Johnnie Johnson, Owen Jones and Leilei Tang

Exploring decision-makers’ use of price information in a speculative market

Abstract

We explore the extent to which the decisions of participants in a speculative market effectively account for information contained in prices and price movements. The horserace betting market is chosen as an ideal environment to explore these issues. A conditional logit model is constructed to determine winning probabilities based on bookmakers’ closing prices and the time indexed movement of prices to the market close. The paper incorporates a technique for extracting predictors from price (odds) curves using orthogonal polynomials. The results indicate that closing prices do not fully incorporate market price information, particularly information which is less readily discernable by market participants.

1. Introduction

There is a large body of evidence which suggests that the combined judgements of decision-makers within financial markets effectively incorporate information concerning historical prices into current prices, to the extent that abnormal returns cannot be made if buy/sell decisions are made on the basis of historical prices. The horserace betting market is one form of financial market that has received considerable scrutiny in this regard. An important reason for this focus is that “wagering markets are especially simple financial markets, in which the pricing problem is reduced. As a result, wagering markets can provide a clear view of pricing issues which are complicated elsewhere” (Sauer, 1998, p. 2021).

Previous studies exploring the extent to which horserace betting markets incorporate closing odds and the time indexed movement of pre-closing odds fall into three broad categories. One set of studies explores the distribution of closing odds in horserace betting markets. These offer
strong evidence for a consistent over/under estimation of the probability of longshots/favourites winning (the favourite-longshot bias). These conclusions have been reached for studies widely dispersed in time and across a variety of countries (e.g. McGlothlin, 1956; Ali, 1977; Tuckwell, 1983; Ziemba and Hausch, 1986; Bird and McRae, 1994; Bruce and Johnson, 2000). However, these studies do not generally detect opportunities for trading profitably on this information\(^1\); suggesting that historical prices are largely discounted in closing odds.

A second set of studies explores the information content associated with changes in odds from the opening of the market to its close. Information held by those with privileged information may be transmitted to the market via their betting behaviour; this may be revealed as bookmakers adjust odds to account for their liabilities or parimutuel odds change to reflect relative volumes of bets. In general, previous studies suggest that betting activity reveals information that is not readily available prior to the formation of the market, but that betting strategies based on these odds adjustments do not yield positive expected returns (e.g. Asch, Malkiel and Quandt, 1982; Bird and McCrae, 1987; Crafts, 1985; Schnytzer and Shilony, 2002; Tuckwell, 1983). Sauer (1998, pp. 2048-2049), reviewing these studies, concludes that “the evidence suggests that an informed class of bettors is responsible for altering prices in these markets . . (but) the opinions of ‘experts’ appear to be fully discounted in market prices”.

A third set of studies attempt to construct arbitrage betting strategies which employ information revealed within one market in a parallel market. Some of these studies have demonstrated that positive returns are possible by exploiting differences in pre-closing odds in independent win pools (e.g. Hausch and Ziemba, 1990; Schnytzer and Shilony, 1995) or between win and place or show pools\(^2\) (e.g. Hausch, Ziemba and Rubinstein, 1981). However,

---

1. The one exception, a study conducted by Ziemba and Hausch (1986) in the USA, identifies a small expected profit from betting on a very small number of extreme favourites (odds < 3/10).
2. In the parimutuel market, separate pools are created for win bets (those which attempt to select the winner), place bets (those which attempt to select the horse to finish first or second) and for show bets (those which attempt to select horses to finish first, second or third). Odds are determined separately in each pool by the
opportunities for profitable betting employing the above strategies are limited and the degree of inefficiency is often small. Consequently, this third set of studies offer little more than a “crease in what is predominantly a smooth pattern of efficiency in the racetrack betting market” (Sauer, 1998).

It is clear from the preceding discussion that various aspects of closing odds and odds movements from the opening of the market to its close, taken alone (e.g. under-betting of favourites (Vaughan-Williams and Paton, 1997); odds assessed by different handicappers (Figlewski, 1979); differences between morning line and closing odds (Crafts, 1985); odds at different time points (Lo, 1994)) do contain information, but that this is discounted in closing odds, to the extent that profitable wagering strategies based on this information cannot be constructed. However, Ceci and Liker (1986) demonstrate that expert horse handicapping requires bettors to combine different types of information in complex, interactive models. We set out, therefore, to develop such a comprehensive model. This model combines closing odds and variables derived from the movement of pre-closing odds to closing odds that capture information which (a) may not be closely associated in the public’s mind with a horse’s success (e.g. the volatility of odds changes) and (b) are less readily discernable by bettors (e.g. odds changes scaled by closing odds). In addition, we will include interaction terms between the variables. Our view is that bettors may not readily discount the combined effect of such variables and their interactions. Consequently, it may be possible to construct profitable wagering strategies based on such a model, and hence demonstrate that the horserace betting market is not weak form efficient.

To achieve our aim the paper proceeds as follows. Section 2 outlines the data employed, provides a rationale for the model which is used to test for market efficiency, and discusses the relative amount of money on each horse.
significant components of the fitted model. In Section 3 the model is tested, to explore the extent to which it demonstrates that information on changes in pre-closing odds is discounted in closing odds. Finally, in Section 4 we summarize our conclusions.

**2. Description of the data and the model**

There are two distinct forms of horserace betting market that operate in parallel at racetracks in the UK; the parimutuel and the bookmaker markets. The latter forms the setting for this study. The odds on offer in the bookmaker market are determined by the decisions of both bettors and bookmakers and bets are settled at the odds available in the market at the time the bet is struck. The more serious bettors and those with access to privileged information are most likely to bet in the bookmaker market (Crafts, 1985; Sauer, 1998; Schnytzer and Shilony 1995) since they can secure their return, without the possibility of a bandwagon effect eroding their gains (which can happen in the parimutuel market). Consequently, since we seek to exploit as much information as possible contained in closing odds and pre-closing odds and their changes over time, it seems appropriate to use bookmakers’ odds.

