ONLINE SPORTS GAMBLING: A LOOK INTO THE EFFICIENCY OF BOOKMAKERS' ODDS AS FORECASTS IN THE CASE OF ENGLISH PREMIER LEAGUE

By: Jasmine Siwei Xu

Thesis Advisor: Professor Roger Craine

Undergraduate Economics Honor Thesis
University of California, Berkeley
May 2011

Acknowledgement: I would like to thank Professor Roger Craine for his advice and guidance in completing this Undergraduate Honors Thesis. I would also like to thank Mr. James Church for assisting me in finding data.

Abstract

This paper aims to examine the efficiency of using bookmakers' odds as forecasts of soccer match outcomes of the English Premier League. The result of analysis shows that bookmakers' odds for Premier League Season 2006-2007 are effective forecasts of soccer match outcomes, but could be improved by incorporating the effect of the number of yellow cards the home team receives in its last match. However, although yellow cards is proven to be statistically significant in Season 2006-2007, its economical significance remains uncertain in the following season.

1. Introduction

Sports betting attracts attention of casual bettors, professionals, and even academic researchers. Wagering in betting markets resembles trading in financial markets in many ways. The betting market consists of bookmakers, who offer certain odds for the outcomes of uncertain future events, and the bettors have to right to decide which outcome to bet on, or whether to bet on those events at all. The bookmakers have a single goal of making a profit out of the odds they offer. This goal ensures the bookmakers to make the odds high enough to be competitive and attractive to bettors, but not so high that they become unprofitable. Hence, the odds offered by bookmakers can be viewed as their probabilistic assessments, or forecasts, of an event's outcomes.

The English Premier League is chosen as the subject of this paper not only because it is one of the most prominent and popular soccer tournaments of the world, but also because it is held annually, which makes its data collectability easier than other comparable soccer competitions such as FIFA world cup, which is held once every four years. In this paper, I try to exam how effective are the odds offered by bookmakers as forecasts for match outcomes in English Premier League.

E. Strumbelj and M. Robnik Sikonja (2009) also strived to answer this question in their work, *Online bookmakers' odds as forecasts: The case of European soccer leagues*. Strumbelj and Sikonja translated bookmakers' odds into Brier score and ranked probability score (RPS) to evaluate the effectiveness of forecasts for a soccer match. They found that odds from some bookmakers are better forecasts than those of other and showed that the effectiveness of bookmakers' odds as forecasts has increased over time.

In this paper, I took a different approach to answer this question. Instead of translating odds into scores to determine their efficiency, I adopt a binary probit model with the real outcome of matches in one season as the dependent variable and bookmakers' odds for the matches of the same season and other variables that I believe may impact the outcome of a soccer match as independent variables.

2. Summary of Results

The result of the probit regression shows that bookmakers' odds are effective as forecasts of English Primer League Season 2006-2007. However, one other variable, the number of yellow cards the home team receives in its last match, is also statistically significant, although its impact on the forecasts is small in comparison to bookmakers' odds. Despite incorporating the additional variable improves the efficiency of forecasts in Season 2006-2007, the model fails to be a profitable betting strategy for the following season—Season 2007-2008.

3. Model

3.1 The Weak Form of Efficiency Market Hypothesis

The goal of this paper is to determine whether online bookmakers' odds are efficient predictors of the real result of matches in the case of English Premier League. Hence, the underlying task is similar to testing for the Weak-form of Efficient Market Hypothesis. The Efficient Market Hypothesis was developed by Professor Eugene Fama of University of Chicago Booth School of Business in the 1960s. There are three versions of the Efficient Market Hyp0thesis (hereafter EMH): the Weak Form, the Semi-Strong Form, and the Strong Form. The Weak Form of EMH specifically states that the information set includes only the history of

prices or returns, which means that the past history of prices or returns of a stock is the only efficient explanatory variable to explain the price of a stock. Putting in context of this paper, the Weak-Form of EMH would suggest that bettors cannot produce a more efficient prediction of the outcome of soccer matches than what is provided by the online bookmakers and that the bookmakers have incorporated all relevant factors that could influence the result of a match into their probabilistic assessments.

