大阪商業大学アミューズメント産業研究所紀要
第12号（平成22年6月）抜刷

テクニカル分析の立場をベースとした複勝馬券の的中確率に関する統計モデル

伊　藤　耕　介

論文

テクニカル分析の立場をベースとした複勝馬券の的中確率に関する統計モデル

伊藤耕介

1 はじめに

1940年代後半から、欧米では競馬を学術的な研究対象として考察するという試みがなされており、経済物理学の発達とともに近年では多くの研究がなされている。競馬の馬券の的中が心理学や統計学、経済学の立場からどのように記述されるかということが真剣に議論されてきたわけである[1)]。そのなかでも、馬券の的中に関する研究は、株式投資の用語を借用して次のように大きく2つに分類される。一つは「ファンダメンタル分析」である。これは、馬の着タイムや走破速度を、過去の着順やタイムなどの要因から説明しようとすることである。そのために、重回帰分析やニューラルネットワーク、あるいはより一般化された共分散構造分析などの統計的手法を用いて、説明変数と馬の走破能力とを関係づけることによって明らかにしようとする立場に立つ。もう一つの試みは「テクニカル分析」的な解析である。これは、馬券購入者は適正に馬の能力をとらえ、支持率に代表される各馬の人気は実力を反映しているとして、これ自身が情報として有用であるという立場に立つ。

ファンダメンタル分析的な手法を用いた解析は、国内の中央競馬にも応用された例がいくつかある[2)]が、テクニカル分析の立場で行う馬券の的中確率予測は、まだ日本ではほとんど知られていない。そこで、本論文では、過去に欧米で行われてきた研究をレビューするとともに、彼らの手法を日本の中央競馬の最新のデータに適用し、馬券の的中確率予測システムの実例を示すことにしたい。

さて、何の情報も与えられず、ランダムに日本で馬券を購入する場合には、回収率は平均25%といわれる控除率を除いた値75%に収束するはずである（厳密にいえば、日本では控除率は券種によって異なる値に収束する）。しかし、統計的知見に基づいて回収

率を向上させることは十分に可能である。単純に言うならば、馬券の的中確率に比して、馬券のオッズが十分に高い場合を選び出せばよいのである。日本の公営競技では、パリ・ミュチュエル方式(pari-mutuel system)が導入されており、競馬のオッズは馬券の支持率によって決まる値となっている。すなわち、馬券の的中確率ではなく、馬券の人気によって、オッズが決まっているのだから、馬券の的中確率が購入者によって過小評価されている場合を統計学的に抽出できれば、回収率の向上を見込めるのである。

テクニカル分析の立場では、単勝馬券の支持率が各馬の実力を適切に反映し、そのうえで、単勝馬券以外の券種に関しては、必ずしもオッズが馬券の的中確率を適切に反映しているものではないと仮定することが多い。これらの仮定の妥当性については、第5節で議論することにするが、結論から言うと、この前提に則って理論を構築しても、回収率の高いと予想される馬券と低いと予想される馬券をそれなりに分類できることが示されている。本論文では、過去の研究にならって、複勝馬券の的中確率に焦点を当てて議論を展開することとする。

さて、回収率が高い馬券と低い馬券を分離するためには、単勝馬券以外の券種の真の的中確率をいかにして推定するかという問題が、非常に重要だということは容易に想像がつくであろう。そこで、的中確率推定にとって最も重要なテーマは、単勝馬券の人気の情報から、単勝馬券以外の的中確率を合理的に推定するシステムを、どのようにして構築するかということになる。ところが、残念なことに、馬券の的中確率に関しては、もっともらしい理屈を立てれば複数のモデル化がありえてしまう。そこで、提案された複数のモデルから、ふるいをかけてよりよいモデルを選び出すということを試みる。そのためには、情報量規準及び最大対数尤度という概念が必要となる。そこで、本論文ではこれらの理論についても簡単に触れておきたい。

本論文の構成は次の通りである。第2節では、今回の統計モデルを議論するのにあたり必要となる「真の確率」の概念に焦点をあてる。第3節では、単勝馬券以外の券種の的中確率に関して過去に提案された計算式を紹介するとともに、対数尤度関数を用いてその比較を行った議論を紹介する。第4節では、第3節の議論を日本の中央競馬のデータセットに適用する。これにより、日本の中央競馬の複勝馬券の的中確率に関しても統計モデルが有用であることが示される。第5節では、本研究のベースとなっている単勝馬券の支持率が単勝馬券の的中確率に等しいという前提の問題点について触れ、さらなる期待値の向上を目指すにはどうしたらよいかについて簡単に述べる。第6節では、本

論文のまとめを述べる。

2 「真の確率」の概念

突然であるが、五つの面に1、残り一つの面に2と書いてある立方体のサイコロについて考える。細工がなければ、1が出る確率は6分の5であり、2が出る確率は6分の1である（出る目の場合の数は2通りであるが、当然のことながらそれぞれの出る確率は2分の1ではない)。このことは、どの面も均等に出るという前提に立ったとき、各面の出る確率は同様に確からしく6分の1となるという根拠に基づいた確率となっている。

競馬における馬券の的中確率は、このような「同様に確からしい確率」として計算することはできない。いうなれば、目の数がどのように配分されたかわからないサイコロを、自分の見えないところで誰かに振られて、自分はその結果だけを知るといったような試行に相当するからである。例を挙げていうと、a、b、c、dという4種類の馬券に対して、aが二面、bが一面、cが二面、dが一面割り当てられたサイコロがあったとして、誰かがそのサイコロを振り、自分は“b”という結果だけを教えてもらう、といったようなことになる。ここで出てくるような不可知なサイコロの割り付け方に基づく確率を、以後、「真の確率」と呼ぶ。

