# An analysis of strategy in the first three innings in test cricket: declaration and the follow-on

Philip Scarf* and Sohail Akhtar*

**Centre for Operations Management, Management Science and Statistics,*
*Salford Business School,*
*University of Salford,*
*Salford, Manchester M5 4WT, UK.*
(email p.a.scarf@salford.ac.uk)

Salford Business School Working Paper Series

Paper no. 337/10

# An analysis of strategy in the first three innings in test cricket: declaration and the follow-on

Philip Scarf*, Sohail Akhtar, University of Salford

**Abstract.** This paper analyses declaration and the follow-on decisions in test cricket. We model the match outcome given the end of first, second and third innings positions; data on 391 test matches, from the period 1997 to 2007, are used to fit the models. We then investigate how declaration strategy should vary from innings to innings, and how the nature and strength of the covariate effects vary. As the match progresses, the explanatory power of the covariates increases (44% at the end of the first innings, to 80% at the end of the third); the effects of team strengths and home advantage decrease. Overs-remaining, or equivalently overs used, and lead remain important throughout. The follow-on decision problem is also briefly considered, and surprisingly, we find that the decision to enforce the follow-on or otherwise has no effect on match outcome.

*Keywords*: cricket; multinomial logistic regression; strategy.

## Introduction

The first official international test match was played between England and Australia in 1877 at Melbourne Cricket Ground, Australia. A rule of the contest was that it be continued until either a team won or both teams had completed each of their two allotted innings. Thus, there was no time limit on the duration of the match. In the event, Australia won the match by 45 runs on the fourth day. Unlimited-time matches were the norm in the early years of test cricket and matches could continue for several days. Even with unlimited time, a win for a team was not guaranteed as the scores could be tied after two innings each, although a tied result has occurred only twice in the history of the game. The first fixed-time test match was played over three days between England and Australia in 1880 at the Oval, England—England won the match by 5 wickets. Subsequently, the durations of matches varied; some were unlimited, some were fixed at 3 or 4 days, with match durations agreed beforehand by the competing teams. The last ever unlimited-time test match was the fifth test (of a series of 5) between South Africa and England at Durban, South Africa, in 1939. A draw was agreed between the captains after 9 days of play. England in their final innings required 696 to win. By the end of the match they had reached 654 for 5; this is the highest fourth innings score in test history. The next day, England sailed home from Durban.

In unlimited-time test matches, a declaration of an innings was rare. There was no reason to declare an innings. Each team had unlimited time, as the description indicates, to chase the target or to bowl out their opponents twice. In the 99 unlimited-time matches played in test cricket history, only 9 innings were declared. Andrew Stoddart of England was the first team captain to declare an innings; this happened at Lord's cricket ground in 1893. This was a three-day match. On the third day at lunch England were 234 runs for 8 wickets in their second innings. A drizzling rain prevented play immediately after lunch. Shortly afterwards England declared their second innings, setting Australia 300 runs to win in 3 hours and 45minutes play. However, no play was possible and the game ended as a draw (see Wisden, 2010).

At the point of declaration, the batting team forfeits the remainder of its innings, and requires the opposition to commence their next innings. Since the fixed-duration format was universally adopted in June 1939, the declaration has become an important aspect of the game. To win a test match, a team needs to bowl out the opponents twice within the time limit of five days. If the

leading team bats for a long time, then a draw becomes more likely. For this reason, if a batting team is in a strong position in an innings then they may declare their innings in order to optimize the match outcome from their point of view. In the match between Sri Lanka and India at Khettarama Stadium, Colombo, in 1997, Sri Lanka in the second innings scored 952 runs—the highest ever innings total in a test cricket match. However, the match was drawn. Sri Lanka wasted a match winning opportunity by not declaring, in what was the first test match of a two-match series. Of course, setting the highest ever total was some compensation.

A decision regarding the timing of a declaration is arguably the most critical decision in the game, in terms of the effect the decision can have on match outcome. However, declarations are not universal. Between 1877 and 2007, 1856 test matches were played, and among these, 836 innings were declared. Therefore, the captain of a batting team makes a decision to declare in every second test match. That the decision made can have a profound effect is exemplified by the following examples. In the second test match between India and South Africa at Eden Gardens, Kolkata in 2010, South Africa scored 296 runs in 85 overs (approximately 1 day) in their first innings. In reply, India declared their first innings at 643/6 runs in 153 overs, leading by 347 runs. India thus gave their bowlers two days to dismiss South Africa in their second innings. South Africa scored 296 runs all out and lost the game. It should be pointed out here that broadly speaking there are 90 overs bowled in a day, and six balls bowled in an over, so that a test match can comprise 2700 balls, each ball being an opportunity for a batsmen to hit the ball and to score runs. The exact figure can vary for a number of reasons; matches may be won within five days; bowlers may get through their overs at a slow or a fast rate; weather and light conditions may bring an early end to play on particular days; particular balls may be deemed to be illegal and therefore re-bowled (MCC, 2010).

The effect of the declaration was also apparent in the first test match between Australia and New Zealand at Basin Reserve, Wellington in 2010. Australia batted first and declared their first innings at 459/4 runs in 131 overs—a very modest total for a first innings declaration. New Zealand replied with 157 runs all out in their first innings. New Zealand, trailing by 302 runs, were forced the follow-on. New Zealand were bowled out for 407 runs in 134.5 overs. Australia easily reached the target of 106 runs on the fifth day and won the match by 10 wickets. In the circumstances, Australia were right to declare their first innings. On the other hand, in the third and final test match between India and England at Kennington Oval, England in 2007, India declared their second innings at 180 runs in 58 overs and set a final innings target of 500 runs with 110 overs remaining. England scored 369/6 in 110 overs by the end of the five days and the match was drawn. The news media at the time were critical of the Indian captaincy both for the late declaration and the slow run-rate in the third innings. However, leading the series 1-0 at the time, a draw was sufficient for India to win the series. Therefore, one would have expected them to set a very conservative target, putting themselves in a position from which they would be very unlikely to lose.

