三田祭論文

因子分析を用いた競走馬データの分析

長倉大輔研究会
池田孝平[1]、斎藤光[2]、篠田星南[3]
[123]慶應義塾大学経済学部

要旨
競馬においては、さまざまな参加者の判断が馬券のオッズに反映される。本稿では、日本中央競馬において競走馬がもつ様々なデータをもとに、統計解析ソフトRを用いて因子分析を行い、単勝と三連単の馬券の違いを生み出している変数の手がかりを探す。

目次

1.はじめに

日本の公営ギャンブルの一つである競馬には、馬の成績やオッズなどの様々なデータがある。そして、そのようなデータは、お互いが複雑に影響しあう為に、順位の予想を難しくさせている。そこで、この論文では、そのようなデータ間にある複雑な関係を統計ソフトRによる因子分析によりデータの背後にある因子を推定し、考察する。

2.先行研究

辺見(2009)は、過去の中山競馬場での勝馬の成績のデータを用いて因子分析と主成分分析を行い、有馬記念の勝馬の予想を行った。

小幡・太宰(2014)は、日本の競馬の三連単における、勝率が低い馬に過剰な人気が集まってしまう大穴バイアスというものについて、単勝と三連単の比較をすることによってその存在を統計的に立証していた。この大穴バイアスは行動経済学におけるプロスペクト理論というもので説明できるとされている。プロスペクト理論というのは、確率が微小であればあるほど人間がそれを過大評価してしまうという理論である。単勝馬券とは単純に一着となる馬を予想する馬券で、三連単馬券とは日本の競馬市場にある、レースの 1 着から 3 着までを予想する難易度の高い馬券である。三連単馬券は当選する確率が非常に低いことで知られている。そのため、プロスペクト理論によって、実際の勝率とオッズの間に歪みが生じていると考えられる。このことから、当選確率に極端な差がある単勝と三連単のオッズでは、各変数から受ける影響が異なることが考えられるため、この二つの比較を中心に研究を進めた。

3.実証分析

3.1 データ及び分析手法について

Ⅰ.因子分析

因子分析とは、多変量データから潜在的ないくつかの共通因子を推定する手法である。
因子分析では、観測される複数の変数が、それらの変数に共通の成分（共通因子）と、それぞれの変数に独自の成分（独自因子）から構成されるというモデルを想定する。仮に、p 個の観測変数に対し、一つの共通因子 f のみを想定する時、それぞれの観測変数 Yj は以下の式で表される。

$$Y_j=\beta_j f+e_j (j=1,2,...,p)$$

ここでは、共通因子 f にかかる係数β_iを因子負荷と呼び、変数 j がその因子をどの程度反映しているかを示す。また、e_jは変数 j に含まれる独自因子を示す。因子負荷β_iと共通因子 f は最尤法により推定する。因子負荷より、それぞれの変数がどの程度、その因子の影響を受けているかを表す。因子数は変数の数だけ算出されるが、分析者が特定の基準を用いて因子数を決定する。そして、共通因子の意味を解釈しやすくするために因子ベクトルを回転する。回転には、因子ベクトルが互いに直交するという性質を保持したまま回転する直交回転と、自由に回転させる斜交回転がある。本稿では、因子ベクトル同士が無相関であることを仮定しない斜交回転の一つである、プロマックス解により回転を行った。また、ある変数における、各因子の因子

負荷の 2 乗和を、共通性と呼ぶ。共通性は、その変数が因子空間によってどの程度説明できているかを表す。

Ⅱ.変数の選定

競馬の参加者が馬券購入時に参考にする情報として、順位、単勝オッズ、3 連単オッズと、馬、騎手、調教師それぞれの勝率（一着率）と三着以内率を変数とした。また、馬券の人気に影響を与えそうな数値として、馬主の名前のグーグル検索でのヒット数も加えた。単勝と三連単両方のオッズのデータを使うことによって、違う種類の馬券の比較をした。