もし、サイコロの目の割り付け方（真の確率）がわかっているならば、勝負は勝ったようなものである。それぞれの確率にオッズをかけて、期待値が1を超えるような馬券をどんどん買っていけばよいのである。しかし、残念ながら、そのようなサイコロは我々に見えないばかりか、サイコロを振る側は、天候状態・距離・馬の能力・騎手などの要因によって、異なったサイコロを使い分けてくる。競馬のレースを確率的事象として見たときの難しさは、真の確率がわからないことに尽きるといっても過言ではないのである。

ただし、何らかの仮定を置くことによって馬券の的中確率を予測することは可能である。ここではある仮定に基づいて構築される統計モデルによって計算される確率を「モデルの確率」と呼ぶことにしよう。モデルの確率は真の確率とは異なるが、統計モデルが状況をよりよく反映しているのであれば、真の確率に近い確率を出力することが期待される。次節では、過去に提案された複数の統計モデルを紹介したのち、モデルの確率が真の確率に近いかどうかを将来予測の観点で評価する対数尤度関数を紹介する。

3　過去の研究のレビュー

第2節で述べたように「真の確率」というものは一般には不可知である。しかしながら、もっともらしい仮定を用いて、それなりに信頼のおける統計モデルを提案することは可能である。導入部でのべた通り、ここでは、過去の研究にならい、単勝馬券の的中確率を単勝馬券の支持率に等しいとする仮定をおくことによって確率モデルを構築した研究に触れることにする。この仮定は、Hausch and Rubinstein (1981)によって提案されたものであり、単勝馬券をどのように選んでも利益が出ないということを意味している[3)]。まず、実際にこの仮定が妥当かどうかを示すために、日本中央競馬会(Japan Racing Association; JRA)の1995年から2007年までのデータ（データの詳細は4-1節で述べる）を用いて、単勝支持率5％ごとにビンを用意し、各単勝馬券を分類した。図1の横軸は各ビンの単勝支持率を表し、縦軸には各ビンに含まれる馬券の的中率を示している。また、図2には、それぞれのビンに含まれる馬の総数を示している。これをみると、単勝支持率65%を超える単勝馬券に関しては若干母集団の勝率を過大評価しているものの、多くの場合では、単勝馬券の勝率が支持率にほぼ等しくなっている。すなわち、単勝馬券の的中確率が単勝馬券の支持率に等しいとする仮定は第一次近似として採用可能なものであると考えられる。単勝支持率が1％以下の場合にバイアスが見られることも

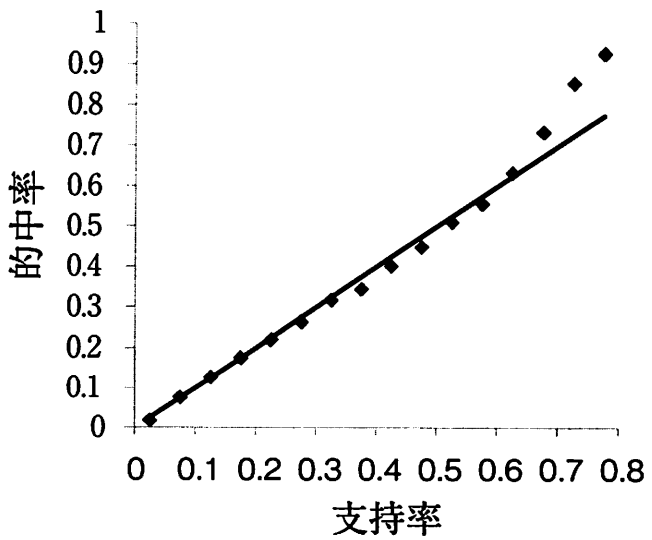

図1　（点）単勝馬券の支持率と各ビンに含まれる馬券の的中率。（実線）支持率＝的中率を表す線。

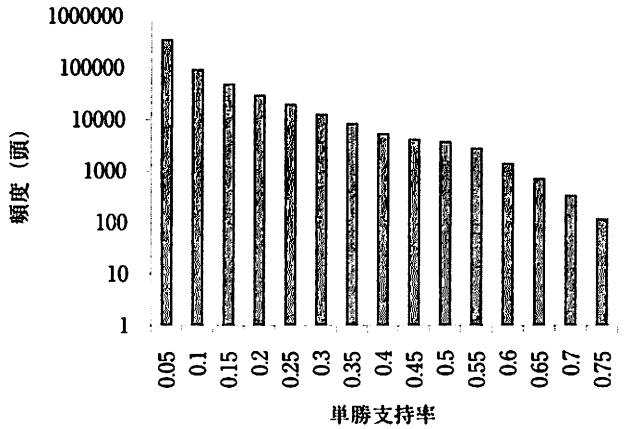

図2　単勝支持率の各ビンに含まれる馬の総数

含めて、より詳細な議論は第5節で行うが、第3節、第4節ではこの仮定を前提として統計モデルの構築を行っていく。

この仮定に従えば、単勝馬券の的中する確率は単勝馬券の支持率から計算可能な量となる。そこで、今度はそれを用いて別の券種の的中確率を評価することを考える。過去に提案されたもののうち、もっともシンプルなものとして挙げられるのは、Harville (1973)の統計モデルであろう[4]。これは、1着目が馬iであるとき、2着目が馬jとなる条件付き確率は、全体から1着馬の票数を除いた票数に占める2着馬の票数の割合で表わされると仮定する。すなわち、馬i、馬jが1着となる確率（＝馬i、馬jの単勝馬券の支持率）をそれぞれπ_i、π_jとするならば、馬iが1着となるときに馬jが2着となる条件付き確率は$\pi_j/(1-\pi_i)$となる。結局、馬iが1着となり、かつ馬jが2着となるモデルの確率π_{ij}は：

$$\pi_{ij} = \pi_i \frac{\pi_j}{1-\pi_i}$$

と書くことができるというものである。彼らのモデルでは、以下同様に1着目が馬iで2着目が馬jとなるとき、3着目が馬kとなる条件付き確率は、全体から1着馬と2着馬の票数を除いた票数に占める3着馬の票数の割合で表わされ、1着がi、2着がj、3着がkとなるモデルの確率π_{ijk}は、

$$\pi_{ijk} = \pi_i \frac{\pi_j}{(1-\pi_i)} \frac{\pi_k}{(1-\pi_i-\pi_j)}$$

と計算されている。本研究で扱う複勝馬券は、ある馬が2着もしくは3着までに入賞したときに払い戻しが受けられる式別である馬券で、2着までの入賞に対して払い戻しが与えられる場合2頭複勝(place)、3着までの入賞に対して払い戻しが与えられる場合3頭複勝(show)と呼ばれるが、その的中確率$\hat{\pi}_i$は、それぞれ、π_{ij}とπ_{ijk}の組み合わせとして書くことができて、それぞれ、モデルの確率は