3.2 Probit Model

Linear probability model is often used to test EMH for its obvious advantage of being simple to estimate and use. However, a linear probability model would be problematic to use in this paper because the fitted probability can be less than zero or greater than one. However, it is obvious that no team can win or lose with a probability outside of the range 0 to 1. This limitation of linear probability model can be overcome by using a binary response model, which could explain the effect of bookmakers' odds and the effects of the X_j on the response probability $\Pr(Y = 1 \mid X_j)$.

$$\Pr(Y = 1 \mid X_j) = \beta_0+\beta_1 Pred+\alpha_1 X_1+ \alpha_2 X_2+...+ \alpha_j X_j$$

Y is the probability of a team winning a English League Premier Soccer match (a team wins a match when Y=1), Pred is online bookmakers' probabilistic forecasts, which are converted from their odds, that a team would win a match, and X_j is a vector of the other explanatory variables that could contribute to a team's probability in winning a match.

I ran a probit regression on the final results of matches in English Premier League Season 2006-2007 against online bookmakers' predictions and selected variables which I believe could

have an effect on the actual result of the game. If the Weak-Form of EMH holds true in this case, that is, the online bookmakers' predictions are very efficient forecast of games and have incorporated all variables I selected, only β_1, the coefficient for bookmakers' probabilistic assessments, should be statistically significant and none of the other coefficients should be statistically significant.

4. Data

4.1 Data Description

The data used in this research paper consist of the 380 games played by the 20 participating clubs during Season 2006-2007 of English Premier League. The data cover match odds from 9 online bookmakers: B365 (b365), Blue Square (bs), Bet & Win (bw), Gamebookers (gb), Interwetten (iw), Ladbrokes (lb), Sportingbet (sb), Stan James (sj), Stanleybet (sy), VC Bet (vc), and William Hill (wh).

It is important to note that the odds used in this paper are published bookmaker odds, that is, odds offered by bookmakers to the public. These odds are a combination of the bookmakers' estimated probabilities, which is unknown to the public, and their adjustments to public expectations and inside trading. Thus, the odds may change when they are first published to the start of the match. However, such changes are usually small and mostly occur as the match nears. In fact, most bets are made on match-day. Hence it is reasonable to assume that the odds collected one or two days before the match is similar to the initially published odds. Since the odds of the 9 above bookmakers used in this paper are collected on Friday for weekend matches and Tuesdays for midweek games, they should be close enough to the initially published odds to be good representations for bookmakers' forecasts.

The data also includes 23 statistics variables[1] for the 20 participating teams. They are: accumulated points for home team [2](ahp), accumulated points for away team (awp), home team shots in last match (hs), accumulated home team shots (ahs), home team shots on targets in last match (hst), accumulated home team shots on targets (ahst), home team fouls in last match (hf), accumulated home team fouls (ahf), home team yellow cards in last match (hy), accumulated home team yellow cards (ahy), home team red cards in last match (hr), accumulated home tea red cards (ahr), away team shots in last match (as), accumulated away team shots (aas), away team shots on targets in last match (ast), accumulated away team shots on target (aast), away team fouls in last match (af), accumulated away team fouls (aaf), away team yellow cards in last match (af), accumulated away team yellow cards (aaf), away team red cards in last match (ar), accumulated away team red cards (tar), and the final result of the match (d_win[3]). All accumulated match statistics only include matches previously played. From this point on, abbreviations will be used when referring to individual explanatory variables.

4.2 Bookmakers' Odds as Forecasts of Games

A regular soccer match has three possible outcomes, either the home or the away team wins, or the game ends with a draw. Hence, each online bookmaker has three odds for each match played: odds for home team winning, odds for home team losing, and odds for draw. For instance, Bet365 has odds for the match played on 8/16/2006 between Arsenal, the home team, and Aston Villa, the away team, as follows: Bet365 home win odds =1.28, Bet365 draw odds =4.5, and Bet 365 away win odds =13. Therefore, if investors bet one dollar on Arsenal winning the match, they will receive a payoff of $1.28 if Arsenal indeed wins the match, or $4.5 and $13

[1] See Table 1 for a summary of the statistics variables
[2] Team receives 3 points for a win, 1 point for a draw, and no point for a loss.
[3] Win=1, draw/lose=0

if the respective outcome is the final result of the match. Furthermore, the fact that odds for betting on Aston Villa, the away team, is higher than that of Arsenal, the home team, implies that Arsenal is favored,

Hence, bookmakers' odds can be viewed as probabilistic forecasts of the match outcomes. E. ˇ Strumbelj and M. Robnik ˇ Sikonja's (2009) method of converting odds to probabilities is adopted in this paper. The odd, 1.28, means that the probability of Arsenal winning is 1/1.28=0.78. Likewise, the probability of Aston Villa winning the match is 1/13=0.08, and the probability that the match ends in a draw is 1/ 4.5=0.22. However, the sum of those three probabilities, 0.78+0.08+0.22=1.08, exceeds 1. The extra 8% is known as the bookmaker margin, the money bookmaker makes regardless the result of the game.

