
Copyright:

This is an Author's Accepted Manuscript of an article published in Chance, 2011, copyright Taylor & Francis, available online at: http://www.tandfonline.com/doi/abs/10.1080/09332480.2011.10739885

Always use the definitive version when citing.

Further information on publisher website: http://www.tandfonline.com

Date deposited: 4th December 2013

Version of article: Author final

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License
Analyzing the placement odds of favorite horses in the thoroughbred racing industry of the British Isles.

Fernando Mata
Sarah Watts

This article describes a logistic regression model to estimate the probability of at least one of the two “favorite to win” horses, being placed in the end of the race (ending up in one of the three first positions), based on the number of horses entered and on the time of the day the race takes place (day, evening), and handicapping (handicapped race or not). The association between placement of favorites and the following variables, was also explored but not found significant: type of race, type of surface, surface conditions, and whether or not it is a stakes race.

The Thoroughbred horse racing industry is a multimillion pound global industry. Annually in the United Kingdom (UK) the sport attracts almost 6 million spectators, not only to view but to place bets on, the 9000 plus races that are run. The
turnover of the horserace betting market in the UK alone in 2006 has been reported to be in excess of £10 billion. However, the money does not only lie within the betting market itself as the racing industry supports the wider economic community not only in the UK but worldwide. The American Horse Council put the impact of the racing industry on the economy in the United States of America (USA) at $26.1 billion, which puts the industry on a par with one of the 75 largest companies in the USA. Similar importance has been stated on the impact of the industry on the Australian economy, and certainly the same kind of information could easily be found for several other different countries.

A lot of research has been done regarding the prediction of final results in thoroughbred horse racing; from predictive modeling (e.g. a E.M. White and coauthors article published in the “International Journal of Forecasting” in 1992) to pre-race behavior (e.g. a G.D. Hutson and M.J. Haskell article published in “Applied Animal Behavior Science” in 1997) and heart size (L.E. Young and coauthors, published in the “Journal of Applied Physiology” in 2005). Competition data has also widely been used, across all equine sports, in genetic evaluation and breeding value studies. Economy and betting markets (M.A. Smith, and L.V. Williams, article published in the “International Journal of Forecasting” in 2010), favorite-long shot bias and bettor preferences (L.D. Brown and coauthors article published in “CHANCE” in 1994) are examples of other topics picked up by research, to predict final results. However, anecdotal information is still, and probably will always be, widely used by bettors to justify the odds.

A simple introduction to the structure of thoroughbred racing

Horse racing can be split in two types: flat racing and national hunt racing. Both types are widespread across the more than 60 racecourses in the British Isles.
Racecourse circuits can be left- or right-handed, flat or undulating, triangular, square, oval, figure of eight shaped or even just straight. The majority of races are run on turf but there are five courses offering the alternative to run flat races on an all weather track (AWT).

A flat race is run over a predetermined distance in which the horses are not required to jump any obstacles. Total distances vary from 5-16 furlongs (1-3.2 km) and are generally split into sprints, mile and classic distances or stayers races.

National hunt (NH) racing is split into steeplechases and hurdle races. They are run over distances of 2-4.5 miles (3.2-7.2 km) and include solid obstacles at a minimum height of 4’6” (~1.37 m). Hurdle races are run over distances of 1.75-3 miles (1.9-4.8 km) and include less dense obstacles at a minimum height of 3’6” (~1.07 m). National hunt racing also includes flat races known as national hunt flat races or bumpers races designed for young horses. Hunter-chases are another classification of NH racing, designed for horses that have regularly hunted throughout the season.

In stake races part of the prize money is put up by the owners of the horse running. Competitors often race against horses of the same age, gender or class. Conditions races may also involve all horses carrying the same weight or with weights being adjusted to reflect age, races of a certain value or number of races won. In a handicap race all horses are allocated weights to be carried based on their previous performance. The principle of handicapping is to allow every horse an equal chance of winning. In other words the horse deemed to have the best chance in a race will carry the most weight and the least fancied runner will carry the least weight.
**Previous research**

Rowe in his book from 2004, “how to win in horseracing” discusses 2 studies with 2,317 and 10,466 races, where the favorite won 33% and 32.6% of the times respectively, and got placed around 50% of the times. Rowe’s data comes from the USA where flat races on dirt tracks are common. In contrast, UK races are more likely to be on turf, and there is mixture of flat races and races that include jumps.

