FACTORS AFFECTING OUTCOMES IN TEST MATCH CRICKET

Paul Allsopp and Stephen R. Clarke
School of Mathematical Sciences
Swinburne University
PO Box 218, Hawthorn
Victoria 3122, Australia
paulalls@bigpond.com, sclarke@groupwise.swin.edu.au

Abstract

Least squares regression is used to model the first innings performances of teams in test cricket in order to establish batting and bowling ratings, a common home advantage and a country effect. Logistic regression techniques are then used to model match outcomes based on a team's first innings lead, innings duration, home advantage, batting and bowling ratings and the country effect. It is shown that the factors that impact most significantly on the outcome of a match are a team's first innings lead home team performance and innings duration. A team's first innings lead is found to more likely shape a win rather than a draw or a loss whereas the longer the duration of the first innings the more likely a match will end in a draw. It is shown that the home team, on average, needs to establish a lead in excess of 93 runs to have a better than even chance of winning, whereas the away team needs to establish a lead in excess of 115 runs to have the same chance. There is a better than an even chance of a draw for a first innings duration in excess of 1165 minutes (or approximately 277 overs). It is also shown that the home team is more likely to win a match rather than lose or draw, which suggests that the home team has a distinct winning advantage over the away team. There is some evidence suggesting that teams gained an advantage by batting last.

1 Introduction

Test match cricket is currently played between the ten International Cricket Committee (ICC) test-playing nations, with Bangladesh being a very recent inclusion. A test match is scheduled to finish within a five-day period and comprises a maximum of two innings per team. There are four possible outcomes: a win, loss, draw or tie. A tied result is an extremely rare event and has only occurred a handful of times throughout the history of test cricket.

Outcomes in a test cricket match are difficult to predict because they are dependent on a wide range of interrelated factors. By applying standard modelling techniques we will initially focus on the factors that affect the performance of teams in their first innings and then determine which of these factors, if any, have an impact on the outcome of a match. The factors to be analysed are a team's first innings lead, home advantage, team batting order, innings duration, attack and defence ratings and the country effect We have considered all 371 completed matches from seasons 1990 through to 2001. By "completed matches", we mean those matches that produced a result independent of weather conditions. Note that as Bangladesh had only played in three matches throughout the study period their results have not been included in the analysis.

2 Exploratory data analysis

Throughout the analysis Team 1 and Team 2 refer to the teams batting first and second respectively in the first innings. Subsequently, unless Team 2 has been forced to follow on it will also be the team batting last in the second innings. Table 1 provides a descriptive summary of the first innings results for each team. Overall, Team 2 won 140 matches and lost 121. There were 110 draws. The results show that Team 2 had a winning advantage over Team 1. However, using a chi-square goodness of fit test to compare the expected and actual number of wins, losses and draws for Teams 1 and 2 suggests that the observed differences are not significant (p-value = 0.251). The observed differences are due solely to random variation. However, the data refute the accepted wisdom that being first in the batting order provides a team with a winning advantage. Conversely, the data show that the team batting first has a tendency to lose rather than win or draw.

The first innings batting performances by India (as Team 1 and 2), on average, are substantially higher than the majority of nations but are much more variable. This underscores India's lack of batting consistency.

Team	Mean first innings score		Standard deviation	
	Team 1	Team 2	Team 1	Team 2
Australia	358	368	131	134
South Africa	354	333	106	106
India	352	342	165	108
England	298	309	121	119
Pakistan	296	326	110	129
Sri Lanka	289	320	107	165
West Indies	284	312	131	133
New Zealand	283	286	111	104
Zimbabwe	262	263	131	70
Overall	312	321	128	125

Table 1: Descriptive summary of the first innings in test cricket.

3 Modelling the first innings

In a typical test match the first innings batting side aims to score as many runs as possible before losing the ten wickets at their disposal, whereas the bowling side aims to dismiss the batting side by taking all ten wickets for a total that is as small as possible. Assuming that both teams are endeavouring to maximise their first innings lead the score that is achieved provides a measure of the relative batting and bowling strength of the two teams. Beyond the first innings, however, playing strategies are harder to predict because teams become more reactionary and tend to customise their style of play. Using techniques similar to those adopted by Harville and Smith (1994), Clarke and Norman (1995), de Silva, Pond and Swartz (2000) and Clarke and Allsopp (2001), a team's first innings score in a test match played between the batting team i and the bowling team j at a location k with home ground l and batting order m is modelled as

