Determining batting production possibility frontiers in One Day International (ODI) cricket[*+]

Scott Brooker

Economics Department, University of Canterbury, Christchurch, New Zealand

Abstract

Individual players in One Day International (ODI) cricket can have substantially different skill sets. In this paper we outline how a dynamic programming model of ODI cricket can provide the necessary information to calculate a production possibility frontier for an individual player. Our method is based on the concept of revealed preference, under the assumption that a rational batsman assesses the impact of scoring rates and survival probabilities on the expected outcome of the innings when deciding on the appropriate level of risk to take.

[*] This paper was prepared especially for the New Zealand Association of Economists (NZAE) Conference.

[+] I would like to acknowledge the outstanding support of Dr. Seamus Hogan, Economics Department, University of Canterbury. I am also grateful to New Zealand Cricket as well as Sport and Recreation New Zealand (SPARC) for providing me with the means, opportunity and required data to conduct this research.

1 Introduction

In One Day International (ODI) cricket[1], a batsman faces a trade off between the rate at which they can score runs and their probability of survival. If a batsman attempts to score at a faster rate then they usually are required to take more risk. Some of examples include the batsman attempting to loft the ball over the fielders (risking being caught), attempting to run with a lower degree of certainty that he will make it (risking being run out) and attempting to hit the ball harder (risking being out in one of a number of ways due to having less control of the bat). In this paper, we outline a method with which we can determine the trade-off between scoring rate and survival for an individual batsman. In order to determine the optimal strategy for a team or individual batsman to apply we need to develop production possibility frontiers (PPFs) to determine the skill set of individual batsmen.

1.1 A hypothetical case

Our goal is to observe the trade off between expected runs and the probability of survival. Unfortunately we cannot observe these variables directly. We are only able to observe the outcome of each ball in terms of number of runs scored and whether or not a wicket fell. To properly observe this trade off, we require information about the risk intentions of batsmen when particular results are achieved. We show the usefulness of this by hypothetical example. Imagine that we observe a batsman; we call him D. Bradman, over the course of his career of several years, facing 10000 balls

[1] We assume that the reader has a basic understanding of the structure of a game of One Day International (ODI) cricket; should this not be the case we recommend the reading of Appendix 1 before continuing.

with the outcomes displayed in Table 1.1. Over this period, our hypothetical D. Bradman scored 9994 runs from 10000 balls faced, a scoring rate of 0.9994 runs per ball. Bradman was also out 100 times over this period, giving him a survival rate of 99.00%. This gives us some overall idea of our hypothetical Bradman's ability but tells us nothing about how his scoring rate and survival probability change given when he adopts different risk strategies.

Table 1.1: Summary of D. Bradman's batting outcomes over observed sample.

Outcome	Number of Occurrences	Percentage of Occurrences
Zero Runs	4656	46.56%
One Run	3344	33.44%
Two Runs	500	5.00%
Three Runs	250	2.50%
Four Runs	1000	10.00%
Five Runs	0	0.00%
Six Runs	150	1.50%
Out	100	1.00%

We now add some more information to our hypothetical model. Imagine that our hypothetical batsman informs us that he only ever played with two strategies, a relatively defensive strategy which we will call strategy *a* and a relatively aggressive strategy which we will call strategy *b*. Our hypothetical D. Bradman is equipped with a powerful memory and he informs us that he played exactly half the balls in his career using each strategy and is able to recall which balls were played under which strategy. This information is summarised in Table 1.2 and Table 1.3.

Table 1.2: Summary of D. Bradman's batting outcomes under strategy *a*.

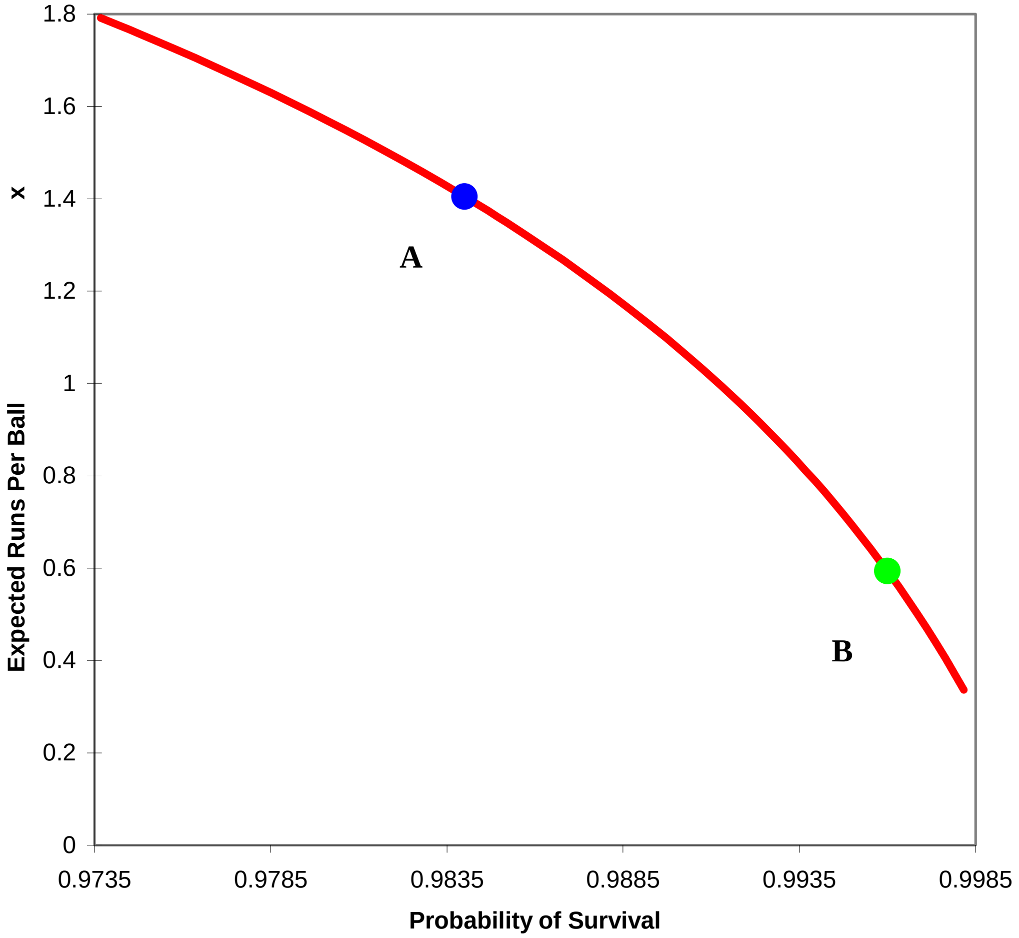
Outcome	Number of Occurrences	Percentage of Occurrences
Zero Runs	3000	60.00%
One Run	1500	30.00%
Two Runs	200	4.00%
Three Runs	50	1.00%
Four Runs	230	4.60%
Five Runs	0	0.00%
Six Runs	0	0.00%
Out	20	0.40%

Table 1.3: Summary of D. Bradman's batting outcomes under strategy *b*.

Outcome	Number of Occurrences	Percentage of Occurrences
Zero Runs	1656	33.12%
One Run	1844	36.88%
Two Runs	300	6.00%
Three Runs	200	2.50%
Four Runs	770	4.00%
Five Runs	0	0.00%
Six Runs	150	3.00%
Out	80	1.60%

When playing strategy *a*, our hypothetical batsman D. Bradman scored a total of 2970 runs from 5000 balls, a scoring rate of 0.5940 runs per ball, while his survival rate was 99.60%. When playing strategy *b,* he scores a total of 7024 runs from 500 balls, a scoring rate of 1.4048 runs per ball, while his survival rate was 98.40%. We now have two points where we know that our hypothetical Bradman was playing different risk strategies and we can infer a basic PPF. Assuming convexity of the production set, that is to say that the higher is the probability of survival, the higher is the marginal cost in terms of expected scoring rate of an additional unit of survival probability, a possible PPF for D. Bradman, if he chose to employ a full range of strategies, is displayed in Figure 1.1. Point "A" represents Bradman's aggressive strategy *a*, while point "B" represents his defensive strategy *b.* It should be noted that with current information, the selection of the points on the PPF other than "A" and "B" are arbitrary; however, this is sufficient for the purpose of illustrating our hypothetical example.

Figure 1.1: A possible PPF for hypothetical batsman D. Bradman



1.2 Inferring the intentions of a batsman

In section 1.1 we outlined a very simple method of finding pairs of scoring rates and survival rates to determine points on the PPF of a hypothetical batsman. In practice, this method meets with a rather large hurdle. It relies on the batsman telling us how much risk he intended to take with each ball. This information is not available to us in any practical way. Even if a batsman wanted to give us this information after a game, the chances of him remembering his exact risk intentions for every single ball of his innings are extremely slim.. One possibility would be to assume that a batsman would take the same level of risk every time he is in exactly the same situation, in terms of number of balls left, number of wickets left and the number of runs required (if the batsman is batting in the second innings). However, we would have a very small set of observations for each "situation" using this method, particularly in the second innings where the additional variable "runs required" results in situations almost never being repeated. We decide to focus our analysis on the first innings and look for some way of grouping similar situations. We cannot say that a batsman is in a similar situation every time he is batting at a certain stage of the innings, as the number of wickets lost certainly plays a role in determining risk strategy. A batsman whose team has lost just four wickets after 45 overs will likely adopt a very aggressive strategy, while if his team has lost eight wickets at the same stage he will be likely to be much more defensive. For similar reasons, we cannot group situations by wickets lost only, ignoring the stage of the innings.

We will show that the key determinant of first innings risk strategy should be the number of runs that a team's expected score falls by if a wicket is lost. The higher is this number, the bigger the potential cost to the batting team of a risky strategy;

therefore, the more defensive is the optimal strategy for the current batsman to employ.

We have stated that each batsman faces a trade off between expected runs and the probability of survival. There also exists a preference trade off, given by some utility function $U(E[r], \eta)$, where $E[r]$ is the expected runs from a particular ball and η is the probability of surviving that ball. Later in this paper, we will show that the indifference curves implied by the utility function can be assumed to be linear (with a few reasonable assumptions) and have a slope that depends on the state of the game. By estimating $E[r]$ and η as functions of that slope we are able to both identify the PPF and test whether a batsman is choosing points on his PPF that are optimal for the game situation.

2 The Theoretical Model

2.1 The Objective Function

The objective function is simple. The rational goal of any team in a game of ODI cricket is to maximise the probability that they win the game.[2] Each team is therefore trying to maximise their probability, π, of winning the game. We write the objective function as

$$\Pr(win)$$

and our optimisation problem as

$$Max(\Pr(win)).$$

[2] There may be rare exceptions, in a multi-stage tournament or league where a team needs to win a game by a particular margin in order to get ahead of another team and qualify for the next round. Alternatively, it might be the case that the team in question simply has to avoid a heavy loss to qualify. We feel that these rare exceptions would not create significant bias.

We assume that, in the range in which first innings totals generally occur, there is a linear relationship between the first innings score and the probability of winning. This means that an extra run is equally valuable regardless of the final score. For example, a score of 261 gives the team batting first the same advantage over a score of 260 as the advantage that a score of 231 would give them over a score of 230. We show evidence of this later in this paper. The implication here is that a team should maximise their expected additional runs, for the vast majority of possible situations that they could be in. We are effectively making current score irrelevant to future decision making. This enables us to revise our objective function for the first innings as

Max(V = Expected Additional Runs)

2.2 The First Innings Value Function

We need to develop a function to calculate the expected additional runs for every possible state. Firstly, however, we must define some variables:

- Let κ = The level of aggression chosen by the batsman.
- Let i = The number of the next ball to be bowled in the innings

  $i \in [1, 2, .., 300]$
- Let j = The current number of wickets lost by the batting team

  $j \in [0, 1, .., 10]$
- Let r_{ij} = The number of runs scored from the i^{th} ball with j wickets lost[3]

  $r_{ij} \in [0, 1, .., \infty]$

- Let λ_{ij} = The probability of losing a wicket on the i^{th} ball with j wickets already lost

  $0 \le \lambda_{ij} \le 1$

- Let γ_{ij} = The probability of a wide or no ball being bowled on the i^{th} ball with j wickets already lost.

