STATISTICAL MODELLING IN LIMITED OVERS INTERNATIONAL CRICKET

Muhammad ASIF

Ph.D. Thesis **2013**

STATISTICAL MODELLING IN LIMITED OVERS INTERNATIONAL CRICKET

Muhammad ASIF

Centre for Sports Business, Salford Business School, University of Salford Manchester, Salford, United Kingdom.

Submitted in Partial Fulfilment of the Requirements of the Degree of Doctor of Philosophy, July 2013

CONTENTS

LIST OF FIGURES

LIST OF TABLES

DECLARATION

I declare that the thesis is my original work. No portion of this work has been previously submitted for another degree or qualification of this or any other University.

AKNOWLEDGEMENT

Firstly, I am more than grateful to Al-mighty Allah (God) to make me able to carry out this research for the degree of Doctor of Philosophy at the Salford University, UK.

Secondly, it is my great pleasure to acknowledge that this research is done under the supervision of Dr. Ian McHale, Reader in Statistics and Director of Centre for Sports Business at University of Salford, UK. I am very thankful to Dr. McHale for his sincere guidance, valuable comments, support and encouragement.

Thirdly, I am grateful to my father Niaz Hussain and to my mother Tahira Niaz (Late) for their unmatchable love, support, encouragement and prayers.

Fourthly, the generous financial support of the Salford Business School UK, Buzz Sports Ltd UK, and University of Malakand Pakistan, is gratefully acknowledged.

Lastly, thanks to, my brothers (Muhammad Atif and Muhammad Arif), my uncles (Nigah Hussain, Javed Tariq, and Khalid Tariq), all friends (especially Mr. & Mrs. Zahoor Khan, Tahir Sharif, Rana Arif and Ayaz Ali), all my colleagues (especially Sohail Akhtar, and Zahid Khan) and all my teachers for their prayers and encouragement.

ABSTRACT

This thesis addresses two areas of research relating to limited overs cricket using statistical analysis. First, we investigate the issue of resetting targets in interrupted matches and propose an alternative, new method to this end. Second, we address the problem of in-play forecasting match outcome.

In regards investigating methods for resetting targets, we provide a thorough overview of methods previously used. These methods also include the official ICC method, Duckworth-Lewis approach, and its alternatives, including the VJD method of Jayadevan (2002). The highly topical debate on which is the best method available, is addressed. Based on statistical analysis, it is shown that the Duckworth-Lewis method is the most viable solution when compared to the currently available alternatives. In the course of our analysis, we develop an estimation method for the Duckworth-Lewis professional edition, a previously unpublished but essential component of the method. Further, we develop a new improved version of the Duckworth-Lewis method which is more flexible than the original Duckworth-Lewis method for resetting targets. Our key modification is to propose a new alternative model for the mean remaining runs at a given stage of the innings. We show that the newly proposed model provides a superior fit to data and has more intuitive properties than the current Duckworth-Lewis method.

Regarding the in-play forecasting match outcome in cricket, we present a model that can be used to estimate match-win probabilities during any stage of a One-Day International match. Our model is a dynamic logistic regression model in that the parameters are allowed to evolve smoothly as the innings progresses. Further, the model utilises our modified Duckworth-Lewis model in measuring the wicket resources available to a team at any moment during the game. The covariates that we use in the model are categorized as either pre-match or in-play. From our dynamic forecasting model, we examine the overall and relative importance of the covariates. We assess how the effects of these covariates vary with respect to the progression of the innings. Further, some cross-validation techniques are used for the model selection and to assess in-play forecasting accuracies. Finally, we compare our ‘in-play’ forecasting model with the betting market. The results show that our newly proposed model, for in-play probability forecasts, is performing well.

CHAPTER 1 INTRODUCTION

1.1 Aims and Objectives

The purpose of this research project is to use statistical analysis to shed light on various issues related to limited overs cricket. First, we aim to develop a statistical model that can be used by the cricketing authorities, for example, the international cricket council (ICC), when resetting targets in interrupted cricket matches, quantitatively and objectively fair. Second, we aim to develop models that can be used to forecast match outcomes while the game is in progress. Such a model could be of use to bookmakers and punters. Team coaches and captains can also use the model to assess the merits of certain strategies of play. Lastly, cricket analysts and media can use the model in post match analysis. We set the following objectives to achieve our aims

- Review the literature on the problem of interruptions and forecasting in cricket.
- Examine some commonly used methods for dealing with cricket interruptions.
- To propose an estimation method for the latest version of the Duckworth-Lewis (D/L) method, the approach currently adopted by the ICC.
- To compare the existing Duckworth-Lewis method with alternative procedures proposed in the literature.
- To develop a new method (model) for resetting targets in interrupted limited overs matches, that provides a superior fit to data and has more intuitive properties than the current D/L method.
- To develop a simple in-play forecasting model that is dynamic and takes account of the stage of the innings.
- To identify factors that is indicators of match outcome during any stage of the game, and to asses and analyse how the effects of these factors vary with respect to as innings progress.

1.2 History of the Limited Overs International (LOI) cricket

The history of cricket dates back to the sixteenth century in England. However, at international level, matches (in the form of test cricket) started around 1877. Cricket's governing body, the International Cricket Council (ICC), has sought to make cricket more popular. In order to achieve, one strategy the ICC adopted was to introduce limited overs cricket (a shorter format of the game) with the intention of making cricket a faster, and more exciting spectacle that might attract a new audience. The limited overs cricket was introduced in the late 1960's, however at the international level the first game of such format were played in 1971. Presently, two types of limited overs international (LOI) matches are played. These are Twenty-20 International (T20I) and One-Day International (ODI).

The idea of limited overs cricket was not appreciated in the early decades after its introduction and therefore only eighty-two international matches were played until 1980. However, in the following decade, the game had achieved some popularity and five hundred and thirteen matches were played during 1980-1990. As of now, at the international level, more than three thousand and six hundred LOIs (One-Day and Twenty-20 International) have been played among the ICC recognized teams (www.Espncricinfo.com).

The International Cricket Council is responsible for organizing cricket matches at the international level. Currently, the ICC full members are Australia, Bangladesh, England, India, New Zealand, Pakistan, South Africa, Sri Lanka, West Indies, and Zimbabwe. The most important tournament in limited overs cricket organised by the ICC, is the world cup. The world cup for One-day International is scheduled once every four years, whilst the Twenty-20 International world cup is held once every two years. Presently, India is the ODI 2011 world champion, whereas West Indies is the T20I 2012 world cup winner. Previously, twice West Indies, once India, four times Australia, once Pakistan, and once Sri Lanka were the world champions for the ODI cricket. For T20I, India, Pakistan, and England, have each been a world champion once.

1.3 The Game of cricket

Cricket is a hugely popular sport around the world. An estimated three billion people are cricket fans, a figure that is larger only for soccer, which has an estimated 3.5 billion

fans (www.digalist.com). Broadly speaking, at international level cricket can be played professionally in two formats: limited overs and non-limited overs games, also known as time limited cricket. A non-limited over matches at the professional level typically last for several days. For example, in the case of international games between major cricket playing countries, a 'test match' lasts for five days. Limited overs matches on the other hand, are designed to start and finish on the same day. For example, ODI matches are limited to fifty overs per side, whilst T20I matches are limited to twenty overs per side. The twenty overs a side cricket is the shortest format of international cricket, with matches typically lasting for three hours, bringing the game closer to the time span of other popular spectator sports, for example football.

Cricket is played between two teams, each of eleven players. Each team has one captain that leads the remaining ten players. Each team bats in succession, known as an innings. A LOI match consists of two innings. However, a time limited match may have several innings, for example, broadly speaking a test cricket match consists of four innings. Regardless of the format, the game starts with tossing a coin between the two captains, a winner of which decides the choice of to bat or to field first.

The game is played on a round or oval-shaped grassy field known as cricket ground. The borderline of the ground is known as a boundary. The central part of the ground is known as pitch. The pitch is a rectangular 22 yards long clay strip with stumps at each end. The stump consists of three standing stakes that are usually made of wood. On top of the stumps are two bails- wooden crosspieces. Each set of three stumps along with the two bails, are known as the wicket. The Pitch should be about 55m from one boundary square of the pitch. Inside the pitch is marked with lines at 1.22m from each wicket, which are known as the creases. Figure 1.1 describes a wicket (right panel) and a standard pitch (left panel) of cricket ground.

Each player of the fielding team takes a location on the ground. One player always takes position as a wicket keeper (behind the wicket of the batsman at the striker's end of the pitch), and one must be selected as a bowler. The remaining nine players take different positions. The team captain is responsible for assigning fielding positions to the players. Figure 1.2 shows a typical set of players' positions on the cricket ground.

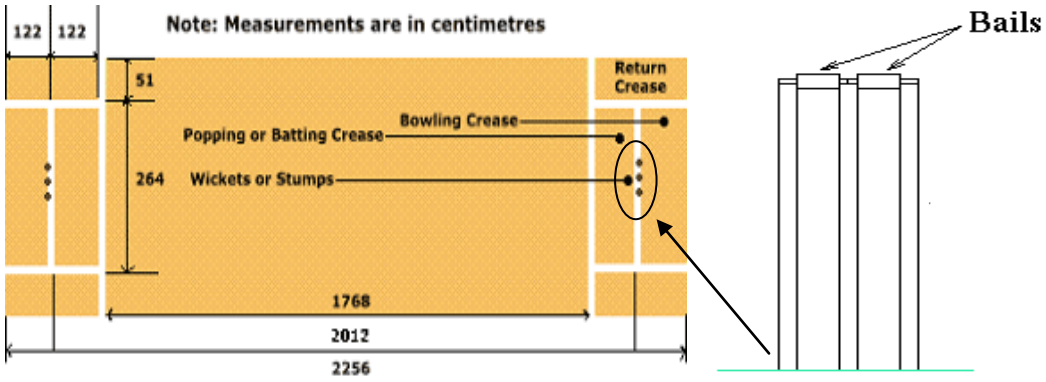

Figure 1.1 The images of the ICC's standard pitch (left panel), and a wicket that stake on each of the pitch (right panel)

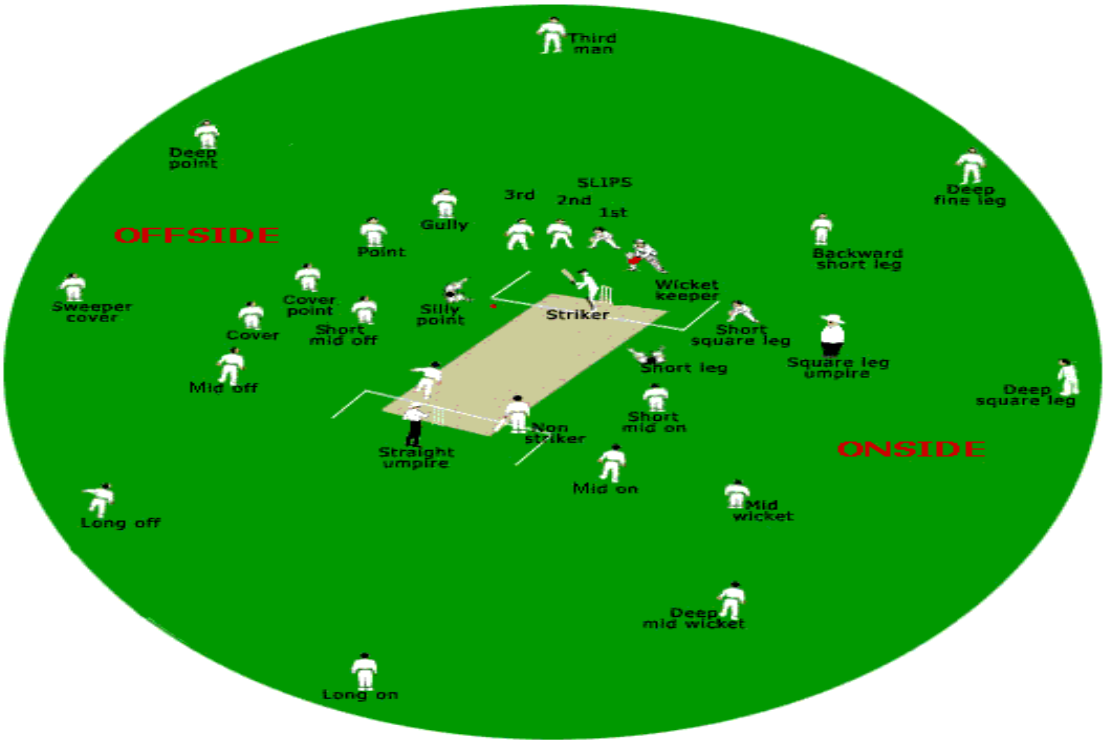

Figure 1.2 A cricket ground show the players and umpires' positions for right handed batsman at the striker end. Note that the mirror image of this figure will show the fielding positions for left hand batsman.

Two players from the batting team, known as batsmen, play in partnership to score runs against the bowling of the fielding side. The fielding side aims to restrict runs scored and to get wickets in one of the ways described in the rules of cricket (details are available on http://icc-cricket.yahoo.net/rules_and_regulations.php). Bowled, caught off the bat, leg before wicket (lbw), stumped by the wicketkeeper, and run-out are the

common ways for a batsman to be 'out'. When a batsman is ‘out’, another player takes his place from the batting team.

From the fielding side a bowler bowls an “over”- of six over-arm deliveries. No bowler can bowl two overs in succession. The maximum number of overs a bowler is allowed to bowl depends upon the format of the game. For example, a bowler can bowl a maximum of ten overs in a one-day international (ODI) , whilst a maximum of four overs can be bowled in T20I cricket. The fielding captain is responsible for appointing bowlers to bowl. Lastly, overs are delivered alternately from each end of the pitch.

Score is counted in the form of “runs”. Runs can be scored by the batting team in different ways. For example, runs are awarded as a result of the number of times the batsmen run from end to end of the pitch. Broadly speaking, the batting team obtains runs by hitting the bowler’s ball with the bat; a hit outside the boundary gives the batting team four runs if the ball touches the ground before crossing the boundary, or otherwise the batting team is awarded six runs.

At the international level, a match consists of one or two innings by each side. In test cricket matches, the side scoring the highest aggregate of runs wins, if the opponent team has completed its two innings of batting. If the match is not played to a finish then the result is a draw. In the case of the limited overs games the winning side is the one that scores most runs during its share of the overs. The innings can be ended in different ways, depending upon the format of the game. For example, innings in limited overs games are ended when all wickets are down, or when the pre allotted overs for batting team have been bowled or when the team batting second passes the target runs.

1.4 Thesis structure and contribution

This thesis is structured as follows. This chapter, CHAPTER 1, contains an introduction to and describes the purpose of our research project. A brief history and some fundamental standard cricket rules to play cricket are described. In the next chapter, CHAPTER 2, we give an overview of the problem of interruptions in limited overs cricket. Some simple and more-advanced methods to tackle the issue are discussed. The major shortcomings of the simple methods and its consequences are highlighted. A brief description of more-advanced methods and their advantages over simpler methods are provided.

In CHAPTER 3 we give overview of the Duckworth and Lewis (1998, 2004) (D/L) method for dealing with cricket interruptions, which has been adopted by the international cricket council (ICC). To our knowledge, it is for the first time in literature that an estimation method for the D/L model is presented. In the course of our analysis, we show that there is little evidence of a difference in the run scoring patterns of One-Day International (ODI) cricket and Twenty-20 International (T20I) cricket. Further, we also discuss the advantages of using a single model for both formats of the game. Some of the contents of this chapter have been published in McHale and Asif (2013)

In CHAPTER 4 we identify some properties that a method to be used for resetting targets in interrupted limited overs cricket should have. Based on these properties, we investigate the appropriateness of some high profile methods for resetting targets following on interruption. We compare the Duckworth-Lewis method, with the methods of Jayadevan (2002), Stern (2009), and Bhattacharya, Gill, and Swartz (2011) and conclude that the D/L method is more viable. We published this work in McHale and Asif (2013).

In CHAPTER 5 we present a new statistical model for resetting targets in interrupted limited overs cricket. We show that the model has a superior fit to data as compare to the existing D/L model. Further, we demonstrate graphically that the new model represents a more intuitive runs scoring pattern than the current D/L model. Again, we published this work in McHale and Asif (2013) .

In CHAPTER 6, we give overview on in-play forecasting in cricket. A Generalized Linear Models (GLMs) are been briefly described. Some model diagnostics and models selection methods are been discussed. For example, some information-criteria and cross-validation based methods are discussed.

In CHAPTER 7, we present a forecasting model for estimating match outcome probabilities during any point of a game. The model is dynamic in its parameters, which are evolving smoothly as the innings progresses. Further, we assess the factors that are indicative of the match outcome during the game. We demonstrate graphically how the effects of these factors vary with respect to the stage of the innings. Finally, we compare our model forecasts with that of betting market.

In CHAPTER 8, we describe the summary of the work done during this research project and description of the future potential research work is provided.

CHAPTER 2 THE PROBLEM OF INTERRUPTION IN LIMITED OVERS CRICKET

This chapter describes the problem of interruptions to play during limited overs cricket. Some standard methods for resetting targets, for the team batting second, following an interruption to play, are presented and discussed. Broadly speaking, these methods are divided into two categories: simple ad-hoc methods and advanced methods. In simpler methods, the targets are revised in an ad-hoc way. On the other hand more advanced methods are based on statistical models. We note that a major shortcoming of the simpler methods is that they do not account the wickets lost when resetting targets. With the help of some real and hypothetical examples, we demonstrate that such methods are easily exploitable by one or both teams.

2.1 Introduction

In comparison to other sports, limited overs cricket is particularly vulnerable to inclement weather – when it rains, or becomes too dark, cricket becomes too dangerous to play. Consequently, when a One-Day International (ODI) or Twenty-20 International (T20I) match is interrupted by rain or bad light, either or both of the competing teams can often not complete their allotted overs. Incomplete games are unsatisfactory for the players and fans alike and, to some extent negate the purpose of the shorter formats since an abandoned match offers minimal levels of excitement. Furthermore, to enable knockout tournament play, such as the ODI and T20I World Cups, games must reach a positive conclusion. Therefore, the cricket authorities have adopted quantitative methods to adjust scores and reset targets in order to ensure interrupted matches are concluded with positive results.

Since the first limited overs match was played in 1962, cricket analysts have searched for a fair method to reset targets in interrupted matches. The issue was elevated to higher importance following the introduction of the ODI world cup tournament in 1975. The ICC has tried several methods. These methods are also known as a rain rule for limited overs cricket. The current rain rule, the Duckworth-Lewis (D/L) method (Duckworth and Lewis, 1998) is now widely accepted as the fairest method available and has been in operation since 1997

In the next section, we briefly overviewed some simple ad-hoc methods for resetting targets following interruptions to play. Section 2.3 provides brief description of the more advanced methods that are proposed in literature. Finally, the summary of the chapter is given in the section 2.4.

2.2 Brief overview of some simple methods

2.2.1 Run rate method

In the past, the average run-rate (runs per over) was a commonly used method by International Cricket Council (ICC) to tackle the issue of interruptions to play. In this method the run-rate of each of the competing sides are compared, and the team with the higher run rate is declared as the winner. The run-rate method is simple to implement, but could unfairly favour either side, depending upon the situation. Other versions of this method, for example the maiden ignored run-rate method and the factored run-rate method, were also experimented by the ICC (CricketArchive, 2012). However, the fundamental problems with the run rate based methods remained unresolved. The major flaws of the run-rate based methods are to ignore the wicket-lost effect and to value (in the term of runs scoring potential) all the overs equally. The subsequent examples show how the method could be exploited as consequence of these anomalies.

Suppose a team batting second (team 2) chases a target 251. After 30 overs of the second innings, this team has scored 155 and lost nine wickets. Rain then interrupts the match and no further play is possible. Clearly, in such a situation team 2 is in weak position given it only has one wicket remaining and would likely lose the match. However, using the run-rate method for resetting targets meant a revised target of 151 in 30 overs was set and therefore team 2 would be declared the winner. In such cases, team 2 has an unfair advantage following the interruption if the target is reset using the run-rate method.

In regards to the situation where the run-rate method favours the team batting first (team 1), suppose team 2 is chasing the same target of 251, and has lost just two wickets in 45 overs. Further, assume that team 2 requires just 28 runs to win in the remaining five overs. Suppose, rain interrupts the match and team 2 is not able to bat for the rest of the innings. In this case, team 2 is in winning position, but using the run-rate method meant to be team 1 is the winner.

In both hypothetical examples, we note that the run-rate method ignored the number of wickets lost at the time of play was halted and therefore favoured team 2 in the first example and team 1 in the second example respectively.

2.2.2 Highest Scoring Overs (HSO) method

To eliminate the shortcomings in run-rate method the ICC adopted the highest scoring overs (HSO) method in the world cup in 1992. The method is also known as most productive overs (MPO) method. In this method team 1's over-by-over runs are arranged in descending order and then the sum of runs in the first x ordered overs is considered as a par score, where x is the number of overs available to team 2 in the second innings. In the implementation of this method, only team 2 could be unfavourably affected by this method. The best example in which team 2 was suffered, was the 1992 World Cup semi-final match between England and South Africa.

In the semi-final of the ICC World Cup 1992, England batted first and scored 252. Play was halted in the second innings when South Africa required 22 runs in the remaining 13 balls with four wickets in hands. Upon resumption the play, only one ball was remaining in the second innings. The HSO method had been applied and the target was revised such that South Africa required 21 runs to win on remaining one ball. Clearly, an impossible target off just one ball. However, before the interruption the required runs to win was not an impossible target.

To some extent, the highest scoring method (HSO) overcomes the shortcomings of the run-rate method. However, the problem of not accounting for a wicket lost effect remains unresolved. In addition, the method is dependent on the run scoring pattern of the team 1, which caused some unwanted consequences. This is especially evident when team 1 scores few runs in some overs and many runs in some others in a given match.

Some modified versions of the highest scoring overs (HSO) method were also experimented. For example, the consecutive highest scoring overs method (CHSO)- compares the maximum runs scored in x consecutive overs of team 1, where x is the number of overs team 2 is deprived, and the adjusted highest scoring overs method (AHSO)- the target is reset by the HSO method, but is then adjusted by reducing it down by a factor 0.5% for each over team 2 is deprived. Despite such modifications to the HSO method, we believe that the fundamental anomalies remain unresolved, for example

number of wickets lost is not been accounted. Therefore, the method is not fair for resetting targets following interruptions in limited overs cricket matches.

2.2.3 Equivalent Point (EP) method

This method was adopted by the England and Wales cricket board (ECB) for their domestic cricket during late 1960's. In this method, team 2's runs are compared to the equivalent point of team 1's runs. For example, on May 18th 1969 in the second innings of the Player's County League match, play was halted after Essex scored 40 runs and lost three wickets in first ten overs. At the equivalent point of the first innings, Derbyshire had scored 38 runs and therefore using EP method, Essex was awarded victory by 2 runs (http://cricketarchive.com/Archive/Scorecards/30/30029.html). Another version of this method is to compare each team's runs per wicket at equivalent points. The EP method is also simple to implement, but can have unwanted consequences. This is especially evident when an interruption happens prior to the start of the second innings or when teams are deprived of some overs in the middle of an innings. Moreover, the method is impossible to use in situation of multiple interruptions in the match.

2.2.4 PARAB method

This method is proposed by do Rego (1995) and is based on the parabola, $f(x) = 7.46x - 0.059x^2$, where $f(x)$ represents the runs obtainable in x overs. This method was adopted by the ICC in the World Cup 1996. In this method the proportion of expected runs obtainable by team 1 is calculated using $R_1 = f(x_1)/f(N)$, where x_1 is the overs available to team 1 and N denotes the number of pre-allotted for each teams. The proportion of expected runs obtainable for team 2 is calculated in a similar manner. The par score, T, for the team batting second is then calculated as $T = SR_2/R_1$.

This method also has the same problems as in the methods discussed above. For example, the number of wickets lost at the time of interruption is not accounted by the *PARAB* method. Further, the parabola has a maximum at about 63 overs (see Figure 2.1), which results in an unintuitive relationship between runs and overs.

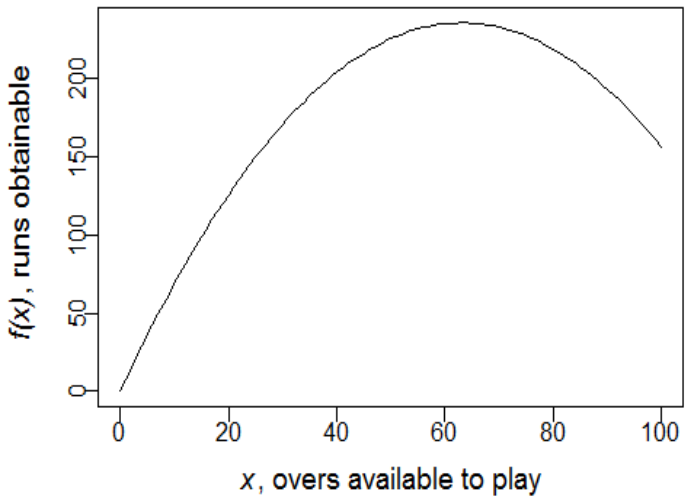

Figure 2.1 Runs obtainable, *f*(*x*), against the number of overs, *x*, in PARAB method

2.3 Brief overview of the advanced methods

2.3.1 The Duckworth-Lewis (D/L) method

In 1997, two British statisticians, Frank Duckworth and Tony Lewis, proposed a method for resetting targets for the team batting in the second innings in interrupted matches. Duckworth and Lewis (1998) describes their method for revising targets that accounts for the situation of the match in terms of number of wickets lost and the overs remaining at the time of interruption. The method is known as D/L method and is currently adopted by the ICC. The fundamental idea behind the D/L method is to estimate the resources available, R, to each team. In an uninterrupted match, each team will have 100% of its resources available and no target adjustment is necessary. However, if there is an interruption and the resources of team 1, R_1, are not equal to team 2's resources, R_2, then the target for team 2 must be adjusted. Let S be the total runs scored by team 1 (the team batting first), then the D/L method states that the par score for team 2 (the team batting second) T, is given by

$$T = \begin{cases} SR_2/R_1, & \text{if } R_2 \leq R_1 \\ S + G(N)(R_2 - R_1), & \text{if } R_2 \geq R_1 \end{cases} \qquad 2.1$$

where $G(N)$ is the average first innings total number of runs in an N-over match (N is typically either 50 or 20). The target for team 2 is then the next integer above T.

To measure each team available resources, Duckworth and Lewis (1998) developed a resources table which is based on exponential type model (will be discussed in the section 3.2). The two dimensional table describes the resources remaining for each overs remaining, u, and given wicket lost, w, and is denoted by $P(u,w)$. Table 2.1 is the extract of the latest D/L resources table published in 2002.

Table 2.1 Extract of the Duckworth-Lewis resources (%) table, published in 2002.

***u*, overs left**	***w*, wicket(s) lost**					
	0	**1**	**3**	**5**	**7**	**9**
50	100	93.4	74.9	49.0	22.0	4.7
45	95.0	89.1	72.5	48.4	22.0	4.7
40	89.3	84.2	69.6	47.6	22.0	4.7
35	82.7	78.5	66.0	46.4	21.9	4.7
30	75.1	71.7	61.5	44.7	21.8	4.7
25	66.5	63.9	56.0	42.2	21.7	4.7
20	56.6	54.8	49.1	38.6	21.2	4.7
15	45.2	44.1	40.5	33.5	20.2	4.7
10	32.1	31.6	29.8	26.1	17.9	4.7
5	17.2	17.0	16.5	15.4	12.5	4.6
0	0	0	0	0	0	0

To understand how the Duckworth-Lewis can be implemented, consider the hypothetical example in the section 2.2.1. That is, while chasing the target of 251, team 2 is deprived of the remaining twenty overs with a score of 155 for the loss of nine wickets at the time of interruption. It can be seen in the Table 2.1, for this example, that team 2 has lost 4.7% resources in the remaining twenty overs and therefore the total resources consumed by team 2, R_2, is equal to 95.4%. Using equation 2.1 the par score for team 2, T, is equal to 238.5 (greater than team 2's score) and hence, using the D/L method means that team 1 is the winner.

2.3.2 The Jayadevan (VJD) method

Jayadevan (2002) proposed a method which takes account of the situation at the time of the interruption in terms of number of overs and wickets. He referred to his method for resetting targets as the VJD system. This method was adopted by the Indian Cricket League (ICL 2007-2009), an Indian domestic cricket league run by private companies.

The VJD system is based on two types of resources, which Jayadevan describes as the 'normal' and 'target' scores. The 'normal' scores are modelled as function of percentage of the *overs* and *wickets* used. Whereas, the 'target' scores is modelled as a function of percentage of *overs* available. Strictly speaking these scores are the proportion of runs; however we refer it as resources (a term that is used in the D/L method). In the VJD system, the resources consumed by the batting team, Q, at the time of an interruption to play, are measured from 'normal' resources. However, when play is resumed, the proportion of the available resources, t, as compare to the total remaining resources (one less 'normal' resources as at the time of interruption) are measured from the 'target' resources. Table 2.2 is an extract of the resources table for the VJD system.

Table 2.2 The extract of the VJD resource table, taken from Jayadevan (2002).

v, percentage of overs (%)	t, 'target' resources (%)	Q, 'normal' resources (%) for give *w*, wicket(s) lost					
		0	**1**	**3**	**5**	**7**	**9**
0	0	0	0	0	0	0	0
10	15.7	8.8	12.0	35.0	60.0	79.0	95.0
20	29.8	16.9	20.8	35.0	60.0	79.0	95.0
30	42.3	24.7	27.2	35.0	60.0	79.0	95.0
40	53.5	32.4	34.7	39.7	60.0	79.0	95.0
50	63.4	40.4	42.0	44.3	60.0	79.0	95.0
60	72.3	49.2	50.1	51.7	64.3	79.0	95.0
70	80.3	59.2	59.7	60.4	69.8	79.0	95.0
80	87.6	70.7	70.7	71.1	74.7	83.6	95.0
90	94.3	84.1	84.1	84.1	84.1	89.2	95.0
100	100	100	100	100	100	100	100

To calculate the par score for the team batting second, Jayadevan (2002) divides the type of interruption into three categories: type A- team 2 is deprived of some overs before the start of the second innings, type B- an interruption in the second innings after team 2 bat for some overs, and type C- the first innings is interrupted. The step-by-step procedure of the application of the VJD system is provided in Appendix I as taken from the Jayadevan (2002) research article. However, in section 4.3 we simplify this method and transform the procedure into a single formula.

2.3.3 The Probability Preservation method

The fundamental notion of this method is to revise the target such that the probabilities of each team winning the match, as calculated before and after the

interruption, are preserved. Preston and Thomas (2002) were the first authors to present a method for adjusting targets that preserves the probability of victory for each team as it stood before the interruption took place. Carter and Guthrie (2004) follow a similar ethos and present algorithms to preserve the probability of victory across interruptions during an ODI game.

