Computer Based Horse Race Handicapping and Wagering Systems: A Report

William Benter
HK Betting Syndicate, Hong Kong

ABSTRACT

This paper examines the elements necessary for a practical and successful computerized horse race handicapping and wagering system. Data requirements, handicapping model development, wagering strategy, and feasibility are addressed. A logit-based technique and a corresponding heuristic measure of improvement are described for combining a fundamental handicapping model with the public's implied probability estimates. The author reports significant positive results in five years of actual implementation of such a system. This result can be interpreted as evidence of inefficiency in pari-mutuel racetrack wagering. This paper aims to emphasize those aspects of computer handicapping which the author has found most important in practical application of such a system.

INTRODUCTION

The question of whether a fully mechanical system can ever "beat the races" has been widely discussed in both the academic and popular literature. Certain authors have convincingly demonstrated that profitable wagering systems do exist for the races. The most well documented of these have generally been of the *technical* variety, that is, they are concerned mainly with the public odds, and do not attempt to predict horse performance from fundamental factors. Technical systems for place and show betting, (Ziemba and Hausch, 1987) and exotic pool betting, (Ziemba and Hausch, 1986) as well as the 'odds movement' system developed by Asch and Quandt (1986), fall into this category. A benefit of these systems is that they require relatively little preparatory effort, and can be effectively employed by the occasional racegoer. Their downside is that betting opportunities tend to occur infrequently and the maximum expected profit achievable is usually relatively modest. It is debatable whether any racetracks exist where these systems could be profitable enough to sustain a full-time professional effort.

To be truly viable, a system must provide a large number of high advantage betting opportunities in order that a sufficient amount of expected profit can be generated. An approach which does promise to provide a large number of betting opportunities is to *fundamentally* handicap each race, that is, to empirically assess each horse's chance of winning, and utilize that assessment to find profitable wagering opportunities. A natural way to attempt to do this is to develop a computer model to estimate each horse's probability of winning and calculate the appropriate amount to wager.

A complete survey of this subject is beyond the scope of this paper. The general requirements for a computer based fundamental handicapping model have been well presented by Bolton and Chapman (1986) and Brecher (1980). These two references are "required reading" for anyone interested in developing such a system. Much of what is said here has already been explained in those two works, as is much of the theoretical background which has been omitted here. What the author would hope to add, is a discussion of a few points which have not been addressed in the literature, some practical recommendations, and a report that a *fundamental* approach can in fact work in practice.

FEATURES OF THE COMPUTER HANDICAPPING APPROACH

Several features of the computer approach give it advantages over traditional handicapping. First, because of its empirical nature, one need not possess specific handicapping expertise to undertake this enterprise, as everything one needs to know can be learned from the data. Second is the testability of a computer system. By carefully partitioning data, one can develop a model and test it on *unseen* races. With this procedure one avoids the danger of overfitting past data. Using this 'holdout' technique, one can obtain a reasonable estimate of the system's real-time performance before wagering any actual

money. A third positive attribute of a computerized handicapping system is its consistency. Handicapping races manually is an extremely taxing undertaking. A computer will effortlessly handicap races with the same level of care day after day, regardless of the mental state of the operator. This is a non-trivial advantage considering that a professional level betting operation may want to bet several races a day for extended periods.

The downside of the computer approach is the level of preparatory effort necessary to develop a winning system. Large amounts of past data must be collected, verified and computerized. In the past, this has meant considerable manual entry of printed data. This situation may be changing as optical scanners can speed data entry, and as more online horseracing database services become available. Additionally, several man-years of programming and data analysis will probably be necessary to develop a sufficiently profitable system. Given these considerations, it is clear that the computer approach is not suitable for the casual racegoer.

HANDICAPPING MODEL DEVELOPMENT

The most difficult and time-consuming step in creating a computer based betting system is the development of the fundamental handicapping model. That is, the model whose final output is an estimate of each horse's probability of winning. The type of model used by the author is the multinomial logit model proposed by Bolton and Chapman (1986). This model is well suited to horse racing and has the convenient property that its output is a set of probability estimates which sum to 1 within each race.

The overall goal is to estimate each horse's current performance potential. "Current performance potential" being a single overall summary index of a horse's expected performance in a particular race. To construct a model to estimate current performance potential one must investigate the available data to find those variables or *factors* which have predictive significance. The profitability of the resulting betting system will be largely determined by the predictive power of the factors chosen The odds set by the public betting yield a sophisticated estimate of the horses' win probabilities. In order for a fundamental statistical model to be able to compete effectively, it must rival the public in sophistication and comprehensiveness. Various types of factors can be classified into groups:

Current condition:
- performance in recent races
- time since last race
- recent workout data
- age of horse

Past performance:
- finishing position in past races
- lengths behind winner in past races
- normalized times of past races

Adjustments to past performance:
- strength of competition in past races
- weight carried in past races
- jockey's contribution to past performances
- compensation for bad luck in past races
- compensation for advantageous or disadvantageous post position in past races

Present race situational factors:
- weight to be carried
- today's jockey's ability
- advantages or disadvantages of the assigned post position

Preferences which could influence the horse's performance in today's race:
- distance preference
- surface preference (turf vs dirt)
- condition of surface preference (wet vs dry)
- specific track preference

More detailed discussions of fundamental handicapping can be found in the extensive popular literature on the subject (for the author's suggested references see the list in the appendix). The data needed to calculate these factors must be entered and checked for accuracy. This can involve considerable effort. Often, multiple sources must be used to assemble complete past performance records for each of the horses. This is especially the case when the horses have run past races at many different tracks. The easiest type of racing jurisdiction to collect data and develop a model for is one with a *closed* population of horses, that is, one where horses from a single population race only against each other at a limited number of tracks. When horses have raced at venues not covered in the database, it is difficult to evaluate the elapsed times of races and to estimate the strength of their opponents. Also unknown will be the post position biases, and the relative abilities of the jockeys in those races.

In the author's experience the minimum amount of data needed for adequate model development and testing samples is in the range of 500 to 1000 races. More is helpful, but out-of-sample predictive accuracy does not seem to improve dramatically with development samples greater than 1000 races. Remember that *data for one race* means full past data on all of the runners in that race. This suggests another advantage of a *closed* racing population; by collecting the results of all races run in that jurisdiction one automatically accumulates the full set of past performance data for each horse in the population.