  $0 \le \gamma_{ij} \le 1$

- Let τ_{ij} = The number of runs scored from a wide or a no ball.

  $\tau_{ij} \in [0,1,..,\infty]$

- Let R_{ij} = The total number of runs already scored by the team at ball i with j wickets already lost.

  $R_{ij} \in [0,1,..,\infty]$

- Let V(i,j,R_{ij}) = The expected number of additional runs for a team currently in the state of being on the i^{th} ball with j wickets already lost and R_{ij} runs already scored.

  $0 \le V(i,j) \le \infty$

We have a state space of 3311 cells, as a team could be on any one of their 301 balls[4] in the innings and they could have lost any number of wickets from zero to ten. It is very unlikely that a team could lose all ten wickets while still being on ball number one[5]. Likewise, it is also unlikely that a team could survive until ball number 300 without having lost any of their wickets. However, we cover the entire state space with our estimated models. This is partly for reasons of completeness, but more importantly because the value of V in any one cell has an effect on the value of V in

[4] Ball 301 is not actually bowled. This simply refers to the state after the 300th ball of the innings has been bowled.

[5] For this situation to occur all ten wickets would have to fall due to batsmen being run out or stumped from a no ball, or run out from a wide. In all other circumstances the ball is counted if a wicket falls.

other cells. Note that our estimation strategy is to model the expected runs and probability of going out under the assumption that batsmen are optimising, rather than directly solving the dynamic programming model. This ensures that the V-functions in the thin data cells takes into account the data in the thick data cells. We could, in theory, calculate the V-functions simply by taking averages, rather than running a dynamic programme. This approach would, however, lead to no information in the zero data cells and unreliable information in the thin data cells.

We define the Bellman equation, which links each state to its predecessor, as follows:

$$V(i,j,R_{ij}) = \max_{\kappa} \sum_{r=0}^{6} \left(\lambda_{ij}(\kappa) V(i+1, j+1, R_{ij} + r_{ij}) + (1 - \lambda_{ij}(\kappa)) V(i+1, j, R_{ij} + r_{ij}) \right) p(r_{ij} | \gamma) + \frac{\gamma_{ij} \tau_{ij}}{1 - \gamma_{ij}}$$

In words, this equation is saying that the expected additional runs scored by the batting team from their current state of being at the i^{th} ball of their allotted 300, having lost j wickets of their allotted ten and having scored R_{ij} runs already is equal to the value function applicable on the next ball *plus* the expected runs scored from extras. The state on the next ball is always one of two possible states; one more ball and one more wicket than the current state (with probability λ_{ij}) or one more ball and the same number of wickets as the current state (with probability $(1 - \lambda_{ij})$. Note that the final term is the infinite sum of a geometric series as a wide or no ball must be bowled again by the bowling side. This means that we could in theory have an infinite number of consecutive extras.

We have made the assumption that a team's probability of winning is linear in their final total and, therefore, their strategy should be to maximise their expected additional runs. It is therefore reasonable to assume that the current score, $R_{ij},$ will not influence future outcomes. Our value function thus reduces to

$$V(i,j)=\max_{\kappa}\sum_{r=0}^{6}\left(r_{ij}+\lambda_{ij}(\kappa)V(i+1,j+1)+(1-\lambda_{ij}(\kappa))V(i+1,j)\right)p(r_{ij}|\gamma)+\frac{\gamma_{ij}\tau_{ij}}{1-\gamma_{ij}}$$

We note that we do not know the relationship between κ and λ_{ij} or κ and r_{ij}, however by assuming that batsmen behave optimally, we can further reduce our value function to the following

$$V(i,j)=\sum_{r=0}^{6}\left(r_{ij}+\lambda_{ij}V(i+1,j+1)+(1-\lambda_{ij})V(i+1,j)\right)p(r_{ij})+\frac{\gamma_{ij}\tau_{ij}}{1-\gamma_{ij}}$$

We can now estimate the values of Pr(r_{ij}), λ_{ij}, γ_{ij} and τ_{ij} as probit regression models (an ordered probit in the case of Pr(r_{ij})). We can take expectations of r_{ij} from our ordered probit to further simplify the value function. We now have

$$V(i,j)=E[r_{ij}]+\lambda_{ij}V(i+1,j+1)+(1-\lambda_{ij})V(i+1,j)+\frac{\gamma_{ij}\tau_{ij}}{1-\gamma_{ij}}$$

We are subsequently able to determine the value of the value function, $V(i,j)$, for each i and j, using a dynamic programming approach. We have two sets of end points as a team cannot score any further runs once their innings is over. This occurs when 300 balls have been bowled or when ten wickets have been lost, whichever is sooner. This means we can define the end points of the dynamic programme as follows:

$V(301,j)=0$, for all j (1)

$V(i,10)=0$, for all i (2)

The remaining values of the V-function can be calculated by backward induction.

2.3 Choosing the level of risk to optimise the value function

The form of the value function described above is used for our dynamic programming work. It is convenient for our optimisation work to take the compliment of the probability of losing a wicket λ_{ij}. We write η_{ij}, the probability of surviving ball *i* given that a team is currently *j* wickets down, as

$$\eta_{ij} = 1 - \lambda_{ij}$$

This way we are describing two desirable goods, scoring rate and survival, on the axes of our production possibility frontiers. Our value function is now as follows

$$V(i,j) = E[r_{ij}] + (1-\eta_{ij})V(i+1, j+1) + \eta_{ij}V(i+1, j) + \frac{\gamma_{ij}\tau_{ij}}{1-\gamma_{ij}}$$

We implicitly differentiate our value function with respect to η_{ij} to obtain

$$\begin{aligned}\frac{\partial V(i,j)}{\partial \eta_{ij}} &= \frac{\partial E[r_{ij}]}{\partial \eta_{ij}} + \frac{\partial(1-\eta_{ij})}{\partial \eta_{ij}}V(i+1,j+1) + \frac{\partial V(i+1,j+1)}{\partial \eta_{ij}}(1-\eta_{ij}) \\ &+ \frac{\partial \eta_{ij}}{\partial \eta_{ij}}V(i+1,j) + \frac{\partial V(i+1,j)}{\partial \eta_{ij}}\eta_{ij} + \frac{\frac{\partial \gamma_{ij}\tau_{ij}}{\partial \eta_{ij}}(1-\gamma_{ij}) - \gamma_{ij}\tau_{ij}\frac{\partial(1-\gamma_{ij})}{\partial \eta_{ij}}}{(1-\gamma_{ij})^2}\end{aligned} \tag{1}$$

As we want to maximise the value function at ball *i*, the value functions at *V(i+1, j+1)* and *V(i+1, j)* can be considered constant terms as we will maximise by starting at the end of the innings and working backwards. Therefore, equation (1) simplifies to:

$$\begin{aligned}\frac{\partial V(i,j)}{\partial \eta_{ij}} &= \frac{\partial E[r_{ij}]}{\partial \eta_{ij}} + (-1)V(i+1,j+1) + (0)(1-\eta_{ij}) + (1)V(i+1,j) + (0)\eta_{ij} \\ &+ \frac{(0)(1-\gamma_{ij}) - \gamma_{ij}\tau_{ij}(0)}{(1-\gamma_{ij})^2}\end{aligned} \tag{2}$$

At the maximum value of *V(i, j)*, the derivative of *V(i, j)* with respect to η_{ij} is equal to zero. Substituting this first order condition into equation (2) gives us the following

$$\frac{\partial E[r_{ij}]}{\partial \eta_{ij}} = V(i+1,j+1) - V(i+1,j) \tag{3}$$

Equation (3) reveals that, in any state of the first innings, the batting team's trade off between run scoring and survival is equal to the negative of the cost of a wicket. Let

C(*i, j*) = the cost to the batting side of losing a wicket on ball *i*, given that they have previously lost a total of *j* wickets.

$$\frac{\partial E[r_{ij}]}{\partial \eta_{ij}} = -C(i, j) \tag{4}$$

Equation (4) holds significant implications for strategy. It is saying that, given the value function, the levels of risk that a batting side should be indifferent between are those where the trade off between run scoring and survival is equal to the negative of the cost of a wicket. For given values of *i* and *j*, the cost of a wicket is a constant; therefore, the batting team's indifference curve is linear.

We have outlined the fact that there are different combinations of runs and survival between which the batting team is indifferent. Now we must address the capabilities of an individual batsman. On every ball, a batsman must decide how much risk he wants to take. Each level of risk results in some number of expected runs and some probability of survival, for that individual batsman. These numbers can be used to form the PPF for that batsman. A batsman will be optimising if he takes the level of risk that places him at the point where his PPF is tangential to the indifference curve of the team, given by the cost of a wicket at that stage of the game. We will assume that a batsman's PPF is continuous, monotonic in η_{ij} and weakly concave.

Our weak concavity assumption means that a batsman will have to give up a higher amount of expected runs in order to get an extra unit of survival, the higher the probability of survival he already has and vice versa. We can theoretically justify our concavity assumption by reasoning that a batsman, over the course of several balls, can reach any point on the straight line between two points of his PPF by mixing between those two strategies.

We illustrate an example of a batsman's PPF and his optimal choices in Figure 2.1, using three potential game situations. For clarity, we do not display the indifference curves that do not contain an optimal point on the PPF. At the beginning of the first innings, this particular batsman wants to be operating at point A, where he is taking a low level of risk, because the cost of a wicket is high here. As this particular innings progresses, assuming the wickets falling are of the batsmen at the other end, this batsmen moves to higher risk point B in the middle of the innings, then to very high risk point C at the end of the innings, where the cost of a wicket is extremely low.

