Optimising Batting Partnership Strategy in the First Innings of a Limited Overs Cricket Match

by

Patrick Brown

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Master of Science
in Statistics.

Victoria University of Wellington
2017

Abstract

In cricket, the better an individual batsman or batting partnership performs, the more likely the team is to win. Quantifying batting performance is therefore fundamental to help with in-game decisions, to optimise team performance and maximise chances of winning. Several within-game metrics exist to summarise individual batting performances in cricket. However, these metrics summarise individual performance and do not account for partnership performance. An expectation of how likely a batting partnership is to survive each ball within an innings can enable more *effective* partnership strategies to optimise a team's final total.

The primary objective of this research was to optimise batting partnership strategy by formulating several predictive models to calculate the probability of a batting partnership being dismissed in the first innings of a limited overs cricket match. The narrowed focus also reduced confounding factors, such as match state. More importantly, the results are of practical significance and provide new insight into how an innings evolves.

The model structures were expected to reveal strategies for optimally setting a total score for the opposition to chase. In the first innings of a limited overs cricket match, there is little information available at the commencement and during the innings to guide the team in accumulating a winning total score.

The secondary objective of this research was to validate the final models to ensure they were appropriately estimating the ball-by-ball survival probabilities of each batsman, in order to determine the most *effective* partnership combinations. The research hypothesised that the more *effective* a batting partnership is at occupying the crease, the more runs they will score at an appropriate rate and the more likely the team is to win the match, by setting a defendable total.

Data were split into subsets based on the batting position or wicket. Cox proportional hazard models and ridge regression techniques were implemented to consider the potential effect of eight batting partnership performance predictor variables on the ball-by-ball probability of a batting partnership facing the next ball without being dismissed. The Area Under the Curve (AUC) was implemented as a performance measure used to rank the batting partnerships.

Based on One-Day International (ODI) games played between 26th December 2013 and 14th

February 2016, the model for opening batting partnerships ranked Pakistani's A Ali and S Aslam as the optimal opening batting partnership. This method of calculating batting partnership rankings is also positively correlated with typical measures of success: average runs scored, proportion of team runs scored and winning. These findings support the research hypothesis. South African's, HM Amla and AB de Villiers are ranked as the optimal partnership at wicket two. As at 28th February 2016, these batsmen were rated 6th equal and 2nd in the world respectively. More importantly, these results show that this pair enable South Africa to maximise South Africa's chances of winning, by setting a total in an optimal manner.

New Zealand captain, Kane Williamson, is suggested as the optimal batsman to bat in position three regardless of which opener is dismissed. Reviewing New Zealand's loss against Australia on 4th December 2016, indicates a suboptimal order was used with JDS Neesham and BJ Watling batting at four and five respectively. Given the circumstances, C Munro and C de Grandhomme were quantified as a more optimal order.

The results indicate that for opening batsmen, better team results are obtained when consecutive dot balls are minimised. For top order and middle order batsmen, this criteria is relaxed with the emphasis on their contribution to the team. Additionally, for middle order batsmen, minimising the occasions where 2 runs or less are scored within 4 deliveries is important.

In order to validate the final models, each one was applied to the corresponding Indian Premier League (IPL) 2016 data. These models were used to generate survival probabilities for IPL batting partnerships. The probabilities were then plotted against survival probabilities for ODI batting partnerships at the same wicket. The AUC was calculated as a metric to determine which models generated survival probabilities characterising the largest difference between IPL partnerships and ODI partnerships. All models were validated by successfully demonstrating the ability of these models to distinguish between higher survival probabilities for ODI partnerships compared with IPL partnerships at the same wicket.

This research has successfully determined ball-by-ball survival probabilities for individual batsmen and batting partnerships in limited overs cricket games. Additionally, the work has provided a rigorous quantitative framework for optimising team performance.

Contents

Chapter 1

Introduction to Sport Analytics

Professional team sport is fuelled by winning. Bigger prize money is earned by teams who win, and win often [86]. Winning teams attract crowds, leading to increased viewership, sponsorship, gate takings and sale of merchandise. But what does it take to win?

Batting first in a limited overs game of cricket has many uncertainties. This starts with what is the total to set for the opposition to chase, which maximises the chance of victory? Secondly, what is the best approach to at least achieve that total?

In this thesis, sports analytics will be used to define strategies that optimise the chances of winning for the team batting first.

In [19, p.1], sports analytics is defined as "the management of structured historical data, the application of predictive analytic models that utilize such data, and the information systems used to inform decision makers and enable them to help their organizations in gaining a competitive advantage on the field of play". Statistical measures within sports have existed for several centuries. In reference to baseball, it was noted in [28, p.1] that "in the mid-19th century, Henry Chadwick is credited with developing the box score and his tabulation of hits, home runs and total bases led to the formulation of metrics such as batting average and slugging percentage". Additionally, in [45], there is an extensive amount of cricket match data, dating back to 1864. However, "the statistics used are not reflective of true performance measures" [84, p.11], while the amount of historical analysis in cricket is scarce. The application of complex analytical methods within sports has only boomed within the last half century [6].

During recent decades, there has been a huge growth in sports analytics together with increased demand for insightful sport related statistics. This is due to the large increase in player salaries and total revenue generated within sports. In the National Football League (NFL), average player salaries increased by approximately 440% between 1989 and 2009. The total revenue generated in the NFL increased by approximately 750% during the same period [94].

The authors in [79] predicted that global sports revenue would grow by US$145.3 billion between 2010 and 2015. Given that teams that win earn bigger prize money, increased revenue and achieve increased economic growth, it is important for managers and coaching staff of sport teams to maximise their chances of success. To succeed, it is essential that player selection and strategic decisions include the use of analytical techniques [20]. Furthermore, increased sports revenue leads to increased spectator appeal and entertainment. In [27], data mining techniques using Advanced Scout software were applied to National Basketball Association (NBA) data from the 1995-1996 NBA season. These techniques enabled NBA coaching staff to discover interesting patterns in basketball game data. This helped to increase the quality of play in the NBA, which increased spectator appeal and entertainment.

Given the extensive amount of numerical data generated by sports, it is important that the data is well utilised to extract insightful information. Numerical data is collected by data collection centres such as Opta [9]. For field sport athletes, the data is generated using advanced Global Positioning System (GPS) technology [43]. Numerical data helps teams make decisions on coaching techniques, player selections and objective strategies. This is important to maximise the chances of winning. Numerical data is also generated to help with in-game betting decisions. The results generated from applying statistical techniques to sport related data are called sport statistics. Sport statistics can be broadly categorised as either performance indicators or performance outputs [29]. Performance indicators are a quantitative measure used to indicate individual performance in a particular area of the game. These are collated during the game. In contrast, performance outputs are summary measures detailing direct result of participation in an event. These are described on a score sheet.

There is a breadth of academic literature applying various statistical techniques to sports, ranging from measuring team performance to individual key performance indicators [29]. In [42], the authors analysed the sprinting activities of different playing positions during European Champions League and UEFA Cup matches, contested between 2002 and 2006. The authors conducted several Kruskal-Wallis analyses to compare positional differences. The results found statistically significant differences in the total number of sprints and total sprint distance covered during explosive sprints made by players in different positions. In [22], the authors challenged the adequacy of existing paired comparison models used for ranking foot-

ball teams. These models focus on either wins and losses or points scored, but not both. As such, the authors suggested that the models fail to produce satisfactory rankings. A hybrid paired comparison model was developed as an alternative. This model incorporates both win and loss records and points totals to give a measure of winning difficulty. The model was applied to two sets of simulated data and performed better than the original models in both cases. In [53], the authors developed a modified least squares system to rank American men's college basketball teams. This system was applied to data from the 1999-2000 basketball season and 1999-2001 football season. The system was used to predict the contestants of 73 post season football games and 93 post season basketball games. The results showed that the modified least squares system predicted these games correctly with 76.3% accuracy, compared with a 74.2% predictive accuracy using basic least squares.

The application of analytical techniques to sport has been studied extensively. Analytical techniques enable a cricket team to better understand factors behind winning. These are used to make more *effective* strategic decisions to increase the team's chances of winning.

1.1 Analytics in Cricket

Cricket is a team sport in which statistics feature heavily. A game of cricket is contested between two teams of eleven players. One team is assigned the batting team. Pairs of batsmen work together as partnerships, with the objective of scoring as many runs towards to the team total as possible. A batsman can typically score 0 (known as a dot ball),1,2,3,4,5 or 6 runs, or can be dismissed. A dismissal is referred to as a wicket. The opposition team is assigned the fielding team. Each phase of play is called an innings and the length of an innings depends on the format. This is discussed further in Section 1.2.

Statistical measures in cricket have existed since the mid-nineteenth century, while the use of analytical techniques within cricket has grown substantially during the last half century [6]. As such, the breadth of recent analytical cricket literature is fairly extensive. It was noted in [70, p.1] that "during the past decade a large number of papers have been published on cricket performance measures and prediction methods". Data mining involves applying statis-

tical techniques to transform data into insightful information. In [84, p.1], it was noted that "data mining is quickly becoming an integral part of the sports decision making landscape where managers/coaches using machine learning and simulation techniques can find optimal strategies for an entire upcoming season". In addition, the recent use of data mining techniques and knowledge management tools within cricket has been successful. For example, in [21], multinomial logistic regression techniques effectively showed that a team's batting and bowling strength, first innings lead, batting order and home advantage were all strong predictors of winning in test matches. Sports analytics is based on team performance, individual performance and forecasting, all of which are relevant to this thesis.

1.1.1 Team Performance

In [37], a dynamic programming model was applied to one-day cricket to calculate, at any stage of an innings, the optimal scoring rate, an estimate of the total number of runs to be scored in the first innings or an estimate of the probability of winning in the second innings. Each team is allocated batting resources in the form of wickets and balls. The objective of the team batting first is to maximise the number of runs scored. The aim is to set a score target that the team batting second does not proceed to achieve. As such, the author produced the following first innings formulation:

$$f_n(i) = \max_R \left[p_d \times f_{n-1}(i-1) + \frac{R}{6} + (1-p_d) \times f_{n-1}(i) \right], \tag{1.1}$$

where f_n(i) represents the maximum expected score in the remaining n balls and i wickets in hand, p_d denotes the probability of dismissal, and R denotes the run rate per over. For each number of balls remaining and wickets in hand, the author calculated the optimal run rate and the expected score in the remainder of the innings. The results suggested that "teams should try to score slightly faster than they expect their average rate for the rest of the innings to be, and if wickets are lost, slow up, rather than the current practice of scoring slower than average and speeding up if wickets are not lost" [37, p.333].

The objective of the team batting second is to maximise the probability of achieving a score higher than that achieved by the team batting first. As such, the author defined a new variable,

s, as the number of runs needed to reach a particular score. The author suggested that on "each ball, a batsman either goes out with probability p_d and the team still has s runs to score with one less wicket in hand and one less ball, or scores X runs with probability p_x and so the team has $s-x$ runs to score with one less ball to go and the same number of wickets in hand" [37, p.334]. As such, the author produced the following second innings formulation:

$$p_n(s,i) = \max_R \left[p_d \times p_{n-1}(s, i-1) + \sum_{0 \leq x \leq 6} p_x \times p_{n-1}(s-x, i) \right] \tag{1.2}$$

where $p_n(s,i)$ is the probability of scoring at least another s runs with i wickets in hand and n balls remaining. For each ball, each wicket in hand and each number of runs to go, the author calculated the probability of winning and the optimal run rate. The author claimed that the first innings conclusions also apply to the second innings. In addition, the author suggested that the team batting second has an advantage as "in practice, the first team would begin their innings with much less knowledge of the state of the wicket than the second" [37, p.335].

In [91], the authors used a combination of simulation, Bayesian log-linear modelling and simulated annealing to investigate optimal batting orders in the Indian national cricket team. For India's 2003 World Cup final batting order, the authors initially considered an algorithm to simulate runs scored in an innings. Ignoring rare occasions when five runs are scored, the authors let X_i denote the outcome of the ith ball for $i = 1, \ldots, 300$, where

$$X_i = \begin{cases} 0, & \text{if batsman scores 0 runs,} \\ 1, & \text{if batsman scores 1 run,} \\ 2, & \text{if batsman scores 2 runs,} \\ 3, & \text{if batsman scores 3 runs,} \\ 4, & \text{if batsman scores 4 runs,} \\ 5, & \text{if batsman scores 6 runs,} \\ 6, & \text{if batsman is dismissed,} \end{cases} \tag{1.3}$$

and $X_{m+1} = \cdots = X_{300} = 0$ if the innings ends on the mth ball, where $m < 300$. The joint

distribution of $X_1, \ldots, X_{300}$ was defined as:

$$[X_1, \ldots, X_{300}] = [X_{300}|X_1, \ldots, X_{299}][X_{299}|X_1, \ldots, X_{298}] \ldots [X_2|X_1][X_1|X_0] \tag{1.4}$$

The algorithm generates $X_i \sim [X_i|X_1, \ldots, X_{i-1}]$ and the number of runs scored. The authors proceeded to estimate the distributions $[X_i|X_1, \ldots, X_{i-1}]$ using estimated batting characteristics. These included the number of wickets lost and number of balls bowled, obtained from 71 ODI games, in which India batted first. These estimates were calculated using a Bayesian log-linear model developed through WinBUGS software [87]. Subsequently, the authors were interested in deriving optimal batting orders in one-day cricket for India. The objective function of this optimisation procedure was defined as the mean number of runs scored per innings. Approximation of the objective function required many simulations to be carried out. Additionally, there were 240 million feasible batting orders for the optimisation procedure to consider. As such, the authors applied a simulated annealing, probabilistic search algorithm, discussed in [66]. This was used to "explore the space of permutations of batting orders" [91, p.1940]. The results suggested two potentially optimal batting orders for India. Based on a comparison with the Indian batting order adopted in the 2003 World Cup final, these batting order suggestions were found to potentially improve ODI performance by approximately six runs.

1.1.2 Individual Performance

Statistical measures have been used to assess player performance in cricket. In [64], the authors investigated the differences between the leading 12 wicket takers in English First-Class cricket during the 1991 season. The authors considered three performance measures: bowling average, economy rate and strike rate. The data are summarised in Figure 1.1. A scatter diagram in Figure 1.2 was constructed to compare and illustrate these differences in player performance.

Bowler	Balls	Runs	Wickets	AV	ER	SR
(WY) Waqar Younis	3492	1656	113	14.65	47.65	30.90
(NF) N.A. Foster	4544	2138	102	20.96	47.05	44.53
(PT) P.C.R. Tufnell	5420	2219	88	25.21	40.94	61.57
(AD) A.A. Donald	3135	1634	83	19.68	52.12	37.73
(FS) F.D. Stephenson	4315	2010	78	25.76	46.58	55.32
(JM) J.N. Maguire	4716	2437	77	31.64	51.68	61.25
(DL) D.V. Lawrence	3091	1790	74	24.18	57.91	41.77
(SW) S.L. Watkin	4373	2175	74	29.39	49.74	59.09
(PD) P.A.J. DeFreitas	3943	1780	73	24.38	45.14	54.01
(TM) T.A. Munton	4159	1863	73	25.52	44.79	56.97
(JE) J.E. Emburey	5397	2170	68	31.91	40.21	79.37
(RP) R.A. Pick	3904	2080	67	31.04	53.28	58.27

Figure 1.1: Statistics for the leading 12 wicket-takers in English First-Class cricket for the 1991 season (Figure obtained from [64])

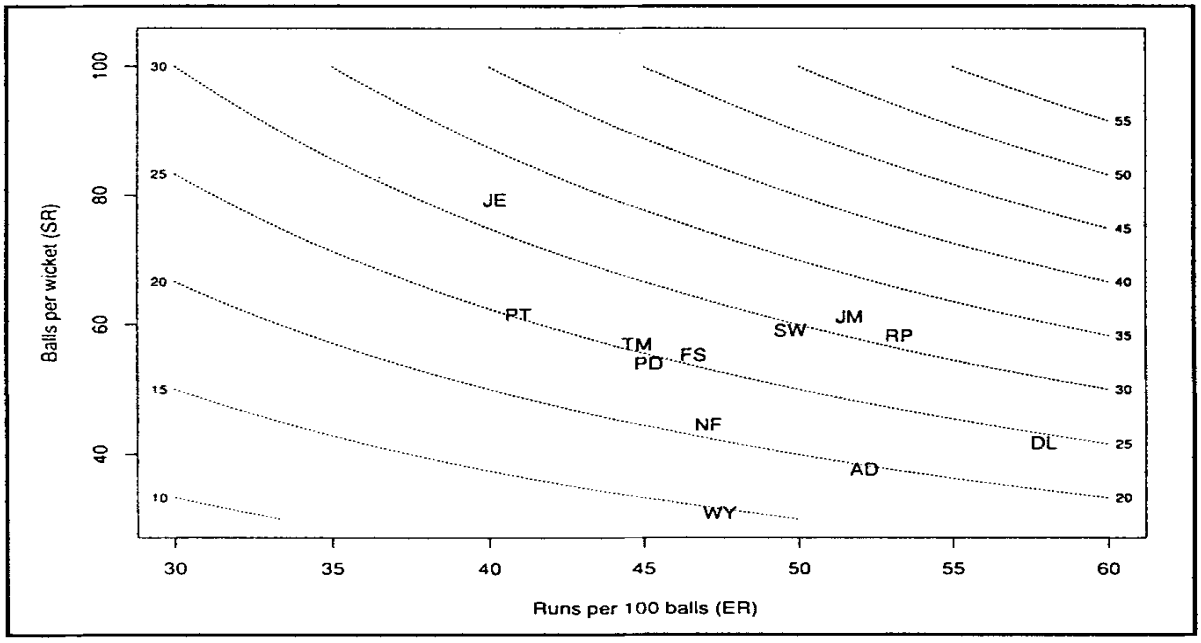

Figure 1.2: Augmented scatter plot of the twelve highest wicket-takers in English First-class cricket in 1991 (Figure obtained from [64])

The bowling average was defined as the total number of runs conceded per wicket taken. The economy rate was defined as the number of runs conceded per 100 balls. The strike rate was defined as number of balls bowled per wicket taken. The scatter diagram illustrated that 'good'

spin bowlers and 'good' fast bowlers were located on separate areas of the graph. In addition, the scatter diagram illustrated that fast bowlers had a low strike rate and a high economy rate. Whereas, spinners had a high strike rate and a low economy rate.

In [30], reject inference methods (RIMs) were applied to estimate the likelihood of a bowler taking a wicket within an innings. Ball-by-ball data from the 2015 IPL was used for analysis. To be consistent with the terminology used in reject inference methods (RIMs), bowlers were categorised according to whether or not they had defaulted and whether they were accepted or rejected for inference. Non-default bowlers were defined to be those with strike rates in the top 25%. All other bowlers were labelled as defaulted. Those bowlers who had not taken any wickets were labelled as rejected. The accepted wicket taking bowlers were referred to as "AccApps" and all bowlers were referred to as "AllApps". Through utilisation of the Generalised Boosted Machine Model (GBM) decision-tree based approach, the top five variables selected from a possible 14 were: economy rate, number of dot balls bowled, runs conceded, number of boundaries and number of bowler-penalised extras. A logistic regression analysis was carried out to test these variables and predict the probability of default. When all five variables were included, a Receiving Operating Characteristic (ROC) curve was constructed and the resulting Area Under the Curve (AUC) was 0.832. When the number of variables was restricted to three, the AUC became 0.833. The inclusion of the additional two variables resulted in no significant model improvement. As such, only the top three variables were retained for analysis. Logistic regression models were used to compare RIMs. Each RIM logistic model was used to predict the probability of default. The outcome variable, strike rate, for rejected bowlers was computed using an observed strike rate regression model. This model was fitted on all accepted bowler observations with the three variables included. To assess the effectiveness of these models, several summary statistics were calculated. These included a modified type II error, AUC and Mean Absolute Difference (MAD) between the predicted probabilities of "AllApps" and the model being tested. The results found that the Memory Based Reasoning logistic model generated the highest AUC and had a reasonably low MAD. This model was selected as the preferred model. All predicted probabilities were transformed to give an estimate of strike rate. The number of wickets per bowler per innings was then estimated by the estimated strike rate divided by the number of balls bowled. The model predicted that bowlers who bowl a high

number of dot balls and have a low economy rate have a higher probability of taking wickets.

In [32], a mixed distribution, called the Ducks 'n' Runs distribution was proposed. The distribution consisted of a beta distribution to model ducks (zero scores) and a geometric distribution to model runs (non-zero scores). Runs scored and the probability of failure to contribute to the team total were used to evaluate individual batsmen in an innings. The probability model was assessed at a macro (similar batsmen) and micro (individual batsmen) level. Data associated with New Zealand first class batsmen, over four domestic seasons between 1994 and 1998, were used for analysis. At the macro level, four groups of batsmen were created, based on batting position. In each group, the scores of the batsmen were grouped into 20 bins $(0, 1-2, 3-6, 7-10, 11-20, 21-30, \ldots, 141-150, 151-200, >200)$ and the observed proportion of scores were compared with expected probabilities. Scatter plots were constructed and illustrated a strong linear relationship between the observed and expected instances of scores. This indicated that the Ducks 'n' Runs model was a good approximation for batting scores. At the micro level, the probability model for individual scores was fitted to all individuals who participated in twenty or more innings. This was used to calculate the proportion of Ducks, number of scores > 50 and number of scores > 100 an individual was expected to score. The results showed that experiment-wise p-values were less than 5% for all three measures, indicating that the Ducks 'n' Runs distribution adequately models individual batting scores. Control charts based on quartiles of individual batting scores, were developed, to monitor individual batting performance. The control charts successfully demonstrated the occurrence of significant batting performance changes. This enabled identification of changes in individual player form and ability. This may help to inform coaches to aid the delivery of higher quality player feedback.

In [82], the differences between great batsmen of different eras were examined. The authors suggested that the traditional method of calculating a batsman's average may be justified assuming the runs scored in several incomplete and complete innings are exponentially or geometrically distributed. In [39], the authors observed that the exponential distribution and geometric distribution are inadequate representations of batting scores. As such, the authors in [82] examined the exponential and Weibull distributions to estimate batting consistency. Data were collected from the ESPN Cricinfo website and consisted of the scores of 25 batsmen who had

scored 8000 or more international test runs. Additionally, four batsmen who had scored more than 6000 international test runs with an average over 55 were included. Figure 1.3 illustrates a histogram resembling the exponential distribution, summarising all innings scores for SR Tendulkar.

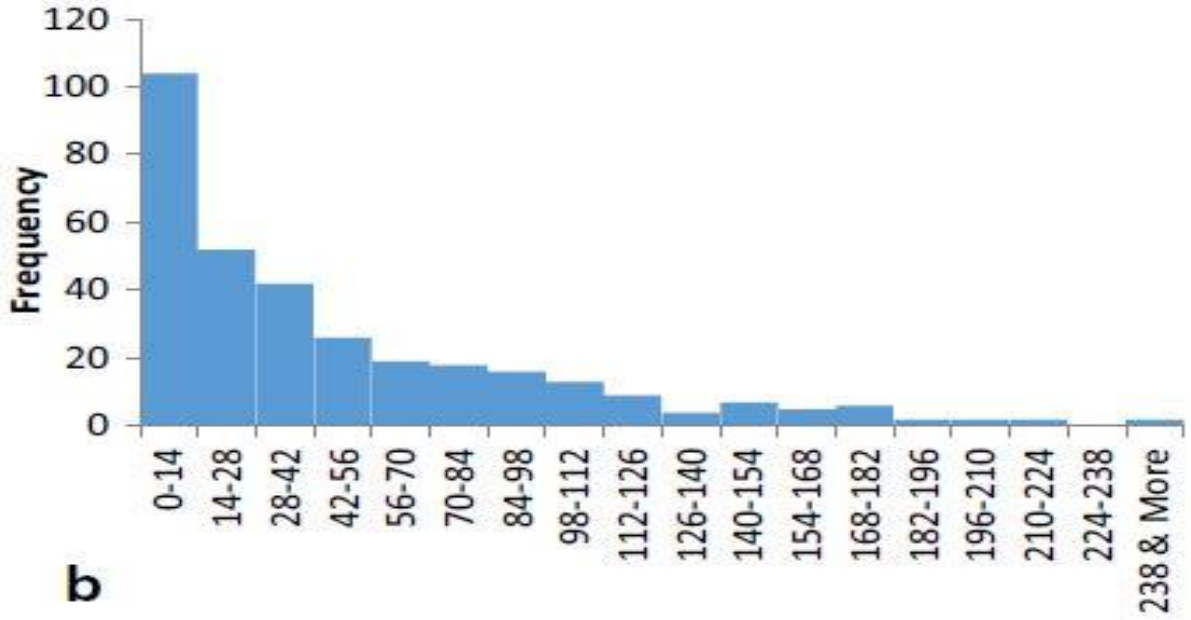

Figure 1.3: Histogram of all innings scores of SR Tendulkar (Figure obtained from [82])

Pearson's chi-square test was applied to assess the goodness of fit for the exponential distribution and the result suggested poor fit. This was repeated for other players and the exponential distribution was found to inadequately model the batting scores in approximately 50% of cases. As an alternative, the authors considered the Weibull distribution. The probability density function of a Weibull(α,θ) distribution was given by:

$$f(z) = \left(\alpha \frac{z^{\alpha-1}}{\theta}\right)\left(\exp^{-\frac{z^{\alpha}}{\theta}}\right), z \geq 0, \alpha > 0, \theta > 0. \tag{1.5}$$

Assuming the batting scores form a random sample from a Weibull(α,θ) distribution, the authors calculated the maximum likelihood estimates (MLEs) of the parameters for each player. Pearson's chi-square test was applied to assess the goodness of fit of the Weibull distribution, with parameters estimated using maximum likelihood methodology. The Weibull distribution was found to provide a better fit than the exponential distribution in 26 out of 32 cases. For the batsmen considered, the MLE of the batting standard deviation was found to provide an

informative estimate of batting inconsistency, while the MLE of the batting mean was found to be slightly higher than the traditional batting average. The batting mean and batting standard deviation estimates were two of five measures used to rank the batsmen. Longevity is a measure used to increase the ranking of batsmen who have performed well over a long period of time. This is often incorrectly ignored in a ranking analysis [81]. The authors in [82] adopted longevity as a third ranking criteria. As a fourth measure, an index for quality-runs scored as a function of opposition strength was generated. The fifth criteria was a measure of diversity of opponent teams encountered by a player. The Mahalanobis distance is a multi-dimensional generalisation of how many standard deviations a point is from the mean of its distribution [71]. The authors in [82] incorporated different combinations of each criteria as a measure of Mahalanobis distance. This was used as an overall ranking metric. DG Bradman was ranked as the best batsman, followed by SR Tendulkar, L Hutton, KF Barrington and J Kallis.

In [31], time series clustering analysis was used to map the test career progression of Australian cricketing legend Sir Donald Bradman, acknowledged as the greatest batsman of all time with an exceptional career batting average of 99.94, from 80 innings. However, during the Second World War, between 1939 and 1945, all international cricket was suspended. Of interest was whether or not Bradman's prime would have occurred during the Second World War period. To assess this, the authors utilised time series clustering to characterise Bradman's test career and compare him with other test players. These players had all batted in at least 70 innings during a career spanning at least 17 years as of 1st January 2009. The selected clustering method was based on global characteristics measures as "it does not require many conditions to be true before it can be used" [31, p.4]. Moreover, this approach clusters global features extracted from individual time series rather than using a distance measure. As such, this approach can be applied on different length time series. The scaled average contribution per calender year was then modelled using weighted least squares regression. The parameters of the model were used to cluster the batsmen into groups. The results showed that Bradman's career progression was most similar to West Indian legend Brian Lara, indicating that Bradman's peak performance would have occurred in the 12th to 14th years of his career (1939-1941), coinciding with World War II. The authors proceeded to impute Bradman's likely scores for 1939-1945 and estimated his batting average to be 105.41. Interestingly, there was insufficient evidence to suggest that

this estimated average was significantly higher than Bradman's actual career average of 99.94.

1.1.3 Forecasting

In [77], a roster based optimisation system that selects an optimal cricket team of eleven players from a list of players that generates the greatest probability of winning, was developed. The authors hypothesised that a team rating system accounting for individual player metrics will perform more effectively than a system that only accounts for variables on the wider team, such as home advantage, opposition strength and past team performances. Random forest techniques were utilised to identify individual cricket player performance metrics that had a significant effect on the player's contribution to the proportion of team wins achieved. The five most important batting metrics were: strike rate, number of balls faced, batting average, total runs scored and percentage of boundaries. The five most important bowling metrics were: economy rate, bowling average, strike rate, percentage of boundaries and percentage of dot balls. The roster based optimisation method required implementation of a binary decision variable, Y, to indicate whether the player was selected or not. This was categorised as:

$$Y = \begin{cases} 1, & \text{if player selected} \\ 0, & \text{otherwise} \end{cases} \tag{1.6}$$

As part of the adaptive rating system, an individual rating system was implemented using a combination of the Product Weighted Measure + Analytical Hierarchy process and Exponentially Weighted Moving Averages technique. The developed optimisation system was applied to the 2015 IPL. The predictive accuracy of the developed system was assessed by comparing the system's performance with that of the New Zealand Totalisator Agency Board (TAB) and the CricHQ algorithm. The results found that the performance of the adaptive rating system was 20% better than the TAB and 13% better than the CricHQ algorithm. The results of this research supported the hypothesis that cricket team ratings based on individual performances are more effective than those based on team performances.

1.2 Formats of Cricket

International cricket games are categorised as either test matches, one-day internationals or Twenty20 (T20) matches. One-day international and T20 formats are categorised as limited overs cricket due to restrictions on the number of overs allocated to each team, and the number of overs an individual may bowl during an innings. In one-day internationals, each batting team bats for one innings and is allocated 50 overs, compared to an allocated 20 overs to the batting team in T20 cricket. In addition, the number of fielders allowed in a particular area of the field at any time is restricted. In contrast, in test matches, each team is given two innings to bat. Test matches may last for five days. There are no restrictions on the number of allocated overs each team is given, the number of overs a bowler may bowl or fielding positions.

1.3 Purpose of Research

Recent developments in cricket have included the introduction of T20 cricket. In this format, teams are restricted to 20 overs to score as many runs as possible. This results in faster, more intensive games compared with those contested in other formats, with the aim of increasing spectator appeal. As such, the amount of revenue generated by cricket globally has increased. It was claimed in [2] that global cricket will generate total revenues of approximately $2.5 billion between 2014 and 2022. In addition, the IPL experienced 4% growth in brand valuation between 2012 and 2013 [47]. The Big Bash T20 League (BBL) is Australia's domestic T20 cricket competition, established in 2011. The average attendance in the BBL increased by 22% between 2015 and 2016, while TV ratings increased by 11% [17]. Furthermore, BBL merchandise sales increased by 44% between 2014 and 2015 [17]. This research was motivated by the rapid growth within cricket, and the importance in understanding the factors behind a cricket team's ability to win.

There exists a substantial amount of previous research into the use of analytical techniques in cricket. However, the amount of research covering in-game analysis is scarce. Given the rise in popularity of in-game betting in the 1990's [58] and further rise more recently [8], the need for in-game analysis is also increasing.

An expectation of how likely a batting partnership is to survive each ball within an innings can aid the development of more *effective* partnership strategies to optimise a team's final total. The primary objective of this research was to optimise batting partnership strategy by formulating several predictive models to calculate the probability of a batting partnership being dismissed in the first innings of a limited overs cricket match. The narrowed focus also reduced confounding factors, such as match state.

The model structures were expected to reveal strategies for optimally setting a total score for the opposition to chase. In the first innings of a limited overs cricket match, there is little information available at the commencement and during the innings to guide the team in accumulating a winning total score.

The secondary objective of this research was to validate the final models to ensure they were appropriately estimating the ball-by-ball survival probabilities of each batsman, in order to determine the most *effective* partnership combinations. The research hypothesised that the more *effective* a batting partnership is at occupying the crease, the more runs they will score at an appropriate rate and the more likely the team is to win the match, by setting a defendable total.

Specifically, the purpose of this research was to address the following two key questions:

1. What are the in-game strategies for optimising the runs scored in the first innings?

2. What are the practical applications of this knowledge?

1.4 Publications Arising

Early work from this thesis, in [36], was peer-reviewed and published in Mathsport 13 Conference proceedings. More detailed analysis, in [35], presented throughout Chapter 8, was peer-reviewed and published in Mathsport International 2017 Conference proceedings.

1.5 Structure of Thesis

Chapter 2 discusses various applications of survival analysis to performance in non-cricket sports and cricket that were identified in the academic literature. Chapter 3 consists of research objectives and methodology, which define the research questions and describes the adopted methodology. In Chapter 4, data extraction and processing procedures are described, before exploratory data analysis is presented in Chapter 5. An overview of ridge regression is provided in Chapter 6. Survival analysis is introduced in Chapter 7. Model application and model validation techniques are presented in Chapters 8 and 9 respectively. Chapter 10 discusses the adopted methodology and the link between the research objectives and previous research. Potential areas of future work are suggested, before a summary of the key findings of this research is presented.

Chapter 2

Literature Review

In Chapter 1, literature focussing on general analytical applications in cricket was discussed. This chapter outlines the academic literature concerned with the application of survival analysis techniques to both non-cricket sports and cricket. The chapter starts with a wider review of literature focussing on the application of survival analysis techniques to team sports. Major sports covered include baseball, basketball and football. The chapter proceeds by addressing the application of survival analysis techniques to cricket. The chapter concludes by linking the academic literature with the objective of this research.

2.1 Survival Analysis in Non-Cricket Sports

The application of survival analysis within a sports context has been studied extensively, with a literature review revealing a number of publications addressing this particular area of sport analytics.

In [85], survival analysis techniques were applied to investigate manager retention rates in baseball, basketball and football. Kaplan-Meier survival curves were fitted, with 95% confidence bands proposed in [51], for managerial survival for each of baseball, basketball and football. The authors examined whether the survival probabilities differed between the three sports. In [83], the log-rank test was suggested as the appropriate test. This test uses the chi-squared statistic to compare the actual and predicted failures for each sport. The result from this test indicated that the survival curves for each sport were statistically different from each other. The authors also investigated which distribution was the most appropriate at describing the survival probabilities. For simplicity, the distributions considered were restricted to exponential and Weibull forms. To determine which distribution was most appropriate, the Weibull regression model was initially implemented with parameters, p and λ, equivalent to respective parameters, α and θ, from Equation (1.5). These were estimated using maximum likelihood

estimation. An estimated shape parameter greater than one indicates that the underlying distribution is Weibull. This distribution reduces to an exponential when the estimated shape parameter is equal to one [52]. The estimated shape parameter, p, was greater than one in all models. This suggested that the Weibull distribution provided a more accurate description of managerial survival rates than the exponential. As a result, the Weibull regression technique was utilised for each sport. In each regression model, managerial efficiency was used as a covariate and regressed against managerial tenure, how long the manager stays with the team. Managerial efficiency was calculated as a comparison of the manager's winning percentage with the manager's maximum win percentage. The results showed a highly significant positive relationship between managerial efficiency and managerial tenure in all three sports: baseball, basketball and football. This suggested that the higher the proportion of games won by a manager, the longer the manager will stay with the team.

In [93], the author investigated the effects of minority status on managerial survival rates in Major League Baseball. Major League Baseball manager data between 1986 and 2005 in which the manager managed five consecutive games were used for analysis. The dependent variable indicated whether the manager returned for another season or not. As an extension to the work implemented in [85], a number of different performance and individual characteristics were implemented. Table 2.1 summarises the explanatory variables used.

Explanatory variable	Definition
Minority status	Equal to one for black and Hispanic managers; zero otherwise
Efficiency	Score based on how efficiently the manager transforms his given resources into wins
Winning percentage	The winning percentage for the team during the portion of the year that the manager was with that team
Play-off wins	The team's winning percentage during League Championship Series games and World Series games
Experience	Number of years spent in baseball as a coach or player
Age	Manager's age

Table 2.1: Explanatory variables

Technical efficiency was calculated using data envelopment analysis, based on the output-oriented technical efficiency proposed in [24]. The output-oriented technical efficiency refers to the maximum output (win percentage), given that level of inputs. The player salaries were divided into offensive and defensive salaries, used as two inputs. "For offense, the salaries of the players who played the greatest number of games at each infield position and the top three outfielders in games played are summed for each team of each year from 1985 to 2006. For defense, the salaries of the top five pitchers in terms of games started and the top six pitchers in terms of relief appearances are summed for each team of each year" [93, p.528]. The level of talent on opposing teams was another potential factor behind a team's winning percentage. As such, the salary of the competition was included as a third input, calculated as the average total salary of the other teams in the team's division. A price index was created by calculating the average player salary for each season from 1985 to 2006, and dividing the 2006 average salary by that of the other seasons. This was used to adjust all player salaries into 2006 baseball dollars. The authors then solved a maximisation problem to maximise θ, the multiple by which the win percentage could be increased using a feasible combination of observed inputs and subject

to several constraints:

1. $\sum \lambda_i W_i \geq \theta W_0$,
2. $\sum \lambda_i O_i \leq O_0$,
3. $\sum \lambda_i D_i \leq D_0$,
4. $\sum \lambda_i C_i \geq C_0$,
5. $\sum \lambda_i = 1$,
6. $\sum \lambda_i \geq 0$,

where 0 identifies values for the team being analysed and i identifies the other teams in the comparison group. In addition, W,O,D and C represent winning percentage, offense, defense and competition, respectively.