These examples demonstrate that declaration decisions can arise in each of the first, second and third innings in a match. Scarf et al. (2005, 2008) have considered the timing of a third innings declaration. Their model is based on the target set and overs-remaining at the end of the third innings. Scarf, Shi and Akhtar (2010) further consider batting tactics during the third innings. We re-analyse the third innings declaration problem here, using a larger dataset, and consider new explanatory variables: average run rate over the first two innings, and team strengths. In addition, we consider the timing of first and second innings declarations. We also consider the decision regarding the enforcement of a follow-on. By way of aside, a fourth and final innings would never

be declared as such a declaration can, by definition, only result in a loss for the declaring team or a tied game.

Declaration strategy in the second innings is similar to the declaration strategy in third innings (if a large lead is obtained and declaration delayed, a draw may become more likely), but also different. If the opponents cover the lead and put on a reasonable target then it will be difficult to chase this target. This is because typically batting in the fourth innings is much more difficult than in the second innings. Declaration in the first innings is a more contentious issue. If the batting team has a good total, it has to decide between either obtaining as large a lead as possible or obtaining a large lead quickly. The decision is not straightforward because the batting captain needs to take account of a number of criteria when setting the first innings score and time at which it is set. Firstly, the captain should keep in mind the time required to bowl out the opponents twice. Secondly, he needs to predict the batting, bowling and weather conditions in the remaining three innings. He will also need to consider his attitude to the risk of losing.

The follow-on decision is somewhat different, although the modelling approach we use is similar. As we have already stated, in a test match each team has two innings. These are taken in turn. However, if the team who bat second trail by 200 runs or more at the end of second innings, they can be required by the opposing captain to bat again immediately. If they bat again immediately, then a follow-on is said to have occurred, and the third innings in the match is played, "out-of-turn". Thus, if each team has batted once, and the team who bat first have a lead of 200 runs or more, then the captain of this leading team has the following decision to make: should he enforce the follow-on or not. We model this decision problem. This decision problem is the subject of debate. This is because typically test pitches deteriorate with the progress of the match, and captains hesitate to make a decision. If the leading team enforces the follow-on, and if the opposition cover the lead and establish a reasonable target, it will be difficult to chase this target. As mentioned above, batting is generally much more difficult in the fourth innings than in earlier innings. On other hand, if the leading team want to increase their existing lead without enforcing follow-on, then they may have less time to bowl out the opponents, and a draw may then follow. Throughout the history of the game, only three matches have been lost by the team enforcing a follow-on. For this reason, we consider a match outcome with two categories (win, draw) in our analysis.

Throughout this paper, we take a quantitative approach, and consider how match outcome probabilities, forecast from covariates that measure to an extent the state of a match, vary with the timing of declaration. Match outcome probabilities are modelled using multinomial regression, with a win, draw, or loss response, and explanatory variables or covariates relating to match state, at the end of the third, second and first innings in turn. These covariates include the score or lead, overs-remaining or used, run-rate, a home advantage factor, and surrogates for the state of the pitch and the pre-match strengths of teams. We investigate how the covariate effects vary from innings to innings. That is, we determine how the set of important covariates changes over the course of a match and how the total strength of the covariate effect varies over the course of a match. Tables of match outcome probabilities (given match position) act as the declaration decision tool.

There is some other published work on test cricket that is related to ours here, the closest of which is perhaps that of Brooks et al. (2002) who use an ordinal response model to predict test match outcomes given the batting and bowling strengths of teams. Other decision problems, that arguably have a less profound effect on match outcome, have been modelled, for example Clarke and Norman (2003) on the use of night-watchmen. Clarke (1998) provides a useful review of the area. Target setting has been considered in one-day matches e.g. Preston and Thomas (2000).

However, test matches are different because, in one-day matches, there is no notion of playing out the time remaining for a draw.

The remainder of the paper is organised as follows. In the next section we briefly describe the dataset we use on test match outcomes. The modelling approach is described. The results of the model fitting are then presented, focussing on questions regarding which explanatory variables are important at each stage of the match, how strong is their influence, and how knowledge of the match outcome given the match position or state might be used to consider the declaration decision. We conclude with a discussion of the limitations of our work, and opportunities for further development.

**Data description**

Data were collected from the Wisden website (Wisden, 2010). Many forms of information can be found here including, for example, ball-by-ball commentary. An extract of the data we use for declaration modelling is presented in table 1. For the third innings analysis, 301 matches from Dec 1997 to Dec 2007 were used. These are all test matches in the period, excluding two triangular series, in which a final innings target was established and where rain did not affect the match in the fourth innings. For the analysis of the first and second innings 391 international test matches from December 1997 to December 2007 were used. The addition of 90 matches is the result of relaxing the condition for the setting of a final innings target, although a number of matches with rain-affected second and third innings were excluded. For the follow-on analysis, we are interested only in those matches where the first innings batting team had a lead of at least 200 runs at end of the second innings. There were only 85 matches from Jan 1988 to Feb 2009 in this category, excluding matches that were rain affected after the end of second innings.

The values of covariates of interest can be calculated from table 1, with the exception of team strengths. To measure team strength, we calculate a win percentage for each team in each year, this win percentage being the ratio of the number of matches won in previous ten years to the number of matches played in the previous ten years. Thus in year *i,* the win percentage for team A is #(matches won by A in years *i*-1,..,*i*-10)/#(matches played by A in years *i*-1,..,*i*-10). These win percentages are calculated for each of the ten test playing nations. We use win percentage difference (W%D), the difference in the win percentages of the competing teams, as a covariate in the match outcome model. One might use the International Cricket Council (ICC, 2010) official ratings as a measure of team strength. These are published monthly. However they are not available prior to June 2003.