Figure 2.1: The PPF with optimal points

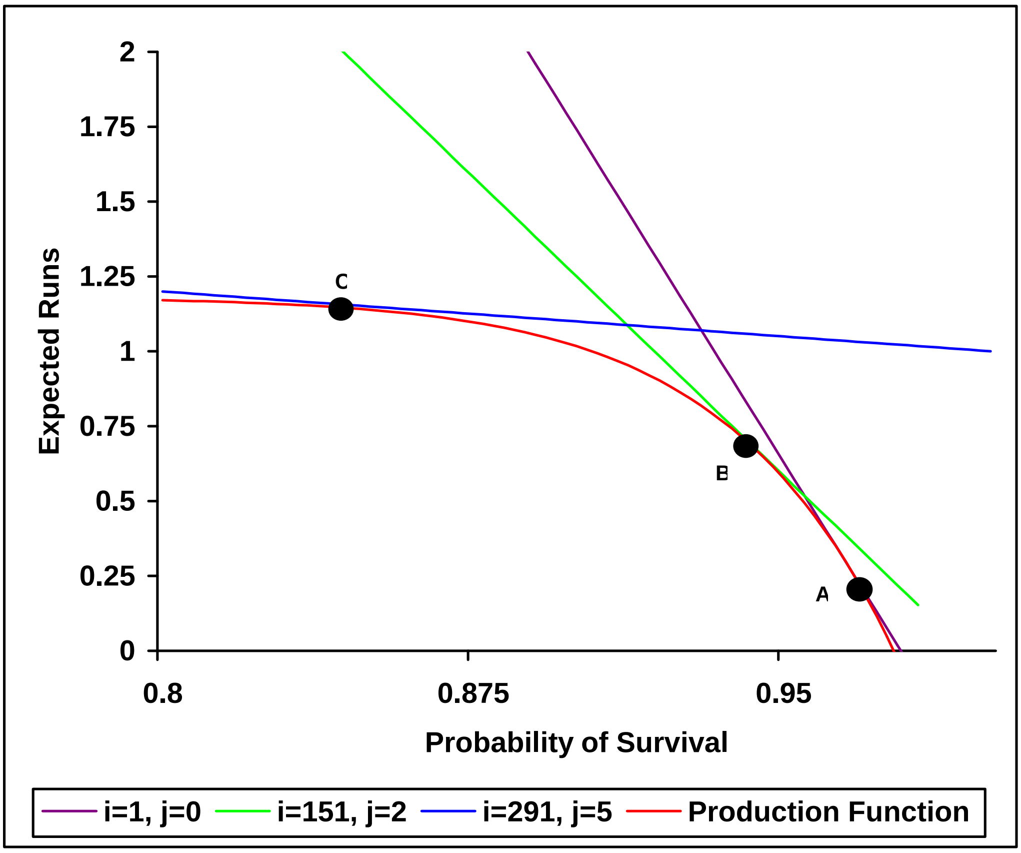
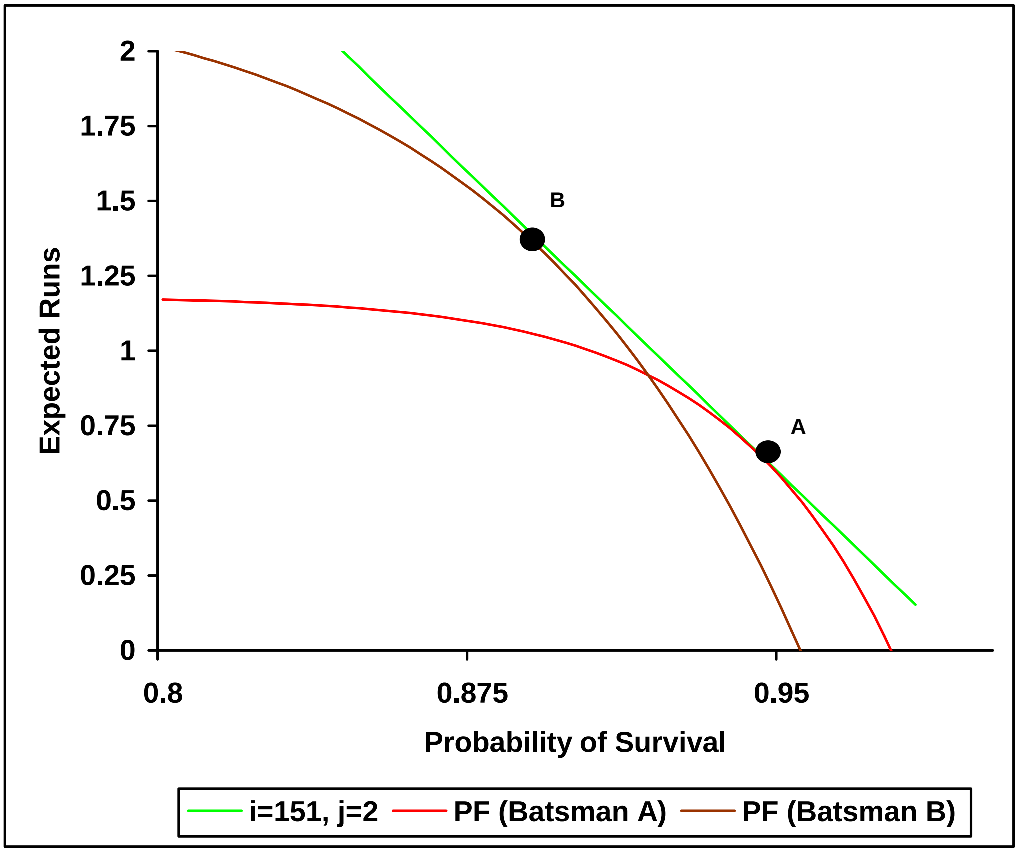


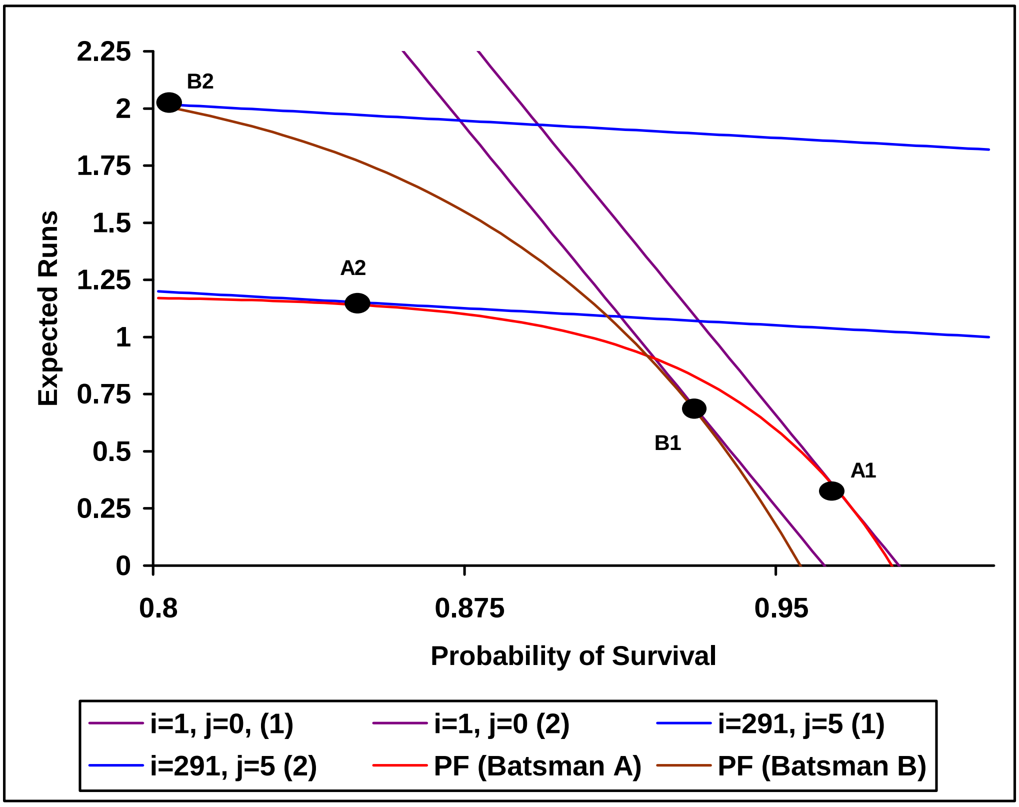
Figure 2.1 showed the PPF of a particular, but imaginary, batsman. We can also illustrate that different batsmen have different optimal points on their PPFs. Some batsmen will have a PPF that is completely inside the PPF of another batsman, indicating that the former is an inferior batsman in every way. Figure 2.2 illustrates an example where two batsmen have different strengths. Batsman A is a relatively better defensive player, as he will score more runs than Batsman B when maintaining a high probability of survival, while Batsman B is a more effective attacking player. As shown on the graph, where the cost of a wicket is at the level implied by this indifference curve, Batsman A will optimally operate at point A, while Batsman B will optimally operate at the relatively higher risk point B.

Figure 2.2: Comparing the PPFs of two batsmen



We note from figure 2.2 that point A is unattainable for Batsman B and point B is unattainable for Batsman A. The optimal point for each batsman appears on the same indifference curve, indicating that the batting team would be indifferent between optimising Batsman A and optimising Batsman B being the current batsman. This would not usually be the case as we show in figure 2.3. If the cost of a wicket was higher than the indifference curve in figure 2.2 implies, the batting side would have a steeper indifference curve and Batsman A's optimal point (A1) would be on a higher indifference curve than Batsman B's optimal point (B1). This is because a more defensive strategy is preferred in this situation and Batsman A is a better defensive player. If, however, the cost of a wicket was lower than the indifference curve in figure 2.2 implies, a more attacking strategy is preferred and Batsman B's (B2) optimal point is on a higher indifference curve than Batsman A's optimal point (A2).

Figure 2.3: Comparing the strengths of two batsmen



3 Analytics

3.1 Data Sources and Timeframe

The data used is a set of 311 matches over the period 20 July 2001 to 25 January 2008**.** It consists of ball-by-ball information collected by New Zealand Cricket.

3.2 A Structural Break

A major rule change occurred in ODI cricket during the period of our data set. All matches played prior to 1 July, 2005 required that the fielding side could place no more than two fielders outside an approximate oval (know as the “circle”) drawn 30 yards from either end of the pitch for the first 90 balls of each innings. This is a fielding restriction as compared to the five fielders allowed outside the circle for the remainder of the innings. In contrast, matches played between 1 July, 2005 and 30 September, 2007 required the above restriction to be in place for the first 60 balls of an innings and for two additional periods of 30 balls, the timing of which were decided by the fielding captain. These 30-ball periods are known as “power plays”. A smaller rule change occurred on 1 October 2007, from when fielding sides were allowed three fielders, rather than two, outside the restricted area during the second power play[6].

The increased presence of fielders close to the batsman and the lack of fielders patrolling the boundary serve to increase both scoring rates and the risk of a batsman getting out. There are generally more runs available but it is more difficult to score

[6] As from 1 October 2008, this rule has significantly changed again as now the batting side is responsible for electing the timing of one of the power plays and both power plays allow three fielders outside the circle.

these runs without hitting the ball over the top of the fielders, rather than along the ground, resulting in the batsman risking hitting a catch. Before we move forward with our analysis, we assume that the minor rule change allowing three fielders in the restricted area during the second power play has no significant effect. By far the more significant rule change is the extension of the fielding restrictions from 90 balls to 120 balls in total. This enables us to divide our data set into two subsets, matches played without the power play rule (Era 1) and with the power play rule (Era 2). Era 1 contains 185 matches while Era 2 contains 126 matches.

3.3 The effect of conditions

Cricket is a sport where the venue conditions on the day of a match can have a marked influence on the amount of runs that a team can score. The variation in scores from conditions is due to three main factors; the size of the venue, the nature of the pitch and the overhead conditions. Smaller venues make it easier to hit the ball to or over the boundary, resulting in higher scores. Pitches can be fast or slow, green or brown, wet or dry, all of which can affect the ease of scoring. Overcast conditions along with humidity can significantly increase the propensity of the ball to swing, making batting more difficult.

In our dataset we do not have a direct measure of conditions; instead we must create the measure. This research is outside the score of this paper. For the purposes of the analysis in this paper, it is sufficient to know that our conditions measure is a normal distribution conditional on the final runs scored in the first innings and the binary outcome of the match. We therefore have a distribution for each match and we draw from this distribution in order to incorporate conditions into our models. We assume that total scores are a function of two distinct factors; these factors having an additive relationship. Let $S = \rho + \chi$, where S is the first innings score, ρ is a measure of "Performance" and χ is a measure of "Conditions". This allows us to observe and model performance independently of conditions; for example, scoring 280 in conditions where the average score would be 250 is equivalent to scoring 230 in conditions where the average score would be 200.

3.4 Testing the linearity assumption

We stated earlier when determining the objective function for the first innings that there is a linear relationship between the first innings score and the probability of winning. This enabled us to define the objective function for the team batting first as to maximise the expected additional runs from any point. In order to text this assumption we look at the relationship between actual first innings scores and the percentage of games won with each score. Since we might have very few observations (in some cases no observations) at each score *S*, we need to smooth the data. Our method is to look at a range of scores in the vicinity of *S*. We use the 41-point interval (*S-20, S+20)* and calculate the percentage of games won by the first team to bat in this interval and we repeat this analysis for each value of *S*.