$$S_{ijklm} = A + a_i - d_j + c_k + h_{il} + b_m + \epsilon_{ijklm}, \tag{1}$$

where the indices i, j, k, $l = 1, \ldots, 9$ represent the nine ICC test-playing nations and $m = 1, 2$ indicates whether a team batted first or second. The response variable s_{ijklm} signifies a team's first innings score; A represents the expected score between average teams on a neutral ground; a_i and d_j signify the first innings batting (attack) and bowling (defence) ratings of teams i and j, respectively;

c_k is the country effect term, which represents the advantage gained by teams playing in a particular country. The common home advantage enjoyed by the batting side is represented by h_{il}, such that

$$h_{il} = \begin{cases} h, & \text{if } l = i, \\ 0, & \text{otherwise.} \end{cases}$$

The team batting first is indicated by b_m, such that

$$b_m = \begin{cases} b, & \text{if } m = i, \\ 0, & \text{otherwise.} \end{cases}$$

Finally, ϵ_{ijklm} is a zero-mean random error. The error term is included because two competing teams will not necessarily repeat their first innings performances the next time they meet. Subsequently, a least squares regression model is fitted to the scores to quantify the parameter estimates for each of the explanatory variables. For convenience, $\sum_{i=1}^{9} a_i = 900$, $\sum_{j=1}^{9} d_j = 900$ and $\sum_{k=1}^{9} c_k = 0$, which assumes that a team's first innings average batting and bowling ratings are each 100 and the country effect rating is 0. Accordingly, a rating greater than 100 signifies that a team has performed above average whereas a rating less than 100 signifies that a team has performed below average.

The batting, bowling and country effect estimates are outlined in Table 2. The parameter estimates associated with the expected score by an average team on a neutral ground, the common home advantage and any advantage gained by batting first are estimated to be 306, 28 and -11 runs, respectively. The p-value for the common home advantage parameter is 0.002, which suggests that the home team, on average, gained a significant first innings runs advantage, whereas the p-value for the batting first parameter is 0.208, which suggests that there is no significant batting order effect.

The long-term dominance of Australia and South Africa in test match cricket is clearly evident, with both teams enjoying batting and bowling ratings substantially above average. All other teams have under-performed in one or both of these areas. Notably, India has performed exceptionally well with the bat but has been let down by relatively poor bowling performances. The negative country effect ratings for India, South Africa and the West Indies suggest that the batting teams playing in these countries were disadvantaged to some degree by the conditions. This latter point highlights India's excellent batting form, particularly when playing at home.

Team	Batting rating	Bowling rating	Country effect rating
India	148	73	−14
Australia	142	145	7
South Africa	132	143	−16
England	96	75	10
Pakistan	94	105	1
West Indies	93	118	−2
Sri Lanka	84	96	8
New Zealand	70	81	3
Zimbabwe	42	63	3

Table 2: First innings ratings.

To show how the model can be applied, assume Australia is playing South Africa at home, with Australia batting first. The model predicts Australia's first innings score to be $306 + 142 - 143 + 28 - 11 + 7 = 329$ runs, whereas South Africa's first innings score is estimated to be $306 + 132 - 145 + 7 = 300$ runs, an advantage to Australia of 29 runs. However, if the match were to be played in South Africa and Australia remains the team batting first, the model predicts Australia's first innings score to be $306 + 142 - 143 - 11 - 16 = 278$ runs, whereas South Africa's first innings score is estimated to be

$306 + 132 - 145 + 28 - 16 = 305$ runs. This time there is an advantage to South Africa of 27 runs. This indicates that with Australia and South Africa being both highly rated teams factors that are indirectly related to a team's batting and bowling performance such as winning the toss or home advantage could have a potentially significant impact on match outcomes.

4 Modelling the second innings

Since the match result is a categorical variable, a logistic regression model is used to model the outcome of a match. The response variable is the match outcome for Team 1 and the explanatory variables are Team 1's lead, the cumulative duration of Team 1's and Team 2's innings, a home team indicator where 1 indicates that Team 1 is the home team and 0 otherwise, the difference in the batting and bowling ratings for competing teams and the country effect. The model is expressed as

$$\ln\left(\frac{\gamma}{1-\gamma}\right) = \beta_0 + \beta_1 l + \beta_2 t + \beta_3 h + \beta_4 d_A + \beta_5 d_B + \beta_6 c + \epsilon, \tag{2}$$

where the response variable γ represents the probability of a win for Team 1. The parameter l signifies the lead enjoyed by Team 1; t represents the innings duration parameter; h indicates whether Team 1 was the home team; d_A and d_B represent the rating differential parameters for the batting and bowling teams; c is the country effect parameter; and ϵ is a zero-mean random error. We will use nominal logistic regression to investigate the three comparisons: win/loss, draw/loss and win/draw. The results are outlined in Tables 3 to 5.