Independent bookmakers (generally 10 to 50) operate at each racetrack and post odds at the commencement of the market before each race. These odds change according to the relative weight of demand, reflecting bettors’ opinions, and according to each particular bookmaker’s subjective view of the horses’ relative prospects. Bettors in the UK are also permitted to bet in off-track betting offices at the pre-closing odds (or the closing odds) prevailing in the on-course market at any given time. An independent organisation, SIS, transmits the evolving racetrack bookmakers’ odds to off-course betting offices. SIS employs assessors who use their judgement to determine a unique value for the odds available on each horse at each moment in time; these are the maximum bookmaker odds available to a ‘substantial wager’ at the track. Accordingly, we collected pre-closing and closing bookmaker odds data supplied by SIS from 1,200 races run at 41 different racetracks in the UK over the period April – June 1998. Only ‘flat’ races of less
than 2 miles were included. The number of horses in each race varies from 2 to 20, with a mode of 11. The betting period for each race lasts from 2.5 to 30 minutes, and averages 12.5 minutes ($\sigma = 4.5$ minutes). The odds for a given horse in the sample can change from 0 to 10 times, with a mode of 2 and an average of 2.7 changes.

The information employed consists of, for each race $i$ and each horse $j = 1, \ldots, k(i)$ in race $i$, a sequence of times and odds $\{(t_{i,j}(1), u_{i,j}(1)), \ldots, (t_{i,j}(n), u_{i,j}(n))\}$. This sequence is unique for horse $j$ in race $i$, being the odds transmitted by SIS. The final pair $(t_{i,j}(n), u_{i,j}(n))$ is always the ‘off time’ and ‘closing odds’. The length of the sequence $n = n(i,j)$ varies from horse to horse, and the number of horses $k = k(i)$ varies from race to race. When we are in the context of a single race we will drop the index $i$ and when in the context of a single horse we will drop the indices $i$ and $j$.

We scale the times so that $t(1) = 0$ and $t(n) = 1$. We regard the odds as a function of time, piecewise constant with jumps where the odds change, and call this function a ‘price curve’. The left-hand diagram of figure 1 displays a typical example of a price curve: the points * are the points $(t(1), u(1)), \ldots, (t(n), u(n))$. Note that the final pair $(t(n), u(n))$ are generated by the start of the race and not by a change in odds, so $u(n) = u(n-1)$.

Taking the price curves as its input, we wish to build a model for $p_i = (p_i(1), \ldots, p_i(k))$, the vector of winning probabilities for race $i$, where $p_i(j) = \Pr(\text{horse } j \text{ wins race } i)$. Suppose that for horse $j$ in race $i$ we have extracted from the price curve, predictors $x_{i,j} = (x_{i,j}(1), \ldots, x_{i,j}(m))$, where $m$ is fixed over all $i$ and $j$. We use a conditional logit model for the $p_i$. That is, for a fixed vector of coefficients $\beta = (\beta(1), \ldots, \beta(m))$, we suppose that

$$
p_i(j) = \frac{\exp(<\beta, x_{i,j}>)}{\sum_{l=1}^{k(i)} \exp(<\beta, x_{i,l}>)}, \quad (1)
$$

where $<\beta, x_{i,j}> = \sum_{h=1}^{m} \beta(h)x_{i,j}(h)$. We justify this choice of model by noting that it allows the exponent $<\beta, x_{i,j}>$ to be interpreted directly as the ability of horse $j$, independent of the race $i$. 

To see this, suppose that $\varepsilon(j)$, $j = 1, \ldots, k(i)$ are independent identically distributed random variables with the double exponential distribution. That is $\varepsilon(j)$ has cumulative distribution function $F_{\varepsilon}(v) = \exp(-\exp(-v))$, for $-\infty < v < \infty$. If we put $W_j(j) = <x_j, \beta>$ then it can be shown that $\Pr(W_j(j) \geq W(l), l = 1, \ldots, k(i)) = p(j)$ (Maddala, 1983). We can interpret $W_j(j)$ as a ‘winningness’ index. That is, the winner of race $i$ is the horse with maximal $W(j)$, and we can interpret the deterministic component $<x_j, \beta>$ of $W_j(j)$ as a direct measure of horse $j$’s ability.

If we observe $N$ races, and the winner of race $i$ is horse $j^*$, then the joint likelihood $L = L(\beta)$ is the probability of observing this set of results, assuming the $p_i$ are as above. That is

$$L(\beta) = \prod_{i=1}^{N} p_j(j^*) = \prod_{i=1}^{N} \frac{\exp(<x_{i,j^*}, \beta>)}{\sum_{j=1}^{k(i)} \exp(<x_{i,j}, \beta>)}. \quad (2)$$

We employ maximum likelihood estimation to choose $\beta$ that maximizes $L(\beta)$.

2.1. Extracting predictors from the price curve: To construct a model for $p_j(j)$ we require a consistent set of predictors, drawn from the price curve for each horse. In doing so we have two aims: to provide a general summary of the shape and other physical characteristics of the price curve, and to pick out particular features that have been identified in the literature or by active gamblers as having an effect on the horse’s winning probability. For all the predictors we consider there is an implicit dependence on the race $i$ and horse $j$, though, to simplify, we will not make this explicit in the notation.

