University of Warwick

MORSE: Final year project

An Analysis of the Duckworth Lewis Method in the context of English County Cricket

Author:
Jack Jewson (1113570)

Supervisor:
Prof. Simon French

March 6, 2017

Acknowledgments

- Firstly I would like to thank Professor Simon French for the support, encouragement and statistical guidance that I received from him throughout the completion of this project.
- I am also grateful to an anonymous County Cricket Club and their performance analyst for providing me with the main data set used throughout the project as well as expert opinions and insights into the game.
- Finally I would like to thank Frank Duckworth, Tony Lewis and Steven Stern for viewing and critiquing sections of this report as well as providing intuition and reasoning behind some of the judgements made when building the model.

Abstract

This project looks at the Duckworth Lewis Method, the incumbent rain rule used to decide the result of a Limited Overs cricket match should it not be able to reach its natural conclusion, and its application within English County Cricket. Some of the key value judgements made be the model builders are examined to see how they impact the fairness of the model when viewed from various different points during a match or a season. The data set was obtained containing complete data from three and a half seasons of County Cricket. This data was examined visually and then, using ideas from Repairable Systems Reliability and Nonlinear Regression, it was modelled. The resulting analysis identified several situations where the model could be considered unfair and discusses how the model parameters could be updated to combat this. Finally, it is recognised that there is no right or wrong answer to the point in time at which the model should be fair. This, along with many other decisions, is simply a value judgement on the part of the model builders and the authorities regulating the game.

Contents

Chapter 1

Introduction

Cricket is a team sport thought to be invented as a gentlemen's social game in England, but now played at international level across the world. Despite only being played at the very highest level in 10 countries, the game stretches across 5 continents and is thought to be the second most popular sport in the world, with an estimated 2.5 billion fans (Wood, 2008). With the introduction of T20 cricket in 2003 the game's popularity reached a new level and since then a number of high profile T20 leagues have been established around the world. As the profile of the game increases more and more money is becoming involved, in terms of players wages, sponsorship and gambling. With so much money now at stake it is more important than ever to have games ending in results and in the event that the game cannot reach its own natural conclusion that a fair method be used to decide upon a winner.

Cricket and Cricketing terminology

A cricket match is played between 2 teams, both consisting of 11 players. At any one time one side will be batting and the other will be bowling. A cricket pitch consists of two sets of three wooden sticks (stumps) set 22 yards apart, a yard and a half in front of each set of stumps is the batting crease where the batsman stands. The bowling (also called the fielding) team bowl the ball from one end of the pitch towards the batsman, standing in front of the stumps at the other end. Its the aim of the batting team to hit this ball and try to score runs either by hitting it far enough to allow enough time for the batsman to run to the other end of the pitch, earning one run, or hitting the ball over the boundary of the playing area, either along the ground earning the batsman 4 runs or without bouncing earning the batsman 6 runs. Balls are bowled in groups of 6 by the same bowler and these are called 'overs'. Batsman bat in pairs, one at each end of the pitch, thus when an odd number of runs are scored the batsman swap over. Whilst trying to score runs the batsman is also trying to foil the bowling team by not getting out. There are 10 ways in which a batsman can get out which are listed in Table 1.1, below. A batting innings can be thought of as that team's turn at batting and once a batsman is out they cannot bat again in that innings, therefore it is the aim of the bowling team to take 10 of the batting wickets (only ten are required as batsman must bat in pairs) and thus end the batting innings. Once the first batting innings is over the roles of the teams reverse.

Table 1.1: Listing the methods of getting out in cricket

Method	Description
Bowled	Where the bowler hits the batsman stumps
Leg Before Wicket	Where the ball is prevented from hitting the stumps by the batsmans pads
Caught	Where the batsman hits the ball and any of the bowling team catch the ball without it bouncing
Run Out	Where the batsman try to complete a run but a member of the fielding team throw the ball at the stumps before the batman can complete the run
Stumped	Similar to Run Out but when the batsman has left his crease, but isnt attempting a run, and a fielder has hit the stumps
Hit wicket	Where the batsman accidently hits his own stumps
Obstructing the Field	Where the batsman deliberately gets in the way of a fielder trying to run him out
Handling the Ball	Where the batsman handles the ball
Double hit	Where the batsman hits the ball twice
Times out	Where the batsman takes longer than the allotted 3 minutes to get to the crease following the dismissal of the batsman before him

Cricket is played professionally in two formats: First Class (Test Match) cricket where the match lasts 5 days and each team have two batting innings. In First Class cricket teams bat until they are out unless they voluntarily end their innings early (declare). If over the course of a teams two inning's they score more than the opposition then they win and if both teams' innings have not finished after 5 days the game is a draw. The other format is Limited Overs cricket. This is slightly different, each team only has one innings and is allotted a maximum number of overs to bat (usually 50, 40 or 20 overs). The team that bat first (known as team 1 from now on) has to try and maximise the number of runs they score in their allotted overs and the team that bat second (team 2) have to try and chase this target by scoring more runs. If team 1 score more runs than team 2, no matter how many wickets either team lose, then team 1 wins and visa versa.

International cricket matches consist of either Test matches, 50 over limited over games called One day internationals (ODIs) and 20 over limited over games (T20s). Domestic county cricket in England is designed to replicate this as closely as possible within the time constraints of the British summer and thus games consist of Test matches, 40 over limited over games called Pro40s, and T20s.

The Duckworth-Lewis Method

Cricket is a summer sport and due to safety concerns with the pitch, plus the reluctance of the players to get wet, the game must stop if it rains to allow the pitch to be protected and the players to stay dry. Rain delays are a very common thing across world cricket especially in countries whose summers suffer from the same volume of rain as the British summer. During a test match a rain break just reduces the duration of the game and as there is no restriction on each teams innings' this causes no issues. Test match cricket is also spread across 5 days making it highly unlikely for the whole match to be stopped because of rain. As a limited overs game only lasts a day and the rules dictate that both teams bat for a set number of overs, the same treatment of rain delays cannot occur. Duckworth and Lewis (1998, p.1) point out that the reason Limited Overs cricket was invented was because too many first class games were ending in draws and players and fans alike craved a shorter format of the game that produced a result in just one day. Therefore considering a rain affected Limited Overs game as a draw is contrary to the reason this format of the game exist, so a fair method of deciding the result of a rain affect Limited Overs game needed to be devised. Many methods were trialled but finally the Duckworth and Lewis method, devised by Frank Duckworth and Anthony Lewis in 1998, was settled upon. Duckworth and Lewis (1998) devised a two factor relationship between the number of overs a team had remaining and the number of wickets they had lost in order to quantify the total resources a team had remaining. This was then divided by the resources each team had at the start of the game to obtain a resource percentage, which Duckworth and Lewis tabulated allowing remaining resource to be calculated for all combinations of overs remaining and resources lost. No matter where a stoppage in play occurred, using this resources table and simple calculations outlined by Duckworth and Lewis its possible to calculate the percentage of their full resources any team has received. And therefore reset a teams target such that both team have to score proportionally the same number of runs in the resources available to them.

Aims

In creating their model, called the D/L model from now on, Duckworth and Lewis made a series of scientific assumptions and value judgements in order to ensure that their model was fair. Many of these are trivial, for example assuming that the altitude at which a match is held or the manufacturer of the cricket ball used has no impact on the way a teams resources are distributed. But many of these come down to judgements that Duckworth and Lewis themselves have made which, in their opinion, ensure that if a rain break occurs, the game is as fair as possible

A lot of these judgements come down to the point in time at which Duckworth and Lewis judge that the game should be fair. Currently Duckworth and Lewis judge that a game of cricket should be fair when viewed from the start of the season. That is before the format of the game has been decided; before the venue of the game or the opposition have been chosen; before the toss of the coin, which decided which team will bat or bowl first and before either team has commenced their innings. By assuming this Duckworth and Lewis are saying that every innings played within a season has its resources distributed in exactly the same way, whether its a 20 over innings or a 50 over inning; whether its a first or a second innings or whether a team is 100-1 or 50-3 their (remaining) resources are distributed in the same way.

Here the aim is to assess whether Duckworth and Lewis judging that a cricket match should be fair at the start of the season is actually making the matches consistently unfair, to either team, when they are played. The specific aim are as follows:

- To examine if Standard D/L is fairly predicting first innings scores in English limited overs cricket

- To see if there exists any correlation between past and future performance in a game of cricket
- To see if resources appear to be distributed in the same way in both the first and second innings of cricket match
- To examine whether it is fair to use one resource table to cover all formats of the limited overs cricket

Chapter 2

Literature Review

Duckworth and Lewis

The Duckworth-Lewis (D/L) method was created by Frank Duckworth and Anthony Lewis in 1998. Their first paper Duckworth and Lewis (1998) highlights the need for a fair way to decide the result of an interrupted game of one day cricket and goes on to propose their method for doing this, which has since been universally adopted by the ICC (International Cricket Council) since 2001 (BBC, 2007). Duckworth and Lewis start by examining previous methods that have been trailed in one day cricket. Most notably the ARR (average runs rate) which requires a team to simply score a higher run rate than the opposition for the overs they have available, is dismissed as favouring the team batting second. Whilst the MPO (most productive overs method), where if team 2 can only bat for x overs then they have to chase the score the team batting first scored in their x highest scoring overs, is criticised for giving an advantage to the team who batted first. Given the perceived failure of all methods tried before Duckworth and Lewis aim to produce a method satisfying 5 criteria they set themselves:

1. The method should be fair for both sides mean the relative position of both teams should not be altered by an interruption
2. Must give a reasonable and sensible target or result in all possible situations
3. Must not depend on the way team 1 scored their runs (as chasing a target in an uninterrupted game isn't)
4. Should be easy to apply requiring just a table and calculator
5. Should be understandable for all people involved in the game

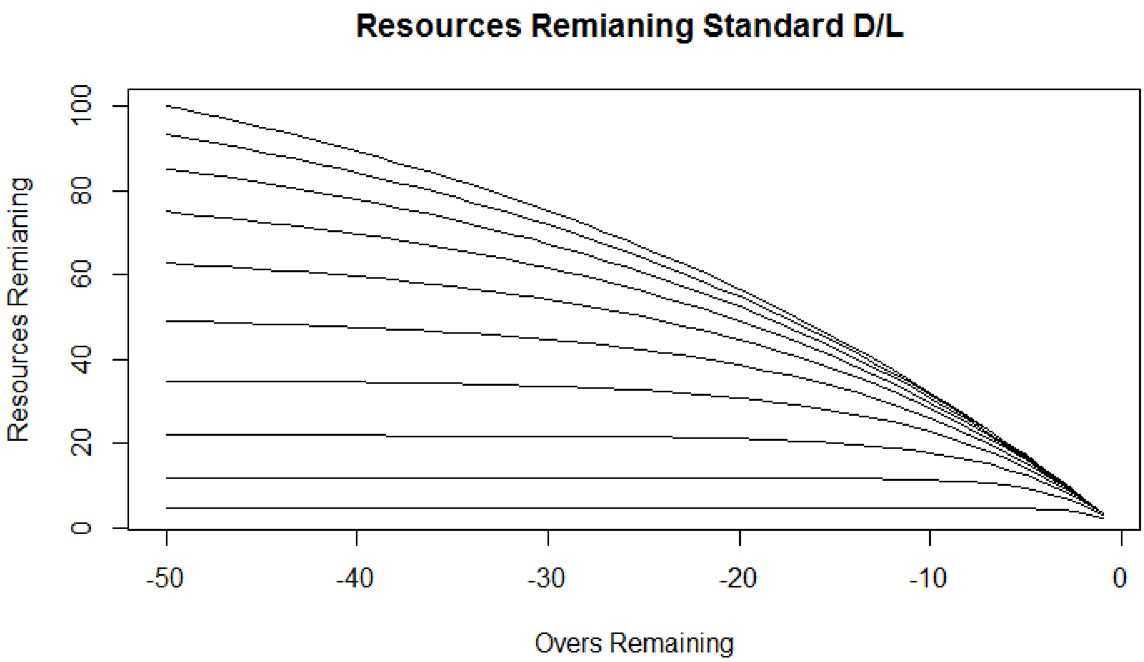
Unlike any of the previous methods, Duckworth and Lewis noticed that when a team bat they have 2 resources with which to score runs: the batsman they have who are yet to get out (wickets remaining) and the overs they have left in the match (overs remaining). When a rain break occurs one of these resources, overs remaining, is reduced where as the number of wickets they have left stays intact. By reducing one resource and not the other the trade off, of maximising runs scored whilst minimising the risk of getting out, that the batting team face, is upset for one side. Duckworth and Lewis realised that if they could construct a two factor relationship equating the number of wickets a team had lost and the number overs a team had remaining to an overall resource percentage, they could then reset the target required to win the match so that the batting team had to score proportionally the same number of runs in the resources they have available to them than the opposition, thus re-balancing the trade off so it was the same for both teams. The two factor relationship they settled on is the exponential decay relationship below:

$$Z(u, w) = Z_0(w)[1 - exp(-b(w)u)]$$

Where u is overs remaining, w is wickets already lost, $Z_0(w)$ is the asymptotic average runs scored by the last $10 - w$ wickets in unlimited overs under one day rules and $b(w)$ is the exponential decay constant which again depends on the number of wickets that have been lost.

Figure 2.1 below was created in R using the 50 over D/L resource table (Appendix 1.1) . The graph shows how the D/L model resources change as overs remaining decrease for different wickets demonstrating the relationship that the D/L model predicts.

Figure 2.1: A demonstration of how the D/L model formula predicts resources to decay as overs remaining decrease for $W = 0, \ldots, 9$



The top curve corresponds to when W=0 and W is increasing as the curves get lower. The constant $Z_0(W)$ governs the asymptotes that these graphs meet as overs remaining tend to infinity. This is D/L acknowledging the limiting factor that wickets impose on runs scoring, even if a team could bat for infinite overs they would still reach a finite score as they would eventually lose all 10 wickets. From the graph its obvious that these asymptotes decrease when W increase as teams who have lost more wickets will be bowled out quicker on average and thus are expected to score less runs in infinite overs. The D/L model then decays this asymptotic average based on the number of overs the team has remaining using decay of the form $1 - exp(-x)$. This was the shape that D/L judged to be appropriate to model resources. The parameter $b(w)$ also depends on the number of wickets the team has lost and this governs the shape of the curves observed above. As wickets decrease the gradient of these curves is decreasing in (modulus) value suggesting that resources decay slower when more wickets have been lost, though due to $Z_0(W)$ they also start lower. This is reflected by an increase in the value of $b(W)$.

Having a decay constant and asymptote both depending on how many wickets the team had lost and multiplying the former by the number of overs the team has remaining allows the D/L resource formula to capture the 2 factor relationship, between overs remaining, wickets lost and resources, that Duckworth and Lewis proposed.

Due to confidentiality Duckworth and Lewis are unfortunately unable to publish the forms that these parameters take, but they explain that they are obtained 'following extensive research and experimentation' (Duckworth and Lewis, 1998, p.4) and $Z(u, w)$ has been tested to check it behaves as expected in all scenarios. In their paper Duckworth and Lewis state that the average score in a 50 over game is 250 it therefore seems reasonable that given this a team might on average score around 350 were they to bat for infinite overs. This puts $Z_0(0)$ around 350 and solving with $Z(50, 0)$=250 gives $b(0)$ approximately equal to 0.025[1]. As mentioned above the Z values are expected to decrease as W increase and the b values are expected to increase.

In order to achieve their criteria 4, Duckworth and Lewis created a table, outlining the percentage resources, $P(u, w)$, a team has remaining for varying u and w, where $P(u, w) = Z(u, w)/Z(N, 0)$ and N is the number of overs the teams had available to them at the start of the match. From these resource percentages the proportion of their full resources that a team has available after a delay in their innings can be calculated. If the team lose $u_1 - u_2$ overs, when the innings is cut short u_2 will be zero, then the proportion of their resources that they actually receive is $R = 1 - P(U_1, W) + P(U_2, W)$.

Therefore if team 1 scored S in R_1 percentage resources available and team 2 has R_2 percentage resources available, then team 2s required target, T, is:

$T = S\frac{R_2}{R_1}$ if $R_1 > R_2$
$T = S$ if $R_1 = R_2$
$T = S + G(N)(R_2 - R_1)$ if $R_1 < R_2$ where $G(N)$ denotes the average runs scored in the first innings by a team batting for N overs.

The final case is required as applying the method when $R_1 > R_2$ to the case where $R_1 < R_2$ involves extrapolating a teams current scoring rate, which could feasibly not be sustainable, and this leads to silly results in some occasions. Duckworth and Lewis believe their method satisfies all 5 of their original criteria and since its incorporation into world cricket, has been used in some form ever since.

Following the full use of D/L for 3 years in international cricket, Duckworth and Lewis (2004) provides a review of how their model is actually performing. Firstly, in order to check the accuracy of their resource percentages, Duckworth and Lewis conduct an analysis of the average runs scored by teams for variable overs remaining and wickets lost. Using data from 330 ODIs and only looking at the first innings, they looked at every possible combination of overs remaining and wickets lost and calculated the average runs teams that achieved this position scored in their remaining resources. Duckworth and

[1]In fact based on the findings of Chapter 7, $Z_0(0)$ was closer to 330 and $b(0)$ was around 0.027

Lewis then compared these observed average runs scored to the runs their D/L model expected teams to score. This was calculated by multiplying all of the D/L resource percentages by $G(50)$, which at the time was 225. This allowed Duckworth and Lewis to see how their estimated resources compared with reality, for example if teams are scoring more for a given overs remaining and wickets lost than D/L expected then the resources percentage must have been under estimated and visa versa. This analysis was conducted by constructing 10 graphs (one for each wicket $w = 0, \ldots, 9$) detailing how average runs score (y-axis) decrease as overs remaining decrease (x-axis). From this analysis Duckworth and Lewis conclude that the shape of their curves associated with their model is consistent with the shape of the observed data. However, occasionally the observed curves lay either constantly just above or just below the D/L curve, indicting the model parameters associated with that wicket were either too large or not large enough, respectively. This suggested that the D/L model was imposing too greater or not great enough penalty on the resource percentage for losing that wicket. They also notice that the observed average first innings runs scored was higher than 225. Following this analysis the D/L model parameters were updated.

Duckworth and Lewis also noticed that their model somewhat broke down when run scoring was well above the average. This is because the runs scoring pattern required to chase these large scores is very different from normal. When chasing a large score the team need to score at a higher rate for much longer than when chasing an average score. Therefore as scores increase the D/L method should actually converge on the ARR method. This lead to D/L giving some seemingly unfair results when teams scored well above the average. To combat this Duckworth and Lewis incorporated two new[2] parameters, λ and n, to create their D/L Professional Edition. With resource equation

$$Z(u, w|\lambda, n(w)) = Z_0 b(w)\lambda^{n(w)+1}[1 - exp(-\frac{bu}{\lambda^{n(w)}b(w)}]$$

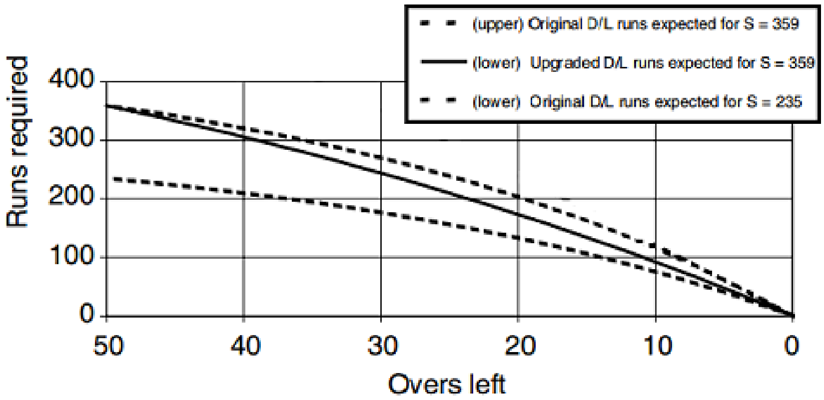
λ allows the D/L model to adjust the targets it sets in light of the fact the team 1 has scored an above average score. In combination with n this allows the D/L curves to be adjusted to allow them to more accurately reflect the pattern of runs scoring required to chase such a large total. This effectively creates an individual resource table for team 2 based on what team 1 has scored. This represents Duckworth and Lewis altering one of their original value judgements, that all innings have the same resource table, in light of evidence that it wasn't fair. This Professional D/L was adopted by the ICC in 2003. Unfortunately, by updating the model in this fashion, Duckworth and Lewis have broken number 4 of the criteria they set themselves as the calculations of λ require a computer. As the programming for this is once again confidential, analysing Professional D/L is unfortunately not possible.

Figure 2.2, below, taken and edited from Duckworth and Lewis (2004, p.9), shows how this new Professional D/L changes the average runs that D/L expects teams to score in their remaining resources when compared with Standard D/L. In this example team 2 are chasing, the well above average, $S = 359$ to win. The lower dotted line shows the Standard D/L average runs obtainable based on the, then, average target 235. The upper dotted line is the average runs obtainable that would be given using the Standard D/L formula but scaled up to chase 359. The dark line in the middle gives the

[2] The parameters $Z_0 b(w)$ and b are just slightly adjusted parameters from thr Standard D/L model

average runs obtainable, given when Professional D/L is used in this situation. All runs required are shown for $W = 0$

Figure 2.2: A demonstration of how the remaining resources predicted by Professional D/L differ from the remaining resources predicted by Standard D/L when chasing the well above average $S = 359$



Duckworth and Lewis (2004, p.10) gives the Professional D/L parameters values for this situation as $n = n(0) = 5$ and $\lambda = 1.149$. In this scenario with λ and n both greater than 1 these parameters combine to increase the asymptote of the Professional D/L resource curve and decrease the value of the exponential decay parameter. Both of these combine to form the the much flatter Professional D/L resource curve that's observed in the graph above. The flatter resource curve achieves the aim that the D/L method should converge towards the ARR method as the target increases. This means team 2 has to score at a higher rate for longer and therefore are expected to score fewer runs in their remaining overs than Standard D/L would have predicted. This is why the Professional D/L curve is always below the Standard D/L curve for this scenario. Had the target S been well below average the parameters, λ and n, would have been less than 1, but greater than 0, reducing the asymptote and increasing the decay parameter. This would act to increase the curvature of the D/L runs required curve allowing teams to score more runs in there remaining overs than Standard D/L currently would.

Criticisms of Duckworth Lewis

The D/L model has come under criticism from players, supporters, journalist and academics and as a result papers have been published outlining possible flaws with the model and attempts to improve upon it. Bhattacharya, Gill and Swartz (2011) question whether the Standard edition of D/L, originally designed to be used in 50 over cricket and created before the invention of 20 over cricket, is suitable for use in T20 cricket. They assert that, due to differing rules and the excessive risk taking encouraged by the short duration of a T20 innings, the runs scoring pattern in T20s may differ from the runs scoring pattern in 50 over cricket and thus D/L, in its current form, may not be suitable. The way the D/L model looks at T20 cricket is it assumes a T20 game is like the last 20 overs of a 50 over game still with 10 wickets in hand. Therefore the D/L resource table for T20s (appendix 1.3) is just a linear transformation of the 50 over resource table and can be created by dividing the resources percentages for the last 20 overs of a 50 over match by 0.566 (the resources left in a 50 over match with 20 overs remaining and 10 wickets in hand).

Bhattacharya, Gill and Swartz identify that the D/L model is parametric but also that there are many non-parametric curves that could be used to fit the data. They therefore suggest that there is possibly some advantage to adopting a non-parametric approach. This resonates well with the objectives of this project. Although the choice of model and model parameters is influenced by observed data ultimately the decision is a judgement made by the model builders. A non-parametric model would be based solely on observed data and thus involve less judgements and could prove fairer. Similarly to Duckworth and Lewis (2004), Bhattacharya, Gill and Swartz use observed data to estimate resources, R, for T20s. R is based on observed data and is therefore obviously non-parametric, however it suffers from the untidiness of observed data and thus Isotonic regression is used to estimate any values missing from the data and to transform R so that resources are at least monotonically decreasing with overs remaining and wickets lost. However in order to create a strictly monotonically decreasing resource table, as a team having lost more wickets or having less overs remaining should have strictly less resources, a Bayesian approach was used. Here estimates of the posterior means of the resources are calculated through Gibbs sampling and these are used to create the final T20 resource table (appendix 1.4). This remains non-parametric as no functional relationship is enforced upon the resources. The fact that this model is non-parametric means it only relies on observed data, therefore in order to create a resource table solely based on observed data would require a reasonable sample of data relating to every combination of balls in a match (300 in a 50 over game) and wickets lost, as is demanded by the current method. This is going to require an awful lot of data and Bhattacharya, Gill and Swartz used just 85 matches[3]. The isotonic regression allows estimates of situations where no data was available to be made, as well as correcting resource estimation where small sample sizes were present, using the data round them, however this lack of data and the subsequent corrections required to fix this could lead to some inconsistencies in the way the resource percentages are calculated. By using the same expression just with differing parameters the D/L model can calculate resources for any combination of balls and wickets and is at least consistent across the board. Bhattacharya, Gill and Swartz are quick to point out that their adjusted resource table is not meant to replace D/L resource table but, just highlight some of the issues the D/L model may posses. Based on the found not to perform that much better than the Standard D/L table does when applied to T20.

Carter and Guthrie (2004) try to create an Iso-probability rule attempting to ensure the way that the target is reset after an interruption is such, that the probability of a team winning a match before an interruption is the same as the probability of them winning after the interruption. They do this by estimating the cumulative distribution function F of the number of runs a team will score when they have n overs remaining and have lost w wickets. In order to calculate F the runs scoring options at each ball are separated, transition probabilities are calculated and then F is recursively put together. The observed frequency is used to estimate the probability that an extra is bowled and a probit model is used to estimate the probability that a team loses a wicket and the probability that various amounts of runs are scored, given the match situation. Combining these with the boundary conditions, that the probability of scoring any runs when all wickets or all overs have been used is 0 and the probability of scoring less than 0 is always 0, F can be calculated.

[3]See graphs in Chapter 5 to observe how averages become unreliable and messy when the sample size that they are averaged over isn't big enough

In the process of creating this model Carter and Guthrie identify some criticisms of D/L that their model addresses. Both D/L and the Iso-probability model work by preserving some quantity before and after an interruption, D/L preserves the ratio of target score to available resources and the Iso-probability model preserves the probability of winning. Carter and Guthries biggest criticism of D/L is the point at which the quantity is conserved, currently D/L preserves the quantity at the start of the match (innings) but Carter and Guthrie argue the quantity should, instead, be preserved at the point where the interruption occurs. Doing this allows the batting team's performance prior to the interruption, something D/L ignores, to be taken into account when resetting the revised target. As an aside Carter and Guthrie suggest a modified D/L (MDL) preserving the runs required in the remainder of the innings to remaining run scoring resources, as an alternative to D/L. So where D/L equates

$$\frac{Target\,required\,prior\,to\,interruption}{Total\,resources\,available\,prior\,to\,interruption} = \frac{Target\,after\,interruption}{Total\,resources\,available\,after\,interruption}$$

MDL equates

$$\frac{Runs\,required\,prior\,to\,interruption}{Resources\,remaining\,prior\,to\,interruption} = \frac{Runs\,required\,after\,interruption}{Resources\,remaining\,after\,interruption}$$

MDL requires a team to score proportionally the same number of runs in their remaining resources before and after the interruption and therefore depends on their score at the time. Allowing a team performance in the first part of the innings to have greater influence on the way in which their target is set. This can still be calculated using the ordinary D/L tables.

What is done is done and therefore can not be altered so the idea of conserving the task that the team still have to do rather than the task over the whole innings seems, to me, to be a sensible course of action. However, doing this would alter one of the D/L model's fundamental principles: If a team are 5 runs ahead before the interruption, they should be 5 runs ahead after the interruption (Lewis, 2015). This is equivalent to saying that teams should perform averagely over any interruption. Conserving the task after the interruption, as MDL does, rather than the whole innings task, would lead to teams becoming further ahead or behind, according to D/L, after an interruption than they were beforehand. This would mean that teams would no longer be assumed to perform averagely over and interruption. This judgement will be analysed in Chapter 4 to see if it has any statistical grounding.

Secondly Carter and Guthrie claim that, unlike D/L, their Iso-probability model is incentive free. The way in which a team bats is a trade-off, teams maximise the number of runs they can score whilst minimising the risk of getting out. Batting more aggressively leads to greater runs scoring but also a greater chance of getting out. At the start of the match the tradeoff faced by the two teams is the same, however if rain is imminent D/L changes this tradeoff. In search of the D/L par score teams can often afford to bat more aggressively as the benefit of the extra runs out ways the risk and penalty of losing a wicket in the short run. It is no secret that these incentives do exist at the highest levels of the game. In international cricket teams will often bat second if they suspect it might rain to take full advantage of these incentives. However, the fact that these incentives exists does not necessarily make D/L unfair. There will be occasions where the batting team has these incentives but there will also be occasions where the reward of taking another wicket outweights the risk of conceding a few more runs and the bowling team receives a similar incentive. As will be discussed in the later chapters the D/L model is only trying to be fair to both sides at the start of the season and thus as these incentives apply to both teams they are not necessarily causing any unfairness.

