International Journal of Forecasting 32 (2016) 34–43

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

In-play forecasting of win probability in One-Day International cricket: A dynamic logistic regression model

Muhammad Asif[a,b,*], Ian G. McHale[c]

[a] Centre for Sports Business, Salford Business School, University of Salford, UK
[b] Department of Statistics, University of Malakand, KP, Pakistan
[c] School of Mathematics, University of Manchester, UK

ARTICLE INFO

Keywords:
Binary
Dynamic
Regression
Sports
Cricket

ABSTRACT

The paper presents a model for forecasting the outcomes of One-Day International cricket matches whilst the game is in progress. Our 'in-play' model is dynamic, in the sense that the parameters of the underlying logistic regression model are allowed to evolve smoothly as the match progresses. The use of this dynamic logistic regression approach reduces the number of parameters required dramatically, produces stable and intuitive forecast probabilities, and has a minimal effect on the explanatory power. Cross-validation techniques are used to identify the variables to be included in the model. We demonstrate the use of our model using two matches as examples, and compare the match result probabilities generated using our model with those from the betting market. The forecasts are similar quantitatively, a result that we take to be evidence that our modelling approach is appropriate.

1. Introduction

Unlike soccer, American football and tennis, relatively little work has been published on forecasting in cricket. This seems especially strange given that there are known to be huge betting markets for cricket. The work that has been done on forecasting in cricket has largely been concerned with pre-match forecasting. However, in recent times, the growth in the popularity of in-play betting in all sports, where punters place bets during a game (or match), has meant that models that enable forecasts to be made as the game progresses are in high demand. Cricket is a sport that lends itself particularly to in-play betting: unlike soccer, for example, the discrete nature of the game means that bookmakers and punters alike have ample opportunities to be active in markets during the game, and as such, cricket attracts extremely large in-play betting volumes. For example, the total amount bet during a typical major One-Day International (ODI) involving Pakistan or India is in the order of \$1bn (according to a personal communication from a betting industry insider in 2013). In this paper, we present an in-play forecasting model for One-Day International cricket, and use the model to estimate the probability of victory for a team at any moment during a game.

Of course, betting is not the only use of a forecasting model. A forecasting model like that presented here could be used for several purposes. Team coaches may use in-play forecasting probabilities to assess the merits of various different strategies or to analyse player and team performances. In addition, the media could use the model to identify key moments in a match and enhance the television coverage further.

Previous work in One-day International (ODI) cricket has focussed largely on the problem of resetting targets in limited overs cricket following interruptions to play.

* Corresponding author at: Centre for Sports Business, Salford Business School, University of Salford, UK.
E-mail addresses: m.asif609@gmail.com, m.asif@uom.edu.pk (M. Asif).

http://dx.doi.org/10.1016/j.ijforecast.2015.02.005
0169-2070/

The most well-known work in this area is of course that of Duckworth and Lewis (1998, 2004). In fact, the Duckworth–Lewis method, as their work has been named, can be used as a forecasting tool itself, as it essentially predicts the number of runs that are still to be scored in an innings, given the number of balls remaining and the number of wickets lost so far. Several other authors have developed alternatives to the Duckworth–Lewis method, and these can also be used for forecasting. For example, Preston and Thomas (2002) use dynamic programming to estimate the probabilities of different match outcomes at any stage of the innings, and use this forecasting model to revise targets in interrupted ODI cricket matches. Similarly, Carter and Guthrie (2004) investigate the distribution of the runs remaining to be scored in an innings at any given stage of the innings. They then use this distribution to estimate match outcome probabilities, and go on to use these probabilities to revise targets in interrupted limited overs cricket matches.

Some work has been done focussing directly on the problem of forecasting. Brooks, Faff, and Sokulsky (2002) estimate test match outcome probabilities using an ordered response model. Similarly, for test cricket, Scarf and Shi (2005) forecast match outcome probabilities using a multinomial logistic regression model, with the specific aim of helping team management to decide on the most appropriate time to declare in an innings. Following on from Scarf and Shi (2005), Akhtar and Scarf (2012) present a multinomial logistic regression based model for forecasting the results of Test matches in which predictions are made after each session of play. They estimate 15 separate multinomial logistic models that could be used at 15 particular stages of the match (at the end/start of each session). Such an approach allows for the covariates and their estimated coefficients to vary, session by session. The work that is related most closely to ours is that of Bailey and Clarke (2006), who develop a forecasting model for predicting the margin of victory in limited overs cricket before the match begins, then, with the help of the Duckworth and Lewis (1998) method, update these predictions whilst the game is in progress. However, our methodology differs fundamentally from theirs, in that the effect of a covariate on the match outcome is allowed to evolve as the match progresses.

In this paper, we present an in-play forecasting model for ODI cricket. The model produces forecasts of the probabilities of different match outcomes (win or lose), both once the match has begun, and at each stage through the match as it progresses. The model that we adopt is a dynamic logistic regression (DLR) model, in that the parameters (the coefficients of covariates) are allowed to evolve smoothly as the match progresses. To the best of our knowledge, no such approach to the production of forecasts of the outcome while the game is in progress exists in the literature.

Before presenting our model, we first describe the data that we obtained and the possible covariates that will be experimented with in the modelling. Transformations of the independent variables and the motivations for these transformations are also given. In Section 3, we describe the procedure for developing our dynamic logistic regression model. In Section 4, we present the final model specification and results. In Section 5, we present model fit diagnostics and use two example matches to compare our predicted probabilities with those of the betting market. Section 6 concludes with some closing remarks and discusses potential future work.

2. Data and covariates

We obtained ball-by-ball data for ODI matches played between January 2004 and February 2010. The data were collected from the commentary logs on the Espncricinfo website. We did not include matches for which the data were incomplete, or in which one of the teams had played fewer than five matches prior to the time of play, or in which play was interrupted. In total, we fit our model to data from 606 ODI matches.

The data set includes several variables that could potentially be used as covariates. We divide these variables into two categories: *pre-match* covariates, which are measured prior to the start of the match, and *in-play* covariates, which are measured only during play. In the next two subsections, we explain how and why we experiment with certain variables and functions of variables as covariates in our models.

2.1. Pre-match covariates

Pre-match covariates are variables that can be measured prior to the start of the game. There are a number of factors that might affect the probability of a match outcome before the play has commenced, for example, home advantage, winning the toss to decide whether to bat first or second, a day–night effect, a team's quality, and a team's current form.

In any format of cricket, it is commonly believed that teams who are playing at home experience some sort of advantage. Amongst other possible explanations, in cricket, this is most likely to be because the home team will typically have played many matches at the venue, and therefore will be more familiar with the conditions.

Similarly, winning the toss in cricket is also considered to be an advantage to a team. However, in the literature, the effect of a binary covariate *toss* on the match outcome has not previously attracted statistical significance (see for example Akhtar & Scarf, 2012, and Bailey & Clarke, 2006). Nonetheless, we experimented with including a *toss* variable. Our results on a *toss* variable agreed with previous findings. However, an interaction term with the binary variable day–night (*dn*) was found to be statistically significant. In addition to experimenting with an interaction effect between the variables *toss* and *dn*, denoted by *dnt*, we also experimented with including all other two-factor interaction effects between categorical variables, but none of these were found to be statistically significant.

In regard to the past performances of the teams, we use the difference between the ICC official ODI ratings (*rd*) for the two teams at the time of the match. The ICC official ratings reflect a team's performance based on the matches they have played over the last three years. These ratings