**Modelling the match outcome**

Scarf et al. (2005) describe a model to explain match outcome probabilities given the position at the end of the third innings. We briefly review their findings here. They used nominal logistic regression to model the multinomial response (win, draw, loss) as a function of match and end of third innings covariates. The match outcome (win, draw, lose) is denoted by (1, 0, -1), covariates by $X$ and taking a draw (0) as a reference category. Nominal logistic regression assumes

$$
\left.
\begin{aligned}
&Y \sim MN(p_1, p_0, p_{-1}; \textstyle\sum p_i = 1),\\
&p_1 = \exp(\alpha_1 + \beta_1^T X)/\{1 + \exp(\alpha_1 + \beta_1^T X) + \exp(\alpha_{-1} + \beta_{-1}^T X)\},\\
&p_0 = 1/\{1 + \exp(\alpha_1 + \beta_1^T X) + \exp(\alpha_{-1} + \beta_{-1}^T X)\},\\
&p_{-1} = \exp(\alpha_{-1} + \beta_{-1}^T X)/\{1 + \exp(\alpha_1 + \beta_1^T X) + \exp(\alpha_{-1} + \beta_{-1}^T X)\}.
\end{aligned}
\right\}
$$

where $p_1$ and $p_{-1}$ represent the probability of a win and a loss, and depend on the covariate $X$ in different ways—through $\beta_1$ and $\beta_{-1}$ respectively.

Table 1. Test match data (extract of 391 test matches, Dec 1997 to Dec 2007). Variables included: match date, teams, venue; runs scored in first three innings; target established at end of third innings; lead obtained at end of second innings; runs scored in the first innings; third, second and first innings declaration indicator; home factor; overs used in the first innings; and estimated overs-remaining at start of third and fourth innings (calculated on basis of 90 overs per day). Match result recorded as 1 (win), 0 (draw), -1 (loss) from point of view of teams set the target, lead or score. Other variables not shown: overs bowled in first; second and third innings; win percentage differences; toss won by each team (Y/N).

| | | | | | | | Third innings | | | | | Second innings | | | | First innings | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Date | Home | Away | Venue | 1st Innings total | 2nd Innings total | 3rd innings total | Home factor | Target set | Declaration Y=1, N=0 | Overs-remaining | Result | Lead | Overs-reaming | Home factor | Result | Score | Overs used | Home factor | Result |
| 13/02/1998 | WI | E | Port of Spain | 159 | 145 | 210 | 0 | 225 | 0 | 207 | -1 | -14 | 307 | 1 | 1 | 159 | 68 | 0 | -1 |
| 27/02/1998 | WI | E | Georgetown | 352 | 170 | 197 | 1 | 380 | 0 | 152 | 1 | -182 | 231 | 0 | -1 | 352 | 128 | 1 | 1 |
| 12/03/1998 | WI | E | Bridgetown | 403 | 262 | 233 | 0 | 375 | 1 | 109 | 0 | -141 | 185 | 1 | 0 | 403 | 154 | 0 | 0 |
| 30/01/1998 | A | SA | Adelaide | 517 | 350 | 193 | 0 | 361 | 1 | 109 | 0 | -167 | 206 | 1 | 0 | 517 | 166 | 0 | 0 |

As there is no system of points in test cricket, match outcome categories do not form a natural order. Also, for example, for the team batting last the difference between losing and drawing is likely to be more dependent on the overs-remaining than on the target faced; the difference between winning and drawing, on the other hand, is likely to depend on both the overs-remaining and the target faced. In this way, the target and overs-remaining influence the match outcome categories in a non-cumulative way. Therefore, it makes sense to regard match outcome categories as nominal.

Nominal multinomial logistic regression is used throughout for describing match outcome probabilities given the end-of-innings position. For the analysis of the follow-on, we use binary logistic regression as the response variable is then just win or draw. Other multinomial regression models might be considered such as probit regression. We would however expect the results to be very similar. This is because the logistic and probit functions are very nearly linearly related over the interval $0.1 \leq p \leq 0.9$ (McCullogh and Nelder, 1989, p.109), and so it is difficult to distinguish the two functions based on goodness-of-fit. Probit regression has been used in football match prediction (e.g. Koning, 2000; Dobson and Goddard, 2003).

To assess model fit, we use the Akaike information criterion (AIC) (Sakamoto et al, 1986) and Nagelkerke $R^2$. This latter statistic is a modified form of Cox and Snell's pseudo $R^2$, and gives

information about the explanatory power of the covariates in the model (Nagelkerke, 1991). Cox and Snell's Pseudo $R^2$ is given by

$$R^2_{CS} = 1 - \exp\{(-2(L_1 - L_0)/n\}$$

where $L_1$ represents the log-likelihood for the model with covariates, $L_0$ represents log-likelihood for the model with no covariates, and $n$ represents number of observations. Nagelkerke's $R^2$ is given by

$$R^2 = \frac{1 - \exp\{(-2(L_1 - L_0)/n\}}{\{1 - \exp(2L_0)/n\}}$$

The maximum value of Cox and Snell's $R^2$ is $1 - \exp(2L_0)/n$ and therefore Nagelkerke's (modified) $R^2$ can vary from 0 to 1. Broadly speaking, Nagelkerke's $R^2$ for a generalized linear model is the percentage of variability in the outcome that is explained by the covariates in the model.

## Model fitting results

*Third innings model*

Outline model statistics for various fitted models are shown in table 2. T represents the target set for the 4th innings; OR, represent overs-remaining in the match assuming 90 overs per day; D, is used as a declaration indicator. This covariate is included for interest, but cannot be used in a model to provide declaration decision support. This is because once the final innings has commenced, a declaration in the third innings ought not to improve a team's chances; the declaration indicator is instead capturing other factors not represented in the data. With the run-rate in the first two innings, $RR_{12}$, we are attempting to capture the quality of the batting conditions. A covariate that represents the condition of a pitch would be of interest and further work and perhaps more data would be beneficial to consider this. Estimates for the highlighted model are shown in table 3. This table indicates that the loss-draw probability ratio depends strongly on both current lead and overs-remaining. However, the win-draw probability ratio depends strongly on overs-remaining only.