It is important to define factors which extract as much information as possible out of the data in each of the relevant areas. As an example, consider three different specifications of a 'distance preference' factor.

The first is from Bolton and Chapman (1986):

> 'NEWDIST' - this variable equals one if a horse has run three of its four previous races at a distance less than a mile, zero otherwise. (Note: Bolton and Chapman's model was only used to predict races of 1 - 1.25 miles.)

The second is from Brecher (1980):

> 'DOK' - this variable equals one if the horse finished in the upper 50th percentile or within 6.25 lengths of the winner in a prior race within 1/16 of a mile of today's distance, or zero otherwise

The last is from the author's current model:

> 'DP6A' - for each of a horse's past races, a predicted finishing position is calculated via multiple regression based on all factors except those relating to distance. This predicted finishing position in each race is then subtracted from the horse's actual finishing position. The resulting quantity can be considered to be the unexplained residual which may be due to some unknown distance preference that the horse may possess plus a certain amount of random error. To estimate the horse's preference or aversion to today's distance, the residual in each of its past races is used to estimate a linear relationship between performance and similarity to today's distance. Given the statistical uncertainty of estimating this relationship from the usually small sample of past races, the final magnitude of the estimate is standardized by dividing it by its standard error. The result is that horses with a clearly defined distance preference demonstrated over a large number of races will be awarded a relatively larger magnitude value than in cases where the evidence is less clear.

The last factor is the result of a large number of progressive refinements. The subroutines involved in calculating it run to several thousand lines of code. The author's guiding principle in factor improvement has been a combination of educated guessing and trial and error. Fortunately, the historical data makes the final decision as to which particular definition is superior. The best is the one that produces the greatest increase in predictive accuracy when included in the model. The general thrust of model development is to continually experiment with refinements of the various factors. Although time-consuming, the gains are worthwhile. In the author's experience, a model involving only simplistic specifications of factors does not provide sufficiently accurate estimates of winning probabilities. Care must be taken in this process of model development not to overfit past

data. Some overfitting will always occur, and for this reason it is important to use data partitioning to maintain sets of *unseen* races for out-of-sample testing.

The complexity of predicting horse performance makes the specification of an elegant handicapping model quite difficult. Ideally, each independent variable would capture a unique aspect of the influences effecting horse performance. In the author's experience, the trial and error method of adding independent variables to increase the model's goodness-of-fit, results in the model tending to become a hodgepodge of highly correlated variables whose individual significances are difficult to determine and often counter-intuitive. Although aesthetically unpleasing, this tendency is of little consequence for the purpose which the model will be used, namely, prediction of future race outcomes. What it does suggest, is that careful and conservative statistical tests and methods should be used on as large a data sample as possible.

For example, "number of past races" is one of the more significant factors in the author's handicapping model, and contributes greatly to the overall accuracy of the predictions. The author knows of no 'common sense' reason why this factor should be important. The only reason it can be confidently included in the model is because the large data sample allows its significance to be established beyond a reasonable doubt.

Additionally, there will always be a significant amount of 'inside information' in horse racing that cannot be readily included in a statistical model. Trainer's and jockey's intentions, secret workouts, whether the horse ate its breakfast, and the like, will be available to certain parties who will no doubt take advantage of it. Their betting will be reflected in the odds. This presents an obstacle to the model developer with access to published information only. For a statistical model to compete in this environment, it must make full use of the advantages of computer modelling, namely, the ability to make complex calculations on large data sets.

CREATING UNBIASED PROBABILITY ESTIMATES

It can be presumed that valid fundamental information exists which can not be systematically or practically incorporated into a statistical model. Therefore, any statistical model, however well developed, will always be incomplete. An extremely important step in model development, and one that the author believes has been generally overlooked in the literature, is the estimation of the relation of the model's probability estimates to the public's estimates, and the adjustment of the model's estimates to incorporate whatever information can be gleaned from the public's estimates.

The public's implied probability estimates generally correspond well with the actual frequencies of winning. This can be shown with a table of estimated probability versus actual frequency of winning (Table 1).

Table 1

PUBLIC ESTIMATE VS. ACTUAL FREQUENCY

range	n	exp.	act.	Z
.000-.010	1343	.007	.007	0.0
.010-.025	4356	.017	.020	1.3
.025-.050	6193	.037	.042	2.1
.050-.100	8720	.073	.069	-1.5
.100-.150	5395	.123	.125	0.6
.150-.200	3016	.172	.173	0.1
.200-.250	1811	.222	.219	-0.3
.250-.300	1015	.273	.253	-1.4
.300-.400	716	.339	.339	0.0
>.400	312	.467	.484	0.6

races = 3198, # horses = 32877

Table 2

FUNDAMENTAL MODEL VS. ACTUAL FREQUENCY

range	n	exp.	act.	Z
.000-.010	1173	.006	.005	-0.6
.010-.025	3641	.018	.015	-1.2
.025-.050	6503	.037	.037	-0.3
.050-.100	9642	.073	.074	0.1
.100-.150	5405	.123	.120	-0.7
.150-.200	2979	.173	.183	1.6
.200-.250	1599	.223	.232	0.9
.250-.300	870	.272	.285	0.9
.300-.400	741	.341	.320	-1.2
>.400	324	.475	.432	-1.6

races = 3198, # horses = 32877

range = the range of estimated probabilities
n = the number of horses falling within a range
exp. = the mean expected probability
act. = the actual win frequency observed
Z = the discrepancy (+ or -) in units of standard errors

In each range of estimated probabilities, the actual frequencies correspond closely. This is not the case at all tracks (Ali, 1977) and if not, suitable corrections should be made when using the public's

probability estimates for the purposes which will be discussed later. (Unless otherwise noted, data samples consist of all races run by the Royal Hong Kong Jockey Club from September 1986 through June 1993.)

A multinomial logit model using fundamental factors will also naturally produce an internally consistent set of probability estimates (Table 2). Here again there is generally good correspondence between estimated and actual frequencies. Table 2 however conceals a major, (and from a wagering point of view, disastrous) type of bias inherent in the fundamental model's probabilities. Consider the following two tables which represent roughly equal halves of the sample in Table 2. Table 3 shows the fundamental model's estimate versus actual frequency for those horses where the public's probability estimate was greater the fundamental model's. Table 4 is the same except that it is for those horses whose public estimate was less than the fundamental model's.