We are interested to include predictors that reflect the general shape of the price curve; which, amongst other things, will reflect the timing of bets by those with privileged information. However, the precise shape of a price curve that is likely to signal a potential winner (or loser) is unclear. We therefore provide a general summary of the shape of a price curve $\{(t(1), u(1)), \ldots, (t(n), u(n))\}$ by using an orthogonal polynomial expansion of order 3. This allows us to measure the height of the curve (closing odds), the linear trend, the curvature (quadratic component) and
change in curvature (cubic component). By using an orthogonal polynomial expansion, we can
measure the size of each component (constant, linear, quadratic and cubic) independently of the
others. Orthogonal polynomials are a classical statistical tool; details of their construction can be
found for example in Wetherill (1981). We summarise the procedure here and it is illustrated in
figure 1.

![Figure 1. Orthogonal polynomial decomposition of a price curve. The first diagram shows the
price curve and its constant, linear, quadratic and cubic approximations. The second diagram
shows the separate constant, linear, quadratic and cubic components, which are added to give
the approximations in the first diagram.]

Given a set of points \{((t(1),u(1)), \ldots, (t(n),u(n)))\}, an orthogonal polynomial basis is a
sequence of polynomials \(f_0, f_1, f_2, \ldots\) such that \(f_i\) is of order \(i\) and

\[
\sum_{l=1}^{n} f_i(t(l)) f_j(t(l)) = 0 \quad \text{for all } i \neq j. \tag{3}
\]

It can be shown that an orthogonal polynomial basis always exists and that there is a unique set of
coefficients \(a_0, a_1, a_2, \ldots\) such that

\[
u(l) = \sum_{i=0}^{\infty} a_i f_i(t(l)) \quad \text{for all } l = 1, \ldots, n. \tag{4}\]

The \(a_i\) can be found by least squares. By restricting ourselves to an order 3 expansion we get an
approximation to \(u\). As the \(f_i\) are orthogonal, we can interpret \(a_i\) as the size of the order \(i\)
component in the price curve. In fact, the equations (3) do not specify a unique basis, and we can
impose further constraints without compromising orthogonality. In our case, because of the importance of the closing odds $u(n)$, we take $f_0(t(n)) = f_0(1) = 1$ and $f_i(1) = 0$ for all $i \geq 1$. The effect of this is to make the constant component $a_0$ equal to $u(n)$. We also norm each $f_i$ so that its leading term is simply $t^i$. In particular, this implies that $a_1$ is the slope of the least squares regression line constrained to pass through $(t(n),u(n))$. We interpret $a_2$ as a measure of the curvature of the price curve and $a_3$ as a measure of the change in curvature.

A potential problem with polynomial expansions is that they are unstable when only a small number of points are used. That is, if $n$ is small, then a small change in one of the $(t(i),u(i))$ can produce a large change in $a_2$ and $a_3$. To mitigate this, we regularise the procedure by introducing a roughness penalty when fitting the $a_i$. Let $F(t) = \sum_{i=0}^3 a_i f_i(t)$, then we choose the $a_i$ to minimize

$$\sum_{j=1}^n \left( (u(j) - F(t(j)))^2 + \lambda G(F) \right)$$

where $G(F)$ is the roughness penalty and $\lambda$ is some constant of proportionality. Typically $G(F)$ is some measure of curvature such as a Sobelov norm (that is, a norm based on first, second and sometimes higher order derivatives of $F$). However, the slope of $F$ at $t(n)$ is of particular interest to us as a measure of late price movement (see below), so we do not want to depress this unnecessarily. So, instead of a Sobelov norm, we put $G(F)$ equal to the area of $F$ above $u_{\text{max}}$ and below $u_{\text{min}} = \min u(i)$. That is,

$$G(F) = \int_0^1 \max(F(t) - u_{\text{max}}, 0) dt + \int_0^1 \max(u_{\text{min}} - F(t), 0) dt.$$  

(6)

In addition to a general summary of the shape of the price curve (discussed above) we also included two predictors to measure the volatility or roughness of the price curve; the number of price changes and the absolute variation. That is, for price curve $\{(t(1),u(1)), \ldots, (t(n),u(n))\}$, we put
\[ b_1 = n - 2 \text{ and } b_2 = \sum_{i=2}^{n-1} |u(i) - u(i-1)|. \] (7)

Clearly there will be some degree of correlation between \( b_1 \) and \( b_2 \).

The remaining three predictors in the model were included since previous literature suggests that they may have some influence on a horse’s probability of success. Firstly, a number of studies have demonstrated a close correspondence between probabilities implied by closing odds and winning probabilities (e.g. Bruce and Johnson, 2000). The closing odds for horse \( j \) in race \( i \) imply a probability of winning \( q_i(j) \), via the relationship

\[ u_{i,j}(n) = \frac{1-q_i(j)}{q_i(j)}, \quad q_i(j) = \frac{1}{1+u_{i,j}(n)}. \] (8)

We call this the ‘track probability’ to distinguish it from the model probability \( p_i(j) \). We already include the closing odds in the model as \( a_0 = u_{i,j} \), which gives \( p_i(j) \propto \exp(\beta(a_0) \cdot (1 - q_i(j)) / q_i(j)) \). However, it is plausible that a more direct relationship between \( p_i(j) \) and \( q_i(j) \) would result in a better fit. Accordingly, we include the predictor \( c_1 = \ln q_i(j) = -\ln(1 + u_{i,j}(n)) \), which gives \( p_i(j) \propto \exp(\beta(c_1) \cdot c_1) = q_i(j)^{\beta(c_1)} \). In fact, Chapman (1994) even found that \( c_1 \) added significant explanatory power in a sophisticated fundamental handicapping model that included 20 variables associated with the horse and its jockey.

Crafts (1985), Tuckwell (1983) and Bird and McCrae (1987), amongst others, have demonstrated that a horse’s enhanced prospects of success are revealed by a large reduction in odds from the start to the completion of the market. Consequently, we take \( c_2 = \text{closing odds} - \text{initial odds} = u_{i,j}(n) - u_{i,j}(1) \), although this will be highly correlated with the slope \( a_1 \).