2頭複勝の場合：

$$\hat{\pi}_i = \pi_i + \sum_{n \neq i} \pi_n \frac{\pi_i}{1-\pi_n} = \pi_i + \sum_{n \neq i} \pi_{ni}$$

3頭複勝の場合：

$$\hat{\pi}_i = \pi_i + \sum_{n \neq i} \pi_n \frac{\pi_i}{1-\pi_n} + \sum_{n \neq m} \sum_{m \neq i} \pi_n \frac{\pi_m}{(1-\pi_n)} \frac{\pi_i}{(1-\pi_n-\pi_m)}$$

$$= \pi_i + \sum_{n \neq i} \pi_{ni} + \sum_{n \neq m} \sum_{m \neq i} \pi_{nmi}$$

と計算される。

このHarville (1973)のモデルは、いわば、条件付き確率が2着目・3着目以下の馬の票数の多寡の情報によって決められると考えるもので、ある程度、合理的であると考えられる。しかし、別の仮定をもとに統計モデルの提案を行うことも可能である。Henery (1981)は各馬の速度分布が分散の等しい正規分布をしていると仮定して、単勝馬券の的中確率と各馬の規格化された速度θ_iを用いた、以下のような関係式を提案している[5]：

$$\pi_i = \int_{-\infty}^{\infty} \prod_{n \neq i} \Phi(u + \theta_i - \theta_n) \phi(u) du$$

ここで、$\phi(\ )$は正規分布の確率密度関数を意味し、$\Phi(\ )$は正規分布の累積分布関数を表している。この式は、簡単に言うと、馬iの速度がuからu+duの範囲にあるときに、馬iが1着となる条件付き確率は、i以外の馬n($\neq$i)の速度がu以下である確率の積として表されるということを意味している。

同様に考えると、1着がiで2着がjとなる同時確率関数は、

$$\pi_{ij} = \int_{-\infty}^{\infty} \Phi(u + \theta_j - \theta_i) \prod_{n \neq i,\ j} [1 - \Phi(u + \theta_j - \theta_n)] \phi(u) du$$

と計算される。また1着がiで2着がjで、3着がkとなる同時確率関数は、

$$\pi_{ijk} = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{u+\theta_k-\theta_j}\Phi\left(u+\theta_j-\theta_i\right)\phi(v)dv\right] \times \prod_{n\neq i,\ j}\left[1-\Phi\left(u+\theta_j-\theta_n\right)\right]\phi(u)du$$

と表わされる。複勝馬券の的中確率は、Hervilleのモデルのときと同様に、π_{ijk}とπ_{ij}から計算することが可能である。ただし、実際にこの手順でπ_{ijk}を計算するには、計算負荷が大きい。というのは、単勝の的中確率π_iの情報からニュートン法などによって各馬の速度θ_iを求め、さらに、各組み合わせについて二重積分と直積の計算をするという作業を繰り返さなくてはならないからである。

HervilleのモデルやHeneryのモデルを包含する、一般化モデルとしてStern (1990)のモデルがあげられる[6]。この一般化モデルでは、速度分布がガンマ分布であることを仮定するが、多重積分を用いて各馬の速度を求め、その計算を基に各馬券の的中確率を計算するため、馬券購入を判断するシステムとしてはHeneryのモデルと同様に計算負荷が大きい。そこで、できるだけ短時間で計算を終了させるために、近似的に確率を求める方法が提案されている[7]。この近似を用いると、Sternの一般化モデルは以下のように簡略化される（以下、この統計モデルを簡略一般化モデルと呼ぶことにする）：

$$\pi_{ijk} = \pi_i \frac{\pi_j^{\lambda 2}}{\sum_{n\neq i}\pi_n^{\lambda 2}} \frac{\pi_j^{\lambda 3}}{\sum_{m\neq i,\ j}\pi_m^{\lambda 3}}$$

ここで、パラメータλ_2とλ_3は、簡略一般化モデルのうち速度分布をどのように選択するかを表す定数である。ここでは、$\lambda_2=\lambda_3=1.0$としたときにHervilleのモデルが現れることは明らかであるが、$\lambda_2=0.76$、$\lambda_3=0.62$としたときにはHeneryモデルを近似したものが現れることに注意していただきたい。本論文で使用したモデルに用いるパラメータは表1にまとめてある。

Lo et al. (1995)は対数尤度が最大となるかどうかを判断基準として、簡略一般化モデルのうちどのようにλ_2とλ_3を代入したものが最良なモデルとなるかを比較している[7]。今回のようにパラメータのチューニングが必要ない場合、赤池情報量規準の観点では、対数尤度を最大とするものが、将来の予測にとって尤もらしいものであるとすることができる（詳細は付録を参照のこと）。彼らは、Meadowland、Hong Kong、Atlanta City

表1　簡略一般化モデルに用いられる
λ_2とλ_3の値（出典：Lo et al., 1995）

r	λ_2	λ_3
1(Harville)	1.00	1.00
2	0.93	0.89
3	0.90	0.84
4	0.88	0.81
5	0.87	0.80
6	0.86	0.78
7	0.86	0.77
8	0.85	0.76
10	0.84	0.75
20	0.82	0.72
30	0.81	0.71
40	0.81	0.70
∞(Henery)	0.76	0.62