"Constraint (1) implies that the combination of other observed winning percentages must be greater than or equal to the observed winning percentage of the manager being evaluated. Constraints (2) and (3) imply that the combination of offensive and defensive inputs must be less than the inputs of the manager under consideration. Equation (4) states that the combination of the negative competition inputs must be at least as great as the competition faced by the manager being evaluated. Constraint (5) implies variable returns to scale by eliminating scaled down versions of one input bundle from the feasibility set. The last constraint simply assures that there are no negative inputs" [93, p.530]. The authors then calculated the inverse of θ to give an efficiency score between zero and one.

The authors proceeded by fitting four different models using different combinations of the predictors described in Table 2.1. These models included an exponential, Weibull, Gompertz and Cox proportional hazards model. Under all models, the results found that winning percentage and play-off wins had a positive effect on managerial survival, as expected. However, these effects were highly insignificant when efficiency was added to the model. The correlation between winning percentage and efficiency was 0.75. Similarly, the correlation between play-off

wins and technical efficiency was 0.27. The authors claimed that these correlations were a possible cause of the observed insignificant effects when efficiency was included as a covariate. As a result, the four models were estimated excluding winning percentage and play-off wins. The final four covariates included minority status, efficiency, experience and age. Tables 2.2 and 2.3 summarise the resulting hazard ratios and p-values.

Predictor	**Exponential**		**Weibull**	
	Hazard ratio	**P-value**	**Hazard ratio**	**P-value**
Minority	0.71391	0.176	0.52267	0.011
Efficiency	0.07379	0.000	0.06400	0.000
Experience	0.99972	0.031	0.99908	0.000
Age	1.03797	0.007	1.05566	0.000

Table 2.2: Hazard ratios and p-values Exponential and Weibull.

Predictor	**Gompertz**		**Cox proportional hazard**	
	Hazard ratio	**P-value**	**Hazard ratio**	**P-value**
Minority	0.60326	0.047	0.63369	0.072
Efficiency	0.06786	0.000	0.07293	0.000
Experience	0.99926	0.000	0.99956	0.006
Age	1.05202	0.000	1.04072	0.005

Table 2.3: Hazard ratios and p-values Gompertz and Cox proportional hazard.

Under all models, minority status, efficiency and experience all had hazard ratios of less than one. This suggested that an increase in these covariates was associated with an increase in the probability of survival. Age had a hazard ratio greater than one suggesting an increase in age was associated with a reduction in probability of the manager surviving. In order to evaluate which model was the most appropriate, each model was plotted with the Kaplan-Meier survival function overlaid. The plots suggested that the Cox proportional hazards model provided the best fit in approximating the Kaplan-Meier survival function. All covariates in this model were statistically significant at the 8% level. The proportional hazards assumption was assessed for this model in two ways. The first involved estimating the model with time by covariate interac-

tions included in the model. None of these interactions resulted in significance, suggesting that the proportional hazards assumption may have held. The second involved constructing a plot of the residuals against time. The slope of the plot was equal to zero, which provided further evidence in support of proportional hazards.

In [76], the author investigated factors that affect the quit behaviour of professional baseball players in Japan. The author considered both pitchers and batters who played between 1977 and 1990 and applied Cox proportional hazard methodology. The dependent variable was defined as the time until the player quit. Duration was defined as the number of years since the player entered the baseball league. The author used wages, productivity and their quadratic terms as explanatory variables. Wages was defined as the wage of the player in the previous year. For batters, productivity was measured as the slugging rate:

$$\text{slugging rate} = \frac{\text{total bases}}{\text{at bats}}. \tag{2.1}$$

For pitchers, productivity was defined in two ways. The first referred to the hit rate:

$$\text{hit rate} = \frac{\text{number of hits given up}}{\text{number of hitters confronted}}. \tag{2.2}$$

The second referred to the strike to walk rate:

$$\text{strike to walk rate} = \frac{\text{number of strikeouts}}{\text{number of bases on balls}}. \tag{2.3}$$

The results found that a higher income discouraged quitting among both batters and pitchers. In addition, the effect of productivity differed between pitchers and batters. For batters, higher productivity was associated with a reduction in probability of quitting. For pitchers, higher productivity was associated with an increase in probability of quitting, regardless of which definition of productivity was assessed. This suggested that there may be other factors at play such as the impact on the body. For example, for batters, high productivity may put more strain on the body. As such, these results could suggest that the higher the impact on the body, the less likely the batter is to quit.

In [61], the author investigated the effect of racial difference on the retention probability of

NBA coaches between 1996-1997 and 2003-2004. The author carried out three separate survival analyses, using a different definition of failure in each. Each coaching spell was used as an observation and failure was firstly defined as a coach exiting. In the second analysis, failure was defined as a coach exiting specifically by quitting. In the third analysis, failure was defined as a coach exiting specifically by being discharged. The dependent variable was the managerial survival time. The author applied hazard function methodology, assuming a log-logistic accelerated failure time model. Table 2.4 summarises the explanatory variables used.

Explanatory variable	Definition
Race	Dummy variable equal to one if coach's race is white; zero otherwise
Winning percentage	The team's current season winning percentage
Payroll	Log of the team's real payroll relative to league average for a given year
Age	Coach's age
NBA head coaching	Coach's number of years of NBA head coaching
College head coaching	Coach's number of years of college head coaching
Non NBA experience	Coach's number of years of professional head coaching experience other than NBA
NBA assistant coach	Coach's number of years spent as assistant coach in NBA
NBA winning percentage	Coach's lifetime NBA head coaching winning percentage
Play-offs	Coach's number of years in NBA play-offs as head coach
Playing experience	Coach's years of NBA or American Basketball Association (ABA) playing experience
Player coach	Number of NBA or ABA all-star teams to which the coach was named as player

Table 2.4: Explanatory variables

The results found that 27% of white coaches and 28% of black coaches had job separations in an average year. In addition, 18% of white coaches and 20% of black coaches were discharged on average, while 9% of white coaches and 8% of black coaches quit in an average year. As such, no statistically significant racial differences on the overall probability of exit, quitting

or discharge were found. The results failed to provide evidence to suggest any presence of retention, hiring or wage discrimination against current black NBA coaches.

In [26], the authors investigated factors that were capable of determining how long head coaches in the German Bundesliga football league would survive between 1981-1982 and 2002-2003. The dependent variable, Y, was defined as a head coach dismissal, categorised as:

$$Y = \begin{cases} 1, & \text{if head coach dismissed} \\ 0, & \text{otherwise} \end{cases} \tag{2.4}$$

The authors applied Cox proportional hazard methodology to model the head coach tenure within the current team. Several predictor variables were considered. These are summarised in Table 2.5.

Explanatory variable	Definition
RSAL	The relative salary of the head coach, calculated as the individual salary divided by the average salary of all head coaches in the respective season
RWB	The relative wage bill of the team, calculated as the team wage bill divided by the average wage bill of all teams in the respective season
Bosman effect	Dummy variable equal to one for the period from when Bosman effect started (from 1995/1996); zero otherwise
CWP	The win percentage of the head coach
RP	The relative number of points won, calculated as the number of points accumulated by the team divided by the average number of points won by the average team in the respective season
CEXP	Experience of the head coach in the Bundesliga, measured in years

Table 2.5: Explanatory variables

It was hypothesised that a head coach's career length depends on the relative salary, relative wage bills of their teams, coaching experience and win percentage. It was also hypothesised that the probability of a head coach dismissal has decreased over time.

In [62], it was found that Cox proportional hazard models are inadequate when it comes to modelling repeated event data. The authors in [26] suggested that many of the coaches investigated have held multiple jobs in the Bundesliga. To account for this, the conditional risk-set model in gap time, proposed in [78], was adopted. In addition, two parametric models, the exponential and Weibull models, were also considered.

Similar results were found in all four models. A positive effect of relative wage on the hazard was found, suggesting coaches of more expensive teams do not last as long. It was also found that the salary of the head coach, head coach experience and career win percentage had no statistically significant effects on the survival of the head coach. However, the Bosman effect had a positive effect on survival. As such, head coaches working during the more recent Bosman effect period are more likely to survive in the Bundesliga. The Bosman effect is based on a decision by the 1995 European Court of Justice which banned restrictions on foreign European Union (EU) players within national leagues. As a result, players in the EU were allowed to move to another club at the end of their contract, without a transfer fee [49].

In [74], the authors investigated the survival rates of German Bundesliga teams who participated between 1981-1982 and 2009-2010. The dependent variable was the duration of firm survival, measured as the team's number of consecutive years after promotion into the Bundesliga. Firm survival ended when the team was relegated or the observed time frame ended. The Kaplan-Meier and Nelson-Aalen estimators were used to illustrate the probability of staying in the Bundesliga for a specific number of seasons. The illustrations showed that the probability of surviving in the Bundesliga more than two years after being promoted was approximately 55%. This suggested that approximately 45% of teams were relegated back to the second Bundesliga only two years after promotion. The authors then modelled the survival of these teams using various predictors. These are summarised in Table 2.6.

Explanatory variable	Definition
Relative budget	The team's annual budget relative to the average annual budget of all teams in the Bundesliga
Average performance	The annual gap between a team's end of the season points total and the corresponding points total of the first team that has been relegated into the second Bundesliga
Local market size	Measured in terms of thousands of residents located in a club's hometown in the year 2006
Number of pre-exits	The number of exits from 1963-1964 to 1980-1981 in the Bundesliga
Newcomer	Dummy variable equal to one if a club participates in the Bundesliga for the first time; zero otherwise
Club age	The year of the current season minus a club's year of foundation
Share of foreign players	The team's average share of foreign, non-German players
Average team age	The average age of players in the team

Table 2.6: Explanatory variables

The functional form of the baseline hazard was found by fitting the model using the exponential, Weibull, log-normal and log-logistic distributions. The results revealed that the log-logistic distribution described the time-to-failure data for relegation in the Bundesliga most accurately. Doubling of the relative budget was associated with an increase in expected time-to-failure of 7.1 years. This suggested that financial resources are highly important in order for teams to avoid relegation. Teams with a better past performance and a higher club age were also found to stay in the Bundesliga for a longer period of time.

In [18], the authors investigated factors that have an effect on the longevity across all Major League baseball players listed in the Baseball Archive v.5 database [14]. Players were categorised according to whether they were inducted into the Hall of Fame while alive, or whether they represented age matched players who were alive at the time of induction, when induction occurred for their age matched cohort. The dependent variable was defined as post induction survival time. The authors used career length, player position and body mass index (BMI) as predictors and carried out a Cox proportional hazards survival analysis. The authors hypothesised that professional Major League Baseball players who are in the Baseball Hall of Fame have longer longevities than players of the same age who are not in the Hall of Fame. The results found that Hall of Fame players died significantly earlier than their controls (non-Hall of Fame players). These results failed to support the research hypothesis. In addition, BMI was found to have a positive effect on post induction survival. The mean BMI for Hall of Fame players was 25.2, compared to 24.7 for their controls. This difference was statistically significant. However, there were no significant relationships found between either the player position or career length and the survival of that player after entering the Hall of Fame.

In [92], the authors investigated factors for an effect on injury among telemark skiers. The authors utilised a population survey of telemark skiers over two ski seasons, between 1996 and 1998, to determine potential risk factors for injury. Respondents to this survey revealed details on their sex, experience, equipment used, injuries and number of days skied in each season. The authors conducted two survival analyses to investigate the impact of injury risk factors on telemark skiers. The first analysis assessed for differences in 'survival without knee injuries' for skiers who wear leather boots compared to skiers who wear plastic boots. The second investigated whether there were differences among skiers using releasable bindings compared to those who use cable bindings and those who use the more traditional three-pin bindings. The survival time referred to the time taken until a knee injury occurred. The results found 'survival without knee injuries' was significantly longer for those skiers wearing leather boots compared to those wearing plastic boots. In addition, 'survival without knee injuries' was longest for those skiers using releasable bindings and substantially less for those using cable bindings and the more traditional three-pin bindings. However, these differences in binding equipment were statistically insignificant.

In [41], the authors analysed the timing of player substitutions during the 2004-2005 Spanish First Division football games. Of interest was whether the first substitution made by each team occurred more often at half time or in the second half of games. The authors estimated a model for the time from kick off to the first substitution and applied several different specifications of the model. These included exponential, Weibull and Gompertz as proportional hazard forms and log-logistic, log-normal, gamma and inverse Gaussian as accelerated failure time forms. Several predictor variables were considered. These are summarised in Table 2.7.

Explanatory variable	**Definition**
Home	Dummy variable equal to one if the substitution is made by the home team; zero otherwise
Result	Goals scored by the team that makes the substitution, minus the goals scored by the team that does not make the substitution, at the moment of the substitution
Home and result interaction	Interactive variable between home and result
Defensive	Dummy variable equal to one if the substitution is defensive; zero otherwise
Neutral	Dummy variable equal to one if the substitution is neutral; zero otherwise
Offensive	Dummy variable equal to one if the substitution is offensive; zero otherwise
Last four matches points	Number of points achieved in the previous four matches by the team that makes the substitution
Last four matches rival points	Number of points achieved in the previous four matches by the team that does not make the substitution

Table 2.7: Explanatory variables

The inverse Gaussian distribution was found to provide the best fit. The results found the predictor with the strongest statistically significant effect on the timing of substitutions, was the score at the time that substitution was made. Statistical evidence was found to suggest that home teams are more likely to make their first substitution at half time than away teams. The results further suggested that defensive substitutions are generally made later in the match, relative to offensive substitutions.

2.2 Survival Analysis in Cricket

The literature review revealed research that either built the foundations for other pieces of work that focus on the application of survival analysis techniques to cricket, or directly focused on survival analysis techniques in a cricket context themselves.

A batsman's innings can be described as a lifespan. When the batsman goes out to bat, he is 'born' and 'lives' for a certain number of balls before he is dismissed. A dismissal is referred to as a batsman 'death'. This form of data can be represented and analysed using survival functions [60].

In [44], the author suggested that batting data could be represented by the geometric distribution. In [38], the author investigated the distribution of batting scores and suggested that the expected score of a batsman would be his true average accounting for all relevant previous innings. To investigate the true average, the author considered data associated with a particular batsman either over a particular period, in a particular position, or against a particular team. The author defined i as the number of innings, n as the number of not out innings, w as the number of dismissals and r as the number of runs scored. The batsman's traditional average, B and true average, A were given as:

$$B = \frac{r}{w}, A = \frac{r}{i} \tag{2.5}$$

respectively. The authors undertook a theoretical approach to show that conditional on certain assumptions, the number of scoring strokes follows a geometric distribution. The number of balls faced was referred to as b, while s referred to the number of scoring strokes made. The

probability of the batsman's innings ending, out or not out, with each ball faced was given by: $p_w = \frac{i}{b}$. The probability that the batsman makes a scoring stroke with each ball faced given the batsman's innings does not end with that ball was given by: $p_s = \frac{s}{b-i}$. The authors also defined $q_w = 1-p_w$ and $q_s = 1-p_s$. These probabilities were assumed to be constant. The authors also assumed that no scoring strokes are made on balls in which the batsman's innings ends. The random variable, X, represented the number of scoring strokes made by the batsman, while j represented the number of balls faced. The probability that the batsman makes k scoring strokes before being dismissed, $Pr(X = k)$, was derived as:

$$\begin{aligned} Pr(X = k) &= \sum_{j=k}^{\infty} q_w^j p_w \binom{j}{k} p_s^k q_s^{j-k} \\ &= \frac{p_s^k p_w q_w^k}{k!} \sum_{j=k}^{\infty} \frac{j!}{(j-k)!} (q_s q_w)^{j-k} \\ &= \frac{p_s^k p_w q_w^k}{(1-q_s q_w)^{k+1}} \\ &= \frac{p_w}{1-q_s q_w} \left(\frac{p_s q_w}{1-q_s q_w} \right)^k \\ &= PQ^k, \end{aligned}$$

where $p = \frac{p_w}{1-q_s q_w}$ and $Q = \frac{p_s p_w}{1-q_s q_w} = 1-p$. Therefore the number of scoring strokes made follows a geometric distribution. Although the findings in [38] have been supported by several studies (i.e. [32], [23]), the results from several studies suggest that the geometric distribution is inadequate at representing batting scores in cricket. As such, these studies contradict the findings in [38].

In [65], the authors adopted survival analysis techniques to investigate the properties of the traditional batting average. The authors considered a sample of batting scores assumed to follow a geometric distribution. If the batting scores are independent geometric random variables, the probability mass function (PMF) for each score was defined as:

$$p_0(x) = \theta(1-\theta)^x, \tag{2.6}$$

where $0 < \theta < 1$ is an unknown parameter. A Pearson chi-squared goodness of fit test found

that the geometric distribution resulted in poor fit. This resulted in an inconsistent batting average metric. As such, an alternative batting average, independent of the geometric assumption, was suggested. M and $M*$ represented the highest completed innings and highest not out score made by a batsman respectively. Assuming $M > M*$, the alternative batting average was defined as: $A = T$, where

$$T = \sum_{x=0}^{M} x(\hat{F}(x) - \hat{F}(x+1)), \tag{2.7}$$

where $\hat{F}(x)$ is the Kaplan-Meier estimator of the survival function $F(x)$ defined by: $F(x) = \sum_{y=x}^{\infty} p(y)$, also known as the Product Limit Estimator ([63], [67], [57]). As the Product Limit Estimator does not depend on parametric assumptions, the alternative batting average is a non-parametric estimator of the mean. Suppose that $M \leq M*$, there are k not out scores at least as big as M, their sum is S and the probability mass unassigned by the Product Limit Estimator is R. The alternative batting average in this situation was defined as:

$$A = \frac{T + \frac{RS}{k}}{1 - R}. \tag{2.8}$$

The authors concluded that a future area of research could be to investigate how additional factors such as scoring rate, opposition bowling strength and pitch state, could be combined with runs scored, to illustrate the qualities of particularly strong batsmen.

In [60], the authors also applied Kaplan-Meier estimation techniques to show how batting careers can be illustrated using survival functions. The batsman's innings was described as a lifespan with a 'death' referred to as a dismissal. Observations during which players were not dismissed were referred to as censored observations. This methodology was used to illustrate the completed career performance of different batsmen. Data were obtained from the 2016 Cricinfo and CricketArchive databases. Statistics for test cricketers, SR Waugh and SR Tendulkar, were used in combination with Kaplan-Meier estimation techniques to illustrate the distinct survival probability curves. These curves are shown in Figure 2.1. The curves show that SR Waugh's survival probabilities were similar than SR Tendulkar's for approximately the first 10 runs scored. However, as the number of runs scored increased beyond 10, SR Tendulkar became more effective, as illustrated by SR Tendulkar's higher survival probabilities.

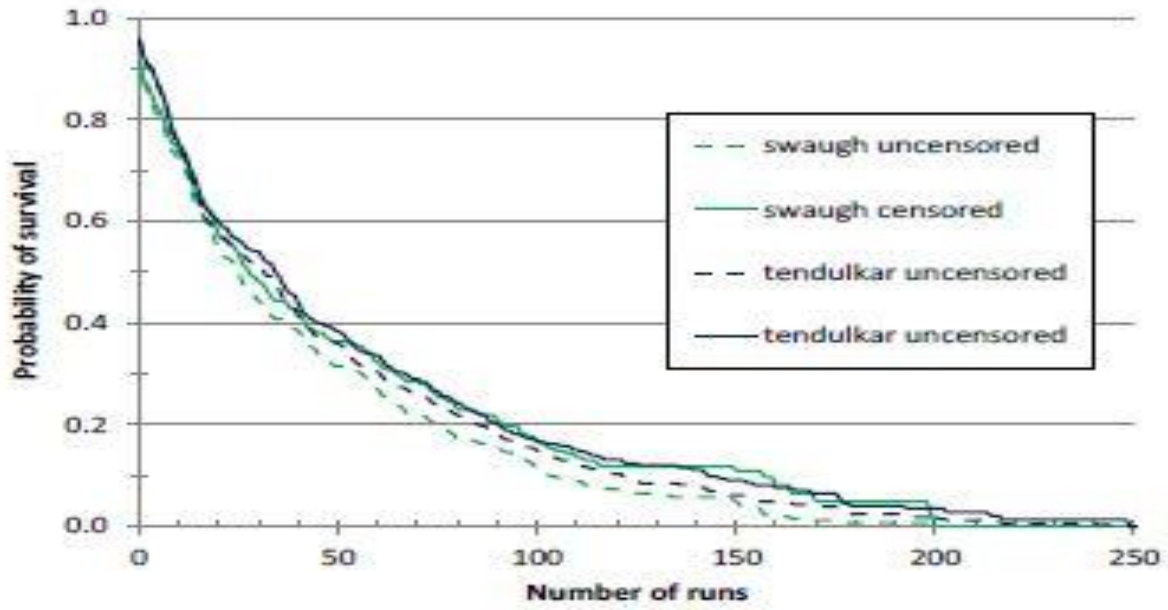

Figure 2.1: Product Limit Estimator survival function for SW Waugh and SR Tendulkar (Figure obtained from [60])

The Greenwood formula, derived in [51], was applied to generate confidence intervals for the survival curves for a number of different batsmen. Australian twins, SR Waugh and ME Waugh, were compared primarily because of the large differences between their batting averages. These curves are shown in Figure 2.2. The curves show that ME Waugh performed as reliably as SR Waugh up until approximately 50 runs had been scored, but then gradually performed worse than SR Waugh as the number of runs scored increased beyond 50.

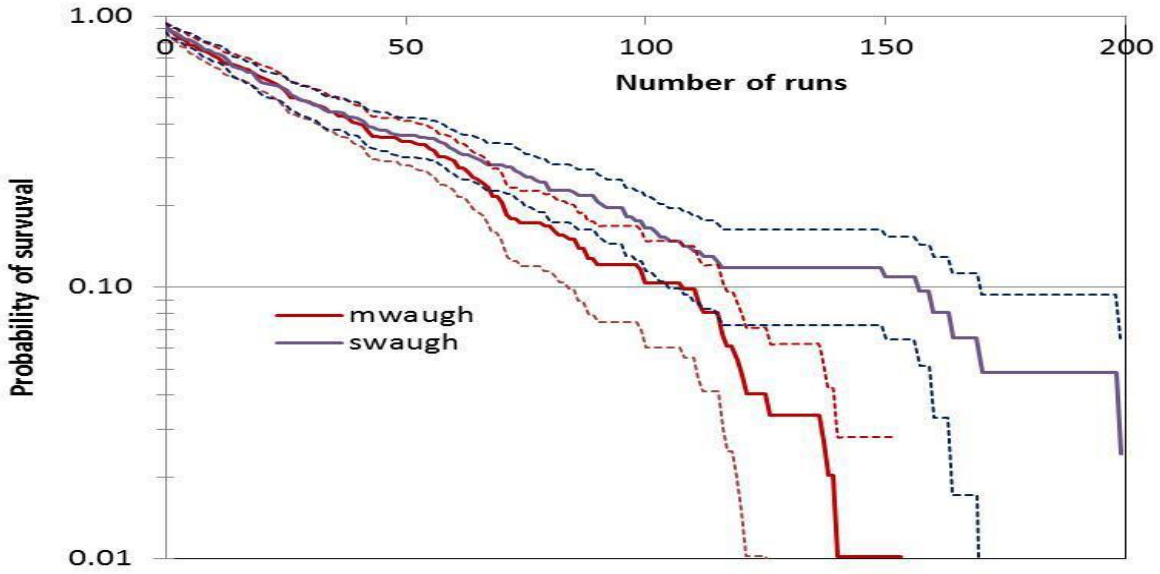

Figure 2.2: Estimated survival functions for SR Waugh and ME Waugh. 95% confidence limits are shown as dotted lines (Figure obtained from [60])

An unweighted log-rank test was used to test for statistical differences in survival between SR Waugh and SR Tendulkar. The results found no statistically significant differences between the two when considering the whole curve. However, when restricting focus to those parts of the curves associated with run totals beyond 100, strong statistical differences were observed. This was extended to construct survival curves for a large number of batsmen split by batting position and innings. These curves are shown in Figure 2.3 and suggest that batting during innings three and four was more difficult than batting during innings one and two.

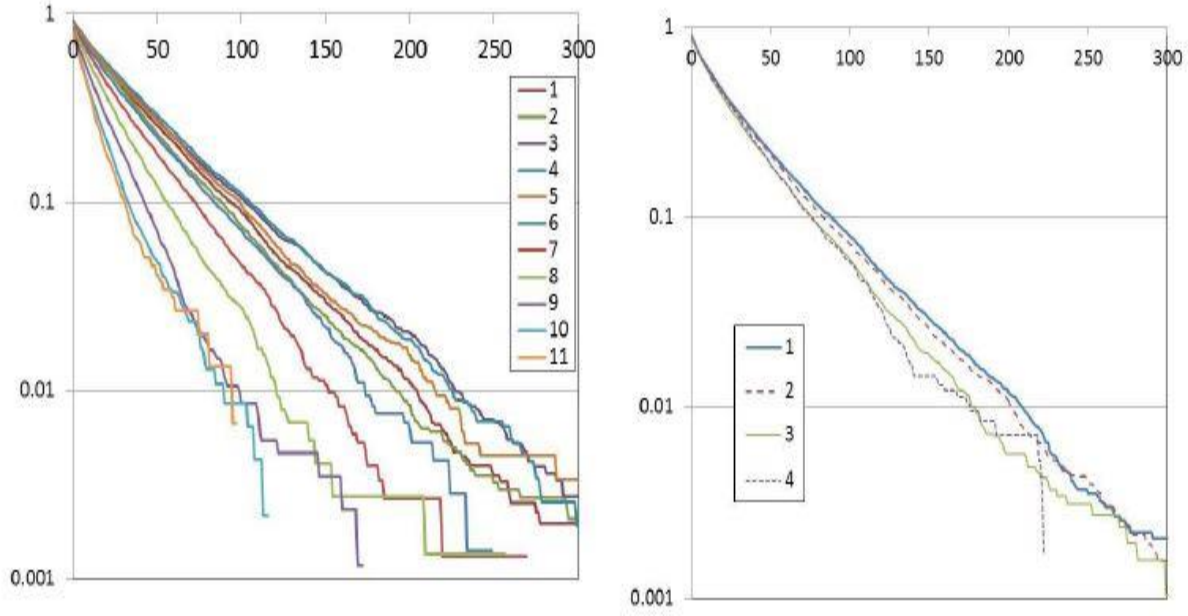

Figure 2.3: PLE survival function for all batsmen in Test matches split by batting position (left) and innings number (right) (Figure obtained from [60])

In [40], the authors challenged the suitability of the Kaplan-Meier approach in estimating the true batting average, and propose an alternative to the traditional batting average estimate. A parametric maximum likelihood approach was adopted and a class of distribution called the Generalised Geometric Distribution (GGD) was built. The authors used several pieces of notation. $X_1, \ldots, X_n$ represented the scores from n innings when a batsman was out and $Y_1, \ldots, Y_m$ represented the scores from m innings when a batsman was not out. K represented the highest score. For $0 < i < K$, f_i referred to the number of times the batsman scored i runs and was dismissed at the score i and f_i^* referred to the number of times the batsman scored i and remained not out at the end of the innings. F_i and $F_{(i)}^*$ denoted the cumulative frequencies of f_i and f_i^*, respectively. K was formally defined as $K = \max(\max_{1\leq i\leq n} X_i, \max_{1\leq j\leq m} Y_j)$. M_i represented the total number of scores from out innings bigger than or equal to i, while N_i represented the total number of scores from out or not out innings bigger than i. Mathematically, this means

$$\begin{aligned} M_i &= \sum_{j=i}^{k} (f_j + f_j^*) \\ &= (n - F_{i-1}) + (m - F_{i-1}^*), \end{aligned}$$

and

$$\begin{aligned} N_i &= M_{i+1} \\ &= (n - F_i) + (m - F_i^*). \end{aligned}$$

Accounting for batting scores during innings when the batsman was dismissed and innings in which the batsman was not dismissed, the authors considered the following form of the Kaplan-Meier estimator of the survival function at time t:

$$\hat{S}_{km}(i) = \prod_{j=0}^{i} \left(1 - \frac{f_j}{M_j}\right), j = 0, 1, 2, \ldots, \tag{2.9}$$

and

$$\hat{S}_{km}(t) = \hat{S}_{km}([t]), \forall t. \tag{2.10}$$

The Kaplan-Meier estimator of the batting average was given by:

$$M_{KM} = \sum_{i=0}^{K} \hat{S}_{KM}(i) = \sum_{i=0}^{K} \prod_{j=0}^{i} \left(1 - \frac{f_j}{M_j}\right). \tag{2.11}$$

While the Kaplan-Meier estimator improves the traditional estimate by accounting for not out innings, there are still limitations when applied in a cricket score context. As the Kaplan-Meier estimator is non-parametric, the survival function is based only on scores where the batsman has been dismissed in the past. As such, the Kaplan-Meier estimator generates a zero probability of dismissal on all occasions when the player has never been dismissed at that particular score. In addition, the authors proposed that the summation in Equation (2.11) should technically go to ∞ but values of $i > K$ are ignored. As such, on occasions when $\max_{1 \leq i \leq n} X_i \leq \max_{1 \leq j \leq m} Y_j$, $f_k \neq N_k$, implying $\hat{S}_{km}(K) \neq 0$, and consequently $\hat{S}_{km}(t) > 0 \forall t$. This means that the Kaplan-Meier survival function may be infinite if the highest score is from a not out innings.

The authors suggested that the geometric distribution provides an unrealistic representation of a batsman's score since it assumes a batsman is equally likely to get out on every ball he faces. The GGD was proposed as an alternative. The GGD is parameterised by a sequence of hazard

rates at score values:

$$\alpha_i = P[X = i | X \geq i] \text{ for } i = 0, 1, 2, \ldots,$$

The GGD has the density $\phi_i = P(X = i)$ given by:

$$\phi(0) = \alpha_0, \phi(i) = \alpha_i \times \prod_{j=0}^{i-1}(1 - \alpha_j), i = 1, 2, \ldots$$

The Kaplan-Meier survival function for the GGD was given by:

$$S(i) = \prod_{j=0}^{i}(1 - \alpha_j), i = 0, 1, \ldots$$

In [90], the authors developed Bayesian survival analysis methodology to predict the test match batting abilities for international cricketers. A Bayesian survival model was proposed to infer a batsman's hazard function from their career batting record. Let $X \in \{0, 1, 2, 3, \ldots\}$ represent the score a batsman scores in a particular innings. The authors defined the hazard function in a cricket context as:

$$H(x) = \frac{P(X = x)}{P(X \geq x)}.$$

This represents the probability that the batsman scores $x(P(X = x))$, given they are currently on score x. The probability distribution for a set of conditionally independent scores $\{x_i\}_{i-1}^{I-N}$ and not out scores $\{y_i\}_{i-1}^{N}$ was defined as:

$$p(\{x\}, \{y\}) = \prod_{i=1}^{I-N}\left(H(x_i)\prod_{a=0}^{x_i-1}[1 - H(a)]\right) \times \prod_{i=1}^{N}\left(\prod_{a=0}^{y_i-1}[1 - H(a)]\right).$$

This gives the likelihood for any proposed model of $H(x; \theta)$. The log-likelihood was given as:

$$\log(L(\theta)) = \sum_{i=1}^{I-N}\log(H(x_i)) + \sum_{i=1}^{I-N}\sum_{a=0}^{x_i-1}\log[1 - H(a)] + \sum_{i=1}^{N}\sum_{a=0}^{y_i-1}\log[1 - H(a)],$$

where θ is the set of parameters controlling for the form of $H(x)$. Given a batsman's batting

average is widely considered to represent a batsman's ability in cricket, the hazard function was parameterised as:

$$H(x) = \frac{1}{\mu(x) + 1},$$

where $\mu(x)$ represents a player's effective batting average. In addition, a batsman is assumed to have an initial playing ability, μ_1, that increases up to a maximum, μ_2, as the number of runs scored increases. The authors used an exponential model to represent the transition from μ_1 to μ_2 as follows:

$$\mu(x; \mu_1, \mu_2, L) = \mu_2 + (\mu_1 - \mu_2) \exp\left(-\frac{x}{L}\right),$$

where L represents the number of runs required for 63% of the transition between the two batting averages to take place. The authors assumed that $\mu_1 \leq \mu_2$ and restricted the value of L to be less than μ_2. As such, they modified the set of parameters, (μ_1, μ_2, L), to be (C, μ_2, D), where $\mu_1 = C\mu_2$, $L = D\mu_2$ and C and D were restricted to the interval $[0, 1]$. Therefore, the hazard function was reparameterised as:

$$H(x) = \frac{1}{\mu_1 - \mu_2(C-1)\exp\left(-\frac{x}{L}\right) + 1}.$$

The proposed model was initially applied to individual players. Individual player data for long test career players during the 1990's and 2000's were analysed using fixed priors for the parameters, C, μ_2 and D, of each player. These players consisted of retired batsmen, all-rounders and a bowler, the same as those used in [33]. The full Bayesian model specification for analysing an individual player was given as:

$$\mu_2 \sim \text{Lognormal}(25, 0.75^2)$$

$$C \sim \text{Beta}(1, 2)$$

$$D \sim \text{Beta}(1, 5)$$

$$\text{log-likelihood} \sim \sum_{i=1}^{I-N} \log(H(x_i)) + \sum_{i=1}^{I-N} \sum_{a=0}^{x_i-1} \log[1 - H(a)] + \sum_{i=1}^{N} \sum_{a=0}^{y_i-1} \log[1 - H(a)].$$

The joint posterior distribution for μ_2, C and D is proportional to the prior times the likelihood function. The authors then sampled from the joint posterior distribution to make inferences about a player's initial and peak batting average and the transition time between the two. The results found that BC Lara and SR Waugh, the batsmen with the highest test career averages, had the highest estimated peak batting average. However, CL Cairns and SM Pollock were found to have the highest initial batting ability, but had lower test career batting averages. The authors generalised their inference to a wider group of players using a hierarchical model structure. The authors introduced hyperparameters, v and σ. v represents a value of μ_2 that the players are clustered around, while σ describes how much μ_2 varies from player to player. The full hierarchical model specification for analysing a group of players was given as:

$$v \sim \text{Uniform}(1, 100)$$

$$\sigma \sim \text{Uniform}(0, 10)$$

$$u_{2,i}|v, \sigma \sim \text{Lognormal}(v, \sigma^2)$$

$$C_i \sim \text{Beta}(1, 2)$$

$$D_i \sim \text{Beta}(1, 5)$$

$$\text{log-likelihood} \sim \sum_i \left(\sum_{i=1}^{I-N} \log(H(x_i)) + \sum_{i=1}^{I-N} \sum_{a=0}^{x_i-1} \log[1 - H(a)] + \sum_{i=1}^{N} \sum_{a=0}^{y_i-1} \log[1 - H(a)] \right).$$

The marginal posterior distribution for the hyperparameters, given all the data, was written in

terms of the expectations over the individual players' posterior distributions:

$$p(v, \sigma|\{d_i\}) \propto p(v, \sigma) \prod_{i=1}^{N} \mathrm{E}\left[\frac{f(\mu_{2,i}|v, \sigma)}{\pi(\mu_{2,i})}\right], \tag{2.12}$$

where $f(\mu_{2,i}|v, \sigma)$ is the Lognormal(v, σ^2) prior applied to μ_2 for the ith player and $\pi(\mu_{2,i})$ is the Lognormal$(25, 0.75^2)$ prior used calculate the posterior of each player. Using Equation (2.12), the posterior samples for each sample were combined to make posterior inferences about hyperparameters, v and σ. The abilities of the next opening batsman to debut for New Zealand were predicted, with μ_1 estimated to be 9.6, μ_2 estimated to be 27.7 and L estimated to be 3.1. In addition, MH Richardson was quantified as New Zealand's best performing test opener since he made his debut in 2001, one proposition that is widely agreed on in cricket.

Early work from this thesis was peer-reviewed and published in Mathsport 13 Conference proceedings. In that work [36], a predictive model capable of calculating the ball-by-ball probability of an opening batsman being dismissed in the first innings of a limited overs cricket game was formulated. A large number of Cox proportional hazard models were implemented with each model fit to a different combination of nine predictors, some of which are modifications to those used in previous work involving cricket analytics (i.e. [77], [30], [32]). These are defined in Table 2.8.

Batting performance metric	Definition
Team total balls	Cumulative number of balls faced by team
Total runs	Cumulative number of runs scored by batsman
Team total runs	Cumulative number of runs scored by team
Strike rate	Batsman strike rate
Dot balls	Cumulative number of dot balls faced by batsman
Consecutive dot balls	Cumulative number of consecutive dot balls faced by batsman
Less than 2 in 4	Cumulative number of balls faced in which less than 2 runs in 4 balls had been scored
Resources	Proportion of resources available to batsman
Pressure	Pressure felt by batsman

Table 2.8: Previous batting performance metrics.

A pragmatic model selection methodology was utilised to find an appropriate set of candidate models. Firstly, the estimated Cox model coefficients were required to make *practical* sense according to whether they were expected to increase or decrease the probability of survival as the number of balls faced increased. The predictors were also required to be *statistically significant*. The last criteria required that the probability of batsman survival either remained constant or decreased at the ball-by-ball level. However, no justification was given for this criteria. The final model selected included: the cumulative number of runs scored, cumulative number of dot balls scored and cumulative number of balls faced in which less than two runs in four balls had been scored. The ball-by-ball survival probabilities for all 43 opening batsmen considered were calculated using:

$$\log\left(\frac{p}{1-p}\right) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n, \tag{2.13}$$

where p represented the probability of survival and $\beta_1, \beta_2, \ldots, \beta_n$ represented the weights for each attribute, $x_1, x_2, \ldots, x_n$, respectively. The results found differences in the roles of MJ

Guptill and BB McCullum and in the roles of DA Warner and AJ Finch. These survival probabilities were used to calculate the average AUC for each batsman. Strong relationships between the average AUC and average ICC rankings for groups of batsmen were found. The correlation between the square root of the average AUC and the square root of the average rankings was -0.87. This suggested higher ranked opening batsmen are more likely to remain in bat than lower ranked batsmen. The correlation between the square root of the average AUC and the square root of the average runs scored was 0.91. This suggested that crease occupation and run scoring are synonymous in successful batsmen.