Duncan’s 2005 book “Winning horse racing formulae”, provides a British perspective, with the analysis broken down and presented according to race type, handicap and age range. When considering the figures given by Duncan for just flat racing, the average percentage for one of the top three winning is 74%. Rowe calculated a top three win percentage of 69%.

Both Rowe and Duncan discuss studies carried out into finishing positions in Thoroughbred horseracing but neither considers in depth the effects of variables such as type of race, age and surface.

G.D. Hutson, and M.J. Haskell, conducted a study into pre-race behavior and its importance in allowing the prediction of race winners in their 1997 article in “*Applied Animal Behavior Science*”.

They found that no single behavior or appearance could improve upon the predictions made by the betting market. By considering multiple behavior variables, however, horses deemed unable to win could be eliminated, and therefore their results were considered of potential high economic value.

A number of other studies have used predictive modeling and forecast combination in order to predict future racing outcomes. S. Lessmann and coauthors in their “*European Journal of Operational Research*” article (2009) use a complex two stage model: First a machine learning technique (support vector machines for classification), estimates the likelihood of a given runner being a winner. Then a conditional logit (CL) model, estimates the winning probability of one horse in conjunction with the other competitors. This is a 2 stage approach based on W. Benter
(Efficiency of Racetrack Betting Markets, 1994), where support vector regression (SVR) was used to model the relationship between variables and horses’ finish positions, being then combined with horses final market prices (odds) using a CL model in a second step. These 2 steps were found to be complementary to each other, once CL accounts for within race competition, whereas SVR uses a large number of variables and automatically models non-linear relationships between these.

E.M. White and coauthors test the accuracy of a forecast combination of judgmental and statistical methods, to predict winning outcomes in horse races. These authors caution that “horse racing possesses a strong element of random chance which makes it impossible to predict every winner with perfect accuracy”.

The bulk of current research into surface and racetrack design is aimed at minimizing the risk of injury to the locomotory system of the horse. Such information is of high economic value if it can extend the racing careers of the horses. Studies into performance and environmental factors are contradictory, such as that by M.D.S. Mota and coauthors in 1998 ("Journal of Animal Breeding Genetics"), reporting varying effects of surface type. It has been suggested that turf tracks are much faster than sand tracks, with light turf producing the best results. The same study also found that drenched sand was faster than smooth, dry sand, but this contradicts results found by R.L. Hintz, and L.D. Van Vleck (published in 1978 in the "Journal of Animal Science") who suggested dry sand to produce faster times due to less resistance.

Most of the race type studies have also been directed at injury risk and reduction. An example of an extensive study into this area was carried out by J.R. Newton and coauthors (published in 2005 in the "Equine Veterinary Journal") who found that the risk of epistaxis increased with an increasing racing effort. This meant that the incidence of blood visible at the nostrils increased between flat and hurdle racers and again between hurdlers and chasers.
Method and Results

To estimate the probability of a favored horse placing, we considered data from The Racing Post for all thoroughbred races in the British Isles between June 1 and October 31, 2008. The date, racecourse location, surface, period of the day the race took place, ground conditions, number of runners, type of race, handicapped or not and the finishing positions of the favorite and second favorite were noted.

Previous research did not use the technique used in this study. We used a logistic regression where the outcome dichotomy was considered to be: at least one of the favorites placed or none of the favorites placed. The predictors (independent variables) used were: the ground condition or “going” (heavy, soft to heavy, soft, good to soft, good, good to firm, firm and hard); the period of the day the race took place (day, evening); handicapped race or not; type of race (flat, national hunt flat, chase, hurdles, stakes); track surface (turf, AWT); and number of runners. All the variables were used as factors, with the exception of the number of runners, which was used as a covariate. A runner is considered to be placed if it ends up in one of the three first positions in the race.