Ⅲ.データの収集方法

JRA のホームページ（http://www.jra.go.jp/）から、2015 年 11 月から 2018 年 10 月の間に、to 東京競馬場で行わ i れた 16 頭、1000 万下、ダート、1600mの条件を満たすレース 32 回分について、単勝の得票率、3 連単の得票率、馬の勝率、馬の三着内率、馬主の知名度、騎手の勝率、騎手の三着以内率、調教師の勝率、調教師の三着以内率、レースでの順位の項目のデータを Python によるウェブスクレイピングによって集めた。

単勝の得票率は、オッズの逆数に単勝の払い戻し率 0.8 をかけることによって求めた。また、三連単馬券については、馬番 i の馬が 1 位、馬番 j の馬が 2 位、馬番 k の馬が 3 位という馬券のオッズをO_{ijk}としたとき、

$$\sum_{j\neq i}\sum_{k\neq i,k\neq j}\frac{1}{O_{ijk}}\times 0.725$$

という計算をすることによって、馬 i が一位になる馬券の得票率を合成し、単勝馬券の得票率と比較できる形にした。ここで、0.725 は 3 連単馬券の払い戻し率である。

また、一部の馬券について、騎手や調教師のデータに欠陥があるものは取り除いて分析を行った。

3.2 分析結果

統計ソフトは R を用いた。”psych”というパッケージをインストールし、因子分析を行う関数として”fa”という関数を用いて分析した。ちなみに、因子の数、因子の回転法、初期会の算出法を自由に仮定できる。

因子数の決定に際し、本稿では、表1にある芝（1981）の基準を用いた結果、変数が 10 種類なので因子数は 2 を採用し、分析を行う。

表 1. 分析変数の数と適切な因子数 出所：芝(1981)

変数の数	因子数
8～13	2
14～18	3
19～25	4
26～31	5
32～38	6
39～46	7
47～53	8

そして、因子分析を行った結果、以下のような因子負荷、共通性、因子寄与率、因子間相関となった。

表 2.

	因子 1	因子 2	共通性
単勝の得票率	1	0	0.995
3 連単の得票率	1.01	−0.03	0.9893
馬の勝率	0.47	0.03	0.226
馬の三着以内率	0.56	0.04	0.327
馬主の知名度	−0.07	0.09	0.0087
騎手の勝率	0.07	0.95	0.9516
騎手の三着以内率	0.07	0.96	0.9812
調教師の勝率	0.18	0.27	0.1361
調教師の三着以内率	0.14	0.3	0.1369
レースでの順位	−0.48	−0.04	0.2423
寄与率	0.29	0.21	
因子間相関	0.33		

3.3 考察

因子 1 に着目すると、因子 1 は得票率に対する因子負荷量が高いため、馬の人気に影響を与える因子であると推測される。また、馬の成績に対する寄与率も比較的高い。よって、馬の人気を予測するためには馬自体の成績に着目すべきだと分かる。次に因子 2 に着目すると、騎手の成績に関する寄与率が高いため、騎手の強さを示す因子であることが推測される。しかし、この因子は馬の人気への影響が単勝で 0、三連単で-0.03 となっている。ここから三連単を選ぶ人々は単勝馬券の購入者と比較して騎手の成績を正しく把握できていない、もしくは把握していないことが推測できる。

また、以上を踏まえた上で因子間相関を見ると、0.33 とあることから、因子 1 と因子 2 には弱い正の相関関係が見られる。

4.終わりに

今後は儲かる馬券の買い方を追求するため、馬の成績に影響を与え、人気に影響を与えない因子を探すことを目標にしたいと考えている。

参考文献

[書籍・論文]

小幡績、太宰北斗(2014)「競馬とプロスペクト理論：微小確率の過大評価の実証分析」行動経済学、第七巻、1-18.

辺見広大 (2009)「主成分分析と因子分析による競馬の勝因の研究」 大阪工業大学
<https://www.oit.ac.jp/is/~shinkai/seminar/thesis/2008henmi/2008_Bthesis_henmi.pdf>

芝祐順(1981)「因子分析法のための会話型プログラム」東京大学教育学部紀要.21, 53-65.

豊田秀樹 (2012)『因子分析入門―R で学ぶ最新データ解析』東京図書 .

[Web 上の資料]

日本中央競馬会(2018)「JRA 日本中央競馬会]
<http://www.jra.go.jp>(2018/11/5 アクセス)