In this paper, however, all probabilities used are normalized. That is, the odds-implied probabilities of home team winning, away team winning, and draw sum up to 1. The margin is eliminated by dividing each probability by 1 plus the margin. Hence, the Bet 365's normalized probability of the home team—Arensal— winning is 0.78/1.08=0.72, the normalized probability for Aston Villa, the away team, winning is 0.07, and the normalized probability for a drawing match is 0.20. From this point on, individual bookmaker's normalized probabilities of the three outcomes of a soccer match will be shown as the bookmaker's abbreviation plus h, d, and a, which stand for home team winning, draw, and away team winning respectively (b365h=0.72, b365d=0.20, b365a=0.07).

Since a binary probit model is used in this paper and the primary interest is to explain the effect of all the explanatory variables mentioned above have on a team's probability of winning an English Premier League soccer match, I chose to focus on the probability of home

team winning, and hence group the probabilities of away team winning and the game resulting in a draw into one probability of the home team *not* winning.

5. Estimates and Results

5.1 Correlation among 9 Online Bookmakers

I first calculated the correlations among the 9 online bookmakers' normalized probabilistic assessments of home team winning and found that the 9 online bookmakers are almost perfect positively correlated, with correlations all greater than 0.99 (see table 2 for correlation matrix). The high correlations among the 9 bookmakers make intuitive sense, because if bookmakers' odds differ by more than the bookmakers' margin, arbitrage opportunities exist. A bettor can make profit by simply betting money on home team winning with one online bookmaker and betting money on the odds of the other two outcomes with another online bookmaker whose odds differ by more than the margin. Since the 9 online bookmakers' probabilistic forecasts for home team winning the match are almost perfectly correlated, I chose Bet365's normalized probabilistic forecasts for home team winning as Pred, and the resulting estimates should be representative of the other 8 online bookmakers.

5.2 Correlation among the explanatory variables

Too many explanatory variables could negatively impact the efficiency of estimates. Hence, I also calculated the correlation among the 23 statistics variables that I believe could affect a team's probability of winning a match, and eliminated those whose correlation is greater than 0.9 (see table 3 for complete correlation matrix for the 23 statistics variables). After eliminating explanatory variables that are highly correlated, the remaining variables are ahp, awp,

hs, ahs, hst, hf, ahf, hy, hr, ahr, as, ast, af, ay, ar, and tar. Table 4 displays the correlation matrix for the remaining variables.

5.3 Estimates

I ran probit regression with d_win on b365h, ahp, awp, hs, ahs, hst, hf, ahf, hy, hr, ahr, as, ast, af, ay, ar, and tar.

$$\Pr(Y = 1 \mid X_j) = \beta_0 + \beta_1 b365h + \alpha_1 ahp + \alpha_2 awp + \alpha_3 hs + \alpha_4 ahs + \alpha_5 hst + \alpha_6 hf + \alpha_7 ahf + \alpha_8 hy + \alpha_9 hr + \alpha_{10} ahr + \alpha_{11} as + \alpha_{12} ast + \alpha_{13} af + \alpha_{14} ay + \alpha_{15} ar + \alpha_{16} tar \quad (1.1)$$

$$H_0: \beta_1 \neq \beta_0 = \alpha_1 = \alpha_2 = \ldots = \alpha_{16} = 0$$

$$H_1: \text{not } H_0$$

The estimated coefficients and corresponding p-values of model (1.1) are presented in table 5. Among the 18 estimated coefficients on the 17 explanatory variables and 1 constant, 3 are statistically significant at 90% confidence level. Those three variables are b365h, ahs, and hy, and their corresponding coefficients are 1.93, 0.00326, and -0.1157176. I thus reject the null hypothesis. The result of matches of English Premier League Season 06-07 cannot be wholly explained by the online bookmaker, Bet 365's odds. According to the result of model (1.1), the accumulated home team shots and the number of yellow cards the home team received in last match also meaningfully contribute to the probability of the home team winning a match.

I then eliminated the statistically insignificant variables and ran the binary probit regression with only the statistically significant variables resulted in model (1.1). The model thus became:

$$\Pr(Y = 1 \mid X_j) = \beta_0 + \beta_1 b365h + \alpha_1 ahs + \alpha_2 hy \quad (1.2)$$

$$H_0: \beta_1 \neq \beta_0 = \alpha_1 = \alpha_2 = 0$$

$$H_1: \text{not } H_0$$

The estimated coefficients and corresponding p-values of model (1.2) are presented in table 6. The estimated coefficients for b365h, hy, and the constant, β_0, are statistically significant, while the estimated coefficient for ahs, which has a p-value of 0.21, is no longer statistically significant. I again reject the null hypothesis that bookmaker's odds alone efficiently explains the probability of the home team winning a match since hy, the number of yellow cards home team receives in last match, affects the team's probability of winning the next match.