Specifically, Carter and Guthrie (2004) estimate the distribution of the runs to be scored in remaining u overs given w wickets already lost. Let, $F(x;u,w)$ be the distribution function for the random variable runs remaining, x, to be scored such that u overs are remaining and w wickets have already been lost. Let S denote the total runs team 1 has scored in the first innings, and y denote the number of runs team 2 has scored at u overs remaining given w wickets already lost. Then team 2's probability of winning the match is given by

$$p = 1 - F(S - y; u, w) \qquad 2.2$$

Suppose, u_1 and u_2 are the overs remaining at and after the interruption respectively such that w wickets already lost. Then the par score (T) of the Carter and Guthrie (2004) method is calculated such that $F(T - y, u_2, w) = F(S - y; u_1, w)$. The functional form for F is given in their paper.

2.4 Summary

We have given an overview of some simple ad hoc methods that have been used by official cricketing authorities, for example the International Cricket Council (ICC). We have examined the run-rate, HSO, EP, and PARAB methods. It is argued that all these methods have undesirable properties and consequently can result in unfair rest targets.

Similarly, in regards to the more advanced methods, we overviewed the Duckworth-Lewis method proposed by Duckworth and Lewis (1998), the VJD system, a similar resources based method proposed by Jayadevan (2002), and the Probability Preservation (PP) method, firstly proposed by Preston and Thomas (2002) and then by Carter and Guthrie (2004). The revised targets using these more advanced methods take account the overs and wickets at the time of interruption to play.

CHAPTER 3 THE DUCKWORTH-LEWIS (D/L) METHOD

In this chapter, we present an estimation method for the latest version (*Professional Edition*) of the Duckworth/Lewis (D/L) method. Further, we analyse and compare the runs scoring pattern of the one-day international (ODI) and Twenty20 International (T20I) formats of cricket. The results suggest that it is reasonable to use a single model for both the formats. Some of the content of this chapter is published in McHale and Asif (2013).

3.1 Introduction

The Duckworth-Lewis method has been through two incarnations. The first was adopted by the ICC in 1997 and is described in Duckworth and Lewis (1998). This version of the D/L method is known is *Standard Edition*. The second version, known as the *Professional Edition*, was introduced in 2003 (see Duckworth and Lewis, 2004) so that the method produced fairer adjusted targets in high scoring interrupted games. Currently, the Duckworth-Lewis Professional Edition is in operation and is being used by the ICC for all interrupted ODI and T20I cricket matches.

Some research in literature on limited overs cricket is closely related to the Duckworth-Lewis method. For example, Clarke and Allsopp (2001) use the D/L method to estimate teams' rankings in a tournament. They measured teams performances in the ICC World Cup 1999. de Silva, Pond, and Swartz (2001) use the Duckworth-Lewis method to estimate the runs margin of victory for the team batting in the second innings. Lewis (2005) proposed a method, based on Duckworth-Lewis model, to estimate a player's contribution in the match. Lewis (2008) further extended this work and proposed a ranking system for players in One-Day International cricket. O'Riley and Ovens (2006) use the Duckworth-Lewis resource table as a forecasting tool to predict total runs in the first innings. They show that the Duckworth-Lewis method has better predictive ability than the following three methods: VJD system of Jayadevan (2002), the Run Rate (RR) method, and the PARAB method of do Rego (1995). Bailey and Clarke (2006) use the D/L method in an ad hoc way in their pre-match forecasting models to forecast match outcomes during the course of a game.

The next section describes the existing D/L method. Section 3.3 describes the data that we have used for estimating the D/L model. In section 3.4 the runs scoring patterns of ODI and T20I are analyzed and compared. In section 3.5 we present the method of estimation for the latest version of the D/L method. In section 3.6, the D/L model fit results are presented. Lastly, the summary of the chapter is provided in section 3.7.

3.2 The Duckworth-Lewis Model

To estimate the resources available to a team, the Duckworth and Lewis (1998) method uses a model of the average runs remaining to be scored, Z. The Duckworth-Lewis model for the expected runs in the remaining *u* overs and given *w* wickets already been lost, is given by

$$Z(u,w) = Z_0 F(w)\{1 - e^{-bu/F(w)}\} \qquad 3.1$$

where Z_0 is the asymptotic average runs with no wickets lost in hypothetically an infinite number of overs. $F(w)$ is a positive decreasing step function with $F(0) = 1$ and is interpreted as the proportion of runs that are scored with *w* wickets lost compared with that of no wickets lost, and hypothetically infinitely many overs available. That is, $F(w) = \lim_{u\to\infty} Z(u,w)/Z(N,0)$. The ratio

$$P_N(u,w) = Z(u,w)/Z(N,0) \qquad 3.2$$

gives the average proportion of runs still to be scored in an innings with *u* overs remaining and with *w* wickets lost, which Duckworth and Lewis (1998) present as the proportion of remaining resources. For brevity, we refer to this as remaining resources, although strictly speaking it is a proportion. Using equations 3.1 and 3.2 to estimate the revise targets in an interrupted match is known as *D/L Standard Edition*.

Duckworth and Lewis (2004) modified the original 1998 model for high runs scoring matches. The idea being that the resources remaining, for a given number of wickets lost, decrease linearly when a team is chasing a well above average target. In other words, each over has equal value and so the over-by-over runs scoring pattern tends to be uniform, if the number of wickets lost remains the same. For this purpose, they include an extra parameter that they call the match factor and is denoted by λ. In matches with well above average targets, the parameter λ scales down the rate parameter *b* and scales up the parameter Z_0. As a result, Z tends to relate more linearly to *u*, overs left. The D/L upgraded model is given by

$$Z(u, w|\lambda) = Z_0 \lambda^{n(w)+1} F(w)\left[1 - \exp\left\{-bu/\lambda^{n(w)} F(w)\right\}\right] \qquad 3.3$$

where $n(w)$ is a positive decreasing function with *n(0)* = 5. The updated version of the D/L method is known as *Professional Edition*. Strictly speaking, we should not be conditioning only on λ. However, to distinguish Z in equation 3.3 from Z in equation 3.1, we follow this notation of Duckworth and Lewis and continue with it throughout the thesis. In innings *i* (*i = 1,2*), following n_i interruptions (the j^{th} interruption stops play when u_{1j} overs remain and w_j wickets have been lost and play is resumed when u_{2j} overs remain), the resources available is given by

$$R_i = 1 - \sum_{j=1}^{n_i} \left\{P_N(u_{1j}, w_j) - P_N(u_{2j}, w_j)\right\} \qquad 3.4$$

Duckworth and Lewis (1998, 2004) did not disclose the estimates and the estimation method for their model parameters. Therefore, in section 3.5 we propose a method of estimating the parameters for the Duckworth-Lewis model.

3.3 Cricket data for the D/L modelling

Estimation of the parameters was facilitated by collecting over-by-over data on 463 ODI uninterrupted matches from January 2008 to October 2011, and 198 uninterrupted T20I matches from the start of these games in February 2005 to September 2011. The data were obtained from the ESPN cricinfo website (www.espncricinfo.com). Purpose written code was used to estimate all parameters of the D/L model using standard optimisation routines in R (R Development Core Team, 2012).

Table 3.1 gives an extract of the average runs remaining to be scored with *u* overs remaining and when *w* wicket have been lost, as denoted by $\bar{x}(u, w)$, for ODIs (right panel) and for T20Is (left panel). Some matches in our original data set were reduced to shorter matches before the first innings started. We include these matches in our estimation sample, as the match was not interrupted during play. As such, the sample sizes for the start of the innings given in Table 3.1 are 458 (not 463) for ODI and 191 (not 198) for T20I.

Before we fit the D/L model in equation 3.1 on the combined data of ODIs and T20Is, first we analyse and compare the runs scoring pattern of the two formats of

cricket. The subsequent section describes whether it is justifiable to combine the data of the two formats and fit a single model for both ODI and T20I cricket interruptions.

Table 3.1 The observed means of remaining runs, $\bar{x}(u, w)$, with corresponding standard deviations, $s(u, w)$, and number of cases, $n(u, w)$, for T20Is (left panel) and ODIs (right panel).

T20I (Feb 2005- September 2011)						ODI (Jan 2008 to October 2011)					
u \ w	0	1	3	5	7	u \ w	0	1	3	5	7
20	151.79	*	*	*	*	50	245.43	*	*	*	*
	34.01	*	*	*	*		63.09	*	*	*	*
	191	0	0	0	0		458	0	0	0	0
15	128.45	115.24	106.33	67.00	*	40	236.26	202.89	147.09	130.67	*
	27.98	28.25	32.32	*	*		38.39	49.99	52.98	24.91	*
	47	79	15	1	0		109	178	44	3	0
10	90.38	88.50	82.16	45.00	58.00	30	189.64	184.95	154.19	99.40	67.25
	17.85	21.84	20.02	21.13	*		26.01	40.18	40.82	40.42	45.10
	13	30	49	12	1		25	81	100	30	4
5	46.50	57.83	47.76	45.81	29.42	20	145.29	143.38	121.84	96.76	65.09
	17.68	14.74	17.17	11.55	11.09		24.34	30.49	29.32	34.19	31.11
	2	6	41	48	12		7	32	114	62	23
2	*	34.00	33.16	27.52	23.65	10	91.00	87.00	85.81	68.18	46.93
	*	9.64	11.92	9.39	11.33		4.24	10.47	19.39	19.82	24.43
	0	3	25	48	20		2	6	64	104	44
3	*	24.00	23.00	21.26	17.58	5	*	45.00	51.77	44.67	33.38
	*	2.83	9.89	7.52	7.63		*	2.83	11.71	15.17	16.26
	0	2	16	46	31		0	2	26	98	61
1	*	8.00	10.00	10.75	9.39	1	*	*	12.14	10.02	10.39
	*	*	5.61	4.33	4.12		*	*	6.59	3.59	5.23
	0	1	8	36	36		0	0	7	50	75

3.4 Runs scoring pattern (ODI and T20I)

To test for whether combining the data of the ODI and T20I is reasonable for estimation purposes, we tested for equality in means between $\bar{x}_{ODI}(u, w)$ and $\bar{x}_{T20}(u, w)$. To do this, at each u overs remaining (ranging from twenty to one), for each value of w (ranging from 0 to 9) we obtained 131 means for T20I. Of these, we have data on 94 means for the corresponding ODI data on means that possibly be tested. Performing 94 independent t-tests produced just three statistically significant differences in means at the 5% level. To further justify combining the ODI and T20I data, we next made the Šidák and Bonferroni corrections (see Abdi (2007)) to the significance level in order to take account of performing multiple independent tests on a data set and found that no cells were significantly different at an overall significance level of 0.05.

It seems there is little evidence of a difference between the scoring patterns in the two forms of the game. In addition to the evidence provided by the statistical tests performed above, we believe it is more appropriate, in an idealistic sense, to have one model for resources in cricket, regardless of the format. For example, suppose a ODI match is reduced to twenty overs per side. If two models are in existence (one for ODI and one for T20), which model would best be suited? In this case, having one overall model for scoring patterns in cricket is more attractive than having separate models.

3.5 Estimation of the Duckworth-Lewis method (*Professional Edition*)

We estimate the Duckworth-Lewis model parameters using the data presented in section 3.3. The D/L parameters are estimated in two stages. First, we estimate Z_0, b and $F(w)$, and next we estimate λ given team 1's total runs, *S*. We note that the parameter λ is estimated on match-by-match basis.

3.5.1 Estimation of Z_0, b and F(w)

Let $x_i(u,w)$ be the observed runs scored in the remaining *u* overs of the first innings of match *i* when *w* wickets have been lost. Similarly, let $\bar{x}(u,w)$ be the observed mean runs scored in the remaining first innings. We use first innings data because the target will affect the scoring pattern in the second innings. The first innings run scoring pattern, on the other hand, represents the true scoring pattern of a team trying to maximise its runs total, rather than a team trying to score enough runs to meet a target and win a game. To estimate Z_0, b and $F(w)$ for each $w = 0,1,..,9$ in equation 3.1, we minimise a weighted sum of squared errors, *WSSE*, given by

$$WSSE = \sum_w \sum_u k(u,w) e^2(u,w) \qquad 3.5$$

where, $e(u,w) = Z(u,w) - \bar{x}(u,w)$ and $k(u,w)$ is a weighting function that is intended to account the heteroskedasticity and consistency of the $\bar{x}(u,w)$. We propose to weight the observations using a weighting function $k(u,w)$, given by

$$\boldsymbol{k(u,w) = \sqrt{n(u,w)}/s(u,w)} \qquad 3.6$$

where, *n* is the number of data points and *s* is the standard deviation of the remaining runs in the innings. Further, for *k* to be finite and $\bar{x}(u,w)$ to be reliable, we discarded means calculated using fewer than five observations.

3.5.2 Estimating λ and n(w)

The Duckworth-Lewis *Professional Edition* requires an estimate of λ when team 1 scores (S) well above average runs. For average and below average of S, $\lambda = 1$. In our experimentation with the resource tables provided by Duckworth and Lewis (2004) we note that $n(w) = \alpha + \beta F(w)$ with $\alpha = 2$ and $\beta = 3$. The λ depends on S, team 1's score, the number of overs allotted before team 1 starts its innings, *N*, and α and β. In a match in which team 1's innings is uninterrupted, λ is estimated such that,

$$g(\lambda) = |Z(N, 0|\lambda) - S| = 0 \qquad 3.7$$

If team 1 faces *n* interruptions in team 1's innings then λ is optimised by minimizing the following function

$$g(\lambda) = \left| Z(N, 0|\lambda) - \sum_{i=1}^{n} \Delta S_i - S \right| \qquad 3.8$$

where, ΔS_i is the expected runs loss in i^{th} interruption and can be defined as

$$\Delta S_i = Z(u_{1i}, w_i|\lambda) - Z(u_{2i}, w_i|\lambda) \qquad 3.9$$

where, w_i are the number of wickets lost, and u_{1i} and u_{2i} are the number of overs remaining at and after the i^{th} interruption respectively.

To our knowledge, no work has been done so far that provides statistical evidence to justify that the D/L *Professional Edition* is an improved version of the D/L method. We test whether using the D/L Professional Edition model for high scoring matches improves the model fit in section 5.3.4. Further, a computer program CODA, only available to the official cricketing authorities, is required to estimate λ in a given match. We developed R code for optimizing λ for any given type of interrupted limited overs cricket match.

3.6 The D/L model fit result

Following the estimation procedure, described in the section 3.5, we fit the Duckworth-Lewis model in equation 3.1. Purpose R code was written using standard optimization function *optim*() to fit the model. Table 3.2 provides the estimated values for the D/L model. It is to be noted that the parameters $F(w)$ are estimated under the constraint $F(0) = 1$ and $(w) \geq F(w+1)\ \forall\ w = 0,..,9$. Further, from Figure 3.1 the fitted curves can be compared with the observed scatter plots. For example, Figure 3.1a shows the curves for the observed mean, $\bar{x}(u, w)$, whereas Figure 3.1b show the corresponding D/L fitted means, Z.

Table 3.2 The Duckworth-Lewis estimated model parameters

Parameter	Z_0	*b*	*F*(0)	*F*(1)	*F*(2)	*F*(3)	*F*(4)	*F*(5)	*F*(6)	*F*(7)	*F*(8)	*F*(9)
Estimate	295	0.03706	1	0.840	0.738	0.577	0.477	0.374	0.279	0.195	0.095	0.033

Some improvements, over the D/L original estimates, are immediately gained by using these updated parameters. For example, the average runs scored in the first innings of the fifty over matches in our sample is approximately 245. Duckworth and Lewis state in their original paper (Duckworth & Lewis, 2004) that the average runs scored in the first innings, as implied by their model parameter estimates, is 235 runs. However, refitting their original model to our updated data set we find the model implies the average runs to be around 247 runs – closer to the observed average.

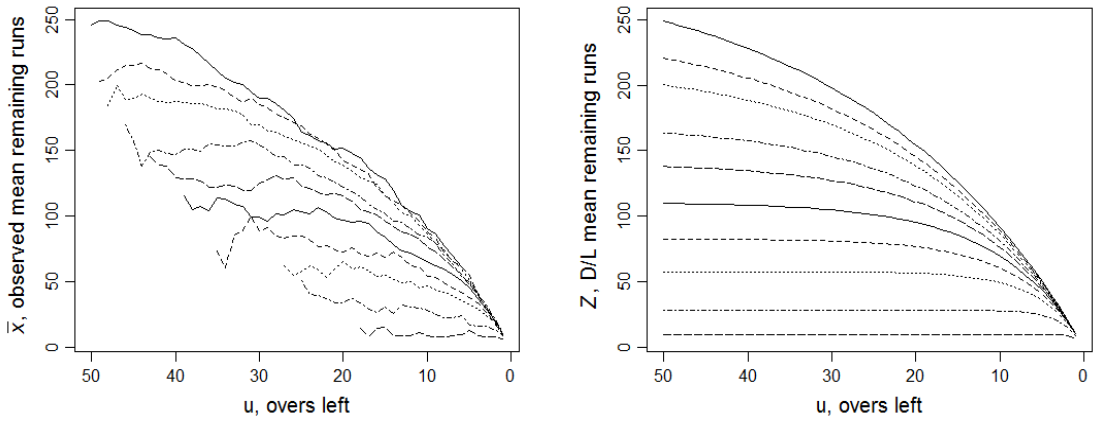

Figure 3.1 The plot of mean remaining runs against *u*, overs remaining, for (a) $\bar{x}(u, w)$, observed means, and (b) $Z(u, w)$, D/L model means. Top line is for zero wickets lost, and the bottom line is for 9 wickets lost.

3.7 Summary

This chapter begin with literature review related closely to the Duckworth-Lewis method for revising targets for the team batting second in interrupted limited overs cricket matches. Further, the latest version of the Duckworth-Lewis method, known as *D/L Professional Edition*, is also overviewed.

In regards to the research contribution in this chapter, we compare the scoring pattern of One-Day and Twenty-20 International cricket formats. The results show that there is no statistical significant difference between the scoring patterns in the two forms of the game. Further, we propose a method of estimation for the *D/L Professional*

Edition. To our knowledge this component of the existing D/L method is unpublished. The Duckworth-Lewis model parameters are estimated by minimizing the weighted sum of squared error. The weight function accounts the heteroskedasticity of the means.

The estimation process for Duckworth-Lewis method is performed in two stages. In the first stage, the D/L model is fitted for the *Standard Edition* of D/L method. Next, we estimate the match factor, λ, for the *D/L Professional Edition* for given estimated parameters of the D/L model for *Standard Edition* and the runs scored by the team batting first in the match. It implies that the parameter, λ, is estimated on match-by-match basis.

Moreover, we fit the D/L model on the combined data of the T20I and ODI data. Apart from statistical justification to combine the data of the two formats and fit a single model, we argue that from an ideological viewpoint it is preferable to have a single model for resetting targets in interrupted matched in both of the formats. The data that facilitate the D/L model fit is obtained from the espncricinfo.com website.

CHAPTER 4 THE DUCKWORTH-LEWIS METHOD COMPARED TO ALTERNATIVES

In this chapter, we contribute to the highly topical debate on which is the best method for resetting targets. Based on statistical analysis, we find that the Duckworth-Lewis method is the most viable solution when compared to some currently available alternatives. We investigate the VJD system of Jayadevan (2002), Stern's adjusted D/L method of Stern (2009) and Bhattacharya's version of the D/L method for T20I as proposed in Bhattacharya et al. (2011). In addition, we identify some standard desirable properties that a method for resetting targets following an interruption should satisfy. Some of the contents of this chapter have been published in McHale and Asif (2013).

4.1 Introduction

The Duckworth and Lewis (1998, 2004) method is heavily scrutinised and academics continue to propose improvements and alternatives. Several academic papers have appeared attempting to improve upon the D/L method and these can be split into two categories: resources based methods and probability-preserving based methods. Possibly the highest profile alternative is the VJD method of Jayadevan (2002) which can be interpreted in terms of resources. Stern (2009) proposes an adjusted D/L method by changing the resources table of the D/L original method in the second innings. The notion of this adjustment is to better reflect how teams batting second are able to adopt a different strategy from the team batting in the first innings. Bhattacharya et al. (2011) present an alternative resources table for the D/L method based on a non-parametric approach for Twenty-20 cricket. Regarding the probability based methods, Preston and Thomas (2002) were the first authors to present a method for adjusting targets that preserves the probability of victory for each team as it stood before the interruption took place. Carter and Guthrie (2004) follow a similar ethos and propose a method for resetting targets which they referred to as an Iso-Probability (IP) method.

In the next section, we present our standard desirable properties for a method to reset targets for the team batting second in interrupted matches. In section 4.3 we test the viability of the Jayadevan (2002) method and compare its performance with that of the Duckworth-Lewis method. Section 4.4 presents and highlights some issues related to the

Bhattacharya et al. (2011) version of the Duckworth-Lewis method. In section 4.5, we investigate Stern's version of the D/L method. In section 4.6 we examine the Iso-Probability (IP) method of Carter and Guthrie (2004). Lastly, the summary of the chapter is given in section 4.7.

4.2 The standard desirable properties of a method to revise targets

Let Z denote the expected runs obtainable in the remaining overs of an innings. Suppose, the number of overs remaining is denoted by *u*, whilst the number of wickets lost is denote by *w*. Let, there is a method $\mathcal{M}$ that fundamentally accounts for the stage and the state of the innings by *u* overs remaining and *w* wickets lost. Then at any given stage and state of the innings the team's expected remaining runs obtainable, Z, by means of method $\mathcal{M}$, should have the following properties.

I. Z should be a non-decreasing function of *u*, overs remaining , so that $Z'_u \geq 0$, provided that all other factors remain constant. Z'_u, is the first order partial derivative of Z with respect to *u*. For example, in the D/L method for any given match factor, λ, and wickets lost, *w*, the mean remaining runs, Z, is decreasing with respect to as the innings progresses (equally, Z is an increasing function of *u*).

II. The rate of change of *Z*, denoted by Z'_u, with respect to *u* should be a non-increasing function of *u* so that $Z''_u \leq 0$ provided all other factors remains constant. For example, in the D/L method for any given match factor, λ, and wickets lost, *w*, the ball-by-ball runs value is increasing with respect to the progression of the innings.

III. Z should be non-increasing function of *w*, wickets lost, provided that all other factors, for example λ and *u* in the D/L model, remain constant. This is intuitively appealing: at any given stage of the innings a team having more wickets in hand should have more (or equal) resources than a team with fewer wickets in hands. This property will be satisfied in the D/L method if the function $F(w)$ is a positive non-increasing function of *w*.

IV. The first order derivative of Z with respect to *u* should be a non-increasing function of *w* provided that all other factors in the method remain fixed. This implies that at any given stage of an innings, a team having more wickets in

hand should have more resources allocated to the current ball (or over). This property also ensures that the resource value lost ($\Delta Z_w = Z(u,w) - Z(u,w+1)$) for the loss of current wicket is decreasing with respect to the progression of the innings. This property can be satisfied if there exists a real number, *r*, such that Z'_u at u=r should be independent of all other factors in the model. For example, in the D/L method Z'_u at u=0 is independent of λ and w.

We use this list of desirable properties as criteria by which to assess each of the alternative methods for resetting targets below. We note that Further, we eventually propose a modification to the D/L method, which satisfies these properties.

4.3 Jayadevan's (VJD) method

Jayadevan (2002) proposed a method also known as the VJD system. In this method, the target to be revised for the team batting second depends on each of the competing teams available resources. Similar to the Duckworth-Lewis method, these resources depend on *overs*, *wickets* and the runs that are scored in the first innings. In print and via electronic media the topic of which method is better, has been extensively discussed. However, to our knowledge it is not been proved which method has more viable solution to the problem. Here we show that the VJD method has some serious flaws and that the D/L method has superior properties over the VJD system.

Before comparing the VJD and D/L methods we simplify the Jayadevan (2002) method by reducing the complicated step-by-step procedure (see Appendix I) into a single formula for calculating the par-score. Suppose, R_1 and R_2, are the resources available to team 1 and team 2 respectively. Let u_1 be the number of overs remaining when the play is halted in the given innings (first or second), and u_2 be the number of overs remaining upon resumption to play. We show that the team available resources (R_i) in the i^{th} innings, using the Jayadevan (2002) approach, can be written simply as

$$R_i = Q(v_1, w) + \{1 - Q(v_1, w)\}t(v_2) \qquad 4.1$$

where $Q(v_1, w)$ is the 'normal' resources corresponding to the v_1 ($= \frac{N-u_1}{N}$) percentage of overs played at u_1 overs to go, whereas $t(v_2)$ is the 'target' resources corresponding to the $v_2 = u_2/u_1$ percentage of available overs (see Table 2.2). The 'normal' resources (Q) is based on two separate regression models with independent variables: *overs* (as measured in percentage) and *wickets* respectively. However, the 'target' resources (*t*) are based on

one regression model with *overs* (as measured in percentage) as an independent variable. Jayadevan refers these resources as normal scores based on 'normal' curves and target scores based on 'target' curve. Jayadevan (2002, 2004) did not provide sufficient information about how the resources table (Table 2.2) for the VJD system has been constructed from such models. However, the availability of these resources table and the detailed procedure for calculating the par-score using the VJD system, means that we are able to investigate the runs scoring pattern implied by the VJD method. Finally, the par score for the team batting second can be determined as $T = SR_2/R_1$. Where S is the total runs scored by team 1 in the first innings.

Jayadevan (2004) updated the method for the well above average runs scoring situation. He constructed separate resources tables for different S, the total runs scored in the first innings. Six independent resources tables are proposed, one for each $S = 100, 200, .., 600$. Fundamentally, in VJD system, the notion of resources' adjustment to well above average runs scoring matches is similar to the Duckworth and Lewis (2004). That is, for well above average target the relationship between the 'normal' resources (Q) and *overs* tends to more linear. Hence, it can be observed that like the D/L method, the resources (R) based on VJD system is a function of S (team 1 total runs), u (overs remaining) and w (wickets lost). Details of testing the viability and fairness of the VJD system are provided in the subsequent sections.

4.3.1 First and Third desirable properties for the VJD system

We contrast Z, the expected remaining runs, of the D/L method (as depicted in Figure 3.1b) with the inferred Z of the VJD system. Suppose, in an N (typically equal to 50 or 20) overs match, team 1 scores, S, an average runs of the first innings (≈ 250). Now suppose no play is possible for the remaining u overs of the second innings at a time when team 2 had lost w wickets. Further, suppose there is no interruption in team 1's innings. Therefore, $R_1 = 1$ (team 1's total available resources), and $t(v_2) = 0$ (the proportion of available resources after the interruption, as compare to the total remaining resources before the interruption). From equation 4.1 the total resources available to team 2, R_2, at the time of interruption is $Q(v, w)$, where $v = \frac{N-u}{N}$. Therefore, using the VJD system, the expected remaining runs in remaining u overs given that w wickets have already been lost, is given by

$$Z(u,w) = S\{1 - Q(v,w)\} \qquad 4.2$$

Figure 4.1 shows the remaining runs, using the VJD system, that one would expect from the team batting second chasing an average target of 250. In other words team 2 is compensated with $Z(u,w)$ runs for the loss of remaining *u* overs provided that *w* wickets already been lost.

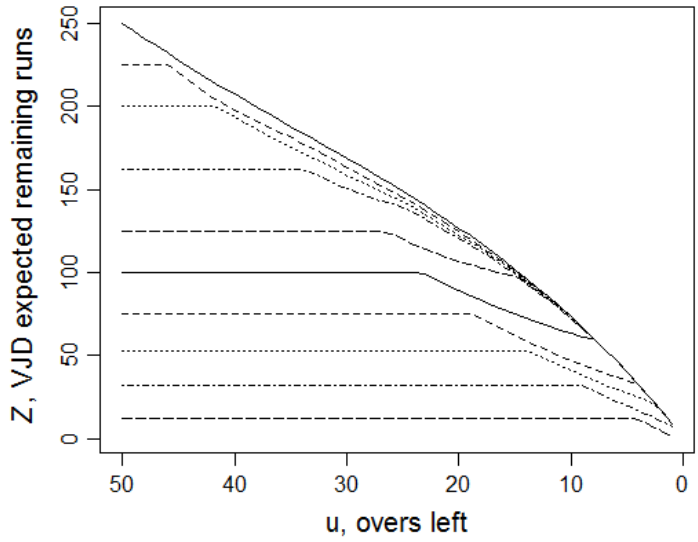

Figure 4.1 Curves of the team 2's expected remaining runs in *u* overs as measured using the VJD system of Jayadevan for *S*=250 (team 1's scores). Top solid line is for no wicket lost and bottom dashed line is for nine wickets lost

It can be observed from Figure 4.1 that the first and third properties are satisfied. However, in contrast to the Figure 3.1b (the Z plot of the Duckworth-Lewis method) the curves in Figure 4.1 are relatively non-smoothed and for most region the curves are flat. This is especially evident for non-zero *w*. This implies that there are many situations where the overs are zero valued in the VJD system. The consequences of this are unattractive. For example, suppose a team is chasing an average target of 250 and has lost six wickets and the innings is ended after 10 overs. This team would be compensated with the same number of runs as a team chasing the same target and lost the same number of wickets but could not play the final twenty overs. It means that with four wickets in hand, the overs 10 to 30 contribute zero resources to the team's innings. It can be argued that with such few wickets in hands (four in this example) there should not be a significant difference in the mean remaining runs in twenty overs and forty overs.

However, contradictorily, given the same number of wickets in hands if a team loses overs 10 to 30 and then play resumes, the same VJD method compensates these overs with 27 runs.

The above contradictory behaviour of the VJD system is because of using additional 'target' curve in the situation if play is possible after the interruption. The 'target' curve is used to estimate the proportion of a team's available resources as compared to the total remaining resources as estimated by 'normal' curve before the interruption to play. Such a shortcoming can be overcome in the VJD system by eliminating the 'target' resources and to use only the 'normal' curve to estimate remaining resources before and after interruption. Further, we note that the VJD system will be reduced to the Duckworth-Lewis method if only 'normal' resources table is used to estimate each team available resources. We further note that one less the VJD's normal resources can be interpreted as D/L's remaining resources. In the next section, we use the above hypothetical example for a single interruption to estimate the over-by-over runs value by VJD method.

4.3.2 Second and Fourth desirable properties for the VJD system

To see if the VJD method satisfies the second and fourth desirable properties, we examine the over-by-over runs value. The VJD system estimates the runs value of the next over (or overs) depending on the type of interruption. For example, suppose *S=250* in an *N* (fifty or twenty) overs cricket match, then in the second innings there are two possible ways in VJD system to estimate the runs value for the u^{th} remaining over, given *w* wickets have already been lost.

First, team 2 is deprived of the next over (u^{th} remaining overs). Then the runs value for the next over by the VJD method can be calculated as

$$\Delta Z_{1u} = S(1 - R_2) \qquad 4.3$$

where R_2 (as described in equation 4.1) is the total resources available to team 2 after the one over interruption. Let *u* and *u-1* denote the remaining overs before and after the interruption such that *w* wickets have already been lost. Assume there is no interruption in the first innings so that $R_1 = 1$. Then, $v_1 = \frac{N-u}{N}$ is the proportion of overs, consumed by team 2, as compared to total allotted overs, *N*, and $v_2 = \frac{u-1}{u}$ is the proportion of overs, available to team 2 after the one over interruption, as compared to the total remaining

overs, u. From equations 4.1 and 4.3 we have the following relation of expected runs value for u^{th} remaining overs

$$\Delta Z_{1u} = S\big(1 - Q(v_1, w)\big)\big(1 - t(v_2)\big) \qquad 4.4$$

where, $v_1 = \frac{N-u}{N}$ and $v_2 = \frac{u-1}{u}$. Given *S=250*, we use equation 4.4 for each $u = 50,49,..,1$ and $w = 0,1,..,9$ to estimate the over-by-over runs value.