2.3 Literature Review Findings

In this chapter, the application of survival analysis techniques to cricket has been reviewed, with apparent disagreements discovered between studies over the adequacy of their respective findings. However, they all appear to be limited in some shape or form. In [65], the authors did not consider any potential covariates in the use of Kaplan-Meier estimation techniques to investigate batsman survival rates. This research extends on the work in [65], by introducing the effect of potential covariates on the historical performance of batsmen. The research in [37] also considered aspects of the work in [65], by exploring the effect of a number of factors associated with optimal batting strategy. These included the required run rate and number of wickets lost. However, these considerations failed to account for within-game events. This research extends on the work in [37], firstly by incorporating additional covariates that were not previously considered and secondly, by specifically focusing on within-game events. These include: dot ball and consecutive dot ball effects, boundary effects and team contribution effects. In [60] and [90], the authors also investigated the historical performance of batsmen. However, those pieces of work are further examples of published work that do not focus on within-game events.

Early work from this thesis was peer-reviewed and published in Mathsport 13 Conference proceedings. In that work [36], survival analysis methodology was successfully applied to investigate the effect of within-game events on the probability of an opening batsman being dismissed in the first innings of a limited overs cricket game. That work extends previous research investigating the historical performance of batsmen to assess within-game events. However, the

research focused only on opening batsmen.

Prior to the work in this thesis, there did not appear to be any published studies that focus on analysing the effect of within-game events on individual batsmen at particular batting positions, or batting partnerships at particular wickets in cricket. Further, there did not appear to be any previous work investigating the impact that optimal batting partnership strategy has on winning. As such, this research attempts to address this gap in the literature. This work, in [35], was peer-reviewed and published in Mathsport International 2017 Conference proceedings. Here, survival analysis methodology was successfully applied to investigate the effect of within-game events on the ball-by-ball survival probabilities of different order batsmen and batting partnerships in limited overs cricket games. This built the foundations to complete the main objective of this research. The objective was to optimise batting partnership strategy in limited overs cricket, in an attempt to increase a team's scoring rate and chances of winning.

Chapter 3

Research Objectives and Methodology

The literature review revealed the extensive amount of published research surrounding survival analysis methodology application, across various sporting disciplines. There have been several pieces of published work that have applied survival analysis methodology to batting performance in cricket. However, this work has focused on individual batsmen across their careers, without particular focus on within-game events. Moreover, prior to the work in this thesis, there did not appear to be any published studies that focus on analysing batting partnerships in cricket. This is likely to stem from the historical difficulty in obtaining ball-by-ball data. More recently, websites like www.espncricinfo.com have enabled much greater access to data. The scarcity of literature surrounding survival analysis applications to the performance of batting partnerships with focus on within-game events in limited overs cricket, has resulted in a gap in the literature. In addition, the growing popularity of in-game betting within the sport ([58], [8]) highlights the potential demand for this research. The real world impact of this research is that it helps teams to increase the number of runs scored at an optimal rate, therefore increasing chances of winning. Given this gap in the literature, research objectives were established. These are discussed in Section 3.1.

3.1 Research Objectives

The primary objective of this research was to optimise batting partnership strategy by formulating several predictive models to calculate the probability of a batting partnership being dismissed in the first innings of a limited overs cricket match. The narrowed focus also reduced confounding factors, such as match state.

The model structures were expected to reveal strategies for optimally setting a total score for the opposition to chase. In the first innings of a limited overs cricket match, there is little information available at the commencement and during the innings to guide the team in accumulating

a winning total score.

The secondary objective of this research was to validate the final models to ensure they were appropriately estimating the ball-by-ball survival probabilities of each batsman, in order to determine the most *effective* partnership combinations. The research hypothesised that the more *effective* a batting partnership is at occupying the crease, the more runs they will score at an appropriate rate and the more likely the team is to win the match, by setting a defendable total.

Specifically, the purpose of this research was to address the following two key questions:

1. What are the in-game strategies for optimising the runs scored in the first innings?

2. What are the practical applications of this knowledge?

3.2 Research Methodology

Given the research objectives, the following research methodology was applied:

3.2.1 Identify and calculate performance metrics

Metrics that significantly affect the ball-by-ball survival probabilities of individual batsmen and batting partnerships were identified. Potential performance predictors were identified leveraging expert opinion from current and former international first class players and coaches. These were then tested for an effect on the probability of a batsman or partnership dismissal using binomial logistic regression analyses. The dependent variable, Y, in these analyses was categorised as:

$$Y = \begin{cases} 1, & \text{if batsman/partnership is dismissed} \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

3.2.2 Apply statistical techniques to batting performance in cricket

Cox proportional hazard models, ridge regression techniques and censoring techniques were applied to determine the effect of in-game events on the survival probabilities of individual batsmen and batting partnerships. Observations where a batsman or partnership was not dismissed were referred to as censored observations. These were taken into account in the modelling procedure using existing functionality as part of the 'survival' package [12] in R [80].

3.2.3 Calculate the ball-by-ball survival probabilities of batsmen and batting partnerships

Cox models for different order batsmen and partnerships at different wickets were formulated. This was carried out by splitting the data into multiple subsets. Individual batsmen were split according to their batting position while batting partnerships were split according to the wicket when the partnership was played. Using each model, Equation (2.13) was implemented to calculate the survival probabilities.

3.2.4 Identify and calculate overall performance and *effectiveness* metrics

The survival probabilities were cumulated to give a total AUC for each batsman and each partnership. The total number of runs scored, proportion of team runs scored and proportion of games won were calculated for each batsman and each partnership. The ODI ICC batting rank for each batsman was also obtained (www.icc-cricket.com/player-rankings/overview). The ranking used was that following the last international in which they batted, on or prior to 14th February 2016. These metrics were assessed for correlations and used as a measure of performance *effectiveness* to complete the objective of optimising batting partnership strategy.

3.2.5 Identify and incorporate model validation methodology

Each final model associated with a subset of ODI data was initially fitted to the corresponding IPL data. For example, the final model associated with third wicket ODI partnerships was

applied to data associated with third wicket IPL partnerships. Survival probabilities for IPL batsmen and partnerships were generated and plotted against survival probabilities for ODI batsmen and partnerships, respectively. The AUC was calculated as a metric to determine which models generated survival probabilities characterising the largest difference between IPL and ODI batsmen and IPL and ODI partnerships.

3.3 Previous Research

This research adopted a Cox proportional hazard modelling approach and censoring methodology. Several pieces of previous research were identified in which such methodology had been applied to individual batting performances ([65], [37], [60], [36], [90]). However, the research methodology outlined in these pieces of work suffered several issues. The following research weaknesses were identified:

1. **Lack of performance metrics as predictors**
   In the research in [65] and [60], no performance metrics were assessed for a possible effect on batsman performance.

2. **Lack of focus on within-game events**
   In the research in [65], [37], [60] and [90], within-game events such as dot ball effects, consecutive dot ball effects, boundary effects and team contribution effects were not taken into account. The work only considered the career statistics from batsmen across games as a whole.

3. **Lack of analysis investigating different order individual batsmen**
   In the research in [65], [37], [60] and [90], no distinction was made between different order batsmen. In [36], the authors applied Cox proportional hazard models to the survival rates of opening batsmen. However, the work did not address the application of survival analysis techniques into different order, non-opening batsmen.

4. **Lack of analysis investigating batting partnerships**
   Prior to the work in this thesis, there did not appear to be any published work which addresses applications of survival analysis techniques into batting partnerships in cricket.

5. **Lack of final model validation**

   Previous published work focussing on survival analysis applications to cricket has failed to include any model validation techniques.

This research has developed an original and rigorous quantitative framework to optimise batting partnership strategy in limited overs cricket. This approach used survival analysis to investigate the survival properties of batsmen and partnerships, while addressing these issues. Early work on this approach, in [36], was peer-reviewed and published in Mathsport 13 Conference proceedings. More detailed analysis, in [35], presented throughout Chapter 8, was peer-reviewed and published in Mathsport International 2017 Conference proceedings.

Chapter 4

Data Extraction and Processing

The analysis conducted throughout this research required ball-by-ball data for One-Day International (ODI) cricket matches contested between 26th December 2013 and 14th February 2016 and Indian Premier League (IPL) matches contested between 9th April and 29th May 2016. Ball-by-ball data provides information on what happened during each ball of a match. This data is accessible from the ESPN Cricinfo website (www.espncricinfo.com). An automated process using the SAS language was applied to extract data from the associated commentary log for each match. This was carried out on a ball-by-ball basis, before this data were converted into a scorecard format and stored in a tabular form, as shown in Appendix C. The SAS code used to carry out this extraction and conversion is presented in Appendix F. Data collected was based on within-game events. Table 4.1 illustrates the contents of a ball-by-ball scorecard.

Player info	**Game info**	**Other descriptors and metrics**
Bowling	Game	Over
Facing	Cricinfo ID	Ball
Batting position	Home	Description
Bowling position	Away	Out (Y)
	Venue	Runs scored
	Dates	Sundry type
	Year	
	Innings	

Table 4.1: Ball-by-ball data elements.

4.1 Data Manipulation

For the ODI data, variables that could potentially have an effect on the prediction of the probability of a batsman or batting partnership dismissal and could be obtained from the original elements of the ball-by-ball data were calculated. These metrics are summarised in Table 4.2, explained in Appendix B and illustrated in Appendix D and Appendix E.

Batsman metrics	**Partnership metrics**
Batsman balls	Partnership balls
Batsman runs	Partnership runs
Batsman dot balls	Partnership dot balls
Batsman consecutive dot balls	Partnership consecutive dot balls
Batsman less than 2 in 4	Partnership less than 2 in 4
Batsman boundaries	Partnership boundaries
Batsman contribution	Partnership contribution
Batsman percentage boundaries	Partnership percentage boundaries
Batsman percentage dot balls	Partnership percentage dot balls
	Wicket

Table 4.2: Performance metrics.

The batting order is important in cricket in order to maximise the score of each batsman and the team. It was noted in [91, p.1939] that "in cricket, there is a long-standing tradition of general strategy that places better batsmen near the beginning of the batting order and the weaker batsmen near the end". As such, the data were then split into multiple subsets, as shown in Table 4.3. Each set consisted of data associated with the first innings of games.

Batsman type	**Batting positions**
Openers	1 and 2
Top order	3 and 4
Middle order	5,6 and 7
Lower order	8 and 9
Tail	10 and 11

Table 4.3: Categorised batsmen.

The data were then split into multiple further subsets. Each subset consisted of data associated with first innings of games. These datasets were created with each separate set consisting of data associated with one wicket. This resulted in ten further datasets, one associated with each wicket.

4.2 Data Limitations

Several data limitations were identified. Firstly, entries associated with incorrect recording of the presence of sundries (extras), type of extra and particular ball of the over were identified. These were manually corrected for. Additionally, there are cricket teams that have players with the same names. The extracted data did not distinguish between different batsmen with the same names. These were manually consolidated and distinguished from one another.

As part of this research, primary emphasis was put on batting partnerships. As the extracted data did not include the partner of the batsman facing each ball, identification and recording of this was incorporated into data consolidation. This was identified by writing a function in R [80] to identify instances when the batsman facing the bowler changed, while the innings and game remained the same. For each instance, the new batsman was imputed as the partner of the previous batsman.

Chapter 5

Exploratory Data Analysis

In this chapter, the characteristics of the extracted data are explored and the assumptions of a logistic regression model fit to these data are assessed. The data outlined in Chapter 4 are investigated for evidence of outliers, multicollinearity and relationships between the predictor variables. A logistic regression analysis was selected, as the dependent variable, whether a batsman or partnership was dismissed, was binary and the probability of a batsman or partnership dismissal was of particular interest. The R code used to investigate the data and carry out the analysis is presented in Section G.1 in Appendix G.

5.1 Outliers

A Cook's distance test was carried out and revealed outlier observations that would have a significant effect on the estimate of the model coefficients in a logistic regression analysis. This would result in unreliable results and conclusions.

5.2 Multicollinearity and Interrelationships

The 'car' package [5] in R [80] was used to produce variance inflation factors (VIFs). These were used to assess for presence of multicollinearity, which will be addressed in Section 5.2.1. The 'asbio' package [1] in R was used to produce correlation matrices. These were used to assess for the presence of statistical relationships between the metrics, which will be addressed in Section 5.2.2.

5.2.1 Variance Inflation Factors

Two binomial logistic regression analyses were carried out. The objective of the first was to determine a selection of metrics that are practically and statistically significant contributors to

the probability of a batsman dismissal. The objective of the second was to determine a selection of metrics that are practically and statistically significant contributors to the probability of a batting partnership dismissal. Using the 'car' package in R, VIFs were fitted to each logistic regression model. In [75], the authors emphasised the VIF rule appearing in many publications: VIFs greater than 10 are a sign of strong multicollinearity.

In the first logistic regression analysis, it was found that batsman balls, batsman runs, batsman dot balls, batsman consecutive dot balls, batsman less than 2 in 4 and batsman boundaries all showed strong evidence of serious multicollinearity with VIFs greater than 10. In the second logistic regression analysis, it was found that partnership balls, partnership runs, partnership dot balls, partnership consecutive dot balls, partnership less than 2 in 4 and partnership boundaries all showed strong evidence of serious multicollinearity, with VIFs greater than 10.

5.2.2 Scatter Plot and Correlation Matrix

A scatter plot and correlation matrix for the batting metrics illustrated strong positive relationships between batsman balls, batsman runs, batsman dot balls, batsman consecutive dot balls, batsman less than 2 in 4 and batsman boundaries.

The relationships between these metrics generated correlation values, r, ≥ 0.90. Table 5.1 summarises the correlation between these metrics.

Pair of variables	Correlation
Batsman balls and batsman runs	0.930
Batsman balls and batsman dot balls	0.952
Batsman runs and batsman boundaries	0.927
Batsman dot balls and batsman consecutive dot balls	0.951
Batsman dot balls and batsman less than 2 in 4	0.901
Batsman consecutive dot balls and batsman less than 2 in 4	0.956

Table 5.1: Metric correlations.

A scatter plot and correlation matrix for the batting partnership metrics illustrated strong positive relationships between partnership balls, partnership runs, partnership dot balls, partnership consecutive dot balls, partnership less than 2 in 4 and partnership boundaries.

The relationships between these metrics generated correlation values, r, ≥ 0.90. Table 5.2 summarises the correlation between these metrics.

Pair of variables	Correlation
Partnership balls and partnership runs	0.945
Partnership balls and partnership dot balls	0.956
Partnership runs and partnership boundaries	0.935
Partnership dot balls and partnership consecutive dot balls	0.958
Partnership dot balls and partnership less than 2 in 4	0.918
Partnership consecutive dot balls and partnership less than 2 in 4	0.966

Table 5.2: Partnership metric correlations.

5.3 Logistic Regression Assumptions

The validity of the binomial logistic regression analysis was tested by examining whether the following assumptions of a logistic regression analysis held:

1. The residuals are independent
2. The residuals are not affected by outliers
3. The relationship between the independent variables and the log odds is linear

In a linear regression analysis, several assumptions hold in addition to (1) and (2) in a logistic regression analysis. The first is that the residuals are normally distributed. The second is that the residuals exhibit constant variance. These are not assumed in a logistic regression analysis [56].

5.3.1 Independence of Residuals

The first assumption was tested by carrying out a Durbin-Watson test, using built in functionality in the 'car' package in R. The Durbin-Watson test assesses for any evidence of autocorrelation among the residuals from a general linear model. The hypotheses tested are $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$, where ρ refers to the autocorrelation. In [46, p.874.], the authors claimed that "a test statistic greater than 2 indicates a negative correlation between adjacent residuals, whereas a value below 2 indicates a positive correlation".

In [48], the authors suggested that there are three conventional levels of statistical significance, 1%, 5% and 10%, commonly used to assess statistical models. The Durbin-Watson test for the first logistic regression analysis, produced a p-value = 0.089. This was statistically significant at the 10% level. This provided evidence against the null hypothesis that no correlation exists among the residuals. The Durbin-Watson test statistic for the logistic regression model was 1.82, indicating that the residuals were positively correlated. This suggested that the residuals may have been linearly dependent.

The Durbin-Watson test for the second logistic regression analysis, produced a significant p-

value at the 10% level (p-value = 0.091). This provided evidence against the null hypothesis that no correlation exists among the residuals. The Durbin-Watson test statistic for the logistic regression model was 1.82, indicating that the residuals were positively correlated. This suggested that the residuals may have been linearly dependent.

5.3.2 Residual Outliers

The second assumption was tested by carrying out a Bonferroni Outlier Test, using existing functionality in the 'car' package in R. This test reports the "Bonferroni p-value for studentised residuals in linear and generalised linear models, based on a t-test for linear models and normal-distribution test for generalized linear models" [3]. In [59], the authors explained that the Bonferroni correction measure tests each of the residuals from a regression model to determine whether or not it is an outlier. The hypotheses tested were H_0 : None of the residuals are outliers vs H_1 : At least one of the residuals is an outlier.

Applying the test to the logistic regression analyses, produced a Bonferroni p-value = 0.001 for the first analysis, and p-value = 0.002 for the second. These were both statistically significant at the 5% level. This provided evidence to reject the null hypothesis and accept that one or more of the residuals from each logistic regression model may have been outliers.

5.3.3 Linearity

The third assumption was tested using the Box-Tidwell approach suggested in [88]. This involves introducing new variables defined as the natural logs of each continuous predictor. The interaction between each new variable and the original variable associated with each predictor is included in the logistic regression model. If the interaction term returns significance, there is evidence to suggest that the relationship between the associated predictor and the log odds is non-linear, violating the assumption.

Applying the test to the first logistic regression analysis, revealed statistically significant interactions between the predictor and the natural log of the predictor for batsman dot balls, batsman consecutive dot balls, batsman contribution, batsman percentage boundaries and batsman per-

centage dot balls. This suggested a non-linear relationship between the log odds and each one of these predictors.

Applying the test to the second logistic regression analysis, revealed statistically significant interactions between the predictor and the natural log of the predictor for partnership balls, partnership runs, partnership consecutive dot balls, partnership less than 2 in 4, partnership contribution, partnership percentage boundaries and partnership percentage dot balls. This suggested a non-linear relationship between the log odds and each one of these predictors.

5.4 Conclusions

The diagnostics from the logistic regression analysis to investigate the probability of a batsman dismissal and the logistic regression analysis to investigate the probability of a batting partnership dismissal suggest that such analyses may result in invalid conclusions. There is evidence against the independent residuals, outlier residuals and linearity assumptions. In addition, the multicollinearity between the metrics of interest suggests that the logistic regression results may be unreliable and inaccurate. Ridge regression is a technique that accounts for multicollinearity in a multiple regression analysis and will be addressed in Chapter 6.

Logistic regression analysis is concerned with the association between several covariates and whether or not an event occurs. As this research is concerned with the association between several covariates and the time to an event, logistic regression may not be the most appropriate technique for analysis. In this research, the particular event of interest is the occurrence of a batsman dismissal. The data used in this research contain observations in which a batsman is not dismissed during the observed time period. These are known as censored observations and a logistic regression analysis does not take into account any possible censoring bias. Survival analysis is a method capable of accounting for these issues and will be addressed in Chapter 7.

Chapter 6

Ridge Regression

6.1 Introduction to Ridge Regression

In Chapter 5, the logistic regression analysis revealed strong multicollinearity among several predictor variables. This chapter will discuss ridge regression as a suitable statistical technique used to analyse multiple regression data that exhibit multicollinearity. Ridge regression is a suitable technique to use when the outcome is binary [25].

The presence of multicollinearity results in unbiased least square estimates. However, the variance of these estimates are large. This means that the estimates may be far from the true value. The ridge regression technique introduces some bias to the regression estimates, reducing the standard errors of these estimates and making them more reliable [11].

6.2 Multiple Regression

A regression analysis is a process which involves investigating the relationship between one dependent variable and one or more independent variables.

Multiple regression involves exploring the relationship between several predictor variables, $x_1, x_2, \ldots, x_k$, and one response variable, y_i. Values of y_i and $x_{1,i}, x_{2,i}, \ldots, x_{k,i}$ are observed for $i = 1, \ldots, n$.

A multiple regression model can be written as [34]:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i} + \epsilon_i, \tag{6.1}$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters estimated from the data and $\epsilon_i \sim N(0, \sigma^2)$. The residuals, ϵ_i, are assumed to be independent of one another, normally distributed with a mean

of zero and constant variance, σ_e^2.

The method of least squares can be used to estimate the parameters in a multiple regression model. This method chooses the best fit by finding estimates of $\beta_0, \beta_1, \ldots, \beta_k$ that minimise the sum of the squared residuals. In multiple regression, the model can be expressed in matrix form as [34]:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{6.2}$$

where

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{k1} \\ 1 & x_{12} & \ldots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \ldots & x_{kn} \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix},$$

and

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

$\boldsymbol{\beta}$ is estimated using least square estimates [34]:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}. \tag{6.3}$$

Under the assumption of normally distributed errors, the least squares estimators are the same as maximum likelihood estimators [7]. Given the design matrix, $\boldsymbol{X}$, and parameters, $\boldsymbol{\beta}$, the likelihood function can be written as [34]:

$$L(\boldsymbol{\beta}, \sigma^2) = f(\boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(\frac{-(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta})}{2\sigma^2}\right), \tag{6.4}$$

and the log-likelihood as [34]:

$$l(\boldsymbol{\beta}, \sigma^2) = \frac{-n}{2}\Big(\log(2\pi) + \log(\sigma^2)\Big) - \frac{1}{2\sigma^2}\Big(\boldsymbol{Y} - \boldsymbol{X\beta}\Big)'\Big(\boldsymbol{Y} - \boldsymbol{X\beta}\Big). \tag{6.5}$$

Maximising $L(\boldsymbol{\beta}, \sigma^2)$ or $l(\boldsymbol{\beta}, \sigma^2)$ analytically, enables derivation of the maximum likelihood estimates.

6.3 Ridge Regression Methodology

A ridge regression analysis is essentially a multiple regression analysis with some modifications.

6.3.1 Standardisation

The first step in a ridge regression analysis is to standardise the dependent and independent variables by subtracting their means and dividing by their standard deviations. All subsequent ridge regression calculations are based on standardised variables. The final regression coefficients are adjusted back to their original scale [11].

6.3.2 Parameter Estimation

As discussed in Section 6.2, the ridge regression coefficients are estimated using the least squares estimator, $\hat{\boldsymbol{\beta}}$. $\hat{\boldsymbol{\beta}}$ is chosen to minimise the sum of squares residuals, $\phi(\boldsymbol{\beta})$. The relationship between $\hat{\boldsymbol{\beta}}$ and $\phi(\boldsymbol{\beta})$ is defined as [54]:

$$\phi(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X'}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{6.6}$$

where

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}. \tag{6.7}$$

The estimator, $\hat{\boldsymbol{\beta}}$, is unbiased such that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and has a minimum variance among all linear unbiased estimators. The variance covariance matrix of the estimates is given by [54]:

$$V(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X'X})^{-1}. \tag{6.8}$$

Ridge regression adds a small value, k, to the diagonal elements of the correlation matrix. The gives the ridge estimator as [54]:

$$\underline{\tilde{\boldsymbol{\beta}}} = (\boldsymbol{X'X} + k\boldsymbol{I})^{-1}\boldsymbol{X'Y}. \tag{6.9}$$

The amount of bias in this estimator is given by [54]:

$$E(\underline{\tilde{\boldsymbol{\beta}}} - \underline{\boldsymbol{\beta}}) = [(\boldsymbol{X'X} + k\boldsymbol{I})^{-1}\boldsymbol{X'X} - \boldsymbol{I}]\underline{\boldsymbol{\beta}}. \tag{6.10}$$

The covariance matrix is given by [54]:

$$V(\underline{\tilde{\boldsymbol{\beta}}}) = (\boldsymbol{X'X} + k\boldsymbol{I})^{-1}\boldsymbol{X'X}(\boldsymbol{X'X} + k\boldsymbol{I})^{-1}. \tag{6.11}$$

The appropriate choice of k depends on the true value of the coefficients being estimated [11]. In [54], the authors suggested using a ridge trace plot. This plot shows the ridge regression coefficients as a function of k. A value is chosen for which the regression coefficients have stabilized.

The presence of multicollinearity results in unbiased least square estimates and inflated standard errors of the regression coefficients. The ridge regression technique introduces some bias to the regression estimates, via the small value, k. This reduces the standard errors of these estimates and makes them more reliable [11].

6.4 Chapter Remarks

Ridge regression techniques are useful when there is multicollinearity evident in multiple regression data. In Chapter 5, the logistic regression analysis revealed multicollinearity among several predictor variables. The application of ridge regression techniques, discussed in this chapter, to the ODI data, will be addressed in Chapter 8.

Chapter 7

Survival Analysis

Logistic regression analysis was performed in Chapter 5. This explored the association between several covariates and event occurrence. As this research is focused on the association between several covariates and the time to an event, logistic regression may not be the most appropriate technique for analysis. The particular event of interest is the occurrence of a batsman dismissal. The data used in this research contain observations in which a batsman is not dismissed during the observed time period. These are known as censored observations and logistic regression analysis does not take into account any possible censoring bias. This chapter will introduce key aspects of survival analysis. This is a suitable approach to use when survival time to an event is of particular interest, and censoring is present [55].

7.1 Introduction to Survival Analysis

Survival analysis is a branch of statistical analysis used to investigate and model the relationship between one dependent variable, the time until a particular event of interest occurs, and several predictor variables [55].

Survival analysis has been used in a range of settings. For example, in a medical setting in [50], three parametric models were applied to model survival data from five clinical trials of adjuvant therapy for Stage II breast cancer. Additionally, survival analysis has been applied to the banking sector. In [69], Cox models were applied to the analysis and prediction of bank failure.

7.2 Censoring

A characteristic of survival data is the possibility of an individual surviving longer than the follow up period. As a result, the observation of survival time for that individual is incomplete.

The process used to produce this particular type of observation is called censoring, while the observation is referred to as a censored observation. Censored observations can take on several forms. Observations that occur at particular times but finish before the outcome of interest occurs are referred to as right censored observations. This is the most common form of censored observation. If the event of interest has already occurred before the first observation is made, the observation is referred to as a left censored observation. If the event of interest is known to be between two particular time points, the observation is referred to as interval censored [68].

7.3 Survival and Hazard Functions

7.3.1 Survival Function

The survival function is the probability of observing a survival time greater than some time, t, defined as [68]:

$$S(t) = Pr(T > t). \tag{7.1}$$

The cumulative distribution function of the survival time, T, is the probability that an individual will have a survival time less than or equal to some time, t, defined as [68]:

$$F(t) = Pr(T \leq t). \tag{7.2}$$

Assuming the time random variable is continuous, the survival function may be expressed as [55]:

$$S(t) = e^{-H(t)}, \tag{7.3}$$

where $H(t)$ is the cumulative hazard function, defined in Section 7.3.3.

7.3.2 Hazard Function

The hazard function refers to the probability of an individual failing after some time, t, conditional on that individual surviving to time, t, and is defined as [68]:

$$h(t) = -\frac{\frac{dS(t)}{dt}}{S(t)}, \tag{7.4}$$

where $S(t)$ is the survival function, defined in Section 7.3.1.

7.3.3 Cumulative Hazard Function

The cumulative hazard function is defined as [68]:

$$H(t) = \int_0^t h(t)dt, \tag{7.5}$$

where $h(t)$ is the hazard function, defined in Section 7.3.2.

7.3.4 Kaplan-Meier Method

The most commonly used estimator of the survival function is called the Kaplan-Meier estimator. This estimator takes into account both uncensored and censored observations and considers the survival time to be a series of declining steps defined at each time point. The Kaplan-Meier estimator, also known as the Product Limit Estimator, is a non-parametric maximum likelihood estimate of $S(t)$ given by [55]:

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i}, \tag{7.6}$$

where n_i is the number of individuals who have survived to time, t_i, and d_i is the number of deaths at time, t_i. This estimator is non-parametric as it is fully based on the data, providing a true estimate of the survival function.

Without censoring, the Kaplan-Meier estimate of the survival function, $\hat{S}(t)$, at any given time, t, is equal to the proportion of individuals in the sample who are still alive [55].

7.4 Cox Proportional Hazards Model

On occasions when the objective of analysis is to assess the effect of continuous covariates on survival, Kaplan-Meier estimation techniques are inadequate. Cox proportional hazard models are used in this situation.

Cox models are used to model the relationship between the hazard function and several covariates. The hazard rate for one individual at time, t, with one independent variable, x, and the hazard rate at baseline levels of covariates, h_0, is defined as [55]:

$$h(t, x) = h_0 \text{exp}(\beta_1 x). \tag{7.7}$$

If there is more than one independent variable, Equation (7.7) can be written in matrix notation:

$$h(t, \boldsymbol{X}) = h_0 \text{exp}(\boldsymbol{\beta}' \boldsymbol{X}). \tag{7.8}$$

Refer to Section 6.2 in Chapter 6 for details on matrix form specification in multiple regression. Equation (7.8) can be generalised to:

$$h(t, \boldsymbol{X}) = h_0(t, \boldsymbol{\alpha}) \text{exp}(\boldsymbol{\beta}' \boldsymbol{X}), \tag{7.9}$$

where $h_0(t, \boldsymbol{\alpha})$ is the hazard function at baseline levels of covariates and is allowed to vary over time and $\boldsymbol{\alpha}$ is a vector of parameters influencing the baseline hazard function. Equation (7.9) is referred to as the Cox model.

Cox models are semi-parametric, consisting of a parametric and a non-parametric component. The parametric component makes assumptions about the effect of covariates, while the non-parametric component makes no assumption about the distribution of the hazard function.

In some situations, the distribution of the survival time has a known parametric form. In these situations, fully parametric regression models may be utilised. These differ from the Cox proportional hazards model in that both the hazard function and the effect of the covariates are specified.

The mathematical theory behind parameter estimation in a Cox model is described in Appendix A. Analytical parameter estimation is discussed in Section A.1, while numerical parameter estimation is addressed in Section A.2. Additionally, several examples of parametric regression models are described in Section A.3.

7.4.1 Cox Model Assumptions

There are two assumptions that the Cox model relies on. The first is based on linearity and the second is based on proportional hazards [55].

1. The effect of each covariate is linear in the log hazard function.

2. The ratio of the hazard function for two individuals with different sets of covariates does not depend on time.

7.5 Chapter Remarks

In general, the results from using a Cox model will closely approximate the results from the correct parametric model [68]. As such, the Cox model is a useful approach and will provide reliable enough results when there is uncertainty around the correct parametric form of the distribution of survival times. The application of censoring techniques, discussed in Section 7.2, and the Cox model methodology, discussed in Section 7.4, to the ODI data, will be addressed in Chapter 8.

Chapter 8

Model Application

In this chapter, a combination of ridge regression techniques, discussed in Chapter 6, and survival analysis techniques, discussed in Chapter 7, are applied to the ODI data. The modelling procedure and results are presented when data is associated with individual batsmen. This is followed by illustration of the equivalent analysis when data is associated with batting partnerships. The chapter proceeds with an illustration of the final models and how they are used to calculate the Area Under the Curve (AUC) as a batting performance metric. Optimal batting partnership combinations from a wide variety of cricketing nations are derived based on a combination of AUC and other robust batting performance metrics: total runs scored, proportion of team runs scored and winning percentage. These optimal partnerships maximise the number of runs scored in the least amount of time possible and the team's chances of winning. Practically, this approach reveals which batsmen are capable of pacing their innings to maximise their team's chances of winning. This is a novel approach which holistically considers strike rate, total runs and the match context. This is particularly difficult to assess in the first innings of a match when setting a total. Additionally, the metrics are used to determine the optimal New Zealand batting order during a selection of ODI games.

8.1 Modelling Methodology

Using existing functionality as part of the 'survival' package [12] in R [80], this research utilised a combination of Cox proportional hazard modelling and ridge regression analysis to model data associated with individual batsmen. There was some uncertainty over the correct parametric form of the survival times of batsmen. Consequently, the Cox modelling approach was preferred over parametric regression modelling, as discussed in Section 7.5 in Chapter 7. A survival object was created and taken to represent the response variable in a Cox model. This consisted of a particular event and the time taken to that event, in this case the event being a batsman dismissal. The total number of balls faced by the batsman was taken to represent

the time to that event. The censoring methodology discussed in Section 7.2 in Chapter 7 and employed in [60], was also adopted. Those observations where batsmen were not dismissed were referred to as censored observations. This research adopted the Efron method, discussed in Section A.1.2.3 in Appendix A, to account for repeated failure times. The Efron method was chosen as it is considered to provide the closest approximation to the exact partial likelihood, and is employed as the default method in R [80].

To assess the first assumption of a Cox model described in Section 7.4.1 in Chapter 7, this research adopted the following approach, described in [89]. Each predictor was split into groups defined at each quantile. A Cox model was fitted to all batting performance predictor groups. For each predictor, the associated estimated coefficients from the Cox model were plotted against the midpoints of each group. For every predictor, the resulting line connecting the midpoints did not follow an approximate straight line. This provided evidence that non-linearity may exist among every predictor. A common transformation used to correct for this situation is the square root transformation [55]. As the data primarily consist of counts, a square root transformation was applied to successfully remove non-linearity in all predictors.

8.1.1 Opening Batsman Modelling

A large number of Cox models were fitted to data associated with openers. To achieve model parsimony, each model was fitted with a maximum selection of four predictor variables from the eight individual batting performance metrics. For each fitted model, the covariates that showed evidence of multicollinearity were specified as ridge regression terms. AIC was initially used to rank these models. From here, two model selection criteria were utilised in order to find an appropriate set of candidate models. To meet the first criteria, the estimated model coefficients had to be *practical*. For example, an increase in predictors such as batsman dot balls was expected to decrease the likelihood of a batsman surviving the next ball. This is due to additional factors such as fatigue, bowling and fielding strategy and batsmen adopting higher risk strategies to increase the run rate. For these predictors, the first criteria was met if the corresponding estimated coefficient was negative. To satisfy the second criteria, the predictors had to be *statistically significant*. In the work in [36], a predictive model capable of calculating the ball-by-ball probability of an opening batsman being dismissed in the first innings of a limited

overs cricket game was formulated. In addition to the criteria previously discussed, an additional criteria was included in the work in [36]. That criteria required that the probability of batsman survival either remained constant or decreased at the ball-by-ball level. Here, that criteria used in [36] is relaxed. The reason for this relaxation is due to metrics which on occasion are smaller when one ball is faced, relative to when the previous ball was faced. For instance, batsman contribution is based on the proportion of team runs scored by the individual batsman. When extras occur in the data, the total number of team runs increases, while the total number of individual batsman runs remains the same. This is due to the rules regarding allocation of extras. This results in a small reduction in contribution and consequently a slight increase in survival probability. The R code used to carry out the opening batsman modelling procedure is presented in Section G.2 in Appendix G. Table 8.1 illustrates the models ranked by AIC that meet the criteria. The Cox model with consecutive dot balls, less than 2 in 4, boundaries and contribution as predictors had the highest AIC.

Rank-ordered model	**Model predictors**			
	P1	**P2**	**P3**	**P4**
1	Consecutive dot balls	Less than 2 in 4	Boundaries	Contribution
2	Runs	Dot balls	Less than 2 in 4	
3	Runs	Consecutive dot balls	Less than 2 in 4	
4	Dot balls	Boundaries	Contribution	
5	Dot balls	Contribution	Percentage boundaries	
6	Consecutive dot balls	Less than 2 in 4	Boundaries	
7	Consecutive dot balls	Boundaries	Contribution	
8	Less than 2 in 4	Boundaries	Contribution	
9	Consecutive dot balls	Contribution	Percentage boundaries	

Table 8.1: Opening batsman ranked models

To assess the second assumption of the Cox model described in Section 7.4.1, this research applied the 'cox.zph' function [13] as part of the 'survival' package [12] in R [80]. This function tests the hypotheses H_0: The proportional hazards assumption holds vs H_1: The proportional hazards assumption does not hold. This function was applied to assess whether each of the predictors in each model from Table 8.1 showed any evidence of violation towards proportional hazards. Through evaluation of this assumption, the candidate set of models was reduced to one model from Table 8.1, with all other models failing the second assumption. Table 8.2 illustrates the predictors included in this final model, together with their estimated coefficients.

Batsman class	Model predictors and coefficients					
	P1	**Coef**	**P2**	**Coef**	**P3**	**Coef**
Openers	Runs	-1.478	Consecutive dot balls	-0.624	Less than 2 in 4	-0.694

Table 8.2: Opening batsman final model predictors

8.1.2 Non-Opening Batsman Modelling

In addition, it was expected that these factors may have a differing effect on the performance of non-opening batsmen in the remaining of the batting innings compared with opening batsmen. As such, the modelling procedure described in Section 8.1.1 was applied to data associated with batsmen categorised as either top order, middle order, lower order or tail, followed by data associated with all batsmen, as suggested in [36]. Data remained associated with first innings of games. The emphasis remained on the first innings due to the challenge of accessing strategies for setting a total. Table 8.3 illustrates the predictors included in the final models, together with their estimated coefficients.