Since ahs is no longer statistically significant, I eliminated it from the regression and obtain a new model:

$$\Pr(Y = 1 \mid X_j) = \beta_0 + \beta_1 b365h + \alpha_1 hy \quad (1.3)$$

$$H_0: \beta_1 \neq \beta_0 = \alpha_1 = 0$$

$$H_1: \text{not } H_0$$

The estimated coefficients and corresponding p-values of model (1.3) are presented in table 7. The two explanatory variables and constant remain statistically significant in this model. We thus obtain the final model,

$$\hat{Z} = -1.045328 + 2.619339\ b365h + -.105469\ hy \quad (1.4)$$

The estimated probability that the binary dependent variable Y takes on the value one is

$$\Pr(Y = 1) = F(\hat{Z}),$$

where F is the cumulative standard normal distribution function.

5.4 Interpreting the Results

Table 1 shows that on average, the home team receives 1.70 yellow cards per match, with a standard deviation of 1.30 yellow cards, and that b365h has a mean of 0.45 with a standard deviation of 0.17.

The change in probability of the home team winning a match when Bet365 forecasts the home team's probability to win the match is 1% higher, given that the team receives 1.70 yellow cards from its previous match, could be calculated by finding the difference in area that takes $\hat{Z}$ or less. For example, the change in probability that a home team actually wins a game when b365h is increased from 0.44 to 0.45 could be calculated as follow:

$$\hat{Z} = -1.045328 + 2.619339\ b365h + -.105469 * 1.70$$

$$Z = -1.045328+2.619339\ (0.44) + -.105469 * 1.70 = -0.072$$

$$Z = -1.045328+2.619339\ (0.45) + -.105469 *1.70 = -0.045$$

The probability that a standard normal distribution takes on a value of -0.072 or less is 47.13% and the probability that a standard normal distribution takes on a value of -0.045 or less is 52.87%. The model predicts that when Bet365's prediction of the probability that home team

wins a match rises from 44% to 45%, the probability that the home team actually wins the match rises by 5.74%.

Similarly, the change in probability that the home team would win a match if it receives different number of yellow card in the previous game, given that Bet365 predicts it has a probability of 0.45 to win the match, could also be calculated by finding the difference in area that takes $\hat{Z}$ or less. For instance, the change in probability that a home team wins a game when it receives 1yellow card and 2 yellow cards during its previous match is calculated as follow:

$$Z = -1.045328+2.619339\ (0.45) + -.105469 * 1= 0.028$$

$$Z = -1.045328+2.619339\ (0.45) + -.105469 *2 = -0.078$$

The probability that a standard normal distribution takes on a value of 0.028or less is 51.12% and the probability that a standard normal distribution takes on a value of -0.078or less is 46.89%. The model predicts if a team receives 2 yellow cards instead of 1 in its previous game, the probability that the team would in its next home game is lowered by 4.23%.

6. Implementation of the Model

The fact that the coefficient of hy is statistically significant implies that the probability of a team winning a match forecasted by online bookmakers could be improved. Model (1.4), incorporating the effect of hy on a team's likelihood to win a match, should produce more efficient forecasts of the result of the match than those offered by online bookmakers.

6.1 Betting on All vs. Bet when Forecasts of Model (1.4) is Greater than B365h

I first computed the rate of return of a passive betting strategy, that is, bet $1 on home teaming winning for all 380 matches of Season 2007-2008. If the home team wins, I would get a payoff that equals the odds offered by B365 for home team winning. If the match results in either a draw or away team winning, I would get a payoff of $0. This strategy serves as a bench mark which I shall later compare the other non-passive betting strategies with as a way to test the economic efficiency of model (1.4). The sum of payoffs or all 380 games is $350.4. The total investment is $380, since the strategy bets on every match. Hence, the rate of return for the passive betting strategy is ($350.4-$380)/$380=-7.79%. This net loss equals the bookmakers' margin, the average profit a bookmaker make if he matches the potential payoff to home team winning against the potential payoff to lose or draw.