The analysis suggests that the first innings lead contributes significantly to the shaping of a win rather than a loss or a draw. There is also strong evidence suggesting that the home team is also significantly more likely to generate a win rather than a loss or a draw. There is also evidence to suggest that the cumulative time taken to complete each of the first innings is more likely to produce a drawn result rather than a win or a loss. There is some marginal evidence suggesting that Team 2, which generally bats last, is more likely to manufacture a loss rather than a drawn result.

	Parameter	Coefficient	p-value
β_0	Intercept term	-0.1598	0.818
β_1	Lead	0.013797	0.000
β_2	Time	-0.0005405	0.499
β_3	Home	1.0362	0.003
β_4	Rating differential (bat 1 − bowl 2)	0.004729	0.401
β_5	Rating differential (bat 2 − bowl 1)	-0.008445	0.123
β_6	Country effect	0.02493	0.241

Table 3: Results for comparison of win/loss.

	Parameter	Coefficient	p-value
β_0	Intercept term	-4.5721	0.000
β_1	Lead	0.007060	0.000
β_2	Time	0.0045660	0.000
β_3	Home	0.9953	0.002
β_4	Rating differential (bat 1 − bowl 2)	0.001076	0.835
β_5	Rating differential (bat 2 − bowl 1)	-0.009724	0.063
β_6	Country effect	0.02504	0.209

Table 4: Results for comparison of draw/loss.

	Parameter	Coefficient	p-value
β_0	Intercept term	4.4123	0.000
β_1	Lead	0.006736	0.000
β_2	Time	−0.0051065	0.000
β_3	Home	0.0409	0.899
β_4	Rating differential (bat 1 − bowl 2)	0.003654	0.475
β_5	Rating differential (bat 2 − bowl 1)	0.001279	0.800
β_6	Country effect	−0.00011	0.995

Table 5: Results for comparison of win/draw.

To get a sense of the effect of batting first we need to restate the model without the time parameter. This is necessary since when the time parameter is set to zero it has little meaning in the context of investigating any perceived advantage for Team 1. The model is re-expressed as

$$\ln\left(\frac{\gamma}{1-\gamma}\right) = \beta_0 + \beta_1 l + \beta_2 h + \beta_3 d_A + \beta_4 d_B + \beta_5 c + \epsilon. \tag{3}$$

Setting all parameters to zero in effect represents Team 1's advantage at the completion of the first innings with all things being equal, i.e. no lead, playing on a neutral ground in a neutral country and equal ratings in the batting and bowling departments. Using nominal logistic regression, the parameter estimates for the comparison of a win/loss, draw/loss and win/draw are, respectively, −0.5446 (p-value = 0.026), −0.2837 (p-value = 0.194) and −0.2610 (p-value = 0.283). With all things being equal, there is a significant batting order effect, with Team 1 more likely to lose a match rather than win. This suggests that generally the team batting last in the match shows a tendency to win rather than lose. Note that this slightly contradicts the notion outlined in Section 2, which suggested that the effect was not significant. This possibly highlights the cumulative effects such as home advantage, a first innings lead and batting and bowling strength may have on a team's overall performance.

5 Analysis of the first innings lead and innings duration

The influence of the first innings lead established by Team 1 can be modelled as

$$\ln\left(\frac{\gamma_m}{1-\gamma_m}\right) = \beta_0 + \beta_1 l + \beta_2 h + \epsilon_m, \tag{4}$$

where γ_m is the probability of a particular match outcome for Team 1 with $m = 1, 2, 3$ for a win, draw and loss respectively. If a loss for Team 1 signifies the reference event, then, using nominal logistic regression, the probability of a win for Team 1 when Team 1 is the home team is expressed as

$$\gamma_1 = \frac{\exp(\beta_0 + \beta_1 l_1 + \beta_2 h_1 + \epsilon_1)}{1 + \sum_{m=1}^{2} \exp(\beta_0 + \beta_1 l_m + \beta_2 h_m + \epsilon_m)}. \tag{5}$$