Table 2. Results of model fitting multinomial logistic regression model to 301 test match outcomes for various sets of predictors: log-likelihood, AIC and Nagelkerke $R^2$. Covariates here are T, target; OR, overs-remaining; $RR_{12}$, average run-rate of first two innings; W%D, win percentage difference; D, declaration indicator.

| Model | Parameters | Log-Likelihood | AIC | Nag.$R^2$ (%) |
|---|---|---|---|---|
| T+OR+$RR_{12}$+W%D+$T^2$ | 12 | -132.6 | 289.2 | 80.0 |
| **T+OR+$RR_{12}$+W%D** | **10** | **-133.8** | **287.6** | **79.7** |
| T+OR+$RR_{12}$+W%D+$OR^2$ | 12 | -132.2 | 288.4 | 80.1 |
| T+OR+$RR_{12}$+W%D+H | 12 | -131.9 | 287.8 | 80.1 |
| T+OR+$RR_{12}$+W%D+D | 12 | -127.7 | 273.9 | 81.7 |
| T+OR+$RR_{12}$ | 8 | -138.9 | 293.7 | 78.5 |
| T+OR+W%D | 8 | -137.7 | 291.4 | 78.8 |
| T+OR (ordinal) | 4 | -198.1 | 404.2 | 61.3 |
| T+OR | 6 | -142.1 | 296.3 | 78.2 |

Table 3. Fitted parameter estimates for minimum AIC logistic regression model (nominal) with covariates lead (T), overs-remaining (OR), run-rate in first two innings ($RR_{12}$), and win percentage difference (W%D), with standard errors and p-values. 301 test matches (Dec 1997 to Dec 2007**).**

| | | coefficient | s.e | p-value |
|---|---|---|---|---|
| win/draw (1/0) | Intercept | -4.8947 | 1.6479 | 0.003 |
| | overs-remaining (OR) | 0.0523 | 0.0088 | 0.000 |
| | target (T) | -0.0068 | 0.0033 | 0.041 |
| | run-rate$_{12}$ ($RR_{12}$) | 0.7121 | 0.4945 | 0.150 |
| | win%diff (W%D) | 0.0086 | 0.0124 | 0.491 |
| loss/draw (-1/0) | Intercept | -1.7028 | 1.9506 | 0.383 |
| | overs-remaining (OR) | 0.0540 | 0.0093 | 0.000 |
| | target (T) | -0.0297 | 0.0039 | 0.000 |
| | run-rate$_{12}$ ($RR_{12}$) | 1.6568 | 0.6204 | 0.008 |
| | win%diff (W%D) | -0.0291 | 0.0155 | 0.060 |

Using the model, highlighted in table 2, we can calculate the probability of win, draw and loss given the end of third innings for different situations (table 4). Our model probabilities are based on target, overs-remaining, run-rate in the first two innings, and the difference in win percentage. Other factors will influence the match result such as the bowling and batting strength, and the weather conditions. These are not quantified in our data. The state of the series will influence the captain's attitude to risk. A captain would be expected to take account of these factors when considering a possible declaration. Note that, for a fixed number of overs-remaining, the win probability increases to a peak and then decreases—if a very large target is set, the team batting last will not attempt to play for a win and a draw becomes more likely.

Table 4. Win (W), draw (D) and loss(L) probabilities at end of third innings position for the team batting third as a function of target set and overs-remaining. $RR_{12}$=average value, W%D=0.

| | | | Overs-remaining | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 60 | 80 | 100 | 120 | 140 | 160 |
| Target set | 200 | W | 0.116 | 0.138 | 0.146 | 0.146 | 0.143 | 0.139 |
| | | D | 0.281 | 0.118 | 0.044 | 0.015 | 0.005 | 0.002 |
| | | L | 0.603 | 0.744 | 0.811 | 0.839 | 0.852 | 0.859 |
| | 250 | W | 0.165 | 0.256 | 0.314 | 0.336 | 0.340 | 0.336 |
| | | D | 0.562 | 0.306 | 0.132 | 0.050 | 0.018 | 0.006 |
| | | L | 0.273 | 0.438 | 0.555 | 0.614 | 0.643 | 0.658 |
| | 300 | W | 0.158 | 0.310 | 0.465 | 0.559 | 0.598 | 0.607 |
| | | D | 0.759 | 0.521 | 0.274 | 0.116 | 0.043 | 0.016 |
| | | L | 0.083 | 0.168 | 0.261 | 0.325 | 0.359 | 0.377 |
| | 350 | W | 0.127 | 0.283 | 0.498 | 0.678 | 0.774 | 0.811 |
| | | D | 0.852 | 0.668 | 0.413 | 0.197 | 0.079 | 0.029 |
| | | L | 0.021 | 0.049 | 0.089 | 0.125 | 0.148 | 0.160 |
| | 400 | W | 0.095 | 0.229 | 0.450 | 0.682 | 0.831 | 0.899 |
| | | D | 0.900 | 0.759 | 0.524 | 0.278 | 0.119 | 0.045 |
| | | L | 0.005 | 0.013 | 0.026 | 0.040 | 0.050 | 0.056 |
| | 450 | W | 0.070 | 0.176 | 0.377 | 0.628 | 0.819 | 0.917 |
| | | D | 0.929 | 0.821 | 0.616 | 0.360 | 0.165 | 0.065 |
| | | L | 0.001 | 0.003 | 0.007 | 0.012 | 0.016 | 0.018 |

*Second and first innings models*

Multinomial logistic regression models considered for the second and first innings are given in tables 5 and 6 respectively. Estimates for the highlighted models are in tables 7 and 8. In the second innings model, lead, overs-remaining, home factor, and win percentage difference were found to be important (based on AIC and Nagalkerke $R^2$). Results suggest that the win percentage difference (the difference in the pre-match strengths of teams) has a bigger effect on match outcome than in the third innings model, and likewise for the home factor, H. In contrast, by the end of third innings, our analysis suggests that the home factor has little or no effect. This is essentially because later in the match, the lead and overs-remaining will reflect to an extent the home advantage.