Carter and Guthrie's Iso-Probaility rule works in a very similar fashion to the D/L model; when a rain break occurs if a team is more than 50% likely to win then they win and visa versa, this is equivalent to saying that if a team is above the D/L par score then they are more likely to win than not win. The difference between the methods comes in how they decide how likely a team is to win. Carter and Guthrie's model calculates an exact probability of winning based on the score the team has at the time, whereas D/L compares the score the team currently have with what the average team is expected to have scored in order to chase this target. Whilst conserving the probability of a team winning is the exact thing every rain rule should attempt to do, my issue with the Iso-probability model is that calculating the probability of a team winning is far more precise than just comparing their current position to the average. In reality the probability of a team winning from any position will depend on much more than just the wickets lost and overs remaining, it will depend on the batsman at the crease, the overhead conditions and many other things. Trying to be so precise, whilst only taking a small subset of the factors affecting the probability of winning into account, could lead to errors in the modelling and the possibility of this model becoming unfair. The D/L model on the other hand just looks at the physical resources the team has at their disposal which, though they may be used differently, represent the same thing in all conditions, and then assumes a team performs averagely over any resources lost. This is taking only the bare minimum, not even the score the team is on at the time, into account and thus is much more likely to be fair universally.

Stern (2009) investigates the D/L model judgement that it is fair to use the same resource table for both innings of a match. Duckworth and Lewis arrived at this judgement by assuming that there exists some optimal run scoring potential at every stage of the game, and that this can be interpreted to produce a remaining resource percentage. Therefore Duckworth and Lewis argue that only first innings data is relevant in producing resources percentages as teams batting second are optimising their chance of winning and therefore not necessarily batting in a way that optimises runs scoring. Thus it is only fair to assume second innings resources are distributed in the same way as first innings resources. Stern however views the D/L resource table as describing the typical pattern of play and therefore sees no reason why, if this patterns is different, that there shouldnt be a different resource table for first and second innings. Stern suggest that targets should be reset in the same way as they currently are for D/L but instead of taking second innings resources, R_2, from the same table as first innings resources ,R_1, a separate second innings resource table should be constructed using the transformation below from the current D/L table.

$$R_2(u, w) = 1 - F[1 - R_{DL}(u, w)]$$

$F(x)$ is the cumulative distribution function of a beta distribution with parameters estimated by applying Tobit regression and random walk boundary crossing analysis. Stern's analysis concluded that D/L in its current form underestimates the importance of very early and very late overs and thus overestimates the resources corresponding to the middle overs of an innings. Stern also discusses the ramification this has when setting a target score for a game that is shortened from the outset. As first innings and second innings resources are no longer the same, if a few overs are lost at the start of a game then, according to Stern, the team batting second has now lost more resources that the team batting first. Therefore team 2's target should be revised down. As the number of overs lost increases, the initial loss of second innings resources is offset by the greater loss of first innings resources in the middle part of the innings and thus the revision of team 2s target should be lessened. Whilst I believe there is a possibility that first and second resources are not distributed in the same way, Sterns idea of non equal targets in equally shortened games, whilst maybe theoretically sound, doesnt work in practice. For example take a 20 over game, which Duckworth and Lewis assumes to be the same as the last 20 overs of a 50 over game. Under Sterns adjusted D/L, when team 1 scores less than the average score, first innings resources would equate to 56.6% whereas second innings resources would equate to 58.3% thus team 2 would require nearly 4%[4] more runs than team 1 to win the game. This is obviously a preposterous way to set up a competition. I believe that the fact that both teams are aware that the game is shortened before it starts nullified this effect but that there is still a possibility that within a shortened game (or a shorter game) first and second innings resources are distributed differently. This is something that will be investigated in this project.

Reliability Data - Ideas and Methodology

Another (cricketing) topic widely discussed in the literature is how to fairly calculate a batsman's average. A batsman's average is an important and widely used statistic in quantifying the ability of that batsman. Currently a batsman's average is calculated by dividing the total number of runs the batsman has scored (in out or not out innings) and dividing it by the number of times the batsman has got out. Das (2011) suggest that the current batting average, as specified above, is not a reasonable representation of a players average performance, especially not when looking at a small set of scores for example over a series or a season. Das notices that Not Out scores are simply examples of right censored data, where its known that had the batsman completed their innings then they would have scored at least as many as they did. Kaplan and Meier (1958) suggested a non-parametric estimator for the survival function $\hat{S}_{km}(t) = P(X > t)$ to be used when a sample contained some data points that were right censored. In the context of Cricket the Kaplan Meier (KM) estimate is defined to be

$$\hat{S}_{km}(i) = \Pi_{j=1}^{i}(1 - f_j/M_j) \text{ for } i = 0, 1, 2,$$

where f_j is the number of times a batsman got out on the score j and M_j is the number of times a batsman finished their innings on j and was either out or not out. Das provides and intuitive description of how the KM estimate works; initially giving each observed score (out or not out) equal weight, then, starting from the smallest observation, individually redistributing the weight of the censored scores to

[4] ((58.8-56.6)/56.6)

the uncensored scores larger than them. Since a batsman score, X, is always a non-negative integer the mean of their discrete distribution $\Sigma_{i=1}^{\infty} iP(X = i)$ is equivalent to $\Sigma_{i=1}^{\infty} P(X > i)$ therefore the Kaplan Meier estimate of the mean is $M_{km} = \Sigma_{i=1}^{K} \hat{S}_{km}(i)$.

Das proves that if scores are distributed according to either an exponential or geometric distribution then the normal batting average is Maximum Likelihood Estimate (MLE) for the mean of the population. However Das suggests neither of these distribution actually make sense in cricketing terms. The exponential distribution is continuous and cricket scores, obviously, must come from a discrete distribution. The geometric distribution has a constant hazard function, suggesting the probability of a batsman getting out for any score is the same. This is unlikely to be the case as batsman are more vulnerable when they have scored very few runs than they will be when they have scored 70 or 80 for example. Das also proves that the KM estimate is the non-parametric MLE if a parametric distribution for the population is not considered. Therefore the KM estimate for the mean constitutes an improvement to the regular batting average.

Whilst endorsing the KM estimate as a 'vast conceptual improvement' (Das, 2011, p.2) Das also outlines some of the pitfalls of using the KM estimate to calculate batting average. Firstly the non-parametric approach used by Kaplan and Meir results in their KM estimate being completely reliant on observed data. The estimate only assigns mass to scores that a batsman has actually got out on and thus if a batsman has never scored 10 and got out the KM estimate will assign a 0 probability of that batsman getting out for 10. This problem becomes greater when only looking at a batsmans scores over a series or season where they will have only got out on a handful of scores. The second problem caused by using the KM estimate arises when a batsman highest not out score (Y) is higher than their highest out score (X), for example ex England captain Michael Athertons highest score was 185 not out (ESPN, 2015), in this case the weight of the not out score(s) greater than X is redistributed above X resulting in $\hat{S}_{km}(i) > 0 \ \forall i$ meaning a positive weight at ∞. This results in an infinite mean which is completely useless. However Das suggests a simple, if rudimentary, way to deal with this by interchanging the highest out score with the highest not out score (X for Y) and continuing as normal. The final issue raised with using the KM estimate here is usually censored data is examined in terms of continuous time; if something is alive at time T we know it will live for a time strictly greater than T. As runs scoring is measured on a discrete set, the natural numbers, a batsman being S runs not out means we know he would have scored at least S. The inequality here is no longer strict as we are working with a discrete time analogue. The actual effect that this could have would only ever be minimal and this is therefore unlikely to cause any noticeable change in the KM estimator.

Kimber and Hansford (1993) agree that the usual batting average is an MLE if a batsman scores are distributed according to a geometric distribution. They also find empirical evidence that many players are more likely to get out on 0 than other scores and that the hazard function decreases a scores increase until it flattens out at a score dependent to the batsman. Again this agree with Das that a Geometric distribution is not suitable. Kimber and Hansford also recognise that batsman's scores are like lifetimes and looking at them over a period of time is like looking at a point process with an event being any time a batsman gets out and interval times being the number of runs the batsman scores in between getting out. They notice that point processes such as this often arise in repairable systems reliability. Repairable systems reliability looks at components in a system that will fail, have to be replaced and then the whole system will start again. The data analysis of this is based around finding trends in the interval times in which the component is working, for example do these increase or decrease with time (Crowder et al., 1981, p.157). Kimber and Hansford use this analogy to model a batsman's scores and see if they can find trends in the way batsmen score runs.

Crowder et al. (1991, pp.157-181) provide intuitive instructions in how to apply the theory of repairable systems to analyse reliability data. They concentrate on modelling repairable systems as non-homogenous Poisson processes (NHPP). NHPPs, like regular Poisson processes, still assume that events occurring independently but now the failure rate is no longer constant. In repairable system reliability the failure rate, which here depends on t, is often referred to as the rate of occurrence of failures (ROCOF) and is defined by $v(t) = \frac{d}{dt}E[N(t)]$ where $N(t)$ is the number of failures of the system in the time interval $(0, t]$. Crowder et al. calculate the likelihood of observing failures at $t_1, t_2, , t_n$ times until the nth failure is seen as:

$$L = \Pi_{i=1}^{n} v(t_i) exp(-\int_0^{t_n} v(t)dt)$$

so once an appropriate form for $v(t)$ has been found its straightforward to calculate MLEs for any of $v(t)$s parameters. Crowder et al. focus on two straightforward, monotonic choices for the ROCOF:

$$v_1(t) = exp(\beta_0 + \beta_1 t)$$

and

$$v_2(t) = \gamma\delta t^{\delta-1}$$

with $\delta > 0$ and $\gamma > 0$

They explain simple graphical methods that can be used to see which $v(t)$, if any fits the data the best. These will be demonstrated in Chapter 6 along with the inference and parameter estimation that occurs once and appropriate $v(t)$ is discovered.

Chapter 3

The data set

Simplifying assumptions

In order to simplify a cricket match and the data relating to it, the following assumption was made. This was done in a way that allowed more intuitive analysis to be conducted without losing too much of the information contained in the data. When Duckworth and Lewis designed their model, one of their requirements was that it could be applied to every feasible situation of a cricket match. This meant the D/L model could assign a resources percentage corresponding to every ball in every over of a match. For the purposes of this project scores will only be looked at on an over by over basis. If runs scoring is viewed as a continuous process throughout the innings this discretises the innings, making it easier to work with and analyse. If a wicket falls in the middle of an over an extra wicket is added to the score at the end of that over. What this does mean is that if an innings ends, either by the rules of the game or by some intervention, just one ball into the over the score will be updated such that it appears the over was concluded. However, the maximum resource change between any 2 overs of a 50 over innings is 3.6% which is believed to be small enough such that ignoring it doesn't make a significance difference to the analysis.

Data Collection and Manipulation

In order to examine how the D/L method is currently performing and to test whether any modifications may improve it, some real world data was required. With cricket being the popular sport that it is, many organisations keep and publish cricketing statistics. Cricinfo (ESPN, 2014) the cricket branch of ESPN's sporting network publishes cricket scores and cricket statistics from matches across the world ranging from the upper echelons of school cricket to the international game. This was where I first turned to collect the data I required. Cricinfo's data is freely available via their website, and it contained all the information that could possibly be required: The date, location and the teams playing in the match followed by the score (runs and wickets) each team had achieved at the end of each over of their innings. Due to the simplifying assumptions made ball by ball data was not required and it was also decided at this early stage to disregard all information about which members of the team were batting. The D/L method, the rain rules that have preceded it and the alternatives discovered in the literature all ignore which batsman are batting as it over complicates the model. Building a model where the D/L target required depended on who was batting could also lead to the ridiculous situation that a team could reduce their target by deliberately losing a wicket so that a batsman with a higher D/L ranking could come in. Cricinfo's data has the advantage that it is viewed by many people across the world and updated regularly, meaning that the data is likely to be as accurate as possible. Cricinfo also contains complete data concerning every county cricket match played for the last 10 seasons and more, so gaining a sufficiently large sample should be easy. Unfortunately, to aid user friendliness, Cricinfo do not display their data as a database and list all the data for each game on separate pages. This made it incredibly time consuming to collect just a season's worth of data and also created more room for human error on my part. Each different statistic required had to be separately copied from different pages onto an excel spreadsheet, creating more opportunity to copy the wrong number or paste it in the wrong place. Once a seasons data was collected some preliminary analysis was conducted and it was quickly established that more than data would be required. Cricinfo ignored a request for their data in a more suitable form so other options were examined.

In order to access more data, without going through the time consuming process of extracting it from Cricinfo, the project's sponsor, were approached to see if they had data I could access. In fact the ECB introduced an initiative in 2012 to document all the data from every game on an online, but not publically accessible, database. I was given access to the database, allowing me to obtain over by over score summaries for every game recorded. These were downloadable as spreadsheets making them much easier to manipulate. Before the data was uploaded, it was cross-referenced by a representative from each side playing in each game. Therefore there is unlikely to be any errors made when recording the data and the only source for human error comes when the scores are inputted. Even if an error has been made its unlikely to be more than 1 or 2 runs which will not have a large impact on the analysis. As all the data from one game was downloaded together, there is minimal chance for any error on my part with the only possible opportunity for error being that I could have missed a game completely, this would not have been done in any deterministic manner and would just result in slightly less data. Only having data from 2012 onwards leads to a maximum of 4 seasons worth of data which is less than what Cricinfo has available. However 4 seasons data is considered enough for the purposes of this project. Further work could be conducted on a data set supplemented with older games extracted from Cricinfo.

Once the spreadsheets had been downloaded it was easy to combine them into one master spreadsheet and using simple excel formula a database of final scores and results and over by over runs scored was created and used throughout the project.

Chapter 4

Exploratory/Preliminary data analysis

Exploratory data analysis is usually concerned with looking at the data itself, getting a general feel for the shape of the data and seeing if any trends or errors exist. This project is looking at a specific model so here the data is used as a tool to explore the model rather than being investigated itself. However the model makes some basic assumptions about the data, and these will also be investigated.

Testing the D/L model as a first innings score predictor

The first stage of analysing the D/L model is to check to see if its performing fairly in the broadest of senses. The way in which the D/L model works is when there is an interruption it takes a teams current position and then assesses it to see if they are more likely to win or not from that position. Therefore if the D/L method is performing fairly it should pronounce teams who were going to win winners and teams who were not going to win losers. But this is nearly impossible to investigate, firstly if a game is uninterrupted we know if a team won or not but that simply makes each score in that inning either a winning or losing score and does not really give any information about how likely a team was to win from there. An interrupted game never actually reached a conclusion so it is impossible to know if the D/L model predicted the right team would win or not. The frequency with which teams won and lost from any position could be analysed, but whether D/L would have regarded these as winning or losing positions (above or below the D/L par score) depends on the score team 1 reached.

A simple way to check whether the D/L method and resource table are performing fairly is to use it as a score predictor for completed games. This involved using the case, outlined by Duckworth and Lewis, when $R_2 > R_1$, randomly selecting a point within team 1's innings, pretending the innings ended there and then seeing what the D/L method would have set team 2 to chase in 20 overs. This could then be compared to what team 2 actually had to chase in the second innings, namely the score team 1 actually scored. If the D/L method and resource table are fair then the predicted scores should lie evenly above and below the actual scores, indicating no bias each way. Here the D/L model's power at predicting scores is not being analysed. An accurate score predictor would need to depend on many more factors than just the teams overs remaining and wickets lost. When the D/L method is used in this way it is supposed to give a fair assessment of what team 1 would have scored in the resources they missed out on, and this is what is being tested here.

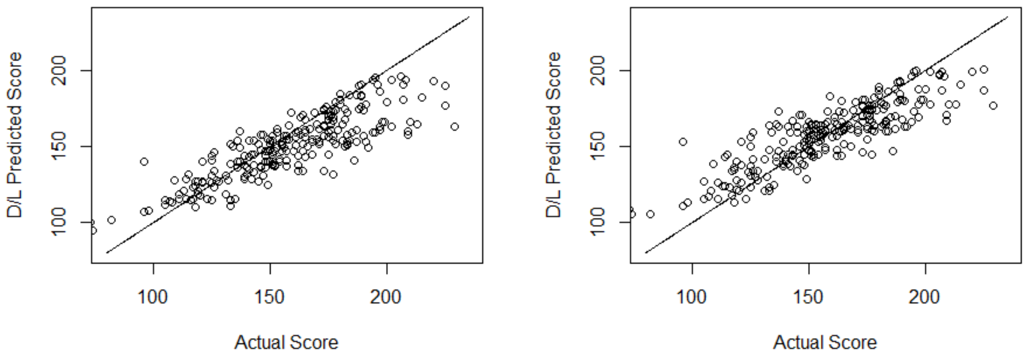
Each game in both the T20 and Pro40 data set was then randomly cut off at the end of an over and the D/L predicted score was then calculated. The random cut-offs were chosen independently by a random number generator in *Microsoft Excel* making sure that the D/L regulations were still adhered to. In T20s no cut-off was made after less than 5 overs and in Pro40s no cut off was made after less than 10 overs, in reality if rain occurs in these scenarios the game is not considered sufficiently long enough for D/L to act fairly so is abandoned. At this stage, only first innings where the team used their full resources, batted all their overs or were bowled out, were used so that a fair comparison could be made between what was predicted and what the team achieved. This method of extrapolating scorers is only ever used during the first innings of a match so applying it to the second innings is not really appropriate. First innings scores often need to be revised upwards due a deficit in the resources they receive. When the resources team 2 receives is lower, the target is adjusted downwards to be proportionally the same as the score team 1 achieved in their resources. Therefore, second innings scores are never adjusted upwards in this manner.

Using the formula, $T = S + G(N)(R_2 - R_1)$, the D/L predicted scores could be calculated. Here S is the score the team had reached at the simulated cut-off, R_1 is the resources team 1 had used at the cut off, R_2 is 100% (as we are seeing what team 2 would have had to chase in their full resources) and $G(N)$ denotes the average runs scored in the first innings by a team batting for N overs. The value G50 is given as $G50 = 245$ and is the only value specified in the 2014 ICC handbook (ICC, 2014). In order to apply the D/L model to Pro40 and T20 matches this value needed to be scaled down, using to D/L resource table, to get values for G40 and G20. The D/L method views a 20 over match, simply, as the last 20 overs of a 50 over game where the batting side has lost no wickets and the same is the case for a 40 over match. Therefore G40 and G20 are calculated by applying the D/L resource percentages for 40 and 20 overs remaining when no wickets had been lost, 89.3% and 56.5% respectively, to $G50$ to yielding $G40 = 219$ and $G20 = 139$. However the observed average score in 40 and 20 overs was 235 and 158 respectively[1]. This are both considerably higher than the D/L table predicts and as this method is so reliant on the parameters G40 and G20, the observed averages are going to be used in tandem with D/L value to ensure a through analysis. This initial discovery rings alarm bells about how D/L will perform for the shorter formats of the game. The parameter G50 is updated yearly so is likely to be accurate but using a table that is thought to correctly model the resource distribution in a shorter game grossly underestimates the average score.

Figure 4.1, below, contains the graphs of actual first innings runs scored vs D/L predicted first innings runs scored for T20 data set. The left graph is when G20=139 and the right is when G20=158. The line y=x is added to show where the predicted scores should lie. If D/L is working fairly an even distribution of points above and below this line should be observed.

[1] all numbers here are rounded to the nearest whole integer for convenience

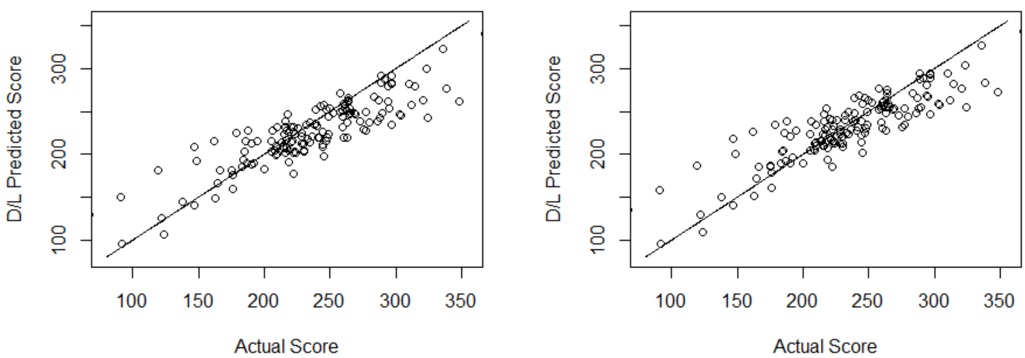
Figure 4.1: T20 Actual vs D/L predicted scores (left: G20=139, right G20=158)



The even distribution of scores above and below the line that would indicate that the D/L model is working fairly is not observed in either of these two graphs. Scores around the set $G20$ value seem to be distributed above and below the line reasonably evenly suggesting that D/L is working fairly if teams score close to what its mean is set at. However in both cases D/L seems to be slightly over predicting the scoring of teams who score less than the average score and under predicting the scoring of teams who score more than the average score, in general. Demonstrated by the predicted scores of scores lower than the average lying above the $y = x$ line and predicted scores of scores higher than the average lying below the $y = x$ line. The points on the graph to the right appear to be slightly more evenly distributed above and below, mainly because the central cluster of points where D/L appears at its fairest is closer to the center of the observed scores. In fact the squared error between predicted and actual scores on the right is 59688 compared to 77140 on the left. It was mentioned before that prediction accuracy is not being tested here however the squared error still allows us to quantify how well D/L is doing.

A similar thing is observed in Figure 4.2, below, for the Pro 40 scores. Again the left hand graph is when G40 = 219 and the right is G40=235 the observed average.

Figure 4.2: Pro40 Actual vs D/L predicted scores (left: G40=219, right G20=235)



A cluster around $y = x$ is again observed around the set value of $G40$ indicating that D/L is appearing fair for scores around its set average. In Pro40s D/L appears to be doing better, with most low scores appearing either side of $y = x$ however there is still evidence of some low scores being largely over predicted by D/L. Similarly to the T20 case there is also evidence that D/L is consistently under predicting scores that are greater than average. The right hand points appear slightly closer to the line than the points on the left hand graph and this better visual fit is somewhat corroborated as the squared error is reduced from 112235 when G40=219 to 98121 when G40=235.

Unfortunately the differing number of games and the differing lengths of these games across formats make it hard to compare the squared error between Pro40s and T20s. However, visually the Pro40 scores seem to fit the desired $y = x$ line slightly better than the T20 scores do. This is likely to be because a Pro40 game of cricket resembles a 50 over game, which D/L was designed for, much more closely than a T20 game.

Comparing the graphs on the left and right have confirmed that adjusting the official average score to the observed average score created a slightly better fit both visually and this was corroborated by the reduction in squared error. This shows that using the D/L tables to adjust the average score for Pro40 and T20 matches is not fair and it would be a fairer to set these independently based on observed data from these matches. In fact, conversation with Stern (2015), who is now responsible to the maintenance of D/L, confirmed that, in the most up to date version of the D/L model, $G40$ and $G20$ are no longer set using the 50 over D/L table and that there current values are similar to the observed values here.

Duckworth and Lewis (2004, pp.8-11) noticed that their model was performing poorly when teams scored above the average score, when analysing it against 50 over international cricket. This initial analysis has demonstrated that D/L also suffers this deficiency when applied to the shorter formats of English County Cricket. This issue was solved by the introduction of Professional D/L which isnt looked at here. However Duckworth and Lewis did not observe that their model was also performing badly when scores were much lower than the average, which the evidence above suggests it might be. This could be caused by the, linear, scaling down of the D/L table for shorter formats. In a 50 over game a team has much more of a chance to recover from a bad start than they do in a shorter format of the game so it makes sense that D/L used in the shorter format may overestimate the score of a team who score below average. This give initial cause for concern that the 50 over D/L table may not be performing fairly when used for Pro40s and T20s.

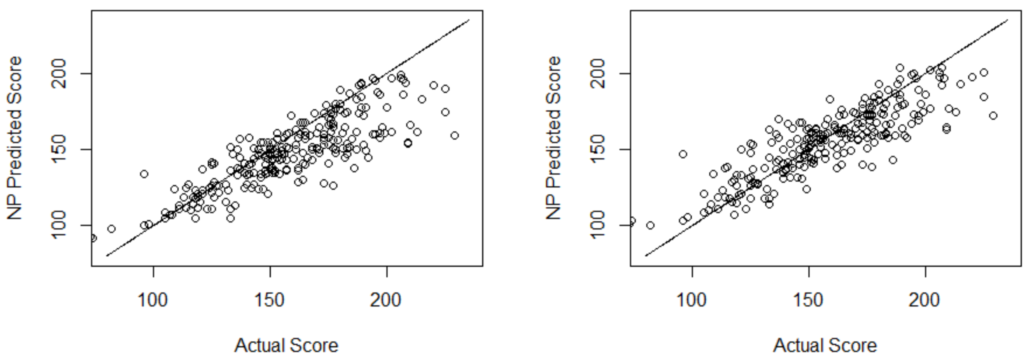
Exploratory testing of alternatives

As well as testing out the fairness of the D/L resource table, this analysis can also be used to test out some of the alternatives and modifications that can be applied to D/L that were mentioned in the literature.

A different resource table for T20 matches

The initial analysis above also appears to suggest that there is some, though only visual, evidence that D/L used for T20 appears less accurate then Pro40. Suggesting the less the format is like the 50 over format the worse D/L does. Bhattacharya, Gill and Swartz (2011) agree with this conclusion and give reasons why they think this might be the case. They produce their own adjusted, non-parametric, version of the D/L resource table to be used for T20 cricket. To test and see if this resource table produces a fairer spread of predicted scores than the D/L resource tale, the analysis was conducted again, using the Bhattacharya, Gill and Swartz's non-parametric (NP) resource table (appendix 1.4). The analysis was conducted using the same games and the same cut off points to ensure any differences were not caused by random fluctuations in the random cut off points. The predicted scores by the NP model and the actual scores are graphed in Figure 4.3 below. The left graph is with G20 =139 and the right with G20 =158.

Figure 4.3: T20 Actual vs NP predicted scores (left: G20=139, right G20=158)



This NP model appears the have solved the issue found that D/L was over predicting scores lower than the average as now, lots of points appear to hug the $y = x$ line with an even distribution above and below. However the left hand graph suffers severely from the under predicting of scores well above the average, that was observed in D/L. This effect is still present in the right hand graphs however it is to a lesser extent . In fact the NP model with the observed average used, as shown on the right, appears to have the fairest distribution of scores seen so far. For the most part, the scores hug the line and there is a reasonably even distribution both above and below the line. There is however, still an issue with consistent under predicting of scores above 160

Despite this observed fairer fit the squared errors of the NP model vs the actual scorers are both greater than the errors observed for D/L. When G20 was set to 139 the NP error was 88227 over 10000 greater than D/L, when G20 was set to 158 it was much closer with the NP squared error of 60489 only around 800 greater than D/L. So when the right G20 value is chosen the fit of the NP model appears to be fairer but its still worse at predicting scores. This, in fact, is warned against by Bhattacharya, Gill and Swartz. They explicitly say that their model is not meant to replace D/L but just demonstrate that improvements could be made. By building a model that appears to have a fairer fit Bhattacharya, Gill and Swartz have shown that D/L is possibly not as fair as it could be and that maybe having a separate table for T20s could help this.

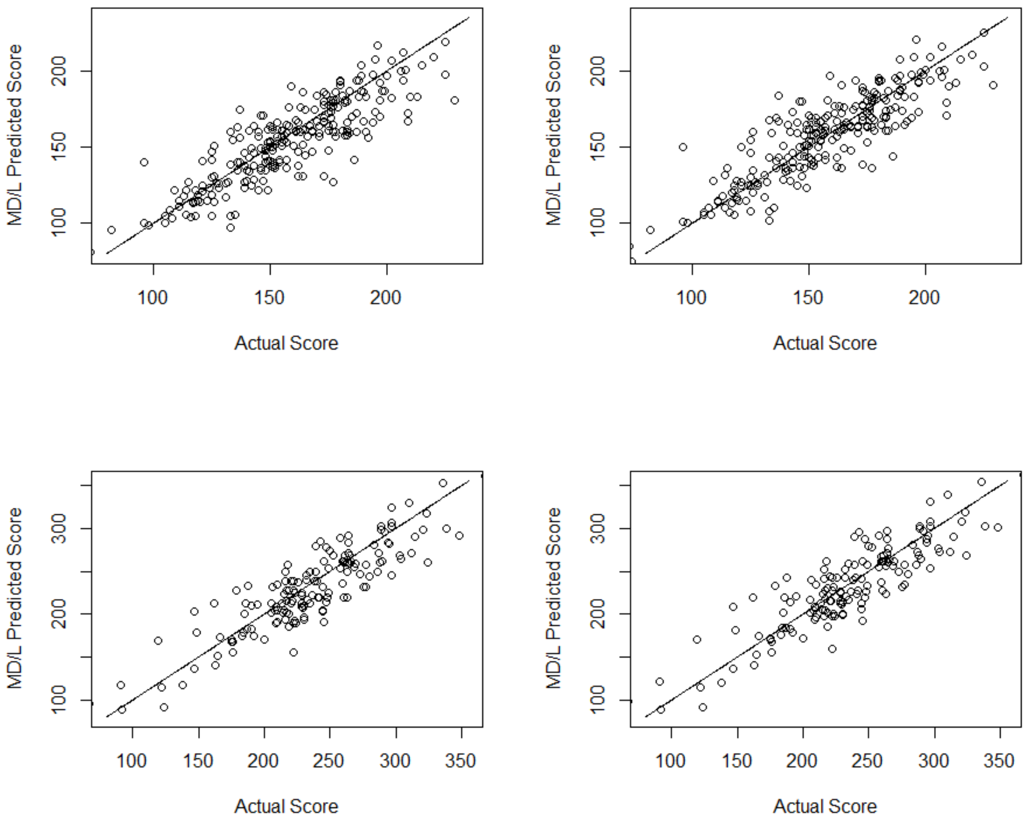
Allowing past performance to have a greater impact

Another simple amendment to D/L, Modified Duckworth Lewis (MDL), suggested by Carter and Guthrie (2004, p.4) can also be investigated in this way to see if it appears to provide a fairer set of predicted scores. Under the D/L model when rain interrupts the second innings of a match the target that team 2 require is reset so that the number of runs they require for their innings is proportionally the same as the runs that the team 1 scored in their resources. The way in which the target is reset takes no account of the score that team has amassed prior to the interruption and this is merely subtracted from the new target to work out how many runs team 2 are still required to score. Carter and Guthrie suggest that instead of preserving the total target to total resources ratio when rain occurs, the remaining runs required to resources left ratio should be preserved. This takes note of the score the team are on when rain occurs and concentrates on ensuring the runs scoring challenge they face after the interruption is the same as it was before the interruption. This allows MDL to compensate teams that have made a good start to their innings further and punish teams that have made a bad start more severely. Carter and Guthrie do not postulate a new resource table and suggest their model can be applied using D/L original resource table.