Table 3

FUNDAMENTAL MODEL VS. ACTUAL FREQUENCY WHEN PUBLIC ESTIMATE IS GREATER THAN MODEL ESTIMATE

range	n	exp.	act.	Z
.000-.010	920	.006	.005	-0.3
.010-.025	2130	.017	.018	0.3
.025-.050	3454	.037	.044	2.1
.050-.100	4626	.073	.091	4.7
.100-.150	2413	.122	.147	3.7
.150-.200	1187	.172	.227	5.0
.200-.250	540	.223	.302	4.4
.250-.300	252	.270	.333	2.3
.300-.400	165	.342	.448	2.9
>.400	54	.453	.519	1.0

races = 3198, # horses = 15741

Table 4

FUNDAMENTAL MODEL VS. ACTUAL FREQUENCY WHEN PUBLIC ESTIMATE IS LESS THAN MODEL ESTIMATE

range	n	exp.	act.	Z
.000-.010	253	.007	.004	-0.6
.010-.025	1511	.018	.011	-2.2
.025-.050	3049	.037	.029	-2.6
.050-.100	5016	.074	.058	-4.3
.100-.150	2992	.123	.098	-4.2
.150-.200	1792	.173	.154	-2.1
.200-.250	1059	.223	.196	-2.1
.250-.300	618	.273	.265	-0.4
.300-.400	576	.341	.283	-2.9
>.400	270	.480	.415	-2.1

races = 3198, # horses = 17136

There is an extreme and consistent bias in both tables. In virtually every range the actual frequency is significantly different than the fundamental model's estimate, and always in the direction of being closer to the public's estimate. The fundamental model's estimate of the probability cannot be considered to be an unbiased estimate independent of the public's estimate. Table 4 is particularly important because it is comprised of those horses that the model would have one bet on, that is, horses whose model-estimated probability is greater than their public probability. It is necessary to correct for this bias in order to accurately estimate the advantage of any particular bet.[1]

In a sense, what is needed is a way to combine the judgements of two experts, (i.e. the fundamental model and the public). One practical technique for accomplishing this is as follows: (Asch and Quandt, 1986; pp. 123-125). See also White, Dattero and Flores, (1992).

Estimate a second logit model using the two probability estimates as independent variables. For a race with entrants $(1,2,\ldots,N)$ the win probability of horse i is given by:

$$c_i = \frac{\exp(\alpha f_i + \beta \pi_i)}{\sum \exp(\alpha f_j + \beta \pi_j)} \qquad (\text{for } j = 1 \text{ to } N) \tag{1}$$

f_i = log of 'out-of-sample' fundamental model probability estimate

π_i = log of public's implied probability estimate

c_i = combined probability estimate

(Natural log of probability is used rather than probability as this transformation provides a better fit)

Given a set of past races $(1,2,\ldots,R)$ for which both public probability estimates and fundamental model estimates are available, the parameters α and β can be estimated by maximizing the log likelihood function of the given set of races with respect to α and β:

$$\exp(L) = \prod c_{ji^*} \quad (j = 1 \text{ to } R) \tag{2}$$

where c_{ji^*} denotes the probability as given by equation (1) for the horse i^* observed to win race j (Bolton and Chapman, 1986 p. 1044). Equation (1) should be evaluated using fundamental probability estimates from a model developed on a separate sample of races. Use of 'out-of-sample' estimates prevents overestimation of the fundamental model's significance due to 'custom-fitting' of the model development sample. The estimated values of α and β can be interpreted roughly as the relative correctness of the model's and the public's estimates. The greater the value of α, the better the model. The probabilities that result from this model also show good correspondence between predicted and actual frequencies of winning (Table 5).

Table 5

COMBINED MODEL VS. ACTUAL FREQUENCIES

range	n	exp.	act.	Z
.000-.010	1520	.007	.005	-1.0
.010-.025	4309	.017	.018	0.1
.025-.050	6362	.037	.038	0.6
.050-.100	8732	.073	.071	-0.5
.100-.150	5119	.123	.119	-0.8
.150-.200	2974	.173	.180	1.0
.200-.250	1657	.223	.223	0.0
.250-.300	993	.272	.281	0.6
.300-.400	853	.340	.328	0.7
>.400	358	.479	.492	0.5

races = 3198, # horses = 32877

By comparison with Tables 1 and 2, Table 5 shows that there is more *spread* in the combined model's probabilities than in either the public's or the fundamental model's alone, that is, there are more horses in both the very high and very low probability ranges. This indicates that the combined model is more informative. More important is that the new probability estimates are without the bias shown in Tables 3 and 4, and thus are suitable for the accurate estimation of betting advantage. This is borne out by Tables 6 and 7, which are analogous to Tables 3 and 4 above except that they use the combined model probabilities instead of the raw fundamental model probabilities.

Table 6

COMBINED MODEL VS. ACTUAL FREQUENCY WHEN PUBLIC ESTIMATE IS GREATER THAN MODEL ESTIMATE

range	n	exp.	act.	Z
.000-.010	778	.006	.005	-0.4
.010-.025	1811	.017	.015	-0.6
.025-.050	2874	.037	.035	-0.7
.050-.100	4221	.073	.073	0.0
.100-.150	2620	.123	.116	-1.0
.150-.200	1548	.173	.185	1.2
.200-.250	844	.223	.231	0.6
.250-.300	493	.272	.292	1.0
.300-.400	393	.337	.349	0.5
>.400	159	.471	.509	1.0

races = 3198, # horses = 15741

Table 7

COMBINED MODEL VS. ACTUAL FREQUENCY WHEN PUBLIC ESTIMATE IS LESS THAN MODEL ESTIMATE

range	n	exp.	act.	Z
.000-.010	742	.007	.004	-0.9
.010-.025	2498	.018	.019	0.6
.025-.050	3488	.037	.041	1.4
.050-.100	4511	.072	.069	-0.7
.100-.150	2499	.123	.122	-0.1
.150-.200	1426	.173	.174	0.1
.200-.250	813	.223	.215	-0.5
.250-.300	500	.272	.270	-0.1
.300-.400	460	.342	.311	-1.4
>.400	199	.485	.477	-0.2

races = 3198, # horses = 17136

Observe that the above tables show no significant bias one way or the other.