In regards to the second possible way of estimating the next over runs value following the VJD approach. We take the difference between the expected remaining runs in u overs and expected remaining runs in *u-1* overs provided that the number of wickets lost, w , remains the same. Let $Z(u, w)$ denote the VJD expected remaining runs, or the runs a team is compensated with after being deprived of the remaining u overs given w wickets already lost, as given in equation 4.2. Then the u^{th} remaining over runs value can be measured as

$$\Delta Z_{2u} = Z(u, w) - Z(u-1, w) \qquad 4.5$$

Figure 4.2a and Figure 4.2a show the plots of over-by-over runs value against u using equations 4.4 and 4.5 respectively. Visual inspection of Figure 4.2 shows that the VJD system does not follow the second and third desirable properties as defined in section 4.2.

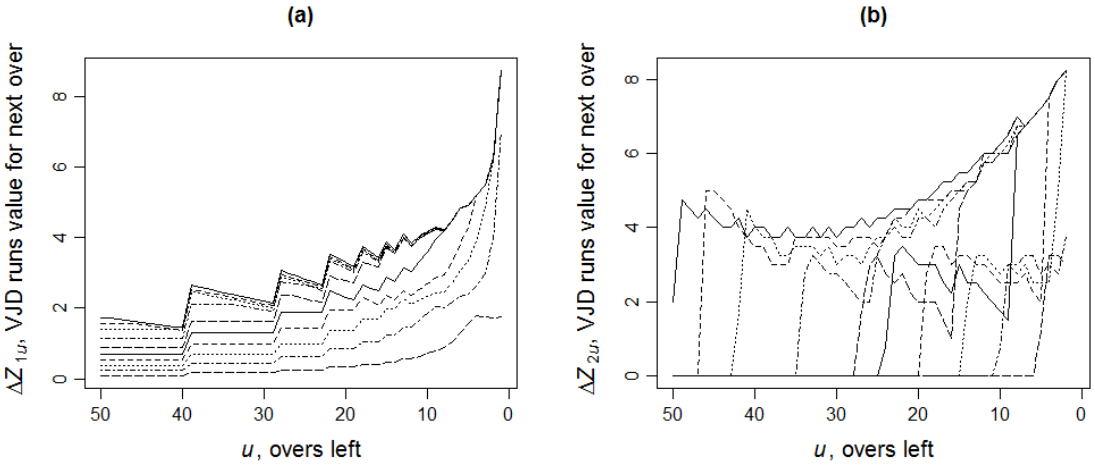

Figure 4.2 Plots for over-by-over expected runs value using the VJD system for a team chasing a target of 250, as measure using (a) equation 4.3 for a type 1 interruption and (b) equation 4.4 for a type 2 interruption, for each given *w=0 (top solid line)*,...,9(*bottom dashed line*)

Despite of not satisfying the third and fourth properties, it is also shown in the Figure 4.2 that resources allocated to the next over (or next few overs) by the VJD method are also unintuitive dependent on the type of interruption. For example, Figure 4.2a describes the next over runs value for the interruption of the next over. In contrast, Figure 4.2b describes the same quantity of next over runs value, but the interruption is taken place for all remaining overs. We refer to these interruptions as type 1 and type 2 interruptions. We note that the reason of such contradictory results is the use of additional 'target' curve for estimating the resources after the interruption to play. Whilst before interruption the remaining resources are estimated by the use of 'normal' curves only. Therefore when there is no play is possible after the interruption the VJD method not requires the use of the 'target' resources in the estimation of resources available to each competing teams. As a consequence, the resources for the similar quantity of next overs are different for a given situation (given u, w and S).

4.4 Bhattacharya's version of the D/L method for T20I

Bhattacharya et al. (2011) claim that the Duckworth-Lewis method is not suitable for Twenty-20 International (T20I) cricket. Their claim was based on a few real examples where the revised targets, by means of their method, seem to be better. However, they could not justify theoretically or empirically for large set of data that how such improvements are achieved in those examples. They proposed an independent resources table that could be used for T20I cricket. In this version of D/L method the resources table is estimated by non-parametric way for Tewenty-20 cricket.

We identify two major shortcomings of the Bhattacharya's version of the D/L method. First, the method does not account for the well above average runs scoring situation. Second, like the VJD system, this method also does not satisfy the second and fourth desirable properties. Figure 4.3 shows the over-by-over runs value, ΔZ_u, as calculated from the resources table given in Bhattacharya et al. (2011) for $S=150$.

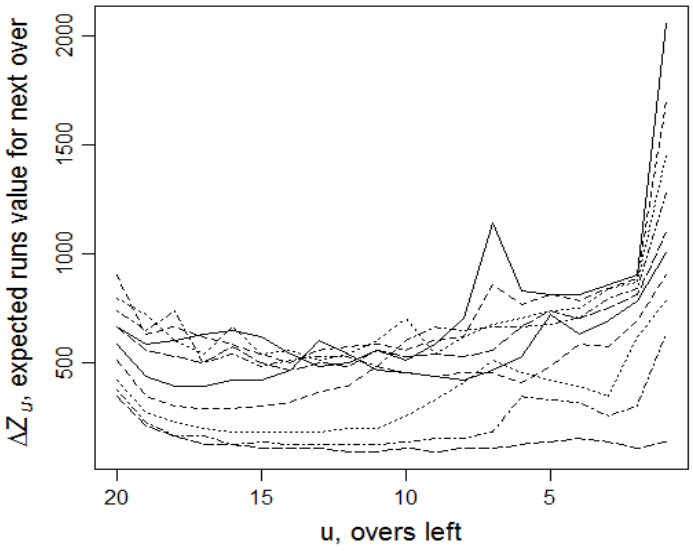

Figure 4.3 Plot of next over runs value, as calculated by Bhattacharya's version of the D/L method for a team batting second and chasing a target of 150 in T20I cricket. Top solid line is for no wicket lost and bottom dashed line is for nine wickets lost

We note that Bhattacharya's version of the D/L resources is estimated under the two constraints: the resources must be non-decreasing with respect to overs remaining, and the resources must be non-increasing with respect to wickets lost whilst no constraint is placed on the resources allocated to the $(N\text{-}u)^{\text{th}}$ over. Consequently, the erratic and unintuitive behaviour of the over-by-over runs value shown in Figure 4.3 results.

4.5 Stern's adjusted D/L method

Stern (2009) proposed an adjusted Duckworth-Lewis method for resetting targets following interruptions in limited overs cricket. He proposed an adjustment to the D/L resources if used for estimating team 2's resources. In contrast to the Duckworth-Lewis remaining resources, Stern's adjusted remaining resources are given by

$$P_{St}(u, w|\lambda) = 1 - F\big(1 - P_{DL}(u, w|\lambda)\big) \qquad 4.6$$

where $P_{DL}(u, w|\lambda)$ is the remaining resources as calculated by the Duckworth-Lewis method, and F is the beta cumulative distribution function with parameters $\alpha(\lambda)$ and $\beta(\lambda)$ to be estimated, and are given in Stern (2009), by

$$\alpha(\lambda) = 1 - 0.238e^{1-\lambda}\text{, and } \beta(\lambda) = 1 - 0.307e^{1-\lambda} \qquad 4.7$$

where λ is the match factor of the Duckworth-Lewis method and could be estimated as we presented in the section 3.5.2.

Our experimentation lead us to believe that Stern's adjustment with the D/L method results in more unintuitive behaviour of the runs scoring pattern during the second innings. Figure 4.4a shows how the runs awarded for the loss of the $(N\text{-}u)^{th}$ over do not behave intuitively with respect to as innings progress. For example, as shown in the top curve of the Figure 4.4a, given no wicket lost, the next over runs value is decreasing with respect to the progression of the innings. This is an undesirable property of Stern's adjusted method.

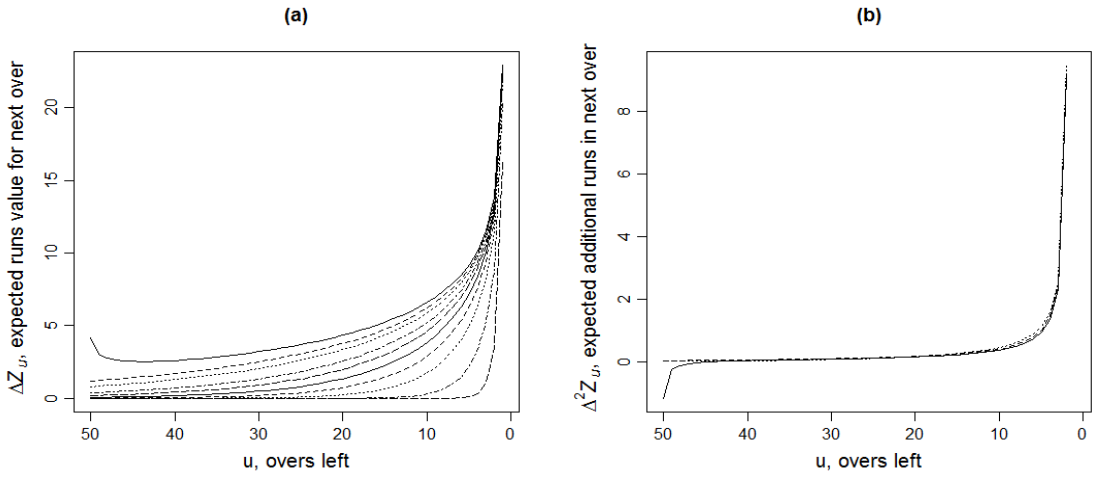

Figure 4.4 (a) The next over runs value for each given *w=0 (topped line),1,..,9(bottom line)* and (b) The average change in the runs value of consecutive overs for given *w=0,2,4* , using the Stern's adjusted D/L method, for a team batting second given *S=250*

Moreover, the change in the over-by-over expected runs value is unreasonably more rapid in the Stern's version of the D/L as compared to the existing D/L method. This is especially evident during the final stage of the second innings (see Figure 4.4b in the region $0 < u \leq 10$) . As a consequence, for example, a team which has lost five wickets is compensated with 12.6 runs for the loss of the 49^{th} over, and this rises to 22.4 runs for the loss of the final over. However, for five wickets lost, the observed average runs scored in the 49^{th} and 50^{th} overs are 10.6 and 10.3 respectively. These means are based on 99 and 86 observations respectively. Hence, we believe that when compared to Stern's adjusted D/L method, the existing Duckworth-Lewis method better represents the scoring patterns.

4.6 Iso-Probability (IP) method

Preston and Thomas (2002) were the first authors to consider the idea of resetting targets so that the match outcome probability is remained same before and after an interruption. After Preston and Thomas, Carter and Guthrie (2004) present a method based on this idea, and claim that their probability based method resets targets for the team batting second in interrupted cricket matches better than does the Duckworth and Lewis (1998) method. In response, Duckworth and Lewis (2005) demonstrate that the Carter and Guthrie (2004) approach for resetting targets produces contradictory revised targets in similar situations. To see this, we overview the subsequent example as given in Duckworth and Lewis (2005).

Suppose on two adjacent grounds A and B, team 1 has scored 250 runs in fifty overs. Let, on both grounds the overs be reduced to forty overs following an interruption after team 2 has batted for twenty overs and have lost three wickets. Suppose on ground A team 2 scores 120 whilst on ground B team 2 scores 50 runs at the time of interruption. The IP method meant on ground A team 2 is awarded 23 runs for the loss of ten overs, whilst on ground B the awarded runs for team 2 are 35 for the loss of ten overs. Hence despite these being similar situations, that is the team batting second is chasing the same target, the stage of the innings and the number of wickets lost for team 2 are same, the IP method compensates team 2 with more runs on ground B.

In response to the Duckworth and Lewis (2005) critical analysis, Carter and Guthrie (2005) extended the example and assume ground C with the same situation as of ground B, but the team batting first (team 1) has scored 180 instead. They argued that in such situation the Duckworth and Lewis (1998) compensates team 2 with 22 runs which is different from ground A and B. Table 4.1 shows the summary of resetting targets for each A, B, and C grounds using the D/L and IP approaches.

We believe that runs should have a different value when chasing different targets. For example, consider the above hypothetical example, we believe that the 22 runs relative to the 180 runs target on ground C should not have significantly different value to the 30 runs relative to chasing the 250 runs target on grounds A and B. For example, the proportion of runs that team 2 is compensated with, as compared to the target runs are same on all three grounds by means of the D/L method. However, using the IP method these runs proportions are different on all grounds of A, B, and C.

Table 4.1 Runs award with corresponding resources lost (in brackets) to the team batting second for the lost of next ten overs interruption after playing first twenty overs on each A, B, and C grounds using both the Duckworth-Lewis method and Iso-Probability methods.

Ground	*S*, team 1's total runs	Position of the team batting second at the time of interruption (runs/wickets)	Runs awarded to team 2 for the loss of 21-30 overs	
			D/L method	*IP method*
A	250	120/3	30 (12%)	23 (9%)
B	250	50/3	30 (12%)	35 (14%)
C	180	50/3	22 (12%)	23 (13%)

Moreover, it is noted that in the IP method the runs awarded to team 2 for the next interrupted overs are inversely proportion to runs scored so far in the second innings. For example, as shown in Table 4.1 that using the IP method meant a team chasing a target 250 and lost three wickets in the first twenty overs, is awarded with more runs for the next ten overs interruption when team 2 scores fifty (35 runs awarded chasing 250 target runs), as compared to the runs award to team 2 for the same interruption but scores 120 (23 runs are awarded chasing 250 target runs). This is somewhat counterintuitive; normally it is expected to perform better in the remaining innings, if the team is performing well so far. On the other hand the Duckworth-Lewis method is independent of how many runs team 2 has scored at the point of an interruption, but rather it depends on how many overs are remaining, how many wickets have already been lost and what is the target.

Apart from the shortcomings that are identified in the IP method of Carter and Guthrie (2004), this method was not adopted by official cricketing authorities; therefore we do not further investigate its appropriateness for resetting targets in interrupted limited overs cricket matches. However, future research might be of interest where this method is tested for the standard properties we have identified (see section 4.2) for a method to be used as resetting targets following cricket interruption.

4.7 Summary

This chapter is started with the identification of four desirable properties for a method to be used to revise targets following interruptions. The existing Duckworth-Lewis method is compared with some high profile alternatives that are existed in

literature. It is shown that the existing Duckworth-Lewis method is a more viable solution for resetting targets following an interruption when compared to other, high profile, resources based methods proposed in the literature.

Firstly, the VJD system of Jayadevan (2002) is investigated. With the help of the graphical demonstration, it is shown that the second and fourth desirable properties are not satisfied. Further, the VJD system produces contradictory results. It is argued that the contradictory behaviour of the VJD system can be resolved by discarding the 'target' resources table, and only the 'normal' resources table is used for estimating each team available resources. By doing such modification in VJD system, the method would reduce to the D/L existing approach for resetting targets.

Secondly, Bhattachary's version of the D/L method for T20I is examined and graphically it is demonstrated that the third and fourth desirable properties are not satisfied by this method. Further, it is argued that this version of the Duckworth-Lewis method does not account for the well above average runs scoring situation. As a result of these deficiencies, we believe that estimating the resources for D/L method in Twenty-20 cricket using the Bhattachary's approach is not appropriate.

Thirdly, the adjusted Duckworth-Lewis method proposed by Stern (2009) is analyzed critically. The results are evident that the runs scoring pattern in the second innings become more unintuitive after the Stern's adjustment to the D/L resources for the second innings. For example, the rate of change in over-by-over resources becomes extremely rapid during the final ten overs of the innings.

Finally, the Iso-Probability (IP) method of Carter and Guthrie (2004) is compared to the Duckworth-Lewis method. It is argued that the IP method for resetting targets produces different results in similar situations. Further, it is noted that the number of runs a team compensated with (after an interruption to play) is inversely related to the team performance (in term of runs scored) so far. Consequently, the interruption to play may have an advantage to a batting team who is performing poorly until the interruption. It is to be noted that we do not further investigate the probability preservation method, as the official cricketing authorities, for example the International Cricket Council, have not adopt it. However, future research might of interest where this method is tested for our identified desirable properties.

CHAPTER 5 A MODIFIED DUCKWORTH-LEWIS METHOD

In this chapter, we present a modified Duckworth-Lewis method for adjusting targets, for the team batting second, in an interrupted limited overs cricket. The key modification is to propose an improved alternative model for estimating a team's resources. Our newly proposed model provides a superior fit to data. Further, we demonstrate graphically that the proposed model has an improved intuitive runs scoring pattern for limited overs cricket. Some of the contents of this chapter are published in McHale and Asif (2013).

5.1 Introduction

The Duckworth-Lewis (D/L) method for adjusting targets in interrupted limited overs cricket matches is widely accepted as the fairest method available and is a great success story of operational research and applied statistics in practice. Despite its widespread use, there remains some doubt about the appropriateness of the D/L method. Firstly, it is because controversial adjusted targets continue to occur. Secondly, with the advent of Twenty-20 cricket, several stakeholders, including players and coaches have questioned whether a further adjustment should be made to the D/L method to adapt it to this shorter format.

In the next section, we identify issues related to the existing Duckworth-Lewis method. Section 5.3 describes a new proposed model for estimating resources available to each team. In section 5.4 avenues for future research in the D/L method is discussed. The summary of the chapter with some concluding remarks are given in section 5.5.

5.2 Issues in Duckworth-Lewis method

The Duckworth-Lewis method is widely accepted as a fair approach for dealing with interruptions in limited overs cricket. However, we believe there is a possibility to improve the latest version of the D/L method. Here we identify some issues in the scoring pattern inferred by existing D/L model.

Firstly, using the nine estimated parameters (one for each w, $w>0$) has consequences on the effect of wicket lost on the remaining runs. For example, we examine the expected runs value of each wicket partnership, as defined by $\Delta Z_w = Z(u,w) - Z(u,w+1)$, for

the D/L model at some given stages of the innings. Figure 5.1 is the graphical demonstration of the behaviour of the expected runs value loss in the remaining innings for lost of the current wicket, for each given $u = 50,45, \ldots ,5$ overs remaining. It can be seen that the first desirable property (see section 4.2) is satisfied by the existing Duckworth-Lewis method. However, the erratic pattern that is evident for the expected runs value lost with respect to successive wickets has some unwanted consequences. For example, until around the five overs left point (first forty-five overs), the second wicket partnership is valued with fewer runs than the first and third wicket partnerships. Similarly, the fourth wicket partnership is valued with fewer runs than the third and fifth wicket partnerships. In brief, for each given stage of an innings the relative importance of wicket partnerships is unintuitive.

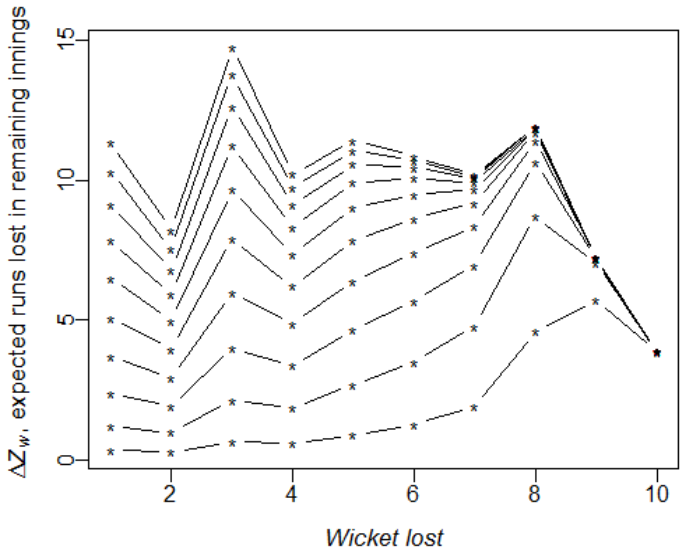

Figure 5.1 : The plot of ΔZ_w, expected runs lost in the remaining inning for the lost of current wicket for u = 50, 45,..,5 using the D/L model for λ =1

Secondly, as a consequence of the exponential type function for the D/L method, the rate of change in over-by-over (or ball-by-ball) expected runs value, as measured by $-\Delta^2 Z_u = \Delta Z_u - \Delta Z_{u-1}$ (where, $\Delta Z_u = Z(u,w) - Z(u-1,w)$), increases exponentially regardless of the situation. Figure 5.2 shows the curves of $-\Delta^2 Z_u$ for $w = 0,2,$ and 4. Hence, the D/L model implies that irrespective of the number of the overs remaining and the number of wickets lost, the batters are expected to score at an ever increasing run-rate.

However, suppose a team has two overs remaining and has lost no wickets. Averaging all other factors, for example the quality difference in two bowlers, the two batsmen are most likely already batting at maximum capacity and it seems unreasonable to expect them to score at an ever-increasing rate.

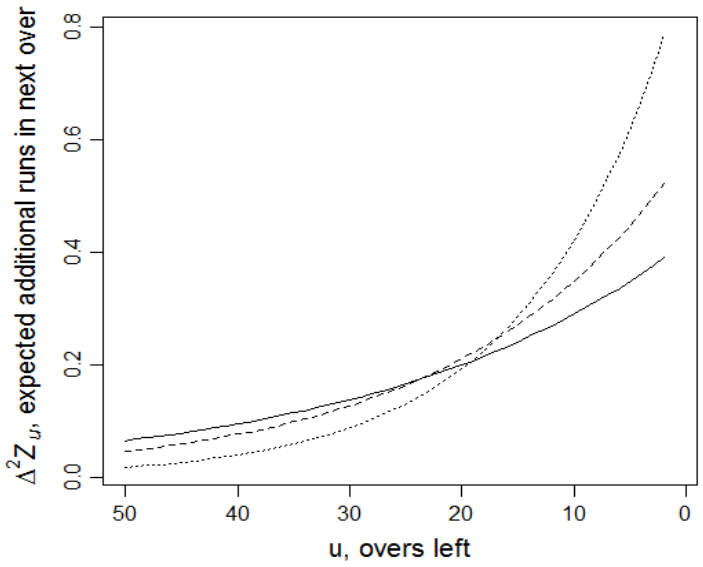

Figure 5.2 Plot for expected additional runs value, $-\Delta^2 Z_u$, against the stage of the innings, *u* overs left, for *w=0, 2,* and 4, using the D/L model in equation 3.1

Thirdly, we believe that the decay towards the asymptotes, $Z_0 F(w)$, is very rapid for the D/L model. As a consequence, in some situations, particularly when a team has lost wickets in the early stage of the innings, losing overs provides very little (sometimes zero) compensation and hence the revised target will remain unchanged. For example, suppose a team is chasing a target of 250 and has lost six wickets after five overs. The existing D/L model provides almost no compensation (actually it provides 0.55 runs) to the team if it is deprived of the next ten overs. With the advent of longer batting line-ups, it may be the case that this level of compensation is no longer reasonable.

Finally, we believe that the method of calculating par score, defined in equation 2.1, for the team batting second should be independent of $G(N)$. Duckworth and Lewis (1998) argue that in situation when team1's resources, R_1, is greater than team 2's resources, R_2, then the direct scaling results in unfair revised targets. We believe that this

is because of how the mean remaining runs are modelled and not because of direct scaling of the resources available to each competing teams in a match.

5.3 A new model for the D/L method

We propose an improved version of the Duckworth-Lewis method, which uses an alternative model for the mean remaining runs. This is done in stages, first we search for a model for $F(w)$ so that the wicket lost effect on the remaining runs is smoothed and intuitive. Next we developed a different model for $Z(u, w)$

5.3.1 Model for the F(w)

In the Duckworth-Lewis model, the function $F(w)$ is interpreted as the proportion of the mean remaining runs in hypothetically infinite overs remaining given that *w* wickets have already been lost. In order to satisfy the third desirable property for a method (see section 4.2), $F(w)$ should be a positive non-increasing function of *w* and intuitively should range from 0 to 1. We note that the properties associated to function $F(w)$ are similar to a truncated survival function. Therefore, we experimented with some survival functions. These are based on the Cauchy, Gamma, Negative Binomial, Normal, Geometric, and , Weibull distributions. However, the survival function based on a truncated normal distribution gives a superior fit to the data for the Duckworth-Lewis model. The function $F(w)$ can be written as,

$$F(w) = \frac{\Phi(10;\ \mu_1, \theta_1) - \Phi(w;\ \mu_1, \theta_1)}{\Phi(10;\ \mu_1, \theta_1) - \Phi(0;\ \mu_1, \theta_1)} \qquad 5.1$$

where Φ is the normal cumulative distribution function, and μ_1 and θ_1 are the location and scale parameters respectively. We refer the D/L model using equation 5.1 for $F(w)$ as Adjusted D/L model.

We note that using a smoothed $F(w)$ does not improve the goodness-of-fit of the model, but of course, the main objective for using a smoothed $F(w)$ function was not to improve the goodness-of-fit, but to produce a more intuitive and well-behaved wicket lost effect on the remaining runs. Figure 5.3 shows the expected runs lost in the remainder of innings for the loss of current wicket using the Adjusted D/L model. It is noticeable that in contrast to the wicket lost effect in the Figure 5.1, the wicket lost effect in Figure 5.3 is well behaved and more intuitive. Further, we note that using equation 5.1 in the D/L model reduces in the number of parameters to be estimated.

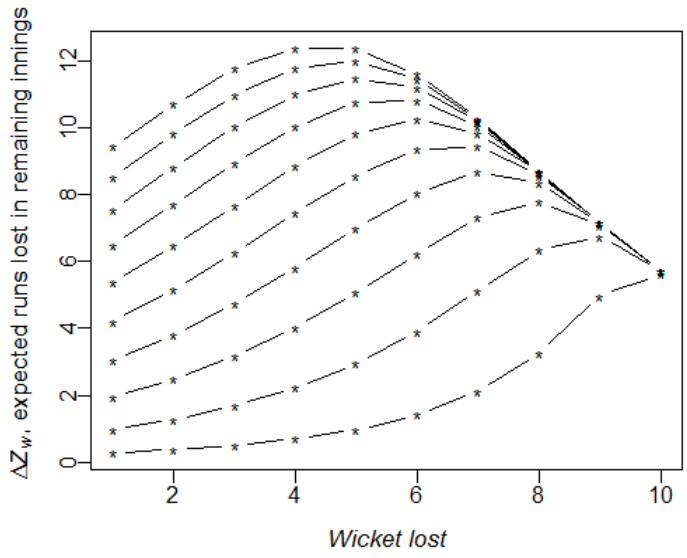

Figure 5.3 The plot of ΔZ_w, expected runs lost in remainder of innings for the loss of current wicket at *u = 50 (top line), 45,..,5 (bottom line)* overs-remaining stage, using the D/L model for our proposed *F*(*w*) in equation 5.1

Finally, as can be seen in Figure 5.3 that for each number of wicket lost the expected runs lost in the remaining innings decreases as the innings progresses. That means that even after smoothing $F(w)$, the method still satisfies the fifth desirable property. However, the relationship between 'expected runs lost in the remaining innings' and *w*, wickets lost, is not only smoothed for each *u* stage of the innings, but also the variation in the shape of relationships with respect to the progression of an innings become more intuitive. For example, it can be seen that in the final stages of the innings (for example, *u=10* and *5*) the runs value lost in remaining innings for lost of the top order wicket partnerships is smaller, as compared to the lower order wicket partnership. This is intuitively reflects a common strategy of limited overs cricket. For example, in limited overs cricket normally a team with enough wickets in hand will play aggressively in final stages of an innings as the risk/reward payoff is attractive during final stages compared to the early stages of an innings.

5.3.2 Model for the Z(u, w)

Having identified an objective way to obtain a smoothed $F(w)$ that produces a well-behaved function, we now propose an alternative functional form for $Z(u,w)$.

Cumulative distribution functions provide a wide range of curves that can be used to model $Z(u,w)$ for given *w*. Since the distribution functions are positive non-decreasing, the fundamental shape of Figure 3.1b can be preserved. This implies that the first desirable property for the new model will be satisfied. In regards to the second desirable property, we need a model such that its second order derivative with respect to *u* should be negative. Here again we have wide range of cumulative distributions of which the second order derivative remains negative in the positive range, for example exponential distribution. The curves need to be truncated at or above zero so that the domain is the positive real line. We note that using the exponential cumulative distribution function leads to the derivation of the existing Duckworth-Lewis model.

In regards to overcoming the second issue (as discussed in section 5.2) in the D/L model, we note that it is because there is no inflection point in the curves of Z'_u in the D/L exponential type model. In other words for the D/L model for any given *w*, the second order derivative of $Z(u,w)$, with respect to *u*, is a maximum for *u*=0 (see Figure 5.2 which reflects the shape of Z''_u for *w*=0, 2, and 4). It is to be noted that the point of inflection is a point in a monotonic curve at which the curve changes from concavity to convexity and vice versa. Hence, the second issue can be resolved if we choose a density function which has a point of inflection in its positive range. For example, there exists point of inflections for Normal and Cauchy distributions. Finally, the third issue can be resolved by selecting the distribution with a tail heavier than exponential distribution.

We experimented with several cumulative distribution functions, including the Normal, ex-Gaussian, t-distribution, Gamma and Cauchy distributions. The following model, based on the Cauchy distribution not only overcame the shortcomings of the D/L model, but also provided a better fit to the data. We propose that the average number of runs scored in the remaining u overs once w wickets have been lost be given by

$$Z(u,w) = Z_0 F(w)\left\{\frac{tan^{-1}\left(\frac{u-\mu_0}{\theta_0 F(w)}\right) - tan^{-1}\left(\frac{-\mu_0}{\theta_0 F(w)}\right)}{\frac{\pi}{2} - tan^{-1}\left(\frac{-\mu_0}{\theta_0}\right)}\right\} \qquad 5.2$$

where $Z_0 > 0, \theta_0 > 0$ and $-\infty < \mu_0 < 0$ are the model parameters. The function $F(w)$ is as described in section 5.3.1. In contrast to the exponential type D/L model, our proposed model might be referred as arc-tangent type model for the mean remaining

runs. We call our proposed model in equation 5.2 a modified form of the D/L model or simply a modified D/L model.

In contrast to Figure 5.2, the rate of change in over-by-over expected runs value is graphically demonstrated in Figure 5.4. It reflects the curves of $-Z_u''$ for our proposed model in equation 5.2, for *w*=0, 2, and 4. Strictly speaking, Figure 5.4 shows the curves of $-\Delta^2 Z_u$, the additional expected runs value allocated to the next over compared with the current over, for 0 (solid line), 2 (dashed line) and 4 (dotted line) wickets lost, for our modified D/L model of $Z(u,w)$. We show these curves as the negative of $\Delta^2 Z_u$, so that the plot provides the change in runs allocated to consecutive overs as the innings progresses from left to right. It is noticeable that for each given number of wickets lost the rate of change of over-by-over runs value, with respect to the progression of the innings, tends to decline at some point of the innings. For example, keeping all other factors constant and given two wickets lost, the change in value of consecutive overs tends to decline after about forty overs (u=*10*) (see the dashed line in Figure 5.4).