Batsman class	Model predictors and coefficients							
	P1	**Coef**	**P2**	**Coef**	**P3**	**Coef**	**P4**	**Coef**
Top order	Consecutive dot balls	-0.285	Less than 2 in 4	-0.273	Boundaries	-0.547	Contribution	-0.740
Middle order	Runs	-0.317	Dot balls	-0.653	Less than 2 in 4	-0.441	Contribution	-1.385
Lower order	Runs	-0.613	Consecutive dot balls	-0.637	Less than 2 in 4	-0.525	Contribution	-1.189
Tail	Runs	-0.667	Dot balls	-0.574	Less than 2 in 4	-0.657		
All batsmen	Runs	-1.459	Less than 2 in 4	-1.209	Contribution	-0.050		

Table 8.3: Categorised batsman final model predictors

For all models described in Table 8.3, evaluation of the proportionality assumption resulted in no statistical evidence to suggest that any model violated this assumption. With model fitting, model checking and model selection procedures successfully completed, the six final models were used to generate results, discussed in Section 8.2.

8.2 Individual Batsman Results

The ball-by-ball survival probabilities for all individual batsmen considered were calculated using Equation (2.13). The survival probabilities for each batsman in each game were plotted to produce survival curves. The intent was to carry out a qualitative analysis of style in individual games.

Table 8.4 describes a selection of batsmen and games. Figures 8.1, 8.2 and 8.3 were constructed to illustrate the ball-by-ball survival probabilities for these batsmen during these games.

Batsman	**Match**	**Date played**
MJ Guptill	New Zealand v Australia	08/02/2016
BB McCullum	New Zealand v Australia	08/02/2016
DA Warner	Australia v South Africa	19/11/2014
AJ Finch	Australia v South Africa	19/11/2014
CJ Anderson	New Zealand v Sri Lanka	14/02/2015
GD Elliott	New Zealand v Sri Lanka	14/02/2015
JP Duminy	Sri Lanka v South Africa	06/07/2014
DA Miller	Sri Lanka v South Africa	06/07/2014
KD Mills	New Zealand v South Africa	21/10/2014
TA Boult	New Zealand v South Africa	21/10/2014
S Senanayake	Bangladesh v Sri Lanka	17/02/2014
SL Malinga	Bangladesh v Sri Lanka	17/02/2014

Table 8.4: ODI batsman match information

Figure 8.1 illustrates a selection of the results from the final model associated with opening

batsmen.

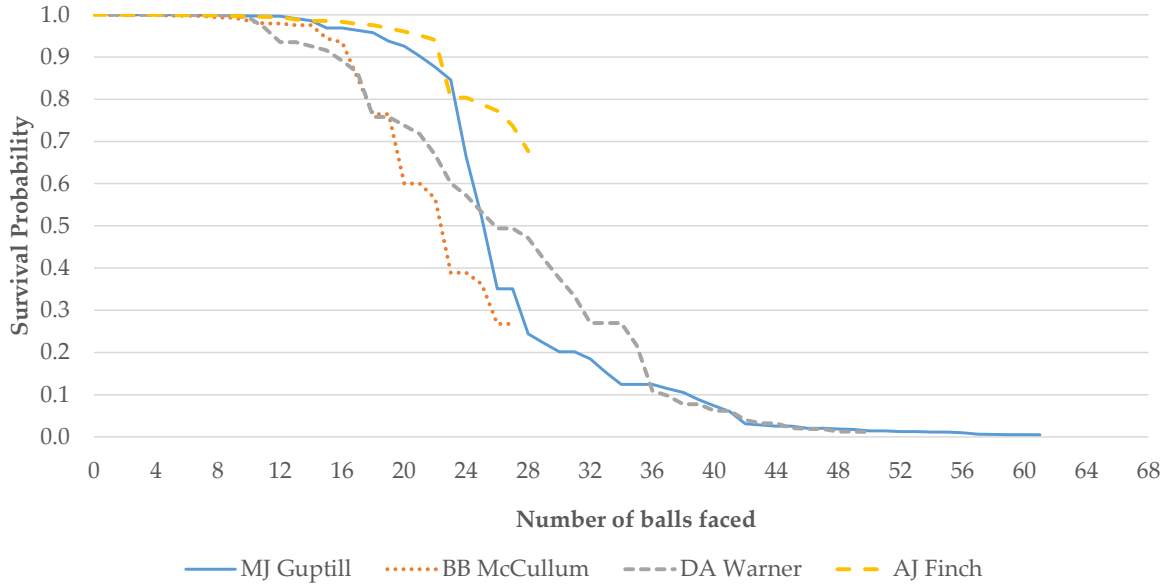

Figure 8.1: Survival probabilities for MJ Guptill, BB McCullum, DA Warner and AJ Finch

The results show that in the ODI between New Zealand and Australia, MJ Guptill had higher survival probabilities than BB McCullum. In the ODI between Australia and South Africa, AJ Finch had higher survival probabilities than DA Warner.

Of interest is the apparently different roles adopted by each batsman. The lower survival probabilities and steeper slopes are indicative of risky behaviour. Given the survival properties and the slope of the curves, the results imply that MJ Guptill and AJ Finch opted for a more conservative style of play, whilst BB McCullum and DA Warner opted for higher risk strategies respectively. MJ Guptill scored 59 from 61 (4x4 runs, 3x6 runs) and BB McCullum scored 47 from 27 (6x4 runs, 3x6 runs). New Zealand won the game by 55 runs. In the second game, AJ Finch scored 109 from 127 (9x4 runs, 3x6 runs) and DA Warner scored 53 from 50 (6x4 runs, 2x6 runs). Australia won the game by 73 runs.

Figure 8.2 illustrates a selection of the results from the final model associated with middle order batsmen.

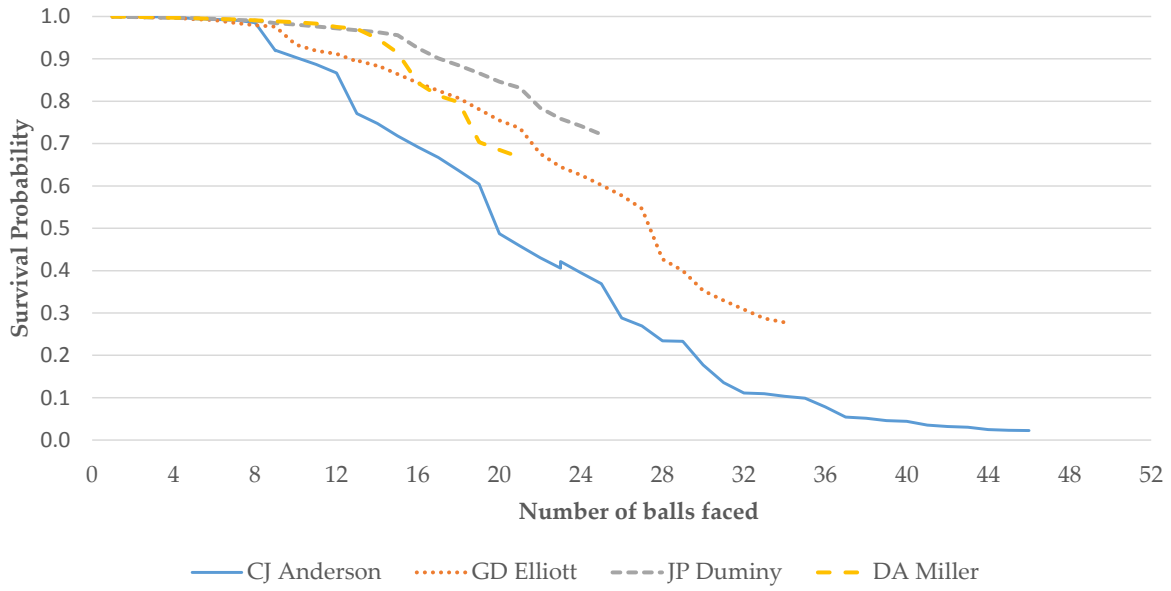

Figure 8.2: Survival probabilities for CJ Anderson, GD Elliott, JP Duminy and DA Miller

GD Elliott and JP Duminy opted for a more conservative style of play, whilst CJ Anderson and DA Miller opted for higher risk strategies respectively. GD Elliott scored 29 from 34 (2x4 runs, 0x6 runs) and CJ Anderson scored 75 from 46 (8x4 runs, 2x6 runs). New Zealand won the game by 98 runs. In the second game, JP Duminy scored 16 from 25 (0x4 runs, 0x6 runs) and DA Miller scored 36 from 21 (2x4 runs, 2x6 runs). South Africa won the game by 75 runs.

Figure 8.3 illustrates a selection of the results from the final model associated with tail batsmen.

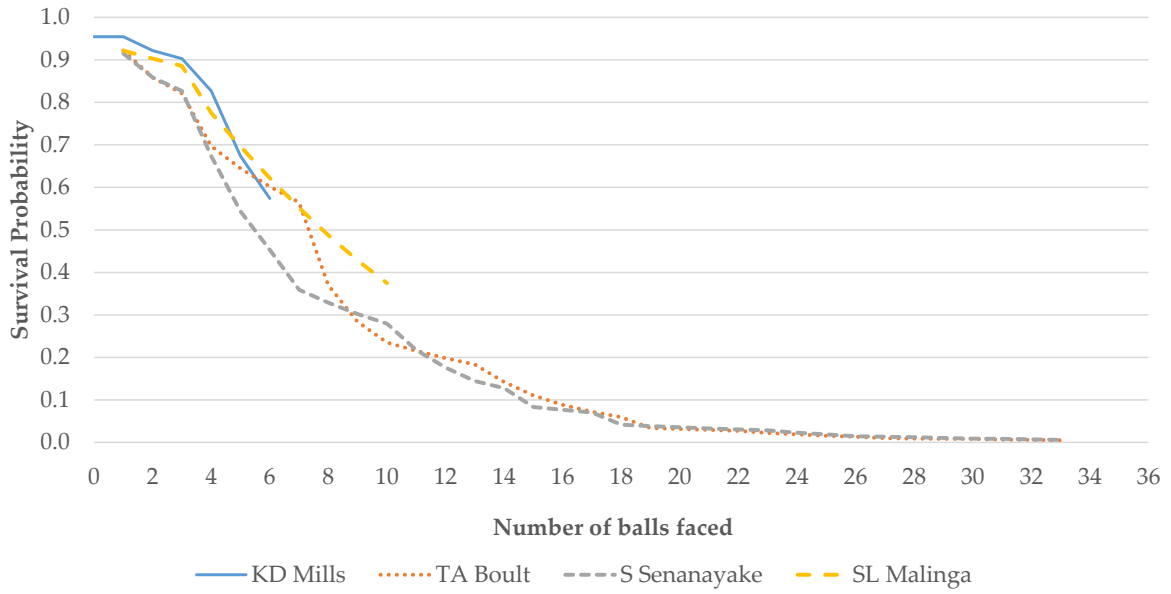

Figure 8.3: Survival probabilities for KD Mills, TA Boult, S Senanayake and SL Malinga

KD Mills and SL Malinga opted for a more conservative style of play, whilst TA Boult and S Senanayake opted for higher risk strategies respectively. However, these differences in style of play are fairly small. KD Mills scored 1 from 6 (0x4 runs, 0x6 runs) and TA Boult scored 21 from 33 (2x4 runs, 1x6 runs). South Africa won the game by 6 wickets. In the second game, SL Malinga scored 0 from 10 (0x4 runs, 0x6 runs) and S Senanayake scored 30 from 48 (2x4 runs, 1x6 runs). Sri Lanka won the game by 13 runs.

The results illustrated in Figures 8.1, 8.2 and 8.3 consider different order batsmen from a variety of cricketing nations, highlighting the ability for this technique to be used to compare players from around the world, which is useful for scouting purposes.

8.2.1 Individual Batsman Performance Measures

The area under each survival curve and the total area under all curves for each batsman were calculated. To account for the differing number of games played by each batsman, the average AUC for each batsman was computed. This was used as a metric for batsman comparison purposes. Another metric, the wins-to-games ratio for each batsman was also calculated.

As the average number of games per batsman was approximately five, results for each batsman were aggregated into groups of five according to their rank ordering, based on the average AUC. This enabled the average AUC statistic to be assessed as a meaningful measure of performance, where a higher AUC indicates longer periods of productive time spent at the crease and is therefore indicative of better performances. This is likely to increase the team's chances of winning.

The ODI ICC batting ranking for each batsman was obtained. The ranking used was that following the most recent international played between 26th December 2013 and 14th February 2016. The average number of runs scored and average proportion of team runs scored for each cohort were also calculated.

8.2.2 Individual Batsman Insights

Table 8.5 summarises the batsmen included in the top cohort associated with each batsman category.

Top cohort	Openers	Top order	Middle order	Lower order	Tail
1	V Sibanda	S Malik	SM Ervine	S Shenwari	SCJ Broad
2	KC Sangakkara	N Jamal	S Haider	LE Plunkett	MA Wood
3	SE Marsh	MR Marsh	Misbah-ul-Haq	NM Coulter-Nile	NM Coulter-Nile
4	MN Samuels	LRPL Taylor	A Ali	MN Waller	J Nyumbu
5	J Anderson	S Ahmed	I Wasim	MS Wade	KMDN Kulasekara

Table 8.5: Top cohort batsmen

Figure 8.4 compares the average AUC with the average ICC ranking per opening batsman in each cohort. The size of the bubble is based on the average number of runs scored per cohort, with that average used to annotate the graph. It is important to note that the ICC ranking considers all innings compared with the method explored here, which assesses first innings only. A moderately strong relationship is illustrated between the average AUC and the average ICC rankings for openers, with a correlation of -0.58.

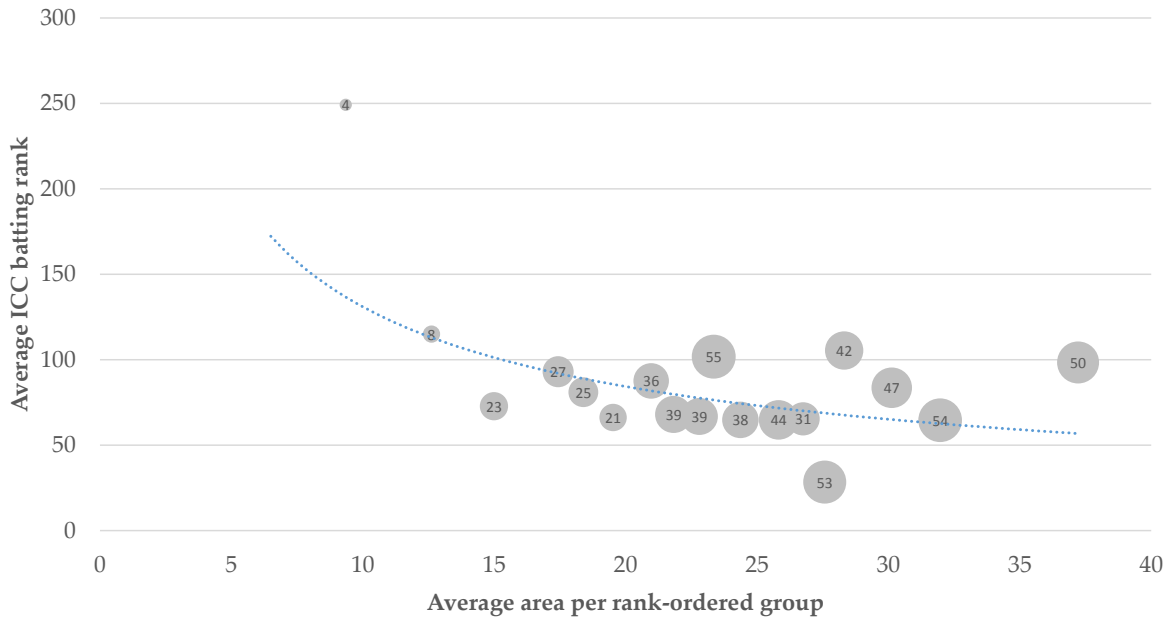

Figure 8.4: Average AUC for rank-ordered cohorts compared with ODI ICC batting ranking for openers and average number of runs scored for the observed time frame (within bubble)

Figure 8.5 compares the average AUC with the average ICC ranking per top order batsman in each cohort. A moderately strong relationship is illustrated between the average AUC and the average ICC rankings for top order batsmen, with a correlation of -0.39.

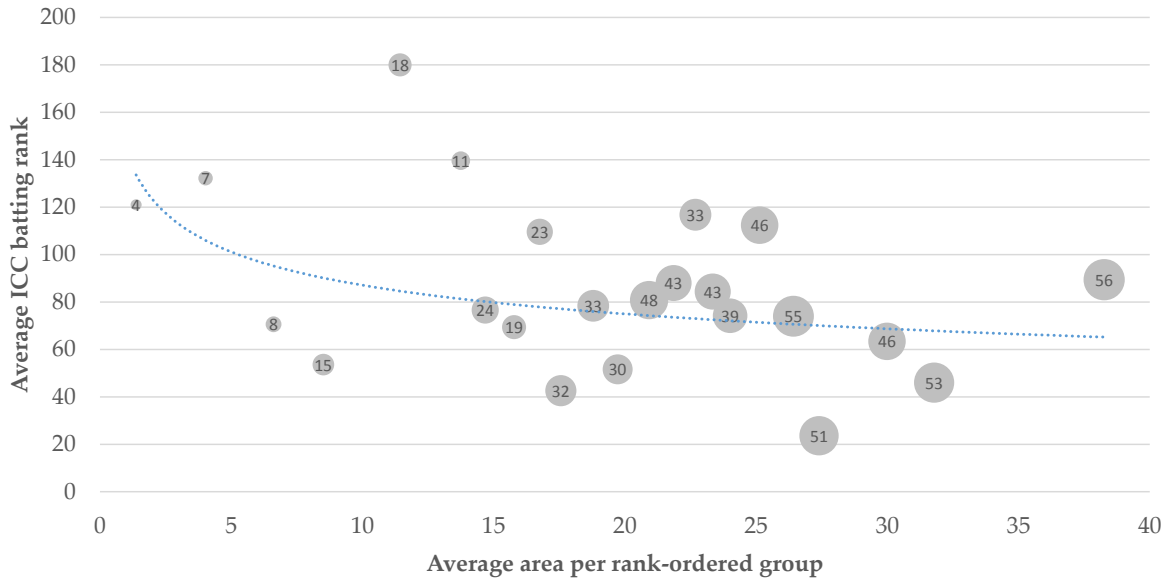

Figure 8.5: Average AUC for rank-ordered cohorts compared with ODI ICC batting ranking for top order batsmen and average number of runs scored for the observed time frame (within bubble)

Figure 8.6 compares the average AUC with the average ICC ranking per middle order batsman in each cohort. A moderately strong relationship is illustrated between the average AUC and the average ICC rankings for middle order batsmen, with a correlation of -0.69.

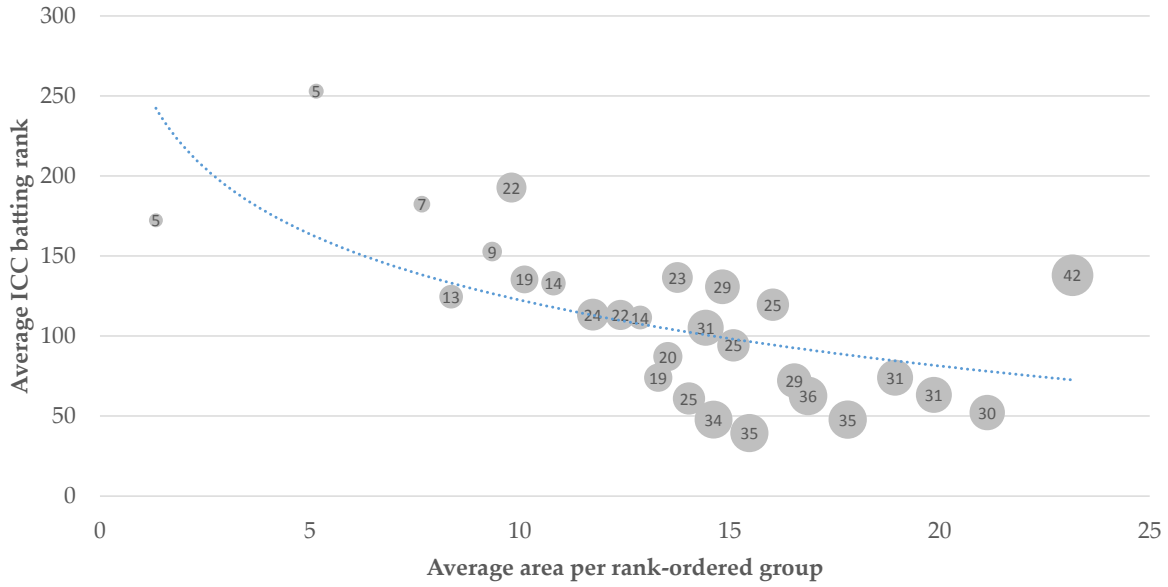

Figure 8.6: Average AUC for rank-ordered cohorts compared with ODI ICC batting ranking for middle order batsmen and average number of runs scored for the observed time frame (within bubble)

Figure 8.7 compares the average AUC with the average number of runs scored for an opening batsman in each cohort.

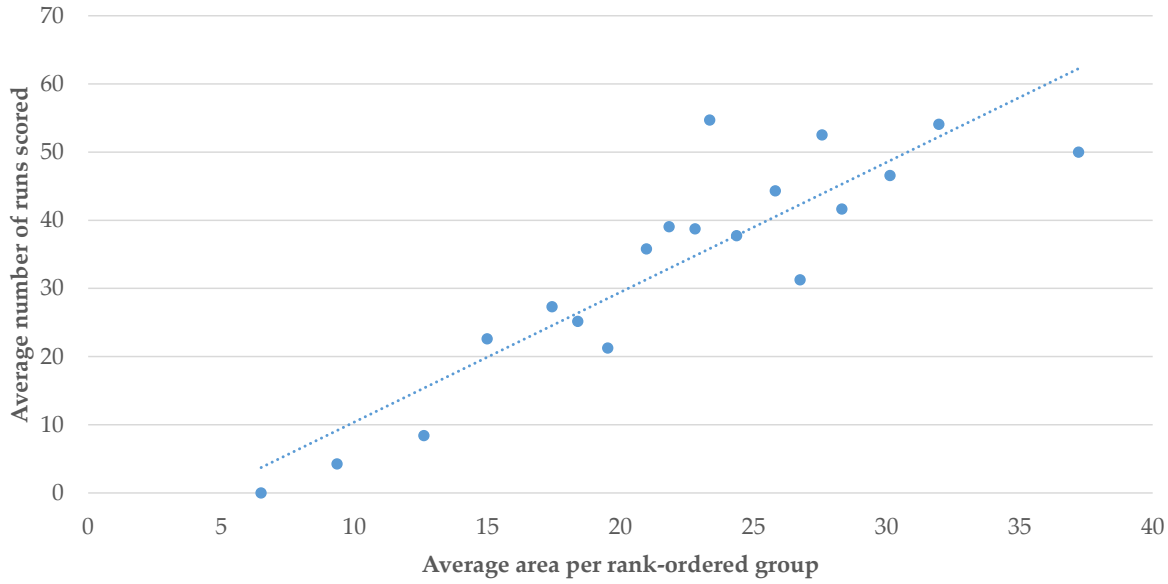

Figure 8.7: Average AUC for rank-ordered cohorts compared with average number of runs scored for openers for the observed time frame

Figure 8.8 compares the average AUC with the logit of the average proportion of team runs scored for an opening batsman in each cohort. A logit transformation of the average proportion of team runs scored was applied to stabilise the variance.

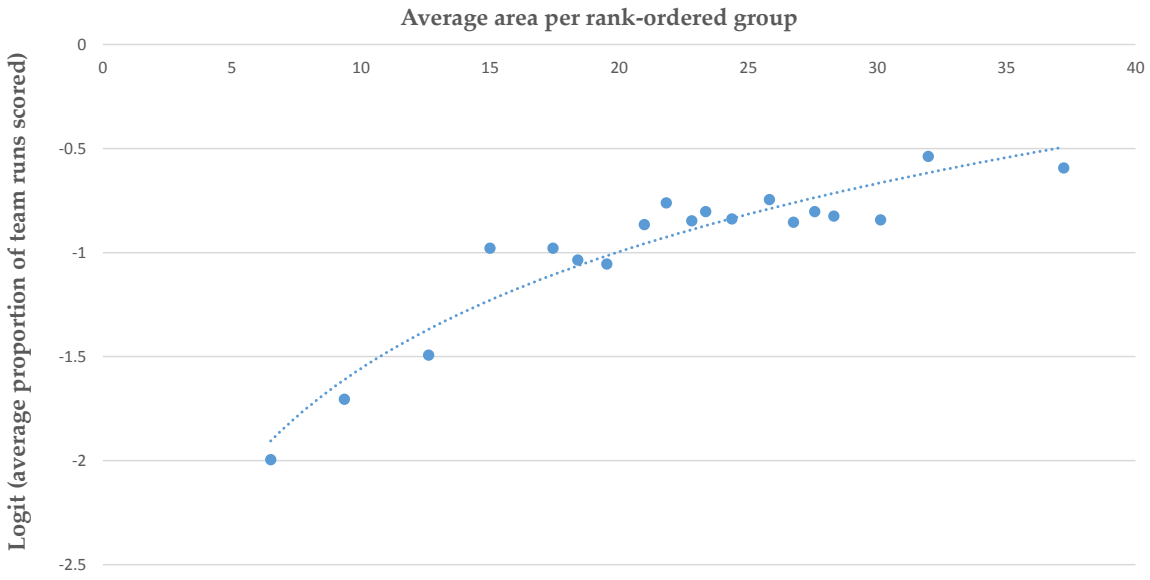

Figure 8.8: Average AUC for rank-ordered cohorts compared with logit of the average proportion of team runs scored for openers for the observed time frame

As illustrated in Figures 8.7 and 8.8, the average AUC is strongly related to both the average number of runs scored and the average proportion of team runs scored. The correlation between the square root of the average AUC and the square root of the average runs is 0.924. Similarly, the correlation between the square root of the average AUC and the square root of the average proportion of team runs scored is 0.929. As expected, these correlations indicate that the more *effective* opening batsmen simultaneously occupy the crease and score runs. Importantly, the nature of the model indicates a steady scoring rate among openers. This starts to set up for investigation into optimal contribution: optimal batting partnerships in low risk conservation and high risk aggression.

Figure 8.9 compares the average AUC with the average number of runs scored for a middle order batsman in each cohort. A strong positive relationship is illustrated between the average

AUC and the average number of runs scored for middle order batsmen, with a correlation of 0.88.

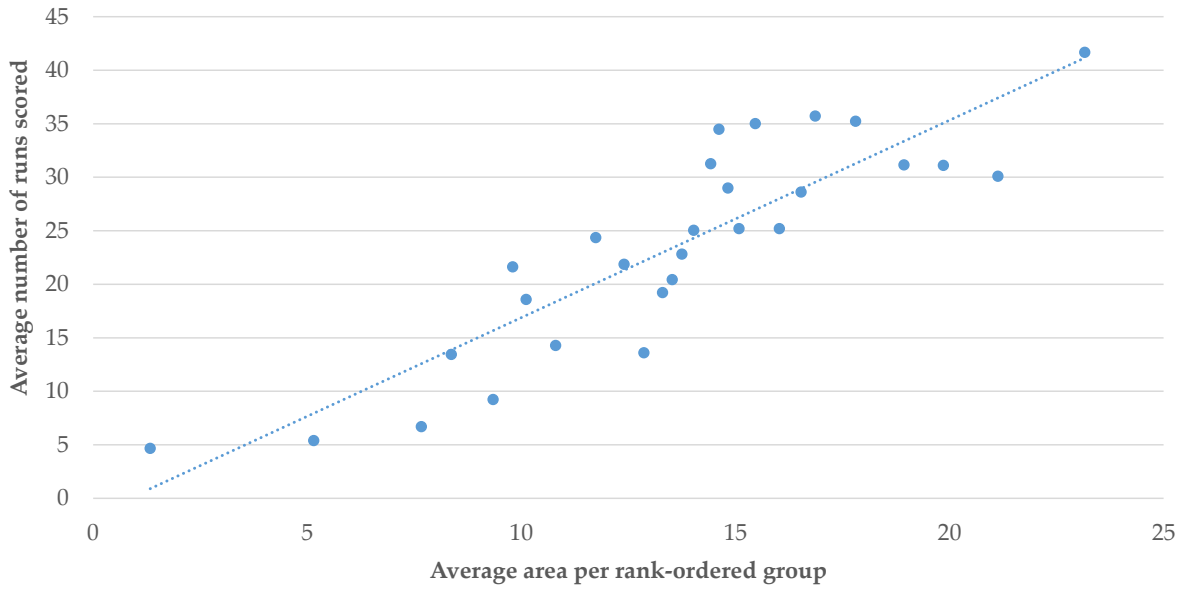

Figure 8.9: Average AUC for rank-ordered cohorts compared with average number of runs scored for middle order batsmen for the observed time frame

Figure 8.10 compares the average AUC with the logit of the average proportion of team runs scored for a middle order batsman in each cohort. A strong positive relationship is illustrated between the average AUC and the logit of the average proportion of team runs scored for middle order batsmen, with a correlation of 0.85.

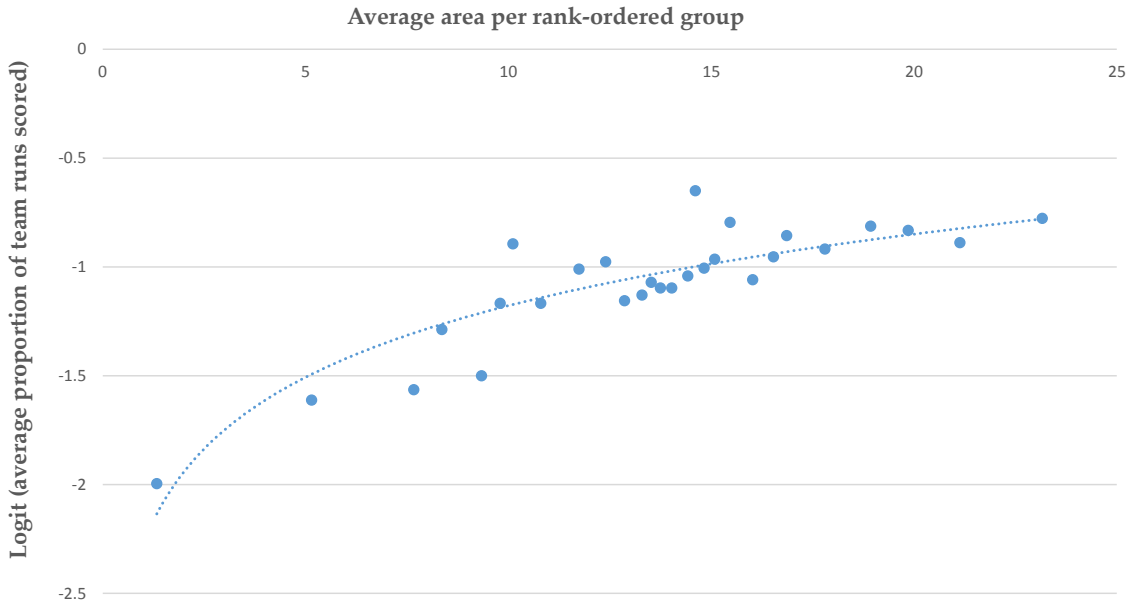

Figure 8.10: Average AUC for rank-ordered cohorts compared with logit of the average proportion of team runs scored for middle order batsmen for the observed time frame

Figure 8.11 compares the average AUC against the average number of runs scored for a tail batsman in each cohort. A strong positive relationship is illustrated between the average AUC and the average number of runs scored for tail batsmen, with a correlation of 0.84.

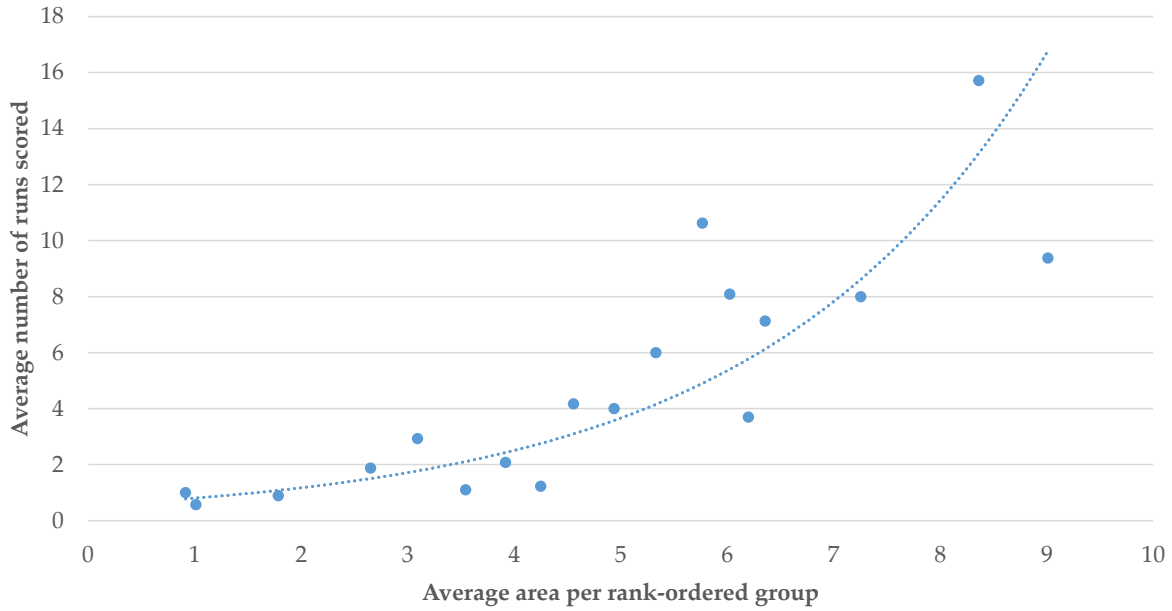

Figure 8.11: Average AUC for rank-ordered cohorts compared with average number of runs scored for tail batsmen for the observed time frame

Figure 8.12 compares the average AUC with the logit of the average proportion of team runs scored for a tail batsman in each cohort. A strong positive relationship is illustrated between the average AUC and the logit of the average proportion of team runs scored for tail batsmen, with a correlation of 0.84.

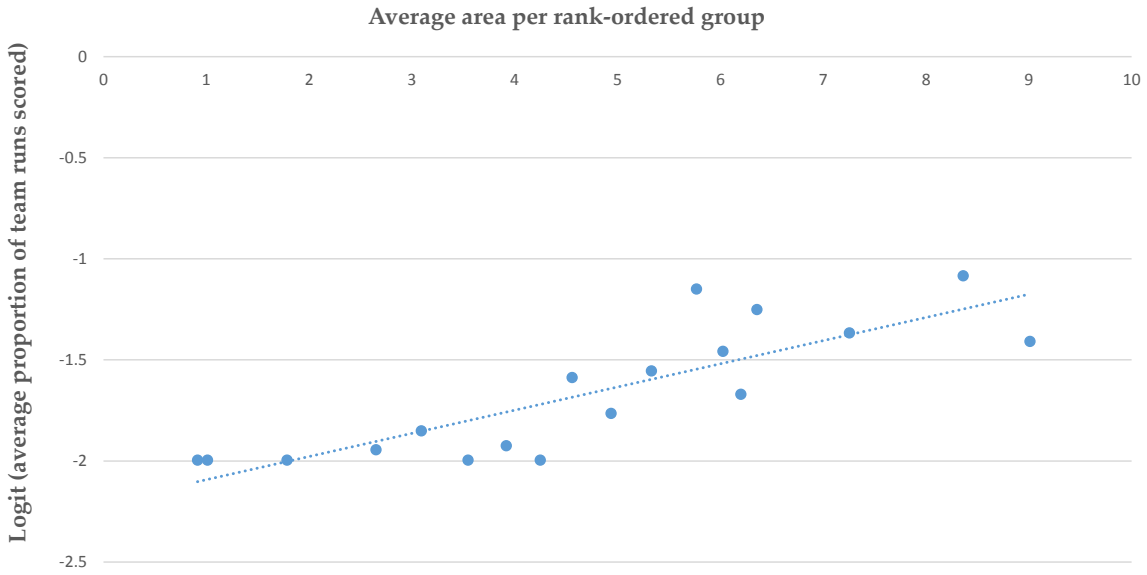

Figure 8.12: Average AUC for rank-ordered cohorts compared with logit of the average proportion of team runs scored for tail batsmen for the observed time frame

From Figures 8.9, 8.10, 8.11 and 8.12, the positive association between a batsman occupying the crease and the number of runs scored and proportion of team runs scored is not restricted to opening batsmen. This association is also evident in lower order batsmen. Importantly, the nature of the model for tail batsmen indicates that the longer tail batsmen remain in bat, the larger the increase in scoring rate becomes. This highlights that even those batting at 10 and 11 must be capable of scoring runs.

8.3 Model Structure Interpretation

In all models, the longer a player bats, the more likely he is to be dismissed. Of interest are the other attributes.

1. For openers, these results suggest that a high number of runs scored, consecutive dot balls faced and balls faced in which less than 2 in 4 runs are scored are associated with an increased probability of being dismissed.

2. For top order batsmen, a high number of consecutive dot balls, balls faced in which less than 2 in 4 runs, boundaries scored and a high contribution are associated with an

increased probability of being dismissed.

3. For middle order batsmen, a high number of runs scored, dot balls faced, balls faced in which less than 2 in 4 runs and a high contribution are associated with an increased probability of being dismissed.