I then implemented model (1.4) on season 2007-2008. I first computed the probability of home team winning the match with model (1.4) for the 380 matches of Season 2007-2008. I then compare the probabilities produced by model (1.4) with those offered by online bookmakers, represented by Bet 365, of the same season (probabilities of online bookmakers are converted from their odds in the same manner as the process described in section 4.2: Bookmakers' odds as forecasts of game). I employed a betting strategy where I invest one dollar on the home team winning the match only if the probability resulted from model (1.4) is higher than that forecasted by Bet365 and I do not bet if the probability produced by model 1.4 is lower than Bet365's forecast. I would generate a payoff that equals to Bet365's odds on home team winning if the home team actually wins the match and $0 if the match results in a draw or the away team wins.

For example, the probability resulted from implementing model (1.4) for the match between Birmingham and West Hampshire on 8/18/2007 is

$$\hat{Z}= -1.045328 + 2.619339 * 0.391387174 + -.105469*1= -0.1256213$$

0.391387174 is the probability of Birmingham, the home team, winning the match and Birmingham received 1 yellow card in its previous match. The probability that a standard normal distribution takes on a value of -0.1256213 or less is 45%, which is higher than that predicted by Bet363—39%. I thus bet one dollar on the home team. However the result of the match shows that West Hampshire is winner. Consequently I earn a payoff of $0.

In another case, the probability of home teaming winning obtained from model (1.4) for the match between Derby and Newcastle on 9/17/2007 is

$$\hat{Z}= -1.045328 + 2.619339 * 0.2168607+ -.105469*1= -0.58276531$$

0.2168607 is the probability of Derby winning forecasted by Bet 365 and Derby received 1 yellow card in its last match. The probability that a standard normal distribution takes on a value of -0.58276531or less is 28%, which is also higher than that of Bet365—22%. I thus bet one dollar on the home team. In this match, Derby, the home team, won, I hence generate a return of 4.33 dollars, which equals the odds Bet365 offered for Derby winning the game.

I calculated my returns all 380 matches played in Season 2007-2008 and summed them to generate the total return. By using model (1.4) and implementing betting strategy as described above, I would win a total of $264.64. Out of 380 matches played in season 07-08, the probability produced by model 1.4 is higher than that produced by Bet365 in 292 matches. Hence, I invest a total of $292 in the whole season. This means that I have a negative rate of return of ($264,64-292)/292=-9.37%.

The result of this betting strategy shows that although incorporating hy would increase the efficiency of forecast of the result of game in season 06-07 of English Primer League, the resulting model may not be particularly helpful in forecasting the result of other seasons, season 07-08 specifically.

6.3 Trimmed the Tails

Empirical studies of betting, especially horse race betting, consistently find evidence for the favorite-longshot bias, an observed phenomenon where on average, bettors tend to overvalue "long shots" and undervalue favorites. Many researchers conclude this anomaly exists in gambling because bettors are risk lovers (Weitzman 1965; Quandt 1986; Kanto, Rosenqvist, and Suvas 1992). However, researchers have also provided counterexamples for this betting anomaly, as well as evidence that bettors are risk-averse (Busche and Hall 1988; Golec and Tamarkin 1998). In the second betting strategy employed, I eliminated both risk-averse and risk-loving betting behaviors, that is, I bet on home team winning only when the probability generated by model (1.4) is greater than b365h, and is in between 0.333 and 0.667. Given that there are only three possible outcomes in a soccer match, by only betting on games where the probability of home team winning is between 0.333 and 0.667, games where home team is too unlikely or likely to win are eliminated. Under this strategy, I earn a payoff equal to the odds if a bet is made and the home team wins the match, and a payoff of $0 in all other situations. The total payoff resulted from this betting strategy is $176.49 and a total of 176 games qualify for a bet under this strategy. Thus, the rate of return is (176.49-176)/176=0.28%.

However, when I use the same "trim the tails" strategy on the b365h, that is, I bet on the home team winning only when b365h is greater than 0.333 and smaller than 0.667, I obtain a total payoff of $235.19 while the total investment is $230. Hence the rate of return for using

"trim the tail" on Bet 365 is (235.19-230)/230= 2.26%. This rate of return is, again, higher than the rate of return when using probabilities produced using model (1.4).

7. Problems and Possible Improvements

In my model, only the statistics variables of the current season of the participating clubs are included. However, game statistics of the previous seasons could as well be good indicators of how well a team would do in the current season, and subsequently impact its probability of winning a match. I did not include the statistics of previous seasons because the participating teams differ each season. Out of the 20 participants of Premier League Season 2006-2007, Reading, Sheffield United, and Watford did not participate in Premier League Season 2005-2006 and , therefore do not have team statistics for the past season. However, it is insensible to input 0 for variables such as goal in last season for those three teams. And omitting those data for teams w did not participate in the previous season would only decrease the number of observations, which would subsequently decrease the efficiency of the estimated coefficients.

