Some Statistical Analysis of Handicap Horse Racing

LAU SIU PING

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Degree of

Master of Philosophy

in

Statistics

July 2001

THE CHINESE UNIVERSITY OF HONG KONG

GRADUATE SCHOOL

The undersigned certify that we have read a thesis, entitled "Some Statistical Analysis of Handicap Horse Racing" submitted to the Graduate School by Lau Siu Ping ( ) in partial fulfillment of the requirements for the degree of Master of Philosophy in Statistics. We recommend that it be accepted.

Prof. M. G. Gu
Supervisor

Dr. K. H. Li

Prof. T. S. Lau

Dr. I. C. Hu
External Examiner

DECLARATION

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my supervisor, Prof. Gu Ming-Gao, for his patience, constant encouragement and guidance during the period of this research program. Further, I would like take this opportunity to give thanks to my husband and my family for their everlasting and hearty support.

ABSTRACT

For so long as there have been horse races, it has given rise to the wide discussion in both academic and public literature. No matter the ordinary people or the academic researchers have intended to find out a profitable wagering system by all means. Some may even spend their whole life span in searching for a wagering strategy with positive return.

My Thesis mainly consists of two parts. First, one of the most popular newspaper supplier was studied and out of which 5 tipsters prediction on horse racing was collected. The accuracy of the tipsters recommendation was firstly accessed by counting the total number of races correctly tipped and by calculating the net profit/deficit for the past two years. It has been proved that no lucre can be gained by following their suggestion. After that, contingency tables were constructed to see if there will be any useful information can be drawn from the tipsters that could help in building our own betting strategies. Second, modeling approach, multinomial logistic model was adopted to predict the winning probability of each horse to see whether a profitable betting strategy can be established. It has been revealed that the statistical model shows a great potential in developing a successful profitable wagering system.

Keywords: Horse races, profitable wagering system, tipsters prediction, multinomial logistic model, winning probabilities

摘要

賽馬， 長久以來， 在學術界及社會上， 都爲人所津津樂道。 無論是學術研究者或普羅大眾， 都試圖以不同的方法， 去尋求一種能獲利的投注系統。 爲了得到這種系統， 有人甚至付上一生的時間去鑽研。

我的論文由兩部份組成。首先，我們 以一份有名的報紙爲資源， 集中研究其中五位馬評家的賽馬預測準確程度。 統計各馬評家在過去兩年所估中的總場數及總利潤/虧損。我們証實了如果只用馬評家的預測將不會獲利的事實。其後， 我們將資料整理成列聯表， 選取有用的部份， 以建立自己的投注策略。 我們用模型方法(多項式羅吉斯模型)去預測每隻馬的勝出機會， 看能否發展出一套成功獲利的投注系統。我們的研究表明：應用統計方法將可以得到一種獲利的投注系統。

Contents

Chapter 1

Introduction

1.1 Pari-Mutuel System

In a pari-mutuel system, punters place wagers on the set of horses in a given race. These wagers form the betting pool and out of which about 17.5%, the HK Jockey Club takes, is deducted. The remainder of the betting pool is then allocated to the winning punters in proportion to their bets. As an example to illustrate, let A_i be the total amount of money betting on the horse i in a given race. A be the total amount of money betting on all horses in that race. Then, the payoff shown in the track's tote board will appear in the exact form of

$$Payoff = \frac{(1 - 0.175) * A}{A_i} * 10$$

per $10 bet. Figure 1.1 shows one of the race run on 24th, Feb 1999.

Win odds there refers to the payoff of "Win" betting.

Note that throughout the whole paper, the value stated will be in $10 base (the minimum wager is $10 per bet as required by the HK Jockey Club) and the word 'odds' will be used as the abbreviation of 'win odds' unless specified. Also, the words 'odds' and 'payoff' will be used interchangeably as they serve the same meaning.

Figure 1.1: Demonstrating the calculation of Win Odds

Turf - "C" Course

Class 4 - (Ratings 56 - 32) - 1650 Meters

Horse	Rank	Draw	Rating	Amount Betted ($'000)	**Win Odds**
Super Conqueror	1	3	50	1,143	20.6
Natural Winner	2	6	43	1,544	15.2
Chiu Shan Glory	3	9	47	8,167	2.9
High Degree	4	2	49	2,465	9.5
Kings Glamour	5	10	50	2,640	8.9
Hollywood Star	6	5	49	2,595	9.1
Lucky Guest	7	4	46	2,154	10.9
Grand Prize	8	11	54	2,025	11.6
Gran Senorum	9	8	45	816	28.8
Win Star	10	1	44	1,085	21.7
Dream Team	11	7	48	709	33.2
Blue Imperial	12	12	52	3,157	7.4
				28,500	

The final track probabilities are proportional to the amounts bet on the horses by all punters. The pari-mutuel probabilities for horse i, ρ_i, can be written as:

$$\rho_i = \frac{A_i}{A}$$

where A_i again is the total amount wagered on horse i by the public and A is the total size of the betting pool. These probabilities represent the public's consensus probabilities as reflected by their wagering preferences.

The values of ρ_i cannot be determined until all the bettors have wagered. However, each bettors's wagering strategy depends on the knowledge of the ρ_i

values to place bets. Therefore, people would always like to wait for the last minutes to bet on their favorite horses.

Define π_i to be the true unknown winning probability associated with the horse i. Then the expected net return of betting on horse i is given by $(\pi_i * payoff - 10)$ per \$10 bet. It is because if we let X_i be the net return on horse i, then X_i would equal to $(payoff - 10)$ with probability π_i and equal to -10 with probability $1 - \pi_i$. Therefore

$$\begin{aligned} E[X_i] &= \pi_i(payoff - 10) + (1 - \pi_i)(-10) \\ &= \pi_i * payoff - 10. \end{aligned}$$

Suppose that the public's consensus probabilities are equal to the true winning probabilities. In such a case, $\pi_i = \rho_i$, it follows that $E[X_i] = -1.75$. Since

$$\begin{aligned} E[X_i] &= \pi_i * payoff - 10 \\ &= \pi_i \frac{(1 - 0.175) * A}{A_i} * 10 \\ &= \pi_i \frac{8.25}{\rho_i} - 10 \\ &= -1.75. \end{aligned}$$

So, if we assume $\pi_i = \rho_i$, it does not matter which horse the bettor wagers on, he will always expected to lose the track takes, 17.5%. In principle, then, it is possible for a bettor to expect to discover a betting procedure that yields positive returns only when the public misestimate the true winning probabilities i.e., when $\pi_i \neq \rho_i$. Positive returns at the track are only possible when $\pi_i * payoff > 10$.

1.2 Different Types of Betting

In HK, there are totally 11 kinds of betting in horse racing. They are namely Win, Place, Quinella, Quinella Place, Tierce, Trio, Double Trio, Triple Trio, Double, Treble and Six Up. Their meanings have been shown in the next page.

In 1998/99 and 1999/00, there were 1264 races carried out altogether during the racing seasons. For "Win" betting only, there was already about $24 million wager per race and it was about $16 and $43 million for "Quinella" and "Quinella Place" betting respectively. The total wager, regardless of the betting types, for only one race would then become $248 million instead. Whereas in a yearly basis, only a "Win" betting already contributed about $15 billion wager to the HK Jockey Club and it was about $10 and $26 billion from the "Quinella" and "Quinella Place" betting as well. What's more incredible is that, it was about $176 billion wager on an entire racing season. Table shown below gives some more information in other betting types for reference.

Table 1.1: Wagers on different betting

Betting	per race ($'000,000)	per year ($'000,000,000)
Win	24	15
Place	16	10
Quinella	43	26
Quinella Place	23	14
Tierce	17	10
Double Trio	23	14
On Course Double Trio	3	2
Triple Trio	36	22
Double	45	28
Treble	6	4
Six Up	12	7
Total	248	176

Table 1.2: Details for different types of betting

betting on 1 horse in a race	
Win	finish in the 1^{st} position
Place	get 1^{st}, 2^{nd} or 3^{rd} in a race of 7 or more declared starters or get 1^{st} or 2^{nd} in a race of 6 declared starters
betting on 2 horses in a race	
Quinella	get 1^{st} and 2^{nd} in either order
Quinella Place	get any two out of the first 3 winning horses in any order
betting on 3 horses in a race	
Tierce	get 1^{st}, 2^{nd} and 3^{rd} in correct order
Trio	get 1^{st}, 2^{nd} and 3^{rd} in any order
betting on 2 nominated races with 3 horses each	
Double Trio	get 1^{st}, 2^{nd} and 3^{rd} in both races in any order
On Course Trio	same as above
betting on 3 nominated races with 3 horses each	
Triple Trio	get 1^{st}, 2^{nd} and 3^{rd} in all 3 races in any order
(Consolation)	get 1^{st}, 2^{nd} and 3^{rd} in the 1^{st} two races in any order
betting on 2 nominated races with 1 horse each	
Double	get 1^{st} in both races
(Consolation)	get 1^{st} in the first race and 2^{nd} in the second race
betting on 3 nominated races with 1 horse each	
Treble	get 1^{st} in all 3 races
betting on 6 nominated races with 2 horses each	
Six Up	get 1^{st} or 2^{nd} in each of the six races
Bonus	get 1^{st} in all 6 races
Five Win Bonus	get 1^{st} in the first five races

1.3 Overview

In this thesis, we investigate two approaches concerning wagering on the popular Hong Kong Handicap horse racing.

The first approach is the tipster's approach. One of the HK's most popular newspaper supplier was studied and out of which 5 tipsters recommendation on horse racing were collected for 2 years. With this approach, we examine whether following a particular tipster one would actually do better than the average. We also compare which tipster gives the "best tips" among them.