We split our data set into two parts; those games played prior to the power play rule change (Era 1) and those games played after the rule change (Era 2). Figure 3.1 shows the relationship between first innings score and the percentage of games won in Era 1, while Figure 3.2 shows the relationship for Era 2. We include a 95% Wilson confidence interval for the estimated proportion of wins, as recommended by Brown et al (2001). In addition, we run a simple linear regression model for each era in order to compare these models with the observed (smoothed) win percentages. The linear model equations are:

For Era 1: Win% = -0.5708 + 0.0045*S

For Era 2: Win% = -0.4235 + 0.0036*S

Figure 3.1: Smoothed relationship between score and win percentage in Era 1

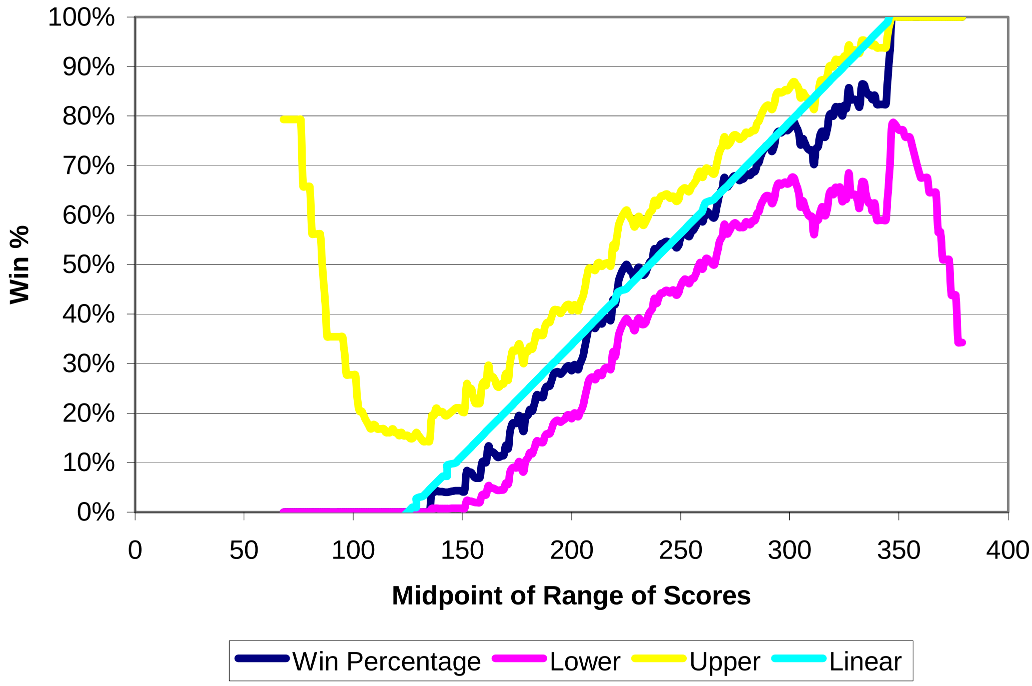
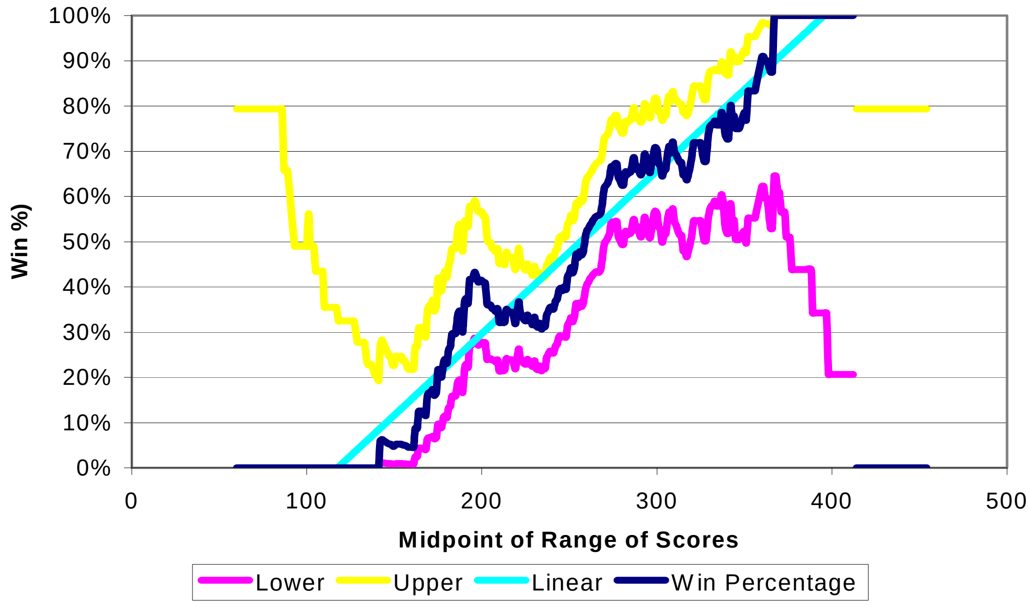


Figure 3.2: Smoothed relationship between score and win percentage in Era 2



It is apparent that assuming a linear relationship between first innings score and the probability of winning is appropriate for the Era 1 data. Any deviations of the linear model from the observed win percentage are well inside the confidence intervals for the majority of observed scores. In Era 2, the linearity assumption does not fit the data as well. Two aspects of Figure 3.2 are of particular note. There is an unexpected decreasing trend in the range of scores (194, 234) and the win percentage falls away to zero, rather than the intuitive level of one, when scores get extremely high. The latter situation is because the sample size is one at these points; the game with the highest score simply happened to result in a loss for the team batting first. The decreasing trend, however, is more difficult to explain but a possible cause would be if conditions were worth relatively low amounts over this range when compared to the scores. When we incorporate conditions into our models we implicitly make this assumption as our conditional distributions for conditions have a lower mean when the game was lost by the team batting first, given a certain value of first innings score.

We want to assess whether we can make the assumption that the probability of winning is linear in the performance of the teams in the first innings. Recall that $S = \rho + \chi$, where S is the first innings score, ρ is a measure of "Performance" and χ is a measure of "Conditions". We sample from the conditional distribution of χ (given S and the result of the match) for each match 100 000 times. We then isolate ρ by subtracting χ from S and calculate a win percentage for each value of ρ. Figure 3.3 (Era 1) and Figure 3.4 (Era 2) show the results. We fit straight lines over the values of performance that appear to have a relationship with the probability of winning that is approximately linear. This occurs for probabilities between approximately 0.1 and 0.8 in Era 1 and between approximately 0.15 and 0.8 in Era 2. In the more extreme ranges the relationship is non-linear. Arguably, it is the games where first innings performances provide a probability of winning in the region of 50% that playing the optimal strategy would have the greatest impact.

Figure 3.3: Relationship between performance and win percentage in Era 1

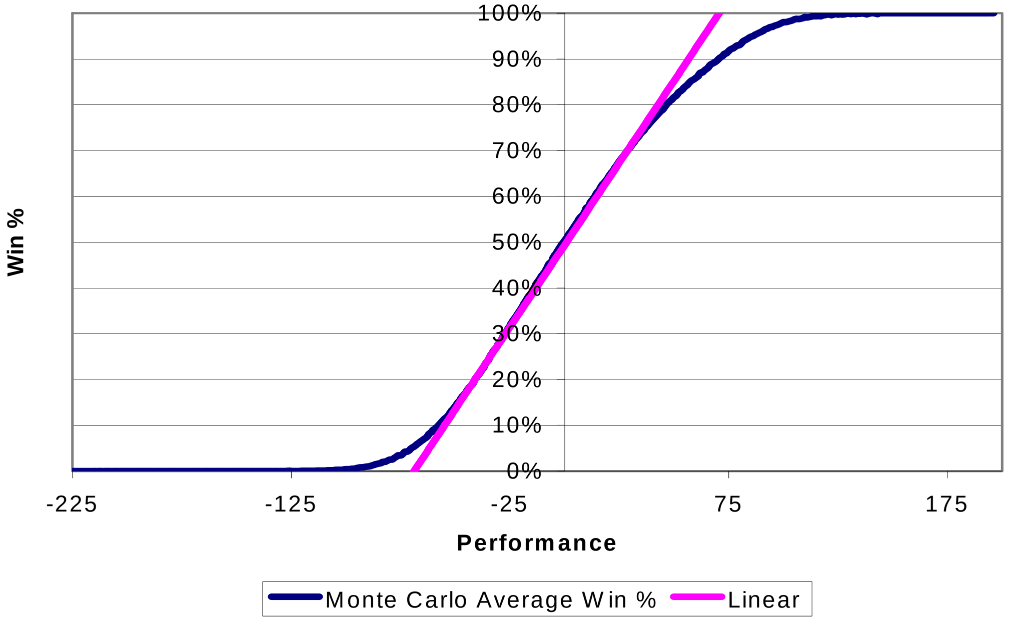
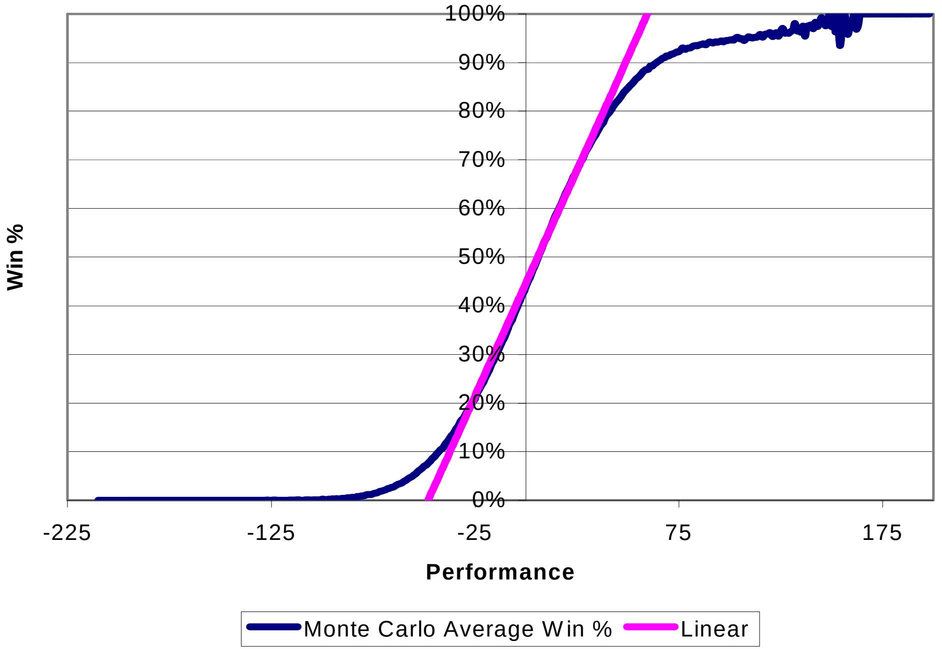


Figure 3.4: Relationship between performance and win percentage in Era 2



3.5 Expected Additional Runs Models without Conditions

Recall our value function for the expected additional runs scored from any point of the first innings.