4. For lower order batsmen, a high number of runs scored, consecutive dot balls faced, balls faced in which less than 2 in 4 runs are scored and a high contribution are associated with an increased probability of being dismissed.

5. For tail batsmen, a high number of runs scored, dot balls faced and balls faced in which less than 2 in 4 runs are scored are associated with an increased probability of being dismissed.

These attributes are consistent with intuition in a cricket context. It is particularly important for openers to score the highest number of runs, by rotating the strike. The more consecutive dot balls faced, the less frequently the openers rotate the strike. As such, the more likely they are to be dismissed. In contrast, top order and middle order batsmen are more affected by their contribution to the team and less affected by consecutive dot balls, compared with openers. Additionally, middle order batsmen are more affected by dot balls and occasions where 2 runs or less are scored within 4 deliveries, compared with openers and top order batsmen respectively.

8.4 Batting Partnership Modelling

The main objective of this research was to investigate the survival rates of batting partnerships in order to optimise batting partnership strategy. Here, batting strategy refers to determining the run rate which provides the best chance of winning when setting a total. As shown in Section 8.3, different batsmen may opt for distinct individual strategies. When focussing on batting partnerships, the event taken for survival analysis was a batsman dismissal. However, the time to that event was modified and taken to be the total number of balls faced by the corresponding partnership. The methodology discussed in Section 8.1 was applied to batting partnership data. A large number of Cox models were fitted to each of the ten partnership datasets associated

with a given wicket. To achieve model parsimony, each model was fitted consisting of a maximum selection of four predictor variables from the eight batting performance metrics at the partnership level. Table 8.6 illustrates the predictors included in each partnership final model, together with their estimated coefficients.

Wicket	Model predictors and coefficients					
	P1	**Coef**	**P2**	**Coef**	**P3**	**Coef**
1	Partnership dot balls	-2.613	Partnership boundaries	-0.915	Partnership contribution	-0.213
2	Partnership runs	-0.744	Partnership dot balls	-0.648	Partnership less than 2 in 4	-0.645
3	Partnership runs	-1.584	Partnership consecutive dot balls	-0.522	Partnership less than 2 in 4	-0.606
4	Partnership runs	-1.175	Partnership dot balls	-2.192	Partnership contribution	-0.467
5	Partnership runs	-0.642	Partnership dot balls	-0.587	Partnership consecutive dot balls	-0.650
6	Partnership runs	-0.797	Partnership consecutive dot balls	-1.391	Partnership contribution	-1.134
7	Partnership runs	-0.543	Partnership less than 2 in 4	-1.054	Partnership contribution	-1.477
8	Partnership dot balls	-0.407	Partnership consecutive dot balls	-0.693	Partnership contribution	-2.242
9	Partnership dot balls	-0.404	Partnership less than 2 in 4	-0.587	Partnership contribution	-1.751
10	Partnership runs	-1.796	Partnership less than 2 in 4	-1.478	Partnership percentage dot balls	-0.172
All wickets	Partnership runs	-1.473	Partnership less than 2 in 4	-1.288	Partnership contribution	-0.057

Table 8.6: Partnership final model predictors

For all models described in Table 8.6, evaluation of the proportionality assumption resulted in no statistical evidence to suggest that any model violated this assumption. With model fitting, model selection and model checking procedures successfully completed, the eleven final models were used to generate results, discussed in Section 8.5.

8.5 Batting Partnership Results

The ball-by-ball survival probabilities for all batting partnerships considered were calculated using Equation (2.13). The survival probabilities for each batting partnership in each game were plotted to produce survival curves.

Table 8.7 describes a selection of batting partnerships and games. Figures 8.13, 8.14, 8.15 and 8.16 were constructed to illustrate the ball-by-ball survival probabilities for these partnerships during these games.

Batting partnership		Match	Date played
P1	**P2**		
MJ Guptill	BB McCullum	New Zealand v Sri Lanka	14/02/2015
CJ Anderson	L Ronchi	New Zealand v Sri Lanka	14/02/2015
AJ Finch	SE Marsh	Australia v England	26/01/2014
GJ Bailey	MS Wade	Australia v England	26/01/2014
MJ Guptill	BB McCullum	New Zealand v Sri Lanka	14/02/2015
AN Cook	IR Bell	Australia v England	17/01/2014
DA Warner	AJ Finch	Australia v Pakistan	12/10/2014
S Dhawan	RG Sharma	India v Pakistan	15/02/2015
EJG Morgan	J Root	New Zealand v England	20/02/2015
GJ Maxwell	MR Marsh	Zimbabwe v Australia	25/08/2014
GD Elliott	LRPL Taylor	England v New Zealand	12/06/2015
AD Mathews	KC Sangakkara	Sri Lanka v England	03/12/2014
H Hassan	S Zadran	New Zealand v Afghanistan	08/03/2015
TA Boult	MJ Henry	New Zealand v Sri Lanka	15/01/2015
E Adil	R Ali	Australia v Pakistan	20/03/2015
PVD Chameera	KMDN Kulasekara	Scotland v Sri Lanka	11/03/2015

Table 8.7: ODI batting partnership match information

Figure 8.13 illustrates survival curves for different order partnerships in the same ODI game. The results show that in the ODI between New Zealand and Sri Lanka, MJ Guptill and BB

McCullum had higher survival probabilities than CJ Anderson and L Ronchi. In the ODI match between Australia and England, AJ Finch and SE Marsh had higher survival probabilities than GJ Bailey and MS Wade.

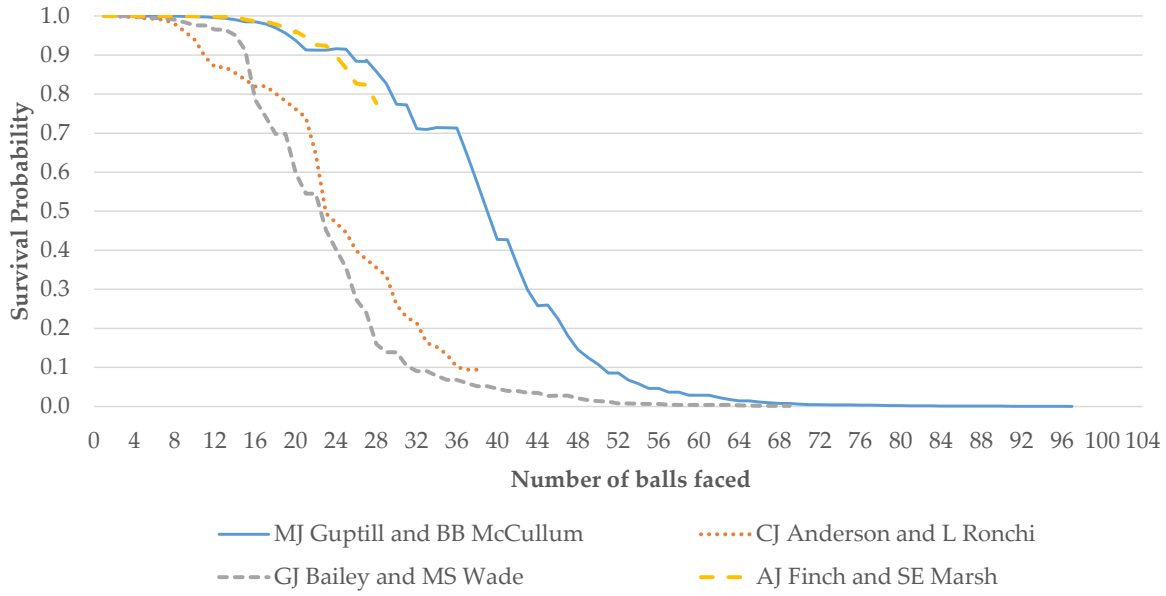

Figure 8.13: Survival probabilities for MJ Guptill and BB McCullum, CJ Anderson and L Ronchi, AJ Finch and SE Marsh, GJ Bailey and MS Wade

Of interest is the apparently different roles adopted by each partnership. The lower survival probabilities and steeper slopes are indicative of risky behaviour. Given the survival properties, the results imply that MJ Guptill and BB McCullum and AJ Finch and SE Marsh, as partnerships, opted for a more conservative style of play, whilst CJ Anderson and L Ronchi and GJ Bailey and MS Wade opted for a higher risk strategy. MJ Guptill and BB McCullum scored 104 from 97 and New Zealand won the game by 98 runs. CJ Anderson and L Ronchi scored 72 from 38. AJ Finch and SE Marsh scored 11 from 28 and Australia won the game by 5 runs. GJ Bailey and MS Wade scored 52 from 69.

Figure 8.14 illustrates a selection of the results from the final model associated with opening partnerships. Given the survival properties, the results imply that DA Warner and AJ Finch opted for a high risk strategy, on route to 48 from 63. Australia won the game by 1 run. Similarly, S Dhawan and RG Sharma opted for a high risk strategy, though not as extreme as DA Warner and AJ Finch. S Dhawan and RG Sharma scored 32 from 45 and Sri Lanka won the game by 76 runs. In contrast, AN Cook and IR Bell and MJ Guptill and BB McCullum played

relatively more conservatively. AN Cook and IR Bell scored 52 from 68 in England's 1 wicket loss. MJ Guptill and BB McCullum scored 104 from 97 in New Zealand's win by 98 runs.

The opening batting partnership model consists of partnership dot balls, partnerships boundaries and partnership contribution. High values of these predictors are associated with a high probability of dismissal. As such, the model suggests that the high risk strategies opted by DA Warner and AJ Finch and S Dhawan and RG Sharma corresponded to a high combination of 'dot balls-to-balls' ratio, 'boundaries-to-balls' ratio and 'contribution-to-balls' ratio.

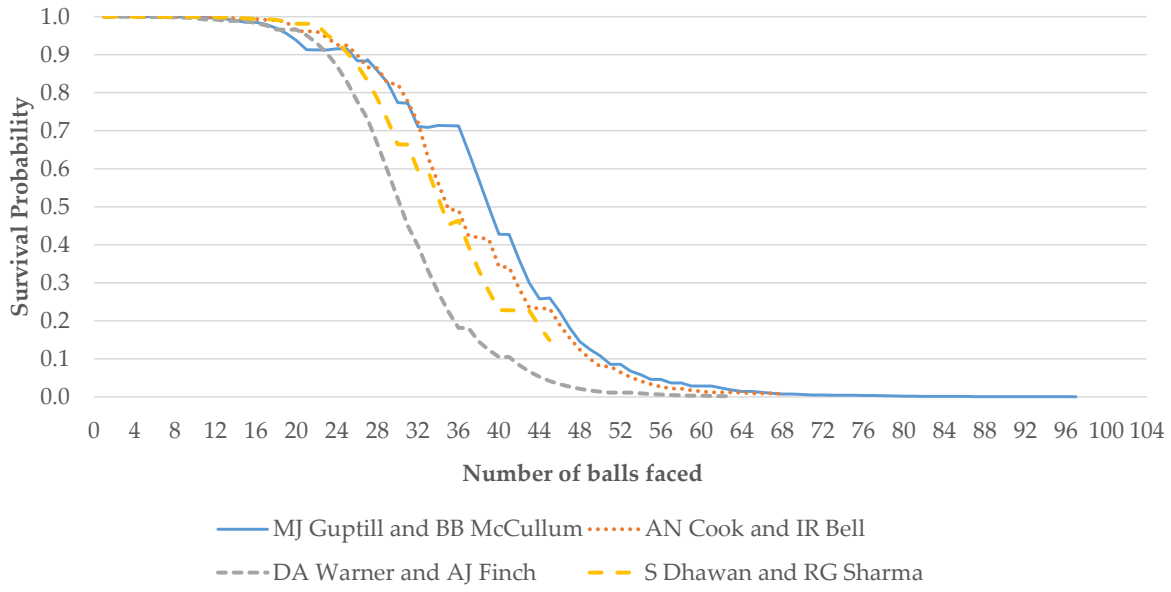

Figure 8.14: Survival probabilities for MJ Guptill and BB McCullum, AN Cook and IR Bell, DA Warner and AJ Finch, S Dhawan and RG Sharma

Figure 8.15 illustrates a selection of the results from the final model associated with fourth wicket partnerships. GD Elliott and LRPL Taylor opted for a high risk strategy on route to 67 from 34. New Zealand won the game by 13 runs. EJG Morgan and J Root, GJ Maxwell and MR Marsh and AD Mathews and KC Sangakkara, as partnerships, all played relatively more conservatively. EJG Morgan and J Root scored 43 from 77 in England's 8 wicket loss. GJ Maxwell and MR Marsh scored 109 from 54 in Australia's win by 198 runs. AD Mathews and KC Sangakkara played the most conservatively after ball 26. They scored 78 from 86 and Sri Lanka lost the game by 5 wickets.

The fourth wicket batting partnership model consists of partnership runs, partnership dot balls

and partnership contribution. High values of these predictors are associated with a high probability of dismissal. As such, the model suggests that the high risk strategies opted by GD Elliott and LRPL Taylor correspond to a high combination of 'runs-to-balls' ratio, 'dot balls-to-balls' ratio and 'contribution-to-balls' ratio.

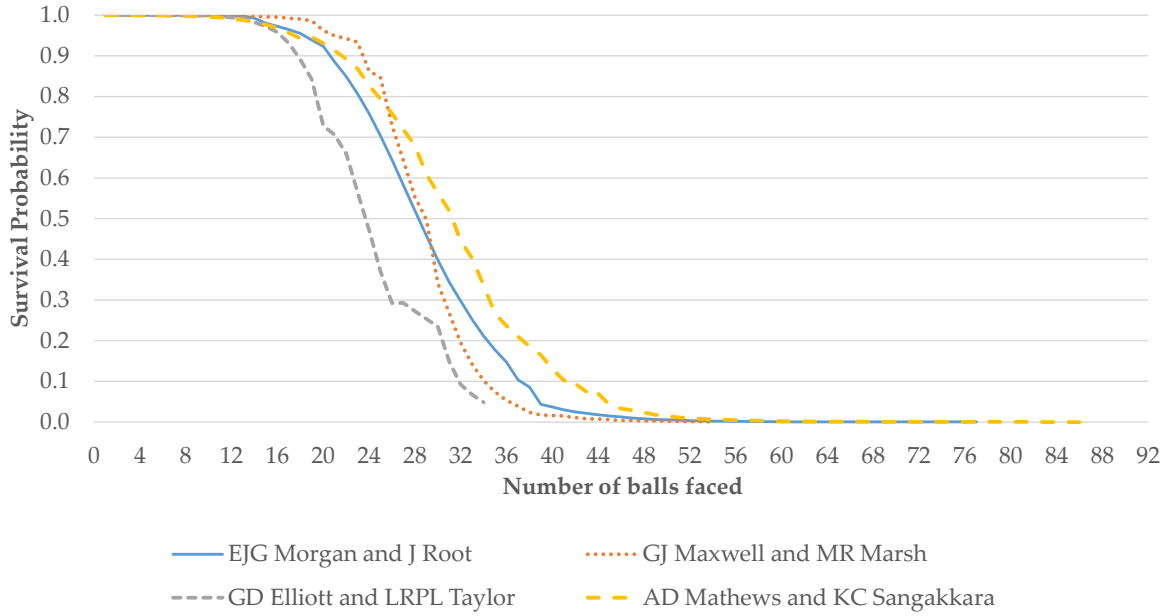

Figure 8.15: Survival probabilities for EJG Morgan and J Root, GJ Maxwell and MR Marsh, GD Elliott and LRPL Taylor, AD Mathews and KC Sangakkara

Figure 8.16 illustrates a selection of the results from the final model associated with tenth wicket partnerships. The survival curves suggest that H Hassan and S Zadran and E Adil and R Ali played a high risk game. H Hassan and S Zadran scored 18 from 14 in Afghanistan's 6 wicket loss. E Adil and R Ali scored 18 from 35 in Pakistan's 6 wicket loss. In contrast, TA Boult and MJ Henry and PVD Chameera and KMDN Kulasekara appear to have played more conservatively. TA Boult and MJ Henry scored 26 from 17 in New Zealand's 6 wicket loss. PVD Chameera and KMDN Kulasekara scored 27 from 19 in Sri Lanka's win by 148 runs.

The tenth wicket batting partnership model consists of partnership runs, partnership less than 2 in 4 and partnership percentage dot balls. High values of these predictors are associated with a high probability of dismissal. As such, the model suggests that the high risk strategies opted by H Hassan and S Zadran and E Adil and R Ali correspond to a high combination of 'runs-to-balls' ratio, 'balls faced in which less than 2 in 4 runs had been scored-to-balls' ratio and

'percentage dot balls-to-balls' ratio.

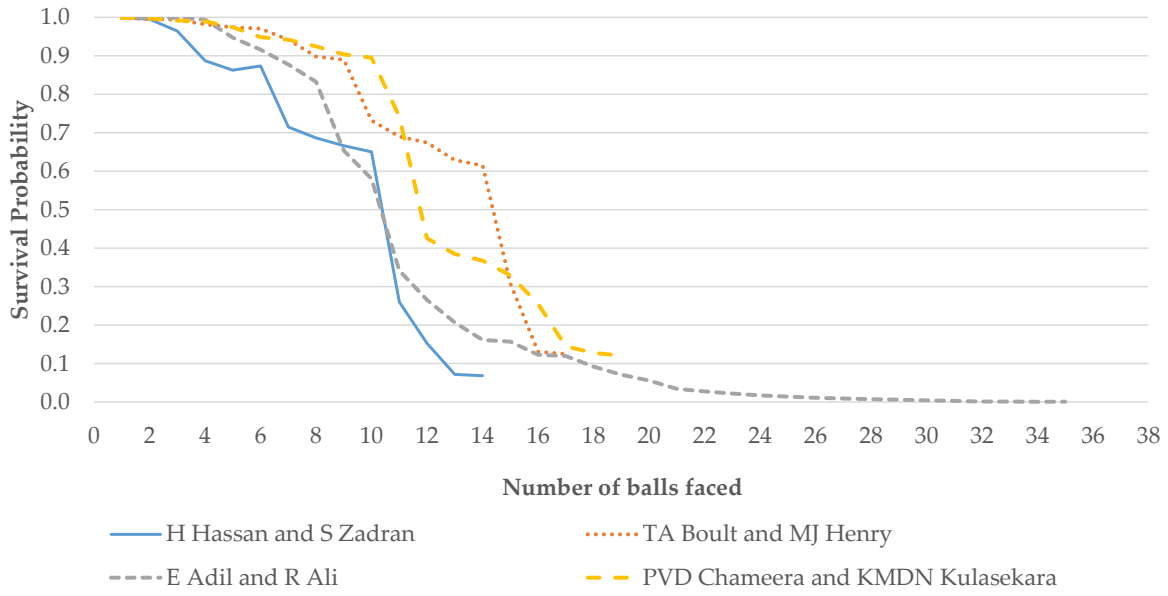

Figure 8.16: Survival probabilities for H Hassan and S Zadran, TA Boult and MJ Henry, E Adil and R Ali, PVD Chameera and KMDN Kulasekara

The results illustrated in Figures 8.13, 8.14, 8.15 and 8.16 consider batting partnerships from a wide variety of cricketing nations, highlighting the ability for this technique to be used to compare partnerships from around the world, which is useful for scouting purposes.

8.5.1 Batting Partnership Performance Measures

The area under each survival curve and the total area under all curves for each partnership were calculated. To account for the differing number of games played by each partnership, the average AUC for each partnership was computed. This was used as a metric for partnership comparison purposes. Another metric, the wins-to-games ratio for each batting partnership was also calculated.

The framework developed in the early work of this thesis, presented in [36], has been leveraged to investigate batting partnerships at a deeper level. Results for each partnership were aggregated into groups of five according to their rank ordering based on the average AUC.

The average number of runs scored and average proportion of team runs scored for each cohort were also calculated. Each rank-ordered cohort consists of five batting partnerships.

8.5.2 Batting Partnership Insights

Table 8.8 summarises the partnerships included in the top cohort associated with wickets one and two.

Top cohort	Wicket 1		Wicket 2	
	P1	**P2**	**P1**	**P2**
1	A Ali	S Aslam	PJ Hughes	SPD Smith
2	A Ali	MB Azam	AK Zazai	NK Mangal
3	DA Miller	Q de Kock	DG Brownlie	KS Williamson
4	A Ali	S Ahmed	A Shehzad	Y Khan
5	IR Bell	AN Cook	MDKJ Perera	HDRL Thirimanne

Table 8.8: Top cohort partnerships wickets one and two

Table 8.9 summarises the partnerships included in the top cohort associated with wickets three and four.

Top cohort	Wicket 3		Wicket 4	
	P1	**P2**	**P1**	**P2**
1	TM Dilshan	HDRL Thirimanne	SMA Priyanjan	KC Sangakkara
2	H Sohail	Misbah-ul-Haq	AB de Villiers	RR Rossouw
3	A Shehzad	A Shafiq	TM Dilshan	M Jayawardene
4	GS Ballance	BA Stokes	LD Chandimal	KC Sangakkara
5	DM Bravo	D Ramdin	A Ali	S Malik

Table 8.9: Top cohort partnerships wickets three and four

Table 8.10 summarises the partnerships included in the top cohort associated with wickets five and six.

Top cohort	Wicket 5		Wicket 6	
	P1	**P2**	**P1**	**P2**
1	F Alam	U Akmal	A Stanikzai	S Shafiqullah
2	TWM Latham	KS Williamson	JF Mooney	SW Poynter
3	KA Pollard	LMP Simmons	LRPL Taylor	MJ Santner
4	N Mangal	S Shenwari	F Berhardien	WD Parnell
5	AD Mathews	SMA Priyanjan	GJ Bailey	MS Wade

Table 8.10: Top cohort partnerships wickets five and six

Table 8.11 summarises the partnerships included in the top cohort associated with wickets seven and eight.

Top cohort	Wicket 7		Wicket 8	
	P1	**P2**	**P1**	**P2**
1	A Zazai	M Ashraf	KJ Abbott	DW Steyn
2	AD Mathews	NLTC Perera	S Ahmed	W Riaz
3	DAJ Bracewell	MJ Santner	NM Coulter-Nile	JP Faulkner
4	R McLaren	DW Steyn	AF Milne	NL McCullum
5	L Ronchi	MJ Santner	NO Miller	R Rampaul

Table 8.11: Top cohort partnerships wickets seven and eight

Table 8.12 summarises the partnerships included in the top cohort associated with wickets nine and ten.

Top cohort	Wicket 9		Wicket 10	
	P1	**P2**	**P1**	**P2**
1	SCJ Broad	AD Hales	TD Chisoro	H Masakadza
2	DJ Willey	CR Woakes	T Muzarabani	J Nyumbu
3	AD Mathews	BKV Prasad	TA Boult	MJ Henry
4	S Ahmed	S Tanvir	PVD Chameera	KMDN Kulasekara
5	MJ McClenaghan	TG Southee	NO Miller	R Rampaul

Table 8.12: Top cohort partnerships wickets nine and ten

Figure 8.17 compares the average AUC with the average number of runs scored for an opening partnership in each cohort.

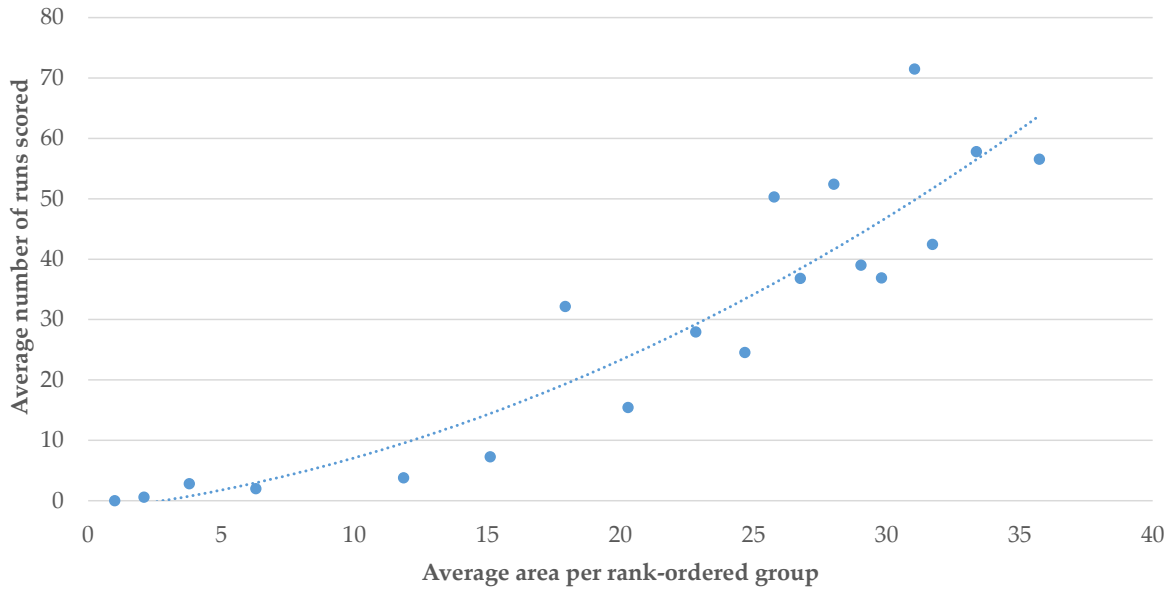

Figure 8.17: Average AUC for rank-ordered cohorts compared with average number of runs scored for opening partnerships for the observed time frame

Figure 8.18 compares the average AUC with the logit of the average proportion of team runs scored for an opening partnership in each cohort.

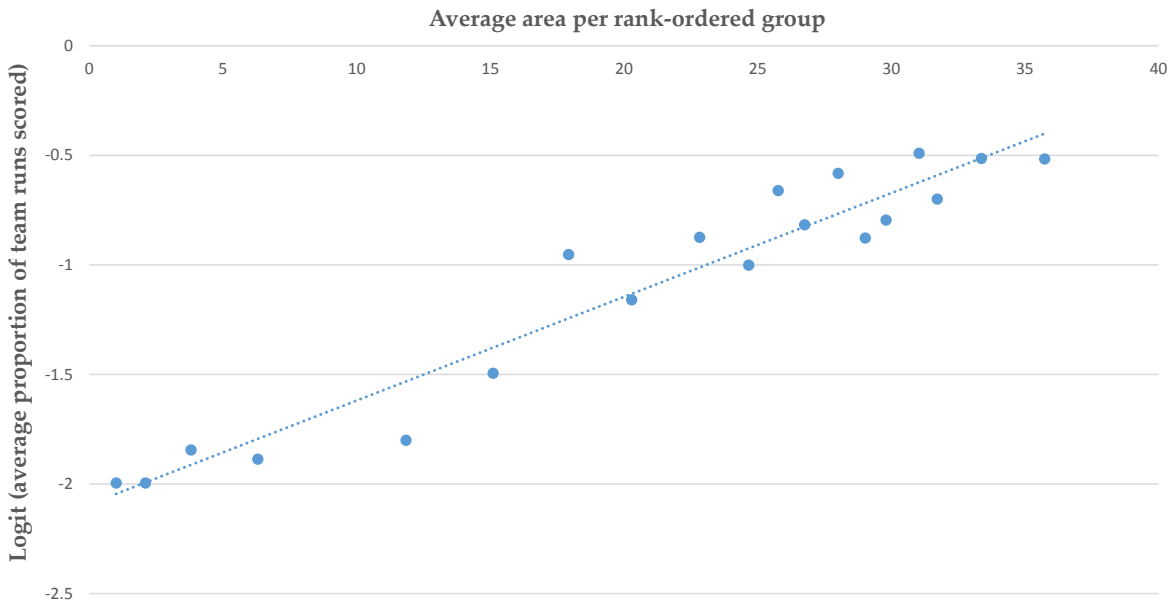

Figure 8.18: Average AUC for rank-ordered cohorts compared with logit of the average proportion of team runs scored for opening partnerships for the observed time frame

As illustrated in Figures 8.17 and 8.18, there are strong relationships between the average AUC and the average number of runs scored and the average AUC and the average proportion of team runs scored for opening partnerships.

For opening partnerships, the correlation between the square root of the average AUC and the square root of the average runs is 0.942. Similarly, the correlation between the square root of the average AUC and the square root of the average proportion of team runs scored is 0.944. As expected, these correlations suggest that more *effective* opening partnerships simultaneously occupy the crease and score runs at a rate that increases the team's chances of winning.

The proportion of games won within each cohort is also explored. The correlation between the average AUC and the logit of the proportion of games won is 0.23. The direction and magnitude of this statistic indicates that the survival time for an opening partnership is positively associated with the partnership team winning the match. However, this relationship is statistically weak and possibly reflects the additional factors involved in winning a ODI game. ODI victory is not solely based on the ability of batsmen to occupy the crease. It is also based on the ability of batsmen to score runs at an appropriate pace, at different stages of the innings. ODI victory also depends on the first innings bowling team defending a total in the second innings.

For non-opening partnerships at subsequent wickets, both the correlation between the square root of the average AUC and the square root of the average runs and the correlation between the square root of the average AUC and the square root of the average proportion of team runs scored are above 0.8. As when investigating opening partnerships, these correlations suggest that *effective* non-opening partnerships simultaneously occupy the crease and produce runs at a rate that increases the team's chance of victory.

For third wicket partnerships, the correlation between the logit of the proportion of team runs scored and the logit of the proportion of games won is 0.31. This suggests a moderate positive association between the proportion of team runs scored by third wicket partnerships and winning. For fifth wicket partnerships, the correlation between the square root of the average runs and the logit of the cohort win percentage is 0.12. This means that the onus is on the top order to optimally score runs.

8.6 Optimisation Procedure

Given the correlations between the four metrics, for each batting partnership at each wicket, the average AUC, average number of runs scored, average proportion of team runs scored and win percentage were combined into an overall measure of *effectiveness*. *Effective* partnerships bat for long periods of time, with a scoring rate that increases the team's chance of defending their total. This measure was used to determine the top three optimal batting partnerships at each wicket for cricketing nations during the time period 26th December 2013 to 14th February 2016. Similarly, the top three optimal batting partnerships at each wicket across all considered nations as a whole were determined. To illustrate these results, Table 8.13 illustrates the top three optimal opening partnerships for New Zealand and Australia. This provides useful insight into working strategies.

	New Zealand		Australia	
	P1	**P2**	**P1**	**P2**
1	DG Brownlie	MJ Guptill	AJ Finch	BJ Haddin
2	BB McCullum	MJ Guptill	AJ Finch	DA Warner
3	DG Brownlie	AP Devcich	AJ Finch	PJ Hughes

Table 8.13: Optimal partnerships wicket one New Zealand and Australia

From Table 8.13, DG Brownlie and MJ Guptill are ranked as the optimal New Zealand opening pair. This is because DG Brownlie is more conservative, while MJ Guptill is more of an aggressive opener. AJ Finch and BJ Haddin are Australia's optimal openers. Similarly, AJ Finch plays aggressively, while BJ Haddin is more conservative.

Table 8.14 illustrates the top three optimal opening partnerships for England and India.

	England		India	
	P1	**P2**	**P1**	**P2**
1	MM Ali	IR Bell	S Dhawan	AM Rahane
2	IR Bell	AN Cook	AM Rahane	M Vijay
3	AD Hales	JJ Roy	S Dhawan	RG Sharma

Table 8.14: Optimal partnerships wicket one England and India

From Table 8.14, MM Ali and IR Bell are ranked as England's optimal opening pair and S Dhawan and AM Rahane are India's optimal openers. IR Bell and S Dhawan play with an aggressive style, while MM Ali and AM Rahane play relatively more conservatively.

Table 8.15 illustrates the top three optimal batting partnerships at wicket two for Sri Lanka and South Africa.

	Sri Lanka		South Africa	
	P1	**P2**	**P1**	**P2**
1	TM Dilshan	KC Sangakkara	HM Amla	AB de Villiers
2	KC Sangakkara	HDRL Thirimanne	HM Amla	RR Rossouw
3	MDKJ Perera	HDRL Thirimanne	Q de Kock	RR Rossouw

Table 8.15: Optimal partnerships wicket two Sri Lanka and South Africa

From Table 8.15, TM Dilshan and KC Sangakkara are ranked as the optimal Sri Lankan pair at wicket two. HM Amla and AB de Villiers are South Africa's optimal pair.

Table 8.16 illustrates the top three optimal batting partnerships at wicket two for West Indies and Bangladesh.

	West Indies		Bangladesh	
	P1	**P2**	**P1**	**P2**
1	CH Gayle	MN Samuels	M Haque	S Rahman
2	DM Bravo	CH Gayle	A Haque	M Haque
3	DM Bravo	DR Smith	M Mahmudullah	T Iqbal

Table 8.16: Optimal partnerships wicket two West Indies and Bangladesh

From Table 8.16, for wicket two, CM Gayle and MN Samuels are ranked as West Indies optimal pair and M Haque and S Rahman are Bangladesh's optimal pair.

Table 8.17 illustrates the top three optimal batting partnerships at wicket three for Zimbabwe and Afghanistan.

	Zimbabwe		Afghanistan	
	P1	**P2**	**P1**	**P2**
1	C Ervine	H Masakadza	A Stanikzai	S Shenwari
2	BB Chari	SC Williams	A Zazai	M Nabi
3	H Masakadza	R Mutumbami	G Naib	N Jamal

Table 8.17: Optimal partnerships wicket three Zimbabwe and Afghanistan

From Table 8.17, for wicket three, C Ervine and H Masakadza are ranked as Zimbabwe's optimal partnership and A Stanikzai and S Shenwari are Afghanistan's optimal partnership.

Table 8.18 illustrates the top three optimal batting partnerships at wicket three for Scotland and Ireland.

	Scotland		Ireland	
	P1	**P2**	**P1**	**P2**
1	KJ Coetzer	MW Machan	A Balbirnie	EC Joyce
2	HJW Gardiner	PL Mommsen	NJ O'Brien	WTS Porterfield
3	RD Berrington	PL Mommsen	NJ O'Brien	PR Stirling

Table 8.18: Optimal partnerships wicket three Scotland and Ireland

From Table 8.18, for wicket three, KJ Coetzer and MW Machan are ranked as Scotland's optimal partnership and A Balbirnie and EC Joyce are Ireland's optimal partnership.

Table 8.19 illustrates the optimal batting partnership at each wicket across, all considered nations as a whole. That is, at each wicket, the partnership that would have maximised their respective team's final score and chances of winning by the largest amount.

Wicket	Batting partnership		Country
	P1	P2	
1	**A Ali**	**S Aslam**	Pakistan
2	**HM Amla**	**AB de Villiers**	South Africa
3	**H Sohail**	S Malik	Pakistan
4	**DJ Bravo**	FH Edwards	West Indies
5	**KM Jadhav**	MK Pandey	India
6	**STR Binny**	**AT Rayudu**	India
7	J Buttler	**AU Rashid**	England
8	**AG Cremer**	H Masakadza	Zimbabwe
9	**T Panyangara**	S Raza	Zimbabwe
10	**TA Boult**	L Ronchi	New Zealand

Table 8.19: Optimal partnerships nationwide

Those batsmen that are in the incoming batting position are emphasised in bold. For example, TA Boult bats at position 11. A Ali and S Aslam are ranked as the optimal opening pair, HM Amla and AB de Villiers are ranked as the optimal partnership at wicket two and TA Boult is ranked as the optimal number 11 batsman. However, with TA Boult ranked as the number one bowler in February 2017, that is of greater importance.

This research has successfully derived the optimal batting partnership strategy at each wicket across all considered nations. That is, those partnerships that would have maximised the score and winning chances of their respective teams by the largest amount. These partnerships were particularly *effective* at occupying the crease, while scoring lots of runs, contributing highly to their team and increasing their team's chances of winning.

This research investigated the optimal partnership strategy in the New Zealand team at a deeper level. Table 8.20 illustrates the optimal batting partnership strategy at each wicket for New Zealand.

Wicket	Batting partnership	
	P1	**P2**
1	**MJ Guptill**	**DG Brownlie**
2	DG Brownlie	**KS Williamson**
3	BB McCullum	LRPL Taylor
4	CJ Anderson	JD Ryder
5	TWM Latham	KS Williamson
6	TWM Latham	L Ronchi
7	HM Nichols	MJ Santner
8	AF Milne	MJ Santner
9	MJ Henry	MJ McClenaghan
10	**TA Boult**	L Ronchi

Table 8.20: Optimal partnership strategy New Zealand

Figure 8.19 shows a decision tree, constructed to illustrate the optimal batting partnership strategy for New Zealand at each wicket, depending on which batsman in the partnership is dismissed.

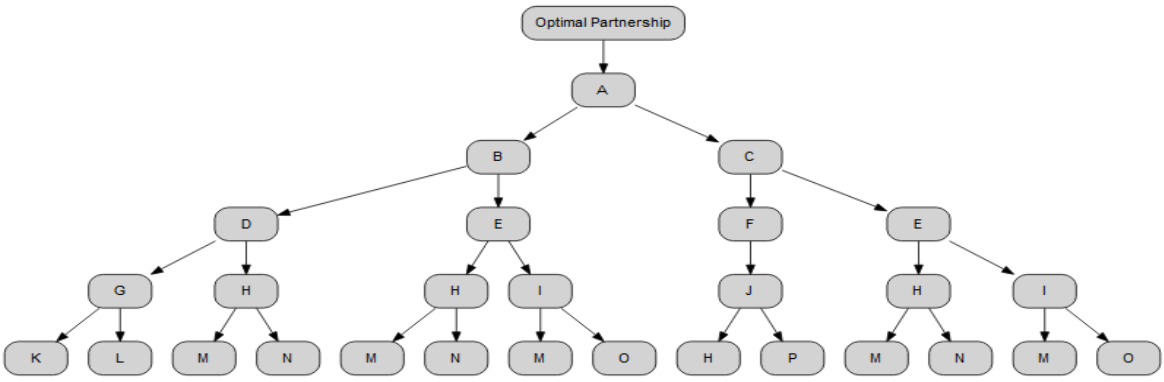

Figure 8.19: Decision tree illustrating optimal New Zealand batting partnership strategy

Table 8.21 describes the partnerships, denoted as letters, in the decision tree in Figure 8.19.