The summary tables on the accuracy of the tipsters suggestion by counting the number of correctly tipped races and by calculating the total money earned/lost throughout a year would be given in Section 2.2. After that, the strategy of betting on the horse according to the tipsters prediction is adopted, under the restriction of wagering one horse per race. The result is then tested against with that of random bet. It would be the content of Section 2.3.

In our second approach, a statistical model based on multinomial logistic regression is developed to predict the outcome of each race. In Chapter 3, we focus on developing this model for the horse races of HK using the data between 98-00. The multinomial logit model proposed by Bolton and Chapman in 1986 is used for this purpose on the data base of 1264 races. The main reason for adopting it is owing to its simplicity and explicity in expression. It explicitly recognizes there are only finite number of horses competing in a race and only one of the entered horses win. The parameters of the model are estimated by the maximum likelihood method and its main output is a prediction of the winning probabilities for each horse. And these probabilities are then used as input to our wagering strategy that determines which horses should be betted on.

In Section 3.3, an out-of-sample simulation is performed to assess the accuracy of the proposed models in predicting the outcomes of horse races.

Finally in Section 3.4, a simple graphical comparison is carried out among the models to see which model and under what constraint the model do can perform the best.

It has been shown that with limited predictors, the statistical model shows great potential for developing a successful profitable wagering system.

Chapter 2

Testing on Tipsters Prediction

2.1 Introduction

One of the most popular newspaper distributor was studied and out of which 5 tipsters prediction on horse racing was recorded for 2 years. Each tipster on each race predicted four horses that they believed would have the greatest chance to win. The indicator '1' is marked beside the horse i under the variable 'Tipster 1' if tipster 1 had firstly suggested horse i to bet on, marked '2' beside the second one he recommended and so on. Similar procedure is proceeded for the other tipsters and the data is described below:

Table 2.1: Tipsters Prediction On Each Race

Date	raceno	horseno	Tipster 1	Tipster 2	Tipster 3	Tipster 4	Tipster 5	Odds
09/06/98	1	1			4	4		100
09/06/98	1	2						170
09/06/98	1	3		3		1		140
09/06/98	1	4	1	1	2	3		69
09/06/98	1	5					4	120
09/06/98	1	6						990
09/06/98	1	7	2				3	180
09/06/98	1	8						970
09/06/98	1	9	4	2	1	2	1	25
09/06/98	1	10		4				420
09/06/98	1	11						93
09/06/98	1	12						780
09/06/98	1	13	3		3		2	81
09/06/98	1	14						260

We will try to see whether it is profitable to follow the 'tips' given by the tipsters. It is done by betting on the horses according to the tipsters prediction for each race.[1] Their performance is firstly accessed by examining their accuracy on prediction by counting the number of correctly tipped races and by calculating the net profit/deficit on a yearly basis. The details will be presented in the next Section. After that, in Section 2.3, it is then accessed by comparing with the strategy of random bet to see whether betting based on the tipsters recommendation can give a better performance. A hypothesis testing will be set to test for the equality between tipsters suggestion and random bet, if H_o is accepted, it will simply mean there is no difference between random betting and following tipster predictions. Before proceeding to the next section, let's have a look to some descriptive statistics for the races during 1998 - 2000 first.

Racing Information for 98-00

There were total of 616 and 658 races taken place in 1998/99 and 1999/00 respectively. About two-third of the races were carried out in Shatin and almost 90% of them were going on the ground 'Turf'. Besides, the distance for running are divided into 11 types: 1000m, 1150m, 1200m, 1400m, 1600m, 1650m, 1800m, 1900m, 2000m, 2200m and 2400m. Half of the races were running with distance less than 1600m. All horses are assigned with a number called "Rating" which reflexes their overall performance in racing. The better their performance is, the higher the number would be. The horses with different ratings will then be allocated to the different classes according to the followings:

Class	1	2	3	4	5	6
Ratings	80+	88 - 64	72 - 48	56 - 32	40 - 16	24 - 0

[1]Only the first horse suggested is used which means only one horse for each race is betted on

Left hand side of Table 2.2 shows the number of races carried in different aspects such as venue, ground and distance. While for different classes, they are displayed in the right side correspondently.

Table 2.2: Number of races carried in different aspects

	98/99	99/00		98/99	99/00
Shatin	397	441	others	48	42
Happy Valley	219	217	Class 1	95	75
Turf	551	587	Class 2	53	59
All Weather Track	65	71	Class 3	140	145
Distance $<$ 1600m	318	344	Class 4	175	205
Distance $\geq$ 1600m	298	314	Class 5	83	112
			Class 6	22	20

"others" here refers to those not belonging to any of the classes. It is the case when the international races occur.

2.2 Summary Tables on Tipsters Performance

Notice that about 70% of the winning horses were having the odds less than 100 and about 90% were less than 200. Therefore, it is not surprising to see the tipsters prediction show the same patten: more than 70% of the horses suggested by the tipsters were with the odds less than 100 and more than 90% were less than 200 as well. Reminded that the odds of a horse refers to the return to the bettor for every $10 bet when that horse got the first place in a race. It indicates that most of the time the tipsters would tend to recommend the favourite horses rather than the others.

In fact, there are several ways to follow the suggestion of the tipsters, two more methods have been considered in checking their profitability over horse racing. They are:

- **Favourite** : Betting on the horse which has the lowest odds.
- **Joint** : Betting on those having two or more tipsters recommended simultaneously.

The overall performance of the tipsters and the above two methods by counting the number of correctly tipped races and total amount earned/lost within these two years are given in Table 2.4 to Table 2.7 accordingly. For the ratio of winning in different aspects are also described in Table 2.8 and Table 2.9. Interested reader can refer to Appendix I for more detail information on each individual tipsters performance.

Apart from concerning the number of correctly tipped races, we are also interested in knowing under what circumstances one tipster can perform better than the others, for example, could tipster 1 do better when the race is carried out in Shatin? In order to determine if such kind of phenomenon really exists, Chi-square tests on 7×2 contingency tables are carried out to investigate the independency among the tipsters win ratio in different aspects.

Table 2.3: Contingency Table for Shatin in 98/99

	Win	Lost	Total
Tipster 1	73	313	386
Tipster 2	66	315	381
Tipster 3	56	330	386
Tipster 4	78	307	385
Tipster 5	61	324	385
Favourite	102	295	397
Joint	91	370	461

Table 2.4: Overall Performance in terms of no. of races for 98/99

98/99	Tipster 1	Tipster 2	Tipster 3	Tipster 4	Tipster 5	Favourite	Joint
No. of betting	597	592	597	597	596	616	707
Win	113	93	93	108	85	137	131
Lost	484	499	504	489	511	479	576
Win Ratio	0.19	0.16	0.16	0.18	0.14	0.22	0.19

Table 2.5: Overall Performance in terms of total money earned/lost for 98/99

98/99	Tipster 1	Tipster 2	Tipster 3	Tipster 4	Tipster 5	Favourite	Joint
Total $ Invested	5970	5920	5970	5970	5960	6160	7070
$ Win	4058	3659	3153	3756	2696	3188	3884
$ Lost	4840	4990	5040	4890	5110	4790	5760
Profit/Deficit	-782	-1331	-1887	-1134	-2414	-1602	-1876
Return per $10 bet	-1.31	-2.25	-3.16	-1.90	-4.05	-2.60	-2.65

Table 2.6: Overall Performance in terms of no. of races for 99/00

99/00	Tipster 1	Tipster 2	Tipster 3	Tipster 4	Tipster 5	Favourite	Joint
No. of betting	655	653	650	655	656	650	783
Win	123	118	119	112	121	167	165
Lost	532	535	531	543	535	483	618
Win Ratio	0.19	0.18	0.18	0.17	0.18	0.26	0.21

Table 2.7: Overall Performance in terms of total money earned/lost for 99/00

99/00	Tipster 1	Tipster 2	Tipster 3	Tipster 4	Tipster 5	Favourite	Joint
Total $ Invested	6550	6530	6500	6550	6560	6500	7830
$ Win	4099	4422	3645	3640	3693	3797	4948
$ Lost	5320	5350	5310	5430	5350	4830	6180
Profit/Deficit	-1221	-928	-1665	-1790	-1657	-1033	-1232
Return per $10 bet	-1.86	-1.42	-2.56	-2.73	-2.53	-1.59	-1.57

Table 2.8: Ratio of Winning in different aspects for 98/99

	Tipster 1	Tipster 2	Tipster 3	Tipster 4	Tipster 5	Favourite	Joint
Overall	0.19	0.16	0.16	0.18	0.14	0.22	0.19
Shatin	0.19	0.17	0.15	0.20	0.16	0.26	0.20
Happy Valley	0.19	0.13	0.18	0.14	0.11	0.16	0.16
Turf	0.18	0.15	0.15	0.17	0.14	0.22	0.18
AWT	0.26	0.22	0.23	0.23	0.14	0.28	0.24
Distance $<$ 1600m	0.22	0.17	0.17	0.20	0.16	0.27	0.20
Distance $\geq$ 1600m	0.15	0.14	0.14	0.16	0.12	0.17	0.16
Odds $<$ 100	0.20	0.17	0.17	0.19	0.15	0.22	0.19
Odds $\geq$ 100	0.11	0.08	0.04	0.11	0.05	0	0.18
others	0.27	0.27	0.20	0.29	0.29	0.35	0.30
Class 1	0.22	0.21	0.14	0.19	0.15	0.28	0.21
Class 2	0.12	0.14	0.15	0.17	0.13	0.23	0.17
Class 3	0.25	0.2	0.16	0.17	0.16	0.17	0.21
Class 4	0.125	0.08	0.17	0.16	0.1	0.18	0.13
Class 5	0.175	0.11	0.11	0.14	0.13	0.28	0.17
Class 6	0.23	0.18	0.14	0.27	0.09	0.14	0.21