Those with access to privileged information have an incentive to bet late in the market (Asch, Malkiel and Quandt, 1983; Schnytzer, Shilony and Thorne, 2003). Consequently we use a predictor to capture late changes in the betting. Let \( U(t) \) be the odds at time \( t \), that is for \( t(i) \leq t < t(i+1) \), \( U(t) = u(i) \). Let \([a, b]\) be a small subinterval of \([0, 1]\), close to 1. We take as our
measure of late change the most extreme slope \((U(1) - U(t))/(1 - t)\) for \(t \in [a, b]\), that is, the slope with the largest absolute value. We let the late change predictor be \(c_3\), and provide two illustrations in figure 2, where \(c_3\) is the slope of the line plotted through \((1, U(1))\).

![Figure 2: Late change predictor.](image)

Values of \(a\) and \(b\) were chosen to make \(c_3\) reasonably robust, so that small changes in \(U(t)\) do not produce large changes in \(c_3\), while minimising correlation with the overall trend \(a_1\). Taking \([a, b] = [0.9, 0.95]\) gave reasonable results.

If the price curve were smooth, then \(c_3\) would simply be an approximation to the derivative at \(t = 1\). Consider again our orthogonal polynomial expansion of the price curve, \(F(t) = \sum_{i=0}^{3} a_i f_i(t)\). \(F(t)\) is a smooth approximation of the price curve, so we expect \(c_3\) to be highly correlated with \(F'(1) = a_1 + 2a_2 + 3a_3\).

Finally, two refinements of the predictor set were incorporated, based on our understanding of how prices behave in practice. Firstly, it is known that on long-odds horses (those with high closing odds), the odds change by larger amounts than for short-odds horses, and we believe that the relative size is more important than the absolute size of any change. Consequently, before calculating predictors \(a_1, a_2, a_3, b_2, c_2\) and \(c_3\), the price curve \(\{(t(1), u(1)), \ldots, (t(n), u(n))\}\) was rescaled by dividing \(u(i)\) by \(u(n)\) for \(i = 1, \ldots, n\). Predictors \(a_0\) and \(c_1\), which are based on the closing odds, were not rescaled, and \(b_1\) is unaffected by this rescaling. Secondly, practicing gamblers interpret the price curve differently when the odds are coming in (decreasing) or going out (increasing). This suggests that we should interpret price changes differently when they are...
changes down rather than up. Accordingly predictors $a_1$, $c_2$ and $c_3$ were split into pairs $x_i^+$ and $x_i^-$, where $x_i^+ = x_i$ if $x_i > 0$ or 0 otherwise, and $x_i^- = x_i$ if $x_i < 0$ or 0 otherwise. Similarly, predictors $b_1$ and $b_2$ were split into two parts, depending on whether or not $a_1 > 0$.

**2.2. Model fitting:** In order to fit and test the model given in Equation (1) the data set was split into two parts. The first 800 races were used to fit the model, and the remaining 400 used to test it. A stepwise fitting procedure was used to select a set of predictors significant at the 5% level. Pairwise interactions of all the predictors were also considered. In the final model the predictors $a_1^+$, $b_2^-$, $c_1$, $c_3^-$ and the interaction $a_1^+ b_1$ were all significant at the 5% level. The estimated coefficients $\beta$ are given in table 1. The log-likelihood ratio of the model over the constant alternative is 857.05, which gives us that the model is significant with a p-value of 0.0000.

Table 1. Estimated coefficients of the model. Note that $a_1^+$, $b_2^-$ and $a_1^+ b_1 \geq 0$; $c_1$ and $c_3^- \leq 0$.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Description</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1^+$</td>
<td>slope up</td>
<td>-2.0493</td>
<td>0.8110</td>
<td>0.0115</td>
</tr>
<tr>
<td>$b_2^-$</td>
<td>absolute variation down</td>
<td>-0.4227</td>
<td>0.1907</td>
<td>0.0266</td>
</tr>
<tr>
<td>$c_1$</td>
<td>$\ln$(track probability)</td>
<td>1.1678</td>
<td>0.0648</td>
<td>0.0000</td>
</tr>
<tr>
<td>$c_3^-$</td>
<td>late change down</td>
<td>-0.0666</td>
<td>0.0331</td>
<td>0.0440</td>
</tr>
<tr>
<td>$a_1^+ b_1$</td>
<td>slope up and number of changes interaction</td>
<td>0.4371</td>
<td>0.1937</td>
<td>0.0240</td>
</tr>
</tbody>
</table>

Firstly we note that, as $a_1$, $b_2$ and $c_1$ are in the model, it is not surprising that $c_2$, $b_1$ and $a_0$ are not, as we knew these predictors were correlated. We interpret this as saying that $a_1$, $b_2$ and $c_1$, respectively, capture more relevant information concerning overall change in odds, volatility and closing odds than $c_2$, $b_1$ and $a_0$. Secondly, some degree of correlation between $c_3$ and $a_1 + 2a_2 + 3a_3$ was expected, so the presence of $c_3$ in the model in part explains why $a_2$ and $a_3$ do not appear.

The predictor $c_1 = \ln q(j)$, has a large influence. Considering just the effect of $c_1$ on the model probability we have $p(j) \propto \exp(1.1678 \ln q(j)) = q(j)^{1.1678}$, which is similar to the utility
function derived in Ali (1977). We interpret this relationship as a reflection of the so-called favourite-long shot bias. That is, the shorter a horse’s closing odds, the nearer the true probability of winning is to the implied track probability. A plot of $\ln u_{i,j}(n) = \ln (1-q_i(j))/q_i(j)$ against $\ln q_i(j)^{1.1678}$ gives a very close match to the analogous plot given in Bruce and Johnson (2000), which was obtained by modelling the favourite-longshot bias directly.