ASSESSING THE VALUE OF A HANDICAPPING MODEL

The log likelihood function of equation (2) can be used to produce a measure of fit analogous to the R^2 of multiple linear regression (Equation 3). This pseudo-R^2 (R^2) can be used to compare models

and to assess the value of a particular model as a betting tool. Each set of probability estimates, either the public's or those of a model, achieve a certain R^2, defined as (Bolton and Chapman, 1986)

$$R^2 = 1 - \frac{L(\text{model})}{L(1/N_j)}. \tag{3}$$

The R^2 value is a measure of the "explanatory power" of the model. An R^2 of 1 indicates perfect predictive ability while an R^2 of 0 means that the model is no better than random guessing. An important benchmark is the R^2 value achieved by the public probability estimate. A heuristic measure of the potential profitability of a handicapping model, borne out in practice, is the amount by which its inclusion in the combined model of equation (1) along with the public probability estimate causes the R^2 to increase over the value achieved by the public estimate alone:

$$\Delta R^2 = R^2_C - R^2_P \tag{4}$$

where the subscript P denotes the public's probability estimate and C stands for the combined (fundamental and public) model of equation (1) above. In a sense, ΔR^2 may be taken as a measure of the amount of information added by the fundamental model. In the case of the models which produced Tables 1,2 and 5 above these values are:

$R^2_P = .1218$ (public)
$R^2_F = .1245$ (fundamental model)
$R^2_C = .1396$ (combined model)

$\Delta R^2_{C\text{-}P} = .1396 - .1218 = .0178$

Though this value may appear small, it actually indicates that significant profits could be made with that model. The ΔR^2 value is a useful measure of the potential profitability of a particular model. It can be used to measure and compare models without the the time consuming step of a full wagering simulation. In the author's experience, greater ΔR^2 values have been invariably associated with greater wagering simulation profitability. To illustrate the point that the important criteria is the gain in R^2 in the combined model over the public's R^2, and not simply the R^2 of the handicapping model alone, consider the following two models.

The first is a logit-derived fundamental handicapping model using 9 significant fundamental factors. It achieves an out-of-sample R^2 of .1016. The second is a probability estimate derived from tallying the picks of approximately 48 newspaper tipsters. (Figlewski, 1979) The tipsters each make a selection for 1st, 2nd, and 3rd in each race. The procedure was to count the number of times each horse was picked, awarding 6 points for 1st, 3 points for 2nd, and 1 point for 3rd. The point total for each horse is then divided by the total points awarded in the race (i.e. 48 * 10). This fraction of points is then taken to be the 'tipster' probability estimate. Using the log of this estimate as the sole independent variable in a logit model produces an R^2 of .1014. On the basis of their stand-alone R^2's the above two models would appear to be equivalently informative predictors of race outcome. Their vast difference appears when we perform the 'second stage' of combining these estimates with the public's. The following results were derived from logit runs on 2313 races (September 1988 to June 1993).

$R^2_P = .1237$ (public estimate)
$R^2_F = .1016$ (fundamental model)
$R^2_T = .1014$ (tipster model)
$R^2_{(F\&P)} = .1327$ (fundamental and public)
$R^2_{(T\&P)} = .1239$ (tipster and public)

$\Delta R^2_{(F\&P)\text{-}P} = .1327 - .1237 = .0090$
$\Delta R^2_{(T\&P)\text{-}P} = .1239 - .1237 = .0002$

As indicated by the ΔR^2 values, the tipster model adds very little to the public's estimate. The insignificant contribution of the tipster estimate to the overall explanatory power of the combined model effectively means that when there is a difference between the public estimate and the tipster estimate, then the public's estimate is superior. The fundamental model on the other hand, does contribute significantly when combined with the public's. For a player considering betting with the 'tipster' model, carrying out this 'second stage' would have saved that player from losing money; the output of the second stage model would always be virtually identical to the public estimate, thus never indicating an advantage bet.

WAGERING STRATEGY

After computing the combined and therefore unbiased probability estimates as described above, one can make accurate estimations of the advantage of any particular bet. A way of expressing advantage is as the expected return per dollar bet:

$$\text{expected return} = er = c * div$$
$$\text{advantage} = er - 1$$

where c is the estimated probability of winning the bet and div is the expected dividend. For win betting the situation is straightforward. The c's are the probability estimates produced by equation (1) above, and the div's are the win dividends (as a payoff for a $1 bet) displayed on the tote board. The situation for an example race is illustrated in Table 8.

Table 8

#	c	p	er	div
1)	.021	.025	.68	33
2)	.125	.088	1.17	9.3
3)	.239	.289	.69	2.8
4)	.141	.134	.87	6.1
5)	.066	.042	1.29	19
6)	.012	.013	.75	61
7)	.107	.136	.64	6.0
8)	.144	.089	1.33	9.2
9)	.019	.014	1.18	60
10)	.067	.066	.86	12
11)	.012	.012	.83	68u
12)	.028	.047	.50	17
13)	.011	.027	.32	30
14)	.009	.019	.41	43

c = combined (second stage) probability estimate
p = public's probability estimate (1-take) / div
er = expected return on a $1 win bet
div = win dividend for a $1 bet

The 'u' after the win dividend for horse #11 stands for *unratable* and indicates that this is a horse for which the fundamental model could not produce a probability estimate. Often this is because the horse is running in its first race. A good procedure for handling such horses is to assign them the same probability as that implied by the public win odds, and renormalize the probabilities on the other horses so that the total probability for the race sums to 1. This is equivalent to saying that we have no information which would allow us to dispute the public's estimate so we will take theirs.

From Table 8 we can see that the advantage win bets are those with an er greater than 1. There is a positive expected return from betting on each of these horses. Given that there are several different types of wager available, it is necessary to have a strategy for determining which bets to make and in what amounts.