そして日本中央競馬会主催のレースについて、π_{ijk}の当てはまりのよさを基準としてモデルの選択を行った結果、彼らは日本の中央競馬以外のレースではHeneryのモデルが最適であり、日本の中央競馬に関してはGamma(4, θ_i)モデル（簡略一般化モデルにλ_2=0.88、λ_3=0.81を代入したもの）がよいと結論付けている。この結果をもとにLo et al. (1995)は、Kelly規準[8)]に基づいて掛け金をコントロールし、Gamma(4, θ_i)モデルを用いて複勝馬券の的中に関するモデルの確率を計算し、実際に期待値が1よりも大きい馬券に対して投資を行っている。その結果、日本の競馬においてもランダムに購入したと仮定した場合に比べると残額は多くなっており、研究で実践された一部のケースについては収益を上げている。

Lo et al. (1995)の研究における一番の問題点は、データが不十分なことである。日本の場合、最適なモデル選択のために使われたサンプル数は600レースであり、π_{ijk}のあてはまりやモデル選択の効果をみるにはデータが少なすぎる。また、モデル選択がうまくいっていたかどうかの検証には最大でも61頭しか使われておらず、システムが健全に動いて収益を上げたのか、それともランダムな要素によってたまたま収益が上がったのかを判断するのは非常に難しい。

今日では、過去のレースの情報はコンピュータのデータベースとして保存され、膨大な量になっており、よりもっともらしいモデルの選択および検証が可能となっている。

次節では、日本中央競馬会の公式データ配信サービスであるJRA-VANが配布しているData Lab.のデータセットを基に、改めて対数尤度関数を用いてLo et al.（1995）の提案したモデル選択の問題について調べ、選択された各確率モデルの検証について議論を展開したい。

4　日本の中央競馬会でのモデル選択問題と確率予測

4-1　使用したデータセットと手法について

本節ではJRA-Van Data Lab. よりダウンロードした1995年１月５日から2007年12月23日までの払戻金データ、確定オッズ及び確定後の票数データを用いる。ただし、この期間中降雪のため開催が中止となった４レースのデータと、複勝馬券が発売されなかった３レースに関しては含んでいない。解析したレースの総数は44774レースである。レース開始前に判明する出走取消・競争除外の馬についてはデータを除き、レース開始後に判明する競争中止についてはデータを採用している。日本の中央競馬会においては、出走頭数が８頭以上の場合に３頭複勝、５頭以上７頭以下の場合に２頭複勝となるため、それを踏まえて計算を行っている。

各確率モデルの対数尤度の計算に用いたのはそのうち1995年１月から2001年６月までのデータ（22261レース分）であり、このデータを用いて先ほどの簡略化された一般化モデルのうちどのようなパラメータを用いるのが最適であるかを対数尤度で比較する(4-2節)。そして、2001年７月から2007年12月までの22513レースを用い、実際に複勝馬券についてモデルの確率を計算し、対数尤度で見られたようなモデルの当てはまりが機能しているかどうか検証を行うとともに(4-3節)、期待値が大きいと予測される馬券と期待値が小さいと予測される馬券で実際に回収率に違いが出るかどうかについて購入のシミュレーションを行った(4-4節)。

今回は確率モデルによる予測を行う際、期待値を計算するために、確定オッズ・票数を入力値として与えている。2010年１月18日現在、JRAでの投票締め切り時刻は発走２分前であり、確定オッズ・確定票数をレース前に知ることは不可能である。そのため、4-4節で行う購入シミュレーションの結果は、実際に運用可能な馬券購入支援システムによるものと若干異なることにご留意いただきたい。

4-2 モデル選択の問題

対数尤度については、Lo et al. (1995)の研究同様π_{ijk}の当てはまりを基準としてモデルの選択を行うことにする。はじめに、第3節に記載したように、簡略一般化モデルにλ_2、λ_3を代入していきπ_{ijk}を計算した。そして、簡略一般化モデルのうち各パラメータで規定されるモデルごとに、対数尤度関数の値を計算する：

$$\log(L) = \sum_{i\neq j,k}^{Ns}\sum_{j\neq k}^{Ns}\sum_{k}^{Ns}\sum_{s=1}^{22161}\left(y_s \log \pi_{ijk,s} + (1-y_s)\log\left(1-\pi_{ijk,s}\right)\right)$$

ただし、ここでlog(L)は対数尤度、sは個々のレースを表すインデックス、Nsは各レースにおける出走頭数、y_sはi、j、kが的中したか外れたかを表すカテゴリカル変数(的中＝1、外れ＝0)である。

計算結果を示しているのが、表2である。この結果をみると、モデルとしてはGamma(r, θ_i)型のr=40が最適なモデルとして選択されており、r>10やr→∞としたモデルで値が大きくなっている。この結果はr=4が最適なモデルであるとしたLo et al. (1995)の結論とは大きく異なっている。原因としては、彼らが解析に用いた600レースというサンプル数が十分でなかった可能性がある。

表2　各パラメータに対して計算された対数尤度
(<<は最大値を取っていることを示している)

r	対数尤度	
1(Harville)	-153455.5449	
2	-152599.1096	
3	-152325.6784	
4	-152192.5480	
5	-152149.2427	
6	-152087.7534	
7	-152067.4318	
8	-152037.4763	
10	-152012.0371	
20	-151966.1394	
30	-151956.8861	
40	-151952.9208	<<
∞(Henery)	-152030.6119	