The fact that participating clubs change each season may also explain why model (1.4) is not economically efficient for Season 2007-2008. Out of the 20 clubs who played in Season 2007-2008, Sunderland, Birmingham City, and Derby County did not participate in Season 2006-2007, and therefore their data was not included in the data set used to derive model (1.4).

I would also like to include more economic variables such as city income level and team payroll. However, finding good data on such economic variables has been unsuccessful. Furthermore, most economical variables have very limited variations within a short period of time. For example, variables such as city income level may only vary from year to year, not game by game in the season and, therefore may not be a powerful explanatory variable for a

team's probability in winning matches within the same season. Nonetheless, I believe more variables from different perspectives could have interesting contributions to the performance of participating clubs.

8. Conclusion

In this paper I investigated the use of online bookmakers' odds as probabilistic forecasts of soccer match outcomes of the English Premier League. Base on the results of the probit regression model, I conclude that online bookmakers' odds are efficient forecasts of soccer match outcomes, but even greater efficiency could be achieved by incorporating the impact exudes by the number of yellow cards the home team receives from its last game. However, under the two betting strategy examined, the same model fails to maintain its effectiveness in forecasting the result of soccer match outcomes in the following season. I reason that this inefficiency may be caused by the inconsistency in the participants of the English Primer League. As part of future work, it would be worth investigating if it is more likely to create an empirical model that can efficiently forecast multiple seasons by including data across seasons and more economical variables.

Table 1: Summary of explanatory variables

Variable	Obs	Mean	Std. Dev.	Min	Max
ahp	380	24.48158	17.9004	0	84
awp	380	25.12632	18.35216	0	91
hs	380	9.931579	4.696662	0	26
ahs	380	206.7711	133.7321	0	599
hst	380	4.976316	2.956495	0	15
ahst	380	104.4684	69.15431	0	322
hf	380	12.45263	4.547071	0	26
ahf	380	234.8553	140.9965	0	549
hy	380	1.697368	1.299683	0	6
ahy	380	30.21316	19.09535	0	82
hr	380	0.097368	0.305609	0	2
ahr	380	1.489474	1.332522	0	7
as	380	11.97368	5.467519	0	32
aas	380	207.4316	134.2746	0	594
ast	380	6.178947	3.344681	0	19
aast	380	105.0395	69.57736	0	322
af	380	11.96053	4.479919	0	25
aaf	380	233.0447	141.4986	0	533
ay	380	1.455263	1.233053	0	6
aay	380	29.9	19.2598	0	84
ar	380	0.036842	0.188622	0	1
tar	380	1.439474	1.339089	0	6
d_win	380	0.478947	0.500215	0	1
b365h	380	0.445049	0.167756	0.077535	0.827338

Table 2: Correlation Matrix for 9 online bookmakers' odds

	b365h	bwh	gbh	iwh	lbh	sbh	whh	sjh	vch
b365h	1								
bwh	0.9947	1							
gbh	0.9977	0.9971	1						
iwh	0.9936	0.9934	0.9955	1					
lbh	0.9909	0.9901	0.9925	0.9907	1				
sbh	0.9968	0.9964	0.9981	0.9942	0.9924	1			
whh	0.9929	0.9929	0.9954	0.9934	0.9929	0.9945	1		
sjh	0.9934	0.9939	0.9952	0.9926	0.9911	0.9947	0.9941	1	
vch	0.9959	0.9949	0.9972	0.9936	0.9937	0.9965	0.9947	0.9946	1