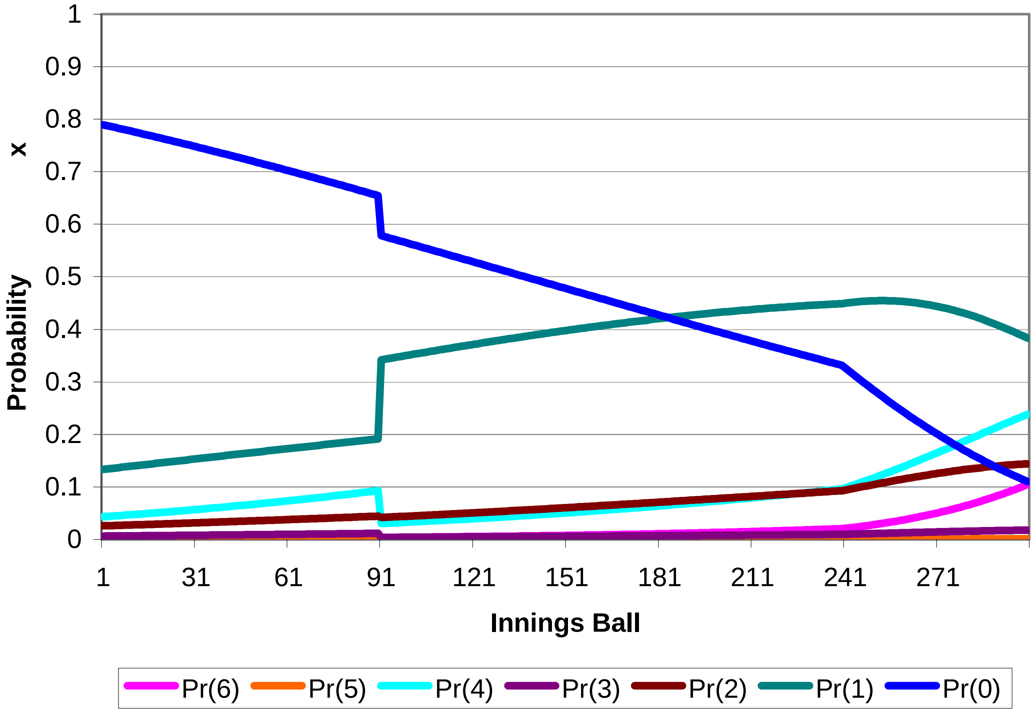
$$V(i,j) = E[r_{ij}] + \lambda_{ij} V(i+1, j+1) + (1-\lambda_{ij}) V(i+1, j) + \frac{\gamma_{ij}\tau_{ij}}{1-\gamma_{ij}}$$

In order to solve the dynamic programme, we need to model r_{ij}, λ_{ij}, γ_{ij} and τ_{ij}. We do this by creating probit regression models. The full modelling process is very detailed and is outside the scope of this paper; however, we briefly outline the models created for the two most important variables, r_{ij} and λ_{ij}, below. The set of possible explanatory variables used in each of models are innings ball (*i*), current wickets lost (*j*) and the rules in place at the time of each match (*Era*). In these models we do not include a variable for conditions; we will show the effect of including this variable in Section 3.4. We are effectively assuming here that all conditions are the same.

For reasons involving the confidentiality of our results, for the remainder of the paper we focus our analysis on Era 1, the games played under the old rules.

The model of r_{ij} is an ordered probit regression based on the variables *i, j* and *Era*. There are seven possible whole numbered values that r_{ij} can take; that is, the whole numbers from zero to six, although five is extremely rare. We show the probabilities of each number of runs being scored from a given ball *i* in Figure 3.5, using the example of being two wickets down. The structural break after ball 90 of the innings occurs due to the fielding restrictions being lifted at that time of the innings under Era 1 rules. The other structural break in the model, a slope change, was imposed after ball 240 to better fit the data. As expected, the probability of scoring a zero is high at the start and steadily drops throughout the innings. The probability of scoring a single increases sharply after the end of the fielding restrictions (ball 90), but falls away at the end of the innings as a team just two wickets down at this point would be more concerned with scoring boundaries than singles.

Figure 3.5: Runs probabilities for Era 1 | *j* = 2

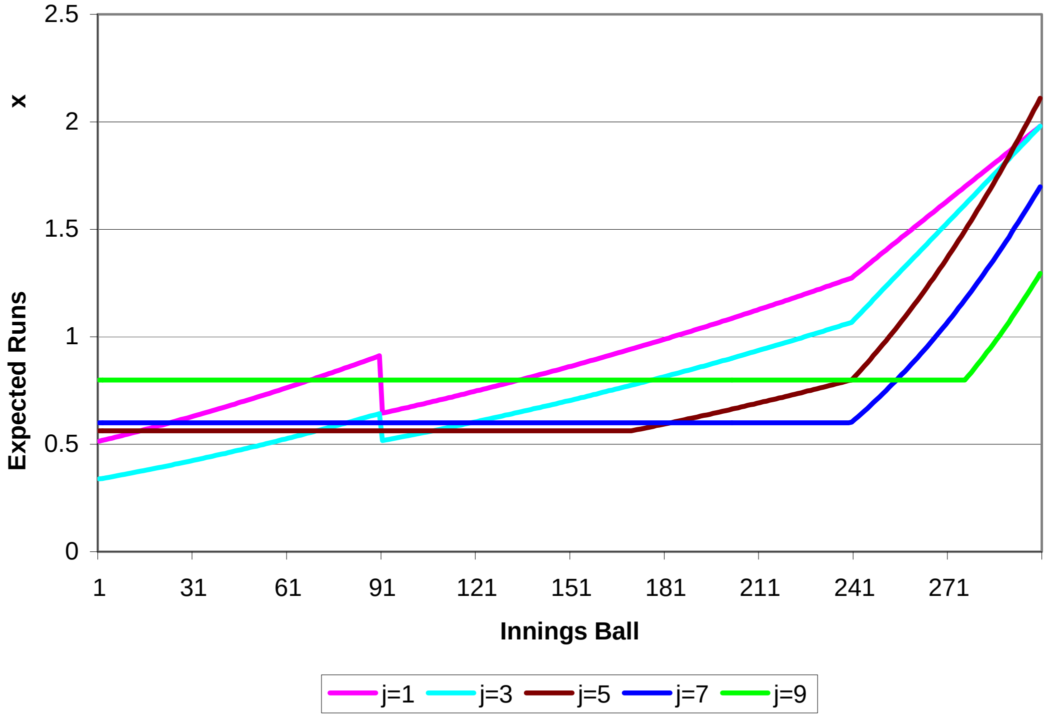


Since our value function does not directly require the probability of scoring a particular number of runs from a given ball, we take the expected value of the runs functions as

$$E(r_{ij}) = 0*\Pr(0) + 1*\Pr(1) + 2*\Pr(2) + 3*\Pr(3) + 4*\Pr(4) + 5*\Pr(5) + 6*\Pr(6)$$

We plot the expected runs functions, for odd-numbered wickets (for clarity), in Figure 3.6.

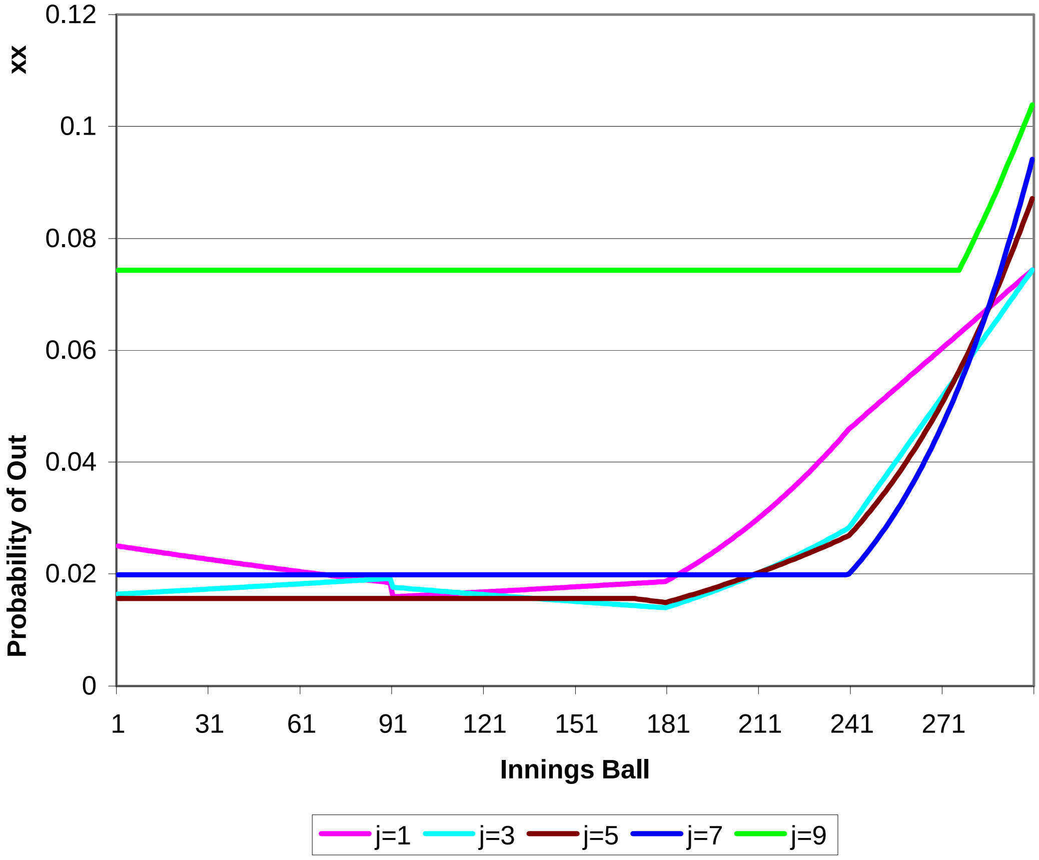
Figure 3.6: Expected Runs Functions for Era 1



Note that our expected runs functions are generally upward sloping. This is what our intuition tells us as teams should want to take more risks and score more quickly as their balls remaining constraint becomes tighter while the wickets remaining constraint remains constant. There is very little crossover of the functions, indicating that teams, for the most part, score at a faster rate the less wickets that they have lost.

The model of λ_{ij} is a probit regression based on the variables *i, j* and *Era*. We show the probabilities of losing a wicket, given a team is a certain number of (odd-numbered) wickets down already, in Figure 3.7. These are clearly more difficult to get an intuitive understanding of than the expected runs functions. With runs, batsmen higher in the order should take more risks (for a given number of wickets lost) and are usually better players. Both these considerations lead to a higher expected runs for the lower numbers of wickets lost than the higher ones. With the probability of getting out, what we should expect is not so clear. On one hand the players earlier in the order have more ability and we would expect them to have a lower chance of going out, but on the other hand for a given value of *i*, the lower is the value of *j*, the more risk it makes sense for the team to take. There are also periods where the functions are downward-sloping; this can be explained by factors beyond the batting team's control such as facing the new ball, which in most cases is more dangerous than the older ball, and the level of aggression in where the fielding captain places his fielders.

Figure 3.7: Probability of a Wicket Functions for Era 1



We have applied two adjustments to the "raw" probit models for r_{ij} and λ_{ij}, using our knowledge of the game of cricket to make corrections in the game situations where data is thin and we believe the model is getting it wrong.

The first adjustment involved forcing the lines for zero to four wickets lost to finish on the same value as we assume that there is no difference in the ability of the batsmen at the crease on ball 300, given that there are two specialist batsmen at the crease.

The second adjustment involved identifying the probability of the team surviving (avoiding losing all ten wickets) until the end of the innings from a given situation. We assume that where this probability is very low (less than 10%), a team will play with the same strategy at all lower values of *i*, for a given value of *j*; that is, the balls remaining constraint is an insignificant factor as the team is unlikely to survive long enough for it to matter. We therefore impose, for each *j*, the value of the expected runs function or probability of a wicket function at the *i* where the survival probability first falls below 10% (when investigating each point in a descending fashion) on all *i* from earlier stages of the innings. We start by applying this process for all *i* for our $j = 9$ function and work backwards through the remaining values of *j*. This is because the adjusted probability of a wicket function for $j = 9$ will affect the survival probabilities for the other values of *j*.

Once we have estimated all the parameters of our value function, we run our dynamic programme, beginning with the two situations that result in the end of the innings and working backwards. Recall that these situations are when 300 balls have been bowled or when 10 wickets have been lost. We plot ten expected additional runs functions *V(i, 0) to V(i, 9)* in Figure 3.8. We notice that the functions are quite flat early in the innings, with large differences between their levels. As the innings progresses, the functions become steeper and the differences between the functions is not as great. This is to be expected as the wickets remaining constraint becomes less important relative to the balls remaining constraint, as more and more balls are bowled in the innings.