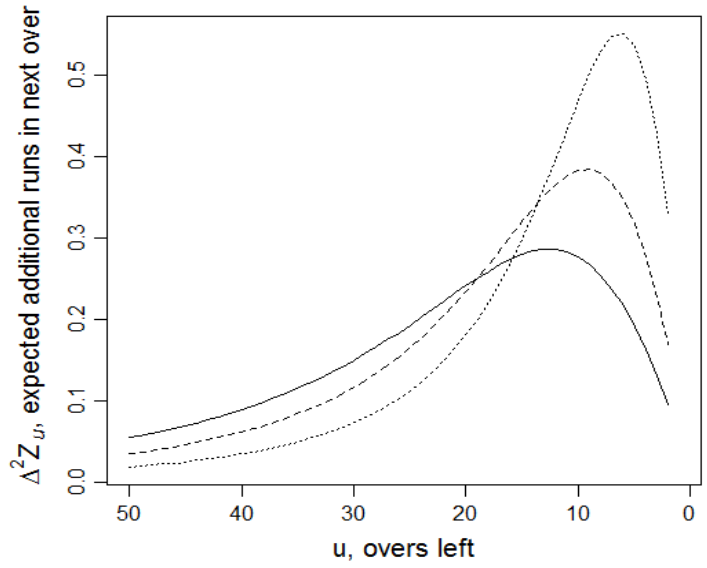

Figure 5.4 The plot for expected additional runs value, -$\Delta^2 Z_u$, against, *u* overs left, for *w=0, 2,* and *4,* using our modified D/L model in equation 5.2

Mathematically, it can be shown for our model that for each given *w*, there exists a point of inflection in the curves for Z_u'. Further, Figure 5.5 shows the curves of expected next over runs, measured as $\Delta Z_u = Z(u,w) - Z(u-1,w)$, that reflect the shape of Z_u' for

the adjusted D/L model (Figure 5.5a) and for our modified D/L model (Figure 5.5b). It is seen in Figure 5.5b that for each of the curves, one associated to each *w*, there exists a point at which the curves of the over-by-over runs-value changes from concavity to convexity as *u* approaches to zero. However, such runs scoring pattern is not observed in Figure 5.5a for the D/L model.

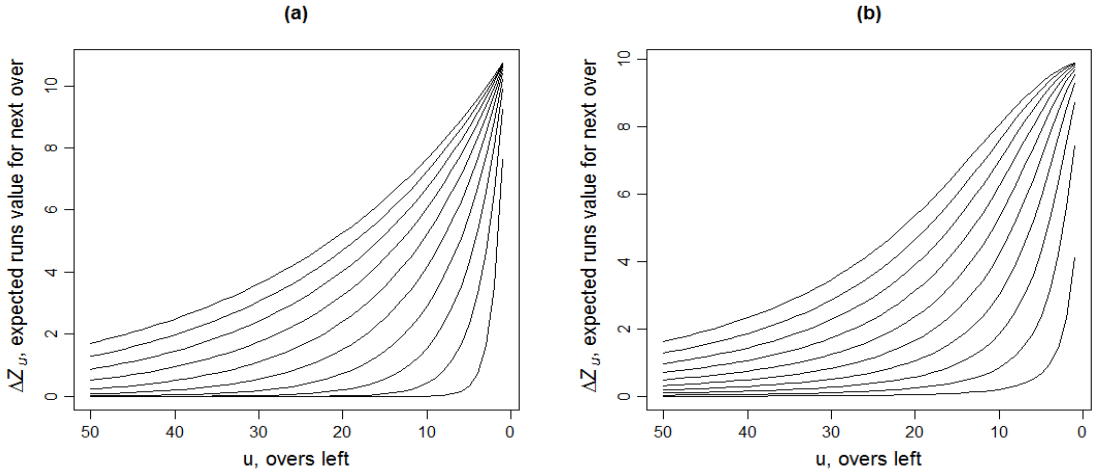

Figure 5.5 A plot of expected runs value , ΔZ_u, for the next over against overs left, u, for *w=0, 1,..,9* using, (a) Adjusted D/L model (b) Modified D/L model.

Moreover, once again it is demonstrated in Figure 5.5, that how our model allows for the rate of increase of the over-by-over runs to slow down for low wickets lost with few overs remaining as overs remaining decreases. Similarly, as seen in Figure 5.5b, there is a slower decay towards asymptotes (see the curves from right to left) that implies a heavier tail, which is especially evident when there are many wickets down in early stage of the innings.

5.3.3 Goodness of fit

In regards to the goodness of fit of the two models: the adjusted D/L model and our modified model, it is observed that our proposed modified D/L model has a lower weighted sum of squared errors (*WSSE*). Recall that weighting function accounts the heteroskedasticity and consistency of the mean remaining runs. Table 5.1 shows the estimated parameters and goodness-of-fit measures for both of these models. Further, Figure 5.6 show plots for observed mean remaining runs, and the fitted lines for the D/L model (solid lines) and our modified D/L model (dashed lines). In addition to the smaller *WSSE*, it can also be observed visually in Figure 5.6 that our modified D/L model more

closely resembles the data. This is especially evident in plots (d), (e), and (f) of the Figure 5.6.

Table 5.1 Estimated parameters for Adjusted and Modified Duckwort-Lewis models.

Parameter	***Adjusted D/L model***	***Modified D/L model***
Z_0	291.9	340.0
θ_0, $(=1/b)$	26.69	22.64
μ_0	NA	-1.46
θ_1	6.027	23.66
μ_1	0.896	-33
WSSE	**1735.9**	**1607.1**

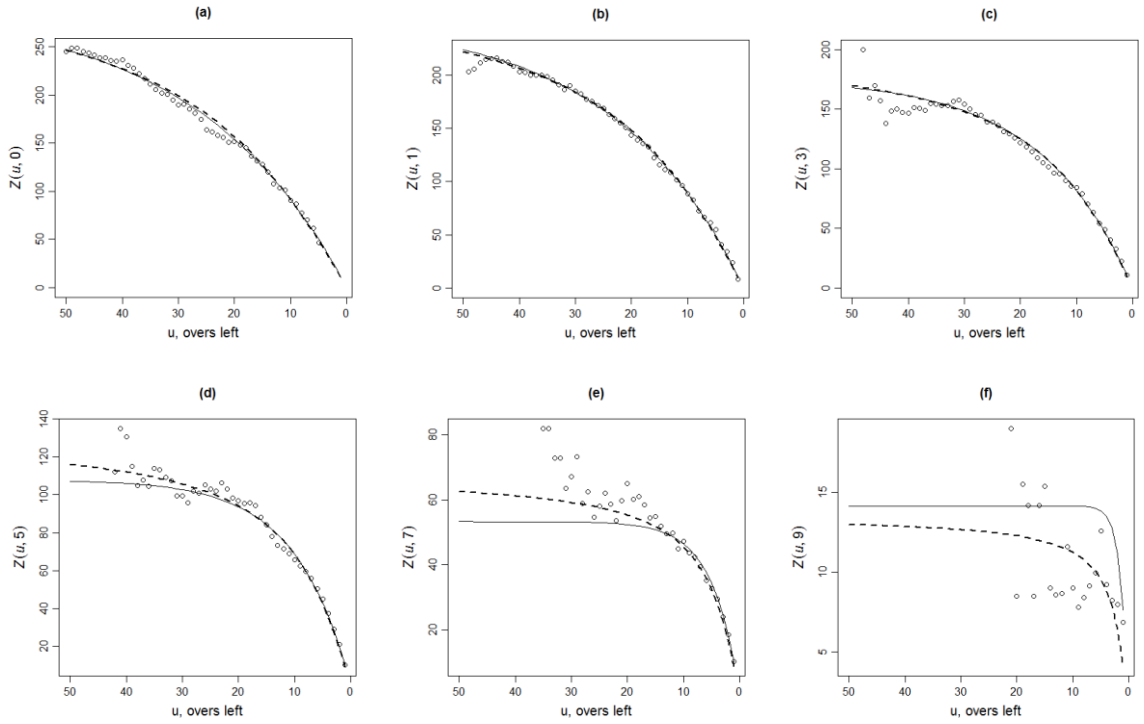

Figure 5.6 Plot of $Z(u,w)$ against u for given (a) w=0, (b) w=1 and (c) w=3 (d) w=5, (e) w=7 and (f) w=9, using the adjusted D/L model (solid lines) and modified D/L model (dashed lines). The circles represent the observed mean remaining runs, denoted by $\bar{x}(u, w)$.

Interestingly, although the choice of the functional form of $F(w)$ changes the *WSSE* a great deal for the original D/L model, we found that it did not have large effect on the *WSSE* for our proposed modified D/L model. For example using different $F(w)$, for example as presented in equation 5.3, in the modified D/L model gives a *WSSE* equal to 1606, whereas using the same $F(w)$ in the D/L model gives a *WSSE* of 1851. Hence, changing the function $F(w)$ in our modified D/L model reduces the value of the error function, *WSSE*, by 1, whilst for the D/L model it is increased by 116. The function, $F(w)$, based on the Weibull distribution is given by,

$$F(w) = \frac{e^{-\left(w/\theta_1\right)^{\theta_2}} - e^{-\left(10/\theta_1\right)^{\theta_2}}}{1 - e^{-\left(10/\theta_1\right)^{\theta_2}}} \qquad 5.3$$

where the parameters, $\theta_1 > 0$ and $\theta_2 > 0$, are scale and shape parameters associated to the Weibull distribution. We experimented with several functions for $F(w)$ and in regards to the goodness-of-fit, it is noted that the existing D/L model is highly sensitive to changes in the functional form of $F(w)$. However, in contrast, we note consistency in the goodness-of-fit when experimenting with different functions, $F(w)$, in our modified D/L model in equation 5.1.

5.3.4 Model adjustment for high scoring matches

Having developed a model which provides an improved fit to the data, and has more intuitive properties than the D/L model, we now incorporate an adjustment to our proposed modified D/L model that accounts for matches with well above average first innings totals. Using the same notation and interpretation of Duckworth and Lewis (2004), we introduce a match factor parameter, λ, to our proposed modified D/L model.

As Duckworth and Lewis (2004) assume, in high scoring matches the relationship between *Z* and *u* tends to become linear and the runs value of each wicket tends to zero. To understand this latter point, consider a match in which a team has 50 overs to bat and is chasing a target of 1800, so that six runs is required from each ball in the innings. This requirement is constant throughout the innings so that each ball has a constant value to the batting team irrespective of the number of wickets lost. In fact, for well above average run scoring matches the D/L method tends to approach the run-rate method.

In order to account for this effect, we scale up the parameters θ_0 and Z_0 in equation 5.2. This scaling allows the relationship between $Z(u,w)$ and u to be more linear in the range $0<u\leq 50$ for well above average runs. Hence, the model in equation 5.2 is altered to the following model given by

$$Z(u,w|\lambda) = Z_0\lambda^{n(w)+1}F(w)\left\{\frac{tan^{-1}\left(\frac{u-\mu}{\theta_0\lambda^{n(w)}F(w)}\right) - tan^{-1}\left(\frac{-\mu}{\theta_0\lambda^{n(w)}F(w)}\right)}{\frac{\pi}{2} - tan^{-1}\left(\frac{-\mu}{\theta_0}\right)}\right\}, \qquad 5.4$$

We estimate the parameter match factor (λ) in a similar fashion as we described in section 3.5.2 for the D/L model. Further, λ depends on team 1's total runs, S. Figure 5.7 shows the visual demonstration of Z curves for different values of $\lambda(S) \geq 1$. Note that we use the function $n(w) = 2 + 3F(w)$, the same as for the existing D/L method. Figure 5.7 shows how Z, the mean remaining runs, changes for different first innings total runs, S. We note that for large S the relationship between Z and u tends to linear.

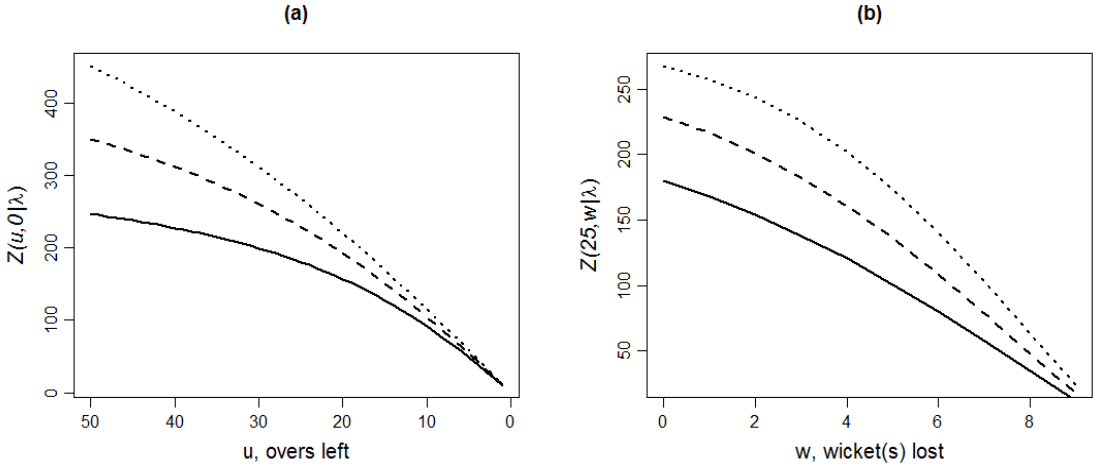

Figure 5.7 Modified D/L model, mean remaining runs (a) against *u* for *w=0*, and (b) against *w* for *u=25*. The solid lines are for λ(246.5)=1 , the dashed lines are for λ(350)=1, and the dotted lines are for λ(450)=1.172

5.3.5 Testing the model adjustment

In ideological point of view, the variation in the resources pattern with respect to S is intuitive. However, to our knowledge no statistical evidences are existed in literature that justifies the ad hoc addition of the parameter λ to the D/L model. To show that the

introduction of the parameter λ significantly improves our modified D/L method, we do the following numerical experiment.

Consider LOI matches, assume $Y_i(u,w)$ is team 1's runs when u overs are remaining and w wickets have been lost, where, $i = 1,2,..,n(u,w)$. Let $n(u,w)$ be the number of innings that are not ended by the time u overs remain given that w wickets have already been lost (see Table 3.1). Suppose in the i^{th} match, $\hat{S}_i(u,w,\lambda)$ is team 1's total runs, as predicted by our modified D/L model with u overs remaining and w wickets lost. We estimate λ for each i and each given u and w, using the methodology described in section 3.5.2 for the modified D/L model. The projected runs is thus given by

$$\hat{S}_i(u,w,\lambda_i) = Y_i(u,w)/\{1 - P_N(u,w|\lambda_i)\} \qquad 5.5$$

where $P_N(u,w|\lambda_i) = Z(u,w|\lambda_i)/Z(N,0|\lambda_i)$, is the remaining resources. Now let S_i be the actual runs the team scores in the completed first innings. Than for each given u and w, the mean absolute error is written as

$$MAE(u,w) = \sum_{i=1}^{n(u,w)} \left|\hat{S}(u,w,\lambda_i) - S_i\right|/n(u,w) \qquad 5.6$$

To estimate the total error we use a weighted sum of the mean absolute error (*WSMAE*)

$$WSMAE = \sum_w \sum_u k(u,w)MAE(u,w) \qquad 5.7$$

where $k(u,w)$ is defined in equation 3.6. Table 5.2 gives *WSMAE* for ODI and T20I matches with forecasts based on the model with λ and without λ (which is equivalent to setting λ = 1). The addition of the λ parameter clearly improves the forecasting power of the modified D/L model since the *WSMAE* is considerably lower for both ODI and T20I cricket.

Table 5.2 Goodness of fit measures for forecasted innings totals with and without λ in our newly proposed modified Duckworth-Lewis model.

Error function	*Without λ, using Eq. 5.2*	*With λ, using Eq. 5.4*	*Number of matches*	*Number of forecasts*
$WSMAE_{ODI}$	5311.1	3039.6	458	19844
$WSMAE_{T20I}$	966.9	823.3	191	2844

Lastly, we address the issue that the D/L method uses an ad hoc way of calculating the par score, *T*, in the situation when team 1's available resources are greater than team 2's total available resources (see section 5.2). We note that in addition to the empirical evidence for using λ, such a modification corrects the shortcoming of revising a target by direct scaling. Consider the example given by Duckworth and Lewis in which a team batting first scores 80 runs for the loss of no wickets in 10 overs when play is interrupted and the match is reduced to ten overs per side. For this example, using the D/L model gives $R_1 = 0.0782$ and $R_2 = 0.3702$. Hence, the revised target $T = 80 \times 0.3702/0.0782 = 378.8$, an impossible high target in just ten overs. As a result of this, Duckworth and Lewis (1998, 2004) suggest that in situations when team 2's available resources are greater than team 1's available resources, the revised target can be determined by $T = S + (R_2 - R_1)G(N)$. Since, in the given example, $R_2 > R_1$, therefore $T = 80 + (0.3702 - 0.0782) \times 245 = 151.4$.

We believe that the above shortcoming is not a consequence of using direct scaling, but rather, it is because no adjustment was made in the Duckworth and Lewis (1998) model to account for high scoring matches. For the above example our estimate of λ is 1.2295. Using this value for our proposed model, gives $R_1 = 0.156$, and $R_2 = 0.236$, and therefore, the revised target is $T = 80 \times 0.236/0.156 = 121.1$, which is a more acceptable revised target in ten overs. Hence, given the addition of the λ parameter to take account of high scoring matches, there is no need to use an ad-hoc scaling as suggested for D/L method in equation 2.1. Thus, we use

$$T = SR_2/R_1 \quad \text{if } R_2 \leq R_1 \text{ or } R_2 > R_1, \qquad 5.8$$

to calculate par scores for the team batting second in interrupted matches.

5.4 Modified D/L model and future research work

Our new model overcomes some shortcomings of the existing D/L model, however there remain some issues with the model that might be addressed in future research.

Firstly, like the existing D/L model, our modified model does not account for fielding restrictions during the power-play overs. As a consequence, a team that does not take the batting power-play before the interruption would be in an unfavourable situation as compare to a team which did take their power-play before the interruption.

Secondly, similar to the existing D/L model, the modified D/L model takes account of the order of the wicket partnerships, but does not account of the order of the striker and non-striker batsmen. Consequently, in some situations, an interruption to play can be advantageous and in some cases, it can be a disadvantage to the batting team. For example, let a team has lost eight wickets and then be deprived of the final five overs. Using the D/L method (or even the modified D/L method) means that the revised target is independent of the orders of the two batsmen. Clearly, in such situation the batting team would get advantage if the two batsmen are playing at numbers 9 and 10, and would be in unfavourable situation if the batsmen were at positions 1 and 10. Future research work might focus on developing a model that differentiates between such cases.

5.5 Summary

This chapter begins with identifying and highlighting issues related to the existing Duckworth-Lewis method. First, it is graphically demonstrated that for each given stage of an innings the relationship between wicket-resource-value and wicket lost is non-smoothed and unintuitive in the current D/L method. Second, for a given number of wickets lost the rate of increase in the runs value of consecutive overs is exponentially increasing irrespective of the situation. Third, the decay towards the asymptotes are very rapid, and as consequence a zero runs value is awarded for the loss of some overs during an innings. Finally, it is argued the par score should always be calculated by the direct scaling of each teams available resources, and therefore the revised target should be independent of $G(N)$, the average first innings total runs.

An improved version of the Duckworth-Lewis method is proposed that uses an alternative model to estimate the resources. The model is based on the Cauchy distribution and is more representative of the data. We refer to the Duckworth-Lewis method that uses our newly proposed model as the modified Duckworth-Lewis method. Further, it is shown that our proposed model provides a superior fit to data and, in addition, has a more intuitive behaviour in regards to the runs scoring pattern of the limited overs cricket. Further, we note that the modified D/L model is more flexible when compared to the existing D/L model in that, if the existing D/L model better reflects the runs scoring pattern in cricket, our modified D/L model would achieve the same, however the reverse is not necessarily true.

Moreover, it is shown empirically that it is appropriate to take account of high scoring matches in a similar way to Duckworth and Lewis (2004) do for their original model. We test the forecasting accuracies of our proposed modified form of the D/L model with and without the ad hoc adjustment to the model. It is observed that the addition of an extra parameter, λ, to the model improves the predictive accuracy of first innings total runs. Hence, it is argued that adjustment in the model for well above runs scoring situation is justifiable, not only on ideological point of view, but also on statistical results. Finally, we highlight future work for further improvement in the Duckworth-Lewis method.

CHAPTER 6 IN-PLAY FORECASTING IN CRICKET AND GENERALIZED LINEAR MODELS

In this chapter, a literature review on forecasting in cricket is given. Further, the class of generalized linear models (GLMs) is briefly described. The GLMs provide basis of our own in-play *dynamic* logistic regression model. Some model diagnostics are also described and we use these later for the model identification in CHAPTER 7.

6.1 Introduction

Unlike soccer, American football, baseball and tennis, relatively little work has been published on forecasting in cricket. This seems especially strange given there is known to be a huge betting market on cricket. The work that has been done on forecasting in cricket has largely been concerned with pre-match forecasting. For example, Brooks, Faff, and Sokulsky (2002) propose a method to estimate test match outcome probabilities (pre match) using an ordered response model. 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 particularly lends itself to betting in-play: unlike soccer for example, the discrete nature of the game means bookmakers and punters alike have ample opportunity to be active in markets during the game and as such, cricket attracts extremely large in-play betting volumes. For example, total volume bet during a major One-Day International (ODI) involving Pakistan or India is of the order of $1bn.

Previous chapters of this thesis, have focussed on the problem of interruption to play. Indeed, this is mirrored in the academic literature, with several papers appearing discussing ways to deal with interruptions in play. Further, a considerable large work in literature is also available on strategies to play in cricket. For example, Clarke (1988), Preston and Thomas (2000), Norman and Clarke (2007), and, Scarf and Akhtar (2010) mainly focus on optimum strategies in cricket. However, there has been relatively little work has done that directly focussing on the problem of forecasting. Of the work that does exist, Preston and Thomas (2002) proposed a method to estimate match outcome probabilities using dynamic programming techniques. Allsopp and Clarke (2004)

forecast teams total runs in an innings to measure each team's relative strengths. Similarly for test cricket, using a multinomial logistic regression model, Scarf and Shi (2005) forecast match outcome probabilities with the specific aim of helping team management to decide on the most appropriate time to declare in an innings. Akhtar and Scarf (2012) further extend this work and developed in-play, session-by-session, forecasting models. The work most related to ours is that of Bailey and Clarke (2006) who developed a forecasting model for limited overs cricket to predict the margin of victory before the match begins, but, with the help of Duckworth and Lewis (1998) method, update these predictions in an ad hoc way whilst the game is in progress.

In the next section we give overview about the Bailey-Clarke and Akhtar-Scarf approaches to estimate match outcome probability forecasts in-play. Section 6.3 describes the basic idea of generalized linear modelling. In section 6.4 some standard model diagnostics are described. Lastly, the summary of the chapter is given in the section 6.5.

6.2 Bailey/Clarke and Akhtar/Scarf approach for in-play forecasts

Bailey and Clarke (2006) propose a method, the B/C method, of estimating in-play probabilities while the game is in progression. They proposed two regression models. One regression model forecasts the margin of victory (MOV), and another forecasts the team total runs (TTR) in total pre allotted overs (N) for an innings. The covariates that were used in these models are, home venue advantage (a categorical variable), reference team past performance against the opposition (difference in the averages), and the current form of the team (based on last ten matches). In case when team team 2 wins the match, the value of the two response variables are than be calculated using the Duckworth and Lewis (1998) model. These regression models only account the pre-match effects and therefore can be considered as pre-match forecasting models. Further, to estimate the in-play MOV forecasts they use the following equation,

$$\text{in-play MOV} = \text{pre-match MOV} + \text{in-play TTR} - \text{pre-match TTR}, \qquad 6.1$$

where the in-play TTR is the predicted team total runs that are estimated using the Duckworth-Lewis model. From equation 6.1, it is clear that the only term that varies with respect to the progression of the innings is the Duckworth-Lewis predicted team total runs. The other two terms, the pre-match MOV and TTR, remain constant as the innings

progresses. The in-play MOV can further be used to estimate the match outcome probability. It is to be noted that the negative value for MOV indicates the reference team (the batting team) loses the match. In regards to the in-play probability forecasts, Bailey and Clarke (2006) suggest dividing the predicted MOV by its standard error and comparing with the standard Normal distribution. However, it is unclear how the standard errors of the in-play MOV can be estimated.

There are two main issues remaining with Bialey-Clarke approach of the in-play forecasts. First, the in-play estimated probabilities are updated, with respect to the progression of the game, by an ad hoc way. Second, the effects of various factors, runs scored for example, on the in-play MOV are constant throughout the game.

Akhtar and Scarf (2012) adopt the approach of fitting a series of independent multinomial logistic models for test cricket with response variable (Win/Draw/Loss). They estimated 15 separate multinomial logistic models that can be used at 15 particular stages of a test match (at the end/start of each session). Such an approach allows the effects of the covariates to vary with respect to the session-by-session progression of the game. A number of in-play and pre-match covariates are used as predictor in each these models. In regards to the pre-match covariates, the win percentage difference (*wd*), the ICC test cricket ratting difference (*rd*), home advantage (*home*) and ground effect (*g*, quantified as the proportion matches that were drawn as compared to total matches played on the ground) were used. Whilst, regarding the in-play covariates, *wickets* and *lead* were used that quantify the position of a team prior to a particular session. Further, the covariate *wickets* was transformed into wicket resources (*wr*). It is noted in the A/S method, the relationship between *wr* and *wickets* is non-smoothed and static with respect to the progression of a test match. We believe that this relationship should account the stage of the innings and therefore should vary with respect to the progression of the match (This point will be explained further in the section 7.2.2). In addition, we note that at any point during a session, the pre-session model of the Akhtar and Scarf cannot be used to estimate win probabilities. Rather one has to wait until the session has ended. Moreover, if such approach is used for limited overs cricket the non-smoothed variation in effects of the covariates can result in unstable probability forecasts in-play (ball-by-ball or over-by-over).

Our approach of modelling for in-play forecasts (described in CHAPTER 7) is different to both the Bailey-Clarke (B/C) and Akhtar-Scarf (A/S). Comparing our approach of estimating in-play probability forecasts with the B/C approach, we note that they fit regression models with continuous response variables; however, we fit binary response models that are dynamic in its parameters. Further, unlike the B/C approach for in-play forecasting, we do not update the estimated probability in an ad hoc way after each ball of the game. Rather, in our dynamic logistic regression model the estimated parameters are allowed to evolve smoothly with respect to the progression of an innings.

In regards to comparing our approach of modelling for in-play ODI cricket with that of Akhtar and Scarf (2012) method for test cricket, we note that partially our approach is similar to the A/S method. However, unlike A/S approach we do not use a series of independent models to forecast probabilities in play, but we use a single dynamic logistic regression model to forecast match outcome probability at any point of given ODI innings. Further, their 15 models estimate match outcome probabilities at fifteen distinct moments during a test match, namely in-between each session. Our two models, one for each innings of ODI cricket, can be used to forecast the match outcome probability after each ball of the game (from first to last ball of an ODI game).

6.3 Generalized Linear Model (GLM)

Nelder and Wedderburn (1972) generalizes the analysis of the variance model and introduces the class of generalized linear models (GLMs). The GLM is an extension of the classical linear model that does not require the assumption of normality for disturbance term. Further, the predictor in GLM needs not necessarily have to relate linearly to the response, but can have linear relationship with some function of the response variable. Moreover, it can be applied to categorical and discrete data where the classical linear regression model is not applicable (McCulloch & Searle, 2001).

The generalized linear model is defined as modelling the conditional mean of response variable that belongs to the exponential family via a link function. This allows for regression modelling for a non-normal and non-continuous response variable with some degree of non-linearity in the model structure (Dobson, 2001).

Trevor. Hastie and Tibshirani (1986) further extend the class of GLMs by transforming the linear form of the predictor into the sum of smoothed splines. This class

of models is known as generalized additive models (GAMs). Generalized additive models usually provide better fit to data as compared to the GLMs. However, it can often result in over-fitting problem. Cross-validation techniques facilitate detection and reduction of over-fitting. However, generally speaking, GLMs may be preferable to GAMs unless the GAM improve the forecasting power significantly for the application under consideration (T. J. Hastie & Tibshirani, 1990; Wood, 2006).

The class of the GLMs can be specified by identifying the three components. A *random component* indentifies the response variable, *Y*, and its distribution function. A *systematic component*, which identifies the covariates included in the model and a *link-function* which relates the mean of the response to the systematic component (Agresti, 2002, 2007). For instance, consider the match outcome; *Y* (Win/Loss) is a binary response variable. Assuming the binomial distribution, let the probability of a success (a win) is denoted by *p*. Taking the logistic function as a *link-function* that equates to the linear predictor $\boldsymbol{X}$ (vector of covariates with first element as 1 if the model involves intercept term). Further, let the response variables are independent, then the class of GLMs known as binary logistic models. The basic mathematical structure for such a model can be written as

$$E(Y|\boldsymbol{X} = \boldsymbol{x}) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = p \qquad 6.2$$

$$log\left(\frac{p}{1-p}\right) = logit(p) = \boldsymbol{x^t\beta} \qquad 6.3$$

where the function $logit(p)$ is the link function and $\boldsymbol{\beta}$ is the vector of parameters. Table 6.1 describes some link-functions that are used commonly in the class of generalized linear regression models (GLMs).

Table 6.1 Some link functions for the GLMs

Family	Link
Binomial	$log\left(\frac{\mu}{1-\mu}\right)$
Gamma	μ^{-1}
Inverse Gaussian	μ^{-2}
Normal (Gaussian)	μ
Poisson	$log(\mu)$

6.4 Model diagnostic measures

6.4.1 Test the significance of association

In the literature, there are various standard statistical procedures available to test the significance of the relationship between predictor and response variables in regression models. For example, D. W. Hosmer and Lemeshow (1980) proposed a procedure that is used to test the goodness of fit for the binary logistic model. The test identifies subgroups as deciles of the predicted probabilities and compares these with the observed rate. The test statistic asymptotically follows the Chi-square distribution (Hosmer and Lemshaw, 2000).

Some other methods, for example the *Wald test* (Wald, 1943) and Log-likelihood ratio test (Wilks, 1935, 1938), also test the statistical significance of association in the GLMs. These methods are based on the log-likelihood function and are valid for large sample sizes. The test statistic for the log-likelihood ratio test is given by

$$LR = -2log(L_0/L_1) = -2(l_0 - l_1) \qquad 6.4$$

where l_0 and l_1 are the maximum log-likelihoods for the null and candidate models respectively. Wilks (1935) shows that *LR* asymptotically follows the null chi-squared distribution.

