©Journal of Sports Science and Medicine (2006) **5**, 480-487
http://www.jssm.org

Research article

The 8th Australasian Conference on Mathematics and Computers in Sport, 3-5 July 2006, Queensland, Australia

PREDICTING THE MATCH OUTCOME IN ONE DAY INTERNATIONAL CRICKET MATCHES, WHILE THE GAME IS IN PROGRESS

Michael Bailey [1] ✉ and Stephen R. Clarke [2]

[1] Department of Epidemiology & Preventive Medicine, Monash University, Australia
[2] Swinburne University of Technology, Melbourne, Australia.

Published (online): 15 December 2006

ABSTRACT

Millions of dollars are wagered on the outcome of one day international (ODI) cricket matches, with a large percentage of bets occurring after the game has commenced. Using match information gathered from all 2200 ODI matches played prior to January 2005, a range of variables that could independently explain statistically significant proportions of variation associated with the predicted run totals and match outcomes were created. Such variables include home ground advantage, past performances, match experience, performance at the specific venue, performance against the specific opposition, experience at the specific venue and current form. Using a multiple linear regression model, prediction variables were numerically weighted according to statistical significance and used to predict the match outcome. With the use of the Duckworth-Lewis method to determine resources remaining, at the end of each completed over, the predicted run total of the batting team could be updated to provide a more accurate prediction of the match outcome. By applying this prediction approach to a holdout sample of matches, the efficiency of the “in the run” wagering market could be assessed. Preliminary results suggest that the market is prone to overreact to events occurring throughout the course of the match, thus creating brief inefficiencies in the wagering market.

KEY WORDS: Linear regression, live prediction, market efficiency, betting .

INTRODUCTION

The first official one day international (ODI) match was played in 1971 between Australia and England at the Melbourne Cricket Ground. Whilst ODI cricket has developed over the past 35 years (2300 matches), the general principles have remained the same. Both sides bat once for a limited time (maximum 50 overs) with the aim in the first innings to score as many runs as possible, and in the second innings to score more than the target set in the first innings. The high scoring nature of ODI matches ensures that team totals and differences between scores can be well approximated by a normal distribution. As shown by (Bailey, 2005), this facilitates the use of multiple linear regression to predict a margin of victory (MOV) prior to the commencement of the match. Using a similar approach, a multiple linear regression is also used to predict the number of runs scored by the team

Table 1. Percentage of resources available for overs remaining and wickets lost.

Overs	Wickets lost									
remaining	0	1	2	3	4	5	6	7	8	9
50	100.0	93.4	85.1	74.9	62.7	49.0	34.9	22.0	11.9	4.7
40	89.3	84.2	77.8	69.6	59.5	47.6	34.6	22.0	11.9	4.7
30	75.1	71.8	67.3	61.6	54.1	44.7	33.6	21.8	11.9	4.7
25	66.5	63.9	60.5	56.0	50	42.2	32.6	21.6	11.9	4.7
20	56.6	54.8	52.4	49.1	44.6	38.6	30.8	21.2	11.9	4.7
15	45.2	44.1	42.6	40.5	37.6	33.5	27.8	20.2	11.8	4.7
10	32.1	31.6	30.8	29.8	28.3	26.1	22.8	17.9	11.4	4.7
5	17.2	17.0	16.8	16.5	16.1	15.4	14.3	12.5	9.4	4.6
1	3.6	3.6	3.6	3.6	3.6	3.5	3.5	3.4	3.2	2.5

batting first. With the use of (Duckworth and Lewis, 1999) approach of converting resources available into runs, as each over is bowled, the current total and the predicted total for the remaining overs are combined to produce an updated predicted total for the batting team. The difference between the pre-match predicted total and the updated predicted total provides a measure of how the batting team is performing through the course of their inning. This difference is then used to provide an updated prediction for the MOV.