Figure 3.8: Expected Additional Runs functions for Era 1

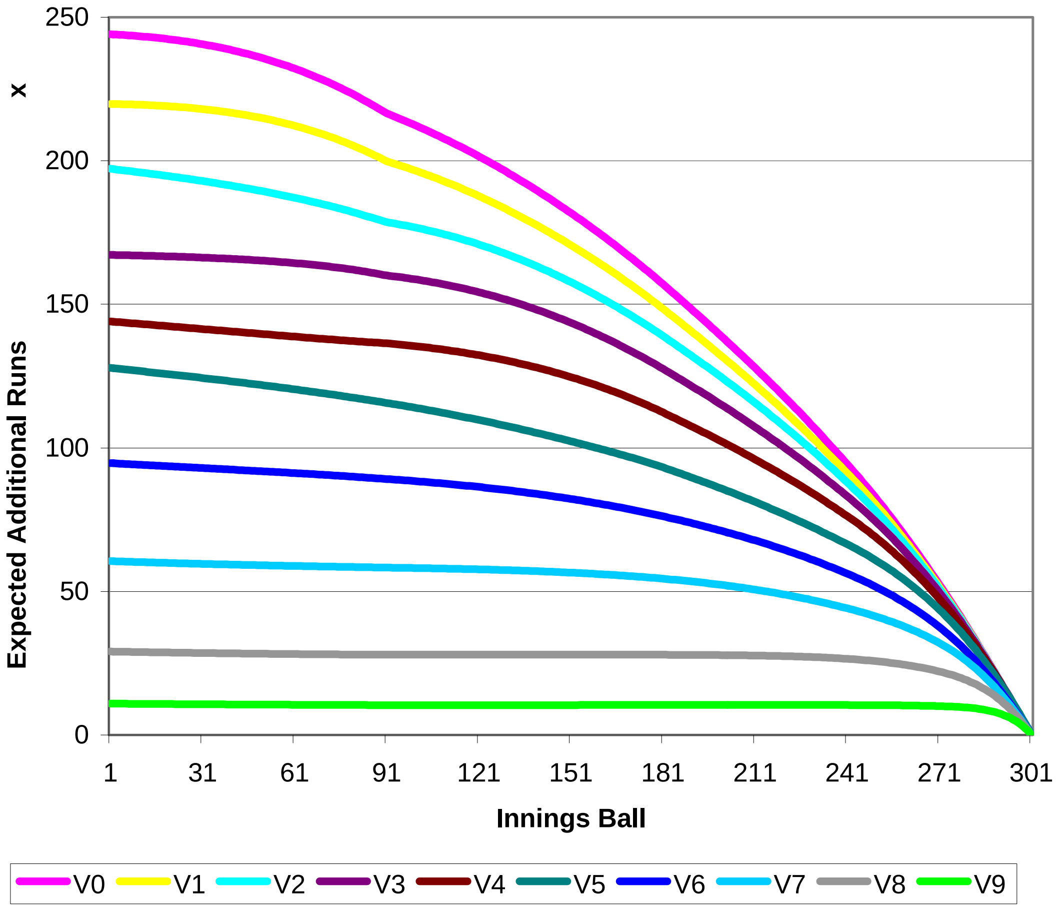
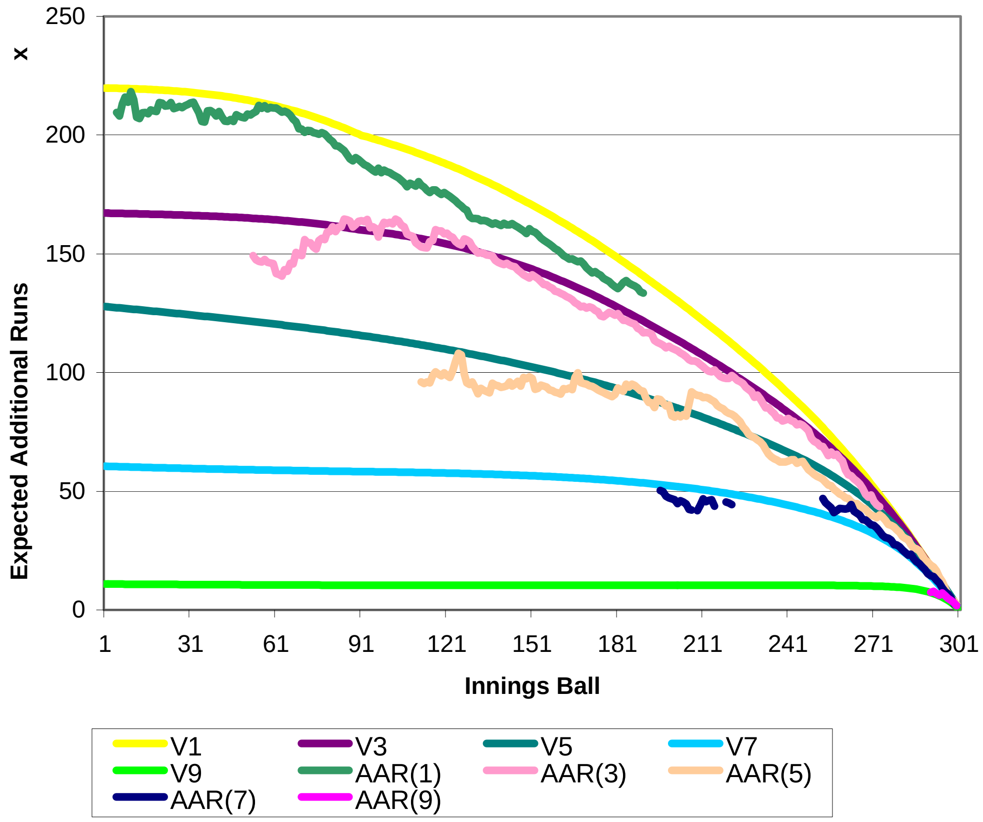


Figure 3.9: V-functions vs actual average additional runs for Era 1



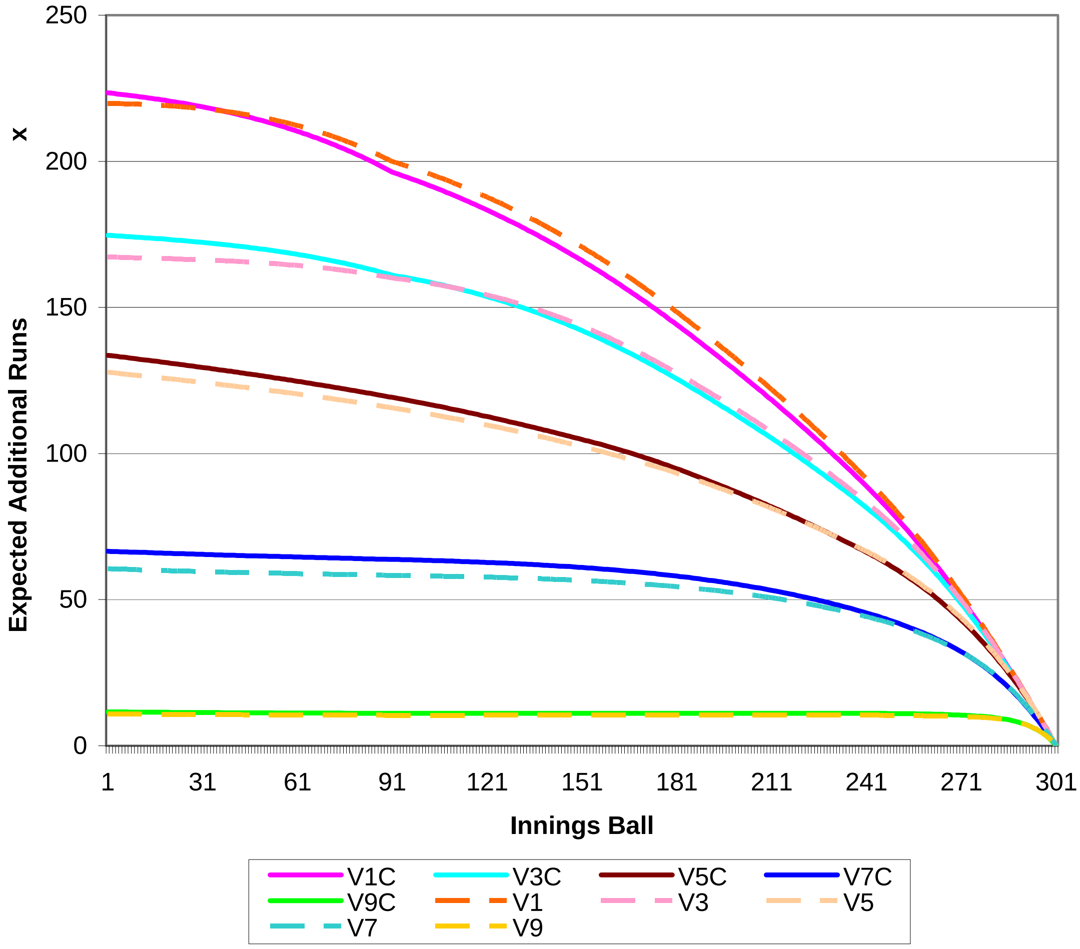
In Figure 3.9 we compare the expected additional runs functions with the actual average additional runs from each point in the innings. We should only (*i, j*) combinations for which we have at least 30 observations and again for clarity we show the odd-numbered values of *j*. There is some over prediction for the early wickets but overall the model fits the data fairly well.

3.5 Expected Additional Runs Models With Conditions

Having constructed and assessed the basic Expected Additional Runs models, we extend this work by including our variable for conditions as an explanatory variable in each of our probit models. We expect this to have two main effects. Firstly, our model in the "average" conditions should differ as previously some of the coefficients on the variables in the probit regressions would be implicitly taking into account the effect of conditions. Secondly, it gives us an insight into how the different the shape of the expected additional runs functions might be under different conditions.

We plot the expected additional runs functions for the odd numbered values of *j* in Figure 3.10. The solid lines refer to the model created with the conditions variable (assuming the conditions expect an average score of 244, which was the score predicted by the model without conditions), while the solid lines refer to the model without the conditions variable. A clear implication of including the conditions variable is that the expected additional runs functions tend to have a slightly larger negative slope. This is an intuitive result. Consider the situation where a team has lost a wicket very early in the innings. This may have been due to poor batting, good bowling, poor batting conditions or simply bad luck; we do not know which combination of these factors caused the early wicket. Given an early wicket has fallen, however, it is slightly more likely than average that we are in worse than average batting conditions. With no further information, the model without the conditions variable predicts a slightly lower expected additional runs value than the model where we know that conditions are in fact average. In the latter case, we are effectively ruling out conditions as the cause of our early wicket. The opposite applies in situations where we have lost a low number of wickets given the point of the innings. The model without the knowledge of conditions then assumes we are on a better-than-average pitch and overstates our future run-scoring potential.

Figure 3.10: V-functions with and without conditions variable for Era 1



We now want to investigate whether the nature of the expected additional runs functions change as we impose different conditions. One method of adjusting for different conditions would simply be to take the model in approximately average conditions and scale it up or down using a constant multiplicative scale factor. In order to test this, we take the functions for conditions worth 250 and apply a scale factor to reduce the V(1,0) cell to 200. We then compare these scaled functions with the functions calculated by imposing conditions of 200 in the dynamic programming model. The results for the odd-numbered values of *j* are shown in Figure 3.11. We repeat this analysis for conditions worth 300 and show these results in Figure 3.12.