Diagram notation	Batting partnership	
	P1	**P2**
A	MJ Guptill	DG Brownlie
B	MJ Guptill	KS Williamson
C	DG Brownlie	KS Williamson
D	MJ Guptill	LRPL Taylor
E	KS Williamson	LRPL Taylor
F	DG Brownlie	BB McCullum
G	MJ Guptill	CJ Anderson
H	LRPL Taylor	GD Elliott
I	KS Williamson	GD Elliott
J	BB McCullum	LRPL Taylor
K	MJ Guptill	GD Elliott
L	CJ Anderson	HM Nichols
M	GD Elliott	CJ Anderson
N	LRPL Taylor	JDS Neesham
O	KS Williamson	TWM Latham
P	BB McCullum	CJ Anderson

Table 8.21: Partnership correspondence

Figure 8.19 shows that MJ Guptill and DG Brownlie are the optimal opening partnership. MJ Guptill and DG Brownlie had a 100% win percentage when batting together as openers. In the event that either of these opening batsmen are dismissed, KS Williamson is suggested as the optimal batsman to bat in position three. DG Brownlie and KS Williamson had a 100% win percentage when batting together at second wicket. MJ Guptill and KS Williamson had a 70% win percentage when batting together at second wicket. Together with the fact that KS Williamson is the current New Zealand captain and was ranked as number five in the ICC ODI player rankings on 8th February 2016, these results suggest that the approach developed to optimise batting partnership strategy is valid.

In addition, in [4], KS Williamson was acknowledged as “the most important player to his

team in the world" with the biggest contribution to success in test cricket, spanning the last three years. KS Williamson is the third top century scorer of any batsman over the last three years, while he is only one of two players to score over 30% of his team's hundreds. Further, KS Williamson has scored 3011 runs in the last three years, considerably higher than the next best New Zealand batsman, TWM Latham, with 2031 [4]. This further highlights the validity of the optimisation procedure.

Based on the optimal partnership strategy illustrated in Figure 8.19, an optimal top six New Zealand batting line up is suggested in Table 8.22.

Batting position	**Batsman**
1	MJ Guptill
2	DG Brownlie
3	KS Williamson
4	LRPL Taylor
5	GD Elliott
6	JDS Neesham

Table 8.22: Optimal New Zealand top six batsmen

8.7 Optimal Batting Partnership Strategy by Risk

In [91], the authors suggested that in one-day cricket there are two extreme batting strategies from which intermediate strategies can be obtained. The first is an aggressive high risk strategy. This is where a batsman attempts to score a high number of runs with a greater risk of dismissal. The second is a conservative low risk strategy. This is where a batsman attempts to preserve wickets by scoring at a lower rate. This is likely to involve rotating the strike with singles rather than hitting boundaries. In addition to the ability of batsmen to occupy the crease, ODI victory depends on the ability of batsmen to score runs at an appropriate pace at different stages of the innings. This is a possible reflection of the weak relationships between the average AUC and win percentage, in Section 8.1. Playing riskily to accumulate quickly may be the best strategy, especially towards the end of the innings if there are wickets in hand. Whereas,

playing cautiously may be more sensible earlier on.

As such, of interest as part of this research was the optimal partnership strategy depending on the style of play adopted. For each batting partnership, the average AUC as a ratio of the combination of the average AUC, average number of runs scored and proportion of team runs scored was calculated. This ratio was used to determine the level of risk taken by the partnership. The larger the ratio, the larger the average AUC compared to the number of runs scored and proportion of team runs scored. This suggests that a large ratio is indicative of an overall low risk strategy opted by the partnership. The smaller the ratio, the smaller the average AUC compared to the number of runs scored and proportion of team runs scored. This suggests that a small ratio is indicative of an overall high risk strategy opted by the partnership.

The risk ratio was used to rank each partnership into a risk strategy category. Ten risk strategy categories were created. For low risk categories, these were defined as extreme low risk, upper low risk, moderate low risk, lower low risk and minimal low risk. For high risk categories, these were defined as extreme high risk, upper high risk, moderate high risk, lower high risk and minimal high risk. In each risk category, the overall *effectiveness* measure was used to rank the partnerships and determine the optimal partnership at each wicket.

Table 8.23 illustrates the optimal batting partnership for an extreme low risk strategy at each wicket across all considered nations. That is, at each wicket, the partnership that would have maximised their respective team's final score and chances of winning by the largest amount, had the team opted for an extreme low risk strategy.

Wicket	Batting partnership		Country
	P1	P2	
1	DA Warner	SR Watson	Australia
2	NK Mangal	U Ghani	Afghanistan
3	AK Zazai	N Jamal	Afghanistan
4	M Jayawardene	KC Sangakkara	Sri Lanka
5	M Jayawardene	SMA Priyanjan	Sri Lanka
6	JL Carter	DJ Sammy	West Indies
7	RD Berrington	J Davey	Scotland
8	E Adil	W Riaz	Pakistan
9	M Tauqir	N Aziz	United Arab Emirates
10	M Haq	I Wardlaw	Scotland

Table 8.23: Optimal extreme low risk partnerships nationwide

Table 8.24 illustrates the optimal batting partnership for an upper low risk strategy at each wicket, across all considered nations.

Wicket	Batting partnership		Country
	P1	P2	
1	TM Dilshan	D Karunaratne	Sri Lanka
2	I Kayes	T Iqbal	Bangladesh
3	MJ Clarke	SE Marsh	Australia
4	A Haque	S Al Hasan	Bangladesh
5	PJ Hughes	SPD Smith	Australia
6	A Stanikzai	M Shafiqullah	Afghanistan
7	R McLaren	DW Steyn	South Africa
8	S Binny	A Mishra	India
9	A Hamza	D Zadran	Afghanistan
10	NO Miller	R Rampaul	West Indies

Table 8.24: Optimal upper low risk partnerships nationwide

Table 8.25 illustrates the optimal batting partnership for a moderate low risk strategy at each wicket, across all considered nations.

Wicket	**Batting partnership**		**Country**
	P1	**P2**	
1	A Balbirnie	J Anderson	Ireland
2	U Tharanga	L Thirimanne	Sri Lanka
3	A Shafiq	Misbah-ul-Haq	Pakistan
4	DA Miller	Q de Kock	South Africa
5	A Haque	M Mahmudallah	Bangladesh
6	N Zadran	S Shenwari	Afghanistan
7	CJ Anderson	MJ Santner	New Zealand
8	KJ Abbott	DW Steyn	South Africa
9	MJ McClenaghan	TG Southee	New Zealand
10	L Gamage	N Kulasekara	Sri Lanka

Table 8.25: Optimal moderate low risk partnerships nationwide

Table 8.26 illustrates the optimal batting partnership for a lower low risk strategy at each wicket, across all considered nations.

Wicket	Batting partnership		Country
	P1	P2	
1	J Ahmadi	N Mangal	Afghanistan
2	MJ Clarke	PJ Hughes	Australia
3	G Naib	N Jamal	Afghanistan
4	TM Dilshan	M Jayawardene	Sri Lanka
5	F Berhardien	JP Duminy	South Africa
6	MM Ali	AD Hales	England
7	DAJ Bracewell	MJ Santner	New Zealand
8	JWA Taylor	CR Woakes	England
9	N Kulasekara	S Prasanna	Sri Lanka
10	E Adil	M Irfan	Pakistan

Table 8.26: Optimal lower low risk partnerships nationwide

Table 8.27 illustrates the optimal batting partnership for a minimal low risk strategy at each wicket, across all considered nations.

Wicket	Batting partnership		Country
	P1	P2	
1	A Ali	S Ahmed	Pakistan
2	K Sadiq	N Zadran	Afghanistan
3	AB de Villiers	K van Wyk	South Africa
4	NR Waller	SC Williams	Zimbabwe
5	DA Miller	F du Plessis	South Africa
6	L Ronchi	MJ Santner	New Zealand
7	L Ronchi	MJ Santner	New Zealand
8	A Raza	S Haider	United Arab Emirates
9	NM Coulter-Nile	MR Marsh	Australia
10	T Muzarabani	J Nyumbu	Zimbabwe

Table 8.27: Optimal minimal low risk partnerships nationwide

Table 8.28 illustrates the optimal batting partnership for a minimal high risk strategy at each wicket across all considered nations. That is, at each wicket, the partnership that would have maximised their respective team's final score and chances of winning by the largest amount, had the team opted for a minimal high risk strategy.

Wicket	**Batting partnership**		**Country**
	P1	**P2**	
1	A Ali	B Azam	Pakistan
2	AJ Finch	SE Marsh	Australia
3	TM Dilshan	L Thirimanne	Sri Lanka
4	IR Bell	EJG Morgan	England
5	WTS Porterfield	GC Wilson	Ireland
6	GS Ballance	BA Stokes	England
7	S Maqsood	W Riaz	Pakistan
8	AF Milne	NL McCullum	New Zealand
9	S Ahmed	S Tanvir	Pakistan
10	TS Chisoro	H Masakadza	Zimbabwe

Table 8.28: Optimal minimal high risk partnerships nationwide

Table 8.29 illustrates the optimal batting partnership for a lower high risk strategy at each wicket, across all considered nations.

Wicket	Batting partnership		Country
	P1	P2	
1	TM Dilshan	L Thirimanne	Sri Lanka
2	MDKJ Perera	L Thirimanne	Sri Lanka
3	H Sohail	Misbah-ul-Haq	Pakistan
4	DM Bravo	D Ramdin	West Indies
5	N Mangal	S Shenwari	Afghanistan
6	LRPL Taylor	MJ Santner	New Zealand
7	A Zazai	M Ashraf	Afghanistan
8	M Haque	R Hossain	Bangladesh
9	SCJ Broad	AD Hales	England
10	D Chameera	S Lakmal	Sri Lanka

Table 8.29: Optimal lower high risk partnerships nationwide

Table 8.30 illustrates the optimal batting partnership for a moderate high risk strategy at each wicket, across all considered nations.

Wicket	Batting partnership		Country
	P1	P2	
1	IR Bell	AN Cook	England
2	DG Brownlie	KS Williamson	New Zealand
3	A Shehzad	A Shafiq	Pakistan
4	LD Chandimal	KC Sangakkara	Sri Lanka
5	TWM Latham	KS Williamson	New Zealand
6	GJ Maxwell	MR Marsh	Australia
7	LD Chandimal	AD Mathews	Sri Lanka
8	NM Coulter-Nile	JP Faulkner	Australia
9	DJ Willey	CR Woakes	England
10	D Chameera	N Kulasekara	Sri Lanka

Table 8.30: Optimal moderate high risk partnerships nationwide

Table 8.31 illustrates the optimal batting partnership for an upper high risk strategy at each wicket, across all considered nations.

Wicket	**Batting partnership**		**Country**
	P1	**P2**	
1	A Ali	S Aslam	Pakistan
2	GJ Bailey	SPD Smith	Australia
3	CR Ervine	H Masakadza	Zimbabwe
4	AB de Villiers	JP Duminy	South Africa
5	JE Root	BA Stokes	England
6	F Berhardien	WD Parnell	South Africa
7	A Ali	U Akmal	Pakistan
8	CJ Jordan	EJG Morgan	England
9	AD Mathews	D Prasad	Sri Lanka
10	AM Guruge	S Anwar	United Arab Emirates

Table 8.31: Optimal upper high risk partnerships nationwide

Table 8.32 illustrates the optimal batting partnership for an extreme high risk strategy at each wicket, across all considered nations.

Wicket	Batting partnership		Country
	P1	P2	
1	A Ali	M Hafeez	Pakistan
2	HM Amla	AB de Villiers	South Africa
3	H Sohail	S Malik	Pakistan
4	DJ Bravo	FH Edwards	West Indies
5	KM Jadhav	MK Pandey	India
6	STR Binny	AT Rayudu	India
7	JC Buttler	AU Rashid	England
8	AG Cremer	H Masakadza	Zimbabwe
9	T Panyangara	S Raza	Zimbabwe
10	TA Boult	L Ronchi	New Zealand

Table 8.32: Optimal extreme high risk partnerships nationwide

8.8 Case Study

On 4th December 2016, New Zealand played Australia in a ODI and lost by 68 runs. Former Black caps all-rounder and Auckland A cricket coach, AR Adams, criticised the New Zealand coaching staff for their batting order changes in that game [15].

AR Adams questioned the choice of JDS Neesham as number four. C Munro batted as number six and C de Grandhomme batted as number eight. Given how successfully Aucklanders, C Munro and C de Grandhomme, batted in domestic cricket in 2016, AR Adams suggested that C Munro and C de Grandhomme should have been played in positions four and five, followed by BJ Watling and JDS Neesham [15].

The objective of this case study was to determine the optimal New Zealand batting order for the game against Australia to assess whether it aligns with AR Adams's suggestions, and demonstrate the practical application of this research.

The New Zealand batting order against Australia is illustrated in Table 8.33.

Batting position	Batsman
1	MJ Guptill
2	TWM Latham
3	KS Williamson
4	JDS Neesham
5	BJ Watling
6	C Munro
7	MJ Santner
8	C de Grandhomme
9	MJ Henry
10	LH Ferguson
11	TA Boult

Table 8.33: New Zealand order against Australia 4th December

As part of this research, the final models were fitted to data from the first innings of limited overs cricket games. In the ODI between Australia and New Zealand, New Zealand batted in the second innings. To account for this, the models were applied to data from the most recent ODI game prior to 4th December 2016, in which New Zealand batted in the first innings. New Zealand's opponents in this game were India, with the game contested on 26th October 2016. The intent was to use the performance of batsmen in the ODI against India as an indication of how these batsmen would have performed in the ODI against Australia. As such, the optimal New Zealand batting order against India could be used as an indicator to suggest the order that would have likely optimised the scoring rates and chances of winning against Australia.

Each batsman was assigned a measure of *effectiveness* based on the average AUC, total number of runs scored, proportion of team runs scored and strike rate.

8.8.1 Bootstrapping

Bootstrapping is a technique used to re-sample data without replacement and allows estimation of the sampling distribution of a statistic [73]. As the analysis in this case study was based on

one game, bootstrapping was used to generate 1000 bootstrap samples of batsman *effectiveness*.

To determine the optimal batting order, the process was repeated with a different batsman removed each time. Table 8.34 illustrates the New Zealand batting order used in the ODI game against India.

Batting position	**Batsman**
1	MJ Guptill
2	TWM Latham
3	KS Williamson
4	LRPL Taylor
5	JDS Neesham
6	BJ Watling
7	AP Devcich
8	MJ Santner
9	TG Southee

Table 8.34: New Zealand batting order against India 26th October

Each bootstrapped sample consisted of the *effectiveness* of each batsman from Table 8.34, with one batsman removed. In this particular game, New Zealand only used nine batsmen. As such, each bootstrapped sample contained eight observations. The bootstrapping procedure was repeated to give 1000 samples. The mean *effectiveness* for the team was calculated for each sample. The mean of those means was used as a rating of *effectiveness* for the team with the batsman excluded from analysis. The rating was then used to determine where that batsman was positioned in the optimal batting order. The optimal batting order is the order that would have maximised the team's final score and chances of winning. The smaller the rating associated with each batsman, the less *effective* the team would have been without the batsman and the higher the batsman was optimally positioned.

8.8.2 Optimal New Zealand Order Against India 26th October 2016

Table 8.35 illustrates New Zealand batsmen based on their position in the optimal order, compared with where they were actually positioned in the ODI against India.

Position	Optimal batsman	Actual batsman
1	MJ Guptill	MJ Guptill
2	TWM Latham	TWM Latham
3	KS Williamson	KS Williamson
4	LRPL Taylor	LRPL Taylor
5	**BJ Watling**	**JDS Neesham**
6	**AP Devcich**	**BJ Watling**
7	**JDS Neesham**	**AP Devcich**
8	MJ Santner	MJ Santner
9	TG Southee	TG Southee

Table 8.35: Suggested New Zealand order compared with actual New Zealand order against India 26th October

Table 8.35 illustrates that optimally, JDS Neesham should have been positioned as number seven, behind both BJ Watling and AP Devcich.

8.8.3 Optimal New Zealand Order Against Australia 4th December 2016

The optimal order suggested in Table 8.35 was used to optimally position New Zealand batsmen in the primary ODI of interest against Australia.

Given that C Munro and C de Grandhomme were not involved in the ODI between India and New Zealand or any previous ODI contested in 2016, a rating of *effectiveness* from the optimisation procedure could not be derived for these batsmen.

As such, a different approach was taken to determine the optimal position for C Munro and C de Grandhomme, relative to BJ Watling and JDS Neesham. The Plunket Shield is New Zealand's domestic first-class cricket competition. The 2016-2017 season is the most recent

competition in which C Munro, C de Grandhomme and JDS Neesham had all batted. As such, in order to compare these batsmen, this research investigated their domestic performances in the Plunket Shield. However, the Plunket Shield does not categorise as limited overs cricket. Consequently, the final models and optimisation procedure could not be applied to games from this competition. The optimisation procedure used to rate the New Zealand batsmen against India accounted for AUC, total runs scored, proportion of team runs scored and strike rate. Given this, batting averages in the 2016-2017 Plunket Shield competition were used to compare the *effectiveness* of C Munro and C de Grandhomme with that of BJ Watling and JDS Neesham. The intent was to compare the most recent performance of these batsmen prior to the ODI between Australia and New Zealand, within the same competition for the same period. This is likely to have been indicative of where to position these batsmen, relative to each other, in the optimal New Zealand order against Australia. Despite being from different competitions and formats, these performances are an adequate proxy of batting performance, primarily due to the timeliness of observations.

During the period of the 2016-2017 Plunket Shield season prior to New Zealand's ODI on 4th December, C Munro and C de Grandhomme were scoring at a considerably higher rate compared with JDS Neesham. C Munro was averaging 84.50, while C de Grandhomme was averaging 54.00 [16]. JDS Neesham averaged 8.00 in the same competition for the same period [10]. BJ Watling had not played any domestic cricket during the 2016-2017 season prior to 4th December 2016. However, in ODIs, BJ Watling was averaging 26.09 compared with 21.75 for JDS Neesham [16]. As such, comparative ratings of *effectiveness* for C Munro and C de Grandhomme would likely have been higher than the actual ratings calculated for BJ Watling and JDS Neesham, based on the optimisation procedure. This supports AR Adams's suggestion to play C Munro and C de Grandhomme as number four and five respectively, with BJ Watling and JDS Neesham as number six and seven respectively.

Similarly, lower order batsmen, MJ Henry, LH Ferguson and TA Boult did not bat in the ODI between India and New Zealand on 26th October. With the exception of LH Ferguson, the other eight batsmen from Table 8.35 batted in the previous game between India and New Zealand on 23rd October 2016. As such, the final models were applied to this game and a measure of *effectiveness* was assigned to each batsman, as discussed in Section 8.8. The bootstrapping

procedure, discussed in Section 8.8.1, was applied to rate each batsman. The rating was then used to position each batsman in the optimal order.

LH Ferguson completed his ODI debut against Australia. Given his lack of ODI experience, LH Ferguson is positioned as number eleven in the optimal order.

Table 8.36 illustrates the New Zealand order that would have likely optimised the scoring rates and chances of winning against Australia. This is compared with the actual batting order.

Position	Optimal batsman	Actual batsman
1	MJ Guptill	MJ Guptill
2	TWM Latham	TWM Latham
3	KS Williamson	KS Williamson
4	**C Munro**	**JDS Neesham**
5	**C de Grandhomme**	**BJ Watling**
6	BJ Watling	C Munro
7	JDS Neesham	MJ Santner
8	MJ Henry	C de Grandhomme
9	MJ Santner	MJ Henry
10	TA Boult	LH Ferguson
11	LH Ferguson	TA Boult

Table 8.36: Optimal New Zealand order compared with actual New Zealand order against Australia 4th December

Based on the findings in Table 8.36, New Zealand appear to have utilised a suboptimal order with JDS Neesham and BJ Watling batting at four and five respectively. Given the circumstances, C Munro and C de Grandhomme were quantified as a more optimal order.

In the ODI between New Zealand and Australia, BJ Watling struggled, only scoring 6 from 13 before being dismissed. JDS Neesham scored 34 from 62. C Munro recorded the highest score, 49 from 59, of the three batsmen. The optimal batting order is consistent with this, suggesting C Munro should have batted before BJ Watling and JDS Neesham, as number four. This would have improved New Zealand's chances of winning, based on previously observed individual

batting strategies.

8.9 Chapter Remarks

This chapter has applied Cox proportional hazard modelling, censoring and ridge regression techniques to investigate the survival properties of individual batsmen and batting partnerships in limited overs cricket games. This framework has been used to successfully optimise batting partnership strategy across global cricket nations. This chapter addressed the first four limitations of previous work. Namely, performance metrics that have a significant effect on the probability of a batsman dismissal were identified. This chapter considered the effects of within-game events on batsman survival with success. Cox models were applied to generate survival probabilities for opening and non-opening batsmen and batting partnerships. These were used to successfully optimise batting partnership strategy. The final limitation outlined in Section 3.3 in Chapter 3 was associated with a lack of final model validation. This will be addressed in Chapter 9.

The optimal opening partnership consists of Pakistani openers, A Ali and S Aslam. The optimal wicket two batting partnership consists of South African pair, HM Amla and AB de Villiers. HM Amla and AB de Villiers had a 100% winning rate when batting together as a second wicket partnership. The optimal wicket six partnership consists of Indian pair, STR Binny and AT Rayudu.

New Zealand captain, KS Williamson is the optimal batsman in position three, irrespective of which opener is dismissed. At the time of New Zealand's loss against Australia on 4th December 2016, KS Williamson was ranked as number five in the ICC ODI rankings. Additionally, in [4], KS Williamson was acknowledged as "the most important player to his team in the world".

Reviewing New Zealand's loss against Australia on 4th December 2016, indicates a suboptimal order was used, with JDS Neesham and BJ Watling batting at four and five respectively. Given the circumstances, C Munro and C de Grandhomme were quantified as a more optimal order. This supported the batting order suggestions made by former Black caps all-rounder and

Auckland A coach, AR Adams.

To ensure complete case analysis, performance from other competitions was used. This demonstrated the wider applicability and usefulness of this methodology for scouting and selection purposes.

The work presented throughout this chapter, in [35], was peer-reviewed and published in Mathsport International 2017 Conference proceedings.

Chapter 9

Model Validation

This chapter discusses the application of the final models to IPL data and draws conclusions based on a comparison with the model application to ODI data. These comparisons highlight the differences in performance between ODI and IPL batsmen and partnerships. These differences are practical, suggesting the associated models are valid.

9.1 Model Validation Methodology

This research applied the data extraction and manipulation methodology discussed in Chapter 4 to data from the 2016 IPL season. Games in this league were contested between 9th April and 29th May 2016.

Each final model was initially fitted to the corresponding IPL data. For example, the final model associated with middle order ODI batsmen was applied to data associated with middle order IPL batsmen. Using methodology discussed in Section 8.2 in Chapter 8, survival probabilities for IPL individual batsmen were generated and plotted against survival probabilities for ODI individual batsmen. The Area Under the Curve (AUC) was calculated as a metric to determine which models generated survival probabilities characterising the largest difference between IPL batsmen and ODI batsmen.

9.2 Individual Batsman Model Validation Results

A selection of batsmen across a variety of games was used to illustrate the results of the process used to validate the final models associated with individual batsmen. Table 9.1 describes each batsman and the respective game they played. Figures 9.1, 9.2 and 9.3 were constructed to illustrate the differences between ball-by-ball survival probabilities for ODI batsmen and IPL batsmen.

Batsman	Match	Date played
IR Bell	Australia v England	24/01/2014
G Gambhir	Kolkata Knight Riders v Mumbai Indians	13/04/2016
KS Williamson	New Zealand v Sri Lanka	14/02/2015
SV Samson	Delhi Daredevils v Mumbai Indians	23/04/2016
CJ Anderson	New Zealand v Australia	06/02/2016
P Ojha	Sunrisers Hyderabad v Kolkata Knight Riders	16/04/2015
H Singh	Mumbai Indians v Rising Pune Supergiants	09/04/2016
LM Jongwe	Zimbabwe v Afghanistan	22/10/2015
P Sahu	Delhi Daredevils v Kings XI Punjab	15/04/2016
AM Guruge	India v United Arab Emirates	28/02/2015
SPD Smith	Rising Pune Supergiants v Gujarat Lions	15/04/2016
EJG Morgan	Australia v England	19/01/2014

Table 9.1: Batsman match information

Figure 9.1 was constructed to illustrate the differences between ball-by-ball survival probabilities for ODI and IPL openers. IR Bell had higher survival probabilities than G Gambhir. IR Bell opted for a conservative approach, on route to 55 from 52 (9x4 runs, 0x6 runs). G Gambhir opted for a relatively higher risk strategy, on route to 64 from 52 (4x4 runs, 1x6 runs).

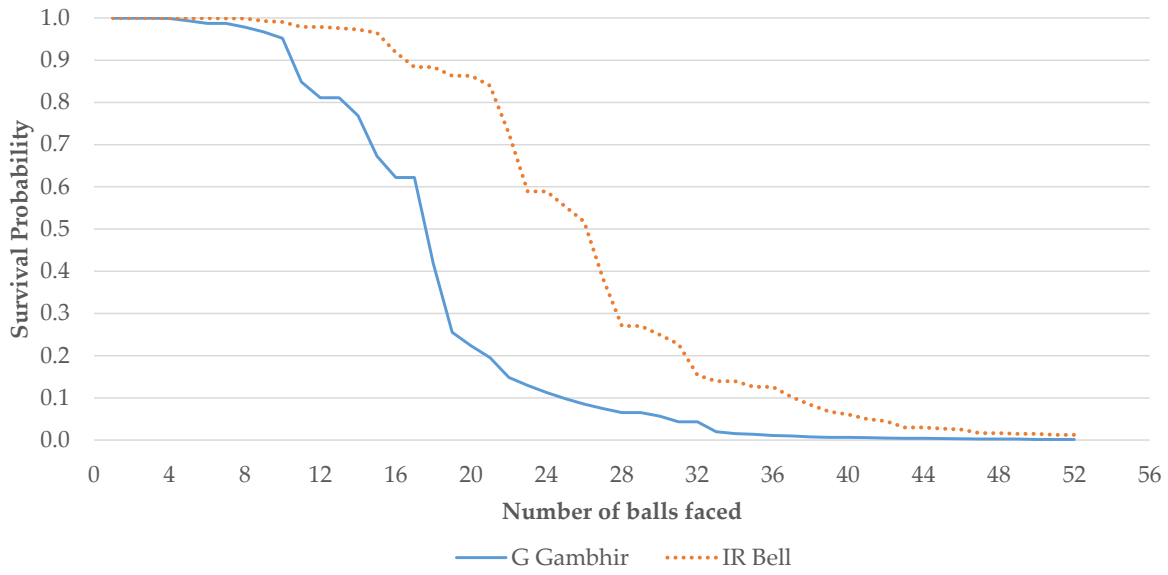

Figure 9.1: Survival probabilities for IR Bell (ODI) and G Gambhir (IPL)

Figure 9.2 was constructed to illustrate the differences between ball-by-ball survival probabilities for ODI and IPL top order batsmen. KS Williamson had higher survival probabilities than SV Samson. KS Williamson opted for a conservative approach, on route to 57 from 65 (5x4 runs, 1x6 runs). SV Samson played a relatively higher risk game, on route to 60 from 48 (4x4 runs, 2x6 runs).

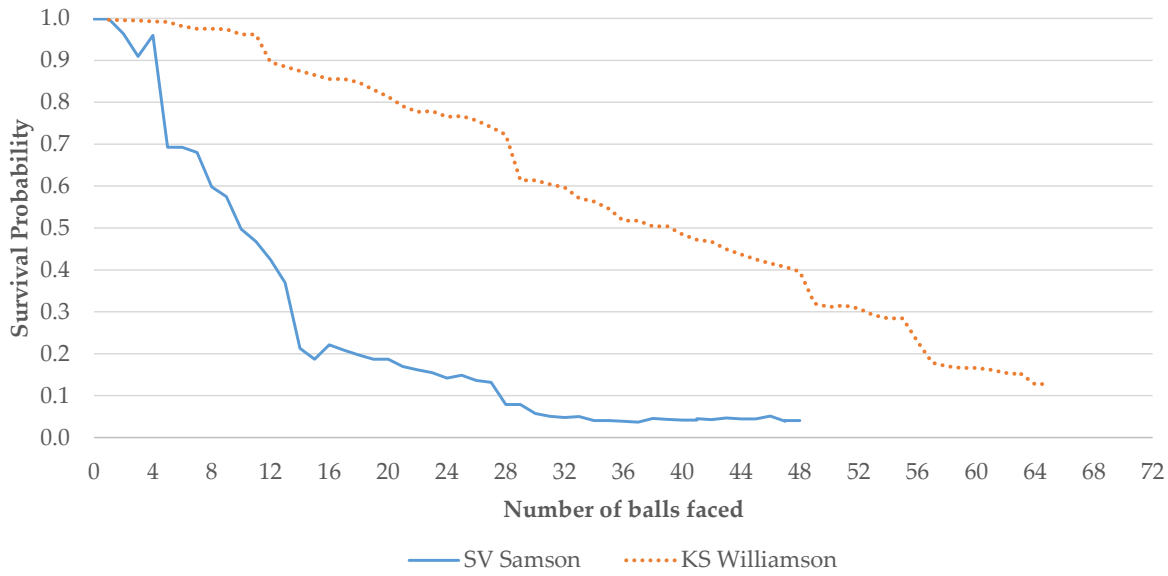

Figure 9.2: Survival probabilities for KS Williamson (ODI) and SV Samson (IPL)

Figure 9.3 was constructed to illustrate the differences between ball-by-ball survival probabilities for ODI and IPL middle order batsmen. CJ Anderson had higher survival probabilities than P Ojha. CJ Anderson opted for a conservative approach, on route to 16 from 28 (0x4 runs, 1x6 runs). P Ojha played a relatively higher risk game, on route to 37 from 28 (2x4 runs, 2x6 runs).

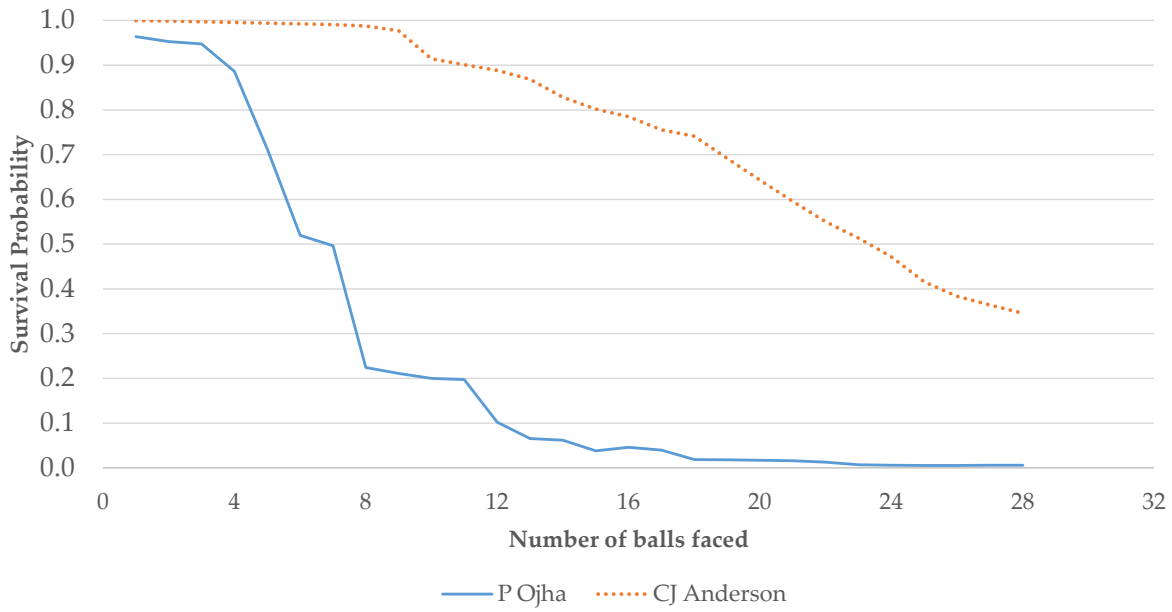

Figure 9.3: Survival probabilities for CJ Anderson (ODI) and P Ojha (IPL)

These results are practical because batsmen in IPL games tend to play a high risk game in order to be successful. Successful IPL batsmen score lots of runs at a rapid rate that increases over the course of the innings. A higher risk approach maximises the chances of scoring a higher number of runs than the opposition with only 20 allocated overs in which to do so. In contrast, batsmen in ODI games tend to play more conservatively in order to be successful. Successful ODI batsmen simultaneously remain in bat for a long period of time and score a high number of runs. A more conservative approach maximises the chances of the batting team using all 50 allocated overs.

9.2.1 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (KS) test is a non-parametric test used to assess whether two samples come from the same distribution. The hypotheses tested are H_0 : The samples come from the same distribution vs H_1: The samples come from different distributions [72].

9.2.1.1 KS Test and Survival Probabilities

This research applied the KS test to assess for statistical evidence of a difference in distribution between the survival probabilities of IPL batsmen and the survival probabilities of ODI batsmen. In each test, one sample consisted of the survival probabilities for an IPL batsman from Table 9.1, and another consisted of the survival probabilities for an ODI batsman from Table 9.1.

Table 9.2 illustrates the p-values from a KS test applied to survival curves in Figures 9.1, 9.2 and 9.3.

IPL batsman	ODI batsman	KS test p-value
IR Bell	G Gambhir	0.007
SV Samson	KS Williamson	1.171e-08
P Ojha	CJ Anderson	6.064e-08
H Singh	LM Jongwe	5.377e-06
P Sahu	AM Guruge	0.1482
SPD Smith	EJG Morgan	0.001343

Table 9.2: Survival Probabilities KS test

From Table 9.2, the results of the KS test suggest that the distribution of survival probabilities differs the most between top order batsmen, SV Samson and KS Williamson.

This suggests that the difference in risk strategy between ODI and IPL batsmen is most evident in those batsmen who play in top order positions. IPL top order batsmen need to ensure they set a high scoring rate in preparation for potentially less *effective* lower order batsmen, while top order ODI batsmen need to give themselves and the team the greatest chance of using all 50 overs.

9.2.1.2 KS Test and AUC

This research also applied the KS test to assess for statistical evidence of a difference in distribution between the cumulative AUC for IPL batsmen and the cumulative AUC for ODI batsmen.

Table 9.3 illustrates the p-values from a KS test applied to the cumulative AUC associated with batsmen from Figures 9.1, 9.2 and 9.3.

IPL batsman	**ODI batsman**	**KS test p-value**
IR Bell	G Gambhir	1.996e-11
SV Samson	KS Williamson	2.2e-16
P Ojha	CJ Anderson	8.493-09
H Singh	LM Jongwe	6.661e-16
P Sahu	AM Guruge	0.001406
SPD Smith	EJG Morgan	4.441e-16

Table 9.3: AUC KS test

From Table 9.3, the results of the KS test suggest that the distribution of cumulative AUC differs the most between top order batsmen, SV Samson and KS Williamson.

Table 9.4 summarises the differences in total AUC between ODI batsmen and IPL batsmen.

IPL batsman	ODI batsman	Difference in AUC
G Gambhir	IR Bell	6.951
SV Samson	KS Williamson	25.267
P Ojha	CJ Anderson	14.609
H Singh	LM Jongwe	8.598
P Sahu	AM Guruge	2.754
SPD Smith	EJG Morgan	16.220

Table 9.4: Difference in AUC between IPL and ODI batsmen

As illustrated in Table 9.4, the results found that the curve that produced the largest difference in total AUC for IPL batsmen compared to ODI batsmen was generated through the model associated with top order batsmen, SV Samson and KS Williamson. This is consistent with the results of the KS test application. These results are also practical, as higher order ODI batsmen tend to play more conservatively than lower order ODI batsmen, and higher order IPL batsmen tend to opt for a higher risk strategy than lower order IPL batsmen.

The ability of the final models to distinguish between higher survival probabilities for batsmen in ODI games and lower survival probabilities for batsmen in IPL games suggests the final models are valid. This is due to the difference in style of play adopted by players in ODI games compared to IPL games. Batsmen in IPL games play more aggressively and are more likely to be dismissed than batsmen in ODI games.

9.3 Batting Partnership Model Validation Results

A selection of batting partnership combinations across a variety of games were used to illustrate the results of the process used to validate the batting partnership models. Table 9.5 describes each IPL batting partnership and the respective game they played. These partnerships were selected, as each pair of partnerships being compared faced a similar number of balls in their respective game.