Table 2.9: Ratio of Winning in different aspects for 99/00

	Tipster 1	Tipster 2	Tipster 3	Tipster 4	Tipster 5	Favourite	Joint
Overall	0.19	0.18	0.18	0.17	0.18	0.26	0.21
Shatin	0.19	0.19	0.20	0.19	0.19	0.28	0.22
Happy Valley	0.19	0.15	0.16	0.14	0.18	0.21	0.20
Turf	0.20	0.18	0.20	0.18	0.19	0.26	0.22
AWT	0.10	0.17	0.04	0.13	0.10	0.21	0.10
Distance $<$ 1600m	0.17	0.18	0.20	0.19	0.19	0.26	0.21
Distance $\geq$ 1600m	0.21	0.18	0.17	0.15	0.17	0.25	0.22
Odds $<$ 100	0.21	0.21	0.20	0.19	0.21	0.26	0.22
Odds $\geq$ 100	0.05	0.07	0.03	0.03	0	0	0.06
others	0.24	0.17	0.24	0.19	0.27	0.29	0.22
Class 1	0.23	0.24	0.27	0.19	0.13	0.24	0.24
Class 2	0.16	0.24	0.14	0.12	0.15	0.24	0.18
Class 3	0.14	0.14	0.17	0.18	0.21	0.27	0.20
Class 4	0.22	0.19	0.21	0.18	0.20	0.27	0.22
Class 5	0.18	0.13	0.11	0.16	0.14	0.23	0.22
Class 6	0.1	0.25	0.1	0.15	0.25	0.2	0.13

It can be seen from Table 2.4 & 2.6 that no more than 20% of the tipsters prediction were accurate. Simply betting on the most favourite horse (with the lowest odds) may have more chance to win. Besides, strategy **Favourite** performs the best in the sense of possessing the highest winning ratio in almost all aspects. Especially when the races were taken place in Shatin or under the track 'Turf' shows a significant difference with the others' win ratio. When the competing horses belonging to the 'others', strategy **Favourite** also gets a significantly high win ratio in predicting the true ones. However, it would be imprudent to make any conclusion without taking 'money' into account. Although it seems the strategy **Favourite** shows a prominent performance, it does not overwhelm the others if they are weighed in terms of total money earned/lost throughout the year. Actually, none of them shows any positive return. On average, they are losing \$2.3 for every \$10 bet instead.

2.3 Tipsters Prediction Vs Random Betting

In this section, the strategy of betting base on tipsters prediction will be tested against with the strategy of random betting. It is our interest to test whether tipsters suggestion is the same as or worse than betting randomly. That is, our null hypothesis becomes

H_0 : Average Return of Tipsters Suggestion = Average Return of Random Betting

It has been observed in several studies that the empirical probability π_j that horse j wins can be accurately estimated by its pari-mutuel probabilities. In Benter's paper (1994), he has shown that the public's betting creates an unbiased and formidably accurate estimate of the horses true winning probabilities.

Range	N	win exp	win act	freq exp	freq act
0.00 - 0.01	4906	32	35	0.007	0.007
0.01 - 0.03	15079	297	333	0.020	0.022
0.03 - 0.06	16512	729	775	0.044	0.047
0.06 - 0.10	15082	1186	1138	0.079	0.075
0.10 - 0.15	11788	1444	1476	0.123	0.125
0.15 - 0.20	6063	1044	990	0.172	0.163
0.20 - 0.25	3394	752	743	0.222	0.219
0.25 - 0.35	2770	800	797	0.289	0.288
0.35 - 0.45	736	288	296	0.392	0.402
0.45 - 1.0	295	152	143	0.517	0.485
0.00 - 1.0	76625	6726	6726	0.088	0.088

Table above is constructed from 6726 Hong Kong races carried out between Sept 1986 and Jan 1999.

Under this assumption, the probability for horse j in race i to win is equal to $\frac{a_{ij}}{A_i}$, where a_{ij} is the total amount wagered on horse j in race i by the public and A_i is the total size of the betting pool in race i. We will first show that under random betting, the expected return per every \$10 bet would be equal to -\$1.75.

Let

X_i be the net return per \$10 bet at ith race

S $= (X_1 + \cdots + X_R)/R$

be the average return on a racing season with R races carried out

A_i be the total size of betting pool for race i

a_{ij} be the amount betted on horse j at race i

$O_i(j)$ be the odds per \$10 bet at race i given horse j wins

$= \frac{A \times (1-0.175) \times 10}{a_{ij}}$

$= 8.25 \frac{A_i}{a_{ij}}$

$P_i(j)$ be the probability for horse j to win at ith race

$h_i = j$ means horse j is picked to bet on at race i

Remember that under random betting, since the horse picked to bet on is randomly selected, $P(h_i = j) = 1/n_i$ for $j = 1, \cdots, n_i$ where n_i is the number of starters in race i and for $i = 1, \cdots, R$.

Therefore, in the case of random betting,

$$\begin{aligned} E(X_i|h_i = j) &= [O_i(j) - 10]P_i(j) - 10[1 - P_i(j)] \\ &= O_i(j)P_i(j) - 10P_i(j) - 10 + 10P_i(j) \\ &= O_i(j)P_i(j) - 10 \\ &= -1.75 \end{aligned}$$

which means

$$E(X_i) = E[E(X_i|h_i = j)] = -1.75.$$

Now consider

$$var(S) = var\{(X_1 + \cdots + X_R)/R\} \quad = \tfrac{1}{R^2} \sum_{i=1}^{R} var(X_i).$$

Since

$$var(X_i) = var[E(X_i|h_i = j)] + E[var(X_i|h_i = j)]$$

where

$$var[E(X_i|h_i = j)] = 0$$

and

$$\begin{aligned} var(X_i|h_i = j) &= E(X_i|h_i = j)^2 - [E(X_i|h_i = j)]^2 \\ &= [O_i(j) - 10]^2 P_i(j) + 100[1 - P_i(j)] - 1.75^2 \\ &= O_i^2(j)P_i(j) - 20O_i(j)P_i(j) + 100 - 1.75^2 \\ &= 8.25O_i(j) - 20(8.25) + 100 - 1.75^2 \\ &= 8.25O_i(j) - 68.0625 \end{aligned}$$

which implies

$$\begin{aligned} E[var(X_i|h_i = j)] &= 8.25E[O_i(j)] - 68.0625 \\ &= 8.25(92.065) - 68.0625 \\ &= 691.477 \end{aligned}$$

where $E[O_i(j)]$ is calculated from an empirical distribution of the odds picked by the tipsters from 2 years data.

Therefore,

$$\begin{aligned} var(X_i) &= 691.477 \\ var(S) &= \frac{691.477}{R}. \end{aligned}$$

By the Central Limit Theorem, under H_o, $S \sim N\{-1.75, var(S)\}$.

Let $s_1 \cdots s_5$ be the test statistics for Tipster 1 to Tipster 5's average return on a racing season respectively. Therefore, we will reject H_o if, and only if $s_i + 1.75 > 1.96^* \sqrt{var(S)}$ when the alternative hypothesis is 'better than random bet' or reject H_o if, and only if $s_i + 1.75 < 1.96^* \sqrt{var(S)}$ when H_a is 'worse than random bet'.

The results of the testing are shown in the table below:

Table 2.12: Result tables on Hypothesis Testing for 98-00

98/99				99/00			
Tipster	Test Statistic	H_a	p-value	Tipster	Test Statistic	H_a	p-value
1	-1.31	better	0.34	1	-1.86	worse	0.46
2	-2.25	worse	0.32	2	-1.42	better	0.37
3	-3.16	worse	0.09	3	-2.56	worse	0.22
4	-1.90	worse	0.44	4	-2.73	worse	0.17
5	-4.05	worse	0.02	5	-2.53	worse	0.22

It can be seen that Tipster 3 and 5 gave even worse performance in 98/99. To conclude, none of them can perform better than a random bet and some may even give poorer results.

Chapter 3

Multinomial Logistic Regression

A prerequisite for having a profitable wagering strategy is to develop accurate prediction on the probability of the horse race outcomes. Therefore, the model which can assign the probabilities accurately would be our utmost concern in this chapter. It is our interest to see whether a profitable betting system could be devised under our proposed models.

3.1 Review

Suppose that the dependent variable Y can only take m responses which we will number with the categorical labels $1, 2, \cdots, m$. Let π_{ij} denote the probabilities that the ith observation gives the jth response, that is, $\pi_{ij} = P(Y_i = j)$, for $j = 1, \cdots, m$.

Assume there are k covariates, $X_i, \cdots, X_k$, on which the π_{ij} depend. Then the multinomial logistic model is given by:

$$\pi_{ij} = \frac{\exp(\beta_{0j} + \beta_{1j}X_{i1} + \cdots + \beta_{kj}X_{ik})}{1 + \sum_{l=1}^{m-1} \exp(\beta_{0l} + \beta_{1l}X_{i1} + \cdots + \beta_{kl}X_{ik})} \qquad for \;\; j = 1, \cdots, m-1$$

$$\pi_{im} = 1 - \sum_{l=1}^{m-1} \pi_{ij} \qquad for \;\; j = m.$$

Therefore, there is one set of parameters, $(\beta_{0j}, \beta_{1j}, \cdots, \beta_{kj})$, for each dependent

variable category except the last one.