The significance of the predictor $a_1^+$ suggests that when odds lengthen (i.e. the least squares regression line constrained to pass through $t(n)$, $u(n)$ has a positive slope) the horse has a significantly lower chance of winning. The effect of the $a_1^+b_1$ interaction is to reduce this effect when the odds worsen in a large number of small steps, as opposed to small number of large steps. The late change down predictor, $c_3^-$ acts as expected. $c_3^-$ is always negative in sign, so when there is a late change down the effect is to increase the probability of winning. The coefficient of $b_2^-$ is negative, indicating that high volatility in the price curve makes a horse less likely to win, but only when the odds have come in.

Each of the predictors with significant coefficients might be described as relatively difficult to discern (relatively opaque) compared with an equivalent, but more transparent predictor excluded from the model; for example, $a_1$ vs. $c_2$, slope of the least squares regression line through $t(n)$, $u(n)$ vs. closing - initial odds; $b_2$ vs. $b_1$, the absolute value of (scaled) odds changes vs. number of odds changes; $c_1$ vs. $a_0$, natural log of the probability implied by closing track odds vs. closing track odds. Whilst the late change predictor $c_3$, which appears in the model, has no directly comparable transparent alternative, we found that the coefficient for the scaled version of $c_3$ is significant whereas that for the non-scaled (more transparent) version of $c_3$ is not. Taken together, the results suggest that more readily discernable (or relatively transparent) information is more efficiently discounted in betting markets than more opaque information.
3. Model testing

In order to test the model we: (i) explore, using log-likelihood ratio tests, whether the model contains more information than a model based solely on closing odds (ii) test whether the additional information contained in the model can be exploited to make profits (iii) use cross-validation and jack-knifing to check that the observed profits do not arise by fortunate selection of the training and validation sets; (iv) use parametric bootstrapping (on the training set employed in (ii)) to test whether the profits we observe arise because our model incorporates more information than that available from closing odds alone or whether the profits arise by chance (v) explore the features of winning bets suggested by the model in order to identify any systematic features of such bets.

3.1. Log likelihood tests: In order to explore the joint importance of the predictors in the model we conduct log-likelihood ratio tests. We compare the log-likelihood (LL) of the full model given in table 1 (LL = -2243.9) with that of a model simply using normed final track probabilities (LL = -2262.7). A LL ratio test comparing the two models has a $p$ value $< 0.0001$. This test confirms that there is significant sample evidence to suggest that the full model incorporates more information than final track probabilities alone. In the following section we explore the extent to which this extra information is substantive, to the extent that we are able to employ it profitably.

3.2. Kelly Betting: To assess whether the observed inefficiency can be exploited sufficiently to make profits, races 801 to 1200, run during May/June 1998, were used to test the predictive ability of the model. As we do not have repeated observations (each race is only run once), we must use indirect methods. We consider a betting strategy based on maximum expected log payoff (Kelly strategy). We use the model probabilities and closing odds as inputs, and analyse the returns produced. Given correct probabilities as inputs (as opposed to estimated probabilities), the strategy gives non-negative expected returns. Thus, if it gives non-negative
returns using our model probabilities \( p_i(j) \) as inputs, we take this as evidence that the \( p_i(j) \) are reasonably accurate. Moreover, a positive return indicates that abnormal returns can be made by simply employing historical price information.

Let \( r_i(j) = 1 + u_{ij}(n) \) be the return on a bet of 1 pound if horse \( j \) wins race \( i \). The Kelly strategy requires that in race \( i \) we bet a fraction \( f_i(j) \) of our current wealth on horse \( j \). Let \( f_i = (f_i(1), \ldots, f_i(k)) \). As usual, we will drop the subscript \( i \) when the context makes it unnecessary. Betting fraction \( f_i \), if horse \( x \) wins then our current wealth will increase by a factor of \( 1 - \sum_{j=1}^{k} f_i(j) + f_i(x)r(x) \). The Kelly strategy consists of choosing \( f_i \) to maximise the expected log payoff, \( F(f) \) where

\[
F(f) = \sum_{x=1}^{k} p(x) \ln\left(f(x)r(x) + 1 - \sum_{j=1}^{k} f_i(j)\right). \tag{9}
\]

This betting strategy was introduced by Kelly (1956). It was later shown to be asymptotically optimal by Breiman (1961), in the sense that it maximises the asymptotic rate of growth for wealth, with 0 probability of ruin. Using the Kelly criterion, the total wealth grows at an exponential rate, though the standard deviation remains proportional to total wealth and thus also grows exponentially. We also note that this strategy only gives 0 probability of ruin if arbitrarily small bets are allowed. In practice this caveat has led some authors to consider modified Kelly strategies (e.g. Benter, 1994; Ziemba and Hausch, 1986), whereby some fixed fraction of \( f \) is bet. As we are interested in the theoretical rather than practical performance of our model, we restrict ourselves to the usual form.

The Kelly strategy tells us which races to bet on, as well as how much to bet on each horse. We can bet on more than one horse in a race, though our bets are restricted to horses that give a positive expected return. In this manner, the Kelly betting strategy makes use of the whole vector of probabilities provided by the model, not just for the horse most likely to win. By betting on a number of horses in a race the risk of losing is reduced at the expense of reducing the expected
return. None the less, bet sizes suggested by the Kelly strategy will be larger when the probability of winning is greater (for the same expected return) and when the expected return is greater (for the same winning probability). Consequently, a Kelly betting strategy makes greater use of information provided by the model than a simple strategy of maximizing the expected return.

Figure 3 plots the natural logarithm of the cumulative wealth, applying the Kelly strategy to the test set, starting with initial wealth 1. Over the out of sample test period, total wealth increased by a factor of 2.4597.