6.4.2 The strength of association

The procedures, as describe in previous section, identify only the statistical significance of the associations. However, in practice one needs to know the strength of such associations, which are expressed normally in percentages. In the classical linear regression modelling, the multiple coefficient of determination (R^2) provides the percentage of explained variability in response variable by the predictor. However, the same statistic is not applicable for the models with categorical response variables. Therefore, Cox and Snell (1989) proposed the pseudo R^2, which could be used to measure the strength of association in categorical response model. The Cox-Snell statistic is based on the maximum likelihood function and is defined for a candidate model as

$$R^2_{cs} = 1 - \left(\frac{L_0}{L_1}\right)^{2/n} \qquad 6.5$$

where, L_1 is the maximum likelihood with all selected covariates, and L_0 is the maximum likelihood with only intercept term. The maximum value of Cox and Snell's pseudo R^2 is

$1-(L_0)^{2/n}$. Nagelkerke (1991) generalizes the definition of the coefficient of determination and proposed a Nagelkerke's R^2 which is a modified form of the Cox-Snell pseudo R^2 that ranges from 0 to 1. It is defined as

$$R^2_{Neg} = \frac{1-\left(\frac{L_0}{L_1}\right)^{2/n}}{1-(L_0)^{2/n}} \qquad 6.6$$

The Nagelkerke's R^2, as given in equation 6.6 can be expressed in percentage by multiplying it by one hundred. In the broad sense, R^2_{Neg} shows the percentage of variability in the response variable that is been explained by the predictor. The R^2_{Neg} can also be interpreted as the percentage of explanatory power of the candidate model as compared to the model with only an intercept term. We will use this statistic to measure the strength of the association between match outcome and the covariates during any point of the game.

6.4.3 Model selection

In statistics, model selection is an important branch of any data analysis. The data can be used for modelling in many different ways. So, what form of a model is best among the class of other available models? Similarly, if there are many covariates which potentially have an effect on the dependent variable, should all of these covariates be included to make a model best? A model that is rich enough to explain the relationships present in the data, but on the other hand is simple enough to *easily* explain these relationships, is known as a parsimonious model. Many procedures for model selection are existed in literature. However, none of them attributed as a best in general. Many of these methods are defined in terms of information criterion (IC). The IC is basically a score associated with the candidate model that is based on data and complexity of the model (Agresti, 2002; Claeskens & Hjort, 2008).

Akaike information criterion (AIC) (Sakamoto, Ishiguro, & Kitagawa, 1986) and Bayesian information criterion (BIC) of (Schwarz, 1978) and (Akaike, 1977, 1978) are the two commonly used procedures that are based on data. Both the AIC and BIC are defined in penalised log-likelihood form. Suppose, there is a candidate model *C* , then the AIC is defined as

$$AIC(C) = 2\,\text{log-likelihood}_{max}(C) - 2\,\dim(C) \qquad 6.7$$

whereas the BIC is defined as

$$BIC(C) = 2\,\text{log-likelihood}_{max}(C) - (\log n)\,\dim(C) \qquad 6.8$$

where $\dim(\mathrm{C})$ is the length of the vector of parameters, $\boldsymbol{\beta}$. The second terms in both criterions are referred as penalty terms. Equations 6.7 and 6.8 show that both procedures are capable of keeping the balance between simplicity and complexity (a model with too many parameters). The model with largest AIC or BIC value is ranked as the preferred model (Claeskens & Hjort, 2008).

The values of the AIC and BIC are meaningless when looked at in isolation. Therefore many authors, for example Agresti (2002) and Congdon (2005) describe the procedures as the negative of equations 6.7 and 6.8. The R built-in functions for these measures also return the value in similar fashion (R Development Core Team, 2012). Therefore, we use these measures in similar way, and would rank 'one' a model with smallest value of AIC/BIC.

There are various advantages and disadvantages that are associated with AIC and BIC. The choice of selection between these criterions is subjective. For example, if simple model is desirable than BIC is preferred to use as it has greater penalty for adding model parameters, provided that sample size is greater than 8 (log 8=2.08). In our analysis, the sample sizes are always greater than 8, therefore BIC would always produce a simple model when compared to the AIC. Claeskens and Hjort (2008, p.70) stated that "the BIC successfully addresses one of the shortcomings of AIC, that the latter will not succeed in detecting 'the true model' with probability tending to 1 when sample size increases". This property of the model selection criteria is known as consistency. On other hand, Claeskens and Hjort (2008) argue that if efficiency is required, the AIC is preferable, as it is more associated with precise prediction than BIC. Yang (2005) tries to combine the consistency strength of the BIC with the efficiency strength of AIC, but fails to do so.

One of the problems related to the above procedures is to use the same data for model fit and model selection criterion. This would especially be an issue if the model is used for future forecasts. Therefore, other commonly used methods are Cross-Validation (CV) techniques. In such methods, the model with the best forecasting ability is selected from the candidate models. In this method, the data is divided into two parts. One part of the observations is used for model fit and is known as training set of data. Another part of

data is used to assess the model forecasting accuracy and is known as validating set of data. The model with the greatest forecasting accuracy among the candidate models is considered to be the best model. There are many types of Cross-Validation techniques available in literature. However, the Delete-d Cross-Validation (CV_d) of Shao (1993) and K-folds Cross-Validation (CV_{KF}), as described in (Trevor Hastie, Tibshirani, & Friedman, 2009), are two commonly used techniques.

Shao (1993) proposes a Cross-Validation method, which he refers as Delete-d Cross-Validation. In this method, a random sub-sample of size *d* is deleted from the sample data prior to fitting model. The deleted *d* observations are then used as the validating set to assess model forecasting accuracy. The process of sub-sampling is repeated many times, and an average prediction error of sub-samples, also known as cross validation score, is calculated. Further, Shao (1997) suggests using $d= n(1-1/(\log n - 1))$.

In regards to the K-folds Cross-Validation (CV_{KF}) method, a sample data are partitioned randomly in K folds or clusters. Then each one-fold data cluster is used as the validating set whilst the remaining data in K-1 folds are used as the training set. This process is continued until all folds are used once as the validating set to measure forecasting errors. Once all the K-fold data have been used exactly once as a validating set then the value of the prediction error is determined. To get a consistent estimated cross validation prediction error, the process of random partitioning of the sampled data should be repeated for some number of times. In regards to the model selection using the CV_{KF} method, generally a one-standard-deviation rule is used. In this procedure of model selection, a most parsimonious model whose cross-validation error is not greater than one standard deviation of smallest cross-validation error of a model is than selected. Further, the leave-one-out cross validation (LOOCV) is a special case of CV_{KF} model selection method, if the number of folds are equal to the size of the sample, *n*. Typical choices for the number of folds are 5 or 10 (Trevor Hastie et al., 2009).

The question of which of the above four methods is best for model selection is dependent upon the purpose of the modelling and the list of candidate models. For example, Shao (1997, p.223) argues that the crucial factor that determines the asymptotic performance of almost every model selection method is whether or not the candidate list of models contain some correct models. Further, if the aim is to use a model as a forecasting tool, and to avoid model complexity and over-fitting, it is appropriate to use a

Cross-Validation technique. The detailed answer to the question, which method to adopt for model identification, is beyond the scope of this thesis. Therefore, we use all the above-mentioned methods for our model selections in the next chapter to decide which covariates should be included in our final forecasting model.

6.5 Summary

This chapter starts with the overview of the work in the literature that mainly focuses on forecasting in cricket. It is noted that in limited overs cricket, most of the research focuses on interruption and strategy to play. However, some work directly relates to forecasting in cricket. Most of it is been related to pre-match forecasting, and little work is available regarding in-play forecasting in cricket. For example, Bailey-Clarke and Akhtar-Scarf attempt to generate in-play forecasts in ODI and 'test' cricket respectively.

Further, the generalized linear model (GLM) is overviewed briefly. The GLM is an extension of the classical linear model that does not require the assumption of normality. Further, in GLM the predictor does not necessarily have to relate linearly to a response variable. Moreover, unlike in classical linear regression models, the response variable in GLMs might be discrete or categorical.

Lastly, in this chapter, some model diagnostics are briefly described. These include, the Hosmer-Lemshow, and Likelihood-Ratio (LR) tests for significance of association. Further, to measure the strength of association between the covariates and response variable, some pseudo R^2 statistics are discussed. The model selection procedures, AIC, BIC, and some Cross-Validation methods are also overviewed. Further, the advantages and disadvantages of these methods are also briefly discussed.

CHAPTER 7 IN-PLAY FORECASTING OF WIN PROBABILITY IN ONE-DAY INTERNATIONAL CRICKET: A DYNAMIC LOGISTIC REGRESSION MODEL

This chapter presents a model for forecasting the outcome of One-Day International cricket matches whilst the game is in progress. This 'in-play' forecasting model is dynamic in that the parameters of the underlying logistic regression model are allowed to evolve smoothly as a match progresses. Using our proposed dynamic logistic regression (DLR) model not only allows the parameters to evolve smoothly, but also less number of parameters are required to estimate as compared to the series of independent models (one for each ball of the game). With the help of our DLR models (one for each first and second innings), we analyse how the effect of covariates vary with respect to the progression of an innings. The model identification is done using the model selection methods as discussed in the previous chapter. Further, the forecasting accuracies of our proposed models are assessed using the cross validation approach. 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 quantitatively similar; in fact, the probability forecasts using our DLR models can be considered to be 'correct' at an earlier point in the games than the probabilities inferred from the betting market. These results we take as additional evidence that our modelling approach is appropriate.

7.1 Introduction

In regards to in-play probability forecasts, two dynamic logistic regression (DLR) models are developed, one for each innings of ODI cricket. Two types of covariates are used as predictors in our DLR models. These are referred to as pre-match and in-play covariates. In regards to our approach of modelling, first, we identify and fit a series of 'best' logistic models, one for each ball of the game. For this purpose, we use four different methods for model selection. Namely, Akaike's Information Criteria (AIC), Bayesian's Information Criteria (BIC), Delete-d Cross-Validation (CV_d) method, and K-folds Cross-Validation (CV_{KF}) methods are used. These model selection methods have been briefly described in section 6.4.3. Second, we transform the series of independent

logistic models into two single DLR models, one for the first innings and one for the second innings.

In the next section, the data and covariates are described. In section 7.3, we present the modelling procedure to develop a dynamic logistic regression (DLR) model that forecasts probabilities of match outcome, in-play (ball-by-ball). Section 7.4 describes the results for model identification, fits, and diagnostics. In section 7.5, we use two recent out-sample matches to compare our predicted probabilities with those of the betting market. In section 7.6, we highlight an issue related to our DLR models and possible future research in this regards. Finally, we provide a summary of the chapter with some closing remarks in section 7.7

7.2 Data and covariates

We obtained ball-by-ball data for ODI matches played from January 2004 to February 2010 from the Espncricinfo website. We do not include matches, for which data were incomplete, or in which one of the teams had played less than five matches, or in which play was interrupted due to rain or bad light. In total, we fit our model to data from 606 ODI matches. In addition to ball-by-ball information, for ground analysis, we use the *statguru* application on the Espncricinfo website to collect summary statistics from data for ODI matches from January 1992 to February 2013.

We have collected data for a number of variables, which can be used as covariates. These covariates are divided into two categories: pre-match covariates (to be measured prior to the start of the match) and in-play covariates (to be measured only during the play). The subsequent sections explain how the covariates describe the pre-match and in-play position of a team quantitatively. Further, the intuition of variable transformation is also been discussed.

7.2.1 Pre-match covariates

Pre-match covariates are those quantitative measures that could be determined prior to the start of the game. There are number of factors that might affect the probability of match outcome before the play has commenced. For example, home venue advantage, winning a toss, day-night effect, team's experience, and team's current form can considered as pre-match situation.

In any format of the cricket game, it is a common opinion that a team might have home venue advantage if playing on the home ground. It is because the home team will typically have played many matches at the home venue, and therefore they are well familiar with the home venues condition. For example, considering the ODI matches played during 1992-2013, Sri Lanka's win percentage on their home ground is about 70%, which is reasonably higher than their win percentage on 'away' and 'neutral' grounds that is approximately equal to 48%. Similarly, during the same period (1992-2013), India's win percentages on 'home', 'away', and 'neutral' grounds are 63.2%, 45.2%, and 54.6% respectively. Therefore, to account for home venue advantage, we use a categorical variable that takes values on *home*, *away*, and *neutral* venues.

Similarly, winning a toss (to decide to bat in first or second innings) in cricket is also considered as an advantage to a team. However, in the literature its effect on the match outcome has been found to be statistically insignificant (see, for example, Allsopp and Clarke (2004), Bailey and Clarke (2006) and Akhtar and Scarf (2012)). We experiment with including the dichotomous variable *toss*, taking the value 1 if a reference team won the toss and 0 otherwise, in our models. The results show that the main effect of the covariate *toss* is found to be statistically insignificant, however, its interaction effect with a binary variable day-night (*dn*), is observed to be statistically significant. In addition to experimenting with an interaction effect of the variables *toss* and *dn*, as denoted by *dnt*, we also experimented with all other two factor interaction effects between the categorical variables, but none of them were found to be statistically significant.

In regards to the general strength of reference team, we measure experience and performance against the opposition team. We use the difference in the ICC official ODI ratings (*rd*) for the two teams, as at the time of the match. The ICC official ratings reflect a team's performance based on the matches that are played in last three years. These ratings are calculated as the total points a team has earned divided by the total number of matches they have played in the last three years. Further, it is noted that matches that are played in the most recent year are weighted more than matches played a year before. In fact, the last three years are weighted as one-third, two-third, and a unit respectively. A team earns points at the end of each match. These points depend on the result of the match, and the strength of the opposition. For example, a team can get higher points if

they win against a higher ranked opposition team (for more details, see the ICC official website).

The ICC official ratings go some way to measuring a team's quality, but do not explicitly indicate team's current form. For example, a weaker team might be in good 'form' and could have high potential to win against some reasonably strong teams. For instance, Bangladesh, with an ICC rating equal to 62 at the time of play, was in good form in the ICC Asia Cup 2012 tournament. They had won against India, (with the ICC ratings 117 at the time of play), had won against the Sri Lanka (with the ICC ratings equal to 113 at that time) and had lost in the final competition in a close match against Pakistan. Therefore, we calculate a team's form as a weighted mean of match outcomes in the last five games. Specifically, let $y_t = 1$ if a team won the match and 0 otherwise, then we define the team's current form as,

$$form = \sum_{t=1}^{5} w(t,\theta) y_t \Big/ \sum_{t=1}^{5} w(t,\theta) \qquad 7.1$$

where, $$w(t,\theta) = \theta(1-\theta)^{t-1}, \quad t = 1,\dots,5 \text{ and } 0 < \theta < 1 \qquad 7.2$$

The *form* of a team as defined above is ranges from 0 to 1. A team will be in 100% form (*form=1*) if they won all the last five matches, and would have in 0% form (*form=0*) if none of the match is won in the last five matches. The function $w(t,\theta)$ is a discounting factor, so that the most recent match could get the highest weight. This implies that for given two teams with same number of wins, in the last five matches, might have different form's value depending upon the order of wins and $\theta>0$. Further, it is noted that as θ tends to unity the weights for most recent matches tends to larger. For example, the reference (a batting team) and opposition team with series of last five results LLLWW (last two matches are won after losing the first three matches) and WWLLL (first two matches are won before losing the last three matches) respectively. For this example, as shown in Figure 7.1a, the form value of the reference team is a positive increasing function, whilst the form value for the opposition team is a positive decreasing function of θ. To incorporate the reference team's current form against the opposition, we use the covariate *fd(θ)*, simply the difference in the forms for the two competing teams, for some suitable choice of θ. The solid line in Figure 7.1a shows the relationship between *fd* and θ for our last hypothetical example. Experimentations led us

to use $\theta=0.2$ (see section 7.4.2). Figure 7.1b demonstrates the geometric decay in 'weights' towards older matches for $\theta = 0.2$. It is to be noted that a win percentage difference (*wd*) is a special case of *fd* for $\theta=0$.

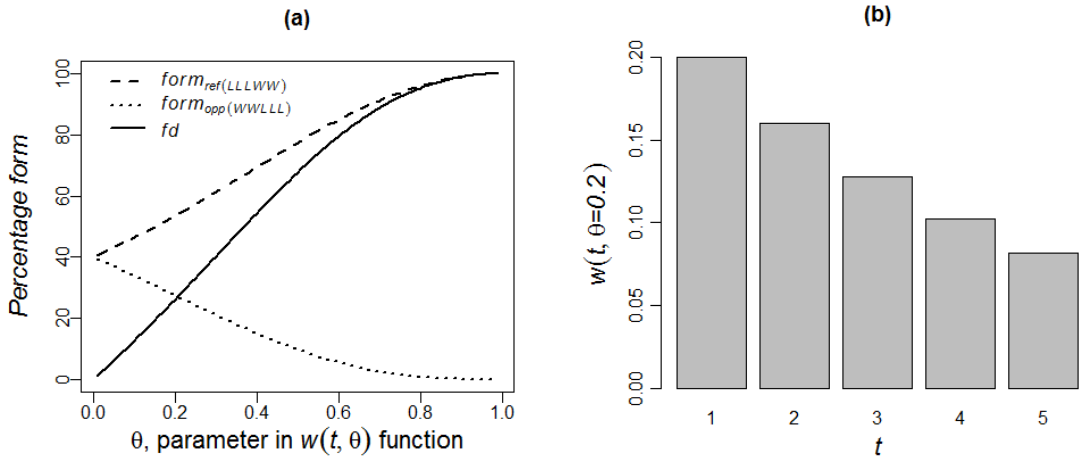

Figure 7.1 (a) Plots of the 'form' against θ and (b) Bar plot of the weighting function $w(t, \theta=0.2)$. Note that the batting team is set as a reference team.

7.2.2 In-play covariates

In regards to the in-play covariates that describe the changing state of play (or the position of a team) with respect to the progression of an innings, we need to incorporate three pieces of information. First, the number of runs being scored, (or the number of runs required to win in the second innings). Second, the number of wickets being lost, and third the number of balls, *k*, (or overs, *u*) remaining. We note that after each ball of the game these variables are changing. Regarding runs being scored, in the first innings, this is described by the run rate (runs per over) and is denoted by *rpo*, whilst in the second innings, the required run-rate (*rrpo*, the number of runs per over needed in remaining innings to win the match) replaces the run rate as the major in-play predictor of match-win probability. Note that *rpo* a function of runs scored and overs played, whilst *rrpo* is a function of runs scored, overs remaining, and the target score.

In regards to *wickets*, we transform the 'number of wickets lost' into the wicket resources lost (*wrl*). We believe that the value of losing a wicket should depend on which wicket has been lost and when in the match the wicket was lost. This is partly a consequence of teams putting higher quality batsmen at the top of the order. Further, as an innings progresses the relative importance of each wicket partnership changes. For

example, suppose there are five overs left and the batting team has already lost eight wickets. Then losing the next wicket should have larger impact on the expected remaining runs (and therefore on their win probability), when compared to a team that has lost only one wicket at the stage when there are just five overs remaining. We believe that in such a situation the value of losing the next wicket should intuitively have different values depending upon which wicket has been lost. In this regard, we define a covariate wicket-resources-lost (*wrl*) as the proportion of the expected runs value lost in the remaining innings for the loss of *w* wickets, as compared to expected runs with no wicket lost in remaining *u* overs. It can be written as

$$wrl = \frac{Z(u,0) - Z(u,w)}{Z(u,0)} \qquad 7.3$$

where $Z(u, w)$ is defined in equation 5.2 and can be interpreted as the expected runs in remaining innings such that *u* overs remaining and *w* wickets has already been lost. From equations 5.2 and 7.3 we have the following relationship between *wrl* and *w*,

$$wrl = 1 - F(w)\left\{\frac{\tan^{-1}\left(\frac{u-\mu}{\theta_0 F(w)}\right) - \tan^{-1}\left(\frac{-\mu}{\theta_0 F(w)}\right)}{\tan^{-1}\left(\frac{u-\mu}{\theta_0}\right) - \tan^{-1}\left(\frac{-\mu}{\theta_0}\right)}\right\} \qquad 7.4$$

where $F(w)$ is given in equation 5.1 and can be interpreted as the proportion of runs that are scored with *w* wickets lost compared to no wickets lost in hypothetically infinite remaining overs.

In regards to further discussion upon the intuition of the covariate *wrl*, we note that it is a continuous variable ranging from zero to one. We multiply it with 10 before using as a covariate in our model, since this has an intuitive meaning, as there are ten wickets available to each team in cricket. Further, we note that the relationship curves of *wrl* and *w* are dynamic and evolve smoothly with respect to the progression of the innings. Figure 7.2a demonstrates the relationship between *wrl* and *w* for each stages *u=50, 45, …,5* overs remaining. It can be seen that in the early stages of an innings (50 overs remaining for example), the relationship between *wrl* and *w* is more linear compared to the later stages of the innings. This implies that losing top order wicket partnerships in the later stages of the innings has a smaller wicket resources value compared to the losing a wicket of a lower order batting wicket partnership. This is somewhat intuitive as in the limited overs cricket a common strategy of the ODI cricket is to play defensive in the

early stages to save wickets, in preparation of playing more aggressively in the later stages of the innings. For example, recall the hypothetical example where a team is losing a wicket at a stage when five overs remaining. Given the stage of five overs remaining, using equation 7.4 meant the ninth wicket partnership resources value is 2.76, which is clearly greater than the second wicket partnership resources value (=0.148) at the stage of five overs remaining. It is implied that in latter case the team should intend to play with an aggressive strategy, as the risk-reward payoff is small. Figure 7.2b shows the plot of wicket resources value against the wicket number at the stage when five overs remaining. It can be seen that at this stage, losing a wicket number in 1-6 partnerships has a value smaller than a single wicket lost, and losing a wicket number in 7-10 partnerships has the value greater than a single wicket lost.

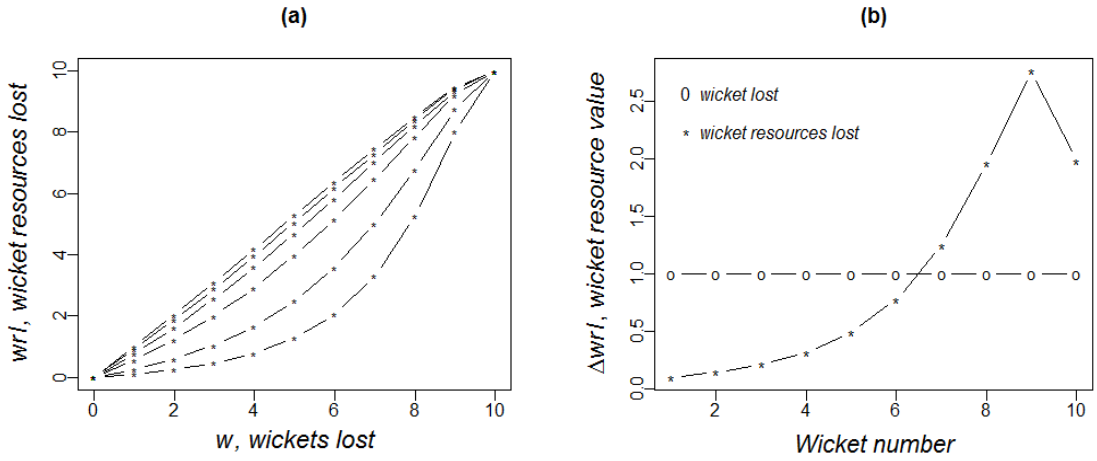

Figure 7.2 Plots of (a) curves for relationship of total wicket resources lost (*wrl*) and wickets lost (*w*) for each *u=50(top line) ,40,..,10,5(bottom line)* overs remaining, and (b) $\Delta wrl = wrl_{w+1} - wrl_w$, a wicket resource value and *wicket number* at *u=5* overs remaining

In addition to the intuition of the *wickets* transformation as discussed above, using *wrl* as a covariate also reduces the correlation between wickets lost (*w*) and runs-per-over (*rpo*). This is especially evident in the later stages of the first innings. The problem of multicollinearity should be taken into consideration, if the model is used to explain the relationship between covariates and response variable (match outcome). However, if the aim is only to predict the match outcome, then the amount of multicollinearity is not a serious issue. Interestingly, the correlations between the *w* and *rpo* are negative for each stage of the first innings. This is somewhat counterintuitive, as a common believe in

cricket is, if a team is intending to play with higher run-rate, this will normally result in a high risk of losing wickets. For example, Preston and Thomas (2002) assume a positive convex increasing relationship of a hazard of dismissal and run rate. In our study, the observed negative correlations, between *rpo* and *w*, show that teams do not have full control over the covariate *rpo*. We believe that the reason we observed the negative correlations between *w* and *rpo* during the first innings, is that both of these measures are in-play performance indicators. We believe that if a team is performing well until some point of an innings, then it is more likely that they will score with a higher run-rate (*rpo*) by losing less number of wickets (*w*).

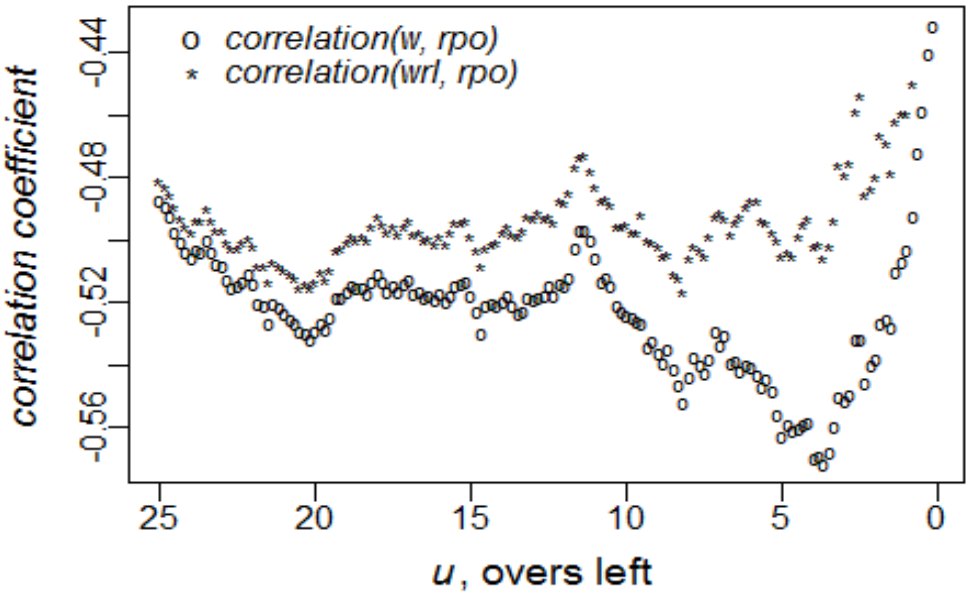

Figure 7.3 Plot for the series of Pearson's correlation coefficients of the number of wickets lost and run-rate during last twenty five overs of the first innings.

In experimenting with transformations of the variables, we note that a position of the reference team (runs scored and wickets lost in 50-*u* overs played) can also been quantitatively described by a single covariate. We denote this variable by *rpr* and is been interpreted as runs per unit of percentage resources lost. Mathematically, we define the *rpr=runs/crl*, where *crl* denotes the combined (*wickets* and *overs*) resources lost and can

be defined as the expected proportion of runs scored in 50-u overs, such that w wickets have already been lost, as compared to the expected total runs in the first innings of ODI cricket. Mathematically, the *crl* is written in simplified form as,

$$crl = 1 - F(w)\left\{\frac{\tan^{-1}\left(\frac{u-\mu}{\theta_0 F(w)}\right) - \tan^{-1}\left(\frac{-\mu}{\theta_0 F(w)}\right)}{\tan^{-1}\left(\frac{N-\mu}{\theta_0}\right) - \tan^{-1}\left(\frac{-\mu}{\theta_0}\right)}\right\} \qquad 7.5$$

where N is the number of pre-allotted overs to each team for a complete innings.

We note that the combined (*wickets* and *overs*) resources lost (*crl*), as defined in equation 7.5 is equal to 100% for $u=0$ or/and $w=10$. Further, the *crl* has a similar intuition to *wrl*, but it also accounts for the quantity of overs played. Figure 7.4 describes the relationship between the combined resources lost (*crl*) and wickets lost (w) for each given $u=50, 40,..,10, 5$ overs remaining.

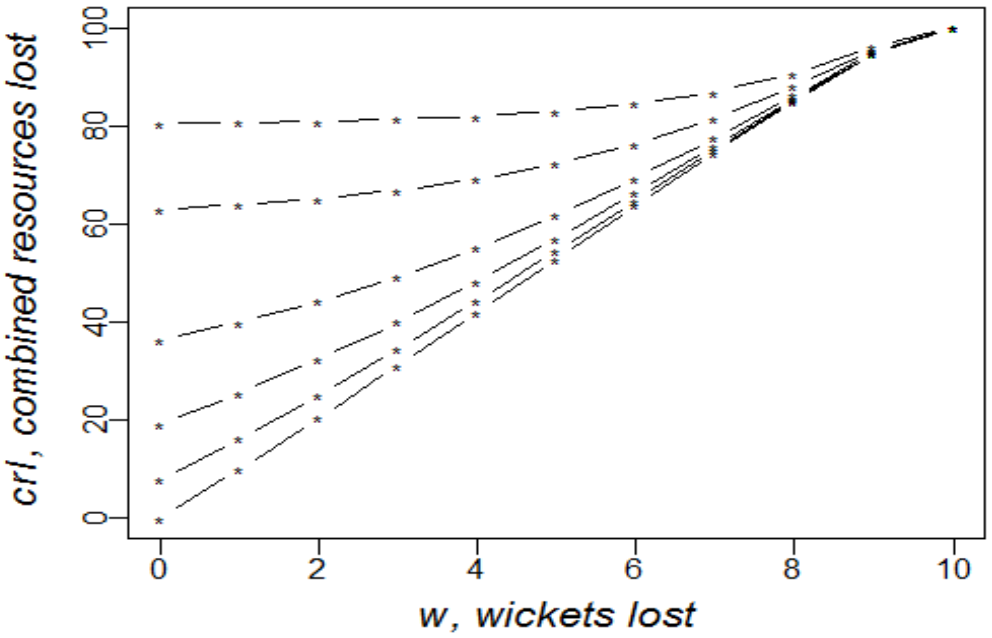

Figure 7.4 Plots of the relationships between the percentage of combined resources lost (*crl*) and wickets lost (*w*) for each *u=50 (bottom line),40,..10,5(top line)* overs remaining.

Similarly, for the second innings the covariate *rrpr*, runs required per unit of the total remaining resources, replaces *rpr* as in-play covariate. The covariate *rrpr* could be

defined as *runs-required-to- win/1-crl*. Note that *rpr* is a function of *runs*, *wickets*, *u*, and *N*, whilst the covariate *rrpr* is a function of *runs*, *wickets*, *u*, *N*, and *target*.