Parameter Estimation

To fit the model to the data, there involves the maximum likelihood method. Since each Y_i takes on its possible values $1, 2, \cdots, m$ with probabilities $\pi_{i1}, \pi_{i2}, \cdots, \pi_{im}$, define indicator variables $W_{i1}, \cdots, W_{im}$, as suggested by Nerlove and Press (1973) with $W_{ij} = 1$ if $Y_i = j$ and $W_{ij} = 0$ if $Y_i \neq j$. Thus

$$\begin{aligned} p(y_i) &= \pi_{i1}^{W_{i1}} \pi_{i2}^{W_{i2}} \cdots \pi_{im}^{W_{im}} \\ &= \prod_{j=1}^{m} \pi_{ij}^{W_{ij}}. \end{aligned}$$

Assume the observations are sampled independently, then their joint probability distribution is given by

$$\begin{aligned} p(y_1, \cdots, y_n) &= p(y_1) \cdots (y_n) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{m} \pi_{ij}^{W_{ij}}. \end{aligned}$$

Let

$$X_{n\times(k+1)} = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix}$$

where

$$x_i' = (1, X_{i1}, \cdots, X_{ik}).$$

The probabilities of the data conditional on X is therefore,

$$p(y_1, \cdots, y_n | X) = \prod_{i=1}^{n} \prod_{j=1}^{m} \left(\frac{\exp(x_i' \beta_j)}{1 + \sum_{l=1}^{m-1} \exp(x_i' \beta_l)} \right)^{W_{ij}}$$

where

$$\beta_j = (\beta_{0j}, \beta_{1j}, \cdots, \beta_{kj})'.$$

The constraint $\sum_{j=1}^{m} \pi_{ij} = 1$ can be imposed by setting $\beta_m = 0$. The log likelihood becomes

$$\begin{aligned} \log L(\beta_l, \cdots, \beta_{m-1}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} W_{ij}\left[x_i'\beta_j - \log\left(1 + \sum_{l=1}^{m-1}\exp(x_i'\beta_l)\right)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m-1} W_{ij}x_i'\beta_j - \sum_{i=1}^{n}\log\left(1 + \sum_{l=1}^{m-1}\exp(x_i'\beta_l)\right) \end{aligned}$$

because $\sum_{j=1}^{m} W_{ij} = 1$.

Differentiating the above function with respect to the parameters and setting them to 0 yields

$$\begin{aligned} \sum_{i=1}^{n} W_{ij}x_i &= \sum_{i=1}^{n} x_i \frac{\exp(x_i'b_j)}{1 + \sum_{l=1}^{m-1}\exp(x_i'b_l)} \qquad for \;\; j = 1, \cdots, m-1 \\ &= \sum_{i=1}^{n} P_{ij}x_i \end{aligned}$$

with b_j's are the MLE of β_j's. The fitted probabilities are

$$P_{ij} = \frac{\exp(x_i'b_j)}{1 + \sum_{l=1}^{m-1}\exp(x_i'b_l)}.$$

Since $\mathbf{b}$ is a MLE, its estimated asymptotic covariance matrix can be achieved by taking the inverse of the information matrix. Let

$$\beta_{(m-1)(k+1)\times 1} = (\beta_1', \cdots, \beta_{m-1}')'.$$

The information matrix is

$$\mathbf{I}(\beta)_{(m-1)(k+1)\times(m-1)(k+1)} = \begin{bmatrix} \mathbf{I}_{1,1} & \mathbf{I}_{1,2} & \cdots & \mathbf{I}_{1,m-1} \\ \mathbf{I}_{2,1} & \mathbf{I}_{2,2} & \cdots & \mathbf{I}_{2,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{I}_{m-1,1} & \mathbf{I}_{m-1,2} & \cdots & \mathbf{I}_{m-1,m-1} \end{bmatrix}$$

where

$$\begin{aligned}\mathbf{I}_{jj(k+1)\times(k+1)} &= -E\left[\frac{\partial^2 \log L(\beta)}{\partial\beta_j \partial\beta_j'}\right] \\ &= \sum_{i=1}^{n} \frac{x_i x_i' \exp(x_i'\beta_j)[1+\sum_{l=1}^{m-1}\exp(x_i'\beta_l) - \exp(x_i'\beta_j)]}{[1+\sum_{l=1}^{m-1}\exp(x_i'\beta_l)]^2}\end{aligned}$$

and

$$\begin{aligned}\mathbf{I}_{jj*(k+1)\times(k+1)} &= -E\left[\frac{\partial^2 \log L(\beta)}{\partial\beta_j \partial\beta_{j*}'}\right] \\ &= -\sum_{i=1}^{n} \frac{x_i x_i' \exp[x_i'(\beta_{j*}+\beta_j)]}{[1+\sum_{l=1}^{m-1}\exp(x_i'\beta_l)]^2}\end{aligned}$$

evaluated at β=**b**.

For an individual coefficient, testing the null hypothesis $H_0 : \beta_j = \beta_j^{(0)}$ can be accomplished by calculating the Wald Statistics

$$Z_0 = \frac{b_j - \beta_j^{(0)}}{A\hat{S}E(b_j)}$$

where $A\hat{S}E(b_j)$ is the estimated asymptotic standard deviation of b_j.

By comparing $\log L_0$ for the model containing only the constant with $\log L_1$ for the full model, we can measure the degree of improvement in predictability by using the independent variables. It is done by evaluating

$$R^2 = 1 - \frac{\log L_1}{\log L_0}$$

which is analogous to R^2 in a linear model.

3.2 Proposed Models for the Horse Racing

Before constructing a multinomial logistic model to predict the winning probabilities for each horse, a function used to measure the utility of each individual horse has to be developed first. Let

$$U_h = V_h + \varepsilon_h \qquad\qquad -\infty < \varepsilon < +\infty$$

where V_h is comprised of independent variables which are thought to be related to or affecting the utility of the horses, whereas ε_h refers to the error term which accounts for those cannot be covered by the model. Each error term is independent and identically distributed according to the double exponential distribution ,i.e $\varepsilon_h \sim F_{\varepsilon_h}(\varepsilon_h) = \exp(-\exp(\varepsilon_h))$. Suppose that the horse h^* is observed to win in a race which means h^* achieves the greatest utility comparatively to all the other competing horses. This implies $U_{h^*} \geq U_h$ for $h = 1, \cdots, n$ where n is the number of starters in that race.

Since

$$\varepsilon_h \sim F_{\varepsilon_h}(\varepsilon_h) \Longrightarrow U_h \sim F_{U_h} = F_{\varepsilon_h}(U_h - V_h)$$

Therefore, the joint distribution function of $(U_1, \cdots, U_n)$ is

$$\prod_{h=1}^{n} F_{\varepsilon_h}(U_h - V_h)$$

Then,

$$\begin{aligned}
P_{h^*} &= P(U_{h^*} \geq U_h; h = 1, \cdots, n; h \neq h^*) \\
&= \int_{\substack{U_h \leq U_{h^*} \\ h \neq h^*}} d\prod_{h=1}^{n} F(U_h - V_h) \\
&= \int_{-\infty}^{+\infty} [\prod_{\substack{h=1 \\ h \neq h^*}}^{n} \int_{U_h \leq U_{h^*}} dF(U_h - V_h)] dF(U_{h^*} - V_{h^*}) \\
&= \int_{-\infty}^{+\infty} \prod_{\substack{h=1 \\ h \neq h^*}}^{n} [F(U_{h^*} - V_h) - F(-\infty - V_h)] dF(U_{h^*} - V_{h^*})
\end{aligned}$$

$$
\begin{aligned}
&= \int_{-\infty}^{+\infty} \prod_{\substack{h=1 \\ h \neq h^*}}^{n} F(U_{h^*} - V_h) dF(U_{h^*} - V_{h^*}) \\
&= \int_{-\infty}^{+\infty} \prod_{\substack{h=1 \\ h \neq h^*}}^{n} \exp[-\exp(-U_{h^*} + V_h)] d\exp[-\exp(-U_{h^*} + V_{h^*})] \\
&= -\int_{-\infty}^{+\infty} \prod_{h=1}^{n} \exp[-\exp(-U_{h^*} + V_h)] d\exp(-U_{h^*} + V_{h^*}) \\
&= \exp(V_{h^*}) \int_{0}^{+\infty} \exp\{-\sum_{h=1}^{n} t \exp(V_h)\} dt \qquad\qquad t = \exp(-U_{h^*}) \\
&= \exp(V_{h^*}) \int_{0}^{+\infty} \exp\{-t \sum_{h=1}^{n} \exp(V_h)\} dt \\
&= \frac{\exp(V_{h^*})}{\sum_{h=1}^{n} \exp(V_h)}.
\end{aligned}
$$

Therefore, it ends up in a specific, closed-form of multinomial logit model with

$$P_{h^*} = \frac{\exp(V_{h^*})}{\sum_{h=1}^{n} \exp(V_h)} \qquad\qquad for \qquad h^* = 1, \cdots, n.$$

Note that this is a special model of the multinomial logistic regression introduced in Section 3.1.