METHODS

In ODI cricket the aim of the team batting first is to score as many runs as possible in the allotted time (usually 50 six ball overs). If the first team scores more runs than the second team, the MOV can readily be expressed in terms of runs difference between the two teams. The aim of the side batting second is to score more runs than the first team. Because the game is deemed to be finished if the team batting second achieves their target, the MOV is usually expressed in terms of resources (wickets and balls) remaining, rather than runs. In order to develop a predictive process for match outcomes, a consistent measure of the MOV is required. This can be achieved by following the work of Duckworth and Lewis (1999) to convert resources available into runs.

Frank Duckworth and Tony Lewis developed a now well-known system for resetting targets in ODI matches that were shortened due to rain. Although this system has undergone several refinements in recent years, the general way in which the Duckworth-Lewis (D-L) method is calculated has not changed, with wickets and balls remaining expressed as resources available and converted to runs. Table 1 shows an abbreviated version of the remained resources (R) for wickets lost and balls remaining. A complete tables and detailed account of the derivation of this table is given by Duckworth and Lewis (1999).

Whilst the D-L approach was specifically designed to improve 'fairness' in interrupted one-day matches, (de Silva et al., 2001) found that when used to quantify the MOV, the D-L approach sometimes overestimated the available resources when the second team to bat won easily, and underestimated the available resources when the second team to bat only just won. By minimizing the Cramer-von Mises statistic for the differences between actual and predicted runs, de Silva derived a formula to reduce bias by modifying the remaining resources. This is given by

$$R_{mod} = (1.183 - 0.006R)R \qquad (1)$$

where R_{mod} = modified resources and R = resources given using D-L (see Table 1).

When an ODI match is won by the team batting first, the MOV is readily determined by the difference in runs scored. When the match is won by the team batting second, the MOV can be found by multiplying the first innings run total by the corresponding modified percentage of resources remaining as given by (1). By referencing the MOV so that a 'home' win has a positive value and an 'away' win has a negative value, it can be seen from Figure 1, that the underlying distribution for MOV can be well approximated by a Normal distribution.

Statistical analysis

All analysis was performed using SAS version 8.2 (SAS Institute Inc., Cary, NC, USA). Multiple linear regression models were constructed using a stepwise selection procedure and validated a backward elimination procedure. To increase the robustness of the prediction models a reduced level of statistical significance was incorporated with all variables achieving a level of significance below p = 0.005. Comparisons between continuously normally distributed variables were made using student t-tests.

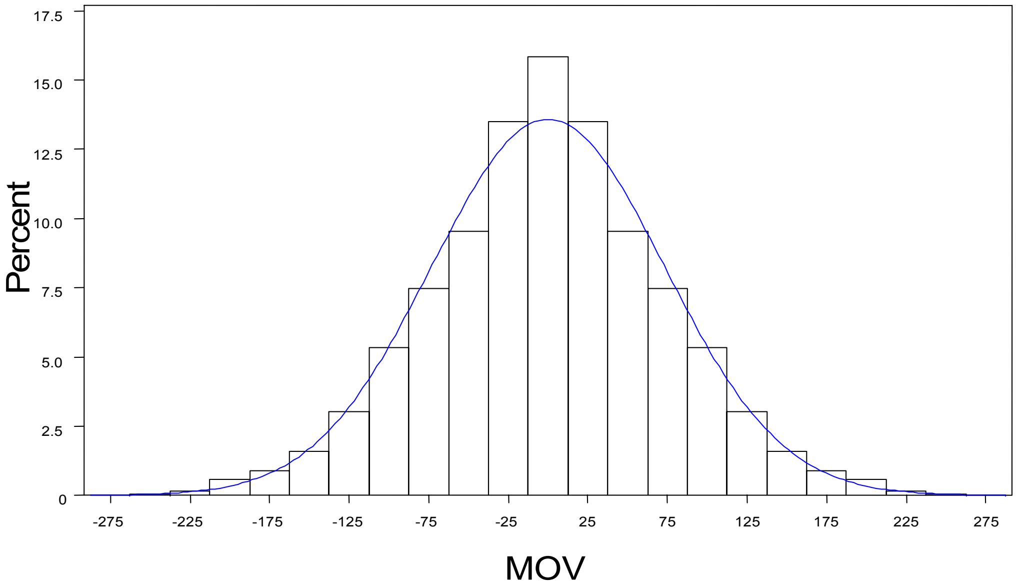

Figure 1. Histogram of MOV referenced against the home team in 2200 matches played prior to Jan 2005.