Figure 3.11: V-functions with conditions worth 200 vs scaled values

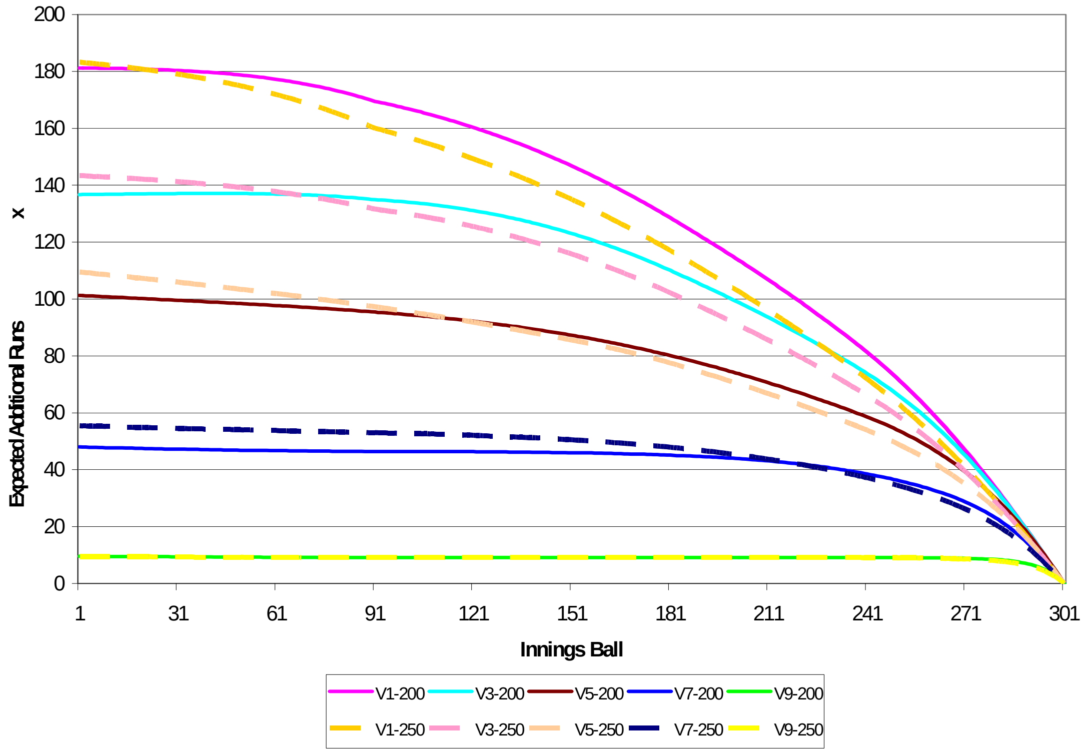


Figure 3.12: V-functions with conditions worth 300 vs scaled values

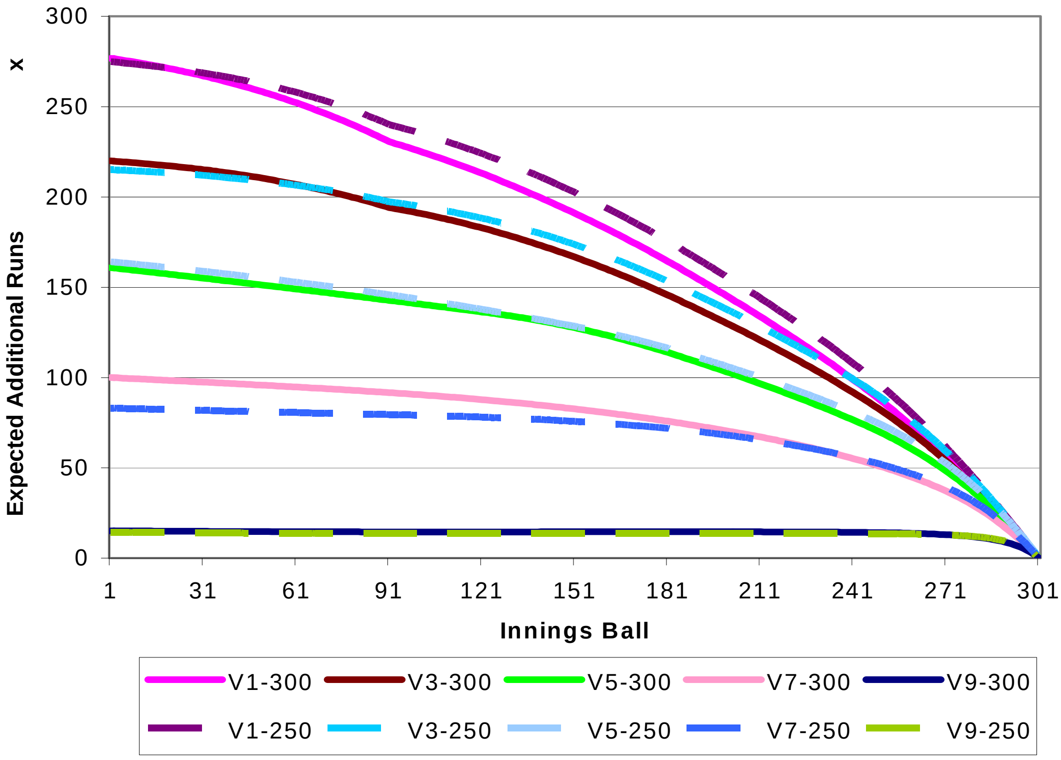


Figure 3.11 shows that for early stages of the innings, relative to the number of wickets lost, the V-functions that include the true conditions value of 200 have a flatter slope than the scaled V-functions where conditions are 250. This implies that is the early stage of the innings that tends to be causing more than its fair share of the difference between scores in 250 conditions and scores in 200 conditions. This is also true for 300 conditions, as seen in Figure 3.12. In this case, the V-functions including the true conditions value of 300 tend to have a steeper slope than the scaled values in the early stages of the innings.

Consider an example of an extreme situation where a team gets a very good start in terms of survival and has survived 90 balls of the innings without losing a wicket. We show in Table 3.1 the score that this team would need to have at this point in order to be still on target for an average score in the conditions. This simple example indicates the need for teams to score a higher percentage of their runs early in the innings, the higher is the value of conditions.

Table 3.1: Zero Wickets Lost at Ball 91

Conditions	Expected Additional Runs	Current Score Required	% of Total Score Required
200	184	16	8.0%
250	215	35	14.0%
300	246	54	18.0%

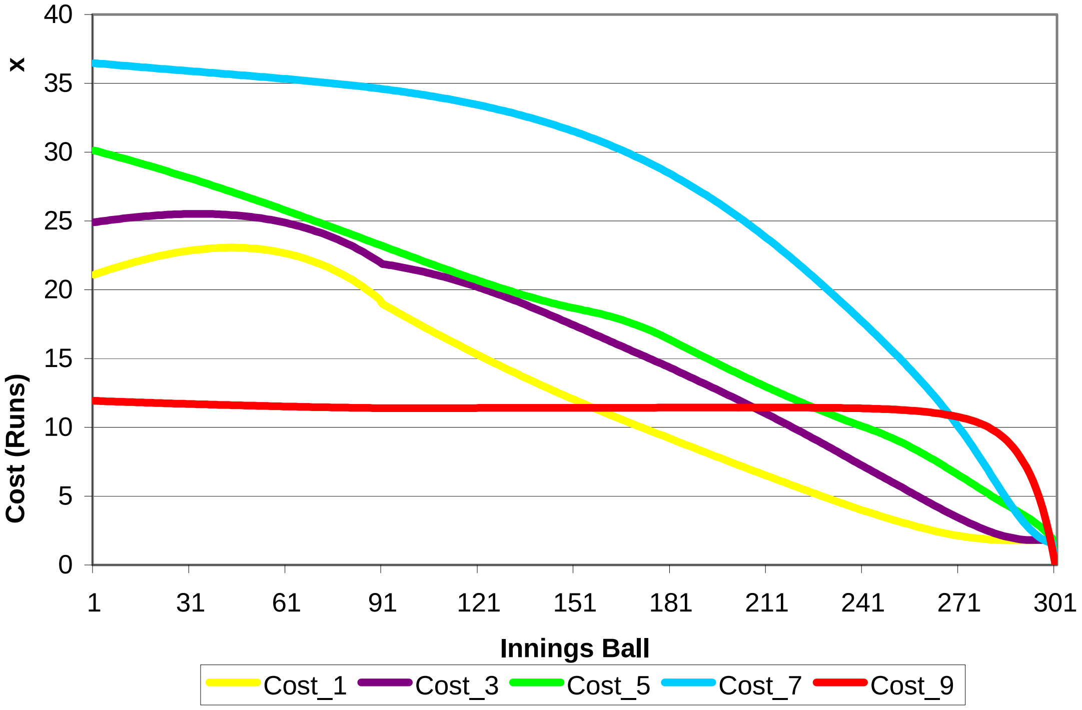
3.6 The Cost of a Wicket Functions

Recall that the key determinant of first innings optimal strategy is C(*i, j*), the cost of losing a wicket on ball *i* given that *j* wickets have already been lost. Having identified our expected additional runs functions *V*(i,j), we can now estimate the cost of losing a wicket as

$$C(i, j) = V(i+1, j+1) - V(i+1, j)$$

As the V-functions vary under different conditions the cost functions also vary; therefore we have a set of cost functions for each type of conditions. In Figure 3.13 we plot the cost functions for odd-numbered values of *j* where conditions are equal to 250. Note that the costs are not ordinal in the number of wickets lost; this is an indication of the complexity of the decision that players have to make on the field in terms of their risk strategies.

Figure 3.13: C-functions with conditions worth 250



3.7 Inferring the Production Possibility Frontiers

We determined earlier that $\frac{\partial E[r_{ij}]}{\partial \eta_{ij}} = V(i+1, j+1) - V(i+1, j)$; that is, assuming concavity in a batsman's PPF, he maximises the expected additional runs function by choosing the level of risk where the slope of his PPF is equivalent to the (negative) cost of getting out. We propose that a rational batsman will adjust their chosen level of risk as the cost of getting out changes. Note that we are not assuming that a batsman will select the optimal level of risk; rather, we are suggesting that the cost of a wicket provides a good reference point for us to compare a batsman's trade-off between scoring rates and the probability of survival. This will enable us to determine what a batsman is capable of and we subsequently can use this information to investigate what the optimal strategy of that batsman should be.

We group our data by individual batsman, era (old rules or new rules) and whether or not in the fielding restrictions period of the innings. For each group we calculate an ordered logit regression model to regress our runs from ball variable, r_{ij}, on our cost of a wicket variable and conditions variable, *C(i,j)* and χ_{ij} respectively. We expect that the cost of a wicket variable will influence the runs variable negatively as batsmen should be more defensive when it is expensive to lose a wicket. We also expect that the conditions variable will influence the runs variable positively as it should be easier to score more quickly in easier batting conditions. We use a logit regression model to regress our binary survival variable, η_{ij}, on *C(i,j)* and χ_{ij}. We expect that the cost of a wicket variable will influence the survival variable positively as survival should be more important to a batsman when the cost of his wicket is high. We also expect that the conditions variable will influence the survival variable positively as survival should be easier in easier batting conditions.

Our runs regression gives us the probabilities of scoring each number of runs; therefore, we take expectations of these values to get expected runs per ball $E[r]$. We then plot the combinations of $E[r]$ and $\Pr(\eta = 1)$ implied by each value of C, for a given value of χ . Some examples of the results are shown in Figures 3.14, 3.15 and 3.16. These results all use games in Era 1, with the fielding restrictions not in place and conditions of 200, 250 and 300 respectively. The functions are for a prominent New Zealand batsman and a prominent Australian batsman. We are able to show their comparative abilities in poor, approximately average and good batting conditions. Note that these functions are currently in the early stages of development.