![Figure 3. Ln of cumulative wealth using the Kelly strategy. On the left we have the wealth for races 1 to 800 (those used to fit the model), and on the right the wealth for races 801 to 1200 (the test data set). Here wealth is given as a multiple of original wealth.](image)

In testing the significance of returns from bets it is common to consider the profit per pound bet $B_i$. However, if we let $X_n$ be the wealth after $n$ races, then we obtain

\[ X_{i+1} = X_i + b_i X_i B_i = (1 + b_i B_i) X_i, \]

where $b_i$ is the amount bet per pound of initial wealth on race $i$. That is, we cannot express the increase in wealth solely in terms of the $B_i$. Let $w_i$ be the profit made per pound of initial wealth on race $i$, so that $B_i = w_i / b_i$ (defining $B_i = 0$ when $b_i = 0$). In the context of Kelly betting, a more natural object to consider than $B_i$ is $W_i = 1 + w_i$, which is the factor by which wealth has increased after race $i$. That is $X_{i+1} = W_i X_i$. Taking
logs we turn the multiplicative form of the cumulative wealth into an additive form, to obtain
\[ \ln(X_n) = \ln(X_0) + \sum_{i=1}^{n} \ln(1 + w_i). \] Let \( A_i = \ln(1 + w_i) \), then \( A_i \) is a natural object to average, and \( X_n \) exhibits long term growth if and only if \( E(A_i) > 0 \). Note that it is not the case that \( X_n \) exhibits long term growth if and only if \( E(B_i) > 0 \), since there is (typically) positive correlation between \( b_i \) and \( B_i \), in which case \( E(1 + b_i B_i) > 1 + E(b_i)E(B_i) \).

From the testing set we estimate \( \mu = E(A_i) = 0.00226 \) and \( \sigma^2 = \text{Var}(A_i) = 0.0477^2 \). When using Kelly betting, \( \mu \) is the asymptotic growth rate for wealth (per race). From the Central Limit Theorem our estimators are approximately Normally distributed, which allows us to calculate that a test of the hypothesis \( \mu = 0 \) against the alternative \( \mu > 0 \) is significant at the 17% level. Using our estimates for \( \mu \) and \( \sigma^2 \) we can estimate the number of races that we would need to consider to ensure that we are 95% certain of making profit (noting that not all races are bet on). This corresponds to the smallest value of \( n \) such that \( \Pr(\sum_{i=1}^{n} A_i > 0) \geq 0.95 \).

Using the Central Limit Theorem to approximate the sum by a Normal random variable, we find that \( n = 1,653 \).

We conclude from these results that using the full model together with the Kelly betting strategy there is some evidence of making profit if betting in a sufficiently large number of races. This in turn suggests that closing odds do not fully incorporate active market odds information; that is the market is not weak form efficient.

An important operational issue is the effect on closing odds of wagers based on the model predictions. With the pari-mutuel system, large wagers automatically reduce the odds at which a bet is settled. However, our model is developed for a bookmaker market and requires that bets are placed close to the start of a race. Although there may be some feedback to closing odds if a large bet is made this would not impact on returns since these are fixed in a bookmaker market at
the time the wager is made. It is possible that an individual bookmaker may refuse a bet or offer reduced odds on a very large bet. However, the odds employed in this study (provided by SIS as indicated in section 2) are the odds available to a ‘substantial wager’ at the track. In addition, in the UK there are upwards of 20 independent bookmakers at any racetrack as well as thousands of independent off-track bookmakers who accept bets at the odds on offer at racetracks at any given time. Consequently, a well organised group of individuals could split a very large wager into smaller amounts and place these simultaneously at several outlets without impacting the odds at which the bet would be settled.

We note that to make a profit in the racetrack betting market it is necessary to overcome the bookmakers’ margin, which is typically between 15% and 20%. That is, betting at random we would expect to see $E(W_i) \leq 0.85$. For our full model, using the test data set we get a 95% CI for $E(W_i)$ of $(0.9993, \infty)$, significantly higher than the 0.85 we would expect from random betting.

**3.3. Cross-validation and Jack-knifing:** We use cross-validation and jack-knifing to explore whether the profits we observe from application of the model arise by fortunate selection of the training and validation sets. In particular we use them to reinforce our estimate of $\mu$.

For the cross-validation we split the data into 12 blocks of size 100. We then consider all possible choices of 8 blocks for the training set and 4 blocks for the testing set. For each of the $\binom{12}{8} = 495$ possible choices we fit the model to the training set and then estimate $\mu$ from the testing set. Let $\hat{\mu}_j$ be the estimate obtained from the jth trial, then the cross-validation estimate of $\mu$ is $\frac{1}{495} \sum_{j=1}^{495} \hat{\mu}_j = 0.00218$. We also obtain that 80% of the cross-validation estimates are positive, which is again in close agreement with the analysis above. As the $\hat{\mu}_j$ are all highly correlated, it is not possible to obtain a confidence interval for $\mu$ directly from the cross-validation estimates.
We also applied a particular variant of cross-validation known as the jack-knife; race $i$ is set aside for testing and the remaining races are used for model fitting. This is repeated for each of the 1200 races. Let $A'_i$ be the observed wealth growth rate obtained by using the fitted model to bet on race $i$. Note that the model will change marginally each time as the training set changes slightly. Let $\mu'_i$ be the mean of $A'_i$. We can view $\mu'_i$ as a continuous non-linear function of the estimated model parameters $\beta$, which now also depend on $i$. Thus, in general, $\mu'_i \neq \mu$, where $\mu$ is the asymptotic growth rate of wealth (per race) for the parameters $\beta$ fitted using the original training set (races 1 to 800). However, maximum likelihood estimation produces consistent estimators, so our estimates of $\beta$ will converge to some limit as the size of the training set increases. Thus it is reasonable to assume that $\mu'_i \approx \mu$.