The implicit assumption within D/L is that when an interruption occurs, a team is expected to perform averagely, for a team with the same resources left, across the overs lost, no matter hows they have performed before the interruption. Therefore a team who are 50-1 after 10 overs and a team who are 10-1 after 10 overs are assumed to score the same number of runs in their next n overs. This acts to maintain the teams position with respect to the D/L par score (see literature review). However, this could possibly be viewed as unfair, if a team has performed above averagely before and interruption assuming they will then perform averagely across an interruption may unfairly disadvantage them. By postulating that prior performance should impact how the target is reset, Carter and Guthrie are no longer assuming teams perform averagely over an interruption, allowing them to move further ahead or behind of the D/L par score.

Whilst Carter and Guthrie demonstrate how to apply their MDL when $R_2 < R_1$ they make no mention of how to apply it whenR_2 outweighs R_1. This case is required to analyse MDL in the same way as the D/L and the NP models have been analysed above. So, keeping Carter and Guthries ideas in mind and still adhering to the format of D/L, a method to be applied in this case needed to be devised. The D/L model deals with this situation by applying the deficit in resources faced by team 1 to the average score ($G50$) and then adding this to team 1s current score i.e. assuming that team 1 would have scored averagely during their lost resources. If team 1 had batted averagely before the interruption then this would make sense but if they had batted either above or below averagely beforehand then this could be seen as unfair. MDL works by giving more weight to prior performance so, instead of applying the total average score to the remaining resources, the average score teams reached from their current position should be used. So instead of assuming teams perform averagely through their lost overs they are assumed to perform averagely for a team who was in their position. Calculating the average score teams score from their position is a job that D/L is currently employed to do in international cricket and although its observed above that it may over predict some low scores and under predict some high scores it is generally doing an alright job. Therefore the original $G50$ is replaced by $G(S)$ where $G(S)$ is the D/L predicted score from S, $G(S) = S + G(N)(R_2 - R_1)$, when $R_2 - R_1$ is the resources deficit faced by team 1. This new method was then applied to the cut off scores in the same way as D/L and NP were, using the same random cut offs to ensure a fair comparison. Figure 4.4 demonstrates this with the predicted scores using the ICC (2013) G50 on the left and the observed values on the right.

Figure 4.4: Above: T20 Actual vs MDL predicted scores (left: G20=139, right G20=158) Below: Pro40 Actual vs MDL predicted scores (left: G40=219, right G20=235)



The first thing to notice here is that there is now very little difference between the left hand and right hand graphs. This means the model is now much less sensitive to one of its inputted parameters, the average score. On both graphs the points seem evenly distributed above and below the line, which is exactly what a fair model should do. There is still some very slight evidence of over predicting low scores and under predicting high scores but this is minimal when compared with the original model. The appearance of increased fairness is backed up by a considerably lower prediction error with the T20 MDL achieving an error 58808 when G20 was 139 and 55936 when G20 was 158 and Pro40 MDL achieving an error of 93946 when G40 was 219 and 91693 when G40 was 235.

Duckworth and Lewis view it as fair to ignore a teams current position when recalculating their target, however MDL produce a fairer set of scores by taking account of a team's position, that is having a method that recalculates based, exactly, on a teams current position. This is similar to the situation faced by insurance companies when looking at the use of life tables. A life table is a table giving the probability that a person of a certain age dies before their next birthday and they are used by actuaries to calculate insurance premiums. The ethical question is how much information about that person is it fair for these life tables to take into account. In this example the value of G is analogous to the life table. D/L is saying every team should have the same $G = G50$ which is like saying the same life table should be used for everyone. MDL is suggesting that it may be fairer to have a specific G for every score which is like saying every different type of person should have a different life table. In the life table case it is unfair to have one single life table as it punishes the healthy people by grouping them with unhealthy people, analogously, punishing teams that score highly early by grouping them with teams that score less highly. Having a more individualised life table also causes problems, to ensure it is fair the model must take into account many more factors surrounding the current situation which could create errors, for example its very hard to estimate a teams average score from their current position. The solution adopted for life tables is to group people who have common factors, for example have a life table for those who smoke and another table for those who are obese and so on. It could be possible to apply this to cricket and have a different value of G for teams grouped in similar positions. It could be as simple as having the current $G50$ value for averagely performing teams, a value lower than G50 for below averagely performing team and a value above $G50$ of above averagely performing teams. However this would require further judgement from the model builders.

Another observation Carter and Guthrie make regarding the implementation of D/L is that when there is the threat of rain there exist incentives to the batting team to bat more aggressively than they usually would. During an innings the batting team face a trade-off between batting aggressively and maximising run scoring and minimising the risk of getting out. When rain is imminent, it is often the case that the reward of batting aggressively for a short period outweighs the risk of losing a wicket and therefore the original trade-off faced by the batting side is upset. The existence of these incentives in County Cricket was confirmed by an expert who said that if rain was imminent his side often send into bat a more aggressive batsman who is more likely to score runs but also more likely to get out than they would usually. A drawback that has been noted with MDL is that by giving more weight to prior performance than D/L, it will only exacerbate this issue. If a team think they are likely to resume play after an interruption they could have even more incentive, than they have already under D/L, to bat more aggressively prior to the interruption to gain a greater benefit from the way MDL recalculates a teams target.

It appears that it may be fairer if a team's prior performance in the innings is used when recalculating their target after an interruption. To see if this is merely a coincidence or if there exists some cause, the data must be examined. By increasing the impact a teams prior performance has on the way in which their target is calculated, MDL is dispensing with Duckworth and Lewis judgement that it is fair to assume that a team performs averagely over an interruption. To see if there is any evidence of this, the data will be analysed to see if any correlations are observed between past and future performance. D/L assumes that these correlations are zero provided two teams have used the same resources at the point at which the comparison is made. i.e. assuming two teams who have used the same resources before an interruption will score the same during it. Obviously it is impossible to observe the runs a team score during an interruption. But, in order to see if any correlation between past and future performance exist, runs scored before and after the $5^{th}, 10^{th}$ and 15^{th}, overs in a T20 and the $10^{th}, 20^{th}$ and 30^{th} overs in a Pro40 will be analysed. Here only innings where 100% of the resources were used for reasons stated in the earlier analysis. Again only first innings were used due to the consequences chasing a target is expected to have on past and future performance. For example it may appear that a team has performed badly after a point when they had perfromed well before it simply because their good performance prior to a point had got them close enough to a target that quick scoring was no longer required. Firstly all of the games were grouped together and the correlations are presented in Table 4.1.

Table 4.1: Correlations between past and present performance in T20s and Pro40s, not sorting by wickets

T20s	**Pearsons Correlation coefficient**
Between the first 5 and last 15 overs	0.321
Between the first 10 and last 10 overs	0.449
Between the first 15 and last 5 overs	0.459
Pro40s	**Pearsons Correlation coefficient**
Between the first 10 and last 30 overs	0.314
Between the first 20 and last 20 overs	0.348
Between the first 30 and last 10 overs	0.174

The density plots supplied in Appendix 2.1 demonstrate that the distribution of all the scores before and after the various cut-offs are a reasonable approximation to a normal distributed (bell shaped and reasonably symmetric, though there is some evidence of skewness) and therefore it appropriate to use Rs correlation test to test the null hypothesis that the pearsosn correlation coefficient is 0 (The R Stats Package, 2014) . The test was conducted resulting in all 6 of these correlations being significantly different from 0 at the 10% significance level. This comes as no surprise as they are all well above 0.15. For the T20 matches the correlations increase as the time after the cut off decrease. This is likely to be because as the time after the cut off decrease the time to make a recovery is less so teams are more dependent on being in a good position beforehand. There is no general trend across the correlations of the Pro40 matches. Whilst all these correlation are proven to be significantly different from 0, none of them are over a half so whilst positive correlation has been observed it does not appear to be strong.

However, this does not, disagree with what D/L have assumed. Here the games have not been separated by how many wickets they have lost and therefore the correlations are not tested whilst keeping resources constant. These correlations could just be caused by teams who have lost less wickets scoring more runs before and after the cut off and team who have lost more wickets scoring less before and after, something that D/L accounts for as a teams remaining resources will be lower if they have lost more wickets. Therefore to investigate this further the games were sorted by wickets lost and the correlations were reanalysed. Splitting the matches by how many wickets were lost at the cut off point lead to some small samples occurring for some wickets. Therefore to ensure that the results seen were as reliable as possible, any correlation calculated over a sample of less than 10 was discounted. The correlations for T20 matches and Pro40 matches are presented in Table 4.2, below, with the sample sizes over which they were calculated in the brackets.

Table 4.2: Correlations between past and present performance in T20s and Pro40s, sorting by the number of wickets the team had lost at the cutoff point

	Correlation coefficients						
T20s	W=0	W=1	W=2	W=3	W=4	W=5	W=6
Between the first 5 and last 15 overs	0.067 (54)	**0.243 (105)**	0.176 (62)				
Between the first 10 and last 10 overs		0.190 (49)	**0.313 (77)**	**0.429 (60)**	0.037 (37)		
Between the first 15 and last 5 overs			**0.315 (39)**	**0.334 (56)**	**0.323 (64)**	0.152 (44)	**0.423 (18)**
Pro40s	W=0	W=1	W=2	W=3	W=4	W=5	W=6
Between the first 10 and last 30 overs	0.238 (32)	0.198 (56)	0.152 (39)	0.160 (14)			
Between the first 20 and last 20 overs		-0.141 (17)	-0.008 (45)	0.258 (39)	0.085 (26)		
Between the first 30 and last 10 overs			-0.256 (16)	**0.426 (27)**	0.015 (43)	0.097 (27)	**0.458 (16)**

Here there is a much greater spread of correlations present in both the T20 and Pro40 data, with most of the values between 0.15 and 0.35 but also many values close to 0 and even a few negative values. Once again the density plots in Appendix 2.2 show that it is reasonable to assume all the scores before and after the cut offs are approximately normally distributed so Rs correlation tests can be used. Any correlation above that is in bold is statistically different from 0 at the 10% significance level. For the Pro40 data there are only 2 situations where it appears that the correlation between past and future runs are significantly different to 0. There is slightly more evidence that there may be correlation between past and future performance in the T20 data with 7 out of the 12 scenarios having correlations significantly different from 0. This is much more ambiguous than the Pro40 case suggesting that prior and future performance are independent in some cases and not in others.

The results from the Pro40 data indicate that the D/L model judgement that it is fair to assume teams perform averagely regardless of how they have performed before seems correct. There does not appear significant correlation between past and future performance and therefore, it would be unfair to assume a team will perform anything other than averagely. This shows that the relationship that was expected to have caused MDL , used in Pro40s, to produces a fairer set of predicted scores than D/L wasnt observed in the data. This then suggests that MDL either appears fairer by chance or that there is something else in the data causing it to do so. The results for the T20 data do not lead to such a straightforward conclusion. It appears that at some point (combination of wickets lost and overs remaining) during a batting innings where the D/L model judgement that teams should be assumed to perform averagely over any lost overs appears to be fair, as past and future runs appear to have no correlation. Whilst, in other cases there is evidence that there may very well be a positive correlation between past and future performance, indicating that average future performance may not be a fair judgement to make. However it would be inconsistent as well as massively over complicating the model to let prior performance influence the way in which the batting target is reset in some scenarios and not in others. This coupled with the fact that even though some of the T20 correlations are significantly greater than 0, they are all still less than 0.5 suggesting that any correlations present certainly are not strong. Therefore there is no real evidence here to conclude that the D/L model judgement that past and future performance are independent, resulting in the D/L model ignoring past performance when recalculating a teams target in a interrupted match, is causing the model to be unfair.

In this section the Standard D/L model has been analysed to see if it makes a fair first innings score predictor. This revealed Standard D/L regularly under predicts large scores, something noticed by Duckworth and Lewis themselves and resolved with Professional D/L, as well as over predicting small scores. An alternative resource table, challenging the D/L model judgement that the resources distribution throughout an innings is the same no matter how long the innings is, was trialed, with some success, to see if a fairer distribution of scores could be achieved. Finally an alternative method for applying the current D/L resource table, but allowing prior performance greater weight in the resetting of targets, was trailed. This, MDL, appeared to give a fairer distribution of scores, but when the data was analysed there appeared very little evidence against the D/L model judgement that the scores should be reset independently of the teams current position. Next to investigate these judgements further the D/L method will be examined further, comparing actual resource usage to what the D/L model predicts.

Chapter 5

Further Exploratory analysis

In their follow up paper, Duckworth and Lewis (2004, pp.5-7) conduct an analysis of average runs scored for overs available and wickets lost using observed data from international cricket. This analysis was then compared to the average runs that the Duckworth Lewis (D/L) method would expect a team to score for the same overs available and wickets lost. This allowed Duckworth and Lewis to see how well their model was fitting the real life data and gave them an idea whether any of the model parameters made by expert judgment may need updating. Duckworth and Lewis did this for ODI cricket but it is the intention here to conduct the same analysis for both Pro40 and T20 domestic cricket using data available.

Methodology

As explain in section 3 the data set, was easily manipulated into a form where for each game it listed, for each over, the runs scored in the remainder of the innings and the number of wickets lost. From this data set, the average runs scored in the remainder of an innings for a given set of resources was calculated. The only way to do this fairly, just using the raw data, is to only use inning's where the team used 100% of their resources (lost all 10 wickets or batted all of their overs) when calculating the average. Innings' where the teams did not use 100% of their resources could be misleading as it may show teams not scoring many runs in their remaining overs simply because they had nearly won the game. As there existed enough first innings where 100% of the batting resources had been used in the sample a simple solution to this for first innings data is to omit any first innings where 100% of the resources were not used. This left a remaining 242 T20 first innings and 149 Pro40 first innings.

Stern (2004) asserts that first innings resources and second innings resources may be distributed in different ways. Further suggesting that early and late second innings overs constitute a higher resource percentage than the corresponding first innings overs do. An expert, who was asked to comment on this, agreed that batting in the second innings, with a target in mind, is certainly different to batting in the first innings suggesting that maybe resources are distributed in a slightly different way. However the expert did also mention that while batting in the second innings is different, bowling in the second innings is also different and that there is a possibility that these two cancel each other out. In light of this and unlike Duckworth and Lewis (2004, p.5-7), it was also decided to study second innings data, as well as data from first innings, in an attempt to see if their resources distribution was, in fact, different. There is an unfortunate caveat to this; it is quite often the case that when a team is chasing a modest total they may win with overs to spare and thus will not have used up all of their resources. When this happens it may appear that a team has scored less runs in their remaining overs when in actual fact they have won the game so have been unable to use all of their resources. In fact this happened in 59/140 Pro40 games and 78/243 T20 games. Similar to what has been done to the first innings data, one solution to this could be to only examine games where the team batting second used 100% of their overs. However this causes two further problems, firstly it radically reduces the sample size of second innings data from 140 to 81 Pro40 games and 244 to 165 T20 games. Secondly if a team batting second has used up 100% of their resources they have either chased the score in the last over or lost because they have been bowled out or batted their overs without reaching team 1's target. Of the 100% resources sample, only 48/165 T20 teams won batting second and for the Pro40 sample, only 16/81 teams won batting second. By only using innings where approximately 3/4 of the teams lose, a bias that the team batting second won't use their resources as well as the team batting first is being inflicted.

When analysing the batting average statistic Das (2011, p.6) notices that a batsman's not out scores are simply examples of randomly right censored data, where its known that had the batsman batted till they got out they would have scored at least as many as they did when they were not out (refer to literature review). Similarities are made here between a batsman's innings being not out and a whole team not using 100% of their resources, which from now on will be referred to as a not out team innings (innings where the team did use 100% resources are now out team innings). Das uses the Kaplan Meier (KM) estimate of the survival function to take account of the batsman not out scores when calculating their batting average and it is the intention here to use the KM estimate to adjust for not out team innings, when calculating the second innings average runs scored in the various combinations of remaining resources. For an explanation of how the Kaplan Meier estimate works and how it is calculated see the literature review.

In order to create the graphs below, a second innings average runs scored in remaining resources was required for every resource combination i.e for all wickets lost and for every over remaining. For each combination of wickets lost and overs remaining, the data set consisted of as list of teams who achieved that resource position and the number of runs they scored in their remaining resources. This data contained runs scored by teams who had used all their resources (uncensored) and runs scored by teams who had not used all of their resources (censored). The Kaplan Meier estimate of the average runs teams score in their remaining resources could then be calculated in the same way as Das calculated an alternative batting average.

Similarly to Das analysis of batting average there are a couple of problems associated with applying the Kaplan Meier estimate in this setting, as discussed in the literary review. These are acknowledged here, however they are though not to influence the averages significantly. There should be enough data present to ensure that the majority of scores were reached at some point for most resource combinations, so the solely parametric nature of the KM estimate shouldn't cause many problems. There are, however, some resource combinations, especially in the Pro40 data set, where the sample of teams who had achieved this combination were low and therefore these averages could be unreliable. It was also observed in the sample that for some combinations of overs remaining and wickets lost, the most runs scored in remaining overs came from not out team innings and thus the KM average is infinite. Obviously this is not the case in reality. When this occurred, the simple solution Das (2011, p.14) suggested was adopted. The largest runs scored in remaining overs from not out team innings was replaced with the largest runs scored in remaining overs from out team innings. This may give rise to some error in the averages for these games however, the difference between the two swapped scores was almost always only a few runs and thus the impact of this is thought to be negligible. Finally the error caused by the fact that runs scoring is measured over a discrete set rather than the continuous time, is again considered not to have a significant effect on the average.

The observed data set is presented in Appendix 3

From the average runs scored data it was possible to create 10 curves (one for each wickets, $W = 0$ up to $W = 9$) detailing the average runs scored (y axis) for variable overs remaining (x axis). When they conducted this analysis in 2004 Duckworth and Lewis applied $G50$, the average runs scored in a full 50 over first innings, to the D/L resource table in order to get an idea of the average runs the D/L method would predict a team to score in their, given, remaining resources. Subsequently, this allowed a comparisons with the observed runs scored to be made. This gave another ten curves (one for each wicket). In order for this to be done with the 40 and 20 over data the G50 value as published by the ICC (2013) needed to be scaled down to create $G40$ and $G20$. As was mentioned in Chapter 4 scaling this $G50$ down using the D/L resource table gave $G40 = 219$ and $G20 = 139$. From the raw data, it is obvious that the observed average first innings score in 40 overs was 235 and in 20 overs was 158. These are both considerably larger than D/L expects. Therefore multiplying by $G40$ and $G20$ is likely to underestimate the number of runs D/L would actually expect teams to score, given the fact that teams score higher than average. In order to overcome this the observed runs scored in remaining resources data was divided by the observed average runs scored in full resources, the observed $G40$ and $G20$ values for the first innings and the KM estimate of the average for the second innings. Doing this produces the percentage of average runs from 100% resources that the team scored in their remaining resources. This has converted observed runs scored back to observed resources remaining and allowed a direct comparison with the values in the D/L resource table. Allowing a better analysis of how the D/L resource table was actually performing, within the 40 and 20 overs.

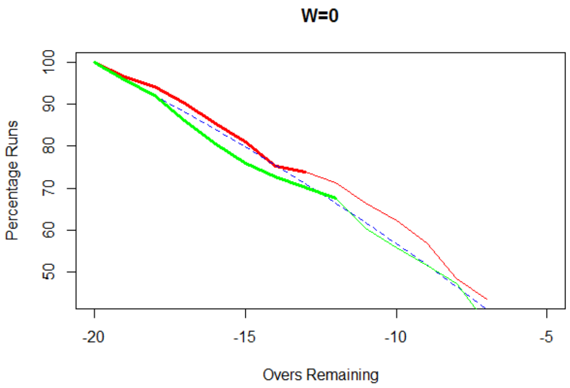
The Graphs

The graphs of the T20 data are in Figure 5.1, below. The red solid line indicates the first innings average resources remaining, the green solid line indicates the Kaplan Meier estimate of the second innings average resources remaining and the blue dotted line indicates the resources remaining as specified by the D/L resource table. Separating these graphs by wickets means each value of x (overs remaining) on each graphs corresponds to a different batting resource percentage from Duckworth and Lewis table, allowing for the comparisons between the two to be made easily.

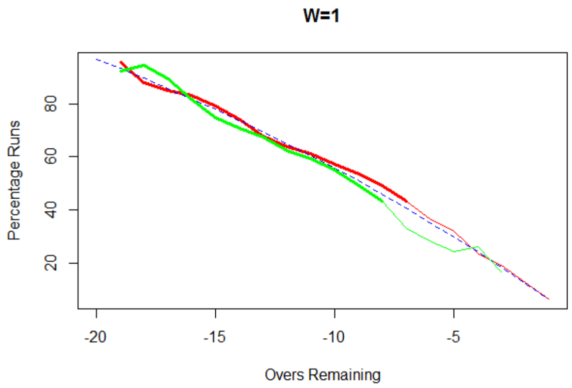
Here every observed data point is plotted, no matter how big the sample size was. Where a sample size of more than 20 is observed then the line is in bold. Samples of size 20 were deemed to be large enough to produce a reliable averages. Looking at the curves when the line is not bold is interesting, but conclusions will not be drawn based on this information as the average could possibly be influenced largely by 1 or 2 large or small scores.

Figure 5.1: T20 Graphs: Comparing observed first and second innings resources with the D/L predicted resources for different wickets

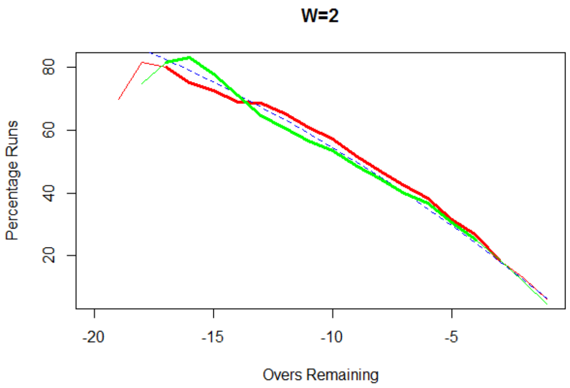
a)



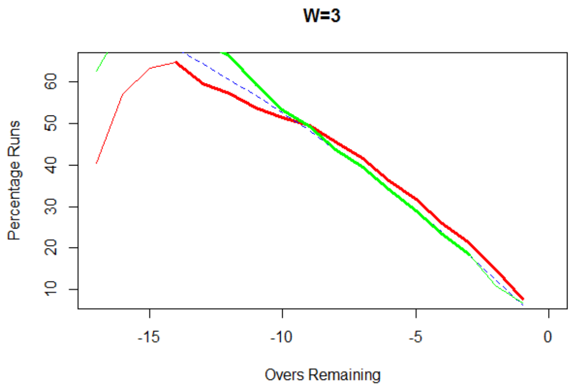
b)



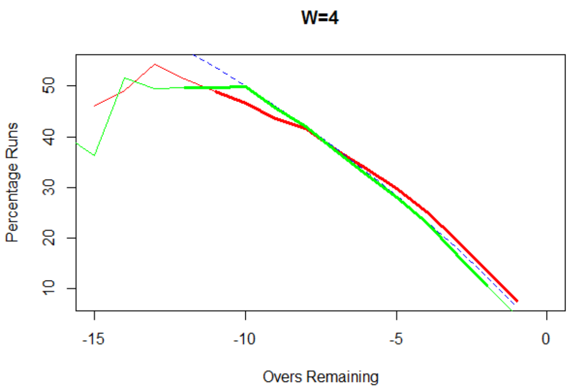
c)



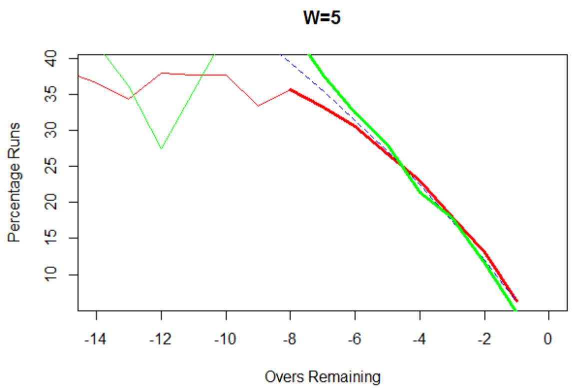
d)



e)

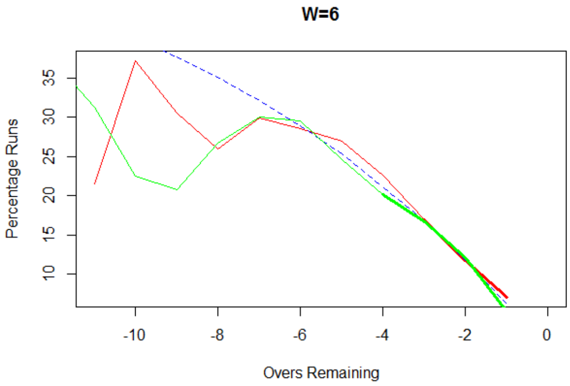


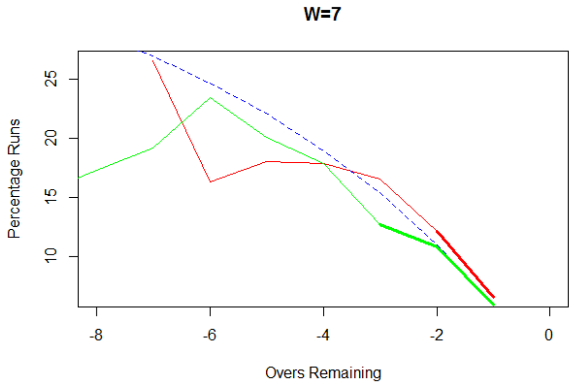
f)



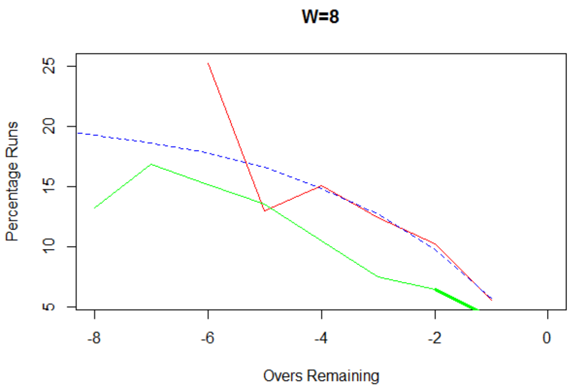
g)

h)

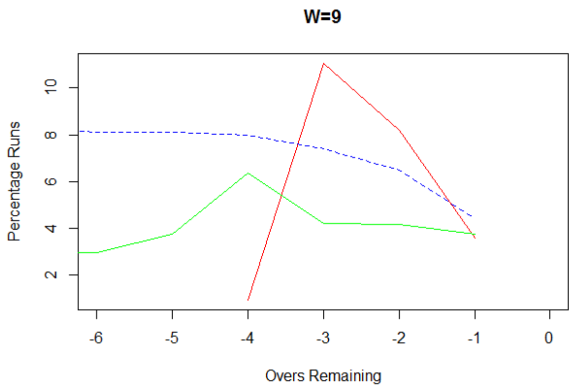




i)



j)



T20 Analysis

Of the sample used to create these graphs 120/242 games were won by the team 1 and 118/242 were won by the team 2 (the remaining 4 games were tied). Therefore half of the games were won by the team batting first and half were won by the team batting second, indicating no winning or losing bias present for either side in this data. Generally, where the sample size was sufficient, the shape of these graphs appear to be similar to the shape of the D/L resources curve suggesting that an exponential decay of the form $1 - exp(-x)$ appears reasonable.

Graph a) shows how resources decay with overs when W=0. Between overs 1 and 5 (19 and 15 overs remaining) the observed first innings resources remaining is higher than D/L predicts. This suggests that for W=0 the overs remaining after these constitute a greater resource to the batting team than D/L currently estimates. This could be explained by the fact that the first 6 overs of the innings are whats known as powerplay overs, where fielding restrictions encouraging aggressive batting are in place, these are not currently accounted for in D/L. In fact the average run rate across all 20 overs is 7.82 runs per over but the average run rate between overs 2 and 6 is marginally higher at 8.05 runs per over. With 15 overs remaining observed first innings resources appear to be what D/L expects however after then observed resources again appear to increase above what D/L expects. Observed second innings resources initially appear to be the same as is estimated by the D/L method however between 18 and 14 overs remaining the observed resources remaining drop below what D/L expects indicting that second innings resources corresponding to these overs are possibly less than D/L is predicting.