Kelly Betting and pool size limitations

Given the high cost in time and effort of developing a winning handicapping system, a wagering strategy which produces maximum expected profits is desirable. The stochastic nature of horse race wagering however, guarantees that losing streaks of various durations will occur. Therefore a strategy

which balances the tradeoff between risk and returns is necessary. A solution to this problem is provided by the Kelly betting strategy (Kelly, 1956). The Kelly strategy specifies the fraction of total wealth to wager so as to maximize the exponential rate of growth of wealth, in situations where the advantage and payoff odds are known. As a fixed fraction strategy, it also never risks ruin. (This last point is not strictly true, as the mimimum bet limit prevents strict adherence to the strategy.) For a more complete discussion of the properties of the Kelly strategy see MacLean, Ziemba and Blazenko (1992), see also Epstein (1977) and Brecher (1980).

The Kelly strategy defines the optimal bet (or set of bets) as those which maximize the expected log of wealth. In pari-mutuel wagering, where multiple bets are available in each race, and each bet effects the final payoff odds, the exact solution requires maximizing a concave logarithmic function of several variables. For a single bet, assuming no effect on the payoff odds, the formula simplifies to

$$K = \frac{\text{(advantage)}}{\text{(dividend - 1)}} . \qquad (5)$$

where K is the fraction of total wealth to wager. When one is simultaneously making wagers in multiple pools, further complications to the exact multiple bet Kelly solution arise due to 'exotic' bets in which one must specify the order of finish in two or more races. The expected returns from these bets must be taken into account when calculating bets for the single race pools in those races.

In the author's experience, betting the full amount recommended by the Kelly formula is unwise for a number of reasons. Firstly, accurate estimation of the advantage of the bets is critical; if one overestimates the advantage by more than a factor of two, Kelly betting will cause a negative rate of capital growth. (As a practical matter, many factors may cause one's real-time advantage to be less than past simulations would suggest, and very few can cause it to be greater. Overestimating the advantage by a factor of two is easily done in practice.) Secondly, if it is known that regular withdrawals from the betting bankroll will be made for paying expenses or taking profits, then one's effective wealth is less than their actual current wealth. Thirdly, full Kelly betting is a 'rough ride', downswings during which more than 50% of total wealth is lost are a common occurrence. For these and other reasons, a *fractional Kelly* betting strategy is advisable, that is, a strategy wherein one bets some fraction of the recommended Kelly bet (e.g. 1/2 or 1/3). For further discussion of fractional Kelly betting, and a quantitative analysis of the risk/reward tradeoffs involved, see MacLean, Ziemba and Blazenko (1992).

Another even more important constraint on betting is the effect that one's bet has on the advantage. In pari-mutuel betting markets each bet decreases the dividend. Even if the bettor possesses infinite wealth, there is a maximum bet producing the greatest expected profit, any amount beyond which lowers the expected profit. The maximum bet can be calculated by writing the equation for expected profit as a function of bet size, and solving for the bet size which maximizes expected profit. This maximum can be surprisingly low as the following example illustrates.

c	div	*er*
06	20	1.20

total pool size = $100,000
maximum *er* bet = $416
expected profit = $39.60

A further consideration concerns the shape of the 'expected profit versus bet size' curve when the bet size is approaching the maximum. In this example, the maximum expected profit is with a bet of $416. If one made a bet of only 2/3 the maximum, i.e. $277, the expected profit would be 35.5 dollars, or 90% of the maximum. There is very little additional gain for risking a much larger sum of money. Solving the fully formulated Kelly model (i.e. taking into account the bets' effects on the dividends) will optimally balance this tradeoff. See Kallberg and Ziemba (1994) for a discussion of the optimization properties of such formulations.

As a practical matter; given the relatively small sizes of most pari-mutuel pools, a successful betting operation will soon find that all of its bets are *pool-size-limited.* As a rule of thumb, as the bettor's wealth approaches the total pool size, the dominant factor limiting bet size becomes the effect of the bet on the dividend, not the bettor's wealth.

Exotic bets

In addition to win bets, racetracks offer numerous so-called *exotic* bets. These offer some of the highest advantage wagering opportunities. This results from the multiplicative effect on overall advantage of combining more than one advantage horse. For example, suppose that in a particular race there are two horses for which the model's estimate of the win probability is greater than the public's, though not enough so as to make them positive expectation win bets.

	c	*div*	*p*	*er*
1)	.115	8.3	.100	.955
2)	.060	16.6	.050	.996

By the Harville formula (Harville 1973), the estimated probability of a 1,2 or 2,1 finish is

$$C_{12,21} = (.115 * .060)/(1 - .115) + (.060 * .115)/(1 - .060) = .0151 .$$

The public's implied probability estimate is

$$P_{12,21} = (.100 * .050)/(1 - .100) + (.050 * .100)/(1 - .050) = .0108 .$$

Therefore (assuming a 17% track take) the public's rational quinella dividend should be

$$qdiv \cong (1 - .17)/.0108 = 76.85 .$$

Assuming that the estimated probability is correct the expected return of a bet on this combination is

$$er = .0151 * 76.85 = 1.16 .$$

In the above example two horses which had expected returns of less than 1 as individual win bets, in combination produce a 16% advantage quinella bet. The same principle applies, only more so, for bets in which one must specify the finishing positions of more than two horses. In *ultra-exotic* bets such as the pick-six, even a handicapping model with only modest predictive ability can produce advantage bets. The situation may be roughly summarized by stating that for a bettor in possession of accurate probability estimates which differ from the public estimates; 'the more *exotic* (i.e. specific) the bet, the higher the advantage'. Place and show bets are not considered exotic in this sense as they are less specific than normal bets. The probability differences are 'watered down' in the place and show pools.[2] Some professional players make only exotic wagers to capitalize on this effect.

First, Second, and Third

In exotic bets that involve specifying the finishing order of two or more horses in one race, a method is needed to estimate these probabilities. A popular approach is the Harville formula. (Harville, 1973):

For three horses (i, j, k) with win probabilities (π_i, π_j, π_k) the Harville formula specifies the probability that they will finish in order as

$$\pi_{ijk} = \frac{\pi_i \pi_j \pi_k}{(1 - \pi_i)(1 - \pi_i - \pi_j)} . \qquad (6)$$

This formula is significantly biased, and should not be used for betting purposes, as it will lead to serious errors in probability estimations if not corrected for in some way.[3] (Henery 1981, Stern 1990, Lo and Bacon-Shone 1992). Its principle deficiency is the fact that it does not recognize the increasing randomness of the contests for second and third place. The bias in the Harville formula is demonstrated in Tables 9 and 10 which show the formula's estimated probabilities for horses to finish second and third given that the identity of the horses finishing first (and second) are known. The data set used is the same as that which produced Table 1 above.