Table 3: Correlation matrix for all statistical explanatory variables

	ahp	awp	hs	ahs	hst	ahst	hf	ahf	hy
ahp	1								
awp	0.6775	1							
hs	0.1916	-0.0047	1						
ahs	0.8966	0.7804	0.1864	1					
hst	0.1741	0.012	0.8339	0.1672	1				
ahst	0.9031	0.7754	0.1876	0.9965	0.1773	1			
hf	-0.004	0.0585	0.0826	0.0065	0.0728	0.001	1		
ahf	0.7498	0.7977	0.0541	0.8634	0.0675	0.85	0.1382	1	
hy	0.1152	0.0998	-0.0596	0.0847	-0.0547	0.0864	0.3742	0.0934	1
ahy	0.74	0.7521	0.0615	0.8326	0.0564	0.82	0.1037	0.9226	0.222
hr	-0.0298	-0.0107	-0.056	-0.055	-0.0763	-0.0547	0.0385	-0.0655	0.0611
ahr	0.4619	0.4227	-0.0128	0.4293	-0.0252	0.4226	0.0792	0.55	0.1574
as	0.083	0.2415	0.102	0.1193	0.0556	0.1136	0.1305	0.148	0.1077
aas	0.7537	0.8979	0.054	0.8596	0.0684	0.8491	0.0429	0.9063	0.0896
ast	0.0988	0.2565	0.0312	0.125	0.0239	0.1248	0.1246	0.1514	0.1066
aast	0.7437	0.9058	0.0457	0.8479	0.0616	0.8375	0.0416	0.895	0.0875
af	0.1071	0.014	0.1348	0.1108	0.0738	0.1029	0.1546	0.1025	0.0913
aaf	0.8025	0.7424	0.0957	0.9072	0.1005	0.8953	0.0638	0.9406	0.0821
ay	0.0022	-0.0564	0.0587	0.0042	-0.0057	0.0037	0.078	0.0349	0.0286
aay	0.775	0.7422	0.0715	0.8775	0.0751	0.867	0.0539	0.8905	0.0622
ar	-0.0248	-0.086	0.0386	-0.0637	0.03	-0.0634	0.002	-0.0628	0.0241
tar	0.4935	0.4333	0.0711	0.5394	0.042	0.5368	0.0158	0.5402	0.0327

	ahy	hr	ahr	as	aas	ast	aast	af	aaf
ahy	1								
hr	-0.033	1							
ahr	0.5997	0.1613	1						
as	0.1555	0.0584	0.0376	1					
aas	0.8432	-0.0337	0.4948	0.2699	1				
ast	0.1673	0.0629	0.0851	0.8534	0.26	1			
aast	0.8326	-0.0302	0.4833	0.271	0.9965	0.2709	1		
af	0.0953	-0.0184	0.0678	0.0192	0.0603	0.0186	0.0529	1	
aaf	0.898	-0.0781	0.539	0.102	0.8632	0.1065	0.8493	0.1781	1
ay	0.0142	0.0291	0.0985	-0.001	-0.0208	-0.0582	-0.0269	0.3911	0.0416
aay	0.8493	-0.0714	0.5003	0.1051	0.8324	0.1207	0.8202	0.1718	0.9277
ar	-0.0469	-0.0624	-0.0404	-0.0835	-0.0794	-0.0356	-0.0801	0.1016	-0.0501
tar	0.5285	-0.0662	0.2769	-0.0143	0.4257	0.0307	0.4189	0.072	0.5604

	ay	aay	ar	tar
ay	1			
aay	0.0968	1		
ar	0.1432	-0.0259	1	
tar	0.0175	0.6064	0.1655	1

Table 4: Correlation matrix for statistical explanatory variables after eliminating variables whose correlation is greater than 0.9

	ahp	awp	hs	ahs	hst	hf	ahf	hy	hr
ahp	1								
awp	0.6775	1							
hs	0.1916	-0.0047	1						
ahs	0.8966	0.7804	0.1864	1					
hst	0.1741	0.012	0.8339	0.1672	1				
hf	-0.004	0.0585	0.0826	0.0065	0.0728	1			
ahf	0.7498	0.7977	0.0541	0.8634	0.0675	0.1382	1		
hy	0.1152	0.0998	-0.0596	0.0847	-0.0547	0.3742	0.0934	1	
hr	-0.0298	-0.0107	-0.056	-0.055	-0.0763	0.0385	-0.0655	0.0611	1
ahr	0.4619	0.4227	-0.0128	0.4293	-0.0252	0.0792	0.55	0.1574	0.1613
as	0.083	0.2415	0.102	0.1193	0.0556	0.1305	0.148	0.1077	0.0584
ast	0.0988	0.2565	0.0312	0.125	0.0239	0.1246	0.1514	0.1066	0.0629
af	0.1071	0.014	0.1348	0.1108	0.0738	0.1546	0.1025	0.0913	-0.0184
ay	0.0022	-0.0564	0.0587	0.0042	-0.0057	0.078	0.0349	0.0286	0.0291
ar	-0.0248	-0.086	0.0386	-0.0637	0.03	0.002	-0.0628	0.0241	-0.0624
tar	0.4935	0.4333	0.0711	0.5394	0.042	0.0158	0.5402	0.0327	-0.0662

	ahr	as	ast	af	ay	ar	tar
ahr	1						
as	0.0376	1					
ast	0.0851	0.8534	1				
af	0.0678	0.0192	0.0186	1			
ay	0.0985	-0.001	-0.0582	0.3911	1		
ar	-0.0404	-0.0835	-0.0356	0.1016	0.1432	1	
tar	0.2769	-0.0143	0.0307	0.072	0.0175	0.1655	1