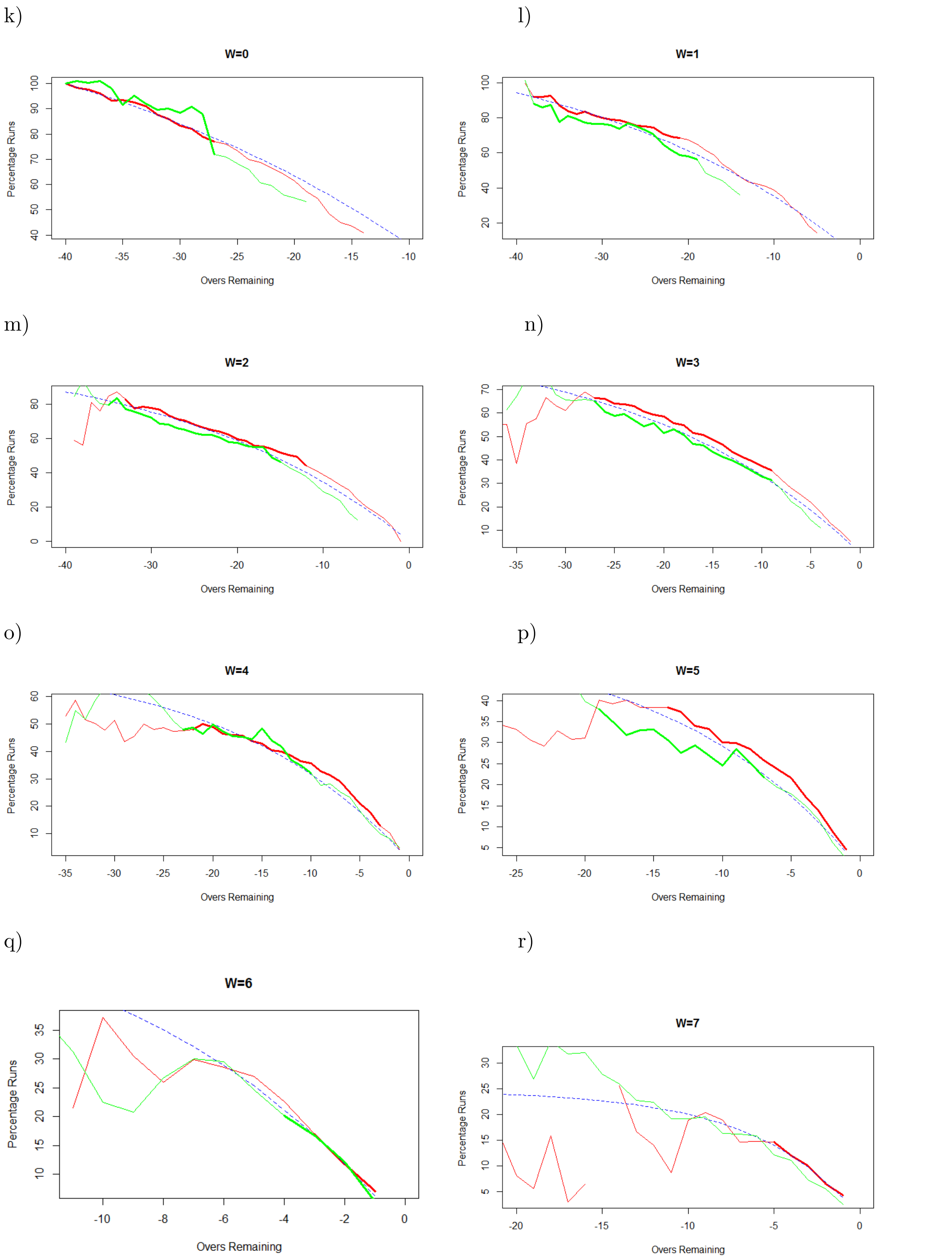
Graphs b), c), d), e) and f) corresponding to W=1,2,3,4 and 5 all show a very similar trend. Initially, where the sample size first becomes reasonable for that wicket, the observed second innings resources remaining appear higher than the value from the D/L resource table and conversely, the observed first innings resources remaining appear lower. As this happens for 5 wickets in total, this is unlikely to have happened by chance and suggests that the resources remaining after the first part of the second innings are higher than the resources remaining after the early part of the first innings, with the D/L resources lying in the middle. After the early part of the innings observed first and second innings resources converge on the D/L percentage and follow it from then onwards. This suggests that for these wickets lost and these overs the D/L method has correctly captured the way in which resources decay as overs remaining decrease and that this decay is the same in both the first and second innings. It does, however, appear that during these overs the observed first innings resources consistently lie just above the second innings resource. The fact that half of the teams in the sample won batting second means that the fact that this is not a result of some losing bias in the sample. This, coupled with the fact the observed second innings resources remaining were higher at the start of the innings, indicates that early second innings constitute greater resources than early first innings resources and to compensate for this first innings resources are then slightly greater for the remainder of the innings.

Graph g) and h) corresponds to the case where W=6 and 7, though there isnt a huge amount of reliable data here, for W=6 it appears that both observed first and second innings resources remaining are the same and that these are consistent with what the D/L table predicts. When W=7 the observed second innings resources remaining appear to agree with what the D/L model predicts however the observed first innings resources remaining appear parallel but higher. This indicates that the decay of resources as overs decrease is the same for both first and second innings and that D/L has modelled this correctly, but the fact that the first inning resources lie above what D/L predicts, suggesting the the D/L parameter associated with losing the 7th wicket is too large. However this is only observed over 2 or 3 overs of reliable data so more data is required to see if this is consistent across a whole T20 innings.

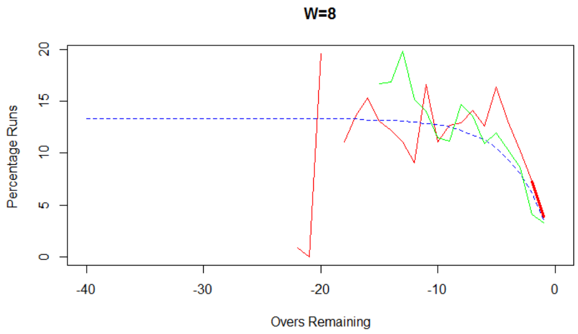
Unfortunately not enough data was present to make any meaningful conclusion in the cases where W=8 and 9. In the sample less than a third (71/242) of T20 first innings lose more than 7 wickets so again more data would be required to conduct a reasonable analysis for these scenarios.

The graphs in Figure 5.2, below, are the corresponding graphs for Pro40 games, again depicting the observed average first innings resources remaining, the KM estimate of the observed average second innings resources remaining and the resources remaining predicted by the D/L resource table.

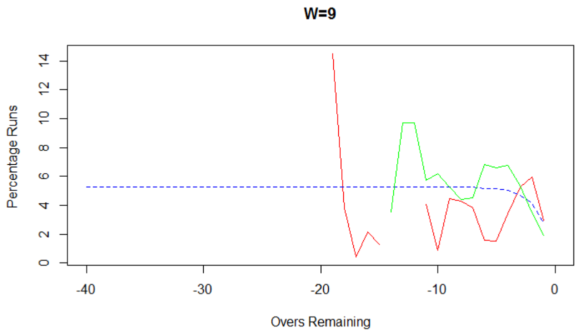
Figure 5.2: Pro40 Graphs: Comparing observed first and second innings resources with the D/L predicted resources for different wickets



s)



t)



Pro40 Analysis

In the Pro40 sample 81/149 teams won batting first and 68/149 teams won batting, this corresponds to less than 46% of teams winning batting second, this doesnt sound like much but this that nearly 9% more games were won by the team batting first. This indicates that there is a slight winning bias in the data meaning that team 1 is often likely to have a slightly higher percentage of resources remaining than team 2. Its also noticed here that the observed second innings resources remaining lines appear much less smooth than the first innings curves. This is likely to be a result of using the KM average to account for the censored second innings and the problems that can be caused using smaller samples. The Pro40 sample size is considerably smaller than the sample size used for the T20 graphs and this explains why this is more evident here. Once again the general shape of these curves is similar to the shape of the predicted D/L resource curve, raising no issues with the general exponential decay form that the D/L model uses to model resources.

Graph k) depicts the observed and predicted decay of resources as overs remaining decrease for the case when W=0. The observed first innings resources remaining appear to be very similar to what the D/L resource table predicts indicating that it is doing a good job of modeling the resource decay here. However the same cannot be said for the observed second innings resources remaining. When between 39 and 36 overs are remaining the observed second innings resources remaining appear higher than what the D/L method is predicting and they appear higher again when between 34 and 28 overs are remaining. This behavior is quite unique to the second innings case when 0 wickets have been lost and therefore no real conclusions can be drawn from this, however it could be investigated further if more data was available.

Graphs l), m) and n) depict the case when W=1,2 and 3 and these three graphs exhibit similar patterns. The observed first innings and second innings resources remaining as well as the D/L predicted resources remaining all appear to have a similar shape, indicating in these cases that D/L has correctly captured the way resources decay as the innings progresses. However observed first innings resources are always above observed second innings resources with D/L resources lying in between. This is unlikely to be the case in reality. Unlike the T20 data, its not the case that second innings resources are higher and then become lower, they are lower throughout the innings and if this was the case in reality it would mean that the second innings is contains less resources than the first innings. Though this is postulated by Stern (2009), it's felt here that this is more likely to either be caused by the slight losing bias present in the sample towards teams batting second or could be caused by the KM average not performing properly. The fact that the first innings line is above D/L could be to do with the winning bias in the data or its possible that the D/L parameter decreasing resources for the loss of the 1st 2nd and 3rd wicket could also be too large.

The case where W=4 is shown in graph o). Here both the observed first and second innings resources remaining agree with what the D/L model predicts during the early stages of the sample. However, as the innings moves on the D/L predicted resources decay at a faster rate than the observed first innings resources indicating that at these stages the batting team possibly have more resources remaining than D/L suggests. Unfortunately not enough data was present to observe anything meaningful for the second innings however from the data that is present, it appears that the observed second innings resources remaining continued to be very similar to the remaining resources that D/L predicts.

Graphs p) and q) show similar trends for when W=5 and 6. On both graphs the observed first innings resources remaining appear consistently higher than what the D/L model is predicting but they do seem to have similar shapes suggesting that the D/L has correctly model the decay in resources. Observed second innings resources remaining appear to sit just below what the D/L model is expecting and the observed first innings resources remaining. As was the case when W=1,2 or 3, this was likely caused by the slight losing bias in the sample. The fact that the first innings resources appear higher could be to do with the first innings winning bias, but could also demonstrate that the D/L parameter reducing a teams resources when they have lost their 5th and 6th wicket appears to be too large indicating too harsher penalty on run potential run scoring.

Very little data was available for the case where W=7 however graph r) does show that the observed first innings data, that was present in a sufficiently large enough sample, appears to be very similar to the remaining resources predicted by D/L. Unfortunately no real conclusions can be drawn from the graphs with W=8 and W=9 and more data would be required to investigate these cases.

Conclusion T20

Its obvious that the population that was used to create these graphs is not representative of every cricket match played around the world. However, it is in fact close the whole population of limited overs games played in England for a couple of years and thus its felt that, using it, some reasonable conclusions can be made . For the most part, where the sample size is adequate, observed first innings resource remaining, observed second innings resources and the D/L predicted remaining resources form curves that are very close together and have a similar shape. The fact that the shape of the curves are similar indicates that the D/L model had correctly captured the way in which the run scoring potential of the batting side decays as overs are used up and also that this decay is constant across both innings. The fact that the observed first innings resources often lie consistently above the D/L resources and rarely below suggests maybe the penalty imposed by D/L for losing that wicket is too harsh. These are the same parameters that are used for the 50 over model, it is not surprising that the 50 over model would over value the loss of a wicket in a 20 over context as with the shorter the time to bat the less valuable the wickets are. When W=1, 2, 3, 4 and 5, and there are still many overs remaining (comparatively for the wicket lost, i.e. when 5 wickets are lost 8 overs is considered a lot to have left) it appears as though teams batting second possess more remaining resources than the D/L model predicts and that teams batting first actually possess less remaining resources than D/L predicts. As the innings progresses both first and second innings resources reach the stage where they are now both very close to what D/L predicts. All of this seems to indicate that when wickets have been lost, the early overs for a team batting second constitute a greater volume of resource than they do for a team batting first and in fact, the D/L model is under predicting how important that section of the innings is to the team batting second and over predicting how important it is to the team batting first. During the remainder of the innings observed second innings resources appear to lie below observed first innings resources and this can be seen as compensating for this effect earlier in the innings, suggesting that the remainder of the inning constitutes slightly greater resources for the team batting first, than it does to the team batting second. This deviation from D/L does resonate with cricketing logic; when a team batting second loses wickets early they are suddenly under increased pressure to score quickly and keep up with the required run rate, thus it is important that, if they lose wickets early, that they also get off to a good start so they get up with the required rate. First innings resources being lower make slightly less sense however, when a team bat first there is not the pressure of chasing a target and therefore they do not need to have got off to quite such a good start.

There are other instances where the observed data has deviated from what D/L has predicted however, these are all isolated instances and therefore it is difficult to make any conclusions based on a sample of 242 matches. Often the data for other scenarios, similar to these, is not available and thus it is not possible to observe a consistent trend. If more data were obtained this analysis could be run again and these instances would be investigated further.

40 overs conclusion

Similarly to the T20 data, it appears that, for the most part, the shape of the observed first innings resources is very similar to the shape of the remaining resources that are predicted by D/L. Again this indicates that D/L has modelled the way in which runs scoring resources decay as overs remaining decrease correctly for the first innings. This decay also appears to be the same for the majority of the second innings with the second innings remaining resource line appearing to have a similar shape to the D/L and first inning lines. It is also the case that, often, remaining first innings resources lie slightly above the remaining resources predicted by D/L, suggesting that the parameter associated with losing the respective wicket may be slightly too large, again this could be expected as wickets in 40 over cricket are likely to be slightly less valuable than they would be in the longer 50 over cricket. This could however partially be caused by the slight bias in the data towards teams batting first. Close to 55% of teams batting first won and for this to have happened they must have, on average, scored more runs in their resources than team 2 and thus their observed resources remaining will appear higher. This is believed to be the reason that the observed second innings remaining resources appear to consistently lie below the remaining resources predicted by the D/L model.

There are no other consistent trends indicating any deviations from the way in which the D/L modeled resource decay. Interestingly the discovery from the T20 data that early, for that wicket, second innings resources appear to constitute a higher resource percentage than D/L predicts and that early first innings overs constitute a lower resource percentage is not seen here when the sample size is sufficient[1]. This maybe unsurprising, as the duration of a Pro40 innings is much more closer to a 50 over innings than a T20 is. The fact that Pro40 resources do not exhibit this deviation from the D/L resource table is also consistent with the hypothesised reason for the deviation given above. A Pro40 is obviously longer than a T20 and thus there is more time for the batting team to recover from a poor start, thus their early overs are comparatively less important than in a T20. Due to the length of the innings, the required run rate for a chasing team is likely to rise less steeply than it would in a T20 so the batting side are also under less pressure if they do get off to a bad start.

Conclusion

All in all its believed that D/L does a reasonable job of capturing the way in which batting resources decay as overs remaining decrease. There is visual evidence that for the most part the shape of both innings are similar to what D/L predicts and also similar to each other. Especially for the Pro40 data the shape appears similar throughout indicating no real need for a different resource table for different innings. However there was evidence that first innings resources remaining were higher than that predicted by the D/L table and also higher than second innings resources remaining. It was suspected that this was caused by the slight bias towards teams batting first in the Pro40 sample. The analysis should therefore be run again using a sample that does not contain this bias. It is also suspected that this bias may not completely explain first innings resources remaining being greater than what D/L predicts. This suggests that the parameters in Duckworth and Lewis model that govern the penalty on runs scoring for losing those wickets should be updated to reflect the fact that in a shortened format of the game losing a wicket has less impact on a teams ability to score runs than it does in a longer format.

Whilst for the majority of T20 innings the first and second innings resources remaining appear to decay at a similar rate with this rate consistent with what D/L is predicting, this is not the case at the start of the innings. When a few wickets have been lost, second innings resources appear to start higher and decay faster than first innings resources which start lower and decay slower. When these resources meet, a different point depending on how many wickets the team has lost, the resources then appear to decay in the way the D/L model has predicted. At this stage, and to compensate for them being lower earlier in the innings, first innings resources are then consistently greater than second innings resources. This agrees with Stern (2004), who built an alternative second innings resource table giving

[1]There are some cases of second innings resources being higher than first innings when the sample size is smaller. These are not considered here to be informative as the averages are thought to be unreliable, but with more data it would be interesting to see if this is still the case.

greater resources to the start of a second innings and fewer resources in the middle. However Stern also observed that late second innings overs also constitute a greater resource percentage, which is not observed here. Stern also observed this trend on 50 over data where as here it is observed in 20 over data but not in 40 over data.

This section has produced some interesting evidence both agreeing with the judgements made by D/L in some cases and disagreeing with their judgements in others. However, the best looking at these graphs can produce is a qualitative assessment of how the model is doing. In order to quantify this and to use inference to draw statistically sound conclusions, these curves need to be modelled.

Chapter 6

Parameters and Likelihood estimation

While graphical comparison and visual judgements are useful it would be more rigorous to reach some statistical conclusions based on the graphs in the previous chapter. If likelihoods for these curves can be established then likelihood based methods such as hypothesis testing could be used to make more formal conclusions on what these graphs show. Constructing likelihoods may also shed some light on whether an exponential decay model correctly models the way in which resources are distributed throughout an innings and whether the parameters Duckworth and Lewis judged themselves, are accurate. This chapter concerns itself with trying to model the remaining resources curves constructed in the previous chapter, dynamically, as a repairable system.

Model setup

Kimber and Hansford (1993, p.3) observe that looking at a batsman scores in order is like looking at a point process; with events being when the batsman gets out and time between events as the runs that the batsman scores in between. It was noticed, whilst completing this project, that it is not just individual batsmans scores that can be analysed in such a fashion. In fact a whole team innings could be viewed as a point process with, for example, events as wickets falling and the time between events being the runs scored in between wickets. Kimber and Hansford also notice that this sort of situation often arise in repairable systems reliability, a subject which Crowder et al (1991, pp.157-181) write about extensively (see literature review). Reliability data analysis usually concerns itself with trying to quantifying the length of performance of components important in the running of larger machine, though the reliability of the machine can also be investigated. Repairable systems reliability is one branch of Reliability data analysis looking at the case where when the component fails, it is instantly replaced or repaired and the system restarts again. It is the plan here to view a batting innings as a repairable system and use simple methods suggested by Crowder et al. to try and generate likelihoods for the curves in the previous chapter.

Unlike some techniques, modelling a batting innings as a repairable system doesn't require the fixing of a model form, like least squares regression does for example. This therefore, allows the D/L judgment in the way in which resources decay to be examined as well, as opposed to just looking at the model parameters. The repairable system framework also models a batting innings dynamically, which is appropriate as, obviously, a batting innings evolves dynamically.

As it was mentioned earlier the most natural way to examine a batting innings in the repairable systems framework is to view the loss of each wicket as a failure and the runs scored in between wickets falling as the time in between failures. However D/L used a two factor relationship using wickets and overs to generate resources. Therefore the best way to try and model the way resources are distributed would be to model a failure as anything that reduced a teams resources and then look at how the runs scored in between these failures are distributed. Due to the simplifying assumptions made, an innings is only being looked at on an over by over basis, so resources will only decrease at the end of an over, and will decrease further if a wicket is lost during that over. Now as the graphs that are being modelled are sorted by wickets, the number of wickets lost remains constant throughout the graph. Therefore for the purposes of this analysis a failure will be taken as the end of an over. This is much less intuitive than the loss of a wicket, which will be discussed later, but for the purposes of modelling the graphs in Chapter 5 it is necessary. In addition to this the resources remaining used in Chapter 5 will be transformed back into runs scored. There is a one to one relationship between runs scored and resources remaining, when wickets are kept constant, so this shouldnt affect the analysis, however using runs not resources allows the length, in terms of runs, of the innings to be increased and should make any trends in the data more evident.

The data from Chapter 5 will be modelled as follows (Crowder et al., 1991, pp.157-159). For each wicket:

Let T_i $i \in (1, 2, \ldots, M)$ be the scores at the end of each over where M is fixed at the length of the innings that is being modelled. The runs scored during each over are $X_i = T_i - T_{i-1}$.

Let $N(r)$ = the number of overs taken to score r runs.

The goal here is to examine any trend that may exist in the number of runs a team score between overs, as the innings progresses. This is done, in a repairable systems framework, by examining the rate at which failures occur, defined as the ROCOF= $v(r) = \frac{d}{dt}E[N(r)]$, and seeing how it behaves as a batting innings progresses.

The first issue that arises when attempting to model the predicted D/L and observed runs scored curves as repairable systems is that they are in terms of runs scored in remaining overs and these are hence decreasing. Therefore, runs being the analogue to time, this is like trying to model a repairable system in inverse time! To resolve this the average runs obtainable data had to be transformed so that it now reflected average runs scored. This was done by taking the highest runs obtainable figure, usually the value corresponding to the maximum resources remaining for which a sample existed, and subtracting the runs scored from the remaining resources less than this maximum. This gave a proxy for the way runs were scored in an average innings. Although this is no longer how the D/L model is intended to work, but there is believed to be no real reason why it should not do a good job. If teams average around 158 runs in 20 overs with no wickets lost and average around 153 with 19 over left and still have lost no wickets, then it is probably fair to say that teams average 5 runs in their first over.

However this raised additional issue of its own; due to the fact that the analysis is only being conducted on observed data there are some resources percentages that no team actually reach (for example no team starts the game having lost 3 wickets!). This problem was especially evident when a few wickets had been lost, for example the first observed incidence of a team having lost 4 wickets in the first innings of a T20 didnt happen until the 5^{th} over. In this case, when transforming the data, the 6^{th} over was considered to be the first and the transformation continued from there. This has no impact on the analysis but may be important when interpreting the results. Throughout this analysis only data points which were reached in 20 or more matches were used, this was to ensure that the averages used were a fair representation of what was happening and that none of the values were being influenced considerably by very large or very small scores. Finally, even using sample sizes of 20 or more ,the highest runs obtainable figure, for a given wicket, was occasionally not the runs obtained from the greatest observed overs. This doesnt make any sense in D/L terms because, keeping wickets constant, you should be able to obtain less runs as your overs decrease. Data points such as these are a consequence of using raw data and for the purposes of this analysis ignored, as including them leads to negative runs scored which is, of course, impossible.

Duckworth and Lewis (1998) modeled resources as a 2 factor exponential decay giving resources

$$Z(u, w) = Z_0(w)(1 - exp(-b(w)u))$$

So keeping wickets constant, fixing $Z_0(w)$ and $b(w)$, the number of runs a team is expected to be able to score decays like $A(1 - exp(-Bx))$, where x is the number of overs remaining. Differentiating this, with respect to x, gives $ABexp(-Bx)$, which as overs remaining decrease, increases. Therefore as the innings moves forward resources are decaying at a faster rate and thus the runs D/L is expecting a team to score in between overs is increasing. Putting this in terms of a repairable system as the innings progresses, overs remaining decreases and runs scored will increase, the rate at which runs are scored should increase and thus on average more runs should be scored every over. Therefore, the failure rate, $v(r)$, should decrease with runs, r.

As wickets are lost the parameters $Z_0(w), (A)$, and $b(w), (B)$, will change. $Z_0(w)$ is the asymptotic average runs from the last $10 - w$ wickets, which will obviously decrease as w increases. $b(w)$ is the exponential decay constant. The exponential decay is expected to decrease as wickets are lost. This is because although resources are already lower, due to Z_0, when wickets have been lost the batsman at the crease are expected to score less runs so runs scoring potential actually decays at a slower rate. As the exponential decay is of the form $1 - exp(-x)$, slower decay is reflected by an increase in the parameter $b(w)$, making the exponential smaller and therefore the multiplier of $Z_0(w)$ closer to 1. This can be seen by the in Figure 2.1 in the literary review. As more wickets are lost the curves starts at lower resources but also decays slower.

These two parameter act in opposite ways and as the D/L parameters are confidential its difficult to know what the overall effect if. However with worse batsman at the crease I expect teams to score less runs per over so overs are expected to arrive at a greater rate so its expected that $v(t)$ should increase as more wickets are lost. As it was established in the previous chapter that D/L does a reasonable job of modeling runs scored in remaining resources for both the first and second innings, a priori it is expected that, for each wicket, the ROCOF will decrease as runs increase and that they will increase with the number of wickets that are lost

The ROCOF

Crowder et al. (1991, pp.164-166) model a repairable system as a Non-homogeneous Poisson process (NHPP) which can be thought of as a standard Poisson process but whose rate is allowed to vary. Instead of the rate being a fixed number, it is now a function of time. Therefore take the ROCOF, which has already been allowed to vary with runs, as the non-constant rate of the Poisson process. This models the number of overs taken to move the score of a batting innings from r_1 to r_2 as a Poisson random variable with mean $\int_{r_1}^{r_2} v(t)dt$.

As mentioned in the literary review, once and appropriate form of $v(t)$ has been decided upon , Crowder et al. demonstrate how to estimate the parameters of that form and provide likelihood based tests to conduct inference.

Crowder et al. (1991, p.166) suggest two monotonic choices for the ROCOF in a repairable system:

$$v_1(r) = exp(\beta_0 + \beta_1 r)$$

and

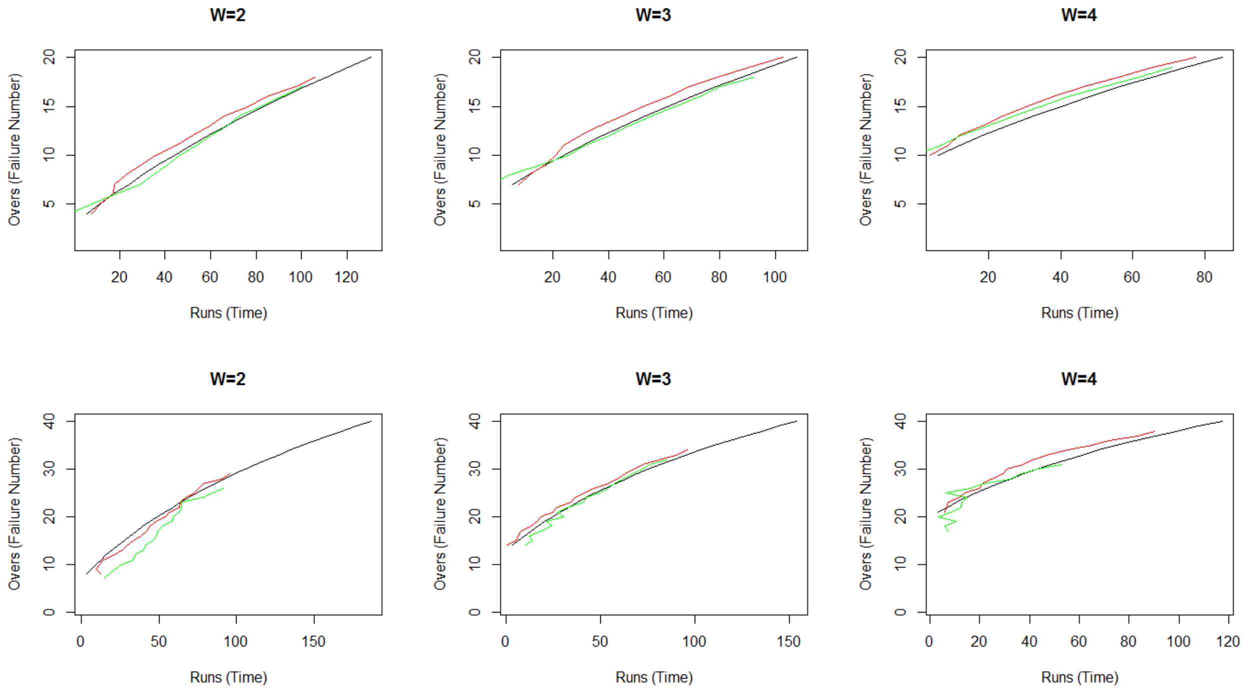
$$v_2(r) = \gamma \delta r^{(\delta-1)} \text{ with } \delta > 0, \gamma > 0.$$

They also advise on graphical ways to choose which of these is most suitable to a given data set. These, ROCOFs, are emphasised as being particularly simple forms and its acknowledged that much more sophisticated ROCOFs do exists. The monotonicity of these does, however, fit with what is expected a priori.

As a consequence of only using data points that were averaged over at least 20 matches, some wickets had little to no overs with reliable data point. When W=8 or 9 there were actually no overs where the runs scored in remaining resources was averaged over 20 or more matches, so the analysis for these wickets is omitted. When W=6 and 7 there are not very many overs (failures) for which a reasonable average was calculated. Not having many overs creates a sample size problem of its own, meaning the trends in the ROCOF are only calculated over a short period of time with a small number of failures. This could lead to possible trends being masked or even misinterpreted by the model.

In order to check that the ROCOF is not, in fact, constant, plots of Runs vs Overs were produced (Crowder et al, 1991, p.159). Some of which are in Figure 6.1 below, the rest show more of the same and are in Appendix 4.1. Here the black line corresponds to D/L predicted runs scored, the red line is observed first innings runs scored and the green line is observed second innings runs scored. The top plots are for the T20 data, with the Pro40 data below.

