統計的方法を用いた変数選択によるプロ野球のチーム成績の包絡分析

2017SS035 北野雄也
指導教員：松田眞一

1　はじめに

好きな野球を分析対象にして，統計学において新たな試みを模索する．そのため，本研究では包絡分析の入力変数の選択手順を考案して，それを用いて包絡分析によってプロ野球の各チームの効率性を比較する．

2　研究の流れ

包絡分析の入力変数の選択に統計的方法を用いる先行研究として尾崎・松田 [2] があるが，包絡分析の出力変数が 2 つ（以下，2 出力と呼ぶ）の場合の手順についての研究であり，出力変数が 1 つ（以下，1 出力と呼ぶ）の場合については触れていない．本研究ではまず，先行研究を参考にしながらプロ野球のデータを分析していく中で，1 出力の包絡分析の場合の入力変数の選択の手順を考案する．そして，2 出力の場合についても，野球成績という膨大なデータを用いるため，先行研究の手順にアレンジを加えてデータを分析していく中で，選択手順を考案する．包絡分析の入力変数の候補には 41 個の変数を用いて，出力変数には 1 出力の場合は勝率とする．2 出力の場合はプロ野球の成績に包絡分析を用いた先行研究の右田 [5] を参考にして，勝率に加えて動員率も出力変数とする．

3　データについて

本研究では一般社団法人日本野球機構 [1] から 2015 年から 2019 年までの各チームのチーム成績及び入場者数のデータを用いる．入場者数のデータについては右田 [5] を参考にして動員率の値を用いるが，修正を加えて本拠地における動員率（以下では本拠地動員率と呼ぶ）の値を用いる．

4　包絡分析

包絡分析とは，銀行，都道府県，学校など様々なものを分析対象（以下，事業体と呼ぶ）として，それらを効率性によって比較する分析方法である．各事業体の活動における産出（出力），投入（入力）に対して，(産出)/(投入) という比を用いて各事業体を比較する．（刀根 [3] 参照）本研究では包絡分析の CCR モデルを用いる．分析には信田 [4] において作られた分析ソフト R 上でのプログラムを用いる．

5　変数選択の手順

考案した変数選択の手順を説明する．変数の元の値の逆数の値にする処理を逆数処理，反転した値にする処理を反転処理と以下呼ぶ．反転した値とは処理対象の値を，その変数の最大値と最小値の和からその値を引いた値である．

5.1　1 出力の場合

包絡分析が 1 出力の場合は重回帰分析を用いる．包絡分析の出力変数を重回帰分析の目的変数，包絡分析の入力変数の候補を重回帰分析の説明変数として分析を行う．手順を以下に示す．

1. VIF 関数と減少法を用いて重回帰分析に用いる説明変数を選択する．
2. 重回帰分析を行い，係数が負の説明変数は元の値に対して逆数処理，または反転処理を行う．
3. 係数の値が負の説明変数がなくなるまで手順 2 を繰り返す．
4. 係数の値で包絡分析における入力変数を選択する．

逆数処理や反転処理は VIF 関数や減少法を用いた後に行っている．それらの処理を先に行うと，重回帰分析の結果に再び，係数の値が負の変数が生じるからである．

5.2　2 出力の場合

包絡分析が 2 出力の場合は先行研究の尾崎・松田 [2] を参考にして，正準相関分析を用いる．また，本研究では野球成績という膨大なデータを扱うため，正準相関分析を用いる前に，VIF 関数や減少法による変数選択を行う．この点は先行研究とは異なり，本研究においてアレンジした部分である．手順を以下に示す．

1. VIF 関数と減少法を用いて，正準相関分析に用いる前に，包絡分析の入力変数の候補に対して変数選択を行う．（減少法は，2 つの出力変数に対してそれぞれ行い，どちらかにおいて削られなかった入力変数の候補は残す．）
2. 手順 1 で選ばれた入力変数の候補と 2 つの出力変数に対して正準相関分析を行う．
3. 入力変数の候補の中で，第 1 正準変量の係数が出力変数の第 1 正準変量の係数と異符号の変数の元の値に対して逆数処理，または反転処理を行う．そして，再び，入力変数の候補と 2 つの出力変数に対して正準相関分析を行う．
4. 手順 3 において入力変数の候補の中で，第 1 正準変量の係数が出力変数の第 1 正準変量の係数と異符号の変数がなくなるまで手順 3 を繰り返す．
5. 入力変数の候補の第 2 正準変量の係数を見て，正と負においてそれぞれ絶対値の大きい変数を包絡分析における入力変数として選ぶ．

ただし，手順 2 において，先行研究では出力変数がその変数が小さい方が望ましい場合，その変数の元の値を逆数処理，または，反転処理を行っている．また，出力変数の係数が第 1 正準変量において同符号になり，第 2 正準変量においては異符号となっている．本研究でもこのことを組み込んで手順 2 を行う．

6　分析結果

紙面の都合上，2 出力の場合の結果のみ載せる．申告敬遠の制度の影響を受ける打撃成績，投手成績のそれぞれの故意四球の成績を入力変数の候補から外した．そして，上記の手順に従って分析した．まず，手順 1 において 20 個に絞られ，手順 3 と手順 4 を行った結果，三塁打，盗塁刺，死球，三振，併殺打，被安打，与四球，暴投の元の値に逆数処理が行われて正準相関分析の結果（表 1）より入力変数に併殺打と奪三振が選ばれた．