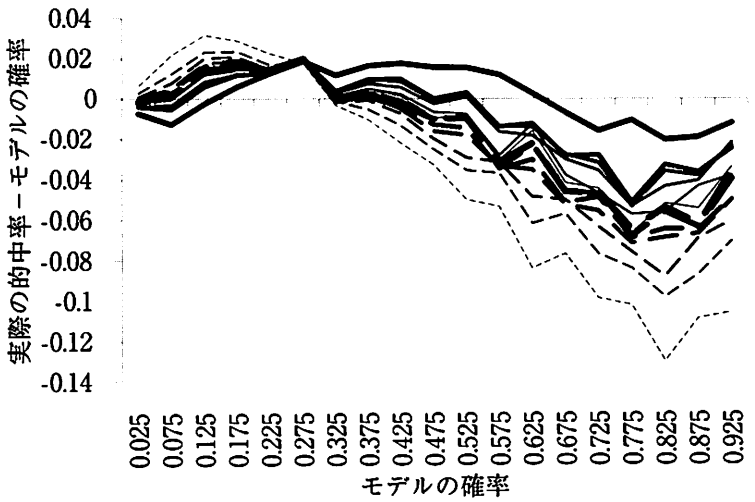

図3　モデルが出力した複勝馬券の的中確率ごとにビンに分類した場合において、実際の複勝馬券の的中率のモデルの出力した確率に対する偏差。破線はr ≦ 7のモデルを示しており、実線はr ≧ 8のモデルであることを意味している。線が太いほどrが大きい。

4-3　複勝馬券の的中確率予測の検証

前小節で、1995年1月から2001年6月までのデータを用いて簡略化モデル中の最適なパラメータを探索した場合、r=40としてパラメータを設定した場合が最適な確率予測を行うモデルとなることが示された。そこで、各モデルで計算された対数尤度が実際の確率予測を反映しているかどうかを確認するため、2001年7月以降のデータを用いて、複勝馬券を各モデルによって計算される的中確率5％ごとのビンに分類し、実際のレース結果における複勝馬券の的中率と比較した。図3の縦軸には、各ビンに含まれる複勝馬券の的中率から、モデルが予測した的中確率のビンごとの中央値を引いた偏差を示している。すなわち、値がゼロに近ければ、モデルの確率は実際の的中率をよく予測したということであり、値が正ならばモデルが出力する確率は複勝馬券の的中率を過大評価し、値が負ならばモデルが出力する確率は複勝馬券の的中率を過小評価していることになる。破線は$1 \leq r \leq 7$、実線は$8 \leq r$を表しており、それぞれ線が太いほどrが大きい。これをみると明らかなように、破線のグループ($1 \leq r \leq 7$)に比べて実線のグループ($8 \leq r$)のほうが値の絶対値は小さくなっており、中でもr＝40としたモデルやHenery型のモデル（$r \to \infty$）で絶対値が小さくなっていることがわかる。これは、4-2節で対数尤度を最大化するパラメータを探索したときに導かれた結論と整合的であり、モデル選択がうまくいっていることを意味している。逆に、Lo et al. (1995)が提案したような

Gamma(4, θ_i)モデルを含め、rが小さい場合には強い馬の複勝馬券の的中確率を過大評価しており、弱い馬の複勝馬券の的中確率を過小評価してしまっている。

4-4 複勝馬券の購入シミュレーション

次にr=40のモデルを採用したときに計算されるモデルの確率と、オッズの情報から、各複勝馬券の期待値Eを予測することを考える。そして、モデルから予測される期待値で、E≦0.2、0.2<E≦0.4、0.4<E≦0.6、0.6<E≦0.8、0.8<E≦1.0、1.0<Eの6つのビンに分け、それぞれの馬券を均等の金額で購入したと仮定したときに得られる回収率を計算した。図4は、その結果を示している。これを見ると、Eが1.0よりも小さい場合には、比較的モデルの期待値が実際の回収率を予測するのに役だっていることがわかる。しかし、Eが1.0より大きい場合には回収率と期待値の対応関係は悪く、モデルの出力する期待値がよい指標となっていないことがわかる。モデルの確率は真の確率とは異なるため、必ずしもモデルで計算された期待値が回収率と等しくなっている必要はない。ただ、問題点を明らかにする意味でも、現在のモデルがどのような特性をもっているかを明らかにしておくことは有用であろう。

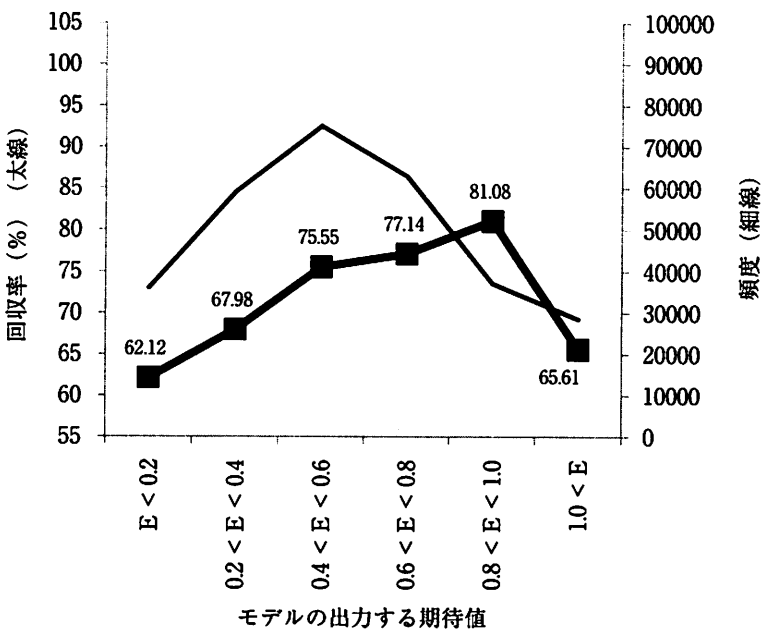

図4 （マーカー付きの太線）モデルの出力する期待値に対して実際に得られた回収率。左軸。マーカー付近の数字は回収率を表す。（細線）各ビンに含まれるサンプル数。右軸。

そこで、一つの試みとして、Lo et al. (1995)にならって、単勝支持率の高い馬のみに限定して値を取り出してみることにする。これは次節で述べるように、単勝支持率が著しく低い場合には、実力を過大評価する可能性があることに関連している。その結果を