Prediction models for MOV

Using match and player information from 1800 ODIs played prior to Jan 2002, (Bailey, 2005) combined measures of recent form, experience, overall quality and home ground advantage (HA), to produce a prediction model that was successfully used to identify inefficiencies the betting market for ODI matches. Using 2200 matches played prior to January 2005 an updated version of this model was created and compared to the original.

Prediction variables of experience, quality and form were derived by developing separate measures for both teams and then subtracting the away team values from the home team values. This effectively references the final result in term of the home team. Indicator variables were created to identify matches played at a neutral venue and matches where the two competing teams were clearly from different class structures (established nation versus developing nation).

From Table 2 it can be seen that the results of ODI matches are becoming more predictable, with the updated model explaining 3.5% more of the variation in ODI outcomes (R-square: 23.4% vs. 19.6% $p < 0.0001$).

Because the MOV in the regression model is nominally structured in favour of the home team, the intercept term in the regression equation reflects HA. It can be seen from Table 2 that HA is equivalent to about 14 runs and is highly statistically significant ($p < 0.0001$). Because one third of all ODI have been played at neutral venues, a binomial indicator variable was created to negate the HA for these games. As the regression process requires a ‘Home’ and ‘Away’ team, when playing at neutral venue, the team with the most experience at the venue was assigned to be the ‘Home’ team. If all matches played at neutral venues were devoid of HA then the binomial variable for a neutral venue would be the exact negative of the intercept term. This was not the quite the case, with the neutral variable equivalent to about eight runs, suggesting a HA in neutral matches equivalent to about six runs. This six run difference could be thought of as a surrogate marker for the difference in familiarity between the competing teams.

Table 2. Multivariate models for MOV constructed with 1800 & 2200 ODI matches.

	Bailey model (n = 1800)			**Updated Model (n = 2200)**		
Variable	**Estimate**	**P-value**	**Partial R^2**	**Estimate**	**P-value**	**Partial R^2**
Intercept / HA	13.4 ± 1.9	<.0001		13.9 ± 1.8	<.0001	
Average Ever	.6 ± .1	<.0001	17.3%	.6 ± .06	<.0001	20.7%
Class	-29.6 ± 6.7	<.0001	1.2%	-25.1 ± 5.9	<.0001	1.0%
Experience	.2 ± .1	.002	0.4%	.2 ± .07	.0003	0.4%
Ave. last 10	.1 ± .04	.003	0.4%	.2 ± .04	<.0001	0.7%
Neutral Venue	-8.6 ± 3.2	.007	0.3%	-8.2 ± 3.2	.005	0.3%
Total R^2			19.6%			23.1%