Batting partnership		Match	Date played
P1	**P2**		
M Vijay	M Vohra	Kings XI Punjab v Gujarat Lions	11/04/2016
AB de Villiers	V Kohli	Royal Challengers Bangalore v Sunrisers Hyderabad	12/04/2016
S Dhawan	Y Singh	Mumbai Indians v Sunrisers Hyderabad	08/05/2016
SW Billings	KK Nair	Delhi Daredevils v Kolkata Knight Riders	30/04/2016
S Al Hasan	YK Pathan	Kolkata Knight Riders v Gujarat Lions	08/05/2016
P Ojha	S Dhawan	Sunrisers Hyderabad v Rising Pune Supergiants	26/04/2016
DJ Bravo	JP Faulkner	Delhi Daredevils v Gujarat Lions	27/04/2016
J Yadav	CH Morris	Delhi Daredevils v Royal Challengers Bangalore	22/05/2016
NM Coulter-Nile	A Mishra	Kolkata Knight Riders v Delhi Daredevils	10/04/2016
NM Coulter-Nile	Z Khan	Kolkata Knight Riders v Delhi Daredevils	10/04/2016
PA Patel	AT Rayudu	Kings XI Punjab v Mumbai Indians	25/04/2016

Table 9.5: Partnership match information

Table 9.6 describes each ODI batting partnership and the respective game they played.

Batting partnership		**Match**	**Date played**
P1	**P2**		
MJ Guptill	JD Ryder	New Zealand v India	31/01/2014
V Kohli	AM Rahane	India v West Indies	17/10/2014
GS Ballance	BA Stokes	Australia v England	19/01/2014
K Khan	R Mustafa	United Arab Emirates v Afghanistan	02/12/2014
GJ Bailey	MR Marsh	Australia v South Africa	16/11/2014
N Hossain	S Rahman	Bangladesh v Zimbabwe	09/11/2015
KMDN Kulasekara	SMA Priyanjan	Ireland v Sri Lanka	06/05/2014
S Khan	W Riaz	Pakistan v Zimbabwe	01/03/2015
MJ McClenaghan	TG Southee	New Zealand v India	25/01/2014
NO Miller	R Rampaul	West Indies v England	02/03/2014
H Masakadza	R Mutumbami	Afghanistan v Zimbabwe	06/01/2016

Table 9.6: Partnership match information

Figure 9.4 was constructed to illustrate the differences between ball-by-ball survival probabilities for ODI and IPL opening partnerships. MJ Guptill and JD Ryder had higher survival probabilities than M Vijay and M Vohra. MJ Guptill and JD Ryder played conservatively, on route to 22 from 46. M Vijay and M Vohra opted for a higher risk strategy, on route to 78 from 50.

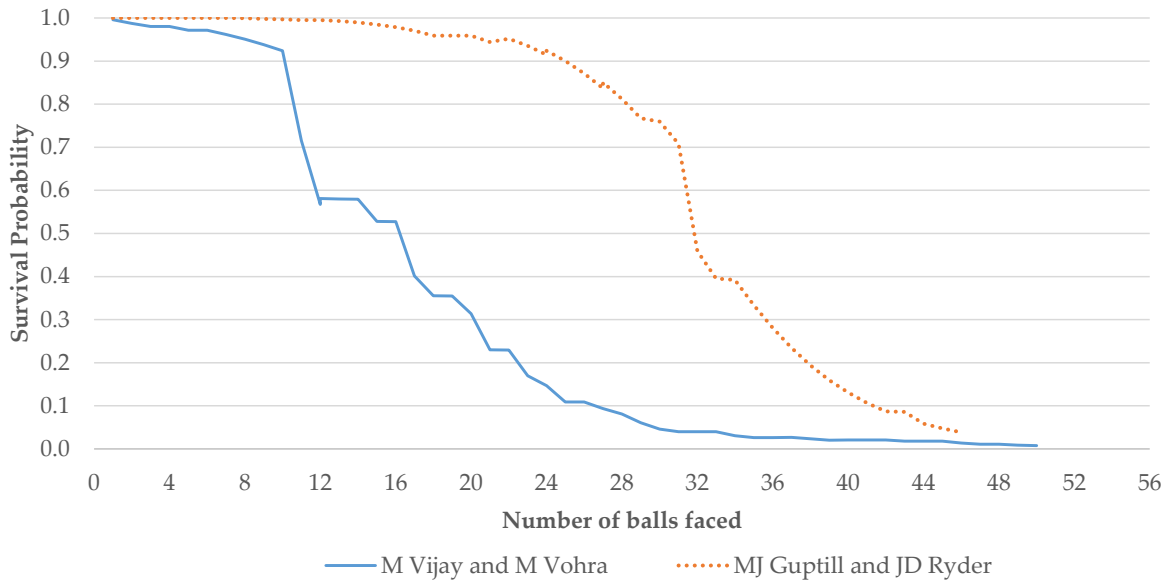

Figure 9.4: Survival probabilities for MJ Guptill and JD Ryder (ODI) and M Vijay and M Vohra (IPL)

Figure 9.5 was constructed to illustrate the differences between ball-by-ball survival probabilities for ODI and IPL wicket two partnerships. From Figure 9.5, V Kohli and AM Rahane had higher survival probabilities than AB de Villiers and V Kohli. V Kohli and AM Rahane played conservatively, on route to 72 from 91. AB de Villiers and V Kohli played a higher risk game, on route to 157 from 87.

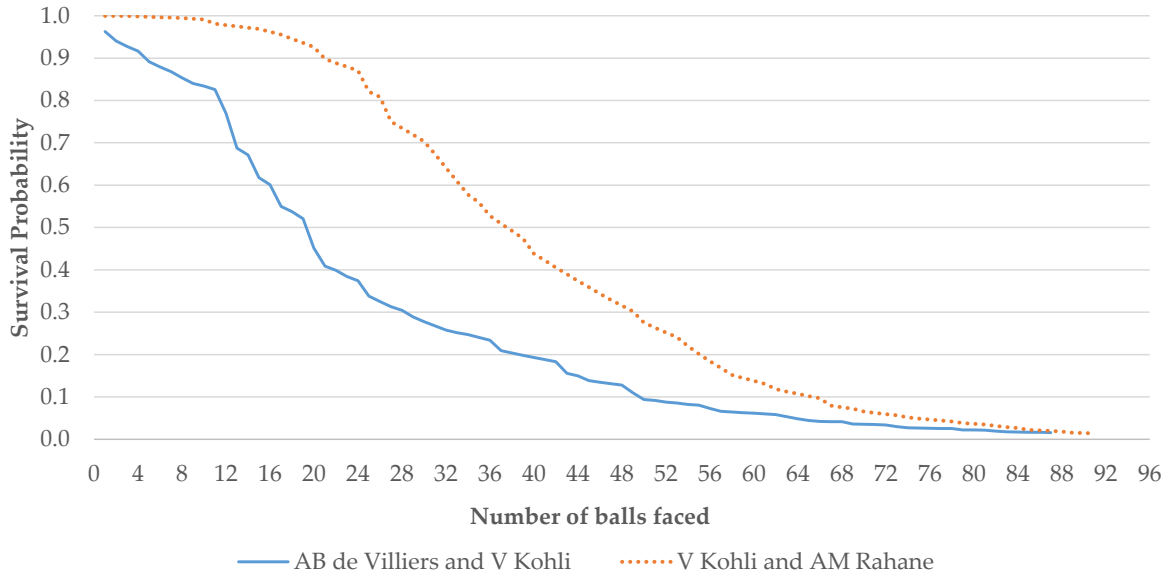

Figure 9.5: Survival probabilities for V Kohli and AM Rahane (ODI) and AB de Villiers and V Kohli (IPL)

Figure 9.6 was constructed to illustrate the differences between ball-by-ball survival probabilities for ODI and IPL wicket three partnerships. From Figure 9.6, GS Ballance and BA Stokes had higher survival probabilities than S Dhawan and Y Singh. GS Ballance and BA Stokes played conservatively, on route to 21 from 51. S Dhawan and Y Singh played a higher risk game, on route to 85 from 49.

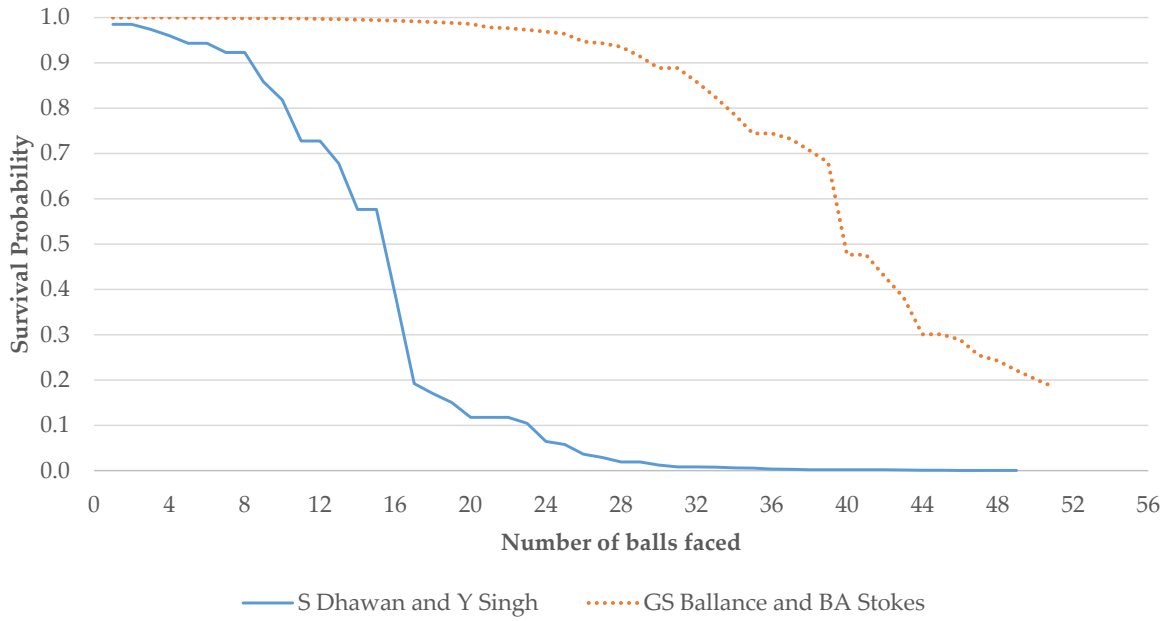

Figure 9.6: Survival probabilities for GS Ballance and BA Stokes (ODI) and S Dhawan and Y Singh (IPL)

9.3.1 KS Test and Partnership Survival Probabilities

This research applied the KS test to assess for statistical evidence of a difference in distribution between the survival probabilities for IPL partnerships and the survival probabilities for ODI partnerships. In each test, one sample consisted of the survival probabilities for an IPL partnership and another consisted of survival probabilities for an ODI partnership.

Table 9.7 illustrates the p-values from a KS test applied to survival curves in Figures 9.4, 9.5 and 9.6.

IPL batting partnership		ODI batting partnership		KS test p-value
P1	**P2**	**P1**	**P2**	
M Vijay	M Vohra	MJ Guptill	JD Ryder	2.453e-05
AB de Villiers	V Kohli	V Kohli	AM Rahane	0.01671
S Dhawan	Y Singh	GS Ballance	BA Stokes	5.459e-10
SW Billings	KK Nair	K Khan	R Mustafa	0.0002322
S Al Hasan	YK Pathan	GJ Bailey	MR Marsh	0.01458
P Ojha	S Dhawan	N Hossain	S Rahman	5.221e-08
DJ Bravo	JP Faulkner	KMDN Kulasekara	SMA Priyanjan	0.08597
J Yadav	CH Morris	S Khan	W Riaz	0.001213
NM Coulter-Nile	A Mishra	MJ McClenaghan	TG Southee	3.609e-06
NM Coulter-Nile	Z Khan	N Miller	R Rampaul	0.001106
PA Patel	AT Rayudu	H Masakadza	R Mutumbami	5.127e-05

Table 9.7: Survival probabilities KS test

From Table 9.7, the results of the KS test suggest the difference in survival probabilities is largest for wicket three partnerships S Dhawan and Y Singh and GS Ballance and BA Stokes.

9.3.2 KS Test and Partnership AUC

This research also applied the KS test to assess for statistical evidence of a difference in distribution between the cumulative AUC for IPL partnerships and the cumulative AUC for ODI partnerships.

Table 9.8 illustrates the p-values from a KS test applied to the cumulative AUC associated with partnerships from Figures 9.4, 9.5 and 9.6.

IPL batting partnership		ODI batting partnership		KS test p-value
P1	**P2**	**P1**	**P2**	
M Vijay	M Vohra	MJ Guptill	JD Ryder	3.533e-11
AB de Villiers	V Kohli	V Kohli	AM Rahane	2.22e-16
S Dhawan	Y Singh	GS Ballance	BA Stokes	8.26e-14
SW Billings	KK Nair	K Khan	R Mustafa	6.661e-16
S Al Hasan	YK Pathan	GJ Bailey	MR Marsh	6.661e-16
P Ojha	S Dhawan	N Hossain	S Rahman	$<$ 2.2e-16
DJ Bravo	JP Faulkner	KMDN Kulasekara	SMA Priyanjan	4.084e-07
J Yadav	CH Morris	S Khan	W Riaz	9.711e-09
NM Coulter-Nile	A Mishra	MJ McClenaghan	TG Southee	3.065e-06
NM Coulter-Nile	Z Khan	N Miller	R Rampaul	0.0003936
PA Patel	AT Rayudu	H Masakadza	R Mutumbami	2.22e-16

Table 9.8: AUC KS test

From Table 9.8, the results of the KS test suggest the distribution of cumulative AUC differs the most between wicket six partnerships, P Ojha and S Dhawan and N Hossain and S Rahman.

Table 9.9 summarises the differences in total AUC between ODI and IPL partnerships.

IPL batting partnership		**ODI batting partnership**		**Difference in AUC**
P1	**P2**	**P1**	**P2**	
M Vijay	M Vohra	MJ Guptill	JD Ryder	17.066
AB de Villiers	V Kohli	V Kohli	AM Rahane	16.695
S Dhawan	Y Singh	GS Ballance	BA Stokes	25.827
SW Billings	KK Nair	K Khan	R Mustafa	22.497
S Al Hasan	YK Pathan	GJ Bailey	MR Marsh	11.026
P Ojha	S Dhawan	N Hossain	S Rahman	15.827
DJ Bravo	JP Faulkner	KMDN Kulasekara	SMA Priyanjan	6.143
J Yadav	CH Morris	S Khan	W Riaz	8.695
NM Coulter-Nile	A Mishra	MJ McClenaghan	TG Southee	11.823
NM Coulter-Nile	Z Khan	N Miller	R Rampaul	9.416
PA Patel	AT Rayudu	H Masakadza	R Mutumbami	18.050

Table 9.9: Difference in AUC between ODI and IPL partnerships

As illustrated in Table 9.9, the results found that the curves that produced the largest differences in total AUC for IPL partnerships compared to ODI partnerships, were generated through the model associated with wicket three partnerships, S Dhawan and Y Singh and GS Ballance and BA Stokes. This is practical, as ODI partnerships at early wickets tend to be more conservative than ODI partnerships at later wickets and IPL partnerships at early wickets tend to play a higher risk game than IPL partnerships at later wickets.

These results are practical because partnerships in IPL games tend to play a high risk game in order to be successful. Successful IPL partnerships score lots of runs at a rapid rate that increases over the course of the innings. A higher risk approach maximises the chances of scoring a higher number of runs than the opposition, with only 20 allocated overs in which to do so. In contrast, partnerships in ODI games tend to play more conservatively in order to be

successful. Successful ODI batsmen simultaneously remain in bat for a long period of time and score a high number of runs. A more conservative approach maximises the chances of the batting team using all 50 allocated overs.

The ability of the final models to distinguish between higher survival probabilities for partnerships in ODI games compared with partnerships in IPL games suggests the final models are valid. This is due to the difference in style of play adopted by partnerships in ODI games compared to IPL games. Partnerships in IPL games play more aggressively and are more likely to be dismissed than batsmen in ODI games.

9.4 Chapter Remarks

This chapter illustrated how each formulated model was able to successfully distinguish between higher survival probabilities for batsmen and partnerships in ODI games compared with those in IPL games. This formed an effective way of validating each model. Consequently, this chapter addressed the model validation limitation in Section 3.3 in Chapter 3. In addition, application of the KS test provided insights into strategical differences opted by ODI and IPL batsmen in particular positions and partnerships at particular wickets. Chapter 10 will include a discussion of this research and how it has addressed the gap in the literature, an overview of potential areas of future research and concludes with a summary of the key results of this research.

Chapter 10

Discussion, Future Research and Conclusion

10.1 Discussion

In cricket, given that teams that win earn bigger prize money, it is important for managers and coaching staff to maximise their chances of success. To succeed, it is essential that player selection and strategic decisions using analytical techniques can provide a competitive advantage. Using analytics optimises the *effectiveness* of player performance and consequently the success of the team, by providing objective transparency into what it takes to win a game of cricket.

This research was motivated by the lack of academic literature surrounding survival analysis applications to the performance of batsmen and batting partnerships, the rapid growth within cricket and the importance in understanding the factors behind a team's ability to win. Particularly challenging is how to optimally set a total. This is, how should batsmen approach an innings to ensure they score as many runs as possible, while minimising the risk, such that all available overs in the first innings are used?

The application of survival analysis techniques to individual batsman performance is not an entirely new area, with several pieces of previous work covering this ([65], [37], [60], [36], [90]). However, there were a variety of shortcomings in these areas of work:

1. Lack of performance metrics as predictors
2. Lack of focus into within-game events
3. Lack of analysis investigating different order individual batsmen
4. Lack of analysis investigating batting partnerships

5. Lack of final model validation

Each of these shortcomings was rigorously addressed in this research. Data were split into multiple subsets. Data associated with individual batsmen were split according to the individual's batting position. Data associated with batting partnerships were split according to the wicket when the partnership was played. Using a combination of Cox proportional hazard modelling, ridge regression techniques and censoring techniques, Cox models were formulated, for each subset of data, with a selection of within-game events included as predictors. Examples of these predictors included runs scored, dot balls, boundaries scored and contribution. These models were used to calculate the ball-by-ball probability of a batsman or batting partnership facing the next ball without being dismissed in the first innings of a limited overs cricket match. Performance metrics considering survival probabilities, run totals and winning were successfully generated to rank batsmen and batting partnerships. These were used to determine optimal partnership combinations across global cricketing nations. Unlike batting in the second innings, where a total is known, the first innings requires the team to score as many runs as possible. This research provides a framework to assess the value of the observed strategies of batsmen, to enable a total that maximises the team's chances of winning to be adopted.

10.2 Future Research

In Chapter 9, the formulated models were successfully validated. However, there are shortcomings of this research:

1. **Distinction between different phases of an innings**

   There may be certain times (i.e. number of balls faced) when the models are able to calculate the probability of partnership dismissal more accurately than at other times. Investigating this issue is a potential area of further research.

2. **Lack of analysis investigating the performance of batsmen and partnerships in the second innings**

   The developed models in this research were fitted to data from the first innings of a selection of ODI games and IPL games. The emphasis on first innings was due to the lack

of previous research into what a match winning total should be. Another area of future work could involve extending this research to investigate the performance of batsmen and partnerships participating in the second innings of limited overs cricket games.

3. **Lack of comparison with different T20 cricket leagues**

   Given that the IPL focuses on Indian matches and conditions, another area of future work could involve extending this research to investigate the performance of individual batsmen and partnerships participating in other T20 cricket leagues such as the Big Bash League in Australia or the T20 cup contested in England and Wales. This may reveal any international differences between the performance of individual batsmen and partnerships.

4. **Lack of investigation into parametric survival analysis**

   Previous work into cricket analytics has compared the use of the exponential and Weibull distributions to model batsman scores in cricket [82]. The general conclusions from this work suggest that the Weibull distribution provides a better representation of batting scores than the exponential distribution. This research could be adapted to undertake a parametric survival analysis approach. The exponential, Weibull and log-logistic regression models could potentially be utilised to model batsman survival rates using similar methodology to that when applying Cox models. Comparisons between the models could then be made to determine which approach is the most suitable.

5. **Lack of model validation using ODI data**

   The model validation methodology involved applying the models to data from the 2016 IPL season. The developed models in this research were fitted to data from ODI games contested between 26th December 2013 and 14th February 2016. Future research could explore the models further by applying them to ODI data from different timeframes close to the current timeframe used.

10.3 Conclusion

The primary objective of this research was to formulate models to calculate the probability of a batting partnership being dismissed in the first innings of a limited overs cricket game. The goal was to optimise batting partnership strategy by determining which pairs of individual batsmen formed the most *effective* partnerships at different stages of the first innings of limited overs cricket games. The research hypothesised that the more *effective* a batting partnership is at occupying the crease, the more runs they will score at an appropriate rate and the more likely the team is to win the match, by setting a defendable total. Survival analysis techniques were the method of choice as they could be applied in a cricket context to investigate the ball-by-ball survival probabilities for batting partnerships. These could subsequently be used to generate insights into the performance of different partnership combinations. Practically, these insights provide valuable information on individual and partnership batting strategies and batting order to maximise the chances of winning when setting a total.

The secondary objective of this research was to validate the final models to ensure they were accurately calculating the ball-by-ball survival probabilities of different partnership combinations. Practically, these insights are useful from a coaching perspective to highlight the importance of minimising dot balls by rotating the strike.

Specifically, the purpose of this research was to address the following two key questions:

1. What are the in-game strategies for optimising the runs scored in the first innings?

2. What are the practical applications of this knowledge?

To answer the first key question, the survival probabilities calculated using each Cox model were cumulated to give a total AUC for each partnership across all games. To account for the differing number of games played by each partnership, the average AUC for each partnership was computed. This measure was used rank the batting partnerships. This method of calculating batting partnership rankings was also positively correlated with typical measures of success: average number of runs scored, proportion of team runs scored and winning. These results support the research hypothesis. These metrics were combined into a measure of perfor-

mance *effectiveness*, used to optimise batting partnership strategy. This completed the primary objective of this research.

To validate each final model, each model that was applied to a subset of ODI data was also applied to corresponding IPL data. All models successfully distinguished between higher survival probabilities for ODI batsmen and partnerships compared with IPL batsmen and partnerships. This completed the secondary objective of this research.

Based on ODI games played between 26th December 2013 and 14th February 2016, the model for opening batting partnerships ranked Pakistani's A Ali and S Aslam as the optimal opening batting partnership. Interestingly, HM Amla and AB de Villiers are ranked as the optimal wicket two partnership and TA Boult is ranked as the optimal number 11 batsman.

New Zealand captain, KS Williamson is the optimal batsman in position three, irrespective of which opener is dismissed. At the time of New Zealand's loss against Australia on 4th December 2016, KS Williamson was ranked as number five in the ICC ODI rankings. Additionally, in [4], KS Williamson was acknowledged as "the most important player to his team in the world".

To answer the second key question, this research reviewed New Zealand's loss against Australia on 4th December 2016. The review indicated a suboptimal order was used with JDS Neesham and BJ Watling batting at four and five respectively. Given the circumstances, C Munro and C de Grandhomme were quantified as a more optimal order. This supported the batting order suggestions made by former Black caps all-rounder and Auckland A coach, AR Adams. This demonstrates a practical application of this research.

The original approach developed in this research may be useful for scouting youth talent. Statistics that demonstrate the positive influence that occupying the crease for batting partnerships has on match outcomes is useful for player development, selection and in-game strategies. This highlights the usefulness of the research as a coaching tool.

This research has utilised survival analysis methodology to develop and validate models to calculate the ball-by-ball probability of a batting partnership being dismissed in the first innings

of a limited overs cricket match. These models were used to derive optimal batting partnership combinations across a wide variety of cricketing nations. Additionally, the models have been used to determine the New Zealand optimal batting order during a selection of ODI games. This research has achieved all research milestones required to develop an original approach capable of optimising batting partnership strategy in the first innings of limited overs cricket games. Additionally, the two key research questions (1) and (2), have been thoroughly examined. The original, robust and validated quantitative framework derived in this research, enables cricket teams to make analytically driven decisions to optimise their performance and increase chances of winning.

Importantly, this novel research provides a unique and objective framework for assessing how teams set totals in the first innings of a limited overs game of cricket, to increase their chances of winning. With pragmatic insights, this research can be deployed by coaches, selectors and other interested parties to gain a deeper understanding of how to dynamically structure a batting order to maximise winning.

Appendix A

Cox Model Theory and Parametric Survival Analysis

A.1 Analytical Parameter Estimation

To estimate the parameters in a Cox model, maximum likelihood estimation techniques are utilised. Suppose:

$$\delta_i = \begin{cases} 1, & \text{if individual } i \text{ is uncensored} \\ 0, & \text{if individual } i \text{ is right censored} \end{cases} \tag{A.1}$$

for $i = 1, \ldots, m$,

then the likelihood function for a general model with some parameters, $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, is defined as [55]:

$$f(\boldsymbol{T}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{X}) = \prod_{i=1}^{m} h(t_i|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x})^{\delta_i} S(t_i|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}). \tag{A.2}$$

For the Cox model, the likelihood function is defined as [55]:

$$f(\boldsymbol{T}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{X}) = \prod_{i=1}^{m} h_0(t_i|\boldsymbol{\alpha})^{\delta_i} \exp(\boldsymbol{\beta}'\boldsymbol{X})^{\delta_i} S_0(t_i|\boldsymbol{\alpha})^{\exp(\boldsymbol{\beta}'\boldsymbol{X})}. \tag{A.3}$$

Taking the logarithm yields:

$$\log f(\boldsymbol{T}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{X}) = \prod_{i=1}^{m} \delta_i \log h_0(t_i|\boldsymbol{\alpha}) + \delta_i \boldsymbol{\beta}'\boldsymbol{X} + \exp(\boldsymbol{\beta}'\boldsymbol{X}) \log S_0(t_i|\boldsymbol{\alpha}). \tag{A.4}$$

To maximise this log-likelihood function requires specification of the form for the baseline hazard, $h_0(t_i|\boldsymbol{\alpha})$. The partial likelihood function may be used as an alternative.

A.1.1 Partial Likelihood for Unique Failure Times

To derive the partial likelihood function conditional on no individuals failing at the same time, define the risk set, $\boldsymbol{R}(t)$, to be the set of individuals who have not died or been censored. Let $\phi_i = \exp(\boldsymbol{\beta}'\boldsymbol{X}_i)$. The partial likelihood for $\boldsymbol{\beta}$ is given by [55]:

$$l_p(\boldsymbol{\beta}, \boldsymbol{X}) = \prod_{i=1}^{m} \left[\frac{\phi_i}{\sum_{j \in R(t_i)} \phi_j} \right]^{\delta_i}. \tag{A.5}$$

A.1.2 Partial Likelihood for Repeated Failure Times

The partial likelihood function differs when two or more individuals are recorded as failing at the same time, depending on the method used. Let t_i be the i^{th} ordered unique failure time. Let $\boldsymbol{D}(t)$ represent the set of individuals who fail at time, t [55].

A.1.2.1 Exact Method

The exact partial likelihood is defined as [55]:

$$l_p(\boldsymbol{\beta}, \boldsymbol{X}) = \prod_{i=1}^{I} \frac{\prod_{j \in D(t_{(i)})} \phi_j}{\sum_{q \in Q_i} \boldsymbol{\Phi}_q}, \tag{A.6}$$

where $\boldsymbol{Q}_i$ is the set of all $|\boldsymbol{D}(t_{(i)})|$-tuples that could be selected from $\boldsymbol{R}(t_{(i)})$ and $\boldsymbol{\Phi}_q$ is the product of ϕ_j for all members j of $|\boldsymbol{D}(t_{(i)})|$-tuple q.

A.1.2.2 Breslow's Method

Breslow's method is one approximation to the exact method. When using Breslow's method, the partial likelihood is defined as [55]:

$$l_p(\boldsymbol{\beta}, \boldsymbol{X}) = \prod_{i=1}^{I} \frac{\prod_{j \in D(t_{(i)})} \phi_j}{(\sum_{j \in R(t_{(i)})} \phi_j)^{|D(t_{(i)})|}}. \tag{A.7}$$

A.1.2.3 Efron's Method

Efron's method is another approximation to the exact method. When using Efron's method, the partial likelihood is defined as [55]:

$$l_p(\boldsymbol{\beta}, \boldsymbol{X}) = \prod_{i=1}^{I} \frac{\prod_{j \in D(t_{(i)})} \phi_j}{\prod_{k=1}^{|D(t_{(i)})|} \left[\sum_{j \in R(t_{(i)})} \phi_j - \frac{k-1}{|D(t_{(i)})|} \sum_{j \in D(t_{(i)})} \phi_j \right]}. \tag{A.8}$$

A.2 Numerical Parameter Estimation

The Newton-Raphson method is an iterative procedure that may be used to numerically estimate the parameters in a Cox model. Let $\boldsymbol{\beta}$ be the a parameter vector of dimension, p. The procedure to find $\hat{\boldsymbol{\beta}}$ that maximises the log-likelihood function, $l(\boldsymbol{\beta})$, is as follows [55]:

1. Let k=0
2. Arbitrarily choose $\boldsymbol{\beta}^{(k)}$
3. Solve

$$I(\boldsymbol{\beta}^{(k)})(\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}) = U(\boldsymbol{\beta}^{(k)}), \tag{A.9}$$

   for $\boldsymbol{\beta}^{(k+1)}$
4. Increment k by one
5. Repeat steps 3,4 and 5 until convergence

$\boldsymbol{\beta}^{(k)}$ represents the value of the parameters at iteration k of the procedure. $U(\boldsymbol{\beta})$ is the score function defined by:

$$U(\boldsymbol{\beta}) = \left(\frac{dl(\boldsymbol{\beta})}{d\beta_1}, \ldots, \frac{dl(\boldsymbol{\beta})}{d\beta_p} \right). \tag{A.10}$$

$I(\beta)$ is the information matrix defined by:

$$I(\boldsymbol{\beta}) = -\begin{pmatrix} \frac{d^2 l(\boldsymbol{\beta})}{d\beta_1^2} & \cdots & \frac{d^2 l(\boldsymbol{\beta})}{d\beta_1 d\beta_p} \\ \vdots & \ddots & \vdots \\ \frac{d^2 l(\boldsymbol{\beta})}{d\beta_1 d\beta_p} & \cdots & \frac{d^2 l(\boldsymbol{\beta})}{d\beta_1^2} \end{pmatrix}. \tag{A.11}$$

The log-likelihood function, $l(\boldsymbol{\beta})$, may be replaced by the partial log likelihood function, $l_p(\boldsymbol{\beta})$.

A.3 Parametric Survival Analysis

Accelerated Failure Time (AFT) models are one type of parametric survival model. These models are survival time models that are linearised by taking logs. One characteristic of AFT models is that the effect of AFT model covariates is said to accelerate survival time. As such, the effect of these covariates are multiplicative on the time scale. Three common AFT models are outlined in Sections A.3.1, A.3.2 and A.3.3 respectively.

A.3.1 Exponential Regression Model

The distribution of survival time to an event, T, can be written as a function of a single covariate as [55]:

$$T = e^{\beta_0 + \beta_1 x} \times \epsilon. \tag{A.12}$$

Assuming one covariate is used to model the survival time, the exponential regression model is written as [55]:

$$\ln(T) = \beta_0 + \beta_1 x + \ln(\epsilon). \tag{A.13}$$

If the error term, ϵ, follows a log-exponential distribution, the survival function for the model in Equation (A.13) is written as [55]:

$$S(t, x, \boldsymbol{\beta}) = \exp\left(\frac{-t}{e^{\beta_0 + \beta_1 x}}\right). \tag{A.14}$$

The hazard function for the model in Equation (A.13) is [55]:

$$h(t, x, \boldsymbol{\beta}) = e^{-(\beta_0+\beta_1 x)}. \tag{A.15}$$

The hazard ratio for a dichotomous covariate is:

$$HR(x = 1, x = 0) = e^{-\beta_1}. \tag{A.16}$$

As a result, the exponential regression model is an AFT model that characterises proportional hazards.

A.3.2 Weibull Regression Model

The Weibull regression model consisting of one covariate is written as [55]:

$$\ln(T) = \beta_0 + \beta_1 x + \sigma + \ln(\epsilon). \tag{A.17}$$

The survival function for the model in Equation (A.17) is [55]:

$$S(t, x, \boldsymbol{\beta}, \sigma) = \exp{-t^{\lambda}}\exp\left[\left(\frac{-1}{\sigma}\right)(\beta_0 + \beta_1 x)\right], \tag{A.18}$$

where $\lambda = \frac{1}{\sigma}$. The hazard function for the model in Equation (A.17) is [55]:

$$h(t, x, \boldsymbol{\beta}, \lambda) = \frac{\lambda t^{\lambda-1}}{(e^{\beta_0+\beta_1 x})^{\lambda}}. \tag{A.19}$$

A.3.3 Log-Logistic Regression Model

The Log-logistic regression model consisting of one covariate is written as [55]:

$$\ln(T) = \beta_0 + \beta_1 x + \sigma + \epsilon, \tag{A.20}$$

where ϵ is logistically distributed. The survival function for the model in Equation (A.20) is [55]:

$$S(t, x, \boldsymbol{\beta}, \sigma) = [1 + \exp(z)]^{-1}, \tag{A.21}$$

where $z = \frac{y-\beta_0-\beta_1 x}{\sigma}$ and $y = \ln(t)$. The hazard function for the model in Equation (A.20) is [55]:

$$h(t, x, \boldsymbol{\beta}, \sigma) = \frac{1}{\sigma} \times \frac{1}{t} \times \frac{e^z}{1 + e^z}. \tag{A.22}$$

Appendix B

Performance Metric Definitions

Performance Metric	Definition
Batsman balls	Cumulative number of balls faced by batsman
Batsman runs	Cumulative number of runs scored by batsman
Batsman dot balls	Cumulative number of dot balls faced by batsman
Batsman consecutive dot balls	Cumulative number of consecutive dot balls faced by batsman
Batsman less than 2 in 4	Cumulative number of balls faced in which less than 2 runs in 4 balls had been scored by batsman
Batsman boundaries	Cumulative number of boundaries hit by batsman
Batsman contribution	Percentage of team runs scored by batsman
Batsman percentage boundaries	Percentage of balls faced by batsman that boundaries were scored off
Batsman percentage dot balls	Percentage of balls faced by batsman that were dot balls
Partnership balls	Cumulative number of balls faced by partnership
Partnership runs	Cumulative number of runs scored by partnership
Partnership dot balls	Cumulative number of dot balls faced by partnership
Partnership consecutive dot balls	Cumulative number of consecutive dot balls faced by partnership
Partnership less than 2 in 4	Cumulative number of balls faced in which less than 2 runs in 4 balls had been scored by partnership
Partnership boundaries	Cumulative number of boundaries hit by partnership
Partnership contribution	Percentage of team runs scored by partnership
Partnership percentage boundaries	Percentage of balls faced by partnership that boundaries were scored off
Partnership percentage dot balls	Percentage of balls faced by partnership that were dot balls
Wicket	Wicket associated with each ball faced

Appendix C

Ball-by-Ball Data Structure

game	cricinfo_id	home	away	venue	dates	year	innings	over	ball	Description	Bowling	Facing	Out	Runs_scored	Sundry_Type	Batting_Pos	Bowling_Pos
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	0	1	full delivery, sw	McKay	Cook	0	0		1	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	0	2	slides this wide	McKay	Cook	0	4		1	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	0	3	tighter line but s	McKay	Cook	0	0		1	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	0	4	full length, Cook	McKay	Cook	1	0		1	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	0	5	wide enough to	McKay	Root	0	0		2	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	0	6	length delivery (	McKay	Root	0	0		2	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	1	1	much too wide (	Coulter-Nile	Bell	0	1	WD	3	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	1	1	full length towa	Coulter-Nile	Bell	0	0		3	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	1	2	touch wider line	Coulter-Nile	Bell	0	0		3	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	1	3	fuller and draws	Coulter-Nile	Bell	0	0		3	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	1	4	this time Bell ge	Coulter-Nile	Bell	0	3		3	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	1	5	full delivery ang	Coulter-Nile	Root	0	0		2	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	1	6	better footwork	Coulter-Nile	Root	0	0		2	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	2	1	full length wide	McKay	Bell	0	0		3	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	2	2	too straight and	McKay	Bell	0	1		3	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	2	3	full at the stump	McKay	Root	0	0		2	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	2	4	short and wide b	McKay	Root	0	0		2	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	2	5	good length and	McKay	Root	0	0		2	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	2	6	full and well wic	McKay	Root	0	0		2	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	3	1	full at the stump	Coulter-Nile	Bell	0	2		3	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	3	2	straighter off st	Coulter-Nile	Bell	0	0		3	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	3	3	very full at the s	Coulter-Nile	Bell	0	1		3	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	3	4	Root slow to cor	Coulter-Nile	Root	0	0		2	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	3	5	straighter line a	Coulter-Nile	Root	0	1		2	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	3	6	full and straight	Coulter-Nile	Bell	0	4		3	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	4	1	good length on (	McKay	Root	0	0		2	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	4	2	straighter and R	McKay	Root	0	2		2	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	4	3	middle and off li	McKay	Root	0	0		2	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	4	4	some width out	McKay	Root	0	0		2	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	4	5	fuller on the stu	McKay	Root	0	0		2	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	4	6	fuller just outsic	McKay	Root	0	0		2	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	5	1	good length on t	Coulter-Nile	Bell	0	0		3	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	5	2	fuller and just b	Coulter-Nile	Bell	0	0		3	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	5	3	length delivery a	Coulter-Nile	Bell	0	1		3	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	5	4	nicely bowled, t	Coulter-Nile	Root	0	0		2	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	5	5	wider line and le	Coulter-Nile	Root	0	0		2	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	5	6	dabbed from th	Coulter-Nile	Root	0	0		2	2
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	6	1	full on the stum	McKay	Bell	0	1		3	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	6	2	firm drive nicely	McKay	Root	0	0		2	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	6	3	some inswing ar	McKay	Root	0	0		2	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	6	4	inswinger again,	McKay	Root	0	0		2	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	6	5	wide, too wide,	McKay	Root	0	1	WD	2	1
4th	636159	Australia	England	Melbourn	12-Jan	2014	1	6	5	full at the pads,	McKay	Root	1	0		2	1