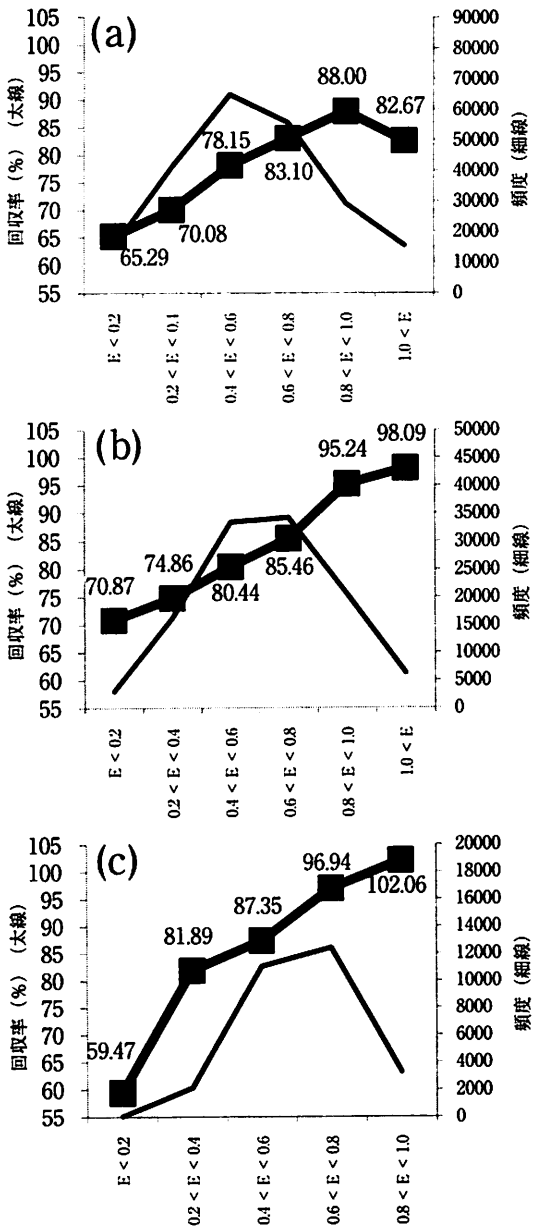

図5　図4と同じ。ただし、単勝支持率が(a)1%以上の馬に限る。(b)5%以上の馬に限る。(c)20%以上の馬に限る。

示したのが図５a、b、cである。単勝支持率が１%以上の馬に限定すると、モデルの出力する期待値と回収率の対応関係はだいぶ改善され、単勝支持率５%以上の馬に限定すると、単調増加の関係となる。この結果をみると、人気馬・中穴の馬に関していえば、モデルの出力する期待値が回収率の良い指標となっており、期待値の増大に伴って、回収率が大きくなっていることがわかる。さらに、単勝支持率が20%を超えるような馬に限定した場合、1.0<Eのカテゴリーで回収率が102.06%に達している。3267回の試行それぞれに100円ずつを賭けたとすると、2047回の複勝馬券が的中し、333,430円の払戻金が得られていたことになる。

今回得られた結果は、統計モデルが期待値の大きな複勝馬券と期待値の小さな複勝馬券をある程度弁別でき、その指標が、回収率が100%を十分超える馬券を見つけ出す手法となりうることを示している。特に、一般にオッズが低い場合には回収率がある値に収束しやすいにも関わらず、場合分けの結果、回収率が非常に高い値となっていた。そのことを踏まえれば、今回の統計モデルによる複勝馬券の確率予測が、ランダム要素によってではなく、本質的に回収率の高い馬券を選び出すシステムとなっていることがわかるだろう。この結果は、一般的に複勝馬券では(1)人気馬の実力が過小評価されており、人気馬の回収率はもともと高いこと、(2)大穴の馬については、モデルの出力する期待値は実力を過大評価することを示している。その意味で、今回選択されたモデルを含めたシステム自体にまだ改善の余地があることがわかる。

以上みてきたように、単勝馬券の人気が第一次近似として実力を反映していることを仮定すると、複勝馬券の的中確率および期待値の大きさをそれなりに分別することができる。このことは、単勝馬券の購入が１着馬を当てればよいという単純なものであるのに対し、複勝馬券では如何に買うべきかという問題を、購入者自身が適切に反映できていない可能性があることを示している。また、本論文の範囲を超えるが、単勝・複勝以外の券種に関しては、ボックス買い・マルチ買いといったまとめ買い行動があり、検討しなければいけない点数が非常に多いため、的中確率に対する人気の偏りが存在する可能性が高い。それらの券種についても、統計的確率モデルによって、実力に対するオッズの偏りが高そうな馬券を抽出することのできる可能性は十分にあるといえるだろう。

5　単勝馬券の支持率は的中率に等しいか？

ここまでは、単勝馬券の的中確率が単勝馬券の支持率に等しいという仮定を用いてき

た。ここでは、その仮定自体が正しいかどうかについて検証しよう。図１に示したように、単勝馬券の支持率が比較的高い馬に関しては、支持率65%以上の馬で若干の過小評価が見られるほかは、母集団のうちの馬券的中率に単勝支持率がほぼ等しく（あるいは若干大きく）なっていることがわかった。今度は単勝支持率を対数スケールのビンで分類して、単勝馬券の支持率が低い馬に注目してみてみることにする（図６）。すると、支持率が著しく小さい場合（１%以下）には、的中率が大きく支持率を下回っており、穴馬を狙った馬券購入心理が過剰にみられることがわかる。4-3節で、人気馬以外の的中確率の予測で失敗していたことは、このことも原因となっていたことが考えられる。さらなる統計モデルの精度向上のためには、支持率と的中率の関係に適切に補正を行うことによって、単勝馬券の支持率が的中確率を推定するためのよりよい関数とすることが必要であろう。