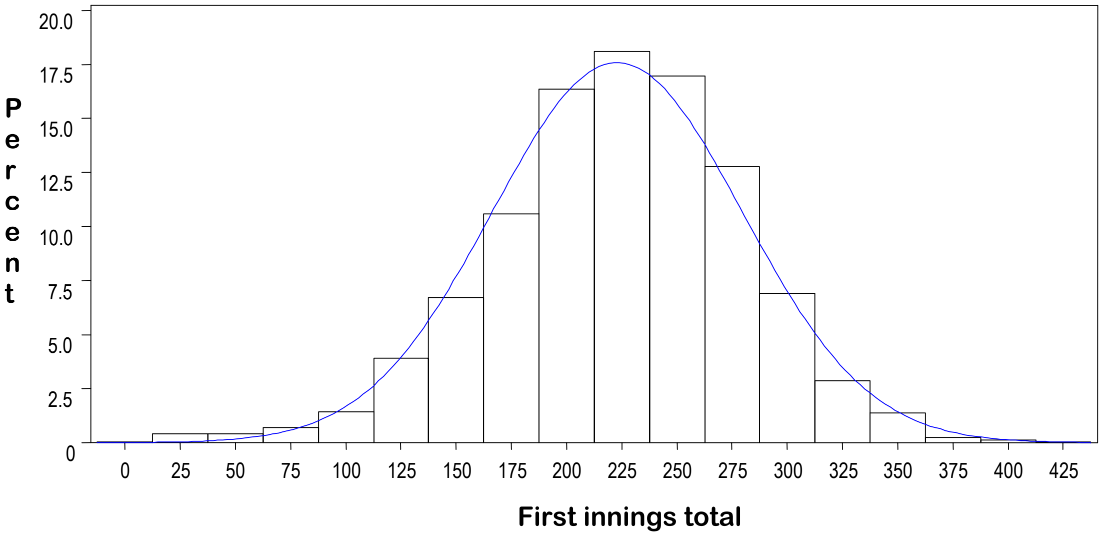

Figure 2. Histogram of first inning scores in 2200 matches played prior to Jan 2005.

The difference in quality, as measured by the difference in averages between the two teams for all past matches, was by far the strongest predictor, explaining 20.7% of the variation in the updated model. The best measure of current form was the difference in averages for the past 10 matches, whilst the difference in overall experience (games played by the country) between the home and away team was also statistically significant. Whilst no statistically significant difference could be found in parameter estimates, the difference in class (when a developing cricket nation played host to an established cricket nation) declined (29.6 runs vs. 25.1 runs) as developing nations gain more experience. Similarly, the effect of HA rose slightly (13.4 runs vs. 13.9 runs) with more data, while the effect of a neutral venue was slightly lower (8.6 runs vs. 8.2 runs). Not surprisingly, all variables in the model achieved a higher level of statistical significant when additional data were used.

Prediction model for team totals

Figure 2 it shows that the total of the team batting first can be well approximated by a normal distribution (mean = 229.7, SD = ± 1.2). When the score of the team batting first was shortened due to rain, (about 13% of matches), the DL method was once again incorporated to determine a projected total.

Using past averages and exponential smoothing, prediction variables relating to past performance were created. Using a multiple linear regression, a six variable model was constructed. The resulting parameter values are given in Table 3.

Interestingly, when using a stepwise selection procedure, the strongest predictor of the total scored by the team batting first was in fact the average of the past MOV between the two teams. The next strongest predictors in the model were derived from the past first innings scores achieved by the batting team as well as scores conceded by the bowling team. HA was the next predictor of importance, with a team playing in it home country scoring an additional 15 runs. A second surrogate marker for the quality of the batting team was given by the average past MOV for the batting team. The final variable that was found to be highly statistically significant (p = 0.0004) was derived from all past first innings played at the venue. This helped account for pitch conditions and venue size.

Whilst over 23% of the variation in MOV could be explained by the multivariate model, the total of the team batting first was not as predictable, with an R-square statistic of 19.1%.

Table 3. Multivariate model predicting the total of the team batting first.

Variable	**Estimate**	**P-value**	**Partial R^2**
Ave. MOV against opposition	.13 ± .04	<.0001	9.7%
Exp. Smooth past totals 1st inning batting team	.25 ± .04	<.0001	3.6%
Ave. total conceded in 1st inning by bowling team	.53 ± .06	<.0001	2.6%
Home Country	15.3 ± 2.3	<.0001	1.6%
Ave. MOV ever	.31 ± .05	<.0001	1.1%
Exp. Smooth past totals 1st inning at venue	.38 ± .05	.0004	.5%
Total R^2			19.1 %