7.2.3 Organizing data for modelling

To facilitate the modelling procedure, we need to organize the ball-by-ball data into a series of data matrices (one for each ball of the game). First, we split the complete ball-by-ball data into two sub-data based on first and second innings. We next divide each sub-data into the series of data matrices by *k*, the balls remaining. For single innings (first or second) of ODI cricket, the data can be organized in 300 data matrices. Table 7.1 is an extract of the data matrix for k=150 (or u= 25 overs) balls remaining in the first innings of ODI.

Table 7.1 The extract of the data matrix for the first innings given *k=150* balls remaining,

ODI#	*win*	*toss*	*home*	*away*	*dn*	*fd*	*rd*	*w*	*wrl*	*rpo*	*rpr*
2075	0	0	0	1	0	100	14	4	3.319840	3.16	1.544753
2158	1	1	1	0	0	80.96	-1	5	4.401591	3.44	1.456323
2248	1	0	0	1	0	100	5	2	1.453567	5.28	3.520864
2322	1	1	1	0	1	3.05	14	2	1.453567	5.12	3.414171
2426	0	1	0	0	1	13.09	-69	5	4.401591	4.4	1.862738
2533	0	0	0	0	0	-29.75	0	2	1.453567	3.96	2.640648
2627	1	1	1	0	0	-72.58	-21	1	0.676281	4.48	3.521384
2707	1	1	0	0	1	-29.75	-3	1	0.676281	4.16	3.269856
2803	1	1	0	1	1	-9.28	-7	1	0.676281	5.16	4.05588
2884	1	1	1	0	1	16.71	-6	4	3.319840	2.72	1.329661
2960	1	1	1	0	1	100	57	2	1.453567	5.56	3.707576

In the data, we note that not all matrices are of the same number of rows (sample sizes). This is because not all innings (first or second) are necessarily ended by playing all the pre-allotted overs (N). Figure 7.5 shows the number of matches, $n(k)$, reaching the stage at k balls remaining. Note that the x-axes for plots in Figure 7.5 are presented reversed (so that the plot view shows a match progressing from left to right). Further, traditionally, cricket analysts and fans think in terms of number of 'overs' and so in the rest of plots we use the balls remaining (k) in a unit of overs remaining, $u=k/6$.

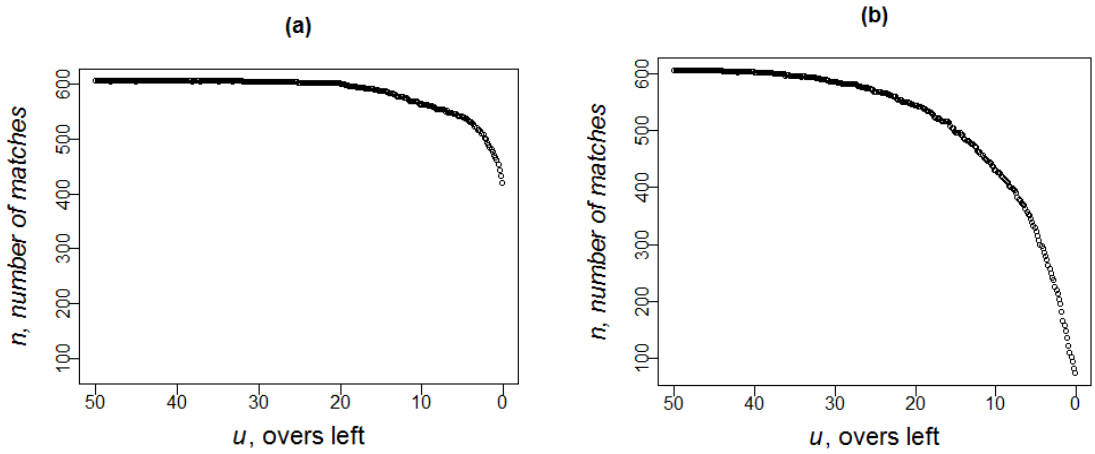

Figure 7.5 Plots of number of matches (*sample sizes*) against overs left for (a) first innings, and (b) second innings.

From Figure 7.5b it is noticeable that roughly the decay in sample sizes with respect to the progression of the second innings is more rapid compared to the first innings. This is because additionally the second innings might be ended with reaching team 2's score to the target set by team 1. Modelling the distribution of number of balls played in an innings might be an interesting problem to address in future.

7.3 Modelling procedure for the DLR models

We adopt a logistic regression model to estimate the probability of the batting team winning the match. However, the model is dynamic in the sense that the parameters are been allowed to vary as the match progresses. We develop two forecasting models, one for each innings of the ODI cricket. The reason we fit separate models, one for each two innings, is twofold: firstly, the batting team (reference team) play with different strategies in different innings. For example, Preston and Thomas (2002) argue that a batting team in the first innings, play with the aim to score as many runs as possible to maximise their win chances. However, a team batting in the second innings play with the aim to reach the target before, either all their wickets down or all the pre allotted overs (N) consumed. Secondly, some covariates, for example runs required per remaining overs (*rrpo*) to win, are only possible to measure for the second innings.

7.3.1 Modelling match outcome

For a given innings (first or second), let Y_k be a response variable for a given k balls remaining. We define the response variable, Y_k, as the match outcome and takes value 1 if the reference team (batting team) wins the match, otherwise takes value 0. For each ball of an innings, we fit a binary logistic model to estimate the probability of the batting team winning the match, p_k. Therefore, for a given $k=300,\ldots,1$ balls remaining, we let

$$logit(p_k) = \boldsymbol{X_k'\beta_k} + \varepsilon_k \qquad 7.6$$

where $\boldsymbol{X_k'\beta_k} = \sum_{m=0}^{M} X_{mk}\,\beta_{mk}$, where β_{mk} is the coefficient on the m^{th} covariate X_{mk}, and X_{0k}=1 for each given $k=300,\ldots,1$. In this way, for a given m, the coefficient β_{mk} is allowed to vary independently with respect to the stage of an innings. Lastly, ε_k is an error term.

Suppose for a given innings (first or second) and $k=300,..,1$ balls remaining, the y_{1k} ,…, y_{ik},…,$y_{n(k)}$ are the data on the response variable. Let $\boldsymbol{x_{1k}}$,,.., $\boldsymbol{x_{ik}}$,…, $\boldsymbol{x_{n(k)}}$ be the data on the corresponding vectors of covariates, where $\boldsymbol{x_{ik}}$= $[x_{0ik}\ldots\ x_{imk}\ldots\ x_{iM(k)}]^T$ and $M(k)$ is the number of covariates. Then from equation 6.3 the probability of the batting team winning the i^{th} match, at the stage when there are k balls remaining, is given by

$$p_{ik} = E(Y_{ik}|\boldsymbol{X_{ik}} = \boldsymbol{x_{ik}}) = \frac{exp(\boldsymbol{x_{ik}'\beta_k})}{1 - exp(\boldsymbol{x_{ik}'\beta_k})} \qquad 7.7$$

where $\boldsymbol{\beta_k} = \left[\beta_{0k} \cdots \beta_{M(k)}\right]^t$ is the vector of parameters. The likelihood for the logistic model for a given k balls remaining in an innings can then be written as,

$$L_k = \prod_{i=1}^{n(k)} (p_{ik})^{y_{ik}}(1 - p_{ik})^{1-y_{ik}} \qquad 7.8$$

where p_{ik} is the probability of the batting team winning the i^{th} (where $i=1,2,..,n(k)$) match of a given innings at the stage when there are k (or $u=k/6$ overs left) balls remaining. From equations 7.7 and 7.8 the log-likelihood is thus written as,

$$l_k = \sum_{i=1}^{n(k)} y_{ik}(\boldsymbol{x_{ik}'\beta_k}) - \sum_{i=1}^{n(k)} log\left(1 + exp(\boldsymbol{x_{ik}'\beta_k})\right) \qquad 7.9$$

In regards to our approach of modelling, we first identify the best subset of all possible pre-match covariates using the following model selection criteria: Akaike information criterion (AIC), Bayesian information criterion (BIC), Delete-d Cross-Validation (CV_d) with random subsamples of size $d= n(1-1/(\log n - 1))$ as the validating set of data, and the K-folds Cross-Validation. These model selection methods have been

briefly described in section 6.4.3. After identifying the best subset of pre-match covariates by fitting all possible models for first innings at *k=300* we next include all of the pre-match covariates in the set of all possible in-play covariates and identify the series of best logistic models, one for each given *k=299,..,1*. A similar procedure is adopted for the second innings and a series of 300 best logistic models are identified. We note that it is appealing to fit the separate logistic regression models for each ball of a given innings. This is because; a value of the response variable (match outcome) does not change with respect to the ball-by-ball given one innings data, whilst the in-play covariates are updated after each ball in the same data. Further, this approach to modelling helps to understand how the effects of covariates vary with respect to as an innings progresses. Further procedure of our proposed modelling approach is described in subsequent section.

7.3.2 Modelling the coefficients on the covariates: A recursive process

Once it is decided what variables are to be included in our final DLR forecasting models, we then estimate the relationship between the series of estimates and the stage of a given innings, *u* (or *k* alike). In this way, the estimates are allowed to evolve smoothly with respect to as innings progresses. To allow the effect of each covariate to depend on the stage of the innings, one could simply use a series of independent separate logistic regression models (one for each ball of an innings) and forecast the probabilities in a standard way. However, this would lead to the unstable variation in probability of match outcome in play.

Hence, instead of using separate models, we estimate the relationship between the estimated coefficients on the covariates (β_{mk}) and overs left (*u=k/6*) for each *m=0,..,M*. We denote this relationship by a function $\beta_m(u, \boldsymbol{\alpha_m})$, where $\boldsymbol{\alpha_m}$ is a vector of parameters to be estimated. In this way, the series of independent logistic models is reduced to a single dynamic logistic regression model in which the estimates themselves become a function of *u*, overs remaining (or *k* alike). The logistic model that uses the fitted functional values, $\beta_0^*(u, \alpha_0), \ldots, \beta_m^*(u, \alpha_m)$, as estimated parameters is then referred to as a dynamic logistic regression (DLR) model. We note that in addition to smooth evolving the estimated parameters, the number of total estimates required to forecast the in-play match outcome are also been reduced dramatically. For example, if *M+1* number of

parameters to be estimated for each *k=300,..,1* model, then we need $(M+1)\times 300$ estimates to forecast in-play match outcome for single innings. In fact, the large number of parameters that are being estimated is not really an issue. This is because, for each given *k* balls remaining model, the parameters are estimated from large sample sizes. This is especially evident during the early part of the both innings. However, any reduction in the number of parameters, at the cost of almost nothing, is an advantage in term of simplicity and usability. Finally, we believe that using a fitted functional values, $\beta_m^*(u, \boldsymbol{\alpha_m})$, to the estimated parameters are more precise and consistent as compared to the non-smoothed original estimates. This because the fitted values not only depended on the data matrix associated to *k* balls remaining, but also depends on the entire sets of data matrices for a given innings.

Suppose for a given innings, we fit a series of *K* independent logistic models, each model has *M* covariates. Further, for each *k=K,K-1,…,1* let the estimated coefficient on the m^{th} (*m=0,1,..,M*) covariate be denoted by $\hat{\beta}_{mk}$. Then for any given *m=0,1,..,M*, we assume normality, $\hat{\beta}_{mk} \sim N(\beta_{mk}, \sigma_{\hat{\beta}_{mk}})$, under the asymptotic property of estimates. Further, since the *K* logistic models are fitted independently, for any given *m* the series of estimates are independent (note that for any given *k* the parameters have not been estimated separately in a model). We then take expectations, so that

$$E(\hat{\beta}_{mk}) = E(\hat{\beta}_m|u) = \beta_m(u, \boldsymbol{\alpha_m}), \qquad 7.10$$

where $\boldsymbol{\alpha_m}$ is a vector of parameters and *u=k/6*, where *k* is the number of balls remaining. Further, also note that $E(\hat{\beta}_{mk}) = \beta_{mk}$. Suppose, in the series of *K* independent logistic models (each with *M* covariates), $\hat{\beta}_{mK}, \hat{\beta}_{m(K-1)}, \dots, \hat{\beta}_{m1}$ are the series of *K* estimates for the coefficients on the m^{th} covariate. Further, let $s.e(\hat{\beta}_{mK}),\ s.e(\hat{\beta}_{m(K-1)}), \dots, s.e(\hat{\beta}_{m1})$ be the corresponding series of estimated standard errors. Then we estimate the vector of parameter, $\boldsymbol{\alpha_m}$, by maximizing the following log-likelihood function

$$l_m = \sum_{k=1}^{K} log\left(\frac{1}{s.e(\hat{\beta}_{mk})\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{k=1}^{K}\left(\frac{\hat{\beta}_{mk} - \beta_m(u, \boldsymbol{\alpha_m})}{s.e(\hat{\beta}_{mk})}\right)^2 \qquad 7.11$$

or equivalently, we minimize the following weighted sum of squared errors.

$$WSSE_m = \sum_{k=1}^{K} \left(\frac{\hat{\beta}_{mk} - \beta_m(u, \boldsymbol{\alpha_m})}{s.e(\hat{\beta}_{mk})} \right)^2 \qquad 7.12$$

where $\beta_m(u, \boldsymbol{\alpha_m})$ is an appropriate function of *u* and will be used as an estimate of $(m+1)^{th}$ parameter in our DLR model to forecast the match outcome in-play.

As mentioned above that for any given *k*, the *M+1* parameters are not estimated separately. Therefore, it would be unappealing if we do not update the remaining *M* parameters, for each *k*, prior to the next fit of the estimates, $\hat{\beta}_{(m+1)k}$. Therefore, we again fit the series of independent logistic models for each *k=K,K-1…,1* but under the parameter constraint $\beta_{mk} = \beta_m^*(u, \boldsymbol{\alpha_m})$. Once the *M* estimates are updated for each *k=K,K-1…,1* logistic models, we fit the next function $\beta_{m+1}(u, \boldsymbol{\alpha_{m+1}})$ on the series of estimates associated to the next parameter. Afterwards, to update the remaining *M-1* parameters for each *k*, we again fit a series of *K* logistic models but this time under the two constraints, $\beta_{mk} = \beta_m^*(u, \boldsymbol{\alpha_m})$ and $\beta_{(m+1)k} = \beta_{m+1}^*(u, \boldsymbol{\alpha_{m+1}})$. We continue this process until all *M+1* parameters are modelled as a function of *u*, overs remaining.

We note that the identification of the functional form for $\beta_m(u, \alpha_m)$ is subjective by examining the scatter plot and testing the statistical significances of the series of estimates. For example, polynomials of varying degrees proved to be an appropriate function to be used to smooth the estimated coefficients. One could also use a spline fit to get a better goodness, however this would lead to more complex model. Further, the spline fit might also cause to the over-fitting issue. In some cases, especially if a covariate is not statistically significant in all the models of an innings, then a polynomial fit or spline fit might not be appropriate to smooth the series of estimates. For example, CV_{KF} based DLR model during the second innings, the covariate *rd* becomes statistically insignificant after about the first ten overs of the second innings. In such a case, we need a curve for estimated coefficients on *rd* such that the magnitudes are insignificant and have smoothed decays to zero with respect to as *u* approaches from 40 to zero. In our DLR model for the second innings, we use a Gamma distribution type function (see section 7.4.4) for estimated coefficients on *rd*.

7.4 The model fit results

7.4.1 A model for estimating pre-match win probability

Before fitting a series of independent logistic models, one after each ball of the first innings, we first identify the best subset of all possible pre-match covariates. The R package 'bestglm' of McLeod and Xu (2011), for each AIC , BIC, CV_d (Delete-d Cross Validation) and CV_{KF} (K-Fold Cross Validation), is used to select the best subset of the covariates. In regards to the CV_d method, the sub-sampling of size $d= n(1-1/(\log n - 1))$ is repeated 1000 times, and in the CV_{KF} method, we use a 10 folds cross validation technique, which is repeated 100 times. The choice of repeating samplings is default in *bestglm()* function of R package.

In our pre-match set of all possible covariates, we include the following variables: *home* (*=1* if the reference team is at home, *0* otherwise), *away* (*=1* if the opposition team is at home, *0* otherwise), *toss* (*=1* if the reference team wins the toss, *0* otherwise), *dn (day-night: a binary variable), fd(θ) (form-difference, where 0<θ <1)*, and *rd (ICC rating difference).* Further, the set of pre-match covariates also includes all possible two factor interactions between the categorical variables, for example *dn:toss (dnt).*

We have written a purpose R code using the package 'bestglm' for model selection and estimating θ in the covariate *fd(θ)*. Further, the argument 'method' in *bestglm()* is set to 'exhaustive' which ensures all possible models are fitted and the best model based on specified model selection criteria is chosen. Moreover, the θ is optimized with respect to the model selection criterion value as obtained from the 'bestglm' routine. For example, using the AIC model selection criteria the best subset of covariates: *dnt, home, rd, and fd(θ=0.23)*). Table 7.2 describes the best subsets of the covariates, along with model diagnostic measures.

Table 7.2 Best subsets of pre-match covariates for a logistic model as obtained by AIC, BIC, CV_d and CV_{KF} model selection methods.

Method	*Best pre-match covariates*	*Model diagnostic measures*
AIC	*home, dnt, rd,* and *fd(θ=0.23)*	AIC=694.4, logL=-343.2
BIC	*dnt, rd,* and *fd(θ=0.24),*	BIC=710.0, logL=-345.40
CV_d	*dnt, rd,* and *fd(θ=0.18)*	CV_{best}=0.2037, CV_{null}=0.2519, logL=-345.44
CV_{KF}	*rd*	$(CV, sdCV, logL)_{null}$= (0.2503, 0.0011, -419.3) $(CV, sdCV, logL)_{best}$= (0.2023, 0.0065, -352.9)

7.4.2 A series of models for estimating in-play win probabilities

Once the best pre-match covariates are identified, we then identify the 'best' subset of all possible covariates (pre-match and in-play) for each given *k* balls remaining of an innings. From Table 7.2, we include the following subset of pre-match covariates: *home, dnt, rd* and *fd(θ)* along with the in-play covariates of a given innings. We use $\theta=0.20$, an average of two optimized values as obtained by Information Criteria (IC) and Cross-Validation Delete-d methods.

First, we identify a series of best independent logistic models, one after each ball of the first innings. To do this we include the above list of covariates (*home, dnt, rd, fd, wrl, rpo,* and *rpr*) as an in-put of all possible covariates in the *bestglm()*. We run the function for each $k=299,298,..,1$ balls-remaining using each method (AIC, BIC, CV_d, and CV_{KF}). To analyse how the importance of covariates vary with respect to the progression of the first innings. We divided a complete innings into five stages, each of that is contained ten overs (sixty balls, and therefore sixty best models). Therefore, the first stage corresponds to $u=50$ *to 40* overs remaining, the second stage corresponds to $u=40$ *to 30*, and so on, with the final stage corresponding to $u=10$ *to 0*.

The results show that the significance of a covariate depends on the stage and model selection method. For example, based on AIC method, the effect of the covariate *home* after about the middle of the first innings becomes insignificant. Therefore, this covariate is appeared only in the first 160 best models for the first innings. Table 7.3 describes the number of time each covariate appeared in the best logistic models for each given stage of the first innings, using the AIC model selection method. It can be seen in Table 7.3 that the covariate *dnt* becomes insignificant during the last two overs of the first innings. Moreover, the covariates *fd* and *rd* are appeared in all the series of best logistic models for the first innings.

Similarly, in regards to in-play covariates, the *rpr* is appeared in the AIC based best models during the first and last stages of the first innings. Contrary to the covariate *rpr*, the covariates *wrl* and *rpo* are appeared in the best models related to the middle stages, and are not appeared in some best models related to the first and last stages of the first innings. We have experimented to identify the best logistic models, based on AIC without listing the covariate *rpr* in the all-possible covariates. It is been observed that the covariates *wrl* and *rpo* are then appeared in all best models for the first innings, except

for the first few balls of the first stage. Similarly, we also experimented by excluding the covariates *wrl* and *rpo*, and it is found that the covariate *rpr* is appeared in all best series of models for the first innings, except for the first few balls of the first innings.

Table 7.3 Number of time a covariate is appeared in the series of best logistic models for each given five stages of the first innings as obtained using the AIC method

	Covariate						
u	*home*	*dnt*	*fd*	*rd*	*wrl*	*rpo*	*rpr*
50-40	60	60	60	60	48	46	25
40-30	60	60	60	60	60	60	0
30-20	40	60	60	60	60	60	0
20-10	0	60	60	60	59	59	1
10-0	0	49	60	60	22	55	42
Total	160	289	300	300	249	280	68

Similar experimentation is performed using the BIC model selection method. It is noted that covariate *home* is not appeared in any of the BIC based series of best logistic models for the first innings. Further, the covariate *dnt* becomes insignificant after about the first seven overs. Moreover, the covariate form-difference (*fd*) is appeared in the best models at almost all given five stages. However, surprisingly for any given ten overs stage the *fd* is not appeared in all 60 best logistic models. For example, as shown in Table 7.4 the covariate *fd* is included in 23 out of total of 60 best models for the first stage (50-40 overs remaining) of the first innings. Similarly, during the last stage of ten overs the covariate *fd* is appeared in 43 out of total 60 best logistic models. Similar to the AIC based series of models, the BIC based series of best models also contained the covariate *rd* in all best logistic models for the first innings. In regards to the in-play covariates, the two covariates *wrl* and *rpo* are appeared during *u=47-20* overs remaining stages. On the other hand the covariate *rpr* is appeared in the best models for the first few overs and again during the last twenty overs. Hence, similar to the AIC based models, the covariate *rpr* and the set of two covariates *wrl* and *rpo* can be used interchangeably.

Table 7.4 Number of time a covariate is appeared in the series of best logistic models for each given five stages of the first innings as obtained using the BIC method

	Covariates					
u	*dnt*	*fd*	*rd*	*wrl*	*rpo*	*rpr*
50-40	38	23	60	40	37	18
40-30	0	20	60	60	60	0
30-20	0	11	60	57	57	3
20-10	0	36	60	4	4	56
10-0	0	43	60	0	17	60
Total	38	133	300	161	175	137

Next, we apply the Cross-Validation Delete-d method to get the series of best logistic models for the first innings. We note that the models that are obtained using the CV_d model selection technique are quite similar to the models that were been obtained using the BIC method. For example, Figure 7.5 that describes the number of time each covariates is appeared in the series of best logistic models for each given five stages, can be compared to Table 7.4.

Table 7.5 Number of time a covariate is appeared in the series of best logistic models for each given five stages of the first innings as obtained using the CV_d model selection method

	Covariates					
u	*dnt*	*fd*	*rd*	*wrl*	*rpo*	*rpr*
50-40	38	17	60	38	36	19
40-30	6	34	60	60	60	0
30-20	5	24	60	41	41	19
20-10	0	29	60	3	3	57
10-0	0	26	60	0	16	60
Total	49	130	300	142	156	155

Similarly, we also use the CV_{KF} method to get the best series of logistic models for the second innings. It is observed that with the exception of few of the models in the first stage, in approximately all of the best models for the first innings, only two covariates *rd* and *rpr* are appeared. Further, if *rpr* is deleted from the set of all possible covariates then using the CV_{KF} method, the best subset of covariates: *rd*, *wrl*, and *rpo* are observed for each best models related to the first innings. Table 7.6 describes the number

of time each given covariate is appeared in the best logistic models for each given five stages.

Table 7.6 Number of time a covariate is appeared in the series of best logistic models for each given five stages of the first innings as obtained using the CV_{KF} method

	Covariates			
u	*rd*	*wrl*	*rpo*	*rpr*
50-40	60	8	18	29
40-30	60	0	11	49
30-20	60	0	0	60
20-10	60	0	0	60
10-0	60	0	0	60
Total	300	8	29	258

Finally, in regards to what covariates to include in our final dynamic logistic regression (DLR) model for the first innings, it is dependent on which model selection method is been adopted. For example, based on AIC model selection method the covariates: *home, dnt, fd, rd, wrl, and rpo* should be included in our DLR model. However, based on BIC or CV_d methods, it is found that the covariate *home* should be excluded from the best set of covariates as obtained by the AIC method. Similarly, based on CV_{KF} with one-standard-deviation rule, the covariates *rd* and *rpr* should be included in our final DLR model for the first innings.

In regards to further discussion on the subject matter, we note that CV_{KF} method provides the most parsimonious series of models whom prediction errors does not exceed one standard deviation of the prediction errors of the models with smallest cross validation errors . This ensures the stability in the forecasting probabilities for each given ball of the first innings is highest in CV_{KF} based models. However, note that the CV_{KF} based DLR model does not ensure the stability in probability forecasts as vary with respect to the progression of the innings. Further, less forecasting accuracy is observed for the CV_{KF} method as compared to the models based on CV_d method. Moreover, we also note that the second best series of models, based on CV_{KF}, are with the covariates: *rd, wrl*, and *rpo*. Hence, we develop two DLR models, one with covariates *rd* and *rpr*, and another with covariates *rd, wrl* and *rpo* for the first innings of ODI cricket. In next

section, we analyse and compare the forecasting accuracies of these models along with the forecasting assessment of the models as obtained using CV_d method.

In regards to the second innings, we perform similar experimentations as we did for the first innings to decide what covariates to be included in our DLR model for the second innings. We include the same pre-match covariates in the list of all possible second innings in-play covariates. The frequencies of the appearance of each covariate in the best logistic model for the second innings are presented in the Table 7.7. Here again to decide which covariates should be included in our final DLR model for the second innings is depended upon which method to be adopted for the model selection. For example, based on AIC method we use the covariates *home*, *dnt*, *fd*, *rd*, *wrl*, and *rrpo* in our DLR model for the second innings. Similarly, based on K-flods Cross-Valication model selection method, we use the covariates *rd* and *rrpr* in our DLR model for the second innings. Similar to the first innings, we develop two forecasting models: one with covarites *rd* and *rrpr*, and another with covariates *rd*, *wrl*, and *rrpo* for the second innings. Before developing our DLR models for the first and second innings, first we assess the forecasting accuracies of the series of models with the covariates that would be included in our final DLR forecasting models.

Table 7.7 Number of times each covariates are appeared in the series of 300 best logistic models during the second innings, using the AIC, BIC, CV_d and CV_{KF} methods

	Covariate						
Method	*home*	*dnt*	*fd*	*rd*	*wrl*	*rrpo*	*rrpr*
AIC	41	39	194	170	241	249	62
BIC	0	0	36	142	109	121	185
CV_d	0	0	23	138	26	33	267
CV_{KF}	0	0	0	62	0	7	293

7.4.3 Assessing forecasting accuracies

In this section, we assess the forecasting accuracies and deterministic power of our proposed models. First, we examine the cross validation forecasting accuracies of CV_d and CV_{KF} based models for first innings. The CV_d based model includes *dnt*, *fd*, *rd*, *wrl*, and *rpo* as covariates, whereas, the CV_{KF} based model uses only *rd* and *rpr* as covariates. Further, we also examine the cross validation forecasting accuracies of the model with

covariates: *rd*, *wrl*, and *rpo*, which is the second best model based on CV_{KF} method. Second, we do similar analysis for the second innings models.

For each ball of the first innings, we examine the cross-validation forecasting errors of the three candidate models as mentioned above. We measure the relative forecasting errors (RFE), as determined by the ratio of leave-one-out-cross-validation (LOOCV) prediction error for each given candidate model (cv_c) to the prediction error of the null model (cv_0). A smaller the value of RFE the better a model would have forecasting accuracy. The RFE=1 show that the candidate model has similar forecasting accuracies to the null model. Figure 7.6 show the scatter plots of RFE against the overs-left for each of the three candidate models. It can be seen clearly in Figure 7.6 that during the last twenty overs of the first innings the average cross validation forecasting errors are approximately same for all the three models. Further, it is noticeable that during about first thirty overs of the first innings the cross validation forecasting errors of the models with covariates *rd* and *rpr* are the largest as compared to the other two candidate models.

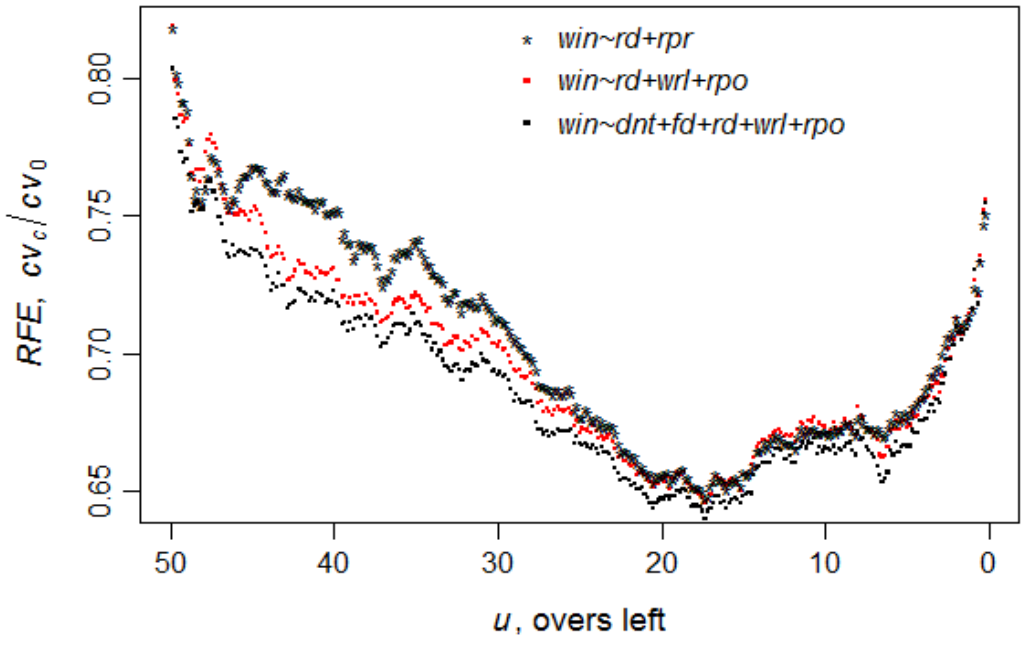

Figure 7.6 The plots of relative forecasting errors (RFE), as determined by the ratio of LOOCV prediction errors of the candidate model as compared to the null model, for the first innings.

In regards to the behaviour of the predictive power of the models as the first innings progresses, we note, for example in Figure 7.6, that the models' cross validation prediction errors are decreasing until about eighteen overs remaining but increasing during the last 18 overs. We believe such behaviour of the models' predictive power is the consequence of two reasons. First, the number of wickets lost becomes less informative during the final stages of an innings. For example, a team's position (winning or losing) can be predicted easily by the number of wickets lost at the stage when there are twenty overs remaining compared to the number of wickets lost on the final ball of the innings. In section 7.4.5 we analyse how the explanatory power or strength of each covariate varies with respect to the progression of the innings. Second, such behaviour of the forecasting accuracies might be because the matches that are more easily predicted will be ones in which the fifty overs allocated are not completed by the batting side. In such matches, the team batting second will tend to have a big advantage and so the prediction is easier to make. Since such matches are not included in the data matrices for the latter stages of the first innings, therefore the average cross validation forecasting errors are increasing during the last eighteen overs. Moreover, in regards to the performance of our model, it is observed that our proposed model perform best during the stage of 22-16 overs remaining as compared to the other stages of the first innings.