The probability of a draw is

$$\gamma_2 = \frac{\exp(\beta_0 + \beta_1 l_2 + \beta_2 h_2 + \epsilon_2)}{1 + \sum_{m=1}^{2} \exp(\beta_0 + \beta_1 l_m + \beta_2 h_m + \epsilon_m)}. \tag{6}$$

The probability of a loss (the reference event) is

$$\gamma_3 = \frac{1}{1 + \sum_{m=1}^{2} \exp(\beta_0 + \beta_1 l_m + \beta_2 h_m + \epsilon_m)} \tag{7}$$

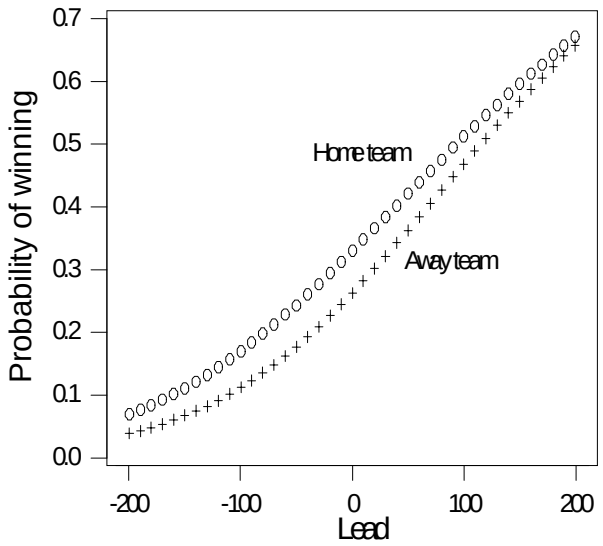

Figure 1: Plot of the probability of winning for the home and away teams.

The results of the analysis are provided in Tables 6 and 7. Notably, the parameter estimate for the lead term in both instances is significant, which confirms that after a team has established a first innings lead they are more likely to win or draw rather than lose a match. Similarly, the home team term is also significant, which suggests that the home team is also more likely to win or draw rather than lose a match. This suggests that the home team has a distinct winning advantage over the away team.

	Parameter	Coefficient	p-value
β_0	Intercept term	-0.4458	0.056
β_1	Lead	0.013040	0.000
β_2	Home	0.8160	0.011

Table 6: Comparison of win/loss.

	Parameter	Coefficient	p-value
β_0	Intercept term	-0.2261	0.285
β_1	Lead	0.007436	0.000
β_2	Home	0.8916	0.002

Table 7: Comparison of draw/loss.

If we let the lead term be zero, so that both teams have the same first innings score, and we apply formulas (5), (6) and (7), then the respective probabilities of a win, draw and loss for Team 1, when Team 1 is the home team, are 0.330, 0.443 and 0.228. These results suggest that after the completion of the first innings, with all things being equal, the home team displays a tendency to win or draw rather than lose a match. A drawn result is the more likely outcome. However, if the lead is increased to 100 runs, say, then the respective probabilities of a win, draw or loss are 0.512, 0.392 and 0.096. As the first innings lead increases, the probability of a win for the home team markedly increases, whereas the probabilities of a draw or loss decrease. To determine the lead the home team needs to establish in order to have a better than even chance of winning we let $\gamma_1 = 0.50$ and the lead in runs be x, such

	Parameter	Coefficient	p-value
β_0	Intercept term	4.2245	0.000
β_1	Time	−0.004312	0.000

Table 8: Comparison of win/draw.

	Parameter	Coefficient	p-value
β_0	Intercept term	4.2242	0.000
β_1	Time	−0.0041396	0.000

Table 9: Comparison of loss/draw.

that

$$0.50 = \frac{\exp(-0.4458 + 0.013040x + 0.8160)}{1 + \exp(-0.4458 + 0.013040x + 0.8160) + \exp(-0.2261 + 0.007436x + 0.8916)}$$

This gives $x = 93$ runs. This suggests that the home team, on average, needs to establish a lead in excess of 93 runs to have a better than even chance of winning. If we repeat this for Team 1 when Team 1 is the away team, this gives $x = 115$ runs. Figure 1 provides a plot of the probabilities of winning for leads up to 200 runs. Clearly, the probability of winning increases for both the home and away teams as the first innings lead increases, with the home team having the upper hand.