Under such specific model, the likelihood function using only the winning horse in each race can be written as

$$L = \prod_{i=1}^{R} P_{ih^*} = \prod_{i=1}^{R} \frac{\exp(V_{ih^*})}{\sum_{j=1}^{n_i} \exp(V_{ij})}$$

where R is the total number of races run and n_i is the number of starters in race i. V_{ij} refers to the utility of the jth horse in the ith race. Consequently, the log likelihood of the model can be obtained as

$$\log L = \sum_{i=1}^{R} \{V_{ih}^* - \log[\sum_{j=1}^{n_i} \exp(V_{ij})]\}.$$

Besides, the asymptotic covariance matrix of the estimated parameters can also be approximated by taking the inverse of the information matrix as follows:

Consider $V_{ij} = \beta_1 X_{1,ij} + \beta_2 X_{2,ij} + \cdots + \beta_k X_{k,ij}$ which implies

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{R} \left\{ X_{,ij} - \frac{\sum_{j=1}^{n_i} X_{,ij} \exp(X_{,ij}^{'}\beta)}{\sum_{j=1}^{n_i} \exp(X_{,ij}^{'}\beta)} \right\}_{k\times 1}$$

where

$$X_{,ij} = (X_{1,ij}, X_{2,ij}, \cdots, X_{k,ij})^{'} \qquad \& \qquad \beta = (\beta_1, \beta_2, \cdots, \beta_k)^{'}.$$

Let

$$\begin{aligned} A_{k\times 1} &= \sum_{j=1}^{n_i} X_{,ij} \exp(X_{,ij}^{'}\beta), \\ B &= \sum_{j=1}^{n_i} \exp(X_{,ij}^{'}\beta). \end{aligned}$$

Then,

$$\begin{aligned} \frac{\partial \log L}{\partial \beta} &= \sum_{i=1}^{R} \{X_{,ih^*} - \frac{A}{B}\}, \\ \frac{\partial^2 \log L}{\partial \beta^2} &= -\sum_{i=1}^{R} \left\{ \frac{(\frac{\partial A}{\partial \beta})}{B} + \frac{\partial(\frac{1}{B})}{\partial \beta} A^{'} \right\}, \end{aligned}$$

where

$$\begin{aligned} \frac{\partial A}{\partial \beta} &= \sum_{j=1}^{n_i} X_{,ij} \exp(X_{,ij}^{'}\beta) X_{,ij}^{'}, \\ \frac{\partial(\frac{1}{B})}{\partial \beta} &= -\frac{\sum_{j=1}^{n_i} X_{,ij} \exp(X_{,ij}^{'}\beta)}{[\sum_{j=1}^{n_i} \exp(X_{,ij}^{'}\beta)]^2}. \end{aligned}$$

Therefore,

$$\mathbf{I}(\beta)_{k\times k} = \sum_{i=1}^{R} \left\{ \frac{\sum_{j=1}^{n_i} X_{,ij} X_{,ij}^{'} \exp(X_{,ij}^{'}\beta)}{\sum_{j=1}^{n_i} \exp(X_{,ij}^{'}\beta)} - \frac{[\sum_{j=1}^{n_i} X_{,ij} \exp(X_{,ij}^{'}\beta)][\sum_{j=1}^{n_i} X_{,ij} \exp(X_{,ij}^{'}\beta)]^{'}}{[\sum_{j=1}^{n_i} \exp(X_{,ij}^{'}\beta)]^2} \right\}$$

After taking the inverse of above matrix and evaluated at $\beta = b$ with b being its MLE, the estimated asymptotic covariance matrix can then be obtained.

3.3 Simulation and Result

Four utility models have been proposed in fitting our 2-year horse racing data from HK Jockey Club. Parameter estimation for each model is accomplished by the maximum likelihood method as described in Section 3.2. After that, the significance of each individual coefficient is tested by the simple standard testing according to the procedure described in Section 3.1. Having finished the variable selection process, the accuracy of these models in predicting the outcomes of the horse races is investigated by an out-of-sample simulation, based on the final models. There are 20 racing months in 2 years racing seasons. In the simulation, four months are held and the model parameters are estimated from the remaining 16 months. Firstly, the model estimated from the last 16 months is used to predict the winning probabilities for the first four racing months. After that, the 5th to 8th months (second four months) are held and predicted by the model estimated from the remaining 16 months. The same procedure is proceeded for the "third four months" (9th to 12th) and so on. Therefore, there are altogether five "four months" being used to perform the simulation.

The strategy of betting is described below:
We will bet \$10 on a horse if the following two conditions are satisfied

1. Its expected return should be greater than r where r is set to be equal to 12, 13, 14 and 15.

   As mentioned in Chapter 1, the expected return on a specific horse j per \$10 bet is equal to (odds $\times \pi_j$) where π_j is its true unknown winning probability. Here this probability will be replaced by our estimated probability p_j. That is, the first condition is (odds $\times p_j$) $> r$. Having deducted the cost, \$10, by setting r equal to the above values will ensure a positive net return.

2. Having satisfied the above condition, the horse will only be betted on as

long as its odds is less than c where c is chosen to be 100, 150, 200, 250, 300, 400, 600, 800 and 1000.

To illustrate the above idea, let's consider an example:
Given that we want our expected return (odds $\times p_j$) > 12 and the constraint on odds is equal to 100. Then we will only bet on the horse $\Longleftrightarrow$ its expected return > 12 with its odds < 100.

For each constraint imposed on odds, the actual average net return is calculated and will be plotted under each value of r used.

Finally, $R^2 = \frac{\log L_0}{\log L_1}$ is calculated for each model in measuring their 'explanatory power' on the data.

Model One

Our first utility model used is:

$$V_{ij} = \beta_1 Hwinper + \beta_2 wt.carried + \beta_3 avesprat + \beta_4 rating + \beta_5 drawing + \beta 6 Jwinper.$$

The model specification is explained in the following text.

Hwinper refers to the percentage of races won by the horse in the past 2 years.

wt.carried is the weight carried by each horse in the competition.

avesprat means an average speed rating for the last four races of each horse.

rating refers to the rating given by the HK Jockey Club.[1]

drawing is the position/gate for the horse to standby during the race.

[1] refer to the table in p.9

Jwinper refers to the percentage of winning rides of the Jockey in the past 2 years.

The three components, **Hwinper**, **avesprat** and **rating** are used as they can account for the horse's self-competitiveness. Whereas **wt.carried**, **drawing** and **Jwinper** represent the outside environment that can affect the horse performance.

Having tested the significance of each individual coefficient, it shows that no variables can be excluded from the model.

Table 3.1: Hypothesis Testing for Model 1

β	$\hat{\beta}$	$\hat{ASE}(\hat{\beta})$	Z_0	p-value
β_1	1.681	0.2273	7.405	0
β_2	0.011	0.0078	5.500	0.073
β_3	0.003	0.0021	0.375	0.075
β_4	0.014	0.0076	1.750	0.028
β_5	-0.025	0.0076	-3.125	0.0005
β_6	5.350	0.5884	9.099	0

where $Z_0 = \frac{\hat{\beta}}{\hat{ASE}(\hat{\beta})}$.

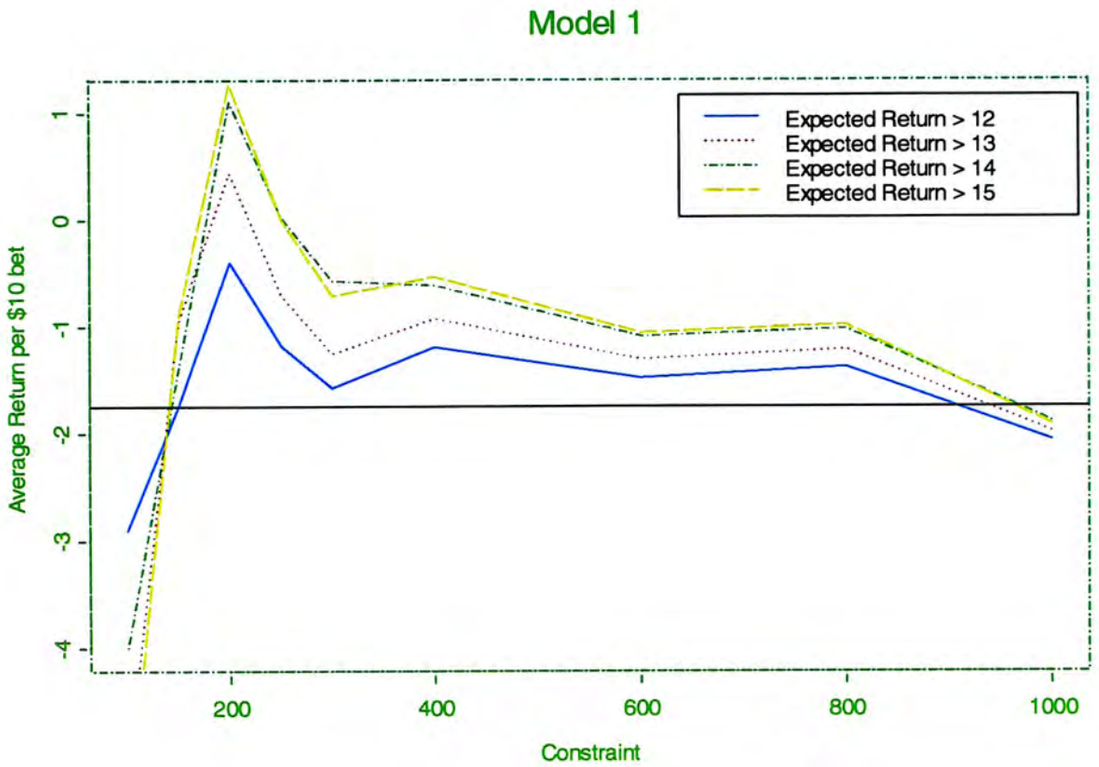
The estimated model is:

$$\begin{aligned} V_{ij} &= 1.681 Hwinper + 0.011 wt.carried + 0.003 avesprat + 0.014 rating \\ &\quad -0.025 drawing + 5.35 Jwinper. \end{aligned}$$

Notice that throughout the whole text, we will reject $H_0 \Leftrightarrow$ p-value > 0.2.

Thus, the simulation done is based on our original model. The graph below shows the actual average return under different constraints imposed in odds for

each value of r used.



Here betting on those with expected return > 15 and under the constraint odds < 200 performs the best.