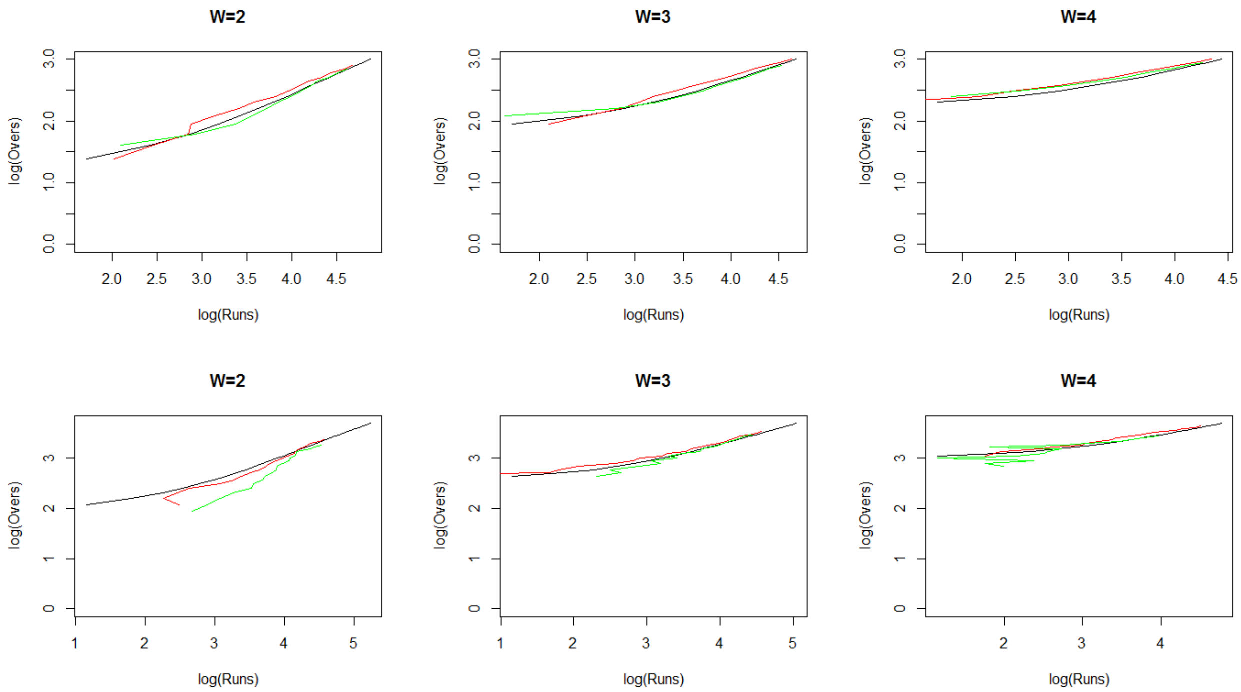
Figure 6.1: Example graphs of Runs (Time) against Overs (Number of Failure). Above: T20 data Below: Pro40 data



If the ROCOF was constant the rate of failures is the same for all runs and therefore the plots of Runs vs Overs would have a constant gradient i.e. be linear. Departures from linearity in these graphs would suggest that the ROCOF is not constant. The lines above appear slightly concave, and therefore non-linear. Hence the ROCOF does not appear to be constant agreeing with the prior expectation. The concavity, meaning that the gradient is decreasing, is a sign that our prior expectation that the ROCOF was decreasing is also correct. The graphs of the Pro40 data also appear more concave than the graphs of the T20 data, this could indicate that the ROCOF is decreasing faster in Pro40s than T20s but is more likely to just be a result of the fact that there are obviously more failures, overs, in Pro40 so the trend appears more pronounced. Interestingly here, the curves associated with the observed data appear higher than the D/L predicted curves, indicating more overs are required to score those runs than D/L would suggest. If this was the case it would lead to teams scoring less in their full overs than D/L predicts, something that is shown not to be the case from Chapters 4 and 5. This is a result of adjusting runs scored in remaining overs to runs scored, something that will be discussed further later in this Chapter. It also appears that the curves depicting the observed data appear more concave than the D/L curves do, this would suggest that the ROCOF of the observed data is decreasing faster than the ROCOF associated with the D/L predicted data.

Now that it appears the ROCOF is non linear $v_1(r)$ and $v_2(r)$ can be checked to see if they are suitable. Crowder et al. (1991, p.175) notice that if $v_2(r)$ is the correct form for the ROCOF then $E[N(r)] = \gamma r^{\delta}$, and therefore, as the number of overs taken to get to r is the best estimate of $E[N(r)]$, taking the logarithm of both sides gives $log(t_r) = log(\gamma) + \delta log(r)$ where t_r is the number of overs taken to score r. So if $v_2(r)$ is the appropriate form for this data, a graph of log(runs) vs log(overs) should be linear. Figure 6.2, below, show log(runs) vs log(overs) for some wickets, again with D/L in black, the first innings in red and the second innings in green and T20 data on top with the Pro40 data below. The complete set of graphs is in Appendix 4.2.

Figure 6.2: Example graphs of log(runs) against log(Overs) to check whether $v_2(r)$ is suitable. Above: T20 data. Below: Pro40 data



These graphs all appear convex, suggesting that $log(t_r)$ vs $log(r)$ is more like a polynomial relationship than a linear one. This means that $v_2(r)$ is not suitable for modeling this system.

Unfortunately Crowder et al. can find no such simple way to graphically test if $v_1(r)$ is suitable. ROCOF = $v_1(r)$ gives the expected number of overs taken to score r runs as

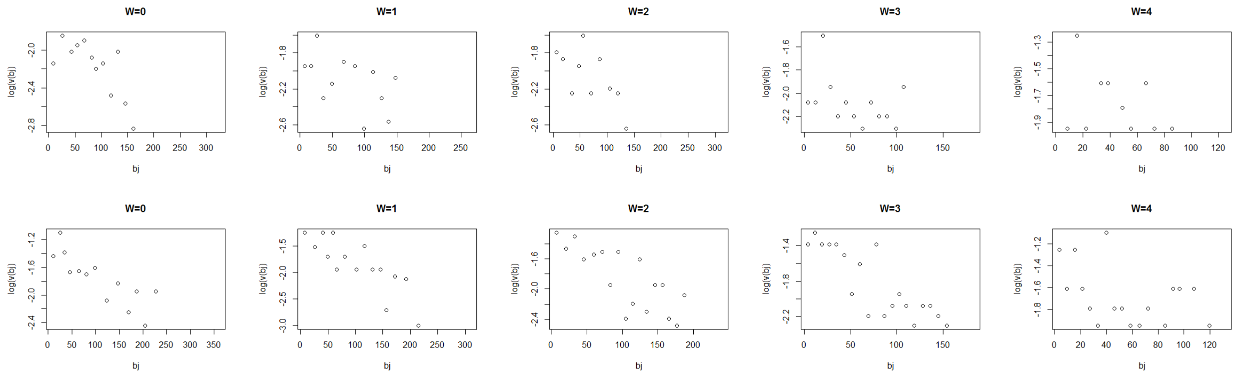
$$E[N(r)] = \frac{e^{(\beta_0)}}{\beta_1} exp(\beta_1 t - 1)$$

this does not give a nice log linear relationship, like $v_2(r)$ did, so other graphical methods must be explored. Now, plots based on the idea of splitting the observed period into intervals are used (Crowder et al., 1991, p.175). The observed period (an innings) from $(0, t_n]$, where t_n is the score when all overs have been used, is divided into k intervals $(0, a_1], (a_1, a_2], , (a_{(k-1)}, t_n]$. The ROCOF is the rate at which failures occur, so an estimate of this in the middle of one of these intervals is the number of failures in the interval, divided by the width of the interval (i.e the number of overs used to score the runs in the interval, divided by the number of runs in the interval). Therefore an approximation of $v_1(r)$ is

$$\tilde{v}(\frac{1}{2}(a_(j-1)+a_j)) = (N(a_j) - N(a_(j-1)))/(a_j - a_{j-1})$$

where $a_0 = 0$ and $a_k = t_n$. Therefore plotting $\tilde{v}(b_j)$ vs b_j gives a good idea of the shape of the ROCOF, with $b_j = \frac{1}{2}(a_{j-1} + a_j)$. Crowder et al. advise that the choices of k and the a_js is up to the user so here the ajs will be calculated by a random simulator in *Microsoft Excel* and then k will be selected accordingly. By the definition of v_1, $\tilde{v}(b_j) \approx exp(\beta_0 + \beta_1 b_j)$ and taking natural logs yields $log(\tilde{v}(b_j)) \approx \beta_0 + \beta_1 b_j$. Therefore if v_1 is the correct form of the ROCOF then a plot of b_j against $log(\tilde{v}(b_j))$ should be approximately linear. The plots using the D/L predicted runs scored data of $log(v(b_j))$ vs b_j for both T20 (above) and Pro40 (below) are presented in Figure 6.3, below. The plots for the first and second innings look much the same and are in Appendix 4.3.

Figure 6.3: Example graphs of b_j vs $log(\tilde{v}(b_j))$ to check whether $v_1(r)$ is suitable. Above: T20 data Below: Pro40 data



These are only a selection of plots but they are reasonably representative of the behavior seen throughout. Some, usually when W is low, of these plots look like they are linear while some of them do not, the plots appear to get less and less linear as the number of wickets increase. As this method is just approximations, using random selection of the interval widths and only plotting a handful of points, its hard to tell anything for sure. It could be that the linearity observed above has happened by chance or that the non-linearity is observed by chance. What is clear from these graphs is that the failure rate $v(r)$ is certainly decreasing as more runs are scored which is what was expected. Despite the evidence that some of these graphs do not appear linear, $v_1(r)$ is going to be assumed correct and its parameters will be estimated. This is done for two reasons, firstly there is evidence that at least for some wickets the relationship is linear, which is stronger evidence than is present for the use of $v_2(r)$. But also, that D/L uses an exponential decay and therefore it may be appropriate to try and exponential here, this does however make me guilty of trying to fit the data to a model rather than visa versa but this is recognised and has been taken note of.

Given the slight evidence above it is now assumed that the ROCOF takes the form $v(r) = exp(\beta_0 + \beta_1 r)$. Given this assumed form and the assumption that overs occur as a NHPP, Crowder et al. (1991, pp.167-169) describe how to construct likelihoods and use these to estimate the parameters β_0 and β_1 as maximum likelihood estimators. The likelihood function is differentiated with respect to β_0 and β_1

then each derivative is set to 0. This gives $\hat{\beta_1}$ such that:

$$\sum_{i=1}^{n} t_i + n\beta_1^{-1} - nt_0(1 - exp(-\beta_1 t_0))^{-1} = 0$$

and using this

$$\hat{\beta_0} = log(\frac{n\hat{\beta_1}}{exp(\hat{\beta_1} t_0) - 1})$$

The observed data was inputted into these functions and, using *Microsoft Excel's* goal and seek function to obtain $\hat{\beta_1}$, the estimates for both parameters were obtained. These are presented in the Table 6.1 below, first for the T20 data and then for the Pro40 data

Table 6.1: Parameters estimates for both T20 and Pro40: D/L, first innings and second innings runs scored.

T20	**D/L**		**First innings**		**Second innings**	
	beta1	beta0	beta1	beta0	beta1	beta0
0	-0.00152	-1.949	0.0197	-2.224	0.0113	-2.302
1	-0.00175	-1.922	0.00291	-2.109	0.00302	-2.313
2	-0.00195	-1.913	-0.00462	-1.721	0.00506	-2.308
3	-0.00168	-1.949	-0.00564	-1.716	0.00323	-2.277
4	-0.00054	-2.020	-0.00845	-1.643	0.00215	-2.140
5	0.00331	-2.156	-0.00633	-1.780	0.0116	-2.590
6	**0.0691**	-3.233	**0.0616**	-3.134	0.0304	-2.584
7	**0.181**	-4.120	**0.165**	-4.241	**0.0358**	-2.268

Pro40	**D/L**		**First innings**		**Second innings**	
	beta1	beta0	beta1	beta0	beta1	beta0
0	**-0.00398**	-1.339	-0.0005	-1.379	-0.0112	-1.471
1	-0.00448	-1.329	**0.0241**	-2.103	-0.0117	-1.021
2	-0.00533	-1.276	-0.0004	-1.452	0.00529	-1.772
3	-0.00607	-1.309	-0.00941	-1.103	-0.00098	-1.457
4	-0.00672	-1.401	-0.0126	-1.098	**-0.0336**	-0.500
5	-0.00676	-1.528	-0.0102	-1.435	**0.268**	-4.748
6	-0.0013	-1.845	-0.0120	-1.564		
7	0.0132	-2.108	0.0177	-2.241		

Here the majority of β_1s are less than 0 indicating that the ROCOFs for these wickets are decreasing. As was expected a priori, this indicates that as the innings progresses the number of runs teams score in between overs increases. There are, however, cases where the value of β_1 is positive. A lot of these come after wickets 6 and 7 have been lost and these are thought to be caused by the small amount of overs for which the sample size was large enough to analyse. There are also some positive values of β_1, when fewer wickets have been lost, occurring in the first innings and throughout the second innings. What these would mean in reality is that the rate at which overs end increases as runs increase so team score less runs in later overs, which looking at the data is not the case. For the first innings T20 data the average runs rate for the first 5 overs is 7.5 runs an over, between overs 5 and 10 it is 7.2 runs an over, between over 10 and 15 its 7.7 and between over 15 and 20 its 8.9. This run rate does initially drop but the general trend across the innings is that the run rate increase as the innings goes on. Therefore these positive values either mean that the method used has not worked, or this is just an issue caused by there not being enough data in the observed sample. Not having enough data means a reliable average for enough overs cannot be calculated for the expected trend to emerge. Alternatively, for second innings, this could be a result of using the KM data.

The D/L data comes from a well-defined model and therefore the expected runs scored produced from this model do not suffer from any of the messiness of working with real world data. For this very reason its unsurprising that the D/L data has produced a set of parameters with a consistent trend running throughout, where the observed data hasn't. Here the values of β_1 are all negative, if only very slightly, and decease, or increase in modulus, as wickets increase. This means that the ROCOF is decreasing with runs and decreasing at a faster rate as wickets are lost. As was explained earlier a decreasing ROCOF was expected but the fact that the ROCOF decreases faster as more wickets fall was not expected. What this means is that overs occur less often for a team who have lost more wickets than for a team who have lost less wickets. More simply, teams who have lost more wickets score at a higher rate than teams who have lost less do. This is completely counter intuitive to cricketing logic, the way a team order its batsman is so that the best batsman bat at the start of the innings and therefore teams who are less wickets down should be able to score quicker.

However, this is probably caused as a result of transforming the data from average runs scored in remaining resources to average runs scored. To perform this transformation all of the scores were subtracted from the largest runs scored in remaining resources for that wicket, as explained above, and this was taken as the starting point for the rest of the innings. However this starting point was almost never when all 20 or 40 overs were remaining as there was often not observed data there, so the starting point for different wickets is now different. For example in the T20 first innings data the 3^{rd} over is the starting point when 2 wickets have been lost and the 9^{th} over is the starting point when 4 wickets have been lost. In the repairable systems framework these are both taken as the same point during the innings, i.e. the first failure, and this is where the error is caused. These parameters are comparing different stages of the innings for different wickets and this is why it appears as though teams score more runs between overs as wickets are lost, they are actually scoring more runs between overs as the innings is progressing.

Whilst some of the parameters above behave the way that was expected, almost all of them are very close to 0. Therefore to check to see if these parameters are actually showing anything, a hypothesis test was conducted. Crowder et al. (1991, p.169) recommend a Laplaces test to test the hypothesis that $\beta_1 \neq 0$. Under the null hypothesis that $\beta_1 = 0$ the test statistic:

$$U = \frac{\sum_{i=1}^{n-1} t_i - \frac{1}{2}(n-1)t_0}{t_0(n-1/12)^{1/2}}$$

is approximately standard normally distributed. Those values above in bold are the values which, according to Laplaces test are significantly different from 0 at the 5% level. Almost all of the estimates of β_1 are not significantly different from 0 indicating that the ROCOF is in fact constant with runs. What this means is that, keeping wickets constant, there is not significant evidence here that teams do not just score at a constant run rate throughout their innings. This is certainly not what is expected by cricketing logic or what is modelled by D/L, therefore this experiment is branded inconclusive.

Though modelling a cricket innings as a repairable system does not appear to have given the results that were expected its noted that this was a very rudimentary first attempt to do something like this. Kimber and Hansford (1993) had success modelling batsman's scores in this way but it seems modelling a whole batting innings like this is more difficult. This could stem from a number of reasons. Firstly, the form, $v_1(r)$, of the ROCOF that was chosen could be causing issues. When the form of $v(r)$ was being investigated only very weak evidence was found in favour of the use of $v_1(r)$ and it was only taken forward as the evidence for it was stronger than the evidence for the only alternative presented. Crowder et al. (1991, pp.166-167) admit that this is only a very simple form and that there exits many more complicated forms throughout the literature.

Secondly, the transformations that had to be made to the D/L table and the observed data to make it suitable to be applied in the reliability data setting could be affecting the analysis. It is said above that the average runs scored obtained from the average runs scored in remaining resources is thought to be a pretty good proxy, but the reference point that the rest of the remaining runs scored were subtracted from could be causing some error.

An issue was also created by only using overs at which a sample size of more than 20 had been observed. This drastically reduced the number of overs that could be considered failures, for each wicket, and made any trends that may have been observable in the ROCOF less evident.

Finally the way in which the repairable system was applied to data could be causing issues. In an attempt to model the curves shown in Chapter 5, the end of an over was taken to be a failure and the loss of a wicket, a more intuitive failure and the type of failure Kimber and Hansford (1993) used, was held constant throughout. This meant that it was the rate at which overs ended that was being analysed and, by the way in which the NHPP was set up, this was allowed to vary with runs. This however is not what the D/L method is saying. The D/L method is expecting the rate at which teams score runs, to increase as overs remaining decrease and therefore analogously the rate at which overs occur should decrease as the number of overs decrease. Although there is, obviously, strong correlation between runs scored and over remaining this is still not the same thing. Put back into the repairable system context, D/L is saying that the rate of failures should change as the number of failures change rather than over time. This is also then a violation of one of the assumptions of an NHPP which is that failures occur independently.

Despite all of this its believed that the repairable systems framework could still be used help investigate the way in which the D/L model models resources. It may, however, require a more sophisticated form for the ROCOF to take and also use wickets as a more intuitive failure. It may also be fruitful to use observed average runs scored at the end of every over instead of trying to work backwards form the average runs remaining. This draws a less direct comparison with the way D/L model resources but could still shed light on how the current model is performing.

Chapter 7

Nonlinear least squares

Following the inconclusive results of attempting to model a batting innings as a repairable system, nonlinear regression is turned to in an attempt to quantify the visual results seen Chapter 5. Though no longer trying to model a dynamic batting innings, the static nonlinear regression will be used to fit curves to the runs scored in remaining resources data. This will provide a reasonable estimate at the confidential D/L model parameters and lead to more concrete conclusions about how the parameters that best fit the observed data may differ from these. This is now, only, testing the D/L model parameters rather than the actual form D/L resources take.

Nonlinear Models

The basic nonlinear model is as follows (Fox and Weisberg, 2010, p.2):

$$y = E[y|x] + \epsilon = m(x, \theta) + \epsilon$$

Where y is the response variable, x is the predictor variable(s) and $m(x, \theta))$ is the mean function, relating the predictor(s) and parameters, θ, to the response. The error, ϵ, are independent with mean 0 and unknown variance ω^2.

In a linear model m is a linear combination of functions of the predictors and the parameters, usually referred to as β. In Nonlinear regression m can be any nonlinear function depending on the known predictors x and unknown parameters (the number of parameters does need to be known).

Applying this setting to the way the D/L model was constructed yields; the response y as the resources remaining for a team who have, the predictor, x overs remaining. The mean function m depends on two parameters, $Z_0(w)$ and $b(w)$, which themselves depend on the number of wickets the team has lost at the time. As has been mentioned before, each curve that will be fitted keeps W constant so though $Z_0(w)$ and $b(w)$ depend on W, they will be a fixed parameters for each curve. Owing to the exponential decay model fitted by Duckworth and Lewis (1998) the mean function here will be of the form

$$m(x, \theta) = Z_0(w)[1 - exp(-b(w)x)]$$

Therefore using this mean function and the observed percentage resources remaining as the response the nonlinear regression model is

$$y = Z_0(w)[1 - exp(-b(w)x)] + \epsilon$$

Nonlinear least squares

As in linear regression, the parameters of a non linear regression are fitted by minimising the residual sum of squares:

$$RSS := \sum_{i=1}^{n}[y_i - m(x_i, \theta)]$$

Unfortunately, unlike in linear modelling, there does not exist a closed form solution to this minimisation problem and therefore an iterative mechanism is required (Fox and Weisberg, 2010, p.3).

Parameter Estimates

The 'nls' function in R can be used to fit the parameters of a nonlinear model by iteratively minimising the residual sum of squares (RSS). The port algorithm, one the of three that R offers, was found to perform the best allowing the necessary positivity constraint on both $Z_0(w)$ and $b(w)$. The port algorithm is an adaptation of the NL2SOL algorithm conceived by Dennis et al (1981). There is some discussion in R forums about whether it is entirely reliable in its R implementation. However, as demonstrated below, it seems to produce reasonable parameter estimates in the most part. Therefore the results are taken at face value, given more time the robustness of these parameters could be checked further.

In order to be consistent with the work done in Chapter 5, the response was taken to be remaining resources. When the D/L model was first built, this exponential decay relationship was fitted using runs obtainable rather than percentage resources remaining. This gives a more intuitive meaning to the model parameters and resources are simply extracted by dividing the runs obtainable by the runs the model expects a team to score in 100% of their resources. This means there is a 1:1 correspondence between percentage resources remaining and runs obtainable so there is no problem with using resources. Again, only data where the average had been calculated over 20 or more games was used to ensure both consistency with the previous analysis and the reliability of conclusions.

D/L parameters

The first stage of the analysis was to estimated the parameters of the D/L models, one set of parameters for each format T20 and Pro40. The D/L resources were produced by a model of this form and therefore the nonlinear regression produced should fit the data almost perfectly meaning the parameters estimates are the best possible guess at the actual D/L parameters. In fact none of the RSS for any of these models exceeded 1 so a near perfect fit was obtained. The parameters are in Table 7.1 below:

Table 7.1: Estimated D/L model parameters for the T20 and Pro40 resources tables.

D/L	**T20**			**Pro40**	
W	Z_0(W)	b(W)		Z_0(W)	b(W)
0	236.836	0.027		150.080	0.027
1	210.822	0.031		132.772	0.031
2	180.629	0.036		114.091	0.036
3	148.989	0.044		94.607	0.043
4	118.046	0.055		75.012	0.055
5	88.992	0.073		56.309	0.073
6	62.019	0.105		39.325	0.105
7	38.874	0.167		24.610	0.168
8	21.082	0.308		13.340	0.310
9	8.301	0.759		5.260	0.764

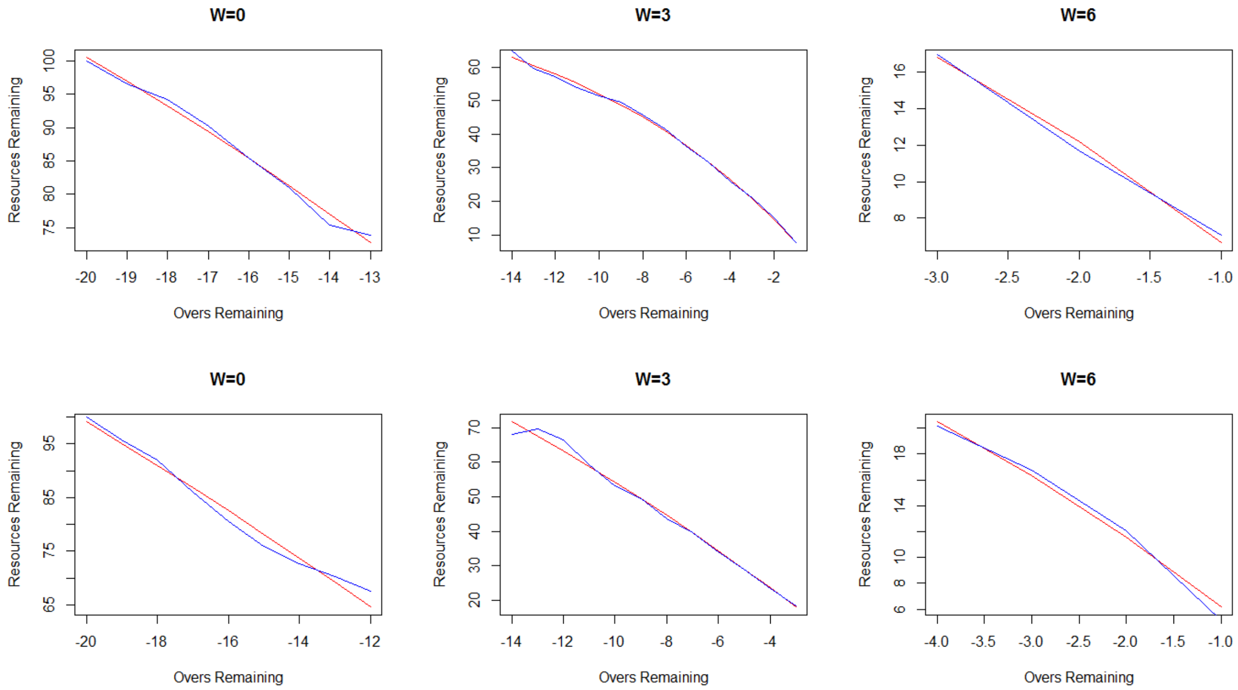
Duckworth and Lewis (1998) gave the interpretation that $Z_0(W)$ is the expected number of runs a team would score in infinite overs when they have lost w wickets. As the parameters here are in terms of percentage resources, rather than runs, the D/L model sets the percentage resources for an infinite overs batting innings, when 0 wickets have been lost, to be 237% of a T20 innings or 150% of a Pro40 innings. So, if a team could bat for infinite overs, the model expects them to score 2.37 times what they score in a T20 and 1.5 times what they score in a Pro40. Obviously, $Z_0(W)$ decreases as W increase as teams are expected to score less runs in infinite overs when they have lost more wickets. The parameter $b(W)$ is the exponential decay parameters, this is multiplied by the number of overs remaining, governing the way in which resources decay as overs remaining decrease. These parameters are almost identical for the T20 resource table and the Pro40 resource table, which is a result of the linear scaling down from the 50 over resource table. This is evidence of Duckworth and Lewis's judgement that resource decay in the same manner no matter how long the match is. The values of $b(W)$ increase as W increase meaning exponential decay is occurring with greater rate. As the mean function is of the form, $1 - exp(-b(W)u)$, the increased exponential decay results in resources decaying more slowly as more wickets are lost. This agrees with what was expected in Chapter 6. As more wickets are lost the batsman at the crease are less able and therefore the runs scoring potential lost in one over is less.

The fact that the D/L resource percentages came directly from a model of this form mean looking at any model diagnostic for these is pretty redundant. Its fair to assume these are a very good guess at the actual model parameters.

T20 parameters

Using an identical method to the D/L resource data, nonlinear regression lines were fitted to the T20 data and the model parameters were estimated. These are now being fitted to observed resources, corresponding to real world data, so the fit of these curves is less perfect, but still pretty good. As with the repairable system, only remaining resources that had been averaged over 20 data points or more were used for the parameter estimation. This was done in the hope to ensure the most reliable estimates possible, however this did result in very few data points being available for W=8 and 9 and thus the parameters estimation for these is omitted. The curves in Figure 7.1, below, demonstrate the fit for some wickets, the remainder of these are in Appendix 5.1.

Figure 7.1: Example graphs showing Overs Remaining vs fitted (red) and observed (blue) Resources. Above: First Innings. Below: Second Innings



The fitted curves seem to be a pretty good approximation of the observed data. The first innings curves seem to fit slightly better, something that is backed up by the RSS given below. A word of caution is given when using the RSS to compare curves, the number of overs for which we have sufficient data varies between first and second innings and therefore there are a slightly different number of fitted values. For example the RSS for W=6 and 7 appear really small but this is just because they are calculated over very few fitted values. The parameter estimates, along with the RSS and the D/L parameters, for these curves are in Table 7.2, below.

