819	0	39	0
福島県	232269	604157	61	823
茨城県	347510	903912	91	1231
栃木県	215167	559673	56	762
群馬県	239597	560976	65	0
埼玉県	500426	1301664	131	1772
千葉県	479610	1247519	126	1699
東京都	614725	1439274	168	0
神奈川県	321201	3753802	98	7031
山梨県	158729	0	29	359
新潟県	263294	684859	69	933
富山県	114539	297928	30	406
石川県	60375	705584	18	1322
長野県	432153	0	79	976
福井県	0	806314	0	1753
岐阜県	225197	585762	59	798
静岡県	380566	989896	100	1348
愛知県	624643	1462495	171	0
三重県	206244	536464	54	730
滋賀県	135813	353264	36	481
京都府	195291	507975	51	692
大阪府	567013	1327563	155	0
奈良県	110697	287936	29	392
和歌山県	174631	0	32	395
兵庫県	405613	1055046	106	1437
鳥取県	105773	0	0	1280
島根県	126613	0	15	761
岡山県	218730	568941	57	775
広島県	256285	666626	67	908
山口県	151518	394116	40	537
徳島県	140412	0	26	317
香川県	113082	294140	30	401
愛媛県	162733	423288	43	576
高知県	140066	0	26	316
福岡県	460949	1079233	126	0
佐賀県	145845	0	28	0
長崎県	144096	374811	38	510
熊本県	204894	532953	54	726
大分県	135458	352342	36	480
宮崎県	224450	0	41	507
鹿児島県	336023	0	61	759
沖縄県	148131	385305	39	525

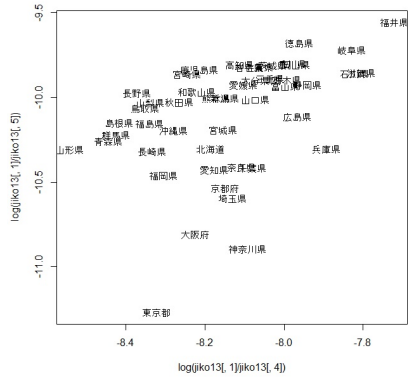

図 3 事故データの変数選択後 2 の貨物車保有台数 1 台あたりの死亡事故者数と 人口 1 人あたりの死亡事故者数

数をみて変数選択することになるが，正負となることが多いため，片方に限ると部分的な解となり，偏りが出る．上田 [4] の方法を改良し，正と負で特色のあるもので絞ると変数選択の意義は出ると考える．その際，第 1 正準相関から正負に分かれるものはそもそも入力と出力の関係を疑ってみるべきかもしれないし，逆数を取るなどの措置や偏相関を確認するなどの細かな措置が必要と考えられる．

7 おわりに

包絡分析法は対象の事業体の特色を残し尊重しつつも，他の事業体より不得手な分野における改善案を示す評価方法であるが，その特異的な評価方法から導入まで検討する企業は少ないと思われる．統計的手法である正準相関分析とともに利用することにより，明確な基準の下，つけた重みに意味があることを示すことが可能であれば，昨今の社会におけるコスト削減や利益向上のよりよい手法になるのではないだろうか．

参考文献

[1] 刀根薫 (1993).「経営効率性の測定と改善-包絡分析法 DEA による-」，日科技連出版社.，東京.

[2] 上田徹 (2003).「DEA における変数選択について」，『日本オペレーションズ・リサーチ学会秋季研究発表会アブストラクト集』，50-51.

[3] 小笠原昭彦 (2006)，「正準相関分析についての解説」， http://homepage3.nifty.com/ogasawara-labo/hanbetsubunseki.pdf .

[4] 森戸 晋 (2014)， 「 『基礎 OR ／ OR 演習』第 4 回 包絡分析法 (DEA)」 ， http://www.morito.mgmt.waseda.ac.jp/kisoor/.

[5] 金 明哲 (2007),「R によるデータサイエンス-データ解析の基礎から最新手法まで-」，森北出版.

[6] 逆瀬川 浩孝,「R サンプルプログラム (数理計画法)」，http://www.f.waseda.jp/sakas/R/Rsample/optimization.html.