Besides, the R^2 calculated in accordance with the above model gives 0.032 which means the independent variables in the model can explain 3.2% of the variability of the data.

Model Two

Our second utility model is:

$$\begin{aligned} V_{ij} &= \beta_1 Hwinper + \beta_2 wt.carried + \beta_3 avesprat + \beta_4 rating + \beta_5 drawing \\ &\quad + \beta 6 Jwinper + \beta_7 Tipster. \end{aligned}$$

There may be some 'inside information' which is not readily included in a statistical model and since the tipsters maybe better informed than us, their recommendation might also be considered. Therefore, a new variable 'Tipster' is added to see if it can really help to improve our model. This new variable is an indicator variable that '1' is marked on the first horse recommended by the 'Tipster 1'[2] and '0' otherwise.

After the process of variable selection, the predictor **wt.carried** is deleted as it shows little contribution to our second model.

Table 3.2: Hypothesis Testing for Model 2

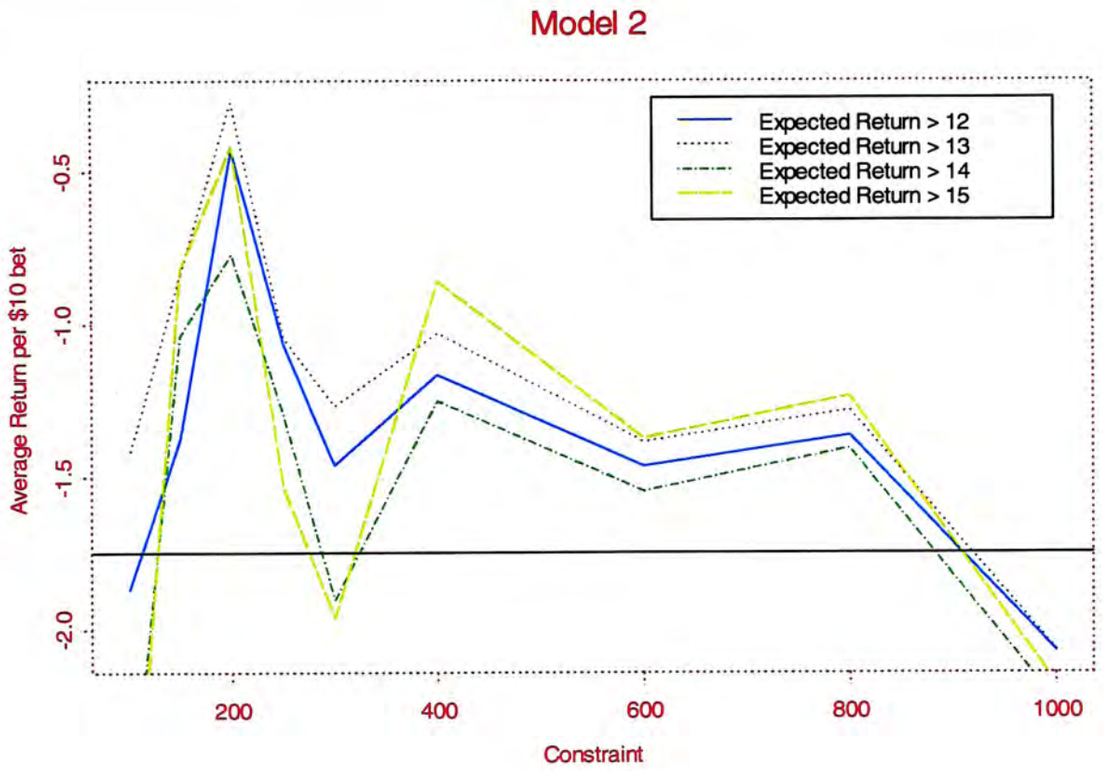
β	$\hat{\beta}$	$\hat{ASE}(\hat{\beta})$	Z_0	p-value
β_1	1.537	0.229	6.712	0
β_2	0.004	0.008	0.500	0.290
β_3	0.002	0.002	1.000	0.139
β_4	0.014	0.008	1.750	0.033
β_5	-0.023	0.008	-2.875	0.001
β_6	4.834	0.596	8.111	0
β_7	0.190	0.019	10.000	0

Then, after re-estimated the parameters under the final model, our model 2 becomes:

$$\begin{aligned} V_{ij} &= 1.539Hwinper + 0.002avesprat + 0.017rating - 0.023drawing \\ &\quad +4.915Jwinper + 0.191Tipster. \end{aligned}$$

The result of simulation based on the above final model has shown in the next page.

[2]refer to Table 2.1



As compare with our first model, though it does not improve our average return, there has increment in the value of R^2 which gives 0.046.

It performs the best under the expected return > 13 and the constraint odds < 200.

Model Three

Our thirdly proposed utility model is:

$$\begin{aligned} V_{ij} &= \beta_1 Hwinper + \beta_2 wt.carried + \beta_3 avesprat + \beta_4 rating + \beta_5 drawing \\ &\quad + \beta_6 Jwinper + \beta_7 Tipster + \beta_8 Tipster4. \end{aligned}$$

where the new variable 'Tipster4' is added. It serves the same function as 'Tipster' with the only difference that '1' is marked on all the four horses recommended by the 'Tipster 1' instead of the first one only and '0' otherwise.

The final model after checking the significance of each variable becomes:

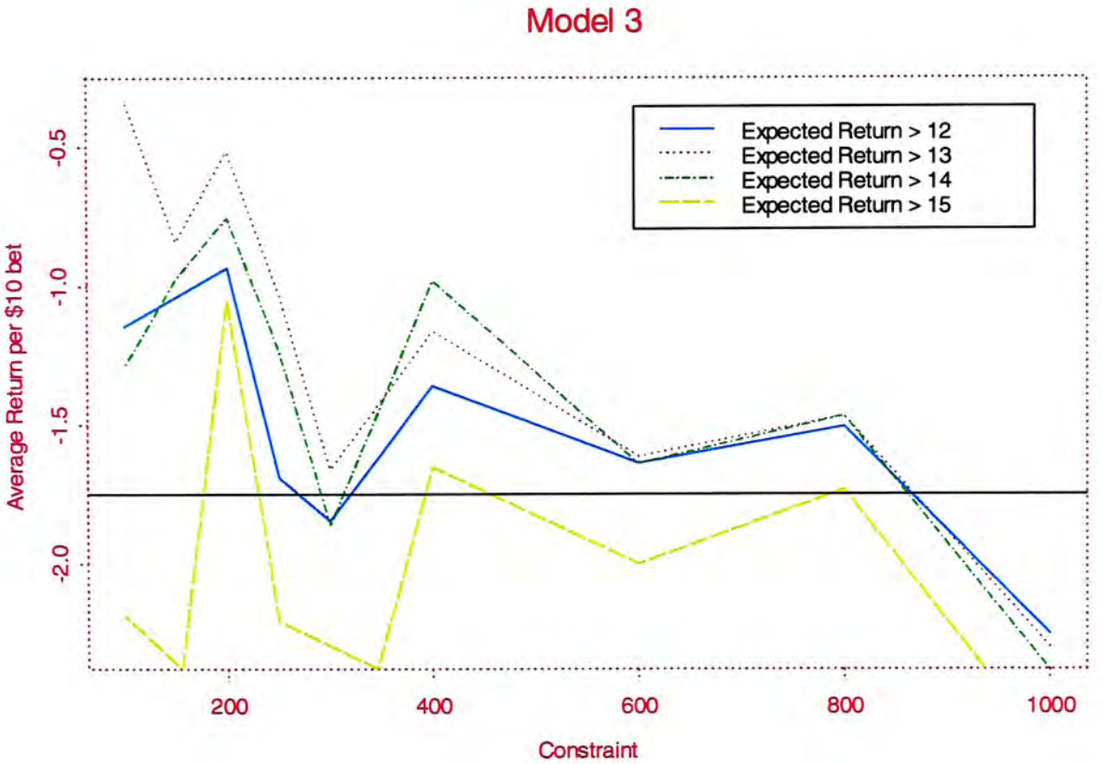
$$V_{ij} = 1.179Hwinper + 0.007rating - 0.021drawing + 4.161Jwinper$$
$$-1.66Tipster + 1.301Tipster4.$$

Now **wt.carried** and **avesprat** have been deleted as they show little significance in our model.

Table 3.3: Hypothesis Testing for Model 3

β	$\hat{\beta}$	$\hat{ASE}(\hat{\beta})$	Z_0	p-value	β	$\hat{\beta}$	$\hat{ASE}(\hat{\beta})$	Z_0	p-value
β_1	1.156	0.232	4.983	0	β_5	-0.021	0.008	-2.625	0.003
β_2	-0.004	0.008	-0.500	0.305	β_6	4.241	0.610	6.952	0
β_3	0.002	0.002	1.000	0.222	β_7	-0.166	0.035	-4.743	0
β_4	0.010	0.008	1.250	0.093	β_8	1.304	0.101	12.911	0

The graph of plotting average return against constraint on odds under different expected return is shown below.



Again, although it shows no improvement in our average return, the value of R^2 has been increased to 0.07 which shows 51.3% increment in its explanatory power. It can perform the best under the same condition as our previous model i.e. with the expected return > 13 under the constraint odds < 200.

Model Four

Our last model proposed is:

$$\begin{aligned} V_{ij} &= \beta_1 Hwinper + \beta_2 wt.carried + \beta_3 avesprat + \beta_4 rating + \beta_5 drawing \\ &\quad + \beta_6 Jwinper + \beta_7 Tipster + \beta_8 Tipster4 + \beta_9 Iage + \beta_{10} log.odds \\ &\quad + \beta_{11} wt.dist + \beta_{12} age.dist. \end{aligned}$$

In this case, four more predictors are added which includes **Iage**, **log.odds**, **wt.dist**, and **age.dist**. Each of these components is discussed below.