By using model estimates we can find the probability of win, draw and loss for given covariate values (lead, overs-remaining, home factor and pre-match strength). These probabilities are illustrated in table 9 for various lead and overs-remaining, at home ground (H=1) and with W%D=0.

Table 5. Results of second innings model fitting: log-likelihood, AIC and Nagelkerke $R^2$. Covariates considered are L, lead; OR, overs-remaining; W%D, win percentage difference; H, home factor.

| Model | Parameter | Log likelihood | AIC | Nag. $R^2$ *(%)* |
|---|---|---|---|---|
| L+ OR | 6 | -274.8 | 561.5 | 54.0 |
| L+OR+$L^2$ | 8 | -271.0 | 558.1 | 55.1 |
| L+OR+$OR^2$ | 8 | -274.2 | 564.4 | 54.2 |
| L + OR+H | 8 | -272.7 | 561.4 | 54.6 |
| L + OR+ W%D | 8 | -261.7 | 539.4 | 57.9 |
| L + OR+ H+ W%D+$OR^2$ | 12 | -257.5 | 539.1 | 59.1 |
| **L + OR+ H+ W%D+ $L^2$** | **12** | **-254.1** | **532.2** | **60.1** |
| L + OR+ H+ W%D | 10 | -258.2 | 635.3 | 59.0 |
| L + OR(ordinal) | 4 | -295.1 | 598.2 | 47.4 |

Table 6. Results of first innings model fitting: log-likelihood, AIC and Nagelkerke $R^2$. Covariates are S, score in first innings; OV, overs used in first innings; W%D, win percentage difference; H, batting team at home (=1)

| Model | Parameter | Log likelihood | AIC | Nag. $R^2$ *(%)* |
|---|---|---|---|---|
| S+OV | 6 | -342.6 | 697.1 | 28.7 |
| S+OV+$S^2$ | 8 | -340.3 | 696.6 | 29.7 |
| S+OV+$OV^2$ | 8 | -342.1 | 700.3 | 28.9 |
| S+OV +H | 8 | -337.7 | 691.3 | 30.9 |
| S+OV+ W%D | 8 | -312.8 | 641.7 | 40.9 |
| S+OV+ H+W%D+ $OV^2$ | 12 | -303.9 | 630.7 | 44.2 |
| S+OV+ H+W%D+ $S^2$ | 12 | -302.6 | 629.2 | 44.7 |
| **S+OV+ H+W%D** | **10** | **-304.4** | **628.7** | **44.0** |
| S+OV(ordinal) | 4 | -364.3 | 736.6 | 18.6 |

Table 7. Fitted parameter estimates for second innings minimum AIC logistic regression model (nominal) with covariates lead (L), overs-remaining (OR), home factor (H) and win percentage difference (W%D), with standard errors and p-values, 391 test matches (Dec 1997 to Dec 2007).

| | | Coefficient | s.e | p-value |
|---|---|---|---|---|
| win/draw (1/0) | Intercept | -3.1133 | 0.7977 | 0.000 |
| | overs-remaining (OR) | 0.0141 | 0.0032 | 0.000 |
| | lead (L) | 0.0071 | 0.0014 | 0.000 |
| | home factor (H) | 0.8057 | 0.3528 | 0.022 |
| | win%diff (W%D) | 0.0308 | 0.0093 | 0.001 |
| | lead*lead ($L^2$) | $1.3\times10^{-5}$ | $5.9\times10^{-6}$ | 0.018 |
| loss/draw (-1/0) | Intercept | -4.1752 | 0.8024 | 0.000 |
| | overs-remaining (OR) | 0.0188 | 0.0033 | 0.000 |
| | lead (L) | -0.0056 | 0.0019 | 0.004 |
| | home factor (H) | 0.0322 | 0.3392 | 0.924 |
| | win%diff (W%D) | -0.0142 | 0.0092 | 0.123 |
| | lead *lead ($L^2$) | $-3.7\times10^{-6}$ | $6.3\times10^{-6}$ | 0.553 |

Table 8. Fitted parameter estimates for first innings minimum AIC logistic regression model (nominal) with covariates score (S), overs used (OV), home factor (H) and win percentage difference (W%D), with standard errors and p-values, 391 test matches (Dec 1997 to Dec 2007).

| | | Coefficient | s.e | p-value |
|---|---|---|---|---|
| win/draw (1/0) | Intercept | 1.9116 | 0.6076 | 0.002 |
| | overs used (OV) | -0.0135 | 0.0077 | 0.079 |
| | score (S) | 0.0003 | 0.0020 | 0.880 |
| | win%diff (W%D) | 0.0221 | 0.0087 | 0.011 |
| | home factor(H) | 0.2422 | 0.3142 | 0.441 |
| loss/draw (-1/0) | Intercept | 5.1360 | 0.6229 | 0.000 |
| | overs used (OV) | -0.0249 | 0.0085 | 0.003 |
| | score (S) | -0.0035 | 0.0023 | 0.130 |
| | win%diff (W%D) | -0.0361 | 0.0091 | 0.000 |
| | home factor (H) | -0.8981 | 0.3375 | 0.008 |

In table 6, for the end of first innings position, we can observe from different models that, largely, pre-match strengths and the home factor play an equal role in explaining outcomes when compared to runs scored and overs used. By the end of the third innings, the effect of lead and overs-remaining dominate. From table 9, it is clear that the probability of win is increasing with increasing lead and overs-remaining. Therefore, the batting team captain should consider both of these factors and the home factor and team strengths in the declaration decision. Table 10 shows the win, draw and loss probabilities for the end of first innings position as a function of score and overs used. Note that the probability of a win initially increases with score and high run-rate while a slow run-rate increases the probability of a draw, as expected. One has to look diagonally, left to right down this table to see winning probability diminish with increasing lead—this is because as the score increases, on average, the overs-remaining decreases. This decreasing win probability is expected since to bowl out the opposition twice one needs sufficient time.