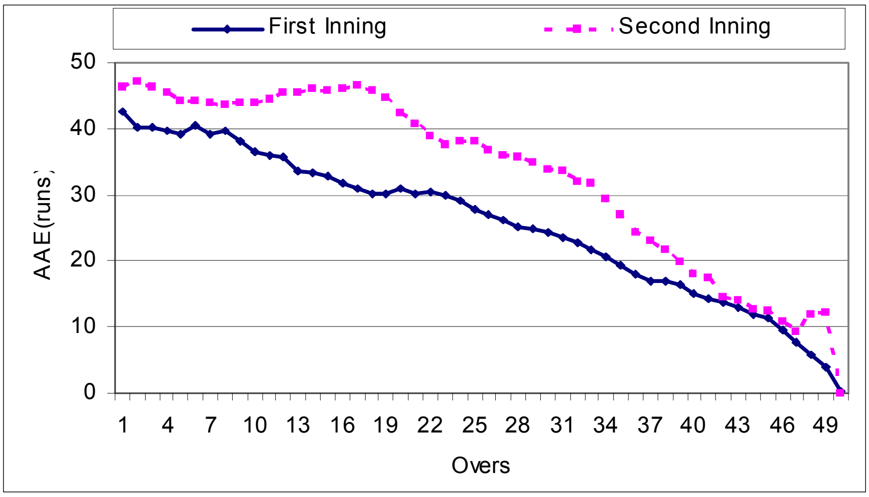

Figure 3. AAE for difference between predicted and actual total.

Using a holdout sample of 100 completed matches played in the year 2005, the regression model successfully predicted the winning team 71% of the time and had an Absolute Average Error (AAE) between the predicted and actual margin of 55.8 ± 4.1 runs. These results compare favourably against the original prediction model of (Bailey, 2005), who accurately identified the winning team 69.6% of the time, and had an AAE of 54.6 ± 0.9 runs for a sample of 336 matches played between 2002 and 2004.

Using the same holdout sample of 100 matches, the AAE for the difference between the predicted and actual totals of the team batting first was 42.5 ± 3.2 runs. By referencing the MOV in terms of the team batting first rather than the home team, a predicted total for the team batting second can be given by

Predicted Total2 = (Predicted Total1) + (Predicted MOVordered) (2)

From the chosen holdout sample of 100 matches, the AAE for the difference between the predicted and actual totals of the team batting second was 47.1 ± 4.0 runs.

RESULTS

With the use of the D-L method to convert available resources into runs, at the completion of each over, an updated total for the team batting first is calculated by combining the actual total with the predicted total for the remainder of the innings.

Updated Total = (existing score) + (projected total for remaining resources) (3)

Using complete over by over information for the 100 match holdout sample, it can be seen from Figure 3 that the accuracy with which the total of the batting team can be predicted, progressively improves throughout the course of the innings, with first innings totals significantly more accurate that those of the second innings.

By subtracting the pre-match predicted total from the updated prediction of the total, a performance indicator can be derived for whether each batting team is performing above or below expectation.

Performance indicator = (updated total) – (pre-match predicted total) (4)

With the use the performance indicator, an updated prediction for the MOV can then be readily obtained

Updated MOV = (Pre-match MOV) + (Performance indicators) (5)

From Figure 4 it can be seen that during the course of the first innings, the AAE for the difference between the predicted and actual MOV reduces by about 10 runs. In the second innings the reduction in AAE is much greater as the game draws nearer to its conclusion.

As shown by (Bailey, 2005), by dividing the predicted MOV by its standard error and comparing with a standard Normal distribution, the approximate probability that either side will win the match can be readily calculated.