Appendix D

Individual Batsman Performance Metrics

b_balls	b_runs	b_dot_balls	b_consecutive_dot_balls	b_less_than_2_in_4	b_boundaries	b_contribution	b_%_boundaries	b_%_dot balls
1	0	1	0	0	0	0	0	100
2	4	1	0	0	1	100	50	50
3	4	2	0	0	1	100	33.33333333	66.66666667
4	4	3	1	0	1	100	25	75
1	0	1	0	0	0	0	0	100
2	0	2	1	0	0	0	0	100
0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	100
2	0	2	1	0	0	0	0	100
3	0	3	2	1	0	0	0	100
4	3	3	2	1	0	37.5	0	75
3	0	3	2	0	0	0	0	100
4	0	4	3	1	0	0	0	100
5	3	4	2	1	0	37.5	0	80
6	4	4	2	1	0	44.44444444	0	66.66666667
5	0	5	4	2	0	0	0	100
6	0	6	5	3	0	0	0	100
7	0	7	6	4	0	0	0	100
8	0	8	7	5	0	0	0	100
7	6	4	2	1	0	54.54545455	0	57.14285714
8	6	5	2	1	0	54.54545455	0	62.5
9	7	5	2	1	0	58.33333333	0	55.55555556
9	0	9	8	6	0	0	0	100
10	1	9	8	7	0	7.692307692	0	90
10	11	5	2	1	1	64.70588235	10	50
11	1	10	8	8	0	5.882352941	0	90.90909091
12	3	10	8	8	0	15.78947368	0	83.33333333
13	3	11	8	8	0	15.78947368	0	84.61538462
14	3	12	9	8	0	15.78947368	0	85.71428571
15	3	13	10	8	0	15.78947368	0	86.66666667
16	3	14	11	9	0	15.78947368	0	87.5
11	11	6	2	1	1	57.89473684	9.090909091	54.54545455
12	11	7	3	1	1	57.89473684	8.333333333	58.33333333
13	12	7	3	1	1	60	7.692307692	53.84615385
17	3	15	12	10	0	15	0	88.23529412
18	3	16	13	11	0	15	0	88.88888889
19	3	17	14	12	0	15	0	89.47368421
14	13	7	3	1	1	61.9047619	7.142857143	50
20	3	18	15	13	0	14.28571429	0	90
21	3	19	16	14	0	14.28571429	0	90.47619048
22	3	20	17	15	0	14.28571429	0	90.90909091
22	3	20	17	16	0	13.63636364	0	90.90909091
23	3	21	17	17	0	13.63636364	0	91.30434783

Appendix E

Batting Partnership Performance Metrics

Partner	Partnership	wicket	p_balls	p_runs	p_dot_balls	p_consecutive_dot_balls	p_less_than_2_in_4	p_boundaries	p_contribution	p_%_boundaries	p_%_dot_balls
Root	Cook,Root	1	1	0	1	0	0	0	0	0	100
Root	Cook,Root	1	2	4	1	0	0	1	100	50	50
Root	Cook,Root	1	3	4	2	0	0	1	100	33.33333333	66.66666667
Root	Cook,Root	1	4	4	3	1	0	1	100	25	75
Bell	Bell,Root	2	1	0	1	0	0	0	0	0	100
Bell	Bell,Root	2	2	0	2	1	0	0	0	0	100
Root	Bell,Root	2	2	0	2	1	0	0	0	0	100
Root	Bell,Root	2	3	0	3	1	1	0	0	0	100
Root	Bell,Root	2	4	0	4	2	2	0	0	0	100
Root	Bell,Root	2	5	0	5	3	3	0	0	0	100
Root	Bell,Root	2	6	3	5	3	3	0	37.5	0	83.33333333
Bell	Bell,Root	2	7	3	6	3	3	0	37.5	0	85.71428571
Bell	Bell,Root	2	8	3	7	4	3	0	37.5	0	87.5
Root	Bell,Root	2	9	3	8	5	3	0	37.5	0	88.88888889
Root	Bell,Root	2	10	4	8	5	4	0	44.44444444	0	80
Bell	Bell,Root	2	11	4	9	5	5	0	44.44444444	0	81.81818182
Bell	Bell,Root	2	12	4	10	6	6	0	44.44444444	0	83.33333333
Bell	Bell,Root	2	13	4	11	7	7	0	44.44444444	0	84.61538462
Bell	Bell,Root	2	14	4	12	8	8	0	44.44444444	0	85.71428571
Root	Bell,Root	2	15	6	12	8	8	0	54.54545455	0	80
Root	Bell,Root	2	16	6	13	8	8	0	54.54545455	0	81.25
Root	Bell,Root	2	17	7	13	8	8	0	58.33333333	0	76.47058824
Bell	Bell,Root	2	18	7	14	8	8	0	58.33333333	0	77.77777778
Bell	Bell,Root	2	19	8	14	8	8	0	61.53846154	0	73.68421053
Root	Bell,Root	2	20	12	14	8	8	1	70.58823529	5	70
Bell	Bell,Root	2	21	12	15	8	8	1	70.58823529	4.761904762	71.42857143
Bell	Bell,Root	2	22	14	15	8	8	1	73.68421053	4.545454545	68.18181818
Bell	Bell,Root	2	23	14	16	8	8	1	73.68421053	4.347826087	69.56521739
Bell	Bell,Root	2	24	14	17	9	8	1	73.68421053	4.166666667	70.83333333
Bell	Bell,Root	2	25	14	18	10	8	1	73.68421053	4	72
Bell	Bell,Root	2	26	14	19	11	9	1	73.68421053	3.846153846	73.07692308
Root	Bell,Root	2	27	14	20	12	10	1	73.68421053	3.703703704	74.07407407
Root	Bell,Root	2	28	14	21	13	11	1	73.68421053	3.571428571	75
Root	Bell,Root	2	29	15	21	13	12	1	75	3.448275862	72.4137931
Bell	Bell,Root	2	30	15	22	13	13	1	75	3.333333333	73.33333333
Bell	Bell,Root	2	31	15	23	14	14	1	75	3.225806452	74.19354839
Bell	Bell,Root	2	32	15	24	15	15	1	75	3.125	75
Root	Bell,Root	2	33	16	24	15	16	1	76.19047619	3.03030303	72.72727273
Bell	Bell,Root	2	34	16	25	15	17	1	76.19047619	2.941176471	73.52941176
Bell	Bell,Root	2	35	16	26	16	18	1	76.19047619	2.857142857	74.28571429
Bell	Bell,Root	2	36	16	27	17	19	1	76.19047619	2.777777778	75
Bell	Bell,Root	2	36	16	27	17	20	1	72.72727273	2.777777778	75
Bell	Bell,Root	2	37	16	28	17	21	1	72.72727273	2.702702703	75.67567568

Appendix F

SAS Code

F.1 Data Extraction and Processing

```
%let innings=1;
%macro ODI_commentary(match,innings);
options nonotes;
%do innings=1 %to &innings;

/*Gathering the Commentary Card*/
filename mydata url
"http://www.espncricinfo.com/ci/engine/match/&match..html?
   innings=&innings.;view=commentary";

data web_data_&match._&innings;
     infile mydata length=len lrecl=32000;
     input record $varying2000. len;
run;

data a1 (keep=match del b2b runs desc);
   set Web_data_&match._&innings;
/*Extract Match ID*/
if index(record,"| Cricket Commentary | ESPN Cricinfo</title>"
   ) ge 1 then do;
   var1 = tranwrd(record,'| Cricket Commentary | ESPN Cricinfo
      </title>','');
   match = compbl(tranwrd(var1,'<title>',''));
end;
```

```
else match="";
drop var1;

/*Extract Delivery*/
if index(record,'<div␣class="commentary-overs">') ge 1 then do
   ;
   var2 = tranwrd(record,'<div␣class="commentary-overs">','');
   del = compbl(tranwrd(var2,'</div>',''));
end;
else del = "";
drop var2;

if del ne "" then header_info=1;
header_info + 0;

if del ne '' then flag=0;
flag + 1;

   if flag = 3 then do;
      b2b = tranwrd(record,'<p>','');
      b2b = tranwrd(b2b,',','');
   end;
   if flag = 4 then do;
      runs = tranwrd(record,'<span␣class="commsImportant">',""
         );
      runs = tranwrd(runs,'</span>',"");
      runs = compbl(tranwrd(runs,',',''));
   end;
   if flag = 5 then desc = tranwrd(record,'/p>','');
                    desc = tranwrd(desc,'<div>','');
                    desc = tranwrd(desc,'<span>','');
```

```
                desc = tranwrd(desc,'</span>','');
                desc = tranwrd(desc,'<','');

retain _match;
            if match ne "" then do;
               _match = match;
            end;
            else do;
            if match = "" then match = _match;
            else _match = match;
            end;

      retain _del;
            if del ne "" then do;
               _del = del;
            end;
            else do;
            if del = "" then del = _del;
            else _del = del;
            end;

      retain _b2b;
            if b2b ne "" then do;
               _b2b = b2b;
            end;
            else do;
            if b2b = "" then b2b = _b2b;
            else _b2b = b2b;
            end;

      retain _runs;
```

```
                if runs ne "" then do;
                   _runs = runs;
                end;
                else do;
                if runs = "" then runs = _runs;
                else _runs = runs;
                end;
          if header_info=0 then delete;
          drop record header_info;
          if flag = 5 then output;
run;

data a2;
   set a1;

   runs1a = strip(scan(runs,1,')'));
   runs1b = strip(scan(runs,2,')'));

   b2b = strip(tranwrd(b2b,',',''));
   b2b = strip(tranwrd(b2b,' to ',' $'));

   bowler = strip(scan(b2b,1,'$'));
   batsman = strip(scan(b2b,2,'$'));

   if index(runs,'OUT') ge 1 then wicket = 1;
      else wicket = 0;

   if runs1b ne '' then runs = runs1b;
   if index(runs,'FOUR') ge 1 then runs1 = 4;
   else if index(runs,'SIX') ge 1 then runs1 = 6;
   else if index(upcase(compress(runs)),"NORUN") ge 1 then
```

```
      runs1 = 0;
   else if index(runs,'OUT') ge 1 and index(runs,"1") = 0 and
      index(runs,"2") = 0
        and index(runs,"3") =0 then runs1=0;
   else runs1 = input(strip(runs),1.0);

   temp = upcase(strip(scan(runs,2,'␣')));

   if temp = 'LEG' then SUNDRY = 'LB';
      if temp = 'NO' then SUNDRY = 'NB';
      if temp = 'WIDE' then SUNDRY = 'WD';
      if runs1a = '(NO␣BALL' then SUNDRY = 'NB';

   drop b2b temp runs runs1a runs1b;

   if b2b = '' then delete;

   rename runs1 = runs;

   n = _n_;

run;

proc sort data = a2 out = batsman (keep = batsman n) nodupkey;
   by batsman;
run;

proc sort data = batsman;
   by n;
run;
```

```
data batsman1;
   set batsman;
    drop n;
   bat_pos = _n_;
run;

proc sort data = a2;
   by batsman;
run;

proc sort data=batsman1;
   by batsman;
run;

data a3;
   merge a2 batsman1;
      by batsman;
run;

proc sort data=a2;
   by n;
run;

proc sort data = a2 out = bowler (keep= bowler n) nodupkey;
   by bowler;
run;

proc sort data=bowler;
   by n;
run;
```

```
data bowler1;
   set bowler;
    drop n;
   bowl_pos = _n_;
run;

proc sort data = a3;
   by bowler;
run;

proc sort data = bowler1;
   by bowler;
run;

data a4;
   merge a3 bowler1;
      by bowler;
run;

proc sort data=a4;
   by n;
run;

data a5;
   set a4;
      drop n;
run;

data ODI&innings;
   set a5;
```

```
    match = tranwrd(match,' v ',' | ');
    game = input(substr(left(match),1,4),$6.);
    cricinfo_id = &match;
    h = strip(scan(match,2,':'));
    home = strip(scan(h,1,'|'));
    a = strip(scan(h,2,'|'));
    a = tranwrd(a,' at ', ' $');
    away = strip(scan(a,1,'$'));
    x = strip(scan(a,2,'$'));
    venue = strip(scan(x,1,','));
    dates = strip(scan(x,2,','));
    year = input(strip(scan(x,3,',')),6.0);

    innings = &innings;

    over = input(strip(scan(del,2,'')),4.0);
    over = floor(over);
    ball = input(strip(scan(del,2,'.')),4.0);

Description = desc;
Bowling = bowler;
Facing = batsman;
Out = wicket;
Runs_scored = runs;
Sundry_Type = sundry;
Batting_Pos = bat_pos;
Bowling_Pos = bowl_pos;
innings = &innings;

drop match h a x del desc bowler batsman wicket runs sundry
    bat_pos bowl_pos;
```

```
run;

proc append data=ODI&innings base=work.game&match force;
run; quit;

proc datasets library=work nolist;
delete a1 a2 a3 a4 a5 batsman batsman1 bowler bowler1 ODI&
   innings
    web_data Web_data_&match._&innings;
run;
quit;

%end;

proc append data=work.game&match base = work.ODI_archive force
   ;
run; quit;

/* Dedupe Data to Ensure that Records are not repeated */
proc sort data=work.ODI_archive nodupkey;
   by cricinfo_id innings over ball description runs_scored
      sundry_type;
run;

proc datasets library=work nolist;
   delete game&match;
run; quit;

options notes;
```

```
%mend ODI_commentary;



%macro bulk_ODI_extract(low_id,high_id);
    %do loop_match=&low_id %to &high_id %by 2;
       %ODI_commentary(&loop_match,2);
    %end;
%mend bulk_ODI_extract;

%bulk_ODI_extract(980901,981019);


proc export
      data=ODI_ARCHIVE
      outfile= "C:\Users\Patrick\SAS␣Extractor\Data\ball_by_
         ball.csv"
      dbms= csv
      replace;
run;
```

Appendix G

R Code

G.1 Exploratory Data Analysis

```
#Read in clean data
data <- read.csv("C:/Users/Patrick/odi_data.csv", header=TRUE,
    sep=",")

#Sort data by match, innings, batting position, over, ball,
   extra and out
data2 <- data[order(data[,"cricinfo_id"],data[,"innings"],data
   [,"Batting_Pos"],data[,"over"],data[,"ball"],-xtfrm(data[,"
   Sundry_Type"]),data[,"Out"],decreasing = F),]

#Subset data to first innings only
finaldata <- data2[which(data2$innings==1),]

#Rename variables for ease of analysis
attach(finaldata)
finaldata$x1 <- b_balls
finaldata$x2 <- b_runs
finaldata$x3 <- b_dot_balls
finaldata$x4 <- b_consecutive_dot_balls
finaldata$x5 <- b_less_than_2_in_4
finaldata$x6 <- b_boundaries
finaldata$x7 <- b_contribution
finaldata$x8 <- b_%_boundaries
finaldata$x9 <- b_%_dot_balls
```

```
finaldata$x10 <- p_balls
finaldata$x11 <- p_runs
finaldata$x12 <- p_dot_balls
finaldata$x13 <- p_consecutive_dot_balls
finaldata$x14 <- p_less_than_2_in_4
finaldata$x15 <- p_boundaries
finaldata$x16 <- p_contribution
finaldata$x17 <- p_%_boundaries
finaldata$x18 <- p_%_dot_balls

#Logistic regression fit to data associated with individual
   batsmen
fit <- glm(Out ~ x1+x2+x3+x4+x5+x6+x7+x8+x9,family=binomial(
   link='logit'),data=finaldata)

#Logistic regression fit to data associated with batting
   partnerships
fit2 <- glm(Out ~ x10+x11+x12+x13+x14+x15+x16+x17+x18,family=
   binomial(link='logit'),data=finaldata)

#Install and load required packages
install.packages("car",dependencies=TRUE)
install.packages("asbio")
library(car)
library(asbio)

#Outliers
cutoff <- 4/((nrow(finaldata)-length(fit$coefficients)-2))
plot(fit, which=4, cook.levels=cutoff)
```

```
#Variance inflation factors
vif(fit)
vif(fit2)

#Correlations
cor(finaldata[,c("x1","x2","x3","x4","x5","x6","x7","x8","x9")
   ], use="complete.obs", method="pearson")
cor(finaldata[,c("x10","x11","x12","x13","x14","x15","x16","
   x17","x18")], use="complete.obs", method="pearson")

#Scatter plot and correlation matrix
pairs(~x1+x2+x3+x4+x5+x6+x7+x8+x9,data=finaldata, main="Simple
   ␣Scatterplot␣Matrix")
pairs(~x10+x11+x12+x13+x14+x15+x16+x17+x18,data=finaldata,
   main="Simple␣Scatterplot␣Matrix")

#Independence of residuals
durbinWatsonTest(fit)
durbinWatsonTest(fit2)

#Residual outliers
outlierTest(fit)
outlierTest(fit2

#Linearity
finaldata$x1 <- finaldata$x1+1
finaldata$x2 <- finaldata$x2+1
finaldata$x3 <- finaldata$x3+1
finaldata$x4 <- finaldata$x4+1
finaldata$x5 <- finaldata$x5+1
finaldata$x6 <- finaldata$x6+1
```

```
finaldata$x7 <- finaldata$x7+1
finaldata$x8 <- finaldata$x8+1
finaldata$x9 <- finaldata$x9+1

finaldata$x1_ln <- log(finaldata$x1)
finaldata$x2_ln <- log(finaldata$x2)
finaldata$x3_ln <- log(finaldata$x3)
finaldata$x4_ln <- log(finaldata$x4)
finaldata$x5_ln <- log(finaldata$x5)
finaldata$x6_ln <- log(finaldata$x6)
finaldata$x7_ln <- log(finaldata$x7)
finaldata$x8_ln <- log(finaldata$x8)
finaldata$x9_ln <- log(finaldata$x9)

fit3 <- glm(Out ~ x1+(x1:x1_ln)+x2+(x2:x2_ln)+x3+(x3:x3_ln)+x4
   +(x4:x4_ln)+x5+(x5:x5_ln)+x6+(x6:x6_ln)+x7+(x7:x7_ln)+x8+(x8
   :x8_ln)+x9+(x9:x9_ln),family=binomial(link='logit'),data=
   finaldata)
summary(fit3)

finaldata$x10 <- finaldata$x10+1
finaldata$x11 <- finaldata$x11+1
finaldata$x12 <- finaldata$x12+1
finaldata$x13 <- finaldata$x13+1
finaldata$x14 <- finaldata$x14+1
finaldata$x15 <- finaldata$x15+1
finaldata$x16 <- finaldata$x16+1
finaldata$x17 <- finaldata$x17+1
finaldata$x18 <- finaldata$x18+1

finaldata$x10_ln <- log(finaldata$x10)
```

```
finaldata$x11_ln <- log(finaldata$x11)
finaldata$x12_ln <- log(finaldata$x12)
finaldata$x13_ln <- log(finaldata$x13)
finaldata$x14_ln <- log(finaldata$x14)
finaldata$x15_ln <- log(finaldata$x15)
finaldata$x16_ln <- log(finaldata$x16)
finaldata$x17_ln <- log(finaldata$x17)
finaldata$x18_ln <- log(finaldata$x18)

fit4 <- glm(Out ~ x10+(x10:x10_ln)+x11+(x11:x11_ln)+x12+(x12:
   x12_ln)+x13+(x13:x3_ln)+x14+(x14:x14_ln)+x15+(x15:x15_ln)+
   x16+(x16:x16_ln)+x17+(x17:x17_ln)+x18+(x18:x18_ln),family=
   binomial(link='logit'),data=finaldata)
summary(fit4)
```

G.2 Opening Batsman Modelling

```
#Read in clean data
data <- read.csv("C:/Users/Patrick/odi_data.csv", header=TRUE,
    sep=",")

#Sort data by match, innings, batting position, over, ball,
   extra and out
data2 <- data[order(data[,"cricinfo_id"],data[,"innings"],data
   [,"Batting_Pos"],data[,"over"],data[,"ball"],-xtfrm(data[,"
   Sundry_Type"]),data[,"Out"],decreasing = F),]

#Subset data to first innings, opening batsmen only
finaldata <- data2[which(data2$innings==1 & data2$Batting_Pos
   < 3),]
```

```
#Rename variables for ease of analysis
attach(finaldata)
finaldata$x1 <- b_balls
finaldata$x2 <- b_runs
finaldata$x3 <- b_dot_balls
finaldata$x4 <- b_consecutive_dot_balls
finaldata$x5 <- b_less_than_2_in_4
finaldata$x6 <- b_boundaries
finaldata$x7 <- b_contribution
finaldata$x8 <- b_%_boundaries
finaldata$x9 <- b_%_dot_balls

finaldata$x10 <- p_balls
finaldata$x11 <- p_runs
finaldata$x12 <- p_dot_balls
finaldata$x13 <- p_consecutive_dot_balls
finaldata$x14 <- p_less_than_2_in_4
finaldata$x15 <- p_boundaries
finaldata$x16 <- p_contribution
finaldata$x17 <- p_%_boundaries
finaldata$x18 <- p_%_dot_balls

#Open libraries in preparation for survival analysis
library(survival)

#Create survival object: failure time is number of balls faced
    by batsman, event is whether batsman is dismissed
object <- survfit(Surv(b_balls,Out)~1)
summary(object)

#Cox model selection using glmulti package
```

```
#Model selection based on smallest AIC
glmulti.coxph <- glmulti(Surv(b_balls,Out,type="right")~sqrt(
   x2)+sqrt(x3)+sqrt(x4)+sqrt(x5)+sqrt(x6)+sqrt(x7)+sqrt(x8)+
   sqrt(x9),method="h",intercept=F,report=T,level=1,minsize=4,
   maxsize=4,data = finaldata,crit = "aic",fitfunction = "coxph
   ")
glmulti.coxph
summary(glmulti.coxph)

#Final model with ridge terms
model <- coxph(Surv(b_balls, Out,type="right") ~ sqrt(x2)+
   ridge(sqrt(x4),sqrt(x5)),method="efron",data=finaldata)
summary(model)

#Testing proportional hazards assumption
out <- cox.zph(model, global = T)
print(out)
par(mfrow=c(2, 2))
plot(out)
```

Bibliography

[1] asbio: A Collection of Statistical Tools for Biologists Package in R. https://cran.r-project.org/web/packages/asbio/asbio.pdf. Accessed: 2017-02-22.

[2] Billions Of Dollars At Stake: Why Is The International Cricket Council Changing Its Revenue Sharing Model? http://www.ibtimes.com/billions-dollars-stake-why-international-cricket-council-changing-its-revenue-sharing-1554078. Accessed: 2016-09-13.

[3] Bonferroni Outlier Test. http://www.inside-r.org/packages/cran/car/docs/outlierTest. Accessed: 2016-09-12.

[4] By the numbers: Why New Zealand superstar Kane Williamson is cricket's most valuable batsman. http://www.foxsports.com.au/cricket/by-the-numbers-why-new-zealand-superstar-kane-williamson-is-crickets-most-valuable-batsman/news-story/241fb534736c1d475d3c28fe053d9754. Accessed: 2017-01-19.

[5] car: Companion to Applied Regression Package in R. https://cran.r-project.org/web/packages/car/car.pdf. Accessed: 2017-02-22.

[6] The emergence of Sport Analytics. http://analytics-magazine.org/the-emergence-of-sport-analytics/. Accessed: 2017-01-12.

[7] Maximum likelihood estimators and least squares. http://people.math.gatech.edu/~ecroot/3225/maximum_likelihood.pdf. Accessed: 2016-09-15.

[8] Olympics Offers Latest Sign of Sports Betting's Growing Popularity. https://www.americangaming.org/newsroom/press-releasess/olympics-offers-latest-sign-sports-betting%E2%80%99s-growing-popularity. Accessed: 2017-02-22.

[9] Opta. http://www.optasports.com/. Accessed: 2016-12-22.

[10] Plunket Shield, 2016/17 - Otago / Records / Highest averages. http://www.stats.espncricinfo.com/plunket-shield-2016-17/engine/records/batting/

highest_career_batting_average.html?id=11507;team=2621;type=tournament. Accessed: 2016-12-20.

[11] Ridge Regression. http://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf. Accessed: 2016-09-13.

[12] survival: Survival Analysis Package in R. https://cran.r-project.org/web/packages/survival/survival.pdf. Accessed: 2017-02-22.

[13] Test the Proportional Hazards Assumption of a Cox regression. https://stat.ethz.ch/R-manual/R-devel/library/survival/html/cox.zph.html. Accessed: 2017-02-22.

[14] The Baseball Archive v.5.0. http://www.seanlahman.com/baseball-archive/statistics/. Accessed: 2016-09-13.

[15] What were they thinking? - Andre Adams on batting line up. http://http:/http://www.nzherald.co.nz/sport/news/article.cfm?c_id=4&objectid=11760500. Accessed: 2016-12-13.

[16] What were they thinking? Andre Adams left dazed and confused by New Zealand's tactics. http://http://www.themercury.com.au/sport/what-were-they-thinking-andre-adams-left-dazed-and-confused-by-new-zealands-tactics/news-story/d9e119fe95695102d0b6c9b38102c187. Accessed: 2016-12-19.

[17] Why Australia's Big Bash League Is Changing The Professional Sports Paradigm. http://www.forbes.com/sites/jasonbelzer/2016/01/22/why-australias-big-bash-league-is-changing-the-professional-sports-paradigm/#f230794693cc. Accessed: 2017-01-12.

[18] ABEL, E. L., AND KRUGER, M. L. The longevity of baseball hall of famers compared to other players. *Death studies 29*, 10 (2005), 959–963.

[19] ALAMAR, B., AND MEHROTRA, V. Beyond 'Moneyball': The rapidly evolving world of sports analytics, Part I. *Analytics Magazine* (2011).

[20] ALAMAR, B. C. *Sports analytics: A guide for coaches, managers, and other decision makers*. Columbia University Press, 2013.

[21] ALLSOPP, P., AND CLARKE, S. R. Rating teams and analysing outcomes in one-day and test cricket. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 167*, 4 (2004), 657–667.

[22] ANNIS, D. H., AND CRAIG, B. A. Hybrid paired comparison analysis, with applications to the ranking of college football teams. *Journal of Quantitative Analysis in Sports 1*, 1 (2005), 3.

[23] BAILEY, M. J., AND CLARKE, S. R. Market inefficiencies in player head to head betting on the 2003 cricket world cup. In *Economics, Management and Optimization in Sports*. Springer, 2004, pp. 185–201.

[24] BANKER, R. D., CHARNES, A., AND COOPER, W. W. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management science 30*, 9 (1984), 1078–1092.

[25] BARKER, L., AND BROWN, C. Logistic regression when binary predictor variables are highly correlated. *Statistics in medicine 20*, 9-10 (2001), 1431–1442.

[26] BARROS, C. P., FRICK, B., AND PASSOS, J. Coaching for survival: The hazards of head coach careers in the German 'Bundesliga'. *Applied Economics 41*, 25 (2009), 3303–3311.

[27] BHANDARI, I., COLET, E., PARKER, J., PINES, Z., PRATAP, R., AND RAMANUJAM, K. Advanced scout: Data mining and knowledge discovery in NBA data. *Data Mining and Knowledge Discovery 1*, 1 (1997), 121–125.

[28] BIRNBAUM, P. A guide to sabermetric research, 2013.

[29] BRACEWELL, P. J. Monitoring meaningful rugby ratings. *Journal of Sports Sciences 21*, 8 (2003), 611–620.

[30] BRACEWELL, P. J., COOMES, M., NASH, J., MEYER, D., AND ROONEY, S. J. Rating the attacking performance of a non-wicket taking bowler in limited overs cricket. In *ANZIAM Mathsport* (2016).

[31] BRACEWELL, P. J., FARINAZ, F., JOWETT, C. A., FORBES, D. G. R., AND MEYER, D. H. Was Bradman Denied His Prime? *Journal of Quantitative Analysis in Sports 5*, 4 (2009), 1–26.

[32] BRACEWELL, P. J., AND RUGGIERO, K. A parametric control chart for monitoring individual batting performances in cricket. *Journal of Quantitative Analysis in Sports 5*, 3 (2009).

[33] BREWER, B. J. Getting your eye in: A Bayesian analysis of early dismissals in cricket. *arXiv preprint arXiv:0801.4408* (2008).

[34] BROWN, J. D. *Linear Models in Matrix Form: A Hands-On Approach for the Behavioral Sciences*. Springer, 2015.

[35] BROWN, P., BRACEWELL, P. J., AND PATEL, A. K. Optimising a batting order in limited overs cricket using survival analysis. In *Mathsport International* (2017).

[36] BROWN, P., PATEL, A. K., AND BRACEWELL, P. J. Real time prediction of opening batsman dismissal in limited overs cricket. In *ANZIAM Mathsport* (2016).

[37] CLARKE, S. R. Dynamic programming in one-day cricket-optimal scoring rates. *Journal of the Operational Research Society 39*, 4 (1988), 331–337.

[38] COHEN, G. L. Cricketing chances. In *Proceedings of the Sixth Australian Conference on Mathematics and Computers in Sport* (2002), pp. 1–13.

[39] DANAHER, P. Estimating a cricketer's batting average using the product limit estimator. *New Zealand Statistician 24*, 1 (1989), 2–5.

[40] DAS, S. On generalized geometric distributions and improved estimation of batting average in cricket. *Communications in Statistics-Theory and Methods 46*, 6 (2017), 2736–2750.

[41] DEL CORRAL, J., BARROS, C. P., AND PRIETO-RODRÍGUEZ, J. The determinants of soccer player substitutions: a survival analysis of the Spanish soccer league. *Journal of Sports Economics* (2007).

[42] DI SALVO, V., BARON, R., GONZÁLEZ-HARO, C., GORMASZ, C., PIGOZZI, F., AND BACHL, N. Sprinting analysis of elite soccer players during European Champions League and UEFA Cup matches. *Journal of sports sciences 28*, 14 (2010), 1489–1494.

[43] DWYER, D. B., AND GABBETT, T. J. Global positioning system data analysis: Velocity ranges and a new definition of sprinting for field sport athletes. *The Journal of Strength & Conditioning Research 26*, 3 (2012), 818–824.

[44] ELDERTON, W., AND WOOD, G. H. Cricket scores and geometrical progression. *Journal of the Royal Statistical Society 108*, 1/2 (1945), 12–40.

[45] ENGEL, M. *Wisden cricketers' almanack 1999*. John Wisden, 1999.

[46] FIELD, A. *Discovering statistics using IBM SPSS statistics*. Sage, 2013.

[47] FINANCE, B. A study by BrandFinance on IPL V to value IPL Brand and its Nine franchisee brands. Report P-21, Brand Finance India, 2013.

[48] FREEDMAN, D. A. *Statistical models: theory and practice*. Cambridge University Press, 2009.

[49] FRICK, B. Globalization and Factor Mobility The Impact of the "Bosman-Ruling" on Player Migration in Professional Soccer. *Journal of Sports Economics 10*, 1 (2009), 88–106.

[50] GAMEL, J. W., VOGEL, R. L., VALAGUSSA, P., AND BONADONNA, G. Parametric survival analysis of adjuvant therapy for stage II breast cancer. *Cancer 74*, 9 (1994), 2483–2490.

[51] GREENWOOD, M. A Report on the Natural Duration of Cancer. *Reports on Public Health and Medical Subjects. Ministry of Health*, 33 (1926).

[52] GUPTA, R. D., AND KUNDU, D. Exponentiated exponential family: an alternative to gamma and Weibull distributions. *Biometrical journal 43*, 1 (2001), 117–130.

[53] HARVILLE, D. A. The selection or seeding of college basketball or football teams for postseason competition. *Journal of the American Statistical Association 98*, 461 (2003), 17–27.

[54] HOERL, A. E., AND KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 1 (1970), 55–67.

[55] HOSMER JR, D. W., AND LEMESHOW, S. *Applied survival analysis: Regression modelling of time to event data*. Eur Orthodontic Soc, 1999.

[56] HOSMER JR, D. W., LEMESHOW, S., AND STURDIVANT, R. X. *Applied logistic regression*. John Wiley & Sons, 2013.

[57] IBRAHIM, J. G., CHEN, M.-H., AND SINHA, D. *Bayesian survival analysis*. Wiley Online Library, 2005.

[58] JACKSON, D. A. Index betting on sports. *The Statistician* (1994), 309–315.

[59] JACOBY, W. G. *Regression III: Advanced Methods*. Michigan State University, September 2014.

[60] KACHOYAM, B., AND WEST, M. Cricket as life and death. In *ANZIAM Mathsport* (2016).

[61] KAHN, L. M. Race, performance, pay, and retention among national basketball association head coaches. *Journal of Sports Economics 7*, 2 (2006), 119–149.

[62] KALBFLEISCH, J. D., AND PRENTICE, R. L. *The statistical analysis of failure time data*. John Wiley & Sons, 2002.

[63] KAPLAN, E. L., AND MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association 53*, 282 (1958), 457–481.

[64] KIMBER, A. A graphical display for comparing bowlers in cricket. *Teaching Statistics 15*, 3 (1993), 84–86.

[65] KIMBER, A. C., AND HANSFORD, A. R. A statistical analysis of batting in cricket. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* (1993), 443–455.

[66] KIRKPATRICK, S., GELATT, C. D., AND VECCHI, M. P. Optimization by simulated annealing. *science 220*, 4598 (1983), 671–680.

[67] KLEIN, J. P., AND MOESCHBERGER, M. L. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

[68] KLEINBAUM, D. G., AND KLEIN, M. *Survival analysis: a self-learning text*. Springer Science & Business Media, 2006.

[69] LANE, W. R., LOONEY, S. W., AND WANSLEY, J. W. An application of the Cox proportional hazards model to bank failure. *Journal of Banking & Finance 10*, 4 (1986), 511–531.

[70] LEMMER, H. H. The single match approach to strike rate adjustments in batting performance measures in cricket. *Journal of sports science & medicine 10*, 4 (2011), 630.

[71] MAHALANOBIS, P. C. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta) 2* (1936), 49–55.

[72] MASSEY JR, F. J. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association 46*, 253 (1951), 68–78.

[73] MOONEY, C. Z., DUVAL, R. D., AND DUVAL, R. *Bootstrapping: A nonparametric approach to statistical inference*. No. 94-95. Sage, 1993.

[74] OBERHOFER, H., PHILIPPOVICH, T., AND WINNER, H. Firm Survival in Professional Sports: Evidence from the German Football League. *Journal of Sports Economics* (2012), 1527002512462582.

[75] O'BRIEN, R. M. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity 41*, 5 (2007), 673–690.

[76] OHKUSA, Y. An empirical examination of the quit behavior of professional baseball players in Japan. *Journal of Sports Economics 2*, 1 (2001), 80–88.

[77] PATEL, A. K., BRACEWELL, P. J., AND ROONEY, S. J. Team rating optimisation for T20 cricket. In *ANZIAM Mathsport* (2016).

[78] PRENTICE, R. L., WILLIAMS, B. J., AND PETERSON, A. V. On the regression analysis of multivariate failure time data. *Biometrika 68*, 2 (1981), 373–379.

[79] PRICEWATERHOUSECOOPERS. Changing the game: Outlook report for the global sports market, Dec 2011.

[80] R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[81] ROHDE, N. An "Economic" Ranking of Batters in Test Cricket. *Economic Papers: A journal of applied economics and policy 30*, 4 (2011), 455–465.

[82] SARKAR, S., AND BANERJEE, A. Measuring batting consistency and comparing batting greats in test cricket: innovative applications of statistical tools. *DECISION* (2016), 1–36.

[83] SAVAGE, I. R. Contributions to the theory of rank order statistics-the two-sample case. *The Annals of Mathematical Statistics 27*, 3 (1956), 590–615.

[84] SCHUMAKER, R. P., SOLIEMAN, O. K., AND CHEN, H. *Sports Data Mining*. Springer, 2010.

[85] SCULLY, G. W. Managerial efficiency and survivability in professional team sports. *Managerial and Decision Economics* (1994), 403–411.

[86] SHROPSHIRE, K. L. *The business of sports*. Jones & Bartlett Publishers, 2011.

[87] SPIEGELHALTER, D., THOMAS, A., BEST, N., AND LUNN, D. WinBUGS Version 1.4 User Manual, Cambridge: Medical Research Council Biostatistics Unit, 2000.

[88] STEVENS, J. P., AND PITUCH, K. A. *Applied multivariate statistics for the social sciences*. Routledge, 2012.

[89] STEVENSON, M., AND EPICENTRE, I. An introduction to survival analysis. *EpiCentre, IVABS, Massey University* (2009).

[90] STEVENSON, O. G., AND BREWER, B. J. Bayesian survival analysis of batsmen in Test cricket. *Journal of Quantitative Analysis in Sports 13*, 1 (2017), 25–36.

[91] SWARTZ, T. B., GILL, P. S., BEAUDOIN, D., AND DESILVA, B. M. Optimal batting orders in one-day cricket. *Computers & operations research 33*, 7 (2006), 1939–1950.

[92] TUGGY, M. L., AND ONG, R. Injury risk factors among telemark skiers. *The American journal of sports medicine 28*, 1 (2000), 83–89.

[93] VOLZ, B. Minority status and managerial survival in major league baseball. *Journal of Sports Economics* (2009).

[94] VROOMAN, J. The economic structure of the NFL. In *The Economics of the National Football League*. Springer, 2012, pp. 7–31.