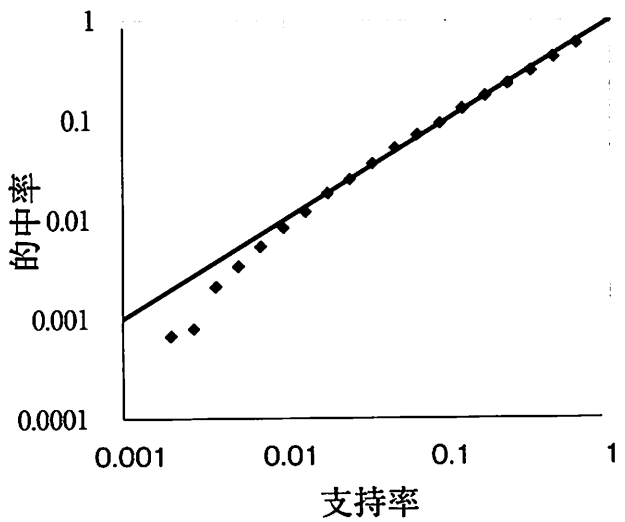

図６　図１に同じ。ただし、両対数軸。

ひとつ注意しておきたいのは、もし単勝馬券の的中率を支持率の関数となるように、適切に補正をかけたとしても、単勝馬券の支持率が真の的中確率を常に適切に反映していることは保証されないということである。というのは、先ほどの図１や図６に示したのはあくまで、ある支持率を持つ馬の母集団にあてはまる統計的な性質であって、個々の馬に対して、単勝の支持率が単勝の的中確率の正しい指標となっているとは限らないからである。単勝の的中確率は、単勝支持率以外のパラメータに依存していたとしても、

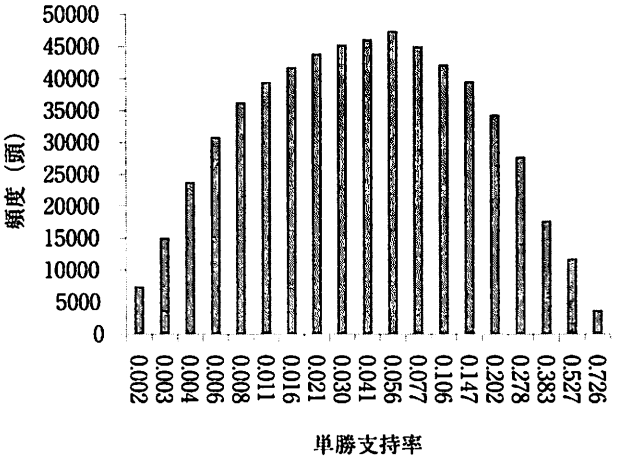

図7　単勝支持率の各ビンに含まれる馬の総数、図6に対応している

その寄与の効果がカテゴライズした集合の中で相殺しあっており、見かけ上現れていない可能性があるわけだ。そして、実際にはどのようなパラメータにどのように依存して値が決まっているかは全くわからないし、より正確に反映させようとするならば、それぞれの指標の相関関係も考慮しなくてはならない。的中確率に影響を与えるパラメータに関しては、それをどのような関数形で的中確率に反映させたらよいかということも含めて、さまざまな提案がありうる。構造方程式を用いた統計ソフトなどである程度参考となる情報を収集することも可能であり、的中確率のモデリングという意味ではさらなるモデルの精度向上を望むことができるだろう。

6　まとめ

本論文では、単勝馬券の支持率が単勝馬券の的中確率に等しいことを仮定するテクニカル分析の立場に立って、複勝馬券の的中確率予測に関する統計モデルについてのレビュー及び日本中央競馬会のデータを用いた再検証を行った。過去に提案されたモデルとしては、Harvilleのモデル・Heneryのモデル、そしてそれらを包含する一般化モデルがあり、このうち、一般化モデルを簡略化した簡略一般化モデルにさまざまなパラメータを与えた複数のモデルから、最適なモデルを選び出す問題について考えた。

過去の研究では、日本の中央競馬についてGamma(r, θ_i)でr=4のときが最適なモデルとされていたが、今回、過去22161レースの結果を用いて再検証した結果、むしろr>10のモデル、もしくは、Henery型モデルのあてはまりがよく、r=40で最適であることが

示された。このことは、r>10とした場合に、22513レース分の実験で予測されたモデルの確率が実際の複勝馬券の的中率をよく反映したことと整合的である。

実際に、モデルが出力する複勝馬券の的中確率を用いて、期待値が大きい馬券と期待値が小さい馬券が分別できるかどうかについて購入のシミュレーションを行った。その結果、単勝支持率が大きなグループに関しては、モデルが予測する期待値が回収率の良い指標となっており、単勝支持率が20%を超える馬の母集団に関しては、モデルが出力する期待値が1.0を超えるグループを購入することによって、回収率が102.06%に達することがわかった。これは、オッズが非常に低い条件の下、3000回以上の試行の結果として得られるものであり、このシステムが本質的に回収率の高い複勝馬券を選びだすことのできるものであることを示している。単勝支持率が著しく低い場合、r=40としたときにモデルの出力する期待値は、適切な回収率の指標とはなっていなかった。このことは、5節で述べたように、著しく支持率が低い場合には、単勝の支持率が単勝馬券の的中確率を過大評価することになってしまうということに大きく関連している。

単勝馬券でも支持率が65%を超える場合や1%を切る場合には、単勝支持率は単勝馬券の的中確率を過大／過小評価する傾向がある。しかし、基本的には単勝馬券の支持率は単勝馬券の的中確率に等しいと仮定することによって、回収率の高い馬券と低い馬券を分別するのに有用であることが示された。単勝馬券の支持率は、的中確率予測に関する統計モデルにとって複勝券種を購入する際の一次情報として、非常に重要な意味を持っていると考えられる。複勝馬券の購入について統計モデルがそれなりに回収率の高低を分別できたのは、複勝馬券の売り上げが必ずしも的中確率を反映していないことに起因していると考えられる。本論文では、複勝馬券以外の式別での議論を展開していないが、馬券の売り上げが実力を必ずしも反映しない券種の場合には、このような統計モデルによる確率予測がうまく働く可能性が考えられる。逆に、単勝馬券のように馬券購入者が情報を適切に反映しており、完全に効率的な市場が形成されていると推測される場合には、谷岡が指摘するように、オッズの高い馬を購入して、当たったら勝ち逃げするという戦略も考えられるだろう[9)]。