Table 9. Win (W), draw (D) and loss (L) probabilities at end of second innings position for team batting second as a function of lead established and overs-remaining. H=1, W%D=0

| | | | Overs-remaining | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 90 | 105 | 120 | 135 | 150 | 165 | 180 | 195 |
| Lead | 100 | W | 0.457 | 0.489 | 0.537 | 0.583 | 0.626 | 0.665 | 0.700 | 0.730 |
| | | D | 0.516 | 0.481 | 0.427 | 0.375 | 0.326 | 0.280 | 0.239 | 0.201 |
| | | L | 0.027 | 0.030 | 0.036 | 0.042 | 0.048 | 0.055 | 0.062 | 0.069 |
| | 150 | W | 0.591 | 0.623 | 0.668 | 0.709 | 0.746 | 0.779 | 0.807 | 0.831 |
| | | D | 0.394 | 0.361 | 0.313 | 0.269 | 0.229 | 0.194 | 0.163 | 0.135 |
| | | L | 0.015 | 0.016 | 0.019 | 0.022 | 0.024 | 0.027 | 0.030 | 0.033 |
| | 200 | W | 0.712 | 0.752 | 0.788 | 0.819 | 0.846 | 0.869 | 0.889 | 0.905 |
| | | D | 0.281 | 0.240 | 0.204 | 0.171 | 0.143 | 0.119 | 0.098 | 0.081 |
| | | L | 0.007 | 0.008 | 0.009 | 0.010 | 0.011 | 0.012 | 0.013 | 0.014 |
| | 250 | W | 0.829 | 0.856 | 0.880 | 0.900 | 0.916 | 0.930 | 0.941 | 0.951 |
| | | D | 0.168 | 0.140 | 0.117 | 0.097 | 0.080 | 0.065 | 0.054 | 0.044 |
| | | L | 0.003 | 0.003 | 0.003 | 0.004 | 0.004 | 0.005 | 0.005 | 0.005 |
| | 300 | W | 0.911 | 0.926 | 0.939 | 0.950 | 0.959 | 0.966 | 0.972 | 0.977 |
| | | D | 0.088 | 0.073 | 0.060 | 0.049 | 0.040 | 0.033 | 0.026 | 0.022 |
| | | L | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.002 | 0.002 | 0.002 |
| | 350 | W | 0.958 | 0.966 | 0.972 | 0.977 | 0.981 | 0.985 | 0.987 | 0.990 |
| | | D | 0.042 | 0.034 | 0.028 | 0.022 | 0.018 | 0.015 | 0.012 | 0.010 |
| | | L | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.001 |
| | 400 | W | 0.982 | 0.986 | 0.988 | 0.990 | 0.992 | 0.994 | 0.995 | 0.996 |
| | | D | 0.018 | 0.014 | 0.012 | 0.009 | 0.008 | 0.006 | 0.005 | 0.004 |
| | | L | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

Table 10. Win (W), draw (D) and loss (L) probabilities at end of first innings position for the team batting first as a function of score established and overs used. Home =1, W%D=0

| | | | Overs used | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 90 | 120 | 150 | 180 | 210 | 240 | 270 |
| First innings score | 450 | W | 0.537 | 0.531 | 0.493 | 0.428 | 0.351 | | |
| | | D | 0.183 | 0.271 | 0.377 | 0.491 | 0.603 | | |
| | | L | 0.281 | 0.197 | 0.130 | 0.080 | 0.047 | | |
| | 500 | W | 0.565 | 0.552 | 0.507 | 0.438 | 0.357 | 0.277 | |
| | | D | 0.190 | 0.278 | 0.382 | 0.495 | 0.604 | 0.702 | |
| | | L | 0.245 | 0.170 | 0.111 | 0.068 | 0.039 | 0.022 | |
| | 550 | W | 0.592 | 0.571 | 0.520 | 0.446 | 0.362 | 0.281 | 0.209 |
| | | D | 0.196 | 0.283 | 0.386 | 0.497 | 0.605 | 0.701 | 0.782 |
| | | L | 0.212 | 0.145 | 0.094 | 0.057 | 0.033 | 0.018 | 0.010 |
| | 600 | W | 0.616 | 0.589 | 0.532 | 0.454 | 0.368 | 0.284 | 0.212 |
| | | D | 0.201 | 0.288 | 0.389 | 0.498 | 0.604 | 0.700 | 0.780 |
| | | L | 0.183 | 0.124 | 0.079 | 0.048 | 0.028 | 0.015 | 0.008 |
| | 650 | W | | 0.604 | 0.542 | 0.461 | 0.373 | 0.288 | 0.214 |
| | | D | | 0.291 | 0.391 | 0.498 | 0.604 | 0.699 | 0.779 |
| | | L | | 0.105 | 0.067 | 0.040 | 0.023 | 0.013 | 0.007 |
| | 700 | W | | 0.618 | 0.552 | 0.468 | 0.378 | 0.292 | 0.217 |
| | | D | | 0.293 | 0.392 | 0.498 | 0.603 | 0.697 | 0.777 |
| | | L | | 0.089 | 0.056 | 0.034 | 0.019 | 0.011 | 0.006 |
| | 750 | W | | | 0.561 | 0.474 | 0.383 | 0.296 | 0.220 |
| | | D | | | 0.392 | 0.497 | 0.601 | 0.696 | 0.775 |
| | | L | | | 0.047 | 0.028 | 0.016 | 0.009 | 0.005 |

An aim of this paper is to support declaration decision-making. For this end, indicative results are presented in the form of tables of probabilities for each innings. These describe how the probabilities of outcomes (win, draw, loss) vary in different circumstances, and to an extent provide a decisions tool to a team captain. The following examples illustrate how these tables might be used.