Figure 3.14: PPFs where χ = 200

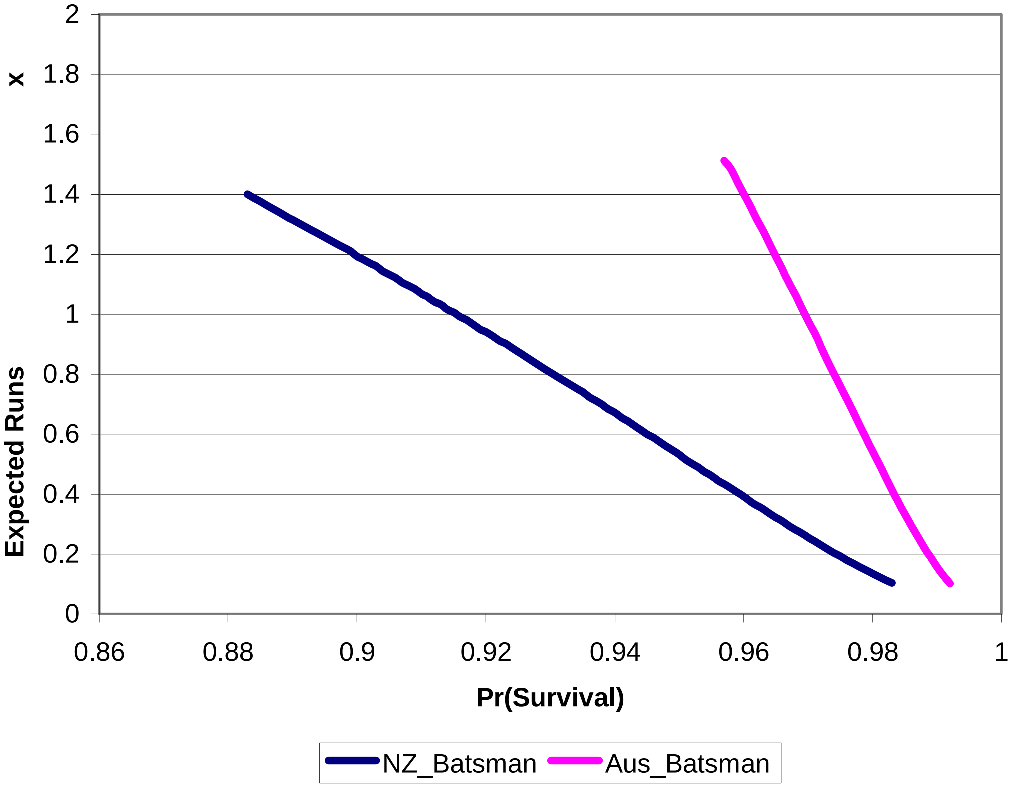


Figure 3.15: PPFs where χ = 250

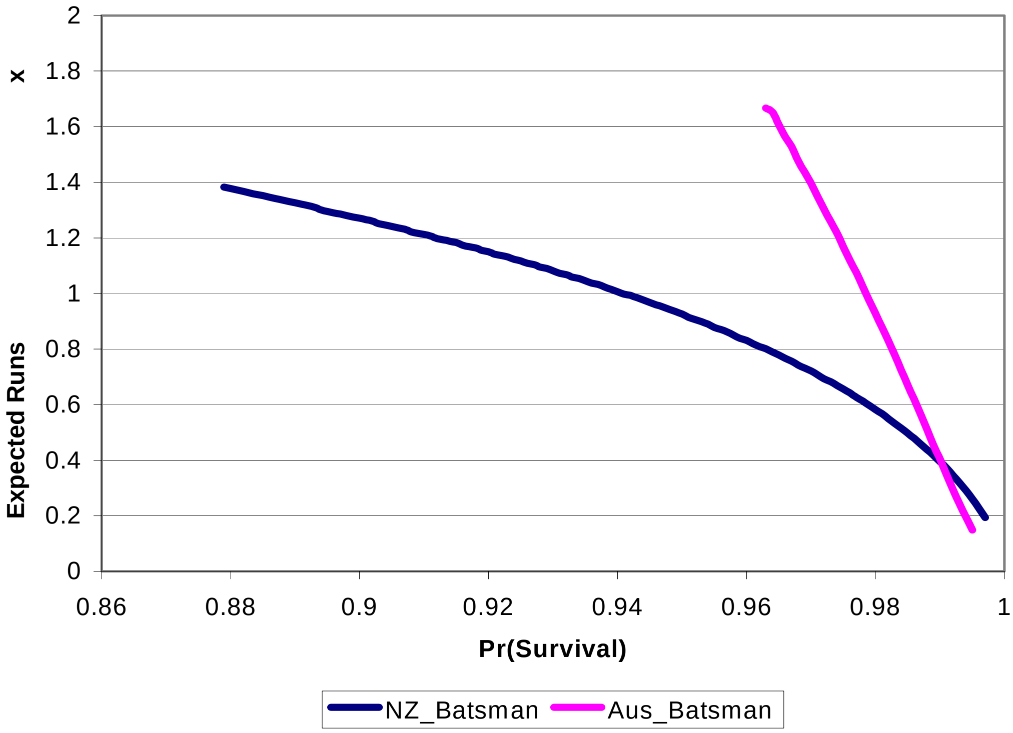


Figure 3.16: PPFs where χ = 300

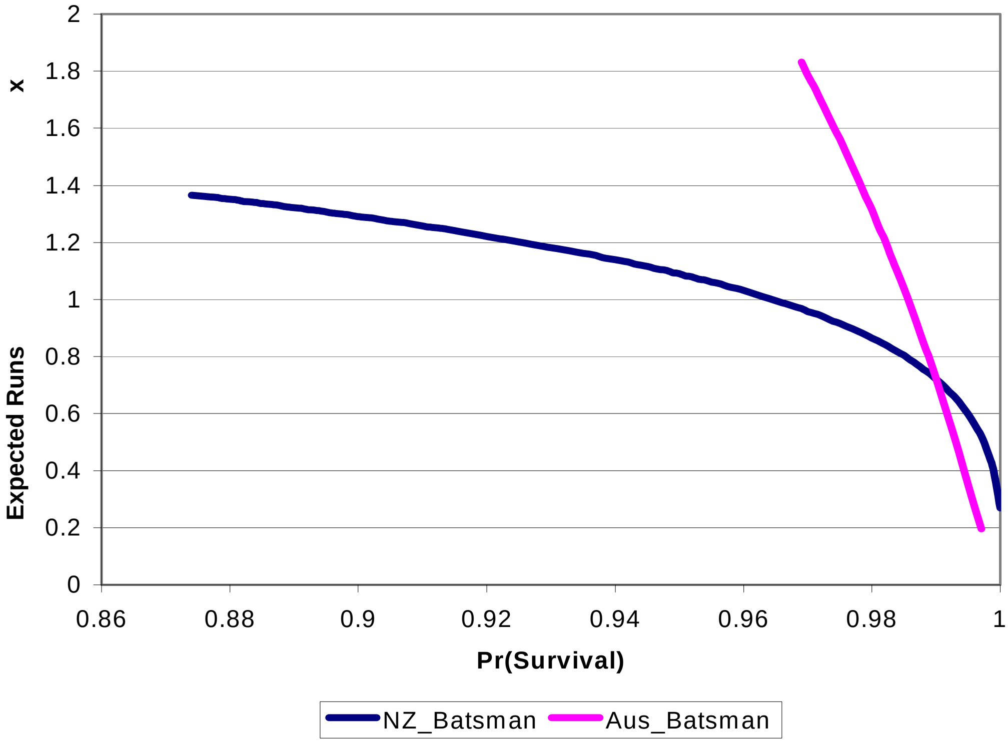


Figure 3.14 shows that the Australian batsman is expected to outperform the New Zealand batsman in all possible game situations when the game is played under poor batting conditions. We consider two factors when making this statement. Firstly, for all rates of run scoring that are common to both functions, the Australian batsman is able to achieve each rate with a higher probability of survival than the New Zealand batsman. Additionally, the Australian batsman's maximum possible scoring rate and maximum possible survival probability are higher than those of his New Zealand rival.

Analysing Figures 3.15 and 3.16 result in different conclusions. In average and good batting conditions, in certain situations the New Zealand batsman would be expected to outperform the Australian batsman. At very high survival rates, the New Zealand batsman is able to score runs at a faster rate than the Australian batsman. The Australian batsman maintains his significant advantage at lower survival rates. The New Zealand batsman would therefore be preferred in situations where survival is of utmost importance; that is, when the cost of a wicket is very high.

4 Conclusions

By developing a dynamic programming model of expected additional runs, we have obtained a new variable, the cost of losing a wicket. We are able to use this variable as a reference point to assess player ability and determine production possibility frontiers. Possible applications of these PPFs include assessing strategic performance and determining the optimal strategy for any point of the innings. This includes determining the appropriate risk that each batsman should take, given the game situation, and the appropriate batsman for a given situation as teams are not required to select their batting order in advance. We hope, with further work, to be able to quantify the levels of overall score that are possible, given that a team as a whole behaves optimally.

References

BROWN, L.D., TONY CAI, T, AND DASGUPTA, A. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science* Vol. 16 No. 2 pp 101-133.

Appendix 1: The necessary basics of the game of cricket

Cricket is a sport played between two teams of 11 players on a large, approximately circular field with a 22-yard-long strip of pressed clay, soil and grass known as a “pitch” in the centre. One team will initially be the bowlers and the other team will be the batsmen. All 11 members of the bowling team are on the field while only two members of the batting team are on the field at any one time. The basic idea of the game is relatively simple. A bowler bowls a ball from one end of the pitch by releasing it with a straight arm action in the direction of the batsman. The ball will usually bounce once before reaching the batsman. The two main goals of a batsman are to score “runs” and avoid getting “out”. A run is scored each time a batsman, having hit the ball with his bat, running to swap ends of the pitch with the other batsman. Alternatively, a batsman may score an automatic four or six runs by hitting the ball so far that it leaves the playing field. These automatic runs are known as “boundaries”, with four being scored if the ball bounces before leaving the playing field and six otherwise. If a batsman is “out” then his turn at batting is over and he must leave the field to be replaced by a team mate.

A batsman may be “out” in a number of ways; however, we outline only the most frequent below:

- Bowled – when the ball, having been bowled by the bowler, hits any of three wooden poles positioned at the batsman’s end of the pitch.
- Caught – when the batsman hits the ball in the air and the ball is subsequently caught by any member of the fielding team.
- Leg Before Wicket (LBW) – We will not discuss the complicated aspects of this method of going “out”; however, in general a batsman is out LBW if the

ball, having not been hit by the bat, strikes the batsman's body (almost always the leg) and would have otherwise hit the wickets.

- Run Out – If the batsman attempt to score runs by running to the opposite end of the pitch and the fielding team hits the wickets at either with the ball before the batsman running to that end reaches a line at the end of the pitch known as a "crease".
- Stumped – When the batsman advances towards the ball, misses it with his bat and the wickets are hit with the ball by the wicketkeeper (the fielder standing behind the wickets) before the batsman can get back to the "crease".

The batting side may continue batting until ten of the 11 members of their side are out, then the two teams switch roles. A team's turn at batting is called an innings and each team will have either one or two innings depending on the type of game. In general, the team that scores the highest number of runs wins the game.

There are three main versions of the game. In test cricket, the traditional form of the game, each team bats for two innings and a match lasts a maximum of five days, with the match being declared a draw if it is not finished in this time. One Day International (ODI) cricket allows each team to bat for one innings but with a limit of 300 balls per innings. The innings finishes when ten batsmen are out or the 300 balls are up. As the name suggests, this type of game is all over in a day, running for approximately eight hours. Twenty20 cricket is the newest form of the game and is similar to ODI cricket except that the limit is 120 balls per innings and the game takes approximately three hours. In this paper, we consider only ODI cricket.