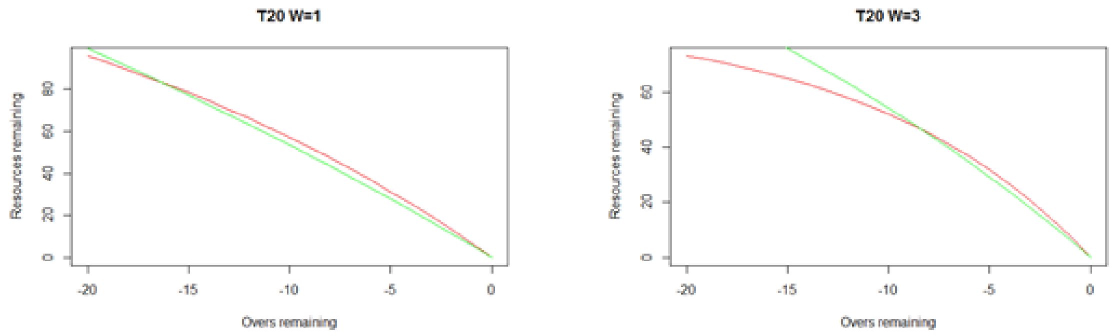
Table 7.2: Estimated T20 parameters for the D/L model, first innings and second innings resources

T20	**First Innings**			**Second innings**			**D/L**	
W	Z_0(W)	b(W)	RSS	Z_0(W)	b(W)	RSS	Z_0(W)	b(W)
0	205.36	0.034	6.565	280.691	0.022	24.03	236.836	0.027
1	177.936	0.039	32.36	**371.893**	**0.016**	47.77	210.822	0.031
2	122.973	0.061	17.83	**205.147**	**0.031**	49.98	180.629	0.036
3	87.564	0.09	9.078	207.791	0.03	30.76	148.989	0.044
4	69.179	0.112	0.7307	**85.61**	**0.08**	30.23	118.046	0.055
5	53.345	0.139	0.5961	123.638	0.051	21.48	88.992	0.073
6	40.549	0.178	0.4337	50.879	0.129	1.11955	62.019	0.105
7	41.895	0.171	0	17.389	0.452	0.415	38.874	0.167

Just like the D/L model parameters, the first innings parameters $Z_0(W)$ and $b(W)$ decrease and the increase with W, respectively. With the exception of W=1, 2 and 4 the second innings parameters behave similarly. $Z_0(1)$ is extremely high, higher then $Z_0(0)$ and whilst $Z_0(2)$ is lower, its also lower than $Z_0(3)$. $Z_0(4)$ is also a lot lower than $Z_0(3)$ and even lower than $Z_0(5)$. Obviously this will not be the case in reality, teams cannot be expected to score many more runs when they are 1 wicket down than they would when they are 0 wickets down, for example. As a result of this the $b(W)$ for these wickets also behave oddly. These strange parameters could be a result of the fitting of the model in *R*, however multiple starting values for the iterations have been tried and they all still converge to the same parameters so its unlikely that the algorithm has got stuck in a local minima. A more likely reason for this is that it just a result of working with messy real world data. These values, shown in bold in Table 7.2, are assumed to be incorrect and omitted.

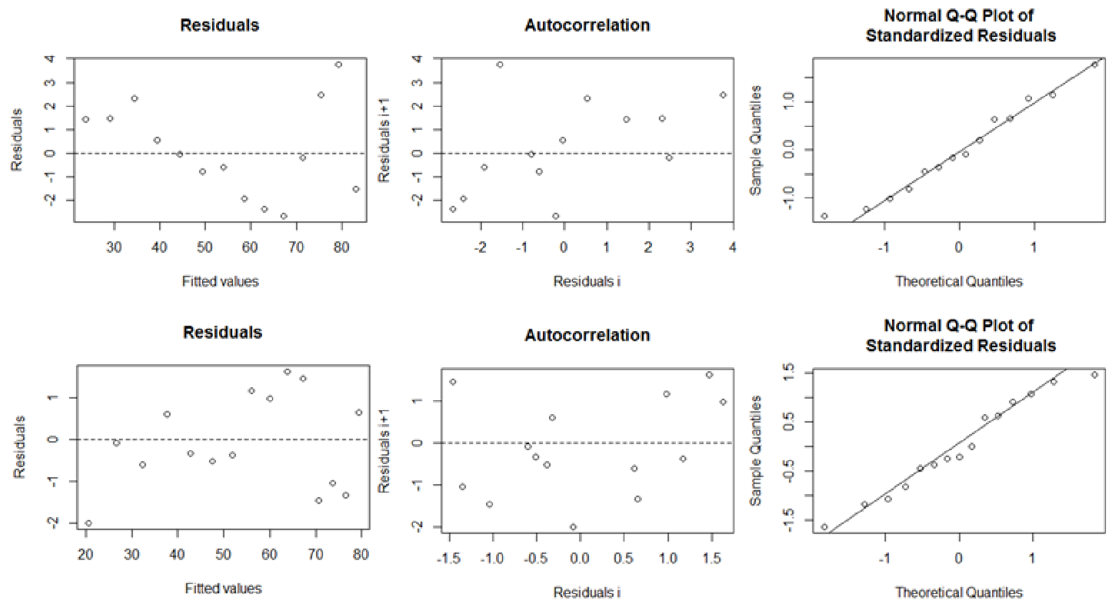
The values that are reasonable can now be compared across first and second innings. In general, the second innings estimates of $Z_0(W)$ appear much higher than the estimates from the first innings and the decay constant, $b(W)$, appear much lower for the second innings than it does for the first. This is a result of the visually observed trend that early second innings resources were greater than early first innings resources. This manifests itself by expecting teams batting second to score more runs in infinite overs but then decaying this at a greater rate. The D/L model parameters lie in between the first and second innings parameters again confirming what was observed visually. The greater $Z_0(W)$ for the second innings initially makes the second innings resources lie above first innings resources. The differing decay parameters then ensure first, second and D/L resources converge on each other as the innings progresses and, as was observed in in Chapter 5, the relationship reverses for the rest of the innings, i.e. first innings resources are higher than D/L resources and second innings resources for the remainder of the innings, to compensate. This is demonstrated in Figure 7.2, below, for W=1 and 3 , the red line indicates fitted first innings resources and the green line indicated fitted second innings resources.

Figure 7.2: Graphs demonstrating the comparisons between fitted first (red) and second (green) innings resources



Unfortunately there was not time to conduct a full diagnostic investigation to check that these models did not violate any of the nonlinear least squares assumptions. Instead, residual plots autocorrelation plots and normal qq plots were briefly examined to check the validity of some of the model assumptions and to see if any obvious violations were present. An example of a few of the plots, when W=2, for the first and second innings are in Figure 7.3, below, the rest are in Appendix 5.2.

Figure 7.3: Example diagnostic plots (Residual plots, Autocorrelation plots and Normal QQ plots) Above: T20 data, Below: Pro40 data

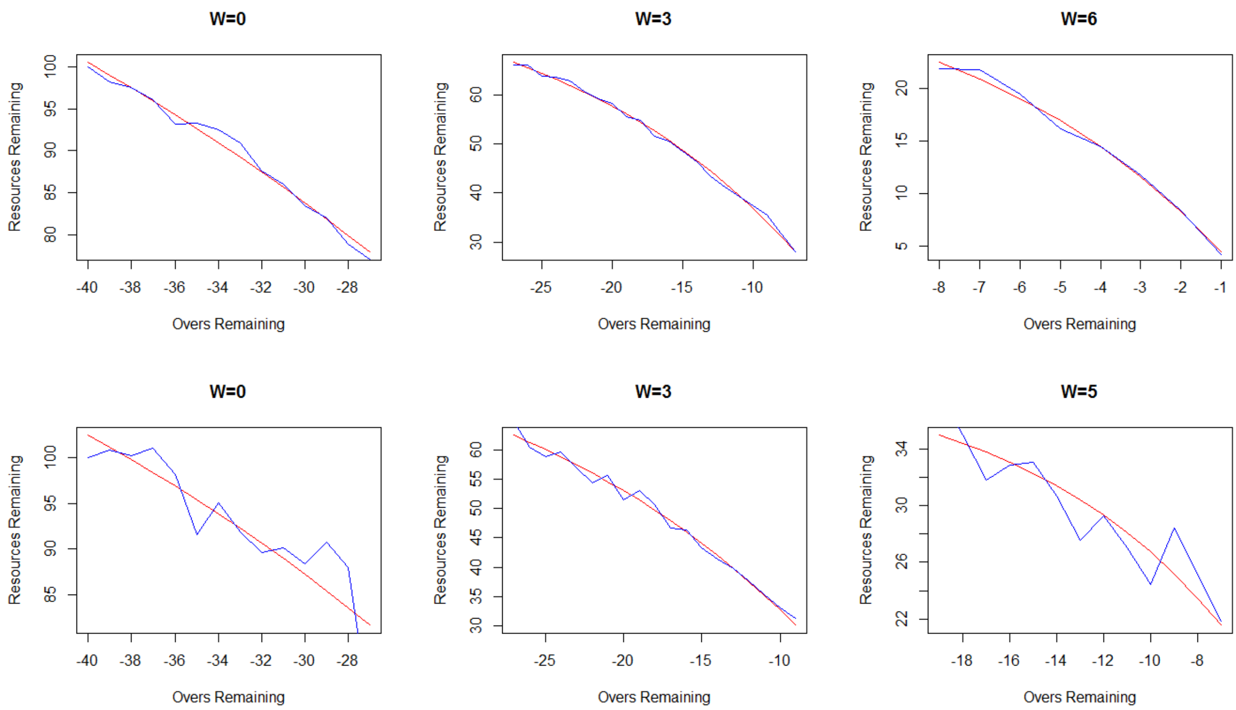


These examples are very similar to that seen for all wickets. The standardised residuals appear to be very close to being normally distributed and the autocorrelation plots show no real trend in the correlation between residual values. However the fitted values vs residuals plots show that the residuals are not independent of the fitted values, the residuals seem to oscillate around zero, almost like a sine curve. This is a trend that runs through that data and raises issues with the fitting of the nonlinear models. There was not time to investigate this further but given more time it should be.

Pro40 parameters

Exactly as was done for the T20 data, non linear regressions were then run in R for the Pro40 data in order to get estimates for the parameters $Z_0(W)$ and $b(W)$. Once again, the response here is taken from observed data so perfect fitting models are no longer expected, though the models do again seem to fit the data pretty well. This is demonstrated using fitted vs observed curves in Figure 7.4, below. Here only three wickets are shown, the rest are in Appendix 5.3. The first innings curves are above and the second innings below. There was not enough data here to fit linear regression for the second innings cases where W=6 and 7 and therefore these are omitted.

Figure 7.4: Example graphs showing Overs Remaining vs fitted (red) and observed (blue) Resources. Above: First Innings. Below: Second Innings



The first innings models here seem to be a good fit. Its hard to tell using RSS as there are obviously differing numbers of overs but they appear to fit as well as the T20 first innings curves do. Unfortunately the second innings curves do not appear to fit as well which is evident from both the graphs and the RSS. It was noticed in Chapter 5 that the observed second innings curves were much less smooth than the other curves. The reduced sample size of the 40 over data and the effect that this can have on the KM estimate of the average was the suspected cause of this. Besides this the second innings curves still appear to follow the general shape of the second innings and do a pretty good job of smoothing out the abnormal looking shape, so no problem was found in using the estimates from these regressions.

Table 7.3: Estimated Pro40 parameters for the D/L model, first innings and second innings resources

Pro40	**First Innings**			**Second innings**			**D/L**	
W	Z_0(W)	b(W)	RSS	Z_0(W)	b(W)	RSS	Z_0(W)	b(W)
0	161.140	0.024	10.410	144.246	0.031	175.900	150.080	0.027
1	111.907	0.044	53.590	114.818	0.037	138.100	132.772	0.031
2	101.696	0.046	52.160	107.673	0.037	92.450	114.091	0.036
3	84.988	0.057	8.276	86.403	0.048	23.340	94.607	0.043
4	55.466	0.099	12.110	55.159	0.101	65.270	75.012	0.055
5	46.925	0.114	8.781	39.708	0.112	41.580	56.309	0.073
6	32.325	0.149	1.985				39.325	0.105
7	29.241	0.136	0.586				24.610	0.168

Above, in Table 7.3 are the parameter estimates of $Z_0(W)$ and $b(W)$ for the observed Pro40 data as well as the Pro40 D/L parameter estimates. Unlike the T20 second inning data there do not appear to be any extraordinarily high or low estimates of both parameters and therefore a consistent trend is observed throughout. As was expected a priori and was observed in the D/L estimates, the parameters $Z_0(W)$ and $b(W)$, decreases and increase, respectively, as more wickets are lost.

Unlike the T20 data, and as was observed in Chapter 5, the first and second innings parameters appear a lot closer together. From W=1 onwards $Z_0(W)$ is slightly lower in the first innings than the second innings and $b(w)$ is conversely slightly higher. Like the T20 data, this means that second innings resources, when viewed for many overs, start higher but decay more quickly than first innings resources do, however here the parameters are much closer together and, as was observed in Chapter 5, actually result in first innings resources being slightly higher than second innings resources for the majority of a 40 over innings. In Chapter 5 this was suspected to be because of the very slight bias in the data towards the teams batting first, and the fact that the parameters are very close suggests that this could be the case. This all results in first and second innings resources appearing very close together for Pro40s, indicating that the judgement that its fair to use the same resource table for both does not appear to be wrong. When W=0 the converse of what is described above occurs, $Z_0(0)$ is greater for the first innings than the second innings and $b(W)$ is lower for the first innings. Unsurprisingly the parameters being opposite resulted in fitted second innings resources being higher than fitted first innings resources for the whole innings. When the winning bias is taken into account this really does not make sense and could be attributed to a failing of the nonlinear least squares model. Though this is opposite this is an isolated incidence so is ignored, though with more time this should receive further investigation.

The fact that these first and second innings parameters are also a lot closer to D/L model parameters is further evidence that the D/L resource table appears fairer for a Pro40 innings than it does for T20 innings, suggesting the closer an innings is to a 50 over innings the better D/L is doing. For W=0 the D/L resource parameters are both in between the first and second innings parameters and thus there is not enough evidence in this data set to suggest the D/L parameters are not reflecting what is observed, however with a fairer sample this could be tested. For W=1 onwards the D/L values of $Z_0(W)$ are all larger than both the first and second parameter estimates and the D/L parameter $b(W)$

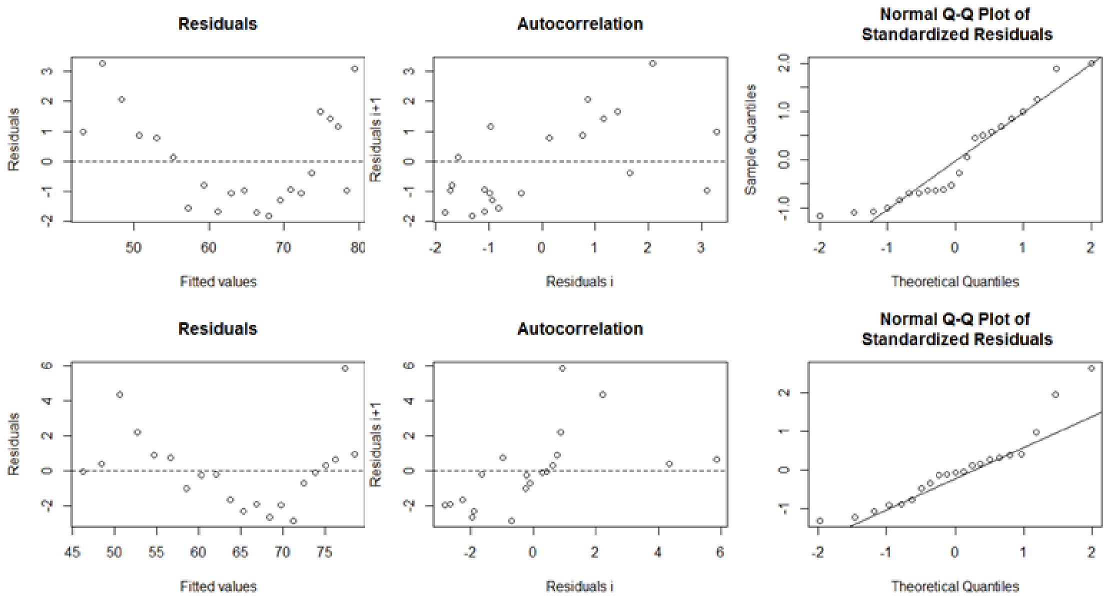
are lower than the first and second innings parameters. So D/L used for Pro40 expects teams to score more runs in infinite overs than the observed data suggest but D/L also decays their resources to fast.

Its hard too know how much impact the first innings winning bias has however, it is suspected in Chapter 5 that even without the bias that first innings resources would appear slightly higher than D/L predicts. The fact that the D/L parameters are closer to the second innings parameters than first innings parameters suggests that this is probably the case.

The effect that these parameters have on resources was expected after looking visually at the graphs in Chapter 5, however it has not manifested itself in the way that was expected. In Chapter 5 it was asserted that the D/L model had, generally, correctly modelled the way resources decay as overs remaining decrease and that any difference between what was observed and predicted was due to the D/L model decreasing runs scoring potential by too greater margin after each wicket was lost. Therefore it was expected that the D/L model's parameter $b(W)$ would be very close to the observed value and the D/L model's value of $Z_0(W)$ would be too low. Whats actually observed is that the D/L value of $b(W)$ is general too low and to compensate the value of $Z_0(W)$ is also too large. The differing values of $b(W)$ is not apparent when looking at the graphs due to the small number of overs involved in the graphs. The values of $b(W)$ differ by an order of 0.001 and therefore when the gradient is observed over 40 overs it will appear pretty similar.

Once again, residual plots, autocorrelation plots and nomall qq plots were briefly examined to get an idea of whether there were any obvious violations of any of the nonlinear least squares assumption present in these models. The first and second innings plots for W=2 are in Figure 7.5, below, the plots for the other wickets are in Appendix 5.4.

Figure 7.5: Example diagnostic plots (Residual plots, Autocorrelation plots and Normal QQ plots) Above: T20 data Below: Pro40 data



Here the standardised residuals appear to be pretty close to being normally distributed, the first inning auto correlations do not seem to have any real trend to them but the second innings ones do. Studying the rest of the second innings autocorrelation plots reveals this is an isolated case. Similarly to the T20 plots there is, however, an identifiable trend present in the fitted values vs residual plots. For W=2 they look like the shape of x^2 though for other wickets they again appear to oscillate around 0 like a sine curve. This shows a violation of the independence of errors assumption that is made in nonlinear regression. There was not time to investigate this further here but, given more time, this should be investigated in the further.

Comparing T20 and Pro40

So far, in this project, it has been very difficult to make comparisons between how the D/L model is performing across the two formats. Qualitative statements have been made based on the fitting of graphs and how far away points or parameters appear from each other but the different lengths of innings and the different sizes of the respective data sets have made it very hard to make more concrete conclusions. However, using the parameters estimates from the nonlinear regression, its hoped to change this. Currently its hard to compare the parameters across formats as they are currently with respect to different 100% resources situations. The T20 resources take a 20 overs with 0 wickets down a 100% resources whereas the Pro40 resources take 40 overs with 0 wickets down as 100% resources, understandably. Now, simply by applying the inverse of the transformation that was applied in Chapter 4 to create the 40 and 20 over resource tables, these can be scaled up so they both reflect resources in terms of a 50 over innings and therefore the parameters will be comparable.

This scaling was done by dividing the T20 resources, both the D/L table and the observed first and second innings, by 0.566, the 50 over resource percentage for the last 20 overs with 10 wickets remaining and dividing the Pro40 resources by 0.893, the 50 over resource percentage for the last 40 overs with 10 wickets remaining. This recovered the 50 over D/L table from the 20 and 40 over tables and scaled the observed first and second innings resources for both formats to show what these observed runs would constitute as part of a 50 over innings.

Using these re-scaled resource tables the nonlinear least squares were then run again in *R*. These are essentially the same models as before but scaled on the y axis, therefore, there is no need to analyse any diagnostic plots or fits of these model as they have been done before.

Table 7.4: Estimated parameters for the scaled up T20 and Pro40: D/L model, first innings and second innings resources

First Innings	**T20**		**Pro40**		**D/L approximation**	
W	Z_0(W)	b(W)	Z_0(W)	b(W)	Z_0(W)	b(W)
0	114.180	0.034	143.898	0.024	132.852	0.027
1	98.932	0.039	99.933	0.044	117.892	0.031
2	68.373	0.061	90.815	0.046	101.156	0.036
3	48.686	0.090	75.894	0.057	83.659	0.044
4	38.463	0.112	49.531	0.099	66.310	0.055
5	29.660	0.139	41.904	0.114	49.882	0.073
6	22.545	0.178	28.866	0.149	34.800	0.105
7	23.294	0.171	26.112	0.136	21.796	0.167

Second Innings	**T20**		**Pro40**		**D/L approximation**	
W	Z_0(W)	b(W)	Z_0(W)	b(W)	Z_0(W)	b(W)
0	156.064	0.022	128.812	0.031	132.852	0.027
1	**206.772**	**0.016**	102.533	0.037	117.892	0.031
2	**114.062**	**0.031**	96.152	0.037	101.156	0.036
3	115.532	0.030	77.158	0.048	83.659	0.044
4	**47.599**	**0.080**	49.257	0.101	66.310	0.055
5	68.743	0.051	35.459	0.112	49.882	0.073
6	28.289	0.129			34.800	0.105
7	9.668	0.452			21.796	0.167

Above, in Table 7.4, are the fitted parameters from the re-scaled resources percentages. As the 20 and 40 over resources were taken from the same table they should give exactly the same resource values and thus parameter values when re-scaled. Unfortunately due to some rounding error in the way these values have been scaled down and then re-scaled up, there were some very slight discrepancies in the parameter values, therefore to correct for this the approximate 50 over D/L parameters are calculated as an average of the parameters given by the scaled up 20 and 40 over models. As the scaling was the same for each wicket the trends as wickets are lost is the same as it was in the unscaled models and therefore here only the differences between the T20 and the Pro40 parameters will be discussed.

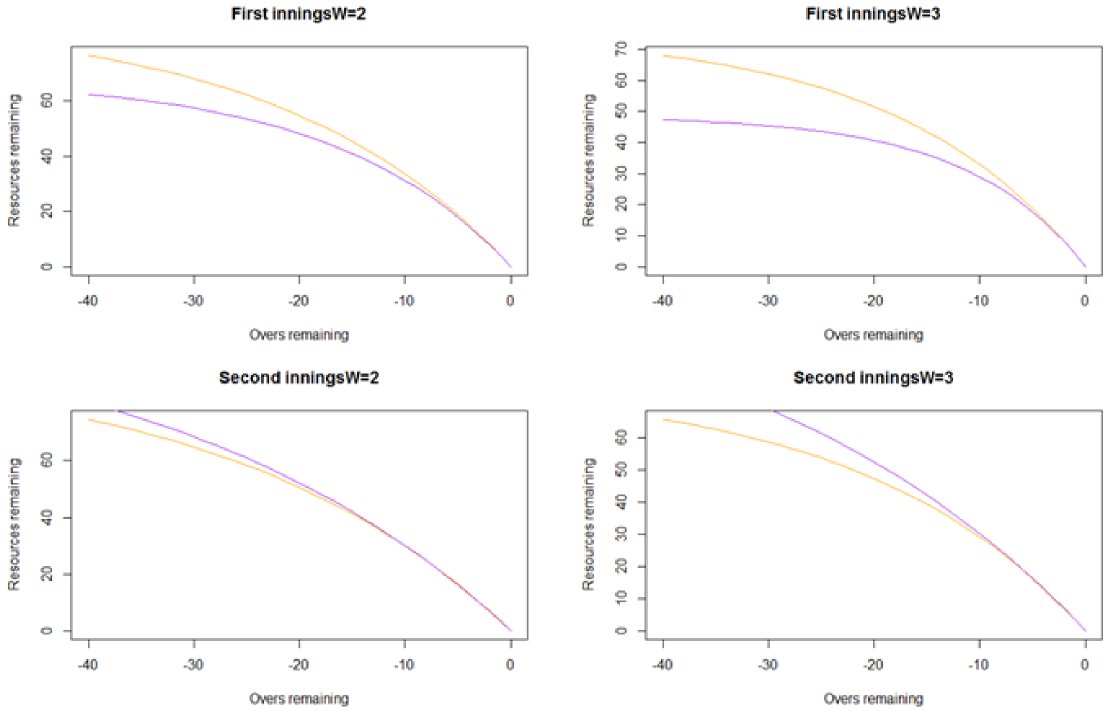
The first thing to notice is that for almost every wicket in both the first and second innings the estimated parameters from the Pro40 data are closer the estimates of the D/L model parameters than the T20 model parameters. This is the first confirmation of something that has been suspected right throughout this project, that the D/L table is better suited to use in Pro40s than it is to T20s. This is unsurprising as the scoring patterns in an innings 4/5 of the length is likely to be closer to the scoring pattern of the 'whole' innings than the scoring pattern of an innings that 2/5 of the length. This suggests that it is less fair to use the 50 over D/L table for Pro40s than it is for T20s (though its still possibly unfair to use 50 over D/l for Pro40).

For the first inning the T20 values of $Z_0(W)$ are always smaller than the Pro40 values and the T20 values of $b(W)$ are (almost) always bigger. This indicates that in infinite overs teams batting in the manner they bat in T20s are expected to score less than teams batting in the manner they bat in Pro40s. This resonates with cricketing logic, during a T20 innings the batting team take considerably more risks to score runs at a greater rate than they do in a Pro40 and they can do this as a T20 innings is shorter so they have to conserve their wickets for less time. However this increased risk taking means that if they batted in this manner for an infinite amount of time they would, likely, lose all ten wickets quicker than a team batting like they do in a 40 over matches, who take less risks so score at less of a rate but as a consequence bat for considerably longer score more runs. This is counter intuitive as if a team had infinite resources to bat then they would not need to bat as aggressively and could take less risks and this is the assumption that D/L make when they use the same parameters (Z_0) and thus the same resource table for 50 over games, Pro40s and T20s. However by doing this the different ways that teams bat in T20 and Pro40 is neglected. What these parameters are showing is that, in fact, the way resources are distributed is much more dependent on the way teams bat within that format of the game and therefore it may be fairer to incorporate this in the model by altering the way $Z_0(W)$ is defined and calculated. This would involve transforming Z_0 from runs scored in infinite overs to runs scored in infinite over batting in the manner teams do in that format. As a result of this the parameter $b(W)$ is greater for the T20 models than it is for the Pro40 models suggesting resources decay at a greater rate in Pro40s than they do in T20s, this makes slightly less sense but is a result of the differing $Z_0(W)$ values.

For the second innings, ignoring the values that are assumed not to be correct, the converse appears to be true. The T20 values for $Z_0(W)$ are greater than the respective Pro40 values and the T20 values of $b(W)$ are lower. This is strange given what we have seen for the first innings. It is mentioned above that having a larger value of $Z_0(W)$ for Pro40s makes more sense in terms of the way teams bat so therefore for this to reverse in the second innings is strange. As was observed earlier the T20 second innings parameter of $Z_0(W)$ being larger than the first innings value is a result of early second innings resources being greater than early first innings ones, it was also noticed that the Pro40 parameter values are very similar for first and second innings. Therefore the fact that runs scored in infinite overs for a second innings T20 is greater than it is for a second innings of a Pro40 is likely to be a manifestation of T20 second innings resources being higher at the start rather than any meaningful relationship between the two.

As is demonstrated in Figure 7.6, below, the overall effect of these 2 parameters is that first innings T20 resources (purple line) are initially, due to their lower $Z_0(W)$ value, comparatively lower than Pro40 resources (orange line) but as overs remaining decrease they converge on each other and the remaining overs are almost identical. Conversely second innings T20 resources (purple line) are initially, due to their higher $Z_0(W)$ value, comparatively higher than Pro40 resources (Orange line) before as overs remaining decrease, converging on each other. This is all attributed to difference between early first and second innings resources in a T20.

Figure 7.6: Graphs demonstrating the comparisons between fitted T20 and Pro40 resources, Above: First Innings. Below: Second innings



Conclusion

In this chapter nonlinear regression have been fitted using the observed data and the resource percentages from the D/L resource table in order to estimate the parameters used in the D/L model and to see how they compare to the parameters, calculated using D/Ls relationship, of the observed data. The fit for the models using the D/L resources percentages were pretty perfect so its fair to assume these parameters are almost exactly the parameters used by D/L. The observed curves obviously do not fit perfectly. However, the fitted curves, visually, seem to do a pretty good job of fitting the general shape of the data and therefore their parameters are thought to be pretty good estimates. Some diagnostic issues were raised when residual plots were examined and these would need to be examined further if more time was available. However, almost all of the parameters were around where they could reasonably be expected to be so its felt that reasonably strong conclusions can be drawn from these parameters. The three parameters from the T20 first innings that did not seem reasonable were ignored.

There was unfortunately not enough time to make this more rigorous by hypothesis testing the parameters, as this would have required the checking of the test assumptions. However this could be done easily with more time in the future to solidify the conclusions.

This analysis was conducted in an attempt to corroborate and, were differences applied, quantify what was suspected after looking at the graphs in Chapter 5. In the most part this was the case however the observed differences did not manifest themselves in the way that was expected.

Firstly it appeared that there is considerable evidence here that first and second innings T20 resources are not distributed in the same way. The model parameter $Z_0(W)$ was much greater in the second innings than it was in the first innings with the decay of second innings resources proceeding to be quicker than the decay of first inning resources. The D/L parameters lie in between the first and second innings parameters throughout. This results in the second innings resources initially being higher than D/L resources with first innings resources lower and then, as overs remaining decrease, the resource curves cross and first innings resources are higher than D/L with second innings resources lower for the rest of the innings. This is exactly what has been seen in the observed curves. It has been discussed previously that early second innings resources being higher than early first innings resources and the subsequent greater decay of second innings resources makes sense. However to model this using the formula given by the D/L model requires a large increase in the value of $Z_0(W)$, the theoretical average runs scored in infinite resources, for the second innings and this makes much less sense in reality. There is no reason why is a team batting first for infinite overs should score less than a team batting second for infinite over. This theoretical parameter does not actually make sense when applied to a second innings as the team batting second would never actually have the opportunity to bat for infinite overs, at some point they would reach team 1's total, something that obviously would not happen to a team batting first. It is therefore, understandable that the D/L model uses the same parameters for each innings. However the differences found come back to the way in which $Z_0(W)$ is defined, which was discussed earlier. Presumably the fact that they have a total in mind suggests that if a team batting second batted for infinite overs they would score more runs that the team batting first and therefore there is some cause here for the values of $Z_0(W)$ to be changed to reflect the differing ways teams bat in the first and second innings. Based on this evidence it could be viewed as unfair to have the same resource table for both innings of a T20 game as the assumption that the resources are distributed in the same way appears to be invalid.