Example: On December 7 2005, Australia played New Zealand in a day/night match at Westpac Stadium in Wellington. After winning the toss and electing to bat Australia proceeded to score a very respectable total of 322. The betting exchange Betfair fielded a betting market for this match, with just over $1,000,000 AUD of matched bets occurring before the start of the game. As betting on

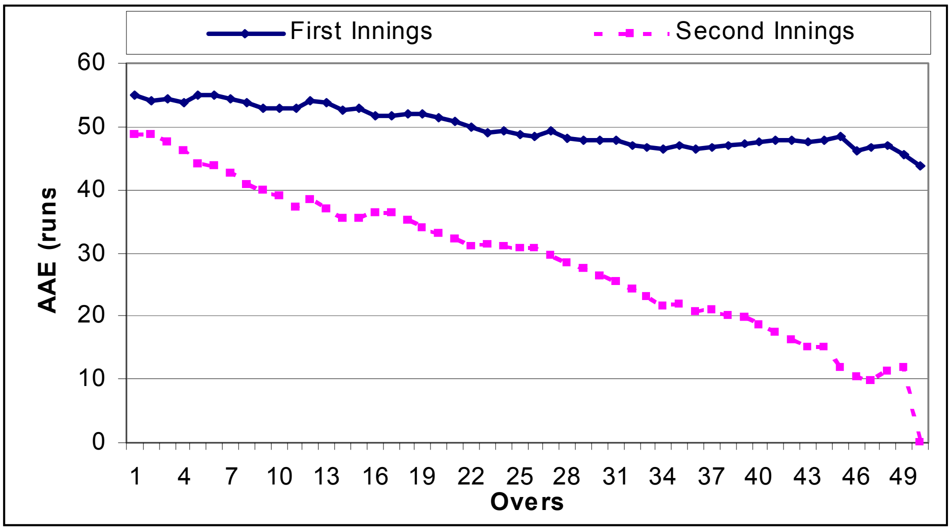

Figure 4. AAE for the difference between the predicted and actual MOV.

this match remains open for the duration of the game, by the completion of the Australian innings, just over $4,000,000 AUD of matched bets had been placed. Figure 5 shows both the volume of bets placed and the price matched. From Figure 5 it can be seen that the opening price for Australia was about $1.38, with the price dropping to $1.30 after Australia won the toss. After losing 3 early wickets, the price drifted out to $1.70, but as Australia rallied, the price continued to drop and by the completion on the 50th over, the best price available for Australia to win was $1.08.

Using prediction models for the team total and MOV, the predicted probability for Australia to win was calculated both before and during the match, and compared with the market price offered by Betfair (market probabilities included 5% for commission). Where the predicted probability can be seen to exceed the market probability, the 'in play' market can be thought to be inefficient. From Figure 6 it can be seen that while Australia was batting, the predicted probability for Australia to win was consistently below the market probability, with only one inefficiency occurring throughout the course of the Australian innings.

Chasing 323 runs to win the match, New Zealand started slowly. With some big hitting towards the end of the innings, the black caps clawed their way into contention and started the final over as favourites, only requiring six runs to win. Unfortunately, two wickets falling in the final over gave victory to Australia by 2 runs. Figure 7 shows that several inefficiencies were present in the betting market with the predicted probability of success often exceeding the market price. By the completion of the 100th over, more than $9,000,000 AUD had been wagered on the outcome of the match.

DISCUSSION

In July 2005 the International Cricket Council (ICC) announced a new set of rules to be applicable to ODI