Because the models used to produce the $A'_i$ are highly dependent, the $A'_i$ are also dependent. None-the-less most of the variation in the $A'_i$ comes from the race $i$ which the model was tested on, and not the model itself. Thus the $A'_i$ will be approximately independent, which allows us to give a confidence interval for $\mu$. The jack-knife sample had mean 0.00309 and standard deviation 0.07055, giving an approximate 95% confidence interval for $\mu$ of $0.00309 \pm 0.00399 = (-0.00090, 0.00708)$. A test of the hypothesis $\mu = 0$ against the alternative $\mu > 0$ is significant at the 6.5% level.

We note that the mean of the $A'_i$ is higher than the mean of the $A_i$ (given in section 3.2), possibly due to a better estimate of $\beta$ resulting from a larger training sample. However, the variance of the $A'_i$ is higher than the variance of the $A_i$, which is probably due to the variation in the models used to calculate each $A'_i$. That is we are getting variance from the model and the test sample point. Overall the results of the jack-knife procedure provide further evidence that the
model can be exploited to earn positive returns; indicating that the market is weak form inefficient.

3.4. Parametric bootstrapping: The likelihood ratio test shows that our full model captures significantly more of the information in the price curve than that available from the closing odds alone, and the application of the Kelly betting strategy suggests that there may be enough actionable information in the model to make a profit. However there is still a question as to whether the ability to make a profit depends only on closing odds (plus some good fortune) or can be attributed to the extra information included in the model. We explore this using a parametric bootstrapping approach; which involves re-sampling from a model fitted to the observed data.

Note that if we apply the Kelly betting strategy using a model based on final track probabilities, then we never actually place any bets as the model never overcomes the bookmakers’ margin. Consequently, in this section we consider a model based on normed final track probabilities which does allow us to apply the Kelly betting strategy, and thus provides a comparison with the full model on the basis of profit earned rather than likelihood.

We norm the track probabilities \( q_i = (q_i(1), \ldots, q_i(k)) \) so that they sum to 1 and call the normed track probabilities \( \bar{q}_i = (\bar{q}_i(1), \ldots, \bar{q}_i(k)) \), where \( \bar{q}_i(j) = q_i(j) / \sum_l q_i(l) \). Under the hypothesis \( H_0 \) that the model \( p_i(j) \) depends only on closing odds, we have \( p_i(j) = \bar{q}_i(j) + \epsilon_i(j) \) where the \( \epsilon_i(j) \) are independent errors, conditional on \( \sum_j p_i(j) = 1 \), that is \( \sum_j \epsilon_i(j) = 0 \). We note that \( \log \bar{q}_i(j) \) is just a scaled version of the predictor \( c_1 \). In order to mimic the form of the conditional logit model, we re-write \( p_i(j) \) as \( \exp(\log \bar{q}_i(j) + \tilde{\epsilon}_i(j)) = \bar{q}_i(j) \exp(\tilde{\epsilon}_i(j)) \) where \( \tilde{\epsilon}_i(j) = \log(1 + \epsilon_i(j) / \bar{q}_i(j)) \) are again independent errors, conditional on \( \sum_j p_i(j) = 1 \). The \( p_i(j) \) are estimated using the full model as given in section 2.2 and \( \tilde{\epsilon}_i(j) \) is a measure of the
discrepancy between $p_i(j)$ and $\bar{q}_i(j)$. The aim of this analysis is to judge whether or not $p_i(j)$ provides more actionable information than $\bar{q}_i(j)$. That is, whether the additional information captured by the full model over closing odds alone, is needed to generate a profit. If profitability depends only on the closing odds then $\tilde{\varepsilon}_i(j)$ is effectively a random adjustment to $\bar{q}_i(j)$, which we can simulate. Consequently, to test whether $p_i(j)$ tells us more than just $\bar{q}_i(j)$ we test whether or not using the real $\tilde{\varepsilon}_i(j)$ is any better than using a randomly generated $\tilde{\varepsilon}_i(j)$.

We estimate the distribution of $\tilde{\varepsilon}_i(j)$ directly from the training data set. Figure 4 gives a histogram of $\tilde{\varepsilon}_i(j) = \log(p_i(j)/\bar{q}_i(j))$ for all horses in races 1 to 800. The sample mean and standard deviation are $-0.1012$ and $1.1865$. The Kolmogorov-Smirnov test for normality gives a $p$-value of 0.1492, indicating no significant cause not to accept this as a normal distribution (this justifies the error form used). Thus under the hypothesis $H_0$ that $p_i(j)$ depends only on closing odds, we obtain

$$p_i(j) = \frac{\bar{q}_i(j)Z_i(j)}{\sum \bar{q}_i(l)Z_i(l)}$$

for $Z_i(j)$ i.i.d. $\text{ln normal}(-0.1012, 1.1865^2)$  \hspace{1cm} (10)

Figure 4. Histogram of the log error $\tilde{\varepsilon}_i(j)$, under the hypothesis that $p_i(j)$ depends only on closing odds
We use Equation 10 to test the hypothesis that the ability of the model to make profit depends only on closing odds. Under $H_0$ our full model is just as effective as that given by Equation 10. We proceeded as follows. For each race in the test data set (races 801 to 1200) we randomly generated sample $p_i(j)$ using Equation 10. These were then used to calculate the asymptotic growth rate for wealth $\mu$, as in Section 3.2. We then repeated this 500 times to obtain an empirical distribution for $\mu$ under $H_0$. From the testing set (in section 3.2) we estimated $\mu = 0.00226$ and we now estimate that under $H_0$, $P(\mu > 0.00226) = 0.006$. That is, under $H_0$ the probability of observing a value of 0.00226 or higher (the value achieved by the full model) is less than 0.006. This suggests that the profit we observed in section 3.2 is likely to be attributable to factors in the model other than closing odds.