Table 9

HARVILLE MODEL CONDITIONAL PROBABILITY OF 2ND

range	n	exp.	act.	Z
.000-.010	962	.007	.010	0.9
.010-.025	3449	.018	.030	5.3
.025-.050	5253	.037	.045	2.8
.050-.100	7682	.073	.080	2.3
.100-.150	4957	.123	.132	1.9
.150-.200	3023	.173	.161	-1.8
.200-.250	1834	.223	.195	-3.0
.250-.300	1113	.272	.243	-2.3
.300-.400	1011	.338	.317	-1.4
>.400	395	.476	.372	-4.3

races = 3198, # horses = 29679

Table 10

HARVILLE MODEL CONDITIONAL PROBABILITY OF 3RD

range	n	exp.	act.	Z
.000-.010	660	.007	.009	0.5
.010-.025	2680	.018	.033	4.3
.025-.050	4347	.037	.062	6.8
.050-.100	6646	.073	.087	4.0
.100-.150	4325	.123	.136	2.5
.150-.200	2923	.173	.178	0.7
.200-.250	1831	.223	.192	-3.4
.250-.300	1249	.273	.213	-4.9
.300-.400	1219	.341	.273	-5.3
>.400	601	.492	.333	-8.3

races = 3198, # horses = 26481

The large values of the Z-statistics show the significance of the bias in the Harville formula. The tendency is for low probability horses to finish second and third more often than predicted, and for high probability horses to finish second and third less often. The effect is more pronounced for 3rd place than for 2nd. An effective, and computationally economical way to correct for this is as follows:

Given the win probability array, $(\pi_{i\ (i=1,2,...N)})$, create a second array σ such that,

$$\sigma_i = \frac{\exp(\gamma \log(\pi_i))}{\sum \exp(\gamma \log(\pi_j)) \qquad (j = 1,2,...N)} \tag{7}$$

and a third array τ such that,

$$\tau_i = \frac{\exp(\delta \log(\pi_i))}{\sum \exp(\delta \log(\pi_j)) \qquad (j = 1,2,...N)} . \tag{8}$$

The probability of the three horses (i,j,k) finishing in order is then

$$\pi_{ijk} = \frac{\pi_i \sigma_j \tau_k}{(1 - \sigma_i)(1 - \tau_i - \tau_j)} . \tag{9}$$

The parameters γ and δ can be estimated via maximum likelihood estimation on a sample of past races. For the above data set the maximum likelihood values of the parameters are $\gamma = .81$ and $\delta = .65$. Reproducing Tables 9 and 10 above using equations (7-9) with these parameter values substantially corrects for the Harville formula bias as can be seen in Tables 11 and 12.

Table 11

LOGISTIC MODEL CONDITIONAL PROBABILITY OF 2ND (γ = .81)

range	n	exp.	act.	Z
.000-.010	251	.008	.012	0.6
.010-.025	2282	.018	.024	1.9
.025-.050	5195	.037	.033	-1.6
.050-.100	8819	.074	.073	-0.4
.100-.150	6054	.123	.125	0.5
.150-.200	3388	.173	.176	0.5
.200-.250	1927	.222	.216	-0.6
.250-.300	973	.272	.275	0.2
.300-.400	616	.336	.349	0.7
>.400	174	.456	.397	-1.6

races = 3198, # horses = 29679

Table 12

LOGISTIC MODEL CONDITIONAL PROBABILITY OF 3RD (δ = .65)

range	n	exp.	act.	Z
.000-.010	4	.009	.000	-0.2
.010-.025	712	.020	.010	-2.7
.025-.050	3525	.039	.035	-1.3
.050-.100	8272	.075	.073	-0.7
.100-.150	6379	.123	.130	1.7
.150-.200	3860	.172	.175	0.5
.200-.250	2075	.222	.228	0.7
.250-.300	921	.271	.268	-0.2
.300-.400	582	.337	.299	-2.0
>.400	151	.480	.450	-0.7

races = 3198, # horses = 26481

The better fit provided by this model can be readily seen from the much smaller discrepancies between expected and actual frequencies. The parameter values used here should not be considered to be universal constants, as other authors have derived significantly different values for the parameters γ and δ using data from different racetracks (Lo, Bacon-Shone and Busche, 1994).

FEASIBILITY

A computer based handicapping and betting system could in princple be developed and implemented at most of the world's racetracks. Today's portable computers have sufficient capacity not only for real-time calculation of the bets, but for model development as well. However, several important factors should be considered in selecting a target venue, as potential profitability varies considerably among racetracks. The following are a few practical recommendations based on the author's experience.

Data availability

A reliable source of historical data must be available for developing the model and test samples. The track must have been in existence long enough, running races under conditions similar to today, in order to develop reliable predictions. Data availability in computer form is of great help, as data entry and checking are extremely time-consuming. The same data used in model development must also be available for real-time computer entry sufficient time before the start of each race. Additionally, final betting odds must be available over the development sample for the 'combined' model estimation of equation (1) as well as for wagering simulations.

Ease of operation

Having an accurate estimate of the final odds is imperative for betting purposes. Profitability will suffer greatly if the final odds are much different than the ones used to calculate the probabilities and bet sizes. The ideal venue is one which allows off-track telephone betting, and disseminates the odds electronically. This enables the handicapper to bet from the convenience of an office, and eliminates the need to take a portable computer to the track and type in the odds from the tote board at the last minute. Even given ideal circumstances, a professional effort will require several participants. Data entry and verification, general systems programming, and ongoing model development all require full-time efforts, as well as the day-to-day tasks of running a small business. Startup capital requirements are large, (mainly for research and development) unless the participants forgo salaries during the development phase.