*Example 1*: Third test, West Indies vs England 2009 at St John's, Antigua. England declared their second innings at 221 runs, setting a target of 503 runs in 135 overs. West Indies scored 370/9 in 128 overs and the match ended as a draw. The last day of the match was reduced to 83 overs, due to bad light. England were in a very strong position in third innings, but West Indies saved the match and led the 5-match series 1-0. The question is where did England make their mistake? According to our analysis, the declaration came too late. The probability of winning the match was much higher in earlier situations. At the given situation, their win and draw probabilities were 0.79 and 0.21 respectively. The win probability could have been increased to 0.89 if England had set a target of 450 runs with 145 overs-remaining. The draw probability would have decreased to 0.10.

*Example 2*: First test between Sri Lanka and India at Khettarama Stadium, Colombo, in 1997. In reply to India's first innings total of 537/8 declared, Sri Lanka scored 952 runs in the second innings—the highest ever innings total in a test cricket match. But the match was drawn. Our analysis suggests that Sri Lanka would have had an outside chance to win the match if they had declared their innings (say 150 runs lead) at lunch on fifth day, with a lead of 150. Their win probability at that point was 0.32.

*Example 3*: The fourth test match, West Indies vs England 2004 at St Johns Antigua. West Indies declared the first innings at 751 runs in 202 overs and the match ended as a draw. The West Indies were in a very strong position in the match but they lost the position due to late declaration. Table 10 suggest that 600 runs in 150 overs was a much better position for the declaration than 751 runs in 202 overs.

*The follow-on decision*

Fir the follow-on analysis, we consider binary match outcome (win, draw) categories and fit the binary logistic regression model. Fit statistics for various fitted models are shown in table 11. Estimates for the highlighted model are in table 12. Table 11 indicates that the win probability depends strongly on overs-remaining and home factor only. Lead, pre-match strength and run-rate are also considered in the analysis and found to be unimportant. Further, the follow-on factor does not appear to have a significant effect here in the analysis. This is supported by table 13. The win-draw ratio is the same whether the follow-on is enforced or not.

Figure1 shows the win probability from the fitted model as a function of overs-remaining. The home advantage effect is quite large here, both in terms of how often the home team are in a position to enforce the follow-on (in 54 of the 85 cases, $p$-value of 0.013 against the hypothesis of no home effect), and how the win probability varies with overs-remaining and home factor. This latter effect indicates that when a team is forced to follow-on, they battle harder if they are at home. The rate of change of win probability peaks at 140 overs-remaining when the leading team is the home team, but at 170 overs otherwise. Furthermore, our analysis broadly indicates that overs-remaining is the determining factor in match outcome. We illustrate this with two further examples.

Table 11. Results of model fitting: log-likelihood, AIC and Nagelkerke $R^2$. Covariates here are L, lead; OR, overs-remaining; $RR_1$, average run-rate of first innings; W%D, win percentage difference; F, follow-on indicator; H, home factor. 85 matches (Jan 1988 to Feb 2009)

| Model | parameters | Log likelihood | AIC | Nag. $R^2$ *(%)* |
|---|---|---|---|---|
| OR | 2 | -31.16 | 66.31 | 40.2 |
| **OR+H** | **3** | -29.31 | **64.61** | **45.1** |
| OR+ H+L | 4 | -29.30 | 66.59 | 45.1 |
| OR+H+F | 4 | -29.25 | 66.50 | 45.2 |
| OR+H+F+OR*F | 5 | -28.18 | 66.36 | 48.0 |
| OR+H+$RR_1$ | 4 | -29.02 | 66.05 | 45.8 |
| OR+H+W%D | 4 | -29.22 | 66.44 | 45.3 |
| OR+H+L+L*F | 5 | -29.22 | 70.43 | 45.3 |

Table 12. Results of second innings model fitting: log-likelihood, AIC and Nagelkerke $R^2$. Covariates considered are OR, overs-remaining; H, home factor.

| | | Coefficient | s.e | p-value |
|---|---|---|---|---|
| win/draw (1/0) | Intercept | -5.3669 | 1.6820 | 0.001 |
| | overs-remaining (OR) | 0.0307 | 0.0086 | 0.000 |
| | home factor (H) | 1.2549 | 0.6626 | 0.058 |

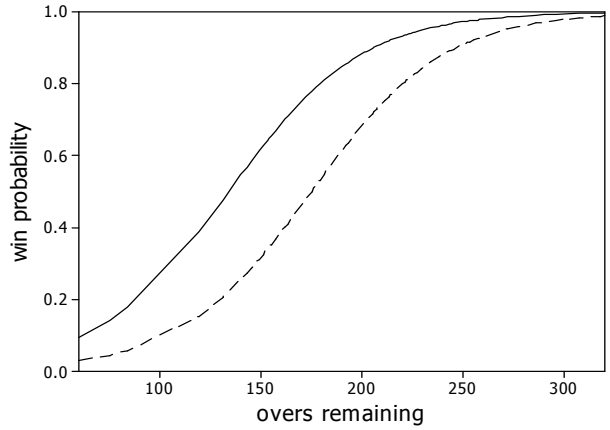


Figure 1. Win probability for the team batting first against overs-remaining. ___ H=1; _ _ _ _ H=0.