In regards to assess the forecasting accuracy for the models during the second innings, we measure the cross validation RFE, based on LOOCV prediction errors, for the each of two candidate models. As seen in the Figure 7.7 the forecasting errors for model with covariates: *rd* and *rrpr* compared to the model with the covariates: *rd*, *wrl*, and *rrpo* are approximately same. Hence, both models are equally suitable (in term of forecasting accuracies) to be used to estimate ball-by-ball probability forecasts during the second innings.

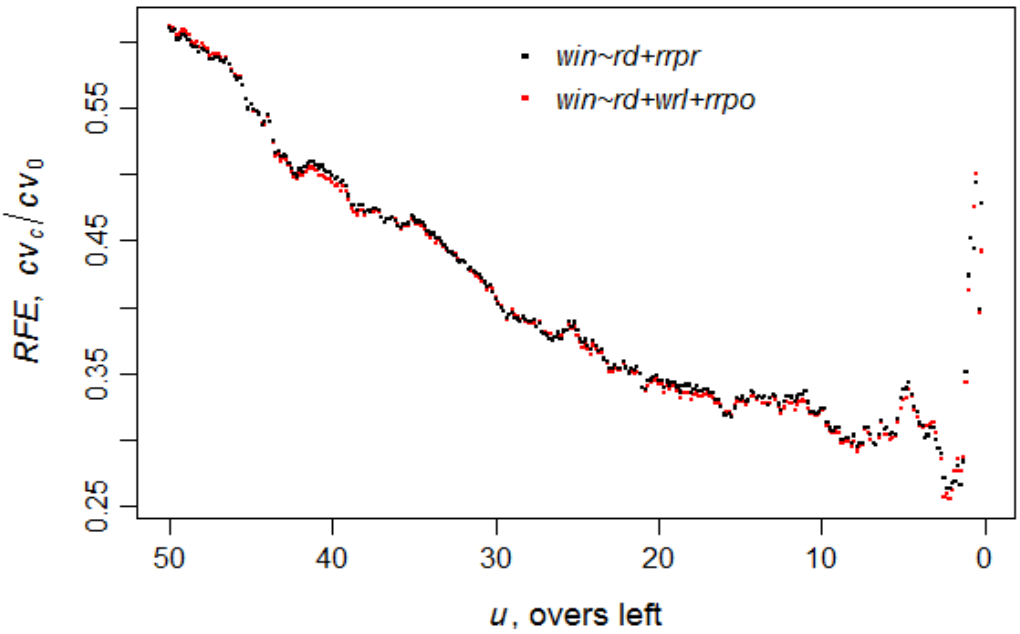

Figure 7.7 The plots of relative forecasting errors (RFE), as determined by the ratio of LOOCV prediction errors of the candidate model as compared to the null model, for the second innings.

The above figure show that the deterministic power of our model during the second innings is even greater than for the first innings, with a RFE of 0.65 at the start of the second innings. Further, it is noted that the cross validation forecasting error decreases throughout the second innings, except to rise during the last few balls. Our explanation for this is that matches which reach the final few balls are the ones in which the outcome is particularly uncertain.

7.4.4 Smoothing the estimated coefficients: A dynamic logistic regression (DLR) model

Once the best subset of covariates is finalized, we now start our recursive procedure to develop dynamic logistic regression (DLR) models, one for each two innings of ODI cricket. To be easily understood our proposed method; firstly, we fit the simplest dynamic logistic regression (DLR) models that are based on CV_{KF} model selection method. For the first innings, we fit a DLR model with covariates: *rd,* and *rpr*. We refer this model as *DLR*(*u*, *rd*, *rpr*). Similarly, we develop *DLR*(*u*, *rd*, *rrpr*) model for the

second innings that is based on CV_{KF} method. Secondly, we apply the same method to develop *DLR*(*u*, *rd*, *wrl*, *rpo*) and *DLR*(*u*, *rd*, *wrl*, *rrpo*) to estimate in-play probability forecasts during first and second innings respectively. Finally, we generalize our proposed approach of modelling for the more complex DLR models, for example, the AIC based DLR models.

We start to fit a series of 299 independent logistic models each with covariates *rd* and *rpr* on the series of data matrices related to the first innings. We than smooth the series estimated coefficients on *rpr* by fitting a weighted polynomial. We use the inverse of the squared standard errors of the estimates as the weights. Figure 7.8a shows the smoothed (fitted polynomial) and non-smoothed (original estimates) plots for the estimated coefficients, and Figure 7.8b shows the corresponding standard errors of the non-smoothed estimated coefficients in the series of 299 logistic models (each with covariates *rd* and *rpr*). It can be seen clearly that there is a strong deterministic evolution of the parameter value associated to the *rpr* covariate in the logistic model.

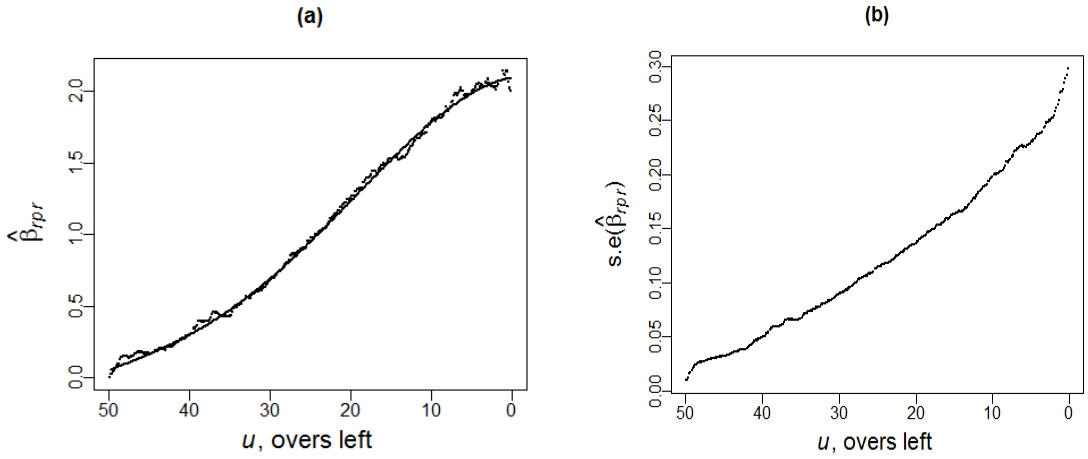

Figure 7.8 The observed estimated (a) coefficients (points) on covariate *rpr* and the fitted polynomial curve (solid lines), and (b) standard errors for the series of 299 first innings logistic regression models with covariates *rd* and *rpr*

As, in each any given k^{th} balls remaining logistic model, the coefficient on *rpr* is not estimated independently from the estimated coefficient on *rd* and intercept term. Therefore, after the weighted polynomial fit on the estimates related to *rpr*, we update the remaining estimates (related to the covariate *rd* and intercept term) by re-fitting the series of logistic models for first innings, but under the parameter constraint related to *rpr*. Note

that the constraint of the first parameter is to set the parametric value as a fitted polynomial value.

Interestingly, we note that smoothing (polynomial fit) the estimated coefficients on *rpr* has approximately no effect on the magnitude of the estimated coefficients related to *rd* and vice versa. Figure 7.9 shows the plot of estimated coefficients on *rd* before (black points) and after (red points) smoothing the estimated coefficients related to the covariate *rpr*. Next, we fit a weighted polynomial on the updated estimated coefficients on *rd*, to smooth the estimates. Afterwards, again we fit a series of 299 logistic models under the two-parameter (coefficients on *rpr* and *rd*) constraint to update the estimated intercept terms. Finally, we fit a weighted polynomial on the latest updated estimates of intercept terms. Purpose-written R code (R Core Team, 2012) utilising the standard *glm*() function has been developed to automate this process.

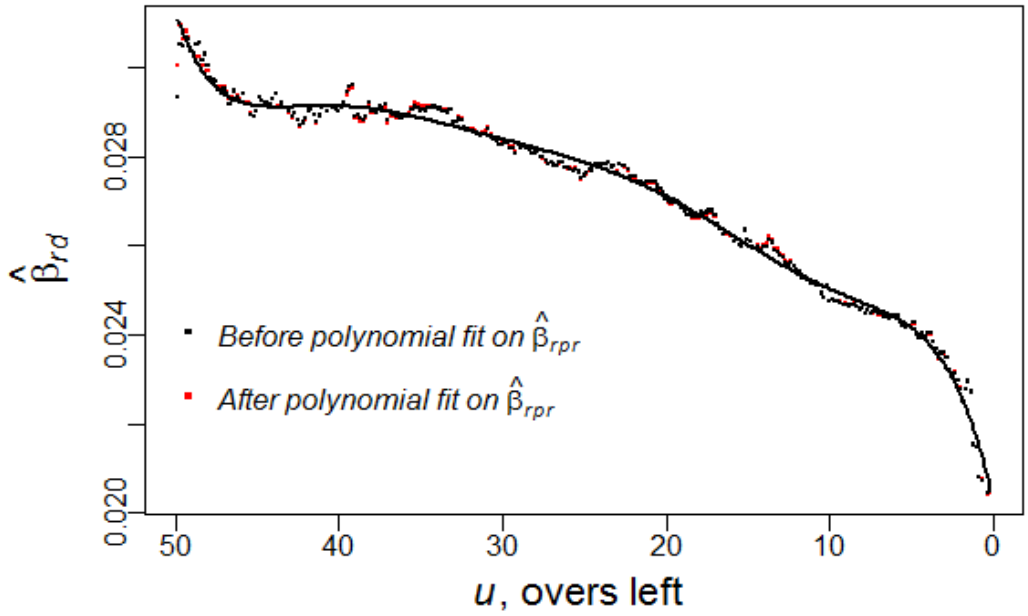

Figure 7.9 The estimated coefficients (points) for the series of 299 first innings logistic regression models with covariates *rd* and *rpr*, and the fitted polynomial curves (solid line).

As a result of the above recursive procedure, we obtain a single logistic model to forecast match outcome in play (ball-by-ball). In such a DLR model the estimates themselves is a function of the stage of the innings. Recall that the stage of the innings is

described by the overs remaining (u). Table 7.8 describes the summary of model fit for the dynamic logistic regression model for first innings. The number of parameters estimated and the diagnostic measures for each three polynomial fits on the series of non-smoothed estimates are given too. By using these fitted polynomials, we reduce the 299 models, with a total 3×299 = 897 parameters (2 covariates plus an intercept term), to a single dynamic logistic regression model with 18 parameters. Clearly, this DLR model is more attractive than the less parsimonious alternative. Despite the smooth evolving of the size of estimated coefficients with respect to the progression of the first innings, there are practical advantages too – our model is easily set up in a spreadsheet for example.

Table 7.8 Summary of the dynamic logistic regression (DLR) model to forecast match outcome in-play during the first innings.

A DLR model for first innings	$p_1(u) = \frac{e^{\beta_0(u)+\beta_1(u)*rd+\beta_2(u)*rpr}}{1+e^{\beta_{0(u)}+\beta_1(u)*rd+\beta_2(u)*rpr}}$		
Fitted function	*Estimates*	R^2	*WMSE*
$\beta_2^*(u)$, coefficient on ***rpr***	5	99.08%	0.00104
$\beta_1^*(u)$, coefficient on ***rd***	8	99.81%	0.00084
$\beta_0^*(u)$, ***Intercept***	5	99.99%	0.00018
Total	**18**		

In regards to the dynamic logistic regression model for the second innings, we follow the similar procedure as for the first innings. However, in the second innings the polynomial fits for smoothing the estimated coefficients on all the covariates are not suitable. This is because; the effect of the covariate *rd* is not found to be statistically significant during all stages of the second innings. In this case, we need a curve such that it has a smoothed tendency towards zero after the stage when *rd* becomes insignificant. For example, after examining the scatter plot of the series of estimated coefficients on *rd*, we fit the following positive non-decreasing function of u on the series of estimates related to *rd*,

$$\beta_{rd}(u) = c(\alpha_0 - u)^{(\alpha_1 - u)} e^{-\alpha_1(\alpha_0 - u)} \qquad 7.13$$

where $\alpha_0>50$, $\alpha_1>1$, $\alpha_2>0$ and $c>0$ are the location, shape, scale, and constant parameters respectively. The intuition of using equation 7.13 is demonstrated in Figure 7.10a, which plots the estimated coefficients on *rd* and the fitted curves using the

equation 7.13 (solid line), a quadratic curve (dashes) and a cubic curve (dots). As discussed in the section 7.4.2, that during the second innings the covariate *rd* is significant only during the first twenty overs. Further, as seen in Figure 7.10a, some instability in evolving the estimated coefficients, with respect to the progression of second innings, is also observed after about the twenty overs. Therefore, we fit equation 7.13 and some polynomials (in contrast) on the estimated coefficients on *rd* that are related to the first twenty overs. Further, we note that using equation 7.13 not only facilitates extrapolation of the estimated coefficients on *rd*, but also after about 20 overs the curve tends to zero as *u* approaches to zero. Similar to the first innings, we found approximately no effect on the estimated coefficients on *rrpr* by smoothing the estimated coefficients on *rd*. Figure 7.10b shows the plots for the series of estimated coefficients on *rrpr*, before (black points) and after (red points) smoothing the estimates related to the covariate *rd*. Further, a smoothed curve in Figure 7.10b shows a fitted polynomial on the estimated coefficients on the covariate *rrpr*.

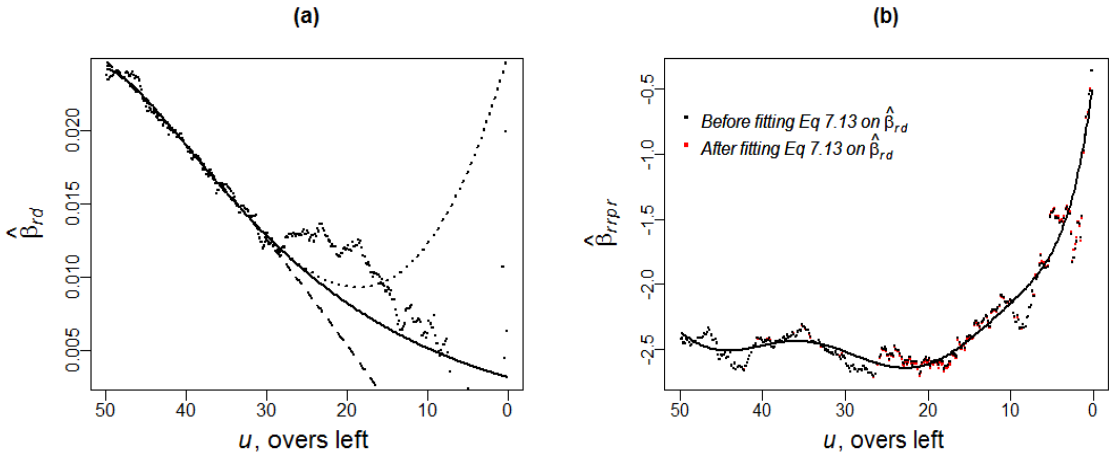

Figure 7.10 The original non-smoothed estimated coefficients (points) and the fitted curves (lines) in the series of independent models each with covariates *rd* and *rrpr*. Note that in (a) the curves are fitted using equation 7.13 (solid line), quadratic (dashed line) and cubic (dotted line).

After performing the recursive process for the second innings, we obtain the DLR model with covariates: *rd* and *rrpr*. Table 7.9 provides the summary of the DLR model fit for the second innings. Similar to the first innings models, the series of 300 independent logistic models are reduced to just a single DLR model with a total 20 parameters that are required to forecasts in-play probabilities during the second innings.

Table 7.9 Summary of the dynamic logistic regression (DLR) model to forecast in-play match outcome during the second innings.

A DLR model for second innings	$p_2(u) = \frac{e^{\beta_0(u)+\beta_1(u)*rd+\beta_2(u)*rrpr}}{1+e^{\beta_{0(u)}+\beta_1(u)*rd+\beta_2(u)*rrpr}}$		
Fitted function	*Estimates*	R^2	*WMSE*
$\beta_2^*(u)$, coefficient on ***rrpr***	8	94.50%	0.01281
$\beta_1^*(u)$, coefficient on ***rd***	4	98.54%	0.00000018
$\beta_0^*(u)$, ***Intercept***	8	99.84%	0.00460
Total	**20**		

As discussed above that smoothing the estimated coefficients on a covariate has very little (approximately zero) effect on the estimated coefficients on the remaining covariates in the series of logistic models. In contrast, relatively larger effects are observed on the intercept terms. We note that after a fit on each given estimated coefficient on covariates, the estimated intercepts tend to become more stable and well behaved against *u*, overs remaining. For example, Figure 7.11 shows scatter plots of the estimated intercepts in the series of independent logistic models before (black points) and after (red points) modelling the coefficients on covariates as function of *u*. Therefore, we recommend fitting a model on the estimated intercept terms once all the covariates have been modelled (smoothed) as a function of *u*.

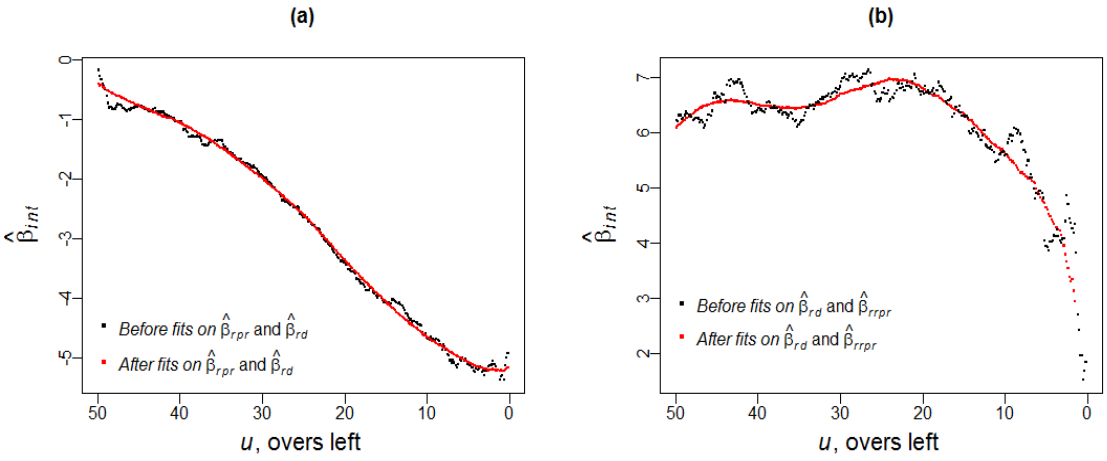

Figure 7.11 The observed estimated intercepts, for (a) first innings, and (b) second innings, in the series of logistic regression models, before (black points) and after (red points) smoothing the estimated coefficients.

We now develop another two DLR models, one with covariates *rd*, *wrl*, and *rpo* for first innings, and another with covariates *rd*, *wrl* and *rrpo* for the second innings of ODI cricket. Figure 7.12 and Figure 7.13 demonstrate graphically the development of these two DLR models, following a similar procedure as discussed above.

Hence, our approach for fitting a dynamic logistic regression model to forecast in-play match outcome for even more complex models. For example, based on AIC model selection criterion, a DLR model for the first innings can be developed with covariates: *home*, *dnt*, *fd*, *rd*, *wrl* and *rpo*. Similarly, for the second innings a DLR model with covariates *home*, *fd*, *rd*, *wrl*, and *rrpo* can be obtained using the AIC model selection criterion.

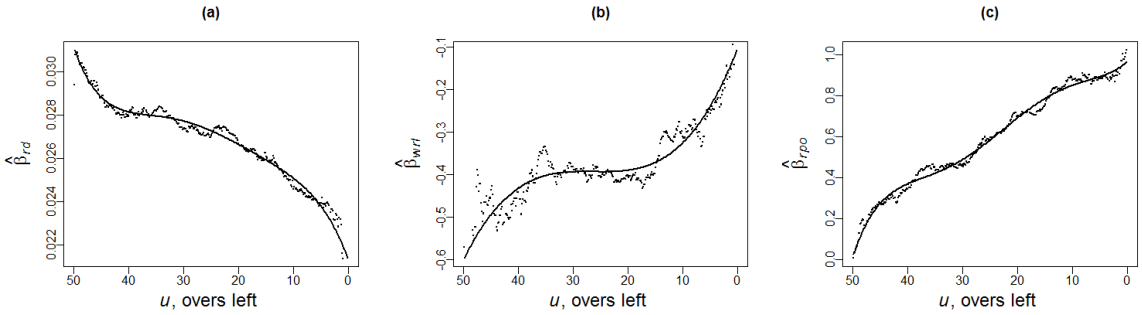

Figure 7.12 The observed estimated coefficients (points) for the series of 299 first innings logistic regression models with covariates *rd*, *wrl*, and *rpo*, and the fitted curves (solid lines).

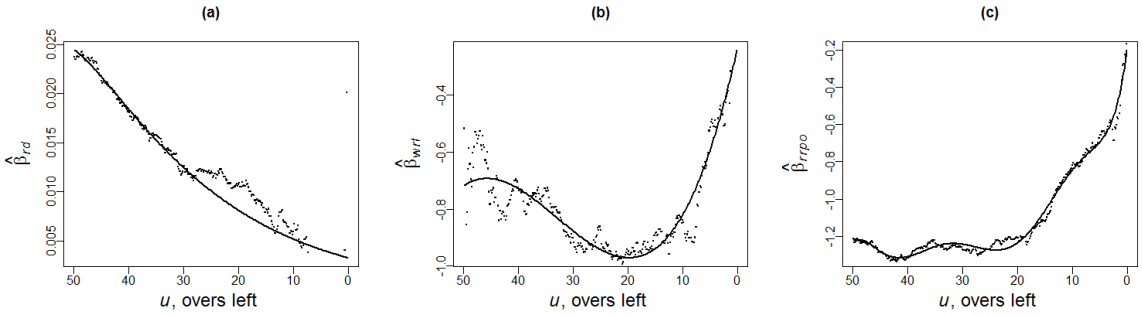

Figure 7.13 The estimated coefficients (pts) for the series of second innings logistic regression models with covariates *rd*, *wrl*, and *rrpo*, and the fitted curves.

7.4.5 Strength of association (Nagelkerke's R^2)

To justify further this DLR model to forecast match outcome in play, we compare the Nagelkerke's R^2 of the DLR model with a series of independent logistic models for

each given k (or u alike) of first and second innings. It can be seen clearly in Figure 7.14a that there is approximately no difference in Nagelkerke's R^2 if the functional fitted values are used, as compared to use the original non-smoothed series of estimates in the series of likelihood functions.

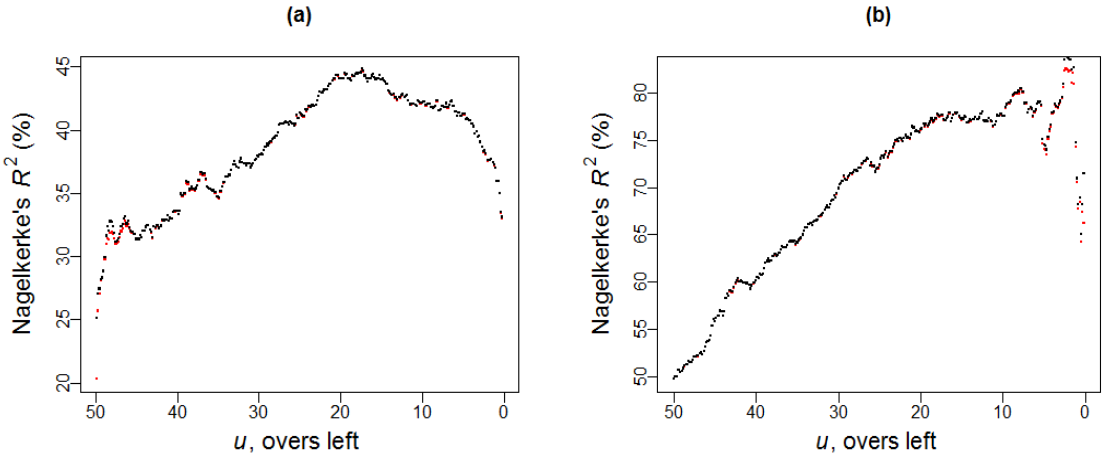

Figure 7.14 Plots of explanatory power, as determined by the Nagelkerke's R^2 using the estimates from the series of independent logistic models (black points) and from our DLR model (red points) for (a) first innings and (b) second innings.

To assess how the strength or explanatory power of each given covariate in our DLR model varies with respect to the progression of an innings, we use the difference between Negelkerke's R^2 of the models with and without the covariate, which we denote by ΔR^2. Figure 7.15 shows the plots of explanatory power of the covariates for the first and second innings models.

It is noticeable in Figure 7.15 that the strength of the pre-match covariate *rd* decreases with respect to as the game progresses. This is because, as the game progresses, the in-play covariates are updating and gathering more information on the state of the current match. For example, rating difference is very informative when making predictions in the early stages of the first innings. However, in the latter stages of the game the contribution of the rating difference made to predictions earlier on has, to some extent, been taken into account by the in-play covariate *rpr* in the first innings and by *rrpr* in the second innings. The decrease in the explanatory power of *rd* even continues in the second innings and by the mid-point of the innings the explanatory power of *rd* becomes approximately zero.

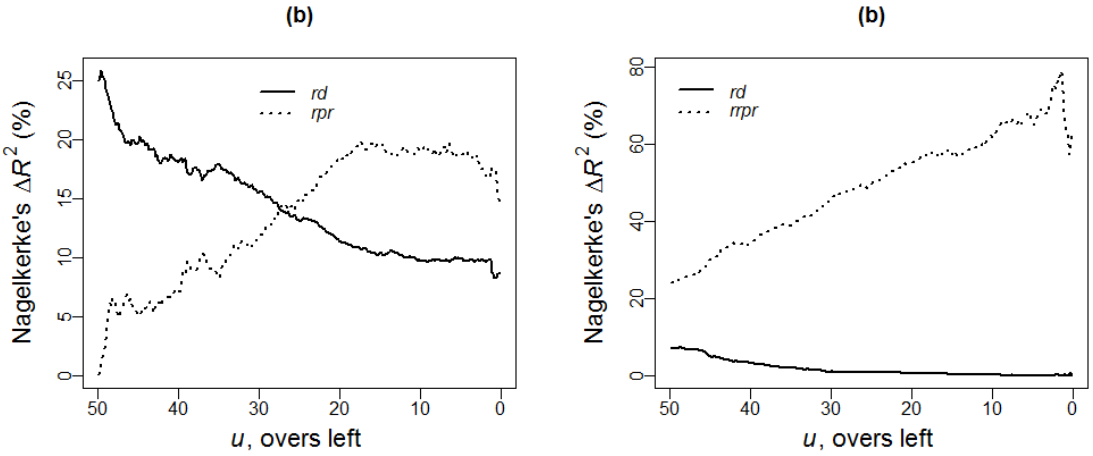

Figure 7.15 The plots of ΔR^2, the additional Nagelkerke's R^2 by covariates (a) *rd* and *rpr* in the first innings, and (b) *rd* and *rrpr* in the second innings, for the DLR forecasting models.

Similarly, in regards to the in-play covariates it can be seen in Figure 7.15a that during the first thirty overs (*u=50 to 20*), roughly the explanatory power of *rpr* rapidly increases as compared to the last twenty overs of the first innings. However, in the second innings, it can be seen in the same figure that the explanatory power of covariate *rrpr* is consistently increasing as the second innings progresses, except for a fall for last few balls.

7.5 Comparison with betting market

Perhaps the sternest test of a forecasting model in sport is to compare it to the betting market. Numerous studies have shown that betting markets are, for the large part efficient in that it is not possible to systematically beat the market, see, for example Sauer (1998). Here, we compare the probability forecasts generated using our dynamic logistic regression models to in-play odds from the betting market (Bet365), for two ODI matches: the second ODI match of the NatWest series between England and South Africa played at the Rose Bowl ground in Southampton on August 28th 2012, and the second ODI of the series between Pakistan and Australia in UAE on August 31st 2012. It is to be noted that the ball-by-ball data of these two matches could be considered as 'test data for the model accuracy' as this data were not included in model fits nor were they included in validating sets during the cross validation for model selection.

For our first example, South Africa won the toss and elected to bat first and set England a target of 287 to win. South Africa went on to win the match. Figure 7.16 shows the predicted probability of England winning the match during the first and second innings.

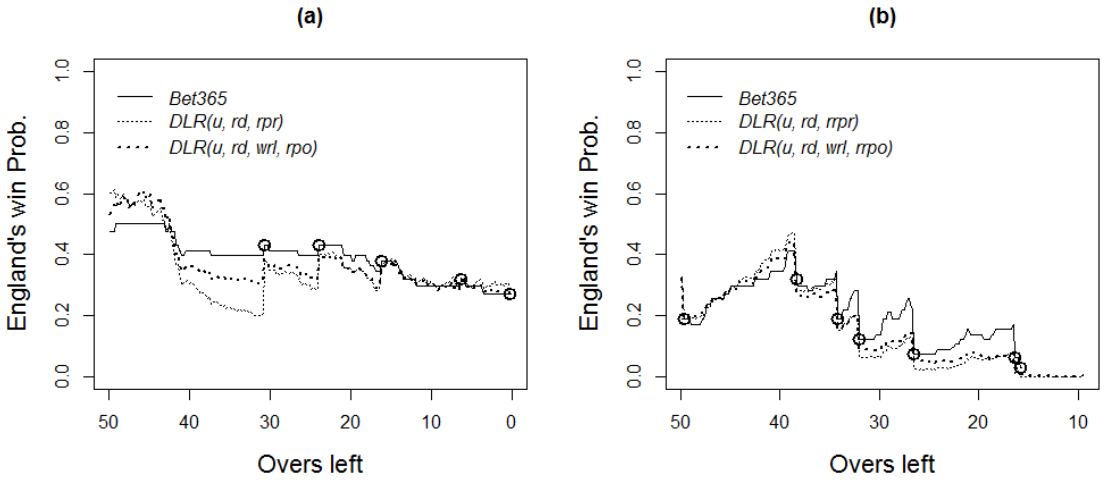

Figure 7.16 Forecast probability of England winning versus South Africa (a) first innings and (b) second innings. The solid line represents the implied bookmaker probabilities, whilst the dotted lines represent the forecast probabilities for our DLR models. The circles indicate the loss of a wicket.

In Figure 7.16, for the large part, our model forecasts follow a similar path to the bookmaker's forecasts, indicating our model is performing as one would hope. In fact, after around ten overs of the game, our model predictions are "more correct" than the bookmakers. Further, it is noticeable that approximately similar forecasts are obtained by the two DLR models (i.e. DLR(*u*, *rd*, *rpr*) and DLR(*u*, *rd*, *wrl*, *rpo*)) after about the first twenty overs of the first innings. However, during the first twenty overs of the game the probability estimates by DLR(*u*, *rd*, *rpr*) are more sensitive to *runs* and *wickets* compared to DLR(*u*, *rd*, *wrl*, *rpo*). Also, recall that during the first twenty overs, the series of models with covariates as used in the latter DLR model have better forecasting accuracies (see Figure 7.6). Therefore, we recommend a DLR model with covariates *rd*, *wrl*, and *rpo* should be used to forecast in-play match outcome probabilities during the first twenty overs of the first innings. Afterwards, both of our DLR models for the first innings are equally efficient. In regards to the second innings, we note that almost similar forecasts are obtained by both of DLR models and follow a similar path, but are "more

correct", than the probability forecasts of the betting market. Of course, this is only a sample of size one.