Beatability of the opposition

Pari-mutuel wagering is a competion amongst participants in a highly negative sum game. Whether a sufficiently effective model can be developed depends on the predictability of the racing, and the level of skill of fellow bettors. If the races are largely dishonest, and the public odds are

dominated by inside information then it is unlikely that a fundamental model will perform well. Even if the racing is honest, if the general public skill level is high, or if some well financed minority is skillful, then the relative advantage obtainable will be less. Particularly unfavorable is the presence of other computer handicappers. Even independently developed computer models will probably have a high correlation with each other and thus will be lowering the dividends on the same horses, reducing the profitability for all. Unfortunately, it is difficult to know how great an edge can be achieved at a particular track until one develops a model for that track and tests it, which requires considerable effort. Should that prove successful, there is still no guarantee that the future will be as profitable as past simulations might indicate. The public may become more skillful, or the dishonesty of the races may increase, or another computer handicapper may start playing at the same time.

Pool size limitations

Perhaps the most serious and inescapable limitation on profitability is a result of the finite amount of money in the betting pools. The high track take means that only the most extreme public probability mis-estimations will result in profitable betting opportunities, and the maximum bet size imposed by the bets' effects on the dividends limits the amount that can be wagered. Simulations by the author have indicated that a realistic estimate of the maximum expected profit achievable, as a percentage of total per-race turnover, is in the range of 0.25 - 0.5 per cent. This is for the case of a player with an effectively infinite bankroll. It may be true that at tracks with small pool sizes, that this percentage is higher due to the lack of sophistication of the public, but in any case, it is unlikely that this value could exceed 1.5 per cent. A more realistic goal for a start-up operation with a bankroll equal to approximately one half of the per-race turnover might be to win between 0.1 and 0.2 per cent of the total track turnover. The unfortunate implication of this is that at small volume tracks one could probably not make enough money for the operation to be viable.

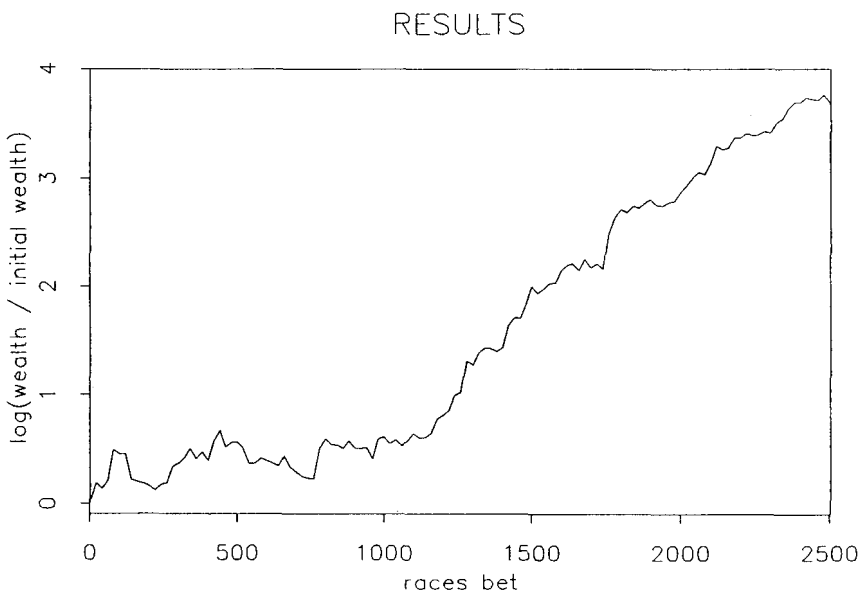
Racetracks with small betting volumes also tend to have highly volatile betting odds. In order to have time to calculate and place one's wagers it is necessary to use the public odds available a few minutes before post time. The inaccuracy involved in using these volatile pre-post-time odds will decrease the effectiveness of the model.

RESULTS

The author has conducted a betting operation in Hong Kong following the principles outlined above for the past five years. Approximately five man-years of effort were necessary to organize the database and develop a handicapping model which showed a significant advantage. An additional five man-years were necessary to develop the operation to a high level of profitability. Under near-ideal circumstances, ongoing operations still require the full time effort of several persons.

A sample of approximately 2000 races (with complete past performance records for each entrant) was initially used for model development and testing. Improvements to the model were made on a continuing basis, as were regular re-estimations of the model which incorporated the additional data accumulated. A conservative fractional Kelly betting strategy was employed throughout, with wagers being placed on all positive expectation bets available in both normal and exotic pools (except place and show bets).[4] Extremely large pool sizes, (> USD $10,000,000 per race turnover) made for low volatility odds, therefore bets could be placed with accurate estimations of the final public odds. Bets were made on all available races except for races containing only *unratable* horses (~5%), resulting in approximately 470 races bet per year. The average track take was ~19% during this period.

Four of the five seasons resulted in net profits, the loss incurred during the losing season being approximately 20% of starting capital. A strong upward trend in rate of return has been observed as improvements were made to the handicapping model. Returns in the various betting pools have correlated well with theory, with the rate-of-return in exotic pools being generally higher than that in simple pools. While a precise calculation has not been made, the statistical significance of this result is evident. Following is a graph of the natural logarithm of [(wealth) / (initial wealth)] versus races bet.



CONCLUSION

The question; "Can a system beat the races?" can surely be answered in the affirmative. The author's experience has shown that at least at some times, at some tracks, a statistically derived fundamental handicapping model can achieve a significant positive expectation. It will always remain an empirical question whether the racing at a particular track at a particular time can be beaten with such a system. It is the author's conviction that we are now experiencing the *golden age* for such systems. Advances in computer technology have only recently made portable and affordable the processing power necessary to implement such a model. In the future, computer handicappers may become more numerous, or one of the racing publications may start publishing competent computer ratings of the horses, either of which will likely cause the market to become efficient to such predictions. The profits have gone, and will go, to those who are 'in action' first with sophisticated models*+.

*An earlier version of this paper was presented at the ORSA/TIMS Joint National Meeting in Phoenix, Arizona on November 1, 1993.

+The author wishes to thank Professor George Miel (University of Nevada, Las Vegas), Paul Coladonato, Randall G. Chapman, and the editors of this volume for many helpful comments, suggestions, and corrections.

NOTES

[1]One technique to alleviate the negative consequences of biases which lead to the over-estimation of advantage is to employ a betting rule which specifies a minimum estimated advantage necessary for making a bet. In Ziemba and Hausch (1987) the authors suggest a mimimum advantage of 10% to account for the bias in their place and show betting model. Also, in their model the authors use place and show probabilities so the often present favorite-longshot win bias tends to cancel with the second and third place reverse bias. For simple probability estimations these schemes can work well, but in exotic bets whose probabilities are the products several individual win probabilities, the calculation of the correct minimum advantage becomes exceedingly complex. The author advocates the practice of correcting the probabilities first and then calculating the betting advantage.