Again agreeing with the visual interpretations of the graphs it appears that the same phenomenon does not happen in the Pro40 data and in fact there is not that much difference in the way resources are distributed in the first and second innings of a Pro40. The estimated parameters for the Pro40 first and second innings data are much closer together than the T20 first and second innings and the slight bias in the data towards teams batting first is blamed for the first innings parameters leading to slightly higher first innings resources throughout. The D/L parameters are always closer to the second innings parameters then they are to the first innings parameters and this is seen to corroborate the visual findings that even when the bias in the data is corrected Pro40 resources will still lie above the D/L predicted resources.

Whilst being slightly above, the first innings resources appeared to decay at the same rate as the D/L resources and therefore it was expected in Chapter 5 that the exponential decay constant $b(W)$ would be the same for the both graphs and that the reason first innings resource lay above D/L resources was that D/L was over punishing teams for losing wickets, i.e. setting $Z_0(W)$ too low. In fact these parameter estimates suggest that in fact the D/L model is decaying resources too quickly and therefore the estimate of $Z_0(W)$ is too high. The gradients appeared similar, when viewed over a 40 over innings because the parameters $b(W)$ are quite close together and very small. Thus when overs remaining are small (even 40 is quite small) the gradients appear very similar. The reason for $Z_0(W)$ being lower than expected is put down to the same reason as why the $Z_0(W)$ for T20 first innings is lower than Pro40 first innings. Risk taking in the shorter format of the game is likely to be greater and thus if you extrapolate that over an infinite number of overs the more aggressively the team bat, the quicker they lose all 10 wickets and therefore they seem to be able to compile less runs.

From this analysis there is no evidence that having the same resource table for the first and second innings of a Pro40 game is leading to any unfairness for either side, there is however some evidence that the parameters estimated using the 40 over data differ from the D/L model parameters and therefore it could be viewed as unfair to use the same resource table for the Pro40 format of the game to the 50 over format.

This analysis has also lead to a conclusion on something that has been suspected throughout the project that the D/L model is performing worse, and could be considered less fair, when used in T20s compared with Pro40s. This conclusion was made based on the fact that the parameters estimated from the scaled Pro40 model are much closer to the D/L parameters than the parameters estimated from the scaled T20 data, suggesting the resources distribution in a Pro40 is closer to that predicted by D/L than it is in a T20. The values of $b(W)$ do not actually differ by very much and as they are so small the difference does not actually have a huge impact on the way resources are distributed over the short number of overs remaining in an innings. However the values of $Z_0(W)$ differ greatly. This is believed to be because of the way they are defined, D/L define it as the score a team would get in infinite overs with $10 - W$ wickets remaining,. As Z_0 is runs in infinite overs the format of the game is redundant and therefore D/L assume this is the same for any length of the game. It appears from the data that this value is not the same across different formats. It is suggested here that this value should be adjusted to incorporate the manner in which teams bat in different formats of the game. For example in a T20 game teams take more risks in order to score runs at a faster rate and therefore in infinite overs will be bowled out quicker and ,likely, for less runs than a team who are batting more conservatively in a Pro40. It therefore may not be fair to use the same D/L resource table in all formats of the game as it is shown here that the differing ways teams bat in different formats influences how their resources are distributed.

Parts of this analysis also raise issues with the fitting of model parameters, does one favour statistical accuracy over real world sense. Duckworth and Lewis used statistical accuracy to fit the D/L model parameters and then they used common sense for example to assume these would be the same in the first and second innings, as it is not unreasonable to assume that how many runs a team score in infinite overs is the same whether their innings is the first or the second. This analysis has relied completely on statistical accuracy. Some of the findings have been explained by cricketing logic, but some of these, whilst producing the desired and expected effect in terms of resources, do not make sense when looked at using the definition the D/L model gives them. For example, even if $Z_0(W)$ is adjusted to incorporate the, previously discussed, effect chasing a total has to the way teams bat in their second innings and the effect differing length of innings has on the way teams bat in Pro40s and T20s. The resulting values of $Z_0(W)$ are such that a team batting like they would in the second innings of a T20 would score more runs in infinite overs than a team batting first in a Pro40. This is caused by the fact that early second innings resources in a T20 are higher than early first innings resources, causing the second innings value of $Z_0(W)$ to be higher than the first innings value. This gives the desired effect on resources, that in a T20 early second innings resources are higher than early first innings ones, but makes no sense in reality. Therefore this comes down to another value judgement from the model builders. Do they want their model to be as statistically accurate as possible and thus compromising interpretability or do they want to adjust the parameters so that the model is less statistically accurate but whose parameters make more sense.

The final thing that can be concluded from this analysis is that the D/L model appears to be fairly insensitive to the value of its parameters. Visually the shape of the all 3 curves generally appears the same indicating that the values of $b(W)$ would be the same. In fact they actually differ quite a lot as a percentage from one another, and this evidently results in very little visual change, though this could just be because they are shown over a small number of overs. Again looking visually, the D/L curves did not appear too far from the observed curves despite the fact that some quite large changes in the estimates values of $Z_0(W)$ were observed. This is the sign of a strong model and means any judgements made by the model builders should not have a too larger impact on the end result.

Chapter 8

Conclusion

The aim of this project was to examine some of the key value judgements made by Frank Duckworth and Anthony Lewis, when they first constructed their D/L model in 1998, to see if there was any evidence in 2 or 3 seasons worth of county cricket data that would indicate any of these judgements result in the D/L method being unfair to either side.

Data analysis Conclusions

Firstly the D/L model assumption that it fair to assume a team performs averagely, independent of how they have performed before, during a stoppage in play was investigated. Very few significant correlations were found between past and future performance and therefore it was concluded that it would be unfair to assume a team performed anything other than averagely during any resources they miss out on. There were however some correlations significantly greater than 0 in the T20 data but these would need more time to be investigated properly.

Secondly there was evidence that it may be unfair to use the same resource table for the first and second innings of a T20 match. It appears visually, and is corroborated by parameter estimates, that the resource distribution in the second innings of a T20 differs considerably from the resources distribution of the first innings and in fact these both differ from the resources distribution predicted by the D/L model. Early second innings resources appear greater than early first innings ones but as the innings moves on, first innings resources become higher than second innings resources and remain that way for the remainder of the innings. Throughout this the D/L predicted resources appear to lie consistently in between both of them. It is therefore concluded here that using the same resource table during both innings of a T20 match could lead to the game being unfair to either side once the match has started. There is no such evidence that this is occurring in Pro40s and therefore using the same resource table for both innings of a Pro40 game does not appear to lead to any unfairness on either side.

Finally there is also evidence throughout this analysis that it may be unfair to use the same resource table for 50 over cricket, Pro40s and T20s, and that actually using the 50 over resource table could be considered to be less fair in a T20 match than it would be in a Pro40 match. Visually observed Pro40 resources appeared to be closer to the resources that D/L predicted than observed T20 resources and this was verified by the fact that the parameters estimates associated with the Pro40 data were closer to the estimates of the D/L parameters than the estimates associated with the T20 data. In fact, combining this with the evidence above, there is evidence here that the resources distribution in both Pro40s and T20s are both different to the resources distribution in a 50 over game and therefore it could be seen to be unfair on one side to use the same resource table for all formats of cricket. It is suggested that this could be changed by altering the parameter $Z_0(W)$ so that instead of representing the universal expected number of runs a team would score in infinite overs, there is a different Z_0 for each format and it now represents the number of runs a team could expect to score if they batted in the manner that the average team bats in that format of the game. Then the decay constants can be modified in order to make the resulting resource table reasonable.

In fact, the most recent adaptation to the D/L model is to adjust the parameters when the model is used for T20s so that they better reflect the way a teams in T20s (Lewis 2015). This is recognition that using the same resource table for T20s and 50 over games was leading to unfairness and it was therefore felt that the D/L model needed changing, agreeing with the findings presented here. No such change has been proposed for used in Pro40s, but as the international game is not played over 40 overs, it receives much less interest from the media and is therefore worked on less. In addition, and contrary to whats been discovered here the same resource table is still being used for the first and second innings in any length game. Whilst there is evidence in the data that has been analysed here that this may be unfair, it is believed, by the people in charge of the D/L model, that the added constraint of the team batting second trying to win makes looking at second innings batting patterns useless (Stern 2015), this in itself represents another value judgement.

Examination of Value judgements

All of these conclusions above are based on empirical findings from observed data, and the only relate to the whether the D/L model appears unfair at differing points in time. In fact there is no right or wrong answer to the point in which the model should be fair to both teams and this is simply and value judgement made by the model builders and the people who regulate cricket.

Originally Duckworth and Lewis set their model up so that it would be fair to both sides at the start of the season, by suggesting that the same resource table should be used for every innings in every format of cricket at every ground ect. The D/L resource table is considered to do a pretty good job of doing this and is thus considered to be fair to every team at the start of the season. By ensuring the model was fair at the start of the season the D/L model uses as little information as possible to produce its resource table, meaning the model does not try to capture anything too complicated reducing the likely error in the model but also meaning that in some circumstance the model could be exploited and prove unfair.

It has been shown in this project that, after the format of the game has been decided, the D/L model is no longer fair to both sides as it appears that the D/L model generally undervalues the remaining resources of teams batting in Pro40 cricket and after the toss, to decide which team will bat or bowl first, has been done in a T20 the D/L model is no longer fair as first and second innings resources are distributed differently.

However there is nothing, except maybe the views of the players and the media, saying that the model should be fair at these points. The reason the toss exists in the first place is to randomise the team that bat first and this is required as predetermining the team that bat first is, evidently, considered unfair. This is especially evident when considering a wet green pitch on a cloudy day, before the toss is done the game is fair to both sides however after the toss when its decided that one team will have to bat in these unhelpful conditions, the game is no longer fair. Therefore it seems reasonable that the D/L model does not need to be adjusted to incorporate the differing way that teams bat in first and second innings.

Though it makes more sense that a game is fair at the start of the match when the format has been decided, a cricket match could also be considered already unfair when the format has been decided. For example one team may be better suited to playing a Limited Overs game and therefore may have an unfair advantage over a team better suited to First Class cricket in a T20. So maybe it is not necessary that the D/L model is fair at this point. In fact the D/L model has recently been changed adopting a different method for T20. This represents a change in the value judgement that the D/L model is fair at the start of the season, such that it is now fair when the format of each game through the season has been established. This judgement has presumably been reversed based on an empirical study, though possibly more thorough, not dissimilar to the one conducted above. Once again there is nothing saying this is new judgement is any more correct than before.

The final word in this dissertation regards the occurrence of value judgements in statistical analysis. It is actually very difficult to construct a model or perform any analyses without making value judgements yourself, and this is something that is often under appreciated by statisticians. I, for example, have made value judgements throughout this analysis, such as that 10% was the correct significance level to use to test whether the correlations were significantly different from 0 and that 20 was a sufficient sample size to calculate averages over. These judgements were made without investigation or statistical justification and were simply made because I felt they led to the fairest possible analysis. In the simplest possible sense these judgements dont differ greatly from the ones made by Duckworth and Lewis in 1998.

Chapter 9

References

BBC, (2007), "A decade of Duckworth-Lewis", [Online], (URL http://news.bbc.co.uk/sport1/hi/cricket /6222943.stm), (Accessed 1 March 2015).

Bhattacharya, R., Gill, P.S. and Swartz, T.B. (2011),"DuckworthLewis and Twenty20 cricket", Journal of the Operational Research Society 62(11), 1951-1957.

Carter, M. and Guthrie, G. (2004),"Cricket interruptus: fairness and incentive in limited overs cricket matches", Journal of the Operational Research Society 55(8), 822-829.

Crowder, M.J., Kimber A.C., Smith, R.L. and Sweeting T.J. (1991),"Statistical Analysis of Reliability Data", 1st ed, London: Chapman & Hall.

Das, S. (2011),"On Generalized Geometric Distributions: Application to Modeling Scores in Cricket and Improved Estimation of Batting Average in Light of Notout Innings", IIM Bangalore Research Paper No. 356.

Dennis, JR., J.E., Gay, D.M. and Welsch R.E. (1981), "An Adaptive Nonlinear Least-Squares Algorithm", ACM Transactions on Mathematical Software, 7(3), 348-368

Duckworth, F.C. and Lewis, A.J. (1998), A fair method for resetting the target in interrupted one-day cricket matches, Journal of the Operational Research Society 49(3), 220-227.

Duckworth, F.C. and Lewis, A.J. (2004), "A successful operational research intervention in one-day cricket", Journal of the Operational Research Society 55(7), 749-759.

ESPN Sports Media Ltd. (2014), "County Cricket 2013" , [Online], (URL http://www.espncricinfo.com /county-cricket-2013/content/series/585678.html), (Accessed 24 October 2014)

ESPN Sports Media Ltd. (2015), "Mike Atherton" , [Online], (URL http://www.espncricinfo.com/eng land/content/player/8579.html), (Accessed 4 March 2015)

Fox, J. and Weisberg, S. (2010, December 13), "Nonlinear Regression and Nonlinear Least Squares in R:An Appendix to An R Companion to Applied Regression, second edition", [pdf] Sage Publications, (URL http://socserv.socsci.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Nonlinear-Regr ession.pdf), (Accessed 3 April 2015)

ICC (2013), “ICC Playing handbook 2013-2014”, [pdf] ICC, (URL http://static.icc-cricket.com/ugc/documents/DOC_469718256094228F17EC0A395AC0E6FC_1385049528133_552.pdf), (Accessed 29 October 2014).

Kaplan and Meier (1958), “Nonparametric Estimation from Incomplete Observations”, Journal of the American Statistical Association 53(282), 457-481

Kimber, A.C. and Hansford, A.R. (1993), “A Statistical Analysis of Batting in Cricket, Royal Statistical Society 156(3), 443-455.

Lewis, A.J. (2015), OR’s solution of problems in rain-interrupted one-day cricket matches; the story of the Duckworth/Lewis method, [Conversation], (Personal Communication, 4 March 2015).

The R Stats Package: Documentation for package stats version 3.3.0 (2015), “Test for Association/Correlation Between Paired Samples”, [Online], (URL https://stat.ethz.ch/R-manual/R-devel/library/stats/html/cor.test.html), (Accessed 15 November 2014).

Stern, S.E. (2009), “An adjusted DuckworthLewis target in shortened limited overs cricket matches”, Journal of the Operational Research Society 60(2), 236-251.

Stern, S.E. (2015), Discussion surrounding the application of Duckworth Lewis, especially in the second innings of a cricket match, [E-mail], (Personal Communication, 2 February 2015).

Wood, R. (2008), “World’s Most Popular Sports by Fans”, [Online], (URL http://www.topendsports.com/world/lists/popular-sport/fans.htm), (Accessed 2 March 2015).

Chapter 10

Appendix

Appendix 1 - Resources Tables

1.1 - D/L 50 over Resource table - ICC (2013)

Resources Remaining	Wickets Lost									
Overs Remaining	0	1	2	3	4	5	6	7	8	9
50	100	93.4	85.1	74.9	62.7	49	34.9	22	11.9	4.7
49	99.1	92.6	84.5	74.4	62.5	48.9	34.9	22	11.9	4.7
48	98.1	91.7	83.8	74	62.2	48.8	34.9	22	11.9	4.7
47	97.1	90.9	83.2	73.5	61.9	48.6	34.9	22	11.9	4.7
46	96.1	90	82.5	73	61.6	48.5	34.8	22	11.9	4.7
45	95	89.1	81.8	72.5	61.3	48.4	34.8	22	11.9	4.7
44	93.9	88.2	81	72	61	48.3	34.8	22	11.9	4.7
43	92.8	87.3	80.3	71.4	60.7	48.1	34.7	22	11.9	4.7
42	91.7	86.3	79.5	70.9	60.3	47.9	34.7	22	11.9	4.7
41	90.5	85.3	78.7	70.3	59.9	47.8	34.6	22	11.9	4.7
40	89.3	84.2	77.8	69.6	59.5	47.6	34.6	22	11.9	4.7
39	88	83.1	76.9	69	59.1	47.4	34.5	22	11.9	4.7
38	86.7	82	76	68.3	58.7	47.1	34.5	21.9	11.9	4.7
37	85.4	80.9	75	67.6	58.2	46.9	34.4	21.9	11.9	4.7
36	84.1	79.7	74.1	66.8	57.7	46.6	34.3	21.9	11.9	4.7
35	82.7	78.5	73	66	57.2	46.4	34.2	21.9	11.9	4.7
34	81.3	77.2	72	65.2	56.6	46.1	34.1	21.9	11.9	4.7
33	79.8	75.9	70.9	64.4	56	45.8	34	21.9	11.9	4.7
32	78.3	74.6	69.7	63.5	55.4	45.4	33.9	21.9	11.9	4.7
31	76.7	73.2	68.6	62.5	54.8	45.1	33.7	21.9	11.9	4.7
30	75.1	71.8	67.3	61.6	54.1	44.7	33.6	21.8	11.9	4.7
29	73.5	70.3	66.1	60.5	53.4	44.2	33.4	21.8	11.9	4.7
28	71.8	68.8	64.8	59.5	52.6	43.8	33.2	21.8	11.9	4.7

Overs Remaining	0	1	2	3	4	5	6	7	8	9
27	70.1	67.2	63.4	58.4	51.8	43.3	33	21.7	11.9	4.7
26	68.3	65.6	62	57.2	50.9	42.8	32.8	21.7	11.9	4.7
25	66.5	63.9	60.5	56	50	42.2	32.6	21.6	11.9	4.7
24	64.6	62.2	59	54.7	49	41.6	32.3	21.6	11.9	4.7
23	62.7	60.4	57.4	53.4	48	40.9	32	21.5	11.9	4.7
22	60.7	58.6	55.8	52	47	40.2	31.6	21.4	11.9	4.7
21	58.7	56.7	54.1	50.6	45.8	39.4	31.2	21.3	11.9	4.7
20	56.6	54.8	52.4	49.1	44.6	38.6	30.8	21.2	11.9	4.7
19	54.4	52.8	50.5	47.5	43.4	37.7	30.3	21.1	11.9	4.7
18	52.2	50.7	48.6	45.9	42	36.8	29.8	20.9	11.9	4.7
17	49.9	48.5	46.7	44.1	40.6	35.8	29.2	20.7	11.9	4.7
16	47.6	46.3	44.7	42.3	39.1	34.7	28.5	20.5	11.8	4.7
15	45.2	44.1	42.6	40.5	37.6	33.5	27.8	20.2	11.8	4.7
14	42.7	41.7	40.4	38.5	35.9	32.2	27	19.9	11.8	4.7
13	40.2	39.3	38.1	36.5	34.2	30.8	26.1	19.5	11.7	4.7
12	37.6	36.8	35.8	34.3	32.3	29.4	25.1	19	11.6	4.7
11	34.9	34.2	33.4	32.1	30.4	27.8	24	18.5	11.5	4.7
10	32.1	31.6	30.8	29.8	28.3	26.1	22.8	17.9	11.4	4.7
9	29.3	28.9	28.2	27.4	26.1	24.2	21.4	17.1	11.2	4.7
8	26.4	26	25.5	24.8	23.8	22.3	19.9	16.2	10.9	4.7
7	23.4	23.1	22.7	22.2	21.4	20.1	18.2	15.2	10.5	4.7
6	20.3	20.1	19.8	19.4	18.8	17.8	16.4	13.9	10.1	4.6
5	17.2	17	16.8	16.5	16.1	15.4	14.3	12.5	9.4	4.6
4	13.9	13.8	13.7	13.5	13.2	12.7	12	10.7	8.4	4.5
3	10.6	10.5	10.4	10.3	10.2	9.9	9.5	8.7	7.2	4.2
2	7.2	7.1	7.1	7	7	6.8	6.6	6.2	5.5	3.7
1	3.6	3.6	3.6	3.6	3.6	3.5	3.5	3.4	3.2	2.5

1.2 - 40 Over scaled down Resource Table

Resources Remaining	Wickets Lost									
Overs Remaining	0	1	2	3	4	5	6	7	8	9
40	100	94.3	87.1	77.9	66.6	53.3	38.8	24.6	13.3	5.3
39	98.5	93.1	86.1	77.3	66.2	53.1	38.6	24.6	13.3	5.3
38	97.1	91.8	85.1	76.5	65.7	52.7	38.6	24.5	13.3	5.3
37	95.6	90.6	84	75.7	65.2	52.5	38.5	24.5	13.3	5.3
36	94.2	89.3	83	74.8	64.6	52.2	38.4	24.5	13.3	5.3
35	92.6	87.9	81.8	73.9	64.1	52	38.3	24.5	13.3	5.3
34	91	86.5	80.6	73	63.4	51.6	38.2	24.5	13.3	5.3
33	89.4	85	79.4	72.1	62.7	51.3	38.1	24.5	13.3	5.3
32	87.7	83.5	78.1	71.1	62	50.8	38	24.5	13.3	5.3
31	85.9	82	76.8	70	61.4	50.5	37.7	24.5	13.3	5.3
30	84.1	80.4	75.4	69	60.6	50.1	37.6	24.4	13.3	5.3
29	82.3	78.7	74	67.8	59.8	49.5	37.4	24.4	13.3	5.3
28	80.4	77	72.6	66.6	58.9	49.1	37.2	24.4	13.3	5.3
27	78.5	75.3	71	65.4	58	48.5	37	24.3	13.3	5.3
26	76.5	73.5	69.4	64.1	57	47.9	36.7	24.3	13.3	5.3
25	74.5	71.6	67.8	62.7	56	47.3	36.5	24.2	13.3	5.3
24	72.3	69.7	66.1	61.3	54.9	46.6	36.2	24.2	13.3	5.3
23	70.2	67.6	64.3	59.8	53.8	45.8	35.8	24.1	13.3	5.3
22	68	65.6	62.5	58.2	52.6	45	35.4	24	13.3	5.3
21	65.7	63.5	60.6	56.7	51.3	44.1	34.9	23.9	13.3	5.3
20	63.4	61.4	58.7	55	49.9	43.2	34.5	23.7	13.3	5.3
19	60.9	59.1	56.6	53.2	48.6	42.2	33.9	23.6	13.3	5.3
18	58.5	56.8	54.4	51.4	47	41.2	33.4	23.4	13.3	5.3
17	55.9	54.3	52.3	49.4	45.5	40.1	32.7	23.2	13.3	5.3
16	53.3	51.9	50.1	47.4	43.8	38.9	31.9	23	13.2	5.3
15	50.6	49.4	47.7	45.4	42.1	37.5	31.1	22.6	13.2	5.3
14	47.8	46.7	45.2	43.1	40.	36.1	30.2	22.3	13.2	5.3
13	45	44	42.7	40.9	38.3	34.5	29.2	21.8	13.1	5.3
12	42.1	41.2	40.1	38.4	36.2	32.9	28.1	21.3	13	5.3
11	39.1	38.3	37.4	36	34	31.1	26.9	20.7	12.9	5.3
10	36	35.4	34.5	33.4	31.7	29.2	25.5	20	12.8	5.3
9	32.8	32.4	31.6	30.7	29.2	27.1	24	19.2	12.5	5.3
8	29.6	29.1	28.6	27.8	26.7	25	22.3	18.1	12.2	5.3
7	26.2	25.9	25.4	24.9	24	22.5	20.4	17	11.8	5.3
6	22.7	22.5	22.2	21.7	21.1	19.7	18.4	15.6	11.3	5.2
5	19.3	19	18.8	18.5	18	17.3	16	14	10.5	5.2
4	15.6	15.5	15.3	15.1	14.8	14.2	13.4	12	9.4	5
3	11.9	11.8	11.7	11.5	11.4	11.1	10.6	9.7	8.1	4.7
2	8.1	8	8	7.8	7.8	7.6	7.4	6.9	6.2	4.1
1	4	4	4	4	4	3.9	3.9	3.8	3.6	2.8

1.3 - 20 Over scaled down Resource Table

Resources Remaining	Wickets Lost									
Overs Remaining	0	1	2	3	4	5	6	7	8	9
20	100	96.8	92.6	86.7	78.8	68.2	54.4	37.5	21.3	8.3
19	96.1	93.3	89.2	83.9	76.7	66.6	53.5	37.3	21	8.3
18	92.2	89.6	85.9	81.1	74.2	65	52.7	36.9	21	8.3
17	88.2	85.7	82.5	77.9	71.7	63.3	51.6	36.6	21	8.3
16	84.1	81.8	79	74.7	69.1	61.3	50.4	36.2	20.8	8.3
15	79.9	77.9	75.3	71.6	66.4	59.2	49.1	35.7	20.8	8.3
14	75.4	73.7	71.4	68	63.4	56.9	47.7	35.2	20.8	8.3
13	71	69.4	67.3	64.5	60.4	54.4	46.1	34.5	20.7	8.3
12	66.4	65	63.3	60.6	57.1	51.9	44.3	33.6	20.5	8.3
11	61.7	60.4	59	56.7	53.7	49.1	42.4	32.7	20.3	8.3
10	56.7	55.8	54.4	52.7	50	46.1	40.3	31.6	20.1	8.3
9	51.8	51.1	49.8	48.4	46.1	42.8	37.8	30.2	19.8	8.3
8	46.6	45.9	45.1	43.8	42	39.4	35.2	28.6	19.3	8.3
7	41.3	40.8	40.1	39.2	37.8	35.5	32.2	26.9	18.6	8.3
6	35.9	35.5	35	34.3	33.2	31.4	29	24.6	17.8	8.1
5	30.4	30	29.7	29.2	28.4	27.2	25.3	22.1	16.6	8.1
4	24.6	24.4	24.2	23.9	23.3	22.4	21.2	18.9	14.8	8
3	18.7	18.6	18.4	18.2	18	17.5	16.8	15.4	12.7	7.4
2	12.7	12.5	12.5	12.4	12.4	12	11.7	11	9.7	6.5
1	6.4	6.4	6.4	6.4	6.4	6.2	6.2	6	5.7	4.4

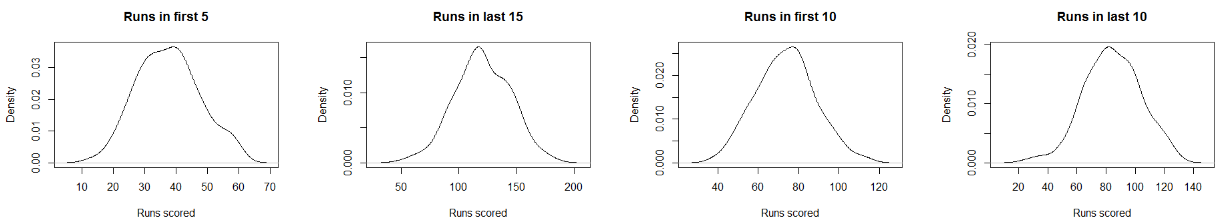
1.4 - Bhattacharya, Gill and Swartz (2011) - Non parametric resource table

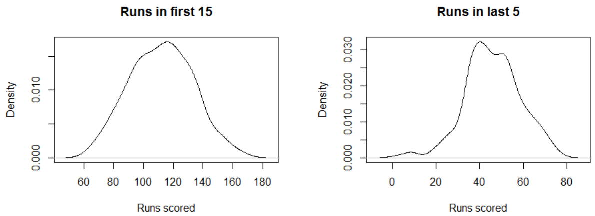
Resources Remaining	Wickets Lost									
Overs Remaining	0	1	2	3	4	5	6	7	8	9
20	100	96.9	93	87.9	81.3	72.2	59.9	44.8	29.7	17.6
19	95.6	90.9	87.7	83	76.9	68.3	56.5	42	27.2	15.3
18	91.7	86.7	82.9	78.7	73.2	65.4	54.2	40.2	25.7	13.9
17	87.7	82.3	78.9	73.8	69.7	62.8	52.2	38.7	24.6	12.8
16	83.5	78.2	75.3	70.5	66.4	60.2	50.3	37.4	23.5	12
15	79.2	74.3	70.9	66.9	62.6	57.4	48.4	36.2	22.7	11.2
14	75.1	70.7	67.3	63.7	59.3	54.6	46.4	35	21.8	10.5
13	71.5	67.4	63.6	60.3	56.2	51.5	44.3	33.8	21	9.8
12	68.3	63.7	60.2	56.8	52.9	47.5	41.9	32.6	20.2	9.1
11	65	59.9	56.6	53.3	49.7	43.9	39.3	31.3	19.4	8.5
10	61.3	56	52.6	50.1	46	40.8	36.1	30	18.6	7.9
9	57.9	52.3	47.9	46.1	42.5	37.8	33.1	28.3	17.7	7.2
8	54	48.3	44.3	41.7	38.9	34.9	30.2	26.1	16.7	6.6
7	49.3	44.2	40.2	37.4	35.4	32.1	27.2	23.4	15.7	5.9
6	41.7	38.5	35.7	33	31.7	29	24.2	20	14.5	5.2
5	36.2	33.4	31	28.6	27.3	25.5	21.5	17	12.2	4.4
4	30.8	28	26.1	24.1	22.4	20.7	18.3	14.2	10	3.5
3	25.4	22.8	21.1	19.4	17.7	16.5	14.4	11.6	7.9	2.5
2	19.7	17.2	15.5	14.1	12.7	11.9	10.6	9.3	6.2	1.6
1	13.7	11.3	9.7	8.5	7.3	6.7	6	5.2	4.2	0.9

Appendix 2 - Density plots of runs scored before and after cutoffs

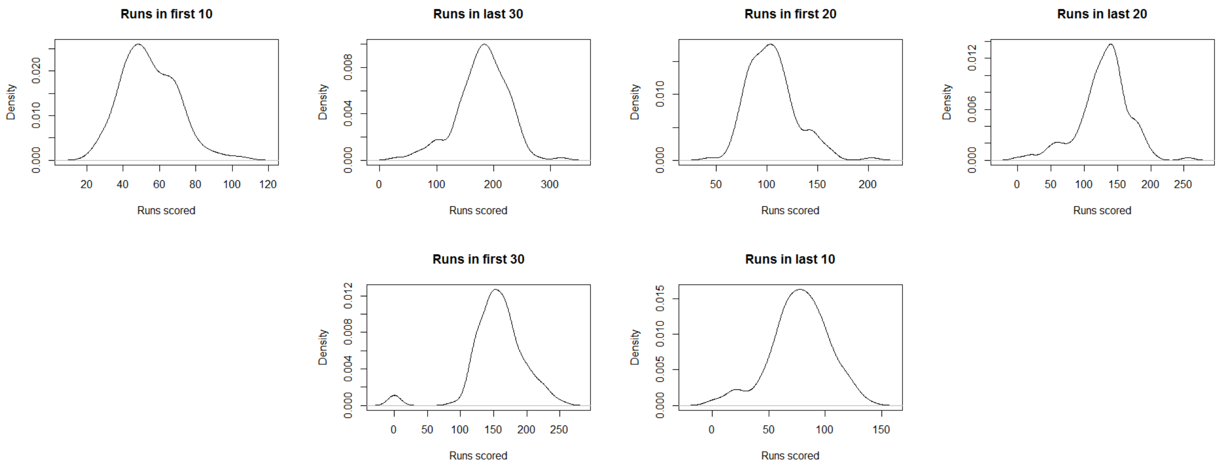
Density plots of the number of runs scored before and after various cutoff in T20 and Pro40 cricket. These are examined to verify the correlation test assumption that both samples are approximately normally distributed.