As the “going” for AWT is classed differently (standard, standard to fast and standard to slow) from the turf, the AWT races were initially not considered in the model, so that the “going” condition could be included. Once the “going” was found to be non significant, AWT races were included in the study and the “going” excluded.

The type of race was simplified as “flat” (flat, NH flat and stakes) or “jump” (hurdles and chase), and stakes was also isolated as a factor.
In total, 4410 races took place during that period, involving the participation of 47847 horses. From these, races where the favorite horses had not finished or started were discounted. Also, races where favorites shared placements were not considered to avoid eventual bias.

We analyzed 3604 racing starts in the five month period. Of these 2918 (81%) were flat group races (flat, NH flat and stakes) and 686 (19%) were NH group races (hurdles and chase). Of the flat group races 542 (19%) were run on AWT and 2376 (81%) were run on turf. All 686 NH races were run on turf, making a total of 3062 runs on turf and 542 on AWT. NH flat or bumpers races accounted for less than 2% of the total races and therefore were included as flat races, as it did not seem appropriate to include them in the jump category. The favorite won 1188 races which equates to 33%, was placed in 1147 races or 32% (overall placement 65%), and was not placed in 1269 (35%). The second favorite won 713 races or 20%, was placed in 1281 races which equals 36% (overall placement 55%), and was not placed in 1610 races (45%). The results are summarized in table 1.

<table>
<thead>
<tr>
<th></th>
<th>1st favorite</th>
<th>2nd favorite</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>%</td>
</tr>
<tr>
<td>Overall placement</td>
<td>2335</td>
<td>65</td>
</tr>
<tr>
<td>Placed won</td>
<td>1188</td>
<td>33</td>
</tr>
<tr>
<td>Placed not won</td>
<td>1147</td>
<td>32</td>
</tr>
<tr>
<td>Not placed</td>
<td>1269</td>
<td>35</td>
</tr>
</tbody>
</table>

Table 1- Number of races taking place and analyzed in the period of analysis, and placements of first and second favorite horses.
A backwards stepwise approach was used to fit the model and several non-significant \((p>0.05)\) variables were excluded (type of race, stakes, and track surface). In the end, two factors (period of the day and handicapping) and the covariate (number of runners) were found to be significant \((p<0.001, p<0.001 \text{ and } p<0.05\) respectively). To finalize, the interactions between the factors were added to the model but were found to be non significant \((p>0.05)\). The reduced adjusted model parameters are summarized in table 2.

Table 2 - Logistic regression model for predicting the placement or not, of favorite horses. Day period has a positive coefficient for evening races, and these have higher odds for favorite horses to be place; Handicap has a negative coefficient for handicapped races, and these have lower odds for favorite horses to be place; Runners has a negative coefficient, and the number of runners correlate negatively with the odds for favorite horses to be placed.

<table>
<thead>
<tr>
<th>Variables in the equation</th>
<th>(\beta)</th>
<th>SE ((\beta))</th>
<th>(p)-value</th>
<th>95% CI ((\beta))</th>
<th>OR ((e^\beta))</th>
<th>95% CI OR ((e^\beta))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.068</td>
<td>0.1937</td>
<td>&lt;0.001</td>
<td>3.688 – 4.447</td>
<td>58.414</td>
<td>39.965 – 85.381</td>
</tr>
<tr>
<td>Runners</td>
<td>-0.152</td>
<td>0.0141</td>
<td>&lt;0.001</td>
<td>-0.180 – -0.125</td>
<td>0.859</td>
<td>0.835 – 0.883</td>
</tr>
<tr>
<td>Handicap</td>
<td>-0.951</td>
<td>0.1073</td>
<td>&lt;0.001</td>
<td>-1.161 – -0.740</td>
<td>0.386</td>
<td>0.313 – 0.477</td>
</tr>
<tr>
<td>Day period</td>
<td>0.231</td>
<td>0.1113</td>
<td>&lt;0.05</td>
<td>0.013 – 0.449</td>
<td>1.260</td>
<td>1.013 – 1.567</td>
</tr>
</tbody>
</table>

SE: standard error; CI: confidence interval; OR: odds ratio. The adjusted final model has a deviance of 274 and a Akaike's information criterion \((\text{AIC})\) of 638.