The first innings duration can be modelled as

$$\ln\left(\frac{\gamma_m}{1-\gamma_m}\right) = \beta_0 + \beta_1 t_m + \epsilon_m, \tag{8}$$

where γ_m is the probability of a particular match outcome at home with $m = 1, 2, 3$ for a win, loss or draw. If a draw signifies the reference event then using nominal logistic regression the probability of a draw is

$$\gamma_3 = \frac{1}{1 + \sum_{m=1}^{2} \exp(\beta_0 + \beta_1 t_m + \epsilon_m)} \tag{9}$$

Tables 8 and 9 provide a summary of the parameter estimates. Both the parameter estimates for duration are highly significant, with the negative coefficients confirming that the combined duration of the first innings is more likely to shape a draw than a win or a loss. Figure 2 provides a plot of the probability of a draw for Team 1 for durations up to 2000 minutes. Clearly, the likelihood of a draw increases as the duration increases. To determine when there is a better than even chance of a draw let the duration be t, such that

$$0.50 = \frac{1}{1 + \exp(4.2245 - 0.004312t) + \exp(4.2242 - 0.0041396t)}$$

giving $t = 1165$ minutes. We calculated the average duration of an over to be 4.2 minutes, with 1165 minutes equating to approximately 277 overs. This suggests that there is a better than even chance of a draw for a duration in excess of approximately 1165 minutes (or about 277 overs). Conversely, there is a better than even chance of a result for durations less than 1165 minutes.

6 Conclusions

Of the many factors shaping test cricket, there is strong evidence to suggest that the factors which impact most significantly on the outcome of a match are a team's first innings lead, home team performance and the duration of the first innings. Clearly, a team is more likely to win a match after they have established a first innings lead, with the probability of winning increasing as the lead increases. To have

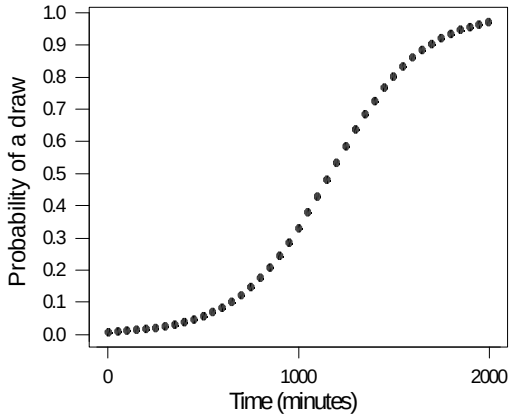

Figure 2: Plot of the probability of a draw against the total time for the first innings.

a better than even chance of winning the home team, on average, needs to establish a lead in excess of 93 runs, whereas the away team needs a lead in excess of 115 runs to have the same chance.

The home team is more likely to win a match rather than lose or draw, which suggests that the home team has a distinct winning advantage over the away team. There is some evidence to suggest that teams gain an advantage by batting second and so are possibly advantaged by batting last in the second innings. This contradicts the accepted wisdom that batting last is a disadvantage. This is an area that encourages more detailed research.

The time taken to complete each of the first innings contributes more to the shaping of a drawn result rather than a win or a loss, with the likelihood of a draw increasing as the duration of the first innings increases. For durations in excess of 1165 minutes (or approximately 277 overs) there is shown to be a better than an even chance of a draw.

References

S. R. Clarke and P. Allsopp (2001), "Fair measures of performance: The World Cup of cricket", *J. Oper. Res. Soc.*, **52**, 471–479.

S. R. Clarke and J. M. Norman (1995), "Home ground advantage of individual clubs in English soccer", *The Statist.*, **44**, 509–521.

B. M. de Silva, G. R. Pond and T. B. Swartz (2000), "Applications of the Duckworth–Lewis method", in *Proceedings of the Fifth Australian Conference on Mathematics and Computers in Sport*, G. Cohen and T. Langtry (editors), University of Technology, Sydney, 113–117.

D. A. Harville and M. H. Smith (1994), "The home-court advantage: How large is it, and does it vary from team to team?", *Amer. Statist.*, **48**, 22–28.

Proceedings of the
Sixth Australian Conference on

MATHEMATICS AND COMPUTERS IN SPORT

Bond University
Queensland

Edited by
Graeme Cohen and Tim Langtry
Department of Mathematical Sciences
Faculty of Science
University of Technology, Sydney

6M&CS

1 – 3 July 2002