さらなるシステムの性能の向上を目指すためにはいくつかのことが考えられる。ひとつは、単勝支持率と単勝馬券の的中確率の関係を適切に補正することであり、また、想定するモデル群をより一般化したものにすることも必要となるかもしれない。そのほか、構造方程式などを用いて、変数間の共分散構造を分析し、単勝支持率以外の因子を

陽に予測のための入力パラメータとして導入することも考えられるが、その場合には赤池情報量基準[10]を積極的に用いて、パラメータの数をモデルのあてはまりのよさに組み込んでいく必要があるだろう。今回示したシステムでは、購入シミュレーションでは確定オッズや確定の支持率を既知として計算しており、実際のシステムの運用とは異なるものになる可能性がある。しかし、本論文の結果は、テクニカル分析の立場に基づいた統計モデルによって、日本の中央競馬会でも、馬券の的中確率予測が行えるということを示すものである。

謝辞

論文執筆のためのデータ利用をこころよく引き受けてくださったJRAシステムサービス株式会社藤倉祐樹様、そして、原稿をチェックしていただいた大阪大学医学系研究科の小幡哲史様にこの場を借りて、感謝申し上げます。

付録　対数尤度関数・情報量規準

ここでは簡単に対数尤度関数と情報量規準についてまとめる。より詳細な導出などに関しては若山ほか[11]などを参照していただきたい。

複数のモデルが提案されたときに、そのモデルの評価基準としていくつかの指標が提案されている。よく用いられるのは、データを発生した真の確率分布との距離をカルバック・ライブラー情報量とよばれる評価基準を用いて予測の観点からモデルの良さをはかるものである。カルバック・ライブラー情報量は真のモデルからランダムにとられたデータの従う分布$g(z)$を、構築した統計モデル$f(z|\hat{\theta})$で予測したときの平均的なよさ、あるいは悪さをはかるものであり（zは真のモデルからとられたデータ; θはパラメータ；$\hat{\theta}$はデータyに依存して推定されたパラメータ）、個々のモデルに依存しない部分を除くと、

$$\int g(z) \log f(z|\hat{\theta}) dz$$

の値が大きなモデルであるほど真の分布に近いことがわかる。ここで、未知の確率分布$g(z)$をn個のデータyに等確率を付与した確率分布であると考える。その結果この量の一つの推定量として、以下のように計算される：

$$\frac{1}{n}\log f(y|\hat{\theta})$$

ここで、log f(y| $\hat{\theta}$)は対数尤度を表している。すなわち、対数尤度がモデルのよさをはかるひとつの指標となっていることを示している。 ただ、厳密に言うと、将来のデータに変えてモデルの推定に用いたデータyを未知の確率分布の推定に再度用いたことからバイアスの補正が必要となる。このバイアスを何らかの方法で評価したものが、情報量規準と呼ばれるものであり、その代表的なものが赤池情報量規準である。この情報量基準では、データを発生した真の確率分布が想定したモデルの中に含まれるという仮定のもと、データ数を無限にするとき、対数尤度のバイアスがモデルに含まれるパラメータの数に近づいていくことを意味しており、パラメータの数をバイアスの補正に用いる。本論文の範囲内では、パラメータの数は変化しないため、対数尤度の大きさをモデルの良さをはかる指標として用いることが可能となっている。

競馬の的中を当てるようなベルヌーイ試行の場合には、pが的中の確率、yがカテゴリカルな確率変数(当たりなら1、外れなら0)を表しているとして確率関数を

$$p^{y}(1-p)^{1-y}$$

のように表わすことができる。X回の試行に対する同時確率関数は

$$\prod_{i=1}^{X} p_i^{y_i}(1-p_i)^{1-y_i}$$

のように計算されるから、対数尤度関数は以下のように計算される。

$$\sum_{i=1}^{X}(y_i \log p_i + (1-y_i)\log(1-p_i))$$

〔注〕

1) Hausch, D. B. (2008), "Efficiency Of Racetrack Betting Markets", World Scientific Publishing Company

2) 月本洋 (1999)、"実践データマイニング—金融・競馬予測の科学"、オーム社。

3) Hausch, D. B., and Rubinstein, M., (1981), "Efficiency of the Market for Racetrack Betting," Management Sci., 27, 1435-1452.

4) Harville, D. A. (1973), "Assigning Probabilities to the Outcomes of Multi-entry Competitions," J. American Statistical Association, 68, 312-316.

5) Henery, R. J. (1981), "Permutation Probabilities as Models for Horse Races," J. Royal Statistical Society B, 43(1), 86-91.
6) Stern, H. (1990), "Models for Distributions on Permutations," J. American Statistical Association, 85, 558-564.
7) Lo, V. S. Y., Bacon-Shone, J., and Busche, K. (1995), "The Application of Ranking Probability Models to Racetrack Betting", Management Science, 41(6), 1048-1059.
8) Kelly, J. L. (1956), "A New Interpretation of Information Rate", Bell System Technical J., 35, 917-926.
9) 谷岡一郎 (1997)、"ツキの法則—「賭け方」と「勝敗」の科学"、PHP新書。
10) Akaike, H., (1973), "Information theory and extension of the maximum likelihood principle", 2nd Inter. Symp. on Information theory, Budapest, 1973.
11) 若山正人編、小西貞則、竹内純一著 (2008)、"統計的モデリング・情報理論と学習理論"、講談社サイエンティフィク