[2]A similar calculation to the one carried out in the quinella pool example above shows that a horse with a positive expected return in the win pool will have a lower expected return as a place or show bet, given that the public bets consistently in the different pools. This effect is different than the one which produced advantages in the place and show pools for Ziemba and Hausch (1987). There the advantages arose because of inconsistencies between the public's estimated win probability for a horse, and the amount bet on that horses in the place or show pools.

[3]The bias in this formula is not as serious when used with win probabilities that show a significant favorite-longshot bias. The favorite-longshot bias often observed at racetracks (Ali, 1977) tends to cancel out the Harville formula bias in estimating second and third place probabilities.

[4]Betting off-track, the author did not have access to real-time show pool betting information. (Place betting in the North American sense is not available in Hong Kong.) Without individual horse show pool betting information, one can always achieve higher advantages by betting in 'exotic' pools such as quinella and trifecta. This follows from the above cited principle of 'the more specific the bet, the higher the advantage'. (See Note 2 above)

APPENDIX

HANDICAPPING REFERENCES*

Ainslie, Tom, *Ainslie's Complete Guide to Thoroughbred Handicapping*, (New York: Simon & Schuster, 1979)

Beyer, Andrew, *Picking Winners*, (Boston, MA: Houghton Mifflin Company, 1975)

Jones, Glendon, *Horse Racing Logic*, (New York: Vantage Press, 1989)

Quinn, J., *The ABC's of Thoroughbred Handicapping*, (New York: William Morrow and Company, 1988)

Quirin, William L., *Winning at the Races: Computer Discoveries in Thoroughbred Handicapping*, (New York: William Morrow and Company, 1979)

Scott, Don, *The Winning Way*, (Sydney: Puntwin PTY Limited, 1982)

*The following is a partial list of references which the author has found helpful in suggesting ideas for significant *factors*. A useful source for difficult to find books on handicapping is 'The Gambler's Book Club' in Las Vegas, Nevada.

REFERENCES

Ali, M., "Probability and Utility Estimates for Racetrack Betting," *Journal of Political Economy*, 85 (1977), 803-815.

Asch, P., R.E. Quandt, *Racetrack Betting: The Professors' Guide to Strategies,* (Dover, MA: Auburn House, 1986)

Bolton, Ruth N. and Randall G. Chapman, "Searching for Positive Returns at the Track: A Multinomial Logit Model for Handicapping Horse Races," *Management Science,* Vol. 32, No. 8, August (1986), 1040-1059.

Brecher, Stephen L., *Beating the Races with a Computer,* (Long Beach, CA: Software Supply, 1980)

Epstein, Richard A., *The Theory of Gambling and Statistical Logic, revised edition*, (New York, NY: Academic Press, 1977)

Figlewski, Stephen, "Subjective Information and Market Efficiency in a Betting Market," *Journal of Political Economy*, Vol. 87, No. 1, (1979), 75-88.

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

Henery, R.J., "Permutation Probabilities as Models for Horse Races,"*Journal of the Royal Statistical Society B* 43, No. 1, (1981), 86-91.

Kallberg, J.G. and W.T. Ziemba, "Pari-mutuel Betting Models," *in this volume* (1994).

Kelly, J., "A New Interpretation of Information Rate," *Bell System Technical Journal.*, 35 (1956), 917-926.

Lo, Victor S.Y. and John Bacon-Shone, "Approximating the Ordering Probabilities of Multi-entry Competions by a Simple Method," *working paper, Department of Statistics, University of Hong Kong,* (1992).

Lo, Victor S.Y., John Bacon-Shone and Kelly Busche "The Application of Ranking Probability Models to Racetrack Betting," *Management Science*, forthcoming (1994).

MacLean, L.C., W.T. Ziemba and G. Blazenko, "Growth Versus Security in Dynamic Investment Analysis," *Management Science*, Vol. 38, No. 11, November (1992), 1562-1585.

Stern, Hal, "Models for Distributions on Permutations," *Journal of the American Statistical Association*, 85, No. 410 June (1990), 558-564.

White, E.M., Ronald Dattero and Benito Flores, "Combining Vector Forecasts to Predict Thoroughbred Horse Race Outcomes," *International Journal of Forecasting* 8 (1992), 595-611.

Ziemba, William T. and Donald B. Hausch, *Betting at the Racetrack,* (Los Angeles: Dr. Z Investments, Inc., 1986)

Ziemba, William T. and Donald B. Hausch, *Dr. Z's Beat the Racetrack, revised edition,* (New York: William Morrow and Company, Inc., 1987)

An Empirical Cross-validation of Alternative Classification Strategies Applied to Harness Racing Data for Win Bets[1]

Larry H. Ludlow
Boston College

ABSTRACT

This paper presents the results of a two year cross-validation of a fundamental handicapping system. Harness race performance data from a single season's entire racing meet were subjected to a discriminant analysis and classification criteria were developed. The Year 1 discriminant function and classification criteria were applied to Year 2 data. The classification techniques are evaluated in terms of percent correct classification and return on investment.

INTRODUCTION

This study was designed to cross-validate the relative efficacy of six alternative classification techniques. In the present study the term discrimination refers to the statistical process of deriving a fundamental harness race handicapping system capable of differentiating between winners and non-winners. The term classification refers to the subsequent application of the initial rules to a second sample of races.

The data consist of performance observations of winning and losing horses racing in two seasons of harness meets (referred to as Year 1 and Year 2). It is not assumed that those two particular meets provided unique data that could not have occurred elsewhere or during a different racing season. The discrimination problem focused on whether a linear function existed that would yield significant separation between winners and non-winners based on the Year 1 data. The classification problem took the Year 1 discriminant function and classification criteria and cross-validated them upon the Year 2 data.

---

[1]Appreciation for assistance and advice is extended to Peter Tommila, Nicholas Bond, Kenneth Krueger, William Ziemba, and Donald Hausch. Correspondence should be addressed to Larry H. Ludlow, Associate Professor, Boston College, School of Education, Chestnut Hill, MA 02167-3813.