Analysis & Conclusion

The results from the current study show that there are certain situations in thoroughbred racing where the winning result anticipated by the allocation of a favorite status is more or less likely to be achieved. In fact, 3 of the variables investigated
(number of runners, handicap and period of the day) appear to help predict whether or not one of the favorites will end up placing.

Several authors refer to the logic of the odds of a favorite placing decreasing with the number of horses running. Fewer runners mean less competition and less potential for interference. The importance of interference is supported by G.S. Martin and coauthors in 1996 in the “Journal of the American Veterinary Association”) who noted that the winning time tends to increase by about 0.23 seconds for each additional runner.

Another logical result is that the favorites are less likely to place in handicap races than in non-handicap races. In handicapped races, more weight is added to better performers, in an effort to give all of the runners about the same chance of winning. In fact, handicapping increases the competitive capacity of less skilled pairs (horse/rider) by the addition of weights to be carried by the favorites, to equalize winning odds. The favorite is, thus, less likely to win in handicapped races, in comparison to non handicapped races.

Perhaps the most interesting result is that the period of the day appeared to be an important factor. After controlling for the number of runners and type of the race, favored horses were more likely to place in evening races than in daytime races. This has not been mentioned before in the literature, and we do not have an intuitive explanation for this effect, and therefore this is a subject in need of further investigation.

Figure 1 displays the estimated probability of the number of runners, for each of the four combinations of the two factors (day period and handicapping). The results are also summarized in table 3. The probability of the favorite placing decreases with the number of runners, with the odds ratio decreasing by about 14.1% on average for each
additional horse. The decline is faster in handicapped races than in non-handicapped races. And for both handicapped and non-handicapped races, the decline is sharper for day races than for evening races.

**Figure 1** – Logistic curves graphing the probability of placement for a favorite horse, in dependency of the number of runners, period of the day and handicapping. Cut off point refers to the point where the probability equals 0.5. In the original data races had between 3 and 29 runners with an average of 11.

<table>
<thead>
<tr>
<th>Table 3 - Numbers of runners corresponding to the given probabilities of placement of a favorite horse, based on the model fitted.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Probability</strong></td>
</tr>
<tr>
<td>0.9</td>
</tr>
<tr>
<td>0.8</td>
</tr>
<tr>
<td>0.7</td>
</tr>
<tr>
<td>0.6</td>
</tr>
<tr>
<td>0.5</td>
</tr>
</tbody>
</table>
Despite trying to analyze a wide number of variables that may affect performance, there were still many influences not measured and therefore not incorporated in the model, which may affect the results. Firstly the race track itself is an uncontrolled environment and performance studies are difficult to accomplish because of the hidden influences that arise. Other variables that may impact upon performance include rider influences such as weight, riding style and use of whip, poor start, wide positioning through a turn and positioning behind a slower horse in the straight. The horse itself is also a never ending source of possible variables to be considered in performance.

The uncountable number of different variables to be taken into consideration in the discussion of horse performance in races is so large that makes it difficult to isolate them for a convenient analysis. Also the eventual interaction between these would obviously result in a biased conclusion, if not considered. Therefore it seems logical to observe that the non inclusion of hidden variables in horse races outcome prediction models leads to an increase in unpredictability and obviously to a decrease in the correct prediction of winning places.

As a final thought, racing possesses a strong element of random chance which makes it impossible to predict every winner with perfect accuracy. Even so, if you have a desire to gamble and want to take the chance of betting on a horse, place your money on a favorite, not handicapped and running in an evening race with a small number of runners. The question now is: do the bookmakers have this information? If so, the payoff will reflect this, otherwise it won’t, and you have here an opportunity to make some easy
money. It’s up to you to find out and decide where to put your money, but please don’t blame us if things go wrong…!

**Further Reading**