表 1　正準相関分析の結果

		1	2
正準相関係数		0.92092	0.75149
入力変数の候補の変数の推定係数	打数	0.05437	0.01059
	三塁打	0.00349	0.02011
	本塁打	0.05432	−0.00717
	盗塁	0.01066	0.00712
	盗塁刺	0.00575	−0.02277
	犠打	0.02338	−0.00245
	犠飛	0.03153	0.00842
	死球	0.00014	0.06886
	三振	0.00881	−0.02294
	併殺打	0.00467	−0.04078
	セーブ	0.04638	−0.01720
	完投	0.00414	0.01576
	完封勝	0.01790	−0.00959
	投球回	0.00402	−0.01989
	被安打	0.02426	−0.02397
	与四球	0.02655	−0.02307
	与死球	0.00044	−0.01985
	奪三振	0.00522	0.08044
	暴投	0.00644	0.06723
	守備率	0.00223	0.00339
出力変数の推定係数	勝率	0.12726	−0.05174
	本拠地動員率	0.00839	0.13712

その 2 つの入力変数と勝率，本拠地動員率を出力変数とした包絡分析の結果が表 2 である．この結果において主要なことを元データに基づいて、そうなった理由を述べる．2017 年において，4 位の巨人はデータを取った 5 年間の全 60 チームの中で併殺打が最も多い中で，本拠地動員率は 3 番目に高いことから効率値が 1 になったと考えられる．また，最下位のヤクルトは全 60 チームの中で併殺打が 3 番目に多い，そして，勝率が最も低いが，本拠地動員率は 26 番目に高いことから効率値が順位に反して高くなったと考えられる．つまり，勝率は低いが本拠地動員率の高さのおかげで効率値が高くなったと考えられる．2015 年の 1 位のソフトバンクは全 60 チームの中で併殺打が 2016 年のソフトバンクと並んで 7 番目に多いという成績の中で，勝率は 2 番目に高いことから効率値が 1 になったと考えられる．

7　まとめ

包絡分析の入力変数の選択手順を先行研究を参考にしながら，データを分析していく中で考案することができた．1 出力の場合は独自の選択手順を考案することができた．2 出力の場合も先行研究の手順に VIF 関数，減少法を用いることを加えて，野球成績という膨大なデータに対応できる手順を考案できた．その手順から選ばれた入力変数を用いて包絡分析を行い，各チームの効率性を比較することができた．その結果においては元データから様々な考察ができた．2 出力の場合は出力変数に本拠地動員率が加わることで，勝率が低くても本拠地動員率が高いことから効率値が高くなるといった 1 出力の場合とは違う考察ができた．

表 2　包絡分析の結果

順位	セ	効率値	パ	効率値
1	巨人 19	0.9420	西武 19	0.9850
2	横浜 19	0.8814	ソフ 19	0.8616
3	阪神 19	0.8851	楽天 19	0.8984
4	広島 19	0.9168	千葉 19	0.7812
5	中日 19	0.8663	ハム 19	0.7615
6	ヤク 19	0.9044	オリ 19	0.6836
1	広島 18	0.9525	西武 18	0.9961
2	ヤク 18	0.9692	ソフ 18	0.8850
3	巨人 18	0.9696	ハム 18	0.8974
4	横浜 18	0.7628	オリ 18	0.7512
5	中日 18	0.9371	千葉 18	0.8617
6	阪神 18	0.8193	楽天 18	0.7697
1	広島 17	1.0000	ソフ 17	0.8756
2	阪神 17	0.8466	西武 17	0.9285
3	横浜 17	0.8361	楽天 17	0.8673
4	巨人 17	1.0000	オリ 17	0.7447
5	中日 17	0.8422	ハム 17	0.7617
6	ヤク 17	0.9194	千葉 17	0.7409
1	広島 16	1.0000	ハム 16	0.9469
2	巨人 16	0.9401	ソフ 16	0.9514
3	横浜 16	0.8022	千葉 16	0.9402
4	阪神 16	0.8112	西武 16	0.7899
5	ヤク 16	1.0000	楽天 16	0.7996
6	中日 16	0.8778	オリ 16	0.7547
1	ヤク 15	0.9106	ソフ 15	1.0000
2	巨人 15	1.0000	ハム 15	0.8895
3	阪神 15	0.8405	千葉 15	0.8707
4	広島 15	0.9347	西武 15	0.8768
5	中日 15	0.8415	オリ 15	0.7441
6	横浜 15	0.7834	楽天 15	0.7223

8　おわりに

本研究では考案した変数選択の手順を用いた包絡分析でプロ野球の各チームを分析したが，他の分野においても試したい．

参考文献

[1] 一般社団法人日本野球機構：NPB.jp 日本野球機構，https://npb.jp，2020 年 8 月閲覧．

[2] 尾崎友彦・松田眞一：正準相関分析を包絡分析に適用する研究，南山大学紀要『アカデミア』理工学編，**20**，21-36，2020.

[3] 刀根薫：『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，1993.

[4] 信田真佑：正準相関分析と包絡分析に関する研究，南山大学大学院理工学研究科修士論文，2015.

[5] 右田光司：包絡分析に基づく球団の運営分析，法政大学経営学部経営戦略学科卒業論文，2017.