Iage which is an indicator variable with the value equal to 1 if the age of the competing horse is less than or equal to 6 and 0 otherwise.

log.odds which is taking the logarithm of the final odds of the horse.

wt.dist refers to the product of **wt.carried** and the distance run.

age.dist is the product of the actual age of the horse and the distance run.

Having performed the hypothesis testing on the necessity of each individual coefficient, the predictor **avesprat**, **drawing** and **age.dist** are deleted.

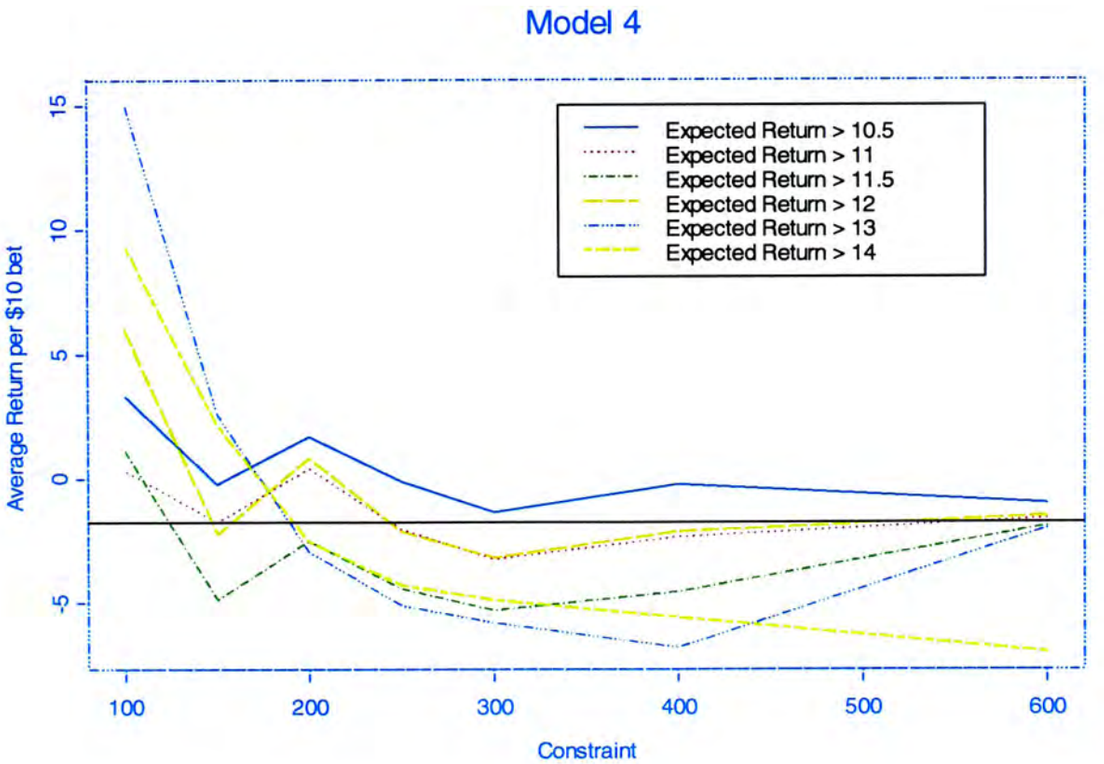
Then the final model turns out to be:

$$\begin{aligned} V_{ij} &= 0.214 Hwinper - .0.67 wt.carried + 0.012 rating + 1.168 Jwinper \\ &\quad - 0.043 Tipster + 0.191 Tipster4 + 0.139 Iage - 2.167 log.odds \\ &\quad + 0.028 wt.dist. \end{aligned}$$

Table 3.3: Hypothesis Testing for Model 4

β	$\hat{\beta}$	$\hat{ASE}(\hat{\beta})$	Z_0	p-value	β	$\hat{\beta}$	$\hat{ASE}(\hat{\beta})$	Z_0	p-value
β_1	0.230	0.238	0.966	0.167	β_7	-0.046	0.035	-1.314	0.096
β_2	-0.067	0.026	-2.577	0.004	β_8	0.205	0.114	1.798	0.037
β_3	0.001	0.002	0.500	0.303	β_9	0.129	0.141	0.915	0.181
β_4	0.012	0.009	1.333	0.081	β_{10}	-2.157	0.118	-18.280	0
β_5	0.004	0.008	0.500	0.319	β_{11}	0.028	0.016	1.750	0.044
β_6	1.177	0.679	1.733	0.042	β_{12}	0.002	0.022	0.091	0.465

The simulation using the above model gives the result below.

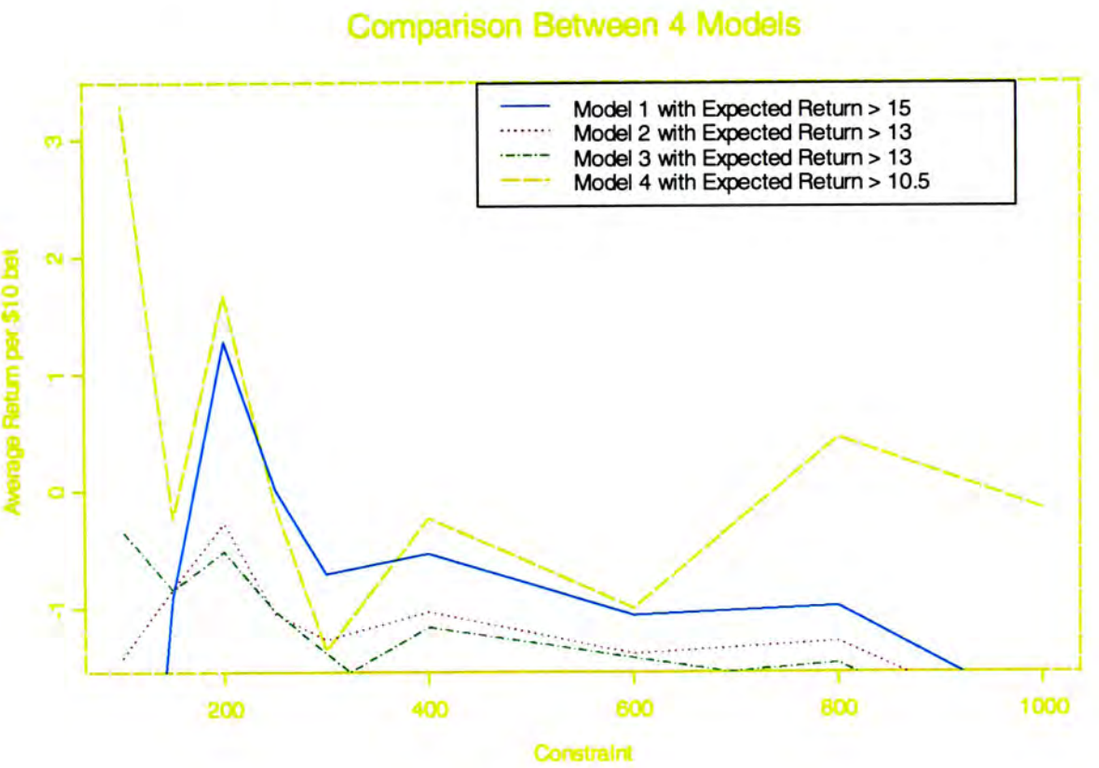


Here shows the greatest augmentation in both average return and explanatory power in this model. The R^2 gives the value of 0.133 which is the largest among all. Furthermore, The black solid line in the graph represents the average loss under the random betting, i.e. -1.75. In this model, having set our expected return $>$ 10.5, no matter what constraint has been posed on the odds, it return

always shows greater than -1.75 which means a tremendous improvement has been made by the model 4 over the random betting. At some particular points, like under the constraint odds < 200, it may even give a positive return over our betting.

3.4 Comparison between four Models

The lines with expected return > 15 from the model 1, with expected return > 13 from the model 2, with expected return > 13 from the model 3 and with expected return > 10.5 from the model 4 are selected for making the comparison. The resulting graph becomes:



It shows that the model 4 gives the best performance (with a positive return) under the constraint 200 among the four models. Besides, when the variables 'Tipster' and 'Tipster4' are added in the model 2 and 3 respectively, they can not improve the model's performance but make it even worse in terms of the average return per \$10 bet.

3.5 Concluding Remarks

As it can be seen that the more the predictors are added, the better the model would perform. It shows a great potential in developing a wagering strategy with positive return as more predictors are included. More variables such as increased/decreased in body weight of the horse, increased/decreased in rating, stakes won and so on can also be considered for further investigation of a profitable betting system.