Table 5: Regression summary of model (1.1)

Probit regression — Number of obs = 380
LR chi2(17) = 58.74
Prob > chi2 = 0.0000
Log likelihood = -233.68673 — Pseudo R2 = 0.1117

d_win	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
b365h	1.930591	0.699361	2.76	0.006	0.559869	3.301312
ahp	-0.00371	0.009627	-0.39	0.7	-0.02258	0.015156
awp	-0.01346	0.00918	-1.47	0.142	-0.03146	0.004529
hs	0.008038	0.028172	0.29	0.775	-0.04718	0.063254
ahs	0.00326	0.001837	1.77	0.076	-0.00034	0.006861
hst	-0.01085	0.042045	-0.26	0.796	-0.09325	0.071562
hf	0.01293	0.017329	0.75	0.456	-0.02103	0.046894
ahf	-0.00173	0.001263	-1.37	0.171	-0.0042	0.000745
hy	-0.11572	0.058568	-1.98	0.048	-0.23051	-0.00093
hr	-0.01939	0.233387	-0.08	0.934	-0.47682	0.438036
ahr	0.061048	0.067277	0.91	0.364	-0.07081	0.192909
as	0.034614	0.025602	1.35	0.176	-0.01556	0.084792
ast	-0.05411	0.041264	-1.31	0.19	-0.13498	0.026767
af	-0.00997	0.016937	-0.59	0.556	-0.04316	0.023228
ay	-0.06832	0.062137	-1.1	0.272	-0.1901	0.053471
ar	0.191535	0.38407	0.5	0.618	-0.56123	0.944298
tar	-0.08946	0.065126	-1.37	0.17	-0.2171	0.038186
_cons	-0.57859	0.447268	-1.29	0.196	-1.45521	0.298043

Table 6: Regression summary of model (1.2)

Iteration 0: log likelihood = -263.05899
Iteration 1: log likelihood = -239.51272
Iteration 2: log likelihood = -239.49807
Iteration 3: log likelihood = -239.49807

Probit regression — Number of obs = 380
LR chi2(3) = 47.12
Prob > chi2 = 0.0000
Log likelihood = -239.49807 — Pseudo R2 = 0.0896

d_win	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
b365h	2.717396	0.433417	6.27	0	1.867914	3.566877
ahs	-0.00064	0.000513	-1.25	0.21	-0.00165	0.000363
hy	-0.09911	0.052208	-1.9	0.058	-0.20143	0.003217
_cons	-0.96786	0.227716	-4.25	0	-1.41418	-0.52155

Table 7: Regression summary of model (1.3)

Iteration 0: log likelihood = -263.05899
Iteration 1: log likelihood = -240.29725
Iteration 2: log likelihood = -240.2857
Iteration 3: log likelihood = -240.2857

Probit regression — Number of obs = 380
LR chi2(2) = 45.55
Prob > chi2 = 0.0000
Log likelihood = -240.2857 — Pseudo R2 = 0.0866

d_win	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
b365h	2.619339	0.425505	6.16	0	1.785364	3.453314
hy	-0.10547	0.051865	-2.03	0.042	-0.20712	-0.00382
_cons	-1.04533	0.219038	-4.77	0	-1.47464	-0.61602

References

1. Busche, Kelly, and Hall, Christopher D. ‘‘An Exception to the Risk Preference Anomaly.’’ *J. Bus*. 61 (July 1988): 337–46.

2. E. ˇ Strumbelj_, M. Robnik ˇ Sikonja “Online bookmakers’ odds as forecasts: The case of European soccer leagues.” *International Journal of Forecasting* 26 (2010) 482-8.

3. Golec, Joseph, and Tamarkin, Maurry “ Bettors Love Skewness, Not Risk, at the Horse Track.” *The Journal of Political Economy* 106 (February 1998): 205-225

4. Kanto, Antti J.; Rosenqvist, Gunnar; and Suvas, Arto. ‘‘On Utility Function Estimation of Racetrack Bettors.’’ J. *Econ. Psychology* 13 (September 1992): 491–98

5. Quandt, Richard E. ‘‘Betting and Equilibrium.’’ *Q.J.E*. 101 (February 1986): 201–7.

6. Weitzman, Martin. ‘‘Utility Analysis and Group Behavior: An Empirical Study.’’ *J.P.E*. 73 (February 1965): 18–26.