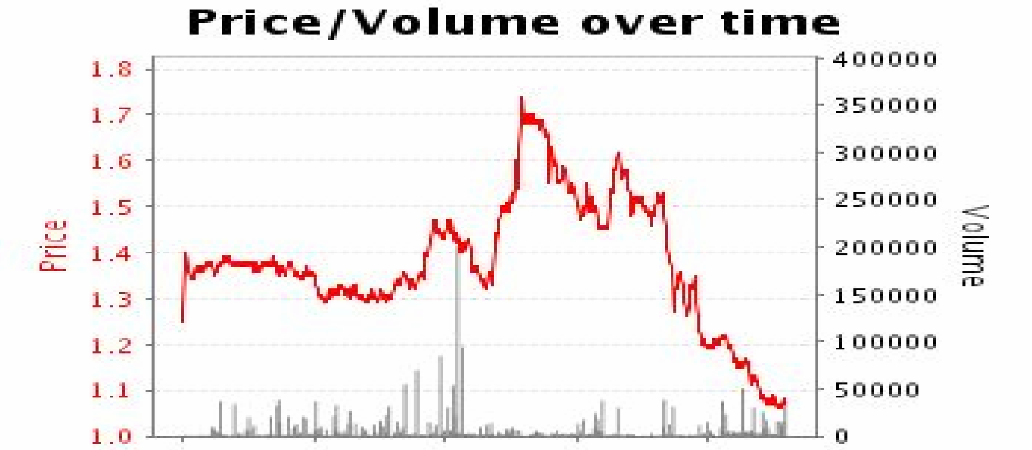

Figure 5. Betfair volume and price for Australia vs. New Zealand ODI 2302 (pre match until end over 50).

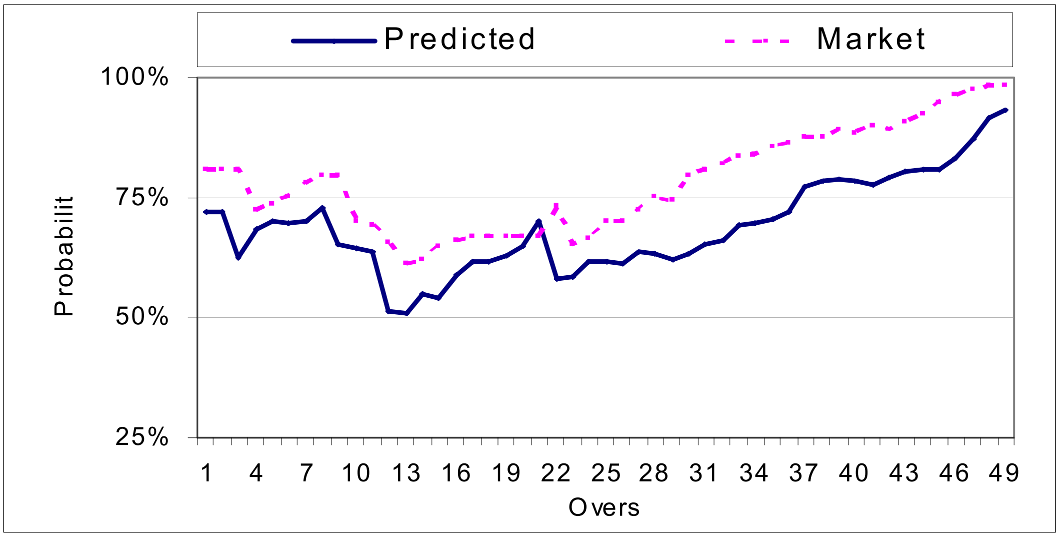

Figure 6. Predicted probability and market price for Australia to win against New Zealand ODI 2302 (pre match until end over 50).

matches. An increase in fielding restrictions and the introduction of a substitute player (super-sub), significantly increased the total achieved by the team batting first by more than 20 runs. (252.7 ± 8.0 vs. 229.7 ± 1.2 p = 0.002). As these changes occurred within the holdout sample of the data used, it is unsure how these modifications would impact upon the prediction process.

Whilst the price and volume of bets traded are available through Betfair (see Figure 5), this information is not time coded by over, ensuring that if the efficiency of the market is to be accurately determined, information must be gathered manually at the completion of each over. This would undoubtedly prove time consuming should a definitive appraisal of the market inefficiency be required.

In Australia, federal laws prevent Australian citizens from placing bets over the internet after a sporting event has commenced. Paradoxically, Australian citizen can place bets ‘in the run’ provided the bets are placed over the phone. This inconvenience causes a greater delay between observing an inefficient price and actually placing a bet.