Appendix I

Table I.1: Performance in different aspects for Tipster 1 in 98/99

	Win	Lost	Win %		Win	Lost	Win %
Total	113	484	0.19	Odds $<$ 100	107	433	0.20
				Odds $\geq$ 100	6	51	0.11
Shatin	73	313	0.19				
Happy Valley	40	171	0.19	others	14	37	0.27
				Class 1	20	70	0.22
Turf	96	436	0.18	Class 2	6	46	0.12
AWT	17	48	0.26	Class 3	33	101	0.25
				Class 4	21	147	0.125
Distance $<$ 1600m	69	241	0.22	Class 5	14	66	0.175
Distance $\geq$ 1600m	44	243	0.15	Class 6	5	17	0.23

Table I.2: Performance in different aspects for Tipster 1 in 99/00

	Win	Lost	Win %		Win	Lost	Win %
Total	123	532	0.19	Odds $<$ 100	119	452	0.21
				Odds $\geq$ 100	4	80	0.05
Shatin	83	356	0.19				
Happy Valley	40	176	0.19	others	10	32	0.24
				Class 1	17	58	0.23
Turf	116	468	0.20	Class 2	9	49	0.16
AWT	7	64	0.10	Class 3	21	124	0.14
				Class 4	44	160	0.22
Distance $<$ 1600m	58	283	0.17	Class 5	20	91	0.18
Distance $\geq$ 1600m	65	249	0.21	Class 6	2	18	0.1

Table I.3: Performance in different aspects for Tipster 2 in 98/99

	Win	Lost	Win %		Win	Lost	Win %
Total	93	499	0.16	Odds < 100	85	411	0.17
				Odds ≥ 100	8	88	0.08
Shatin	66	315	0.17				
Happy Valley	27	184	0.13	others	14	37	0.27
				Class 1	19	70	0.21
Turf	79	449	0.15	Class 2	7	44	0.14
AWT	14	50	0.22	Class 3	26	107	0.20
				Class 4	14	153	0.08
Distance $<$ 1600m	53	254	0.17	Class 5	9	70	0.11
Distance $\geq$ 1600m	40	245	0.14	Class 6	4	18	0.18

Table I.4: Performance in different aspects for Tipster 2 in 99/00

	Win	Lost	Win %		Win	Lost	Win %
Total	118	535	0.18	Odds < 100	109	418	0.21
				Odds ≥ 100	9	117	0.07
Shatin	85	352	0.19				
Happy Valley	33	183	0.15	others	7	34	0.17
				Class 1	18	57	0.24
Turf	106	477	0.18	Class 2	14	45	0.24
AWT	12	58	0.17	Class 3	21	124	0.14
				Class 4	39	164	0.19
Distance $<$ 1600m	62	279	0.18	Class 5	14	96	0.13
Distance $\geq$ 1600m	56	256	0.18	Class 6	5	15	0.25

Table I.5: Performance in different aspects for Tipster 3 in 98/99

	Win	Lost	Win %		Win	Lost	Win %
Total	93	504	0.16	Odds < 100	91	452	0.17
				Odds ≥ 100	2	52	0.04
Shatin	56	330	0.15				
Happy Valley	37	174	0.18	others	10	41	0.20
				Class 1	13	77	0.14
Turf	78	454	0.15	Class 2	8	44	0.15
AWT	15	50	0.23	Class 3	22	112	0.16
				Class 4	28	140	0.17
Distance $<$ 1600m	53	256	0.17	Class 5	9	71	0.11
Distance $\geq$ 1600m	40	248	0.14	Class 6	3	19	0.14

Table I.6: Performance in different aspects for Tipster 3 in 99/00

	Win	Lost	Win %		Win	Lost	Win %
Total	119	531	0.18	Odds < 100	117	461	0.20
				Odds ≥ 100	2	70	0.03
Shatin	86	352	0.20				
Happy Valley	33	179	0.16	others	10	31	0.24
				Class 1	20	54	0.27
Turf	116	463	0.20	Class 2	8	50	0.14
AWT	3	68	0.04	Class 3	24	120	0.17
				Class 4	43	159	0.21
Distance $<$ 1600m	67	273	0.20	Class 5	12	99	0.11
Distance $\geq$ 1600m	52	258	0.17	Class 6	2	18	0.10

Table I.7: Performance in different aspects for Tipster 4 in 98/99

	Win	Lost	Win %		Win	Lost	Win %
Total	108	489	0.18	Odds < 100	102	438	0.19
				Odds ≥ 100	6	51	0.11
Shatin	78	307	0.20				
Happy Valley	30	182	0.14	others	15	36	0.29
				Class 1	17	73	0.19
Turf	93	439	0.17	Class 2	9	43	0.17
AWT	15	50	0.23	Class 3	23	111	0.17
				Class 4	27	141	0.16
Distance $<$ 1600m	62	249	0.20	Class 5	11	69	0.14
Distance $\geq$ 1600m	46	240	0.16	Class 6	6	16	0.27

Table I.8: Performance in different aspects for Tipster 4 in 99/00

	Win	Lost	Win %		Win	Lost	Win %
Total	112	543	0.17	Odds < 100	110	477	0.19
				Odds ≥ 100	2	66	0.03
Shatin	83	358	0.19				
Happy Valley	29	185	0.14	others	8	34	0.19
				Class 1	14	60	0.19
Turf	103	481	0.18	Class 2	7	52	0.12
AWT	9	62	0.13	Class 3	26	119	0.18
				Class 4	36	169	0.18
Distance $<$ 1600m	66	276	0.19	Class 5	18	92	0.16
Distance $\geq$ 1600m	46	267	0.15	Class 6	3	17	0.15

Table I.9: Performance in different aspects for Tipster 5 in 98/99

	Win	Lost	Win %		Win	Lost	Win %
Total	85	511	0.14	Odds < 100	82	450	0.15
				Odds ≥ 100	3	61	0.05
Shatin	61	324	0.16				
Happy Valley	24	187	0.11	others	15	36	0.29
				Class 1	13	76	0.15
Turf	76	455	0.14	Class 2	7	45	0.13
AWT	9	56	0.14	Class 3	21	114	0.16
				Class 4	17	150	0.10
Distance $<$ 1600m	50	259	0.16	Class 5	10	70	0.13
Distance $\geq$ 1600m	35	252	0.12	Class 6	2	20	0.09

Table I.10: Performance in different aspects for Tipster 5 in 99/00

	Win	Lost	Win %		Win	Lost	Win %
Total	121	535	0.18	Odds < 100	121	466	0.21
				Odds ≥ 100	0	69	0.00
Shatin	83	356	0.19				
Happy Valley	38	179	0.18	others	11	30	0.27
				Class 1	10	65	0.13
Turf	114	471	0.19	Class 2	9	50	0.15
AWT	7	64	0.10	Class 3	30	115	0.21
				Class 4	40	165	0.20
Distance $<$ 1600m	66	275	0.19	Class 5	16	95	0.14
Distance $\geq$ 1600m	55	260	0.17	Class 6	5	15	0.25

Table I.11: Performance in different aspects for Favourite in 98/99

	Win	Lost	Win %		Win	Lost	Win %
Total	137	479	0.22	Odds < 100	137	479	0.22
				Odds ≥ 100	0	0	0.00
Shatin	102	295	0.26				
Happy Valley	35	184	0.16	others	18	33	0.35
				Class 1	26	68	0.28
Turf	119	432	0.22	Class 2	12	41	0.23
AWT	18	47	0.28	Class 3	23	115	0.17
				Class 4	32	143	0.18
Distance $<$ 1600m	85	233	0.27	Class 5	23	60	0.28
Distance $\geq$ 1600m	52	246	0.17	Class 6	3	19	0.14

Table I.12: Performance in different aspects for Favourite in 99/00

	Win	Lost	Win %		Win	Lost	Win %
Total	167	483	0.26	Odds < 100	167	483	0.26
				Odds ≥ 100	0	0	0.00
Shatin	122	318	0.28				
Happy Valley	45	165	0.21	others	12	30	0.29
				Class 1	18	56	0.24
Turf	152	427	0.26	Class 2	14	45	0.24
AWT	15	56	0.21	Class 3	39	103	0.27
				Class 4	55	148	0.27
Distance $<$ 1600m	90	251	0.26	Class 5	25	85	0.23
Distance $\geq$ 1600m	77	232	0.25	Class 6	4	16	0.20

Table I.13: Performance in different aspects for Joint in 98/99

	Win	Lost	Win %		Win	Lost	Win %
Total	131	576	0.19	Odds < 100	127	558	0.19
				Odds ≥ 100	4	18	0.18
Shatin	91	370	0.20				
Happy Valley	40	206	0.16	others	19	45	0.30
				Class 1	21	81	0.21
Turf	112	515	0.18	Class 2	11	55	0.17
AWT	19	61	0.24	Class 3	32	119	0.21
				Class 4	25	172	0.13
Distance $<$ 1600m	76	296	0.20	Class 5	17	81	0.17
Distance $\geq$ 1600m	55	280	0.16	Class 6	6	23	0.21

Table I.14: Performance in different aspects for Joint in 99/00

	Win	Lost	Win %		Win	Lost	Win %
Total	165	618	0.21	Odds < 100	163	586	0.22
				Odds ≥ 100	2	32	0.06
Shatin	118	424	0.22				
Happy Valley	47	194	0.20	others	13	45	0.22
				Class 1	21	65	0.24
Turf	156	538	0.22	Class 2	14	64	0.18
AWT	9	80	0.10	Class 3	33	136	0.20
				Class 4	54	190	0.22
Distance $<$ 1600m	85	327	0.21	Class 5	27	98	0.22
Distance $\geq$ 1600m	80	291	0.22	Class 6	3	20	0.13

References

Agresti, A. (1990) *Categorical data analysis.* New York: John Wiley

Benter, W. F. (1994) *Computer based horse race handicapping and wagering systems.* San Diego CA: Academic Press.

Bolton, R. N. & Chapman, R. G. (1986) Searching for positive returns at the track: A multinomial logit model for handicapping horse races, *Management Science*, 32(8), pp. 1040-1060.

Fox, J. (1997) *Applied regression analysis, linear models, and related models.* Thousand Oaks, Calif: Sage Publications.

Mukhtar, M. A. (1998) Probability models on horse-race outcomes, *Journal of Applied Statistics*, 25(2), pp. 221-229.

Venables, W. N. & Ripley, B. D. (1994) *Modern applied statistics with S-Plus.* New York: Springer-Verlag.

Windle, D. (1993) Multiple criteria decision making applied to horse racing, *Teaching Statistics*, 15(2), p. 34-37.