In regards to our second example (Pakistan versus Australia), Australia won the toss and decided to bat first, setting a target of 249 for Pakistan to win. Pakistan went on to win the game by seven wickets. Figure 7.17 shows the estimated ball-by-ball probabilities.

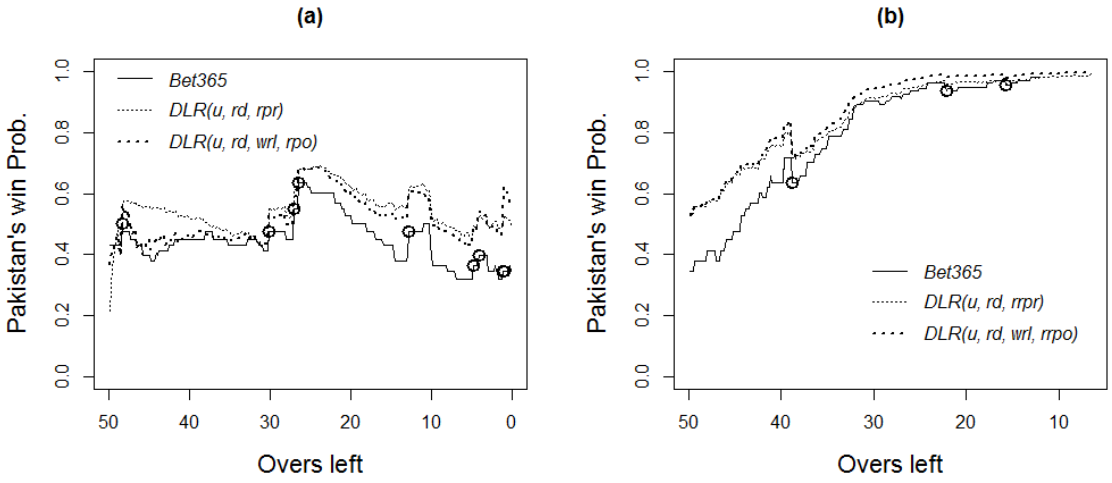

Figure 7.17 Forecast probability of Pakistan winning versus Australia for (a) the first innings and (b) the second innings. The solid line represents the forecast probabilities for implied bookmaker, whilst the dashed and the dotted lines represent probabilities as obtained by our DLR models.

As for the forecast probabilities for our second example, it is testament to the model that the two predictions follow similar trajectories. In fact, what is noticeable in this example is how the model suggests Pakistan's win probability is higher than that implied from the bookmaker's odds from around the midway point of the first innings. Similar to the first example, here again it is observed that during the first twenty overs the probabilities as obtained using the DLR model with covariates *rd* and *rpr* is less stable than the DLR model with covariates *rd*, *wrl*, and *rpo*.

Although we only look at two matches, we believe there is enough evidence to suggest our model is performing well, and that events occurring during a match (like a wicket, or a period of high scoring by the batting team) are appropriately incorporated into the model. It would be interesting in the future to experiment with our model as a tool for betting on a large number of games to form the basis of a more complete test of market efficiency.

7.6 The DLR models and future research

In regards to the shortcoming in our DLR models, we note an absentee: the pitch effect on runs scored. In cricket it is a commonly believed that some pitches are good for batting and some for bowling. One way to account for this effect is to use a covariate *rrp*, run-rate relative to pitch. The *rrp* could be defined as *rpo*/*average rpo in that pitch* or *country*. We initially experimented with such a variable, but found that there was not enough data to do this easily since not all teams have played on every ground in every country which would possibly result in a bias forecasts. For example, during 1992-2013 there are 167 uninterrupted matches played in New Zealand. Out of these 167 matches, only seven matches were played on neutral grounds. Similarly, for the same period in Sri Lanka, out of total 197 uninterrupted matches, only 47 matches were played on the neutral grounds (*statsguru*, www.espncricinfo.com). Future work might look at developing a dynamic logistic model in which the covariate that describes the scoring ability of a team also takes account of the pitch effect.

In addition to estimating probability forecasts, models like the one presented here, could also be used to help inform strategy during a game. Future research work in which our model is used to explain the optimum strategy of limited overs cricket would be an interesting project. Moreover, our DLR model can also be used as part of probability preservation methods for resetting targets in interrupted cricket matches, for example similar to the approach as proposed by Preston and Thomas (2002). However, in this regard the model should be tested for the standard properties that are presented in section 4.2. Finally, our modelling approach could also be used to develop a team and/or players ranking system for ODI cricket.

7.7 Summary

In this chapter, an in-play model for forecasting the winner of One-Day International cricket matches during any point of a game has been presented. The modelling approach that has been taken is one in which the estimated coefficients on covariates are allowed to evolve smoothly as the game progress. We refer this model a dynamic logistic regression (DLR) model.

In regards to the DLR model fit approach, first we fit a series of independent logistic models: one for each ball of the game. Four different methods for model selection are

applied. These methods are Akaike information criterion (AIC) (Sakamoto et al., 1986), Bayesian information criterion (BIC) of (Schwarz, 1978 and Akaike, 1977, 1978), CV_d (Delete-d Cross-Validation with random subsamples, Shao, 1997) and CV_{KF} (Hastie et al., 2009, K-fold Cross Validation). Once it is decided which covariates are to be included in our final DLR model for a given innings, than each of the estimated coefficients on the included covariates are modelled as function of *u*, overs left, by our proposed recursive procedure. It is noted that the CV_{KF} based DLR model is the most parsimonious with only two covariates in the model . On the other hand, AIC based DLR models are relatively more complex with five number of covariates in the model. The AIC based models are the *DLR*(*u, home, dnt, fd, rd, wrl, rpo*) for the first innings and the *DLR*(*u, home, fd, rd, wrl, rrpo*) for the second innings of ODI cricket.

For further justification of our approach of transforming the series of independent logistic models into a single DLR model for a given innings, we compare the Nagelkerke's R^2, for each ball of the game. The results show that an approximately similar explanatory power of the covariates is obtained using the series of independent logistic models as compared to our single DLR models.

Further, in regards to the predictive power of our DLR models, it is observed that the cross validation forecasting accuracies and the explanatory power of our DLR model depends on the stage of an innings. We use the leave-one-out-cross-validation method to assess the forecasting accuracies with respect to the progression of the innings. Broadly speaking, the forecasting accuracies and the explanatory power of covariates increase with respect to the progression of the game. We also examine how the strength or explanatory power of each covariate varies with respect to the progression of innings. To measure the strength of a covariate, we use the difference in Nagelkerke's R^2 of the models with and without that covariate. It has been demonstrated graphically that in our DLR models the predictive power of pre-match covariates decreases, whilst the strength of the in-play covariates increases, with respect to the progression of the game.

Lastly, we compare the ball-by-ball probability forecasts, as obtained by the proposed DLR models with those from the betting market, for two example matches. Roughly, similar forecasts are obtained from our DLR models and the betting market.

CHAPTER 8 SUMMARY AND FUTURE WORK

8.1 Summary of the thesis

The problems of interruptions to play and in-play forecasting are addressed in this thesis. Using statistical analysis, we investigate the issue of resetting targets in interrupted matches and propose an alternative, new method to this end. Further, we address the problem of in-play forecasting of match outcomes and propose a new approach of modelling in which the estimated parameters of the underlying models evolve smoothly with respect to the progression of the game.

We start with CHAPTER 1 to describe our aims and to set the objectives to achieve those aims. A brief historical background of limited overs international cricket is given. Further, some standard rules for cricket, the equipment, and the ground are described briefly. Finally, we end the chapter by describing the structure of this thesis and the research contribution in each chapter.

In CHAPTER 2, we give an overview of the literature about dealing with the problem of interruptions in limited overs cricket. The methods for resetting targets in interrupted limited overs cricket are broadly categorised in two parts: simple ad-hoc methods and more advanced methods that are proposed in the academic literature. For example, the run-rate based methods, the highest scoring overs (HSO) methods, equivalent point (EP) methods, and the PARAB method are investigated. It is argued that with the help of some real and hypothetical examples, that all of these methods are seriously flawed and can favour either of the two competing teams, depending upon situation. There are fundamentally two shortcomings in such methods. First, these methods do not account for the number of wickets the batting team has already lost at the time of interruption. Second, the stages of the interruptions are not been accounted for, and therefore all overs are considered to be of equal value in terms of run scoring potential. Similarly, a brief overview is given on more advanced methods for resetting targets in interrupted matches. For example, we describe briefly the Duckworth and Lewis (1998) method, Jayadevan (2002), and Carter and Guthrie (2004) method. It is highlighted that some of the major shortcomings of the simple ad hoc methods are resolved by these more advanced methods.

In CHAPTER 3, we present a procedure of estimating model parameters for Duckworth-Lewis Professional Edition for resetting targets. In regards to the D/L model estimation, Duckworth and Lewis (1998, 2004) did not disclose the parameter estimates or the estimation method. Further, to our knowledge there is no estimation method available in literature for the current D/L method. Moreover, we compare the runs scoring pattern of the Twenty-20 International and One-Day International cricket, and conclude that there is little evidence of a difference between the mean remaining runs for each given u, overs remaining and w, wickets lost. Further, it is argued that in addition to a statistical justification, it also appropriate from an ideological point of view that one single model is used for both formats of international cricket.

In CHAPTER 4 we investigate and compare the performance of some high profile methods including the existing Duckworth-Lewis method. The results have suggested that the current Duckworth-Lewis possesses more attractive properties than some other advanced methods that have been proposed in the literature. Further, we identified some standard desirable properties a method to reset targets should satisfy. In regards to this, first we investigate the runs scoring pattern of the VJD system of Jayadevan (2002) and showed that the second and fourth desirable properties, as presented in section 4.2, are not satisfied. Further, we demonstrated graphically that VJD system produces contradictory revised targets. Second, we examined Bhattacharaya's version of the Duckworth-Lewis method for T20I as proposed in Bhattacharya et al. (2011). This method also does not satisfy the second and fourth desirable properties and consequently results in unintuitive runs scoring pattern. Third, we investigated the runs scoring pattern of Stern's adjusted D/L method, as presented in Stern (2009). It is observed that in this method the rate of increase in the over-by-over runs value with respect to the progression of the innings is extremely rapid and therefore has some serious consequences. Finally, we overviewed the probability preservation method also known as Iso-Probability method of Carter and Guthrie (2004). The concept of probability preservation method was first proposed by Preston and Thomas (2002). Brief analysis of IP method shows that it compensates the teams for the interruption unreasonably different in similar situations. We note that further investigation of the IP method for appropriateness is beyond the scope of this thesis, partly because it has not been adopted by the ICC.

However, in this regards future research might be test this method for the properties we identified and presented in section 4.2.

In CHAPTER 5 we present a new modified Duckworth/Lewis method for resetting targets following interruptions in limited overs cricket. The fundamental notion of the Duckworth-Lewis method remains the same, that is to estimate each teams' available resources in a complete innings. However, we propose a new model for estimating these resources. We proposed a model to estimate the mean remaining runs as function of u, overs remaining, and w, wickets lost. It was shown that our newly proposed model provides a superior fit to data. Further, we have demonstrated graphically that our model reflects a more intuitive runs scoring properties than the existing model used in the Duckworth-Lewis method. In the course of our analysis, we also have shown that the ad hoc model adjustment for well above average runs situation considerably improves the forecasting accuracies of predicting first innings total runs. Finally, some issues related to the newly proposed method have been highlighted; indeed these same issues exist for the current D/L model.

In CHAPTER 6, we give an overview of in-play forecasting in cricket. It is highlighted that regarding the in-play forecasting, little work exists in the literature. A brief overview of generalized linear models is given. These models provide the basis for us to develop our in-play dynamic logistic models in chapter seven. Further, some model diagnostics and model identification criteria are discussed. Two types of model selection methods are been presented. First, the penalized log-likelihood function based methods, for example AIC and BIC are overviewed. Second, the cross validation forecasting accuracy based methods, for example, Delete-d Cross-Validation and K-folds Cross-Validation are overviewed.

In CHAPTER 7, a new approach to modelling is presented for forecasting in-play match outcome in ODI cricket. The method of modelling that is adopted is the one in which the estimated coefficients on covariates in an ordinary logistic model are allowed to evolve smoothly as the game is in progression. One single such model could be developed for a complete innings (first or second). We refer to this model as a dynamic logistic regression (DLR) model. Two types of covariates, pre-match and in-play, are used in these DLR models. Different model selection procedures provide different set of best covariates that could be used in our final DLR model. For example, based on AIC, a

DLR model with covariates: *home*, *dnt*, *fd*, *rd*, *wrl* and *rpo*, is obtained for the first innings in-play forecasts. However, using K-folds Cross-Validation method a DLR model with covariates: *rd*, and *rpr* is obtained. Similarly, to improve the leave-one-out-cross-validation (LOOCV) forecasting accuracies in the latter DLR model, during the first twenty overs of the first inning, the covariate *rpr* should be replaced with two covariates: *wrl* and *rpo*. Further, in our study it is to be noted that the DLR models that are based on BIC or CV_d methods are similar. The meanings and descriptions of the covariates, that have been experimented with in our modelling, are provided in Appendix II. In regards to the explanatory power of the covariates, it is observed that the explanatory power of pre-match covariates is decreasing and the explanatory power of the in-play covariates is increasing, with respect to game progression. Further, the overall forecasting power of the DRL models is increasing. Finally, we compare the ball-by-ball probability forecasts for match outcome as obtained by our DLR models with that of the betting markets. Our forecasts are similar to those of the betting market, a testament to the accuracy of our model.

8.2 Future work

In this thesis, the statistical analysis that we have performed to tackle the two areas of research, resetting targets in interrupted limited overs matches, and estimating in-play probability forecasts for match outcome, suggests several avenues for future research work.

In regards to our work on the resetting targets, first, our four desirableproperties could be used to assess a future method for resetting targets in interrupted matches. Further, our proposed modified D/L method could be used to estimate the runs margin of victory for a team batting second. Traditionally, in cricket a margin of victory for a team batting second is measured in term of wickets, whilst the margin of victory for a team batting first is determined in term of runs. Hence, it is not possible to compare the two margins of victories as both of these measures are in different units. Future research might be of interest where our proposed modified D/L model is used to estimate runs margin of victories for the teams batting second and compare the results with the margin of victories for the teams batting first. As such, the modified D/L model could also been used to rank teams and players performance. One of the shortcomings of the traditional

measures of performance of the players is that a batsman's performance cannot be compared with that of a bowler or an all-rounder. Our proposed modified D/L model facilitates comparison of the performances of batsmen, bowlers, and all-rounders. For example, our modified D/L model could be used to examine how a player in a match utilizes the resources. Future research might be of interest in this regard.

Our proposed modified D/L model can further be improved by taking account of the power-play overs and the order number of the two batsman. The power-play overs are those overs in which field positions are restricted. For detail, see the ICC official web site. For instance, it is a common opinion that the runs scoring potential in power-play overs are higher than the non-power-play overs. Similarly, a team would be in a better position if the minimum order of the two batsmen, in the current wicket partnership, is lower. This would especially be of greater importance during the final stages of an innings. For example, a team would be in stronger position if the current two batsmen were playing at numbers 1 and 11 rather than a team whose batsmen are playing at numbers 10 and 11. It can be observed that both teams have lost nine wickets ($w=9$), however, the former team has more wicket resources as a well-set quality batsman is at the crease. Future research might be of interest to modify our proposed model such that it accounts the power-play overs as well as the orders of the current batsmen.

In regards to our work for in-play, ball-by-ball, forecasting match outcome, it is noted that in future such models could be used for revising targets in interrupted matches. However, further research is required to test this model for the properties presented in section 4.2. Lastly, our model could be used to assess different strategies during an innings. In this regard, future work is required where the effect of each covariate on the probability of match outcome is examined for each ball of the game.

Appendix I

The Step-by-step description of application of the VJD system taken as it is from the 'Appendix 1' of the Jayadevan (2002)

"The whole problem of fixing target scores is broadly categorized under three cases:
Case-A: The interruption is after team-1 has completed its innings and before team-2 begins its innings.
Case-B: The interruption comes after team-2 has batted through some overs in its innings.
Case-C: The interruption is during the batting of team-1itself.
Any problem related to fixing target scores can be included in one of the three categories or can be treated as a combination of two or all of these cases.

Step-by-step procedure for case-A

1. Find out the percentage of overs team-2 gets.
2. Find out the corresponding target score percentage from the target table.
3. Multiply the score made by team-1 with the value obtained in #2.

Illustrative example-1: Team-1 scores 264 runs in 50 overs. Before team-2 starts batting, an interruption occurs and the match is reduced to a '42-over' one. Target score
for team-2 is found as follows.

Solution

- 'Percentage overs' to be played by team-2 = 42/50 ×100 = 84.
- From target table, corresponding to 84% of overs, target percentage = 90.3.
- Hence, the target score = 90.3 × 264 = **239 runs**.

Step-by-step procedure for case-B

1. Find out the percentage of overs played up to the interruption.
2. Find out the normal percentage of runs corresponding to #1 and the wickets fallen.
3. Find out the PAR score (say PAR-1) as, normal score percentage multiplied by the score of team-1.
4. Find out the percentage remaining overs with respect to the *total overs remaining*.
5. Find out the corresponding target percentage.
6. Multiply the target percentage of #5 with 'the total score of team-1 minus PAR-1' to get the target score in the remaining overs.
7. Add PAR-1 with the target obtained in #6 to get the net target.

Illustrative example-2: LOI# 1442: Australia vs West Indies (WI). Australia 252 in 50 overs; WI, after 29 overs, 138/1. Ten overs are lost. What is the target for WI in 40 overs.

Solution

- Percentage of overs played by WI at the time of interruption = 58.
- Corresponding normal score = 48.3%.
- PAR-1 = 48.3 × 252 = 121.7 → (1).
- Percentage of the remaining overs wrt the total remaining overs = 11/ 21 × 100 = 52.4.
- Corresponding target percentage = 65.6.
- Target score for the remaining overs = 0.656 × (252–121.7) = 85.5 → (2).
- Net target in 40 overs (1) + (2) = 121.7 + 85.5 = 207.2 = **208 runs**.

Step-by-step procedure case-C

1. Find out the percentage of overs played up to the interruption.
2. Find out the normal percentage of runs corresponding to #1 and the wickets fallen.
3. Find out the percentage of *remaining overs* with respect to the *total overs, which was originally remaining.*
4. Find out the corresponding target percentage.
5. Multiply the target percentage obtained in #4 with the
6. Remaining score percentage (i.e. 100 – normal score calculated in #2).
7. Add the percentages obtained in #2 and #5 to get the *effective normal score (ENS) of team-1* in total percentage of overs played.
8. Find out the target percentage for the total percentage of overs played.
9. Target percentage in #7 divided by the ENS percentage in #6 will give the multiplication factor (MF). It is proposed to keep the lower limit of this MF as 1 for game-related reasons.
10. Multiply the score made by team-1 with MF to get the target of team-2.

Illustrative example-4: (Single interruption) LOI #1485
Sri Lanka vs Australia. Australia were 110/3 in 23.1 overs when the interruption took place. Seven overs were lost. Australia make 206 in 43 overs. What is the target for Sri Lanka in 43 overs.

Solution

- Percentage of overs played at the interruption = 46.2.
- Normal percentage with 3 wickets lost = 42.8.
- Remaining over percentage = 19.84/26.84 × 100 = 73.9.
- Corresponding target percentage = 83.2.
- ENS of Australia in 43 overs = 42.8 + (100–42.8) ×83.2% = 90.39%.
- Target score percentage for 43 overs (86%) = 91.6.
- MF = 91.6/90.39 = 1.0134.
- Target for Sri Lanka in 43 overs = 1.0134 × 206 = 208.76 = **209 runs.**"

Appendix II

A table describes the covariates, experimented for DLR models in CHAPTER 7

Covariate	*Meaning*	*Description*
home	*home-vanue*	A binary variable taking the value 1 if the reference team is playing on home venue ground, otherwise 0.
away	*Away-vanue*	A binary variable taking the value 1 if the reference team is playing on away venue ground, otherwise 0.
dn	*day-night*	A binary variable taking the value 1 if the match is a day-night game, otherwise 0.
fd	*form-difference*	A continuous variable, ranging from -100% to 100% and describes the percentage difference between the form of the reference team and the opposition team. The 'form' of a team is determined as, $form = \sum_{t=1}^{5} w(t,\theta) y_t / \sum_{t=1}^{5} w(t,\theta)$, where $w(t,\theta) = \theta(1-\theta)^{t-1},\ t = 1,\ldots,5$ and $0 < \theta < 1$. This covariate describes a performance difference, based on last five matches, between the two competing teams.
rd	*ratings-difference*	The difference in most recently available ICC official ratings of the reference and opposition teams. This covariate describes quantitatively the performance difference (as at the time of play), based on matches played in last three years, between the two competing teams.
wrl	*wicket-resources-lost*	A continuous positive increasing function, ranging from 0 to 10, of two variables *w*, wickets lost, and *u*, overs remaining. It is defined as the proportion of runs lost in remaining innings for the loss of *w* wickets, compare to the expected remaining runs with no wicket lost, given *u* overs are remaining. Mathematically, $wrl = 10 \times \{1 - R(w\|u)\}$, where $R(w\|u) = F(w)\left\{\frac{\tan^{-1}\left(\frac{u-\mu}{\theta_0 F(w)}\right) - \tan^{-1}\left(\frac{-\mu}{\theta_0 F(w)}\right)}{\tan^{-1}\left(\frac{u-\mu}{\theta_0}\right) - \tan^{-1}\left(\frac{-\mu}{\theta_0}\right)}\right\}$
rpo	*runs-per-over*	Runs scored by the reference team, divided by number of overs played.
rrpo	*required-runs-per-over*	Runs required to win for reference team, divided by total number of overs remaining.
rpr	*runs-per-unit-resources-consumed*	Runs scored by the reference team divided by the percentage of combined (*wickets* and *overs*) resources lost, *crl*. Where $crl = 100 \times \{1 - R(w\|u,N)\}$, where $R(w\|u,N) = F(w)\left\{\frac{\tan^{-1}\left(\frac{u-\mu}{\theta_0 F(w)}\right) - \tan^{-1}\left(\frac{-\mu}{\theta_0 F(w)}\right)}{\tan^{-1}\left(\frac{N-\mu}{\theta_0}\right) - \tan^{-1}\left(\frac{-\mu}{\theta_0}\right)}\right\}$
rrpr	*required-runs-per-unit-resources-remaining*	Runs required to win for the reference team, divided by the percentage of combined (*wickets* and *overs*) resources remeaining, *1-crl*.

References

Abdi, H. (2007). *Bonferroni and Šidák corrections for multiple comparisons* In NJ Salkind (ed.). Thousand Oaks, CA: Sage.

Agresti, A. (2002). *Categorical Data Analysis* (Second ed.). Hoboken, NJ: A John Wiley & Sons, Inc.

Agresti, A. (2007). *An Introduction to the Categorical Data Analysis* (Second ed.). Hoboken, NJ: John Wile & Sons, Inc.

Akaike, H. (1977). *On entropy maximization principle*. Paper presented at the Applications of Statistics (Proceedings of Symposium, Wright State University, Dayton, Ohio, 1976), North-Holland, Amsterdam.

Akaike, H. (1978). A new look at the Bayes procudure. *Biometrika, 65*, 53-59.

Akhtar, S., & Scarf, P. (2012). Forecasting test cricket match outcomes in play. *International Journal of Forecasting, 28*(3), 632-643. doi: http://dx.doi.org/10.1016/j.ijforecast.2011.08.005

Allsopp, P. E., & Clarke, S. R. (2004). Rating Teams and Analysing Outcomes in One-Day and Test Cricket. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 167*(4), 657-667.

Bailey, M., & Clarke, S. R. (2006). *Predicting the match outcome in one day international cricket matches, while the game is in progress*. Paper presented at the The 8th Australasian Conference on Mathematics and Computers in Sport, 3-5 July, Queensland, Australia. Research Article retrieved from

Bet365. Retrieved August 2012, from http://www.bet365.com

Bhattacharya, R., Gill, P. S., & Swartz, T. B. (2011). Duckworth–Lewis and twenty20 cricket. *Journal of Operational Research Society, 62*(11), 1951-1957. doi: doi:10.1057/jors.2010.175

Brooks, R. D., Faff, R. W., & Sokulsky, D. (2002). An Ordered Response Model of Test Cricket Performance. *Applied Economics, 34*(18), 2353-2365.

Carter, M., & Guthrie, G. (2004). Cricket Interruptus: Fairness and Incentive in Limited Overs Cricket Matches. *The Journal of the Operational Research Society, 55*(8), 822-829.

Carter, M., & Guthrie, G. (2005). Reply to the Comments of Duckworth and Lewis. *The Journal of the Operational Research Society, 56*(11), 1337-1341.

Claeskens, G., & Hjort, N. L. (2008). *Model Selection and Model Averaging* New York: Cambridge University Press.

Clarke, S. R. (1988). Dynamic Programming in One-Day Cricket-Optimal Scoring Rates. *The Journal of the Operational Research Society, 39*(4), 331-337.

Clarke, S. R., & Allsopp, P. (2001). Fair Measures of Performance: The World Cup of Cricket. *The Journal of the Operational Research Society, 52*(4), 471-479.

Congdon, P. (2005). *Bayesian Models for Categorical Data*. West Sussex: John Wiley & Sons, Ltd.

Cox, D. R., & Snell, E. J. (1989). *The analysis of binary data* (Second ed.). London: Chapman and Hall.

CricketArchive. (2012). Rain Rule Methods Retrieved March 2012, 2012, from http://cricketarchive.com/Miscellaneous/Rain_Rule_Methods.html

de Silva, B., Pond, G., & Swartz, T. (2001). Estimation Of the Magnitude of Victory in One-Day Cricket. [Article]. *Australian & New Zealand Journal of Statistics, 43*(3), 259.

do Rego, W. (1995). Wayne's Systm. *Wisdon Cricket Monthly*.

Dobson, A. J. (2001). *An Introduction to Generalized Linear Models* (Second ed.). London: Chapman and Hall/CRC.

Duckworth, F. C., & Lewis, A. J. (1998). A Fair Method for Resetting the Target in Interrupted One-Day Cricket Matches. *The Journal of the Operational Research Society, 49*(3), 220-227.

Duckworth, F. C., & Lewis, A. J. (2004). A Successful Operational Research Intervention in One-Day Cricket. *The Journal of the Operational Research Society, 55*(7), 749-759.

Duckworth, F. C., & Lewis, A. J. (2005). Comment on Carter M and Guthrie G (2004). Cricket Interruptus: Fairness and Incentive in Limited Overs Cricket Matches. *The Journal of the Operational Research Society, 56*(11), 1333-1337.

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science, 1*(3), 297-310.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Data Mining, Inference and Prediction* (Seccond ed.). New York: Springer-Verlag.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized Additive Models*: Chapman & Hall/CRC.

Hosmer, D., & Lemeshaw, S. (2000). *Applied Logistic Regression* (Second ed.). New York: John Wiley & Sons, Inc.

Hosmer, D. W., & Lemeshow, S. (1980). Goodness-of-fit tests for the multiple logistic regression model. *Comm. Statist. Theory Meth., A A9 (10)*, 1043–1069.

Jayadevan, V. (2002). A new method for the computation of target scores in interrupted, limited-over cricket matches. *Current Science 83*(5).

Jayadevan, V. (2004). An improved system for the computation of target scores in interrupted limited over cricket matches adding variations in scoring range as another parameter. *Current Science, 86*(4).

Lewis, A. J. (2005). Towards Fairer Measures of Player Performance in One-Day Cricket. *The Journal of the Operational Research Society, 56*(7), 804-815.

Lewis, A. J. (2008). Extending the range of player-performance measures

in one-day cricket. *The Journal of the Operational Research Society*

59, 729–742.

McCulloch, C. E., & Searle, S. R. (2001). *Gneralized, Linear and Mixed Models*. New York: John Wiley & Sons, Inc.

McHale, I. G., & Asif, M. (2013). A modified Duckworth–Lewis method for adjusting targets in interrupted limited overs cricket. *European Journal of Operational Research, 225*(2), 353-362. doi: http://dx.doi.org/10.1016/j.ejor.2012.09.036

McLeod, A. I., & Xu, C. (2011). Best Subset GLM. R package version 0.33. Retrieved from http://CRAN.R-project.org/package=bestglm

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

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General), 135*(3), 370-384.

Norman, J. M., & Clarke, S. S. (2007). Dynamic Programming in Cricket: Optimizing Batting Order for a Sticky Wicket. *The Journal of the Operational Research Society, 58*(12), 1678-1682.

O'Riley, B., & Ovens, M. (2006). IMPRESS YOUR FRIENDS AND PREDICT THE FINAL SCORE: An analysis of the psychic ability of four target resetting methods used in One-Day International Cricket. *Journal of Sports Science and Medicine 5*, 488-494

Preston, I., & Thomas, J. (2000). Batting Strategy in Limited Overs Cricket. *Journal of the Royal Statistical Society. Series D (The Statistician), 49*(1), 95-106.

Preston, I., & Thomas, J. (2002). Rain Rules for Limited Overs Cricket and Probabilities of Victory. *Journal of the Royal Statistical Society. Series D (The Statistician), 51*(2), 189-202.

R Development Core Team. (2012). *R: A language and environment for statistical computing.* Vienna, Austria: The R Foundation for Statistical Computing.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike Information Criterion Statistics*. Tokyo: KTK Publishin House.

Sauer, R. D. (1998). The economic of wagering markets. *Journal of Economic Literature 36*, 2021-2064.

Scarf, P., & Akhtar, S. (2010). An analysis of strategy in the first three innings in test cricket: declaration and the follow-on. *Journal of Operational Research Society*.

Scarf, P., & Shi, X. (2005). Modelling match outcomes and decision support for setting a final innings target in test cricket. *IMA Journal of Management Mathematics, 16*(2), 161-178. doi: 10.1093/imaman/dpi010

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464.

Shao, J. (1993). Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association, 88*(422), 486-494. doi: 10.2307/2290328

Shao, J. (1997). AN ASYMPTOTIC THEORY FOR LINEAR MODEL SELECTION. *Statistica Sinica, 7*, 221-264.

Stern, S. E. (2009). An Adjusted Duckworth-Lewis Target in Shortened Limited Overs Cricket Matches. *The Journal of the Operational Research Society*, *60*(2), 236-251.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc. , 54*, 426-482.

Wilks, S. S. (1935). The likelihood test of independence in contingency tables. *Ann. Math. Statist., 6*, 190-196.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist., 9*, 60-62.

Wood, S. N. (2006). *Generalized Additive Models: an introduction with R*: Chapman & Hall/CRC.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? *Biometrika*, *92*(937-950).