CONCLUSIONS

Multiple linear regression provides a useful way to assign the winning probabilities to the competing teams in ODI matches. With the use of D-L approach, this process can be readily modified to produce ‘in the run’ predictions. Whilst a definitive analysis of the efficiency of the betting market is yet

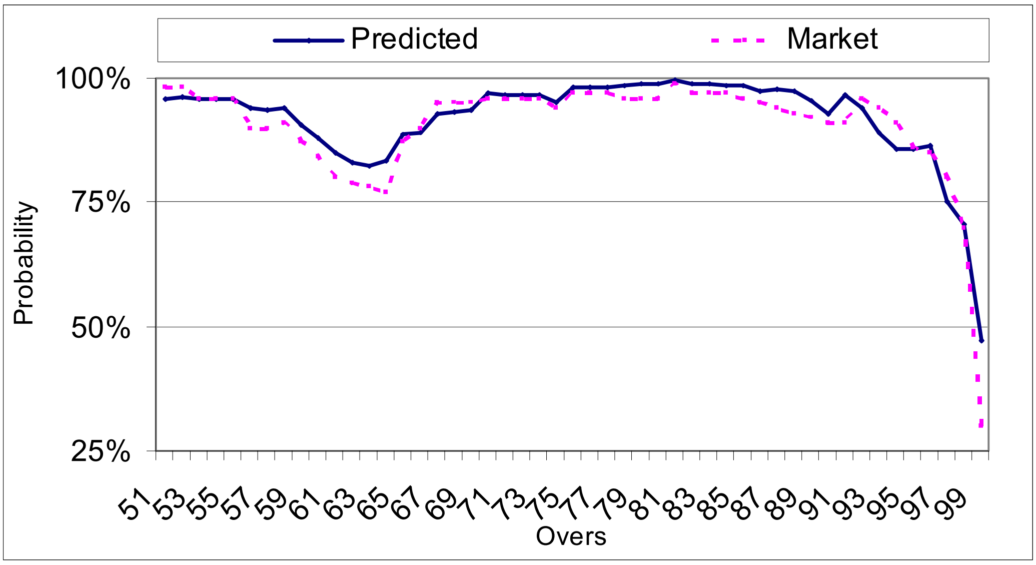

Figure 7. Predicted probability and market price for Australia to win against New Zealand ODI 2302 while New Zealand batted (overs 51-100).

to be conducted, preliminary evidence suggest punters may be prone to over or under estimate the true probability of the competing teams as the game progresses.

REFERENCES

Bailey M., (2005) *Predicting sporting outcomes: A statistical approach*: PhD thesis, Swinburne University, Melbourne. 212.

de Silva, B., Pond, G. and Swartz, T. (2001) Estimation of the magnitude of victory in one-day cricket. *Australian & New Zealand Journal of Statistics*, **43,** 259-268.

Duckworth, F. and Lewis, T. (1999) *Your comprehensive guide to the Duckworth/Lewis method for Resetting targets in one-day cricket,* University of the West of England.

KEY POINTS

- In excess of 80% of monies wagered on the outcome of ODI matches are placed after the match has commenced.
- Using all past data from ODI matches, multiple linear regression models are constructed to predict team totals and margin of victory.
- By combining match information with prediction models, an 'in the run' prediction process is created for ODI matches.

AUTHORS BIOGRAPHY

Michael J. BAILEY

Employment

Statistician, Department of Epidemiology & Preventive Medicine, Monash University, Australia.

Degrees

PhD, MSc (Statistics), BSc(Hons).

Research interests

Health, sport, gambling.

E-mail: Michael.Bailey@med.monash.edu.au

Stephen R. CLARKE

Employment

Professor, Swinburne University of Technology, Australia.

Degrees

PhD, M.A., B.Sc(Hons), Dip Ed..

Research interests

Modelling in sport, gambling.

E-mail: sclarke@swin.edu.au

✉ **Michael J. Bailey**

Department of Epidemiology & Preventive Medicine, Monash University, Australia.