Table 13. Contingency table for match outcome (win, draw) and follow-on decision. p-value for Fisher's exact test of association equals 1.

| | | follow-on | | total |
|---|---|---|---|---|
| | | no | yes | |
| results | draw | 4 | 14 | 18 |
| | win | 17 | 50 | 67 |
| | total | 21 | 64 | 85 |

*Example* 4: The third test match between West Indies vs England 2009 at St John's, Antigua. England batted first and scored 566 runs in 166 overs. West Indies scored 285 runs in 90 overs in their first innings. England did not enforce the follow-on and set a target of 503 runs in 135 overs. West Indies scored 370 runs for 9 wickets in the fourth innings and saved the match.

*Example* 5: The first test match between Bangladesh vs England 2010 at Chittagong. England batting first again scored 599 runs in 139 overs. Bangladesh scored 296 runs in 91 overs. England made the same decision not to enforce the follow-on and set a target 513 runs with 165 overs remaining. Bangladesh were all out for 331 runs in 124 overs, and England won the match.

England faced the same situation in both matches but they achieved different results. In the first of these examples above (example 4), however, there were fewer overs-remaining at the end of second innings and England's win probability was 0.62. In the second example, England's win probability was 0.77 at end of second innings. In media reports following the West Indies-England match, England were blamed for not enforcing the follow-on. Our analysis suggests the reason for the draw lay elsewhere—with the lack of overs-remaining—the enforcement of the follow-on or otherwise was irrelevant.

**Discussion**

The purpose of this paper is to model match outcome at the end of each innings in test cricket. Looking at the end of innings positions has two benefits. Firstly, some progress can be made with the quantitative analysis of the problem. Secondly, the models can provide decision support to a batting captain who is considering a declaration. In essence, the models provide information about how the match outcome (probability of win, draw and loss) changes according to the match situation, and we consider the optimal score and ideal time to declare an innings. It is our view that the methodology described has the potential for implementation as a decision aid.

Where the match outcome models are most interesting is in situations in which they show win probabilities at first increasing with the size of a lead or target and then later decreasing as the lead or target set increases further still. This is because on average a larger lead or target set implies that the match will be further on in time terms, and there will be fewer overs-remaining in which to dismiss the opposition. Also, teams are less likely to chase a large target and bat defensively as a consequence. Furthermore, the extent to which such probabilities change with changing position, and the absolute values of such changes can be dramatic; for example, when considering a declaration position with a lead of 400 and 120 overs-remaining, a 10% increase in the target set can lead to a 20% decrease in the probability of a win.

When considering the follow-on decision, we find that enforcing the follow-on or otherwise has no effect on match outcome. For analysis of declaration decisions, we investigated a number of covariates in the match outcome models: scores, and hence lead; overs (assuming 90 overs per day are bowled); run-rates in earlier innings in the match; pre-match team strengths (expressed as a win percentage difference over the previous years); and a home factor. Interestingly, the fitted models indicate that the "best" set of covariates given the end of third innings position explain 80 percent of the variability in match outcomes. For the end of second innings position, this drops to 60 percent, and for the end of first innings position to 42 percent. Therefore decisions regarding declarations that are considered early in a match (first and second innings) are subject to much more uncertainty than a later decision (third innings). The nature of the covariates that influence the match outcome

changes as the match progresses. Early in the match, pre-match team strengths have a large effect. This reduces as the match progresses. The home effect appears small and exists only early on. This is because these effects translate into lead and overs-remaining as a match progress, so that later in a match, lead and overs-remaining dominate. Of course other factors are influential, such as the weather and the state of the pitch on the last day. The strength of teams is calculated somewhat crudely, and there may be scope to refine the model to consider short-term and the long-term strength measures.

The models developed here should be viewed as tools for supporting decision-making. A captain would be expected to take account of a multitude of factors that are not, and could not be captured in the data that we analyse. Furthermore, as there is no single decision criterion, it is necessary to scrutinize outcome probabilities, and this in itself will be a challenge for the implementation of decision support.

**References**


Brooks R D, Faff R W and Sokulsky D (2002). An ordered response model of test cricket performance. Applied Economics. 34: 2353-2365.

Clarke S R (1998). Test statistics. In: Bennett R (ed). Statistics in Sport (1998). London: Arnold, pp 83–103.

Clarke S R and Norman J M (2003). Dynamic programming in cricket: choosing a night watchman: Journal of the Operation Research Society. 54: 838-845.

Dobson S and Goddard J (2003). Persistence in sequences of football match results: a Monte Carlo analysis: European Journal of Operational Research. 148, 247–256.

ICC (2010) Reliance Mobile Test Championship. http://icc-cricket.yahoo.net/match_zone/team_ranking.php, accessed 30.3.2010.

Koning R H (2000). Balance in competition in Dutch soccer. The Statistician. 49, 419–431.

McCullagh P and Nelder J A (1989). Generalized Linear Models. London: Chapman & Hall.

MCC (2010). The laws of cricket. http://www.lords.org/laws-and-spirit/laws-of-cricket/laws/ accessed 01 Feb. 2010

Nagelkerke N J D (1991). A note on a general definition of the coefficient of determination, Biometrika. 78: 691-692

Preston I and Thomas J (2000). Batting strategy in limited overs cricket: Statistician. 49: 95–106.

Sakamoto Y, Ishiguro M and Kitigawa G (1986). Akaike Information Criterion Statistics. Tokyo: KTK Publishing House.

Scarf P A and Shi X (2005). Modelling match outcomes and decision support for setting a final innings target in test cricket: IMA J. Management Mathematics. 16: 161-178.

Scarf P A, Shi X and Akhtar S (2008). Modelling batting strategy in test cricket, In: ECMI Conference Proceedings (in press).

P A Scarf, X Shi and S Akhtar (2010). The distribution of runs scored and batting strategy in test cricket. Salford Business School technical report 336/10 (under review).

Wisden (2010). Test match archives. http://www.stats.cricinfo.com/ci/content/records/307847.html accessed 01 Feb. 2010.