3.5. Features of winning bets: We explore under what circumstances the model produces successful bets in order to identify any systematic features of such bets, and, in particular, to identify what variables/combination of variables are/is needed to distinguish the horses we should bet on. To achieve this, for each race in the holdout sample we determine, for the winning horse, its probability of winning given by the model, the normed track probabilities, and whether we successfully backed the horse. We plot the results in figure 5; to spread the points we plot $\ln$ (model probability) against $\ln$ (normed track probability). Winning horses which we backed are depicted by triangles, horses we failed to back because we backed another horse in the race (thereby losing money) by crosses and in races where no horse was backed, the winner is depicted by a dot.

We see from figure 5 that we only backed horses when the model probability was greater then the normed track probability. The favourite-longshot bias means that this happens more frequently for short odds horses, and the model is better at picking short odds winners than long odds winners. In fact the results suggest that by restricting our bets to horses where $\ln$ (model
probability) > -2.5 (i.e. model probability > 0.08) profits would improve. We also note that for winning horses the differences between the normed track probabilities and the model probabilities were generally not very large, though there are a few exceptions. The relative importance to overall profit can be gauged by plotting the profit made on the z-axis. To visualize this rather than produce a 3D plot we projected the above points onto the main diagonal, and then plotted these against the profit made per pound of wealth (figure 6). From this we see that profit does not depend on occasional very large profits on long-odds horses or on bets in a particular odds range (other than suggesting that bets be restricted to where the model probability >0.08).

To gain further insight into how the model works we consider the effects of our predictor variables for all winning horses in the holdout sample (set A, 400 horses) and for winning horses we successfully bet on (set B, 61 horses). For each set we give box plots of $-2.0493*a_1^+ + 0.4371*a_1^-b_1,$ $-0.4227*b_2^-,$ $1.1678*c_1$ and $-0.0666*c_3^-$. The sum of these four gives the "winningness index" $\langle \beta, x_{i,j} \rangle$ for horse $j$ in race $i$. Box-plots provide a graphical representation of the distributions of each factor, including the interquartile range (boundaries of the larger box).
and the more extreme values of the distribution. From Figure 7 we see that there is little difference between the factors from set A to set B, other than that winning horses which were wagered upon in the sample start at slightly lower odds than winning horses in general and that there is a preference for selecting lower odds variance horses. However the differences are not large enough to conclude that there is any significant difference between the four factors from set A to set B, indicating that in general the model requires a combination of the factors for it to distinguish which horses to bet on.

Figure 7. Influences of predictor variables for winning horses.

Col. 1: \(-2.0493*a_1 + 0.4371*a_1 + b_1\); Col. 2: \(-0.4227*b^2\); Col. 3: \(1.1678*c^1\); Col. 4: \(-0.0666*c^3\).

If we look at winning horses for which the model probability is much larger than the track probability (by a factor of 1.1 or more) then there are some common characteristics. There were 14 such horses, for 13 of them the odds increased and for 7 there was a late change down. A late change down increases \(\beta, x_{i,j} >\) and thus the model probability \(p_i(j)\). However, increasing odds decrease \(\beta, x_{i,j} >\), and have a larger impact than a late change (resulting from the sizes of the coefficient and the variables). We conclude that in these 13 cases the model probability has only improved because of the relative change in \(\beta, x_{i,j} >\) for the winning horse compared to the
other horses in the race. Again this indicates that information from many sources must be carefully weighed to be able to successfully model the probability of winning.

In summary, this analysis suggests that the profits derived from the model are not dependent on a few long-odds winners and no one predictor can be used to decide when to bet; the model achieves its success by combining a variety of factors. Overall, the results of model testing suggest that the model predictors contain significantly more information than that contained in closing odds. In addition, there is something in excess of an 80% chance that a betting strategy based on the model will produce a long term increase in wealth. Whilst this is not conclusive proof that the horserace betting market is weak form inefficient to the extent that it can be exploited by an expert to his/her advantage, the results suggest that this may be the case.

4. Conclusions

In this paper we set out to explore whether the horserace betting market fully incorporates a variety of historical bookmaker price information variables, including interaction effects. We conclude that there is valuable information contained in odds and pre-closing odds movements which is not fully discounted in closing odds; suggesting that the market is weak form inefficient. Our attempts to use a betting strategy to exploit this inefficiency are also suggestive, if not conclusive, that our model could be employed to make profits. We believe this is because, as Ceci and Liker (1986) observe, expert handicapping requires the ability to combine different types of information in complex, interactive models. Our results suggest that betting market participants as a whole do not achieve this to the extent of our model, which incorporates a range of variables covering different aspects of information associated with closing odds and the movement of pre-closing odds to the closing odds, including their interactions. The results also suggest that market participants are largely effective in discounting readily discernable (more transparent) information concerning a horse’s enhanced prospects in their decisions but they do not appear to incorporate less readily discernable (more obscure) information. We conclude that
closing odds and the manner in which pre-closing odds move, are rich but subtle information sources, which bettors do not fully utilize.

The model has served our purpose in exploring the weak form efficiency of horserace betting markets. However, we have not been tempted to undertake a real-life test of the model, since this would involve considerable effort in terms of real time data capture and operation of the model and moreover, the estimated growth rate for wealth, while positive, is small and may not warrant the potential risks.

In summary, this paper adds to our knowledge of the degree to which different types of information are discounted in decisions made in betting markets. It also introduces a technique for extracting predictor variables from price curves using orthogonal polynomials and a variety of approaches for testing a model that produces probabilities. Future work exploring other financial markets, using the techniques introduced here, may yield interesting conclusions regarding market efficiency and the manner in which information is employed by market participants.

**Acknowledgements**

We are grateful to the anonymous referees, the Associate Editor and the Department Editor, Detlof von Winterfeldt for their insightful comments on an earlier draft of the paper.

**References**