2.1a - T20 grouped density plots

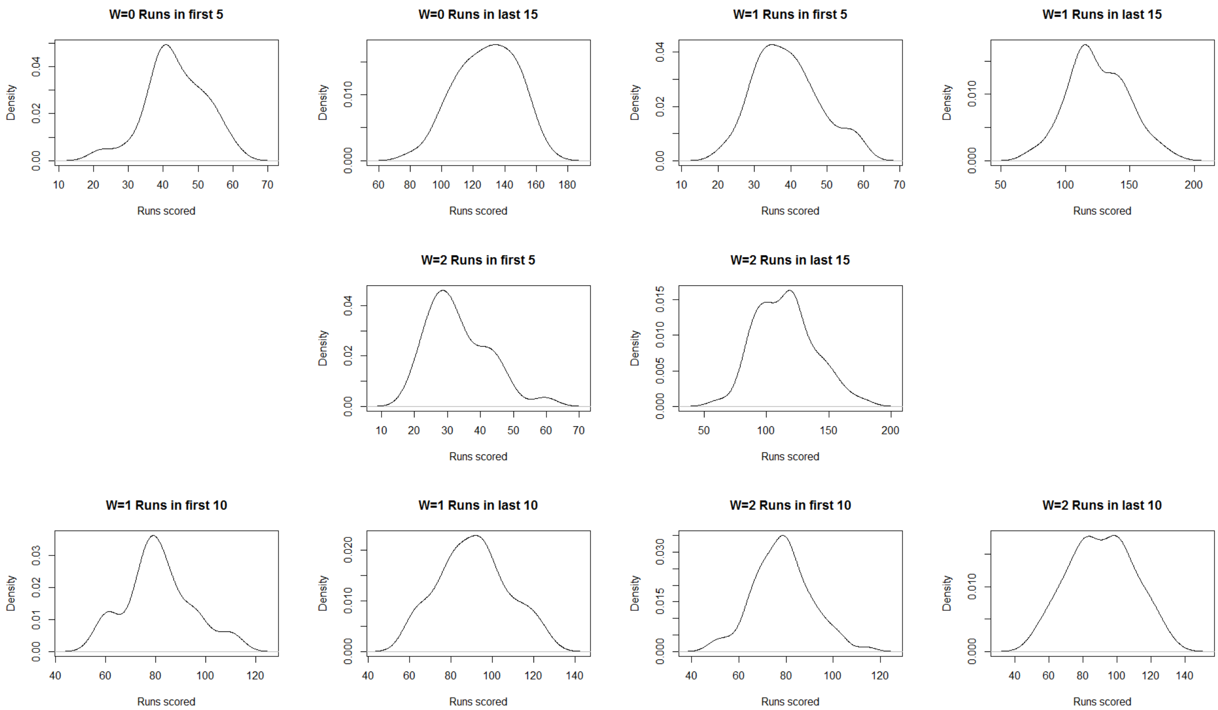


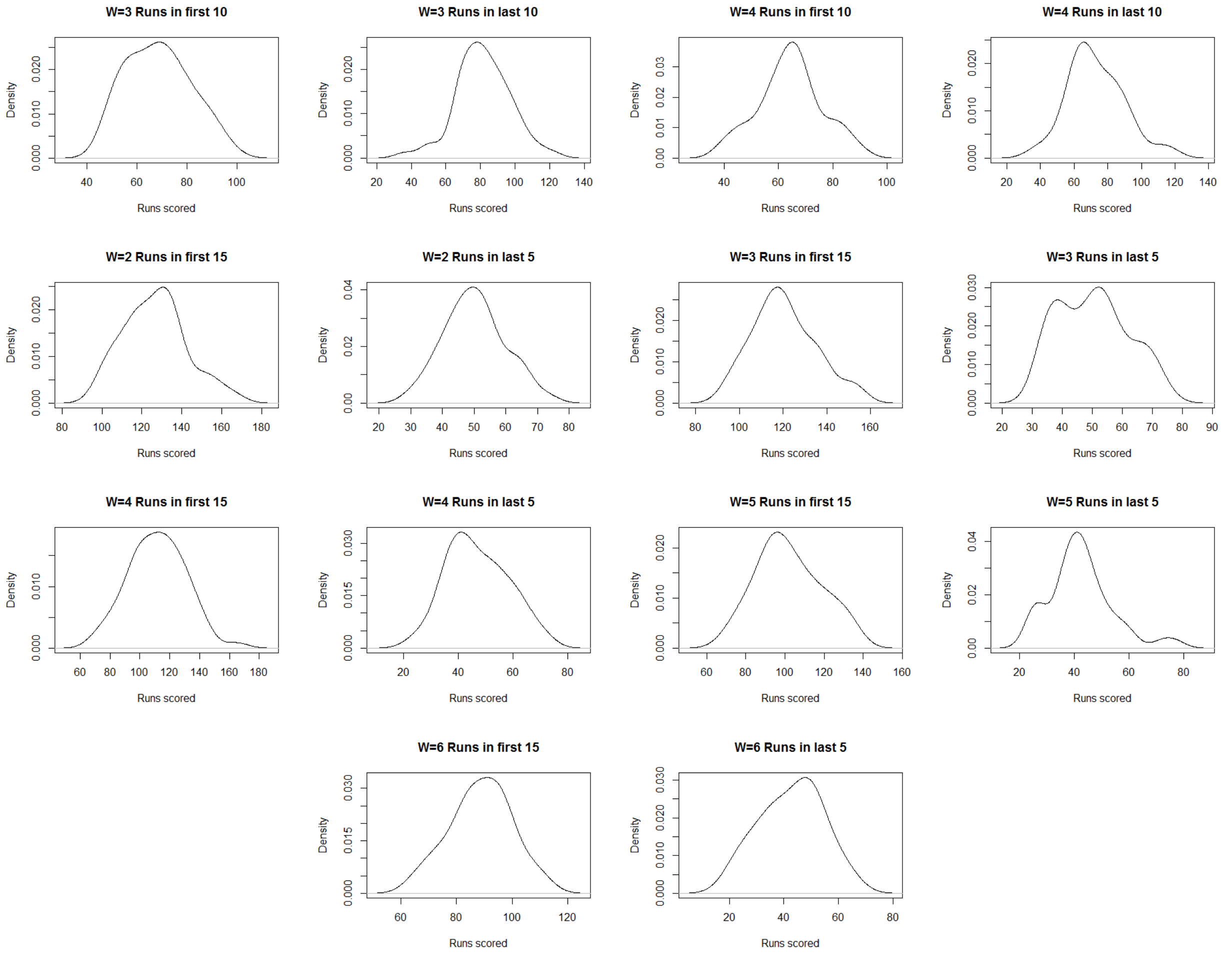


2.1b - Pro40 grouped density plots

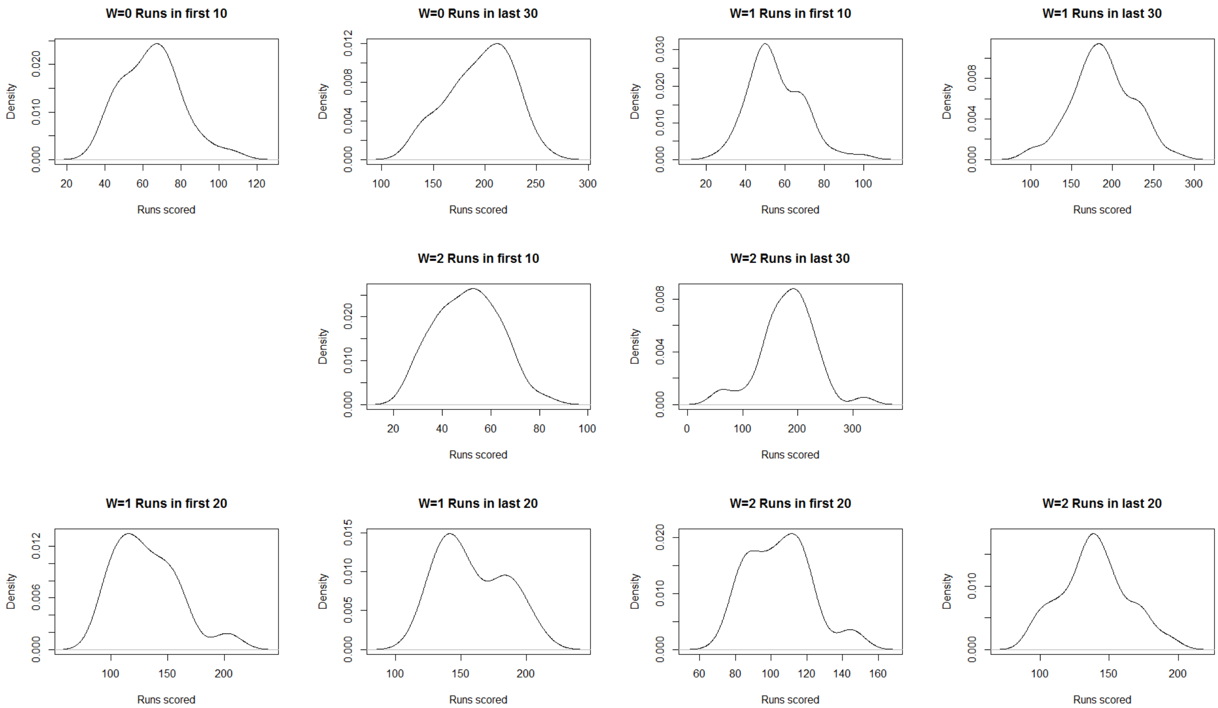


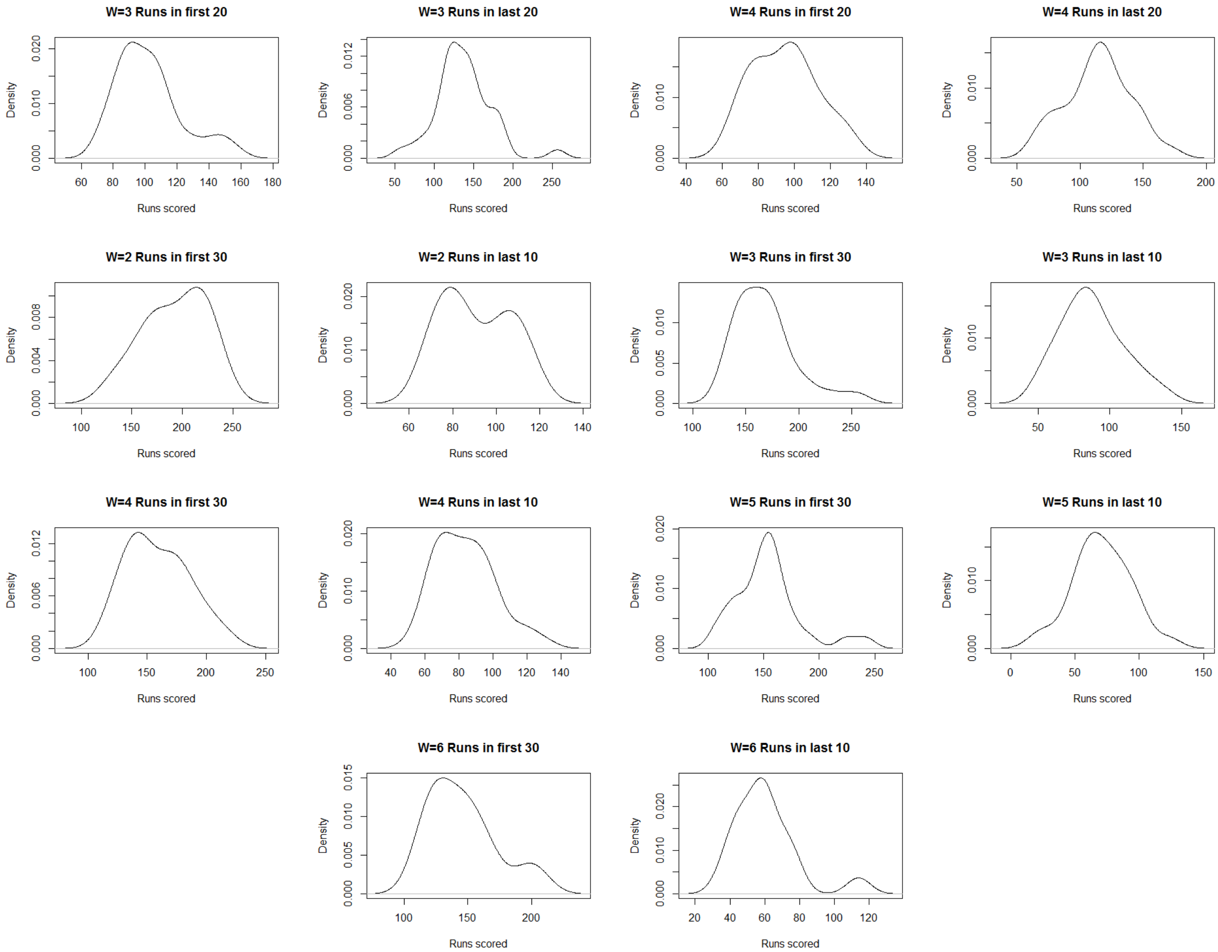
2.2a - T20 sorted density plots





2.2b - Pro40 sorted density plots




W=3 Runs in first 20
Density
0.000 0.010 0.020
60 80 100 120 140 160 180
Runs scored
W=3 Runs in last 20
Density
0.000 0.006 0.012
50 100 150 200 250
Runs scored
W=4 Runs in first 20
Density
0.000 0.010
40 60 80 100 120 140
Runs scored
W=4 Runs in last 20
Density
0.000 0.010
50 100 150 200
Runs scored
W=2 Runs in first 30
Density
0.000 0.004 0.008
100 150 200 250
Runs scored
W=2 Runs in last 10
Density
0.000 0.010 0.020
60 80 100 120 140
Runs scored
W=3 Runs in first 30
Density
0.000 0.005 0.010
100 150 200 250
Runs scored
W=3 Runs in last 10
Density
0.000 0.010
50 100 150
Runs scored
W=4 Runs in first 30
Density
0.000 0.006 0.012
100 150 200 250
Runs scored
W=4 Runs in last 10
Density
0.000 0.010 0.020
40 60 80 100 120 140
Runs scored
W=5 Runs in first 30
Density
0.000 0.010 0.020
100 150 200 250
Runs scored
W=5 Runs in last 10
Density
0.000 0.010
0 50 100 150
Runs scored
W=6 Runs in first 30
Density
0.000 0.005 0.010 0.015
100 150 200
Runs scored
W=6 Runs in last 10
Density
0.000 0.010 0.020
20 40 60 80 100 120
Runs scored

Appendix 3 - The Observed Data set

Observed average scores with variable overs remaining and wickets lost. For T20 and Pro40 first and Second innings'. Sample sizes over which this values are averaged are in brackets.

3.1 - T20 First Innings Data

	Wickets Lost									
Overs Remaining	0	1	2	3	4	5	6	7	8	9
20	158.4/242									
19	152.9/200	151.4/40	110.5/2							
18	149.2/156	139.4/76	129.3/10							
17	143.0/112	134.4/104	126.6/25	64/2						
16	135.3/82	131.7/106	119.1/46	90.3/7	62/1					
15	128.3/54	125.3/105	115/62	100.4/18	73/1	60.5/2				
14	119.3/39	117.1/96	109.4/65	102.6/33	77.7/3	58/3				
13	116.9/25	107.7/90	108.9/76	94.5/41	86/7	54.3/3				
12	112.8/19	100.8/79	103.6/75	90.6/48	81.4/17	60/3				
11	105.3/13	96.9/62	96.6/77	85.2/61	77.6/23	59.6/5	34/1			
10	98.6/9	90.9/49	90.7/77	81.6/60	74/37	59.8/8	59/2			
9	90.2/6	85.0/37	81.6/72	78.4/63	68.9/48	52.8/11	48.4/5			
8	76.7/3	78.2/32	74.3/36	72.2/62	65.8/49	56.4/24	41.1/9			
7	69/1	69/23	67.1/59	65.8/58	59.3/55	52.5/29	47.4/15	42/2		
6		58.1/15	60.6/51	57.5/62	53.5/60	48.4/29	45.3/18	25.8/6	40/1	
5		50.9/9	50.2/39	50.3/56	47.1/64	42.4/44	42.7/18	28.6/7	20.6/5	
4		37.5/6	42.0/32	41.2/40	39.8/68	36.4/56	35.8/19	28.2/13	23.8/6	1.5/2
3		30.8/4	29.4/25	33.6/29	31.1/61	28.6/58	26.9/34	26.2/17	19.7/10	17.5/2
2		20/2	20.6/18	23.7/25	21.7/50	20.7/57	18.5/47	19.3/24	16.2/12	13/3
1		10.5/2	9.9/12	12.3/23	12.0/31	10.0/58	11.2/42	10.4/38	8.8/24	5.7/7

3.2 - T20 Second Innings Kaplan Meier adjusted Data

	Wickets Lost									
Overs Remaining		1	2	3	4	5	6	7	8	9
20	156.6/243									
19	149.9/199	144.2/45								
18	144.1/61	147.9/69	117.1/13							
17	134.8/128	139.7/8	127.7/31	98/4						
16	126.1/94	127.9/100	130.1/39	114.9/7	63/4					
15	118.7/75	116.8/89	121.9/63	113.2/13	56.7/3					
14	113.7/49	116̇/77	111.4/83	106.6/28	80.7/3	65.7/4				
13	109.9/35	105.5/69	101.1/79	109.2/47	77.4/9	56.7/4	49/1			
12	15.8/28	97.4/6	94.8/81	104/49	77.7/21	43/3	58.5/2			
11	94.5/17	93/47	88.7/85	92.9/59	77.8/28	55.4/5	49/2			
10	87.3/11	86/43	83.6/76	83.5/61	78.1/34	68/12	35.2/5			
9	81/8	77.4/33	76/69	77.3/62	71.4/41	66.8/21	32.5/4	24/3		
8	74/7	67.8/24	69.5/54	68.5/68	65.7/50	69.4/19	41.8/13	27/2	27̇/3	
7	59/5	52/15	62.7/48	62/68	58.4/51	58.9/29	47/11	37	26.3/3	4.5/2
6	51/2	44.3/14	57.4/36	53.5/63	51.1/56	50.8/32	46.3/15	36.7/9	23.8/4	4.7/3
5		38.5/9	47.8/27	45.3/49	43.8/59	44.1/41	38.5/14	31.4/19	21.3/4	5.9/8
4		41/6	39.2/21	36.6/39	36/47	33.5/54	31.6/23	28/19	16.4/5	14
3		26/3	29.8/15	28.9/23	26/47	28.1/45	26.2/31	19.9/31	11.8/8	6.6/7
2			18.5/12	17.1/17	16.6/37	18.1/4	18.9/37	17/27	11̇/12	6.5/14
1			7.5/5	17̇/8	7.3/19	7.3/4	8.1/3	9.3/32	6.4/2	5.9/8

3.3 - Pro40 First Innings Data

	Wickets Lost									
Overs remaining	0	1	2	3	4	5	6	7	8	9
40	234.544/149									
39	230.37/135	233.692/13	138/1							
38	228.88/117	215.258/31	131/1							
37	225.313/96	215.574/47	189.4/5	130/1						
36	218.471/87	217.36/50	177.636/11	129/1						
35	218.93/71	204.583/60	198.467/15	90/2	124/1					
34	216.985/65	197.213/61	204/14	130/7	137.5/2					
33	213.389/54	192.333/60	193.577/26	134.6/5	120.75/4					
32	205.311/45	195.833/54	181.6/35	155.909/11	117.5/4					
31	201.658/38	190.8/55	184/38	147.917/12	111.833/6					
30	195.563/32	187.643/56	181.923/39	143.571/14	120.25/8					
29	192.345/29	185.583/48	179.643/42	154.632/19	101.889/9	91/2				
28	184.923/26	184.395/43	171.894/47	161.684/19	106.636/11	72/3				
27	180.591/22	181.293/41	167.18/50	155.286/21	117.273/11	68/2	18/1	75/1		
26	178.176/17	177.487/39	164.28/50	155.077/26	112.636/11	80/4		44.5/2		
25	172.308/13	176.079/38	160.021/47	149.969/32	114.077/13	77.5/4		42.5/2		
24	164.154/13	174.424/33	155.255/47	149.257/35	110.929/14	72/5		38/2		
23	161.273/11	166.724/29	151.787/47	147.444/36	111.421/19	68.6/5		35/2		
22	156/10	162.741/26	149.614/44	142.368/38	112.667/21	76.8/5	51.5/2	54/1	2/1	
21	150.333/9	160.78/20	145.292/48	138.763/38	117.318/22	72.25/8	70/41	37.5/1	0/1	
20	144/9	158.235/17	139.6/45	136.897/39	114.423/26	72.714/7	54/3	19/1	46/1	0/1
19	134.75/8	152.75/16	137.179/39	130.422/45	108.593/27	94/6	42.8/5	13/1		34/1
18	127.625/8	144.571/14	130.703/37	128.767/43	107.926/27	91.833/12	36.25/4	37/1	26/1	9/1
17	113.75/4	138.25/16	129.813/32	120.8/45	107.25/28	93.846/13	61.2/5	7/1	32/3	1/1
16	105.5/4	125.462/13	126.143/28	118.891/46	102.625/32	89.8/15	66.5/4	15/2	36/2	5/1
15	102.333/3	118.5/12	121/26	113.674/46	100.6/30	89.765/17	65.875/8		30.667/3	3/1
14	96/1	109.25/12	118.217/23	108.826/46	94.71/31	89.85/20	63.125/8	60/41	28.667/3	
13		101.833/12	115.105/19	101.667/45	93.484/31	87.417/24	59.111/9	39/2	26/3	
12		99.778/9	103.55/20	96.795/44	89.813/32	79.75/24	63.3/10	33/3	21.333/3	
11		96.5/8	97.765/17	92.371/35	85.075/40	77.92/25	64.583/12	20.2/45	39/2	9.5/2
10		91.429/7	90.75/16	87.519/27	83.628/43	70.593/27	60.313/16	44.333/4	26/4	2/2
9		82.2/5	84.438/16	83.19/21	76.909/44	70/30	54.944/18	47.667/3	29.8/5	10.5/2
8		68.75/4	76.923/13	74.087/23	73.317/41	66.741/27	51.238/21	44.25/8	30.25/4	10/2
7		59.667/3	70.222/9	65.619/21	67.85/40	60.688/32	51.056/18	34.308/13	33.25/4	9/3
6		44/2	57.444/9	59.118/17	58.763/38	55.545/33	45.632/19	34.688/16	29.5/4	3.75/4
5		34/1	47.857/73	51.364/11	49.303/33	50.667/36	37.909/22	34.2/20	38.5/6	3.5/4
4			40.4/5	41.923/13	42.682/22	40.559/34	34.034/29	28.208/24	30.625/8	8/2
3			31.333/3	30.4/10	30.667/21	32.281/32	27.609/23	23.296/27	23.867/15	12.4/5
2			20/2	22/8	24.077/13	20.742/31	19.704/27	15.37/27	17.053/19	14/6
1				12.5/8	10.143/7	10.923/26	10/24	10.042/24	9/29	6.923/13

3.4 - Pro40 Second Innings Kaplan Meier Data

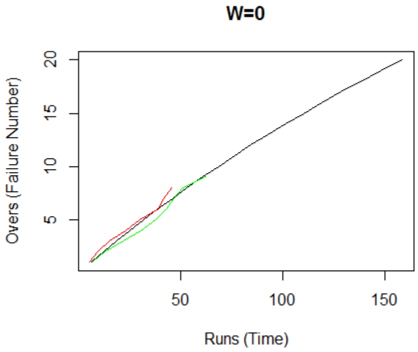
	Wickets Lost									
Overs Remaining		1	2	3	4	5	6	7	8	9
40	227.115/14									
39	229.059/122	230.629/16	191.5/2							
38	227.503/16	200.199/28	212.944/6							
37	229.491/85	195.147/43	193.96/12							
36	222.867/77	198.407/42	181.995/18	139.333/3						
35	208.025/66	176.581/44	180.541/23	152.333/6	98/1					
34	216.028/57	183.946/47	189.159/25	169.75/8	124.333/4					
33	208.797/54	180.486/44	174.719/29	175.222/1	117.333/3					
32	203.54/49	175.628/43	171.218/3	172.727/12	132.833/6					
31	204.686/39	174.286/46	167.535/31	154.006/16	144.208/8					
3	200.776/33	173.848/47	163.225/35	148.714/15	150.25/8	149.5/2				
29	206.047/27	172.1/46	155.374/39	148.505/15	140.709/11	146.5/2				
28	199.748/24	167.597/38	154.187/46	149.37/18	143/1	127.25/4				
27	163.485/21	174.422/39	149.34/43	147.48/22	141.956/9	113.5/6				
26	160.886/18	170.841/35	147.684/45	137.402/25	134.056/9	113.75/8				
25	154.857/14	165.803/37	143.297/4	133.458/29	125.875/11	118.156/8	97/1			
24	150.167/13	160.809/34	141.082/41	135.361/26	115.952/17	112.656/8	93/1			
23	138.167/9	147.481/32	140.529/43	129.418/24	108.621/20	119.704/9	80.5/2			
22	135.5/7	140.055/31	136.492/4	123.46/26	110.169/19	113.333/12	92.333/3			
21	127/7	133.296/25	130.745/4	126.299/29	105.209/17	101.571/17	95/2	79/1		
20	124/7	131.578/23	130.438/36	117.07/30	112.777/19	90.5/17	87.6/5	76/1		
19	121/6	128.325/22	126.303/33	120.414/28	107.396/21	85.947/2	78/6	61/2		
18		110.481/18	124.738/33	115.04/25	103.481/26	79.493/2	65.286/7	77.667/3		
17		105.118/18	124.805/28	106.079/28	102.994/27	72.277/17	66.75/8	72/6		
16		99.938/15	110.984/25	105.095/28	100.713/27	74.696/16	53.417/12	72.571/7		
15		91/13	104.802/2	98.26/30	109.772/27	75.112/17	47/12	63.1/10	38/2	
14		82/11	97.703/18	93.843/30	99.854/28	69.675/15	59.135/13	58.727/11	38.25/4	8/1
13			91.846/15	90.268/30	94.698/27	62.571/13	59.767/15	51.583/12	45/5	22/2
12			85.671/16	85.723/24	83.76/26	66.554/16	60.076/15	50.692/13	34.5/6	22/3
11			76.257/12	80.035/25	79.028/24	61.362/18	66.088/13	43.413/15	32/7	13/3
10			65.857/9	75.003/22	72.747/23	55.571/20	67.248/14	43.533/15	26.167/6	14/5
9			60.75/1	71.197/20	63.031/19	64.516/19	54.679/15	44.337/13	25.273/11	12/7
8			53.063/13	62.488/14	64.021/19	57.231/21	54.741/12	37.09/13	33.308/13	10/6
7			38/7	50.289/17	57.058/19	49.587/21	39.283/12	36.583/12	30.867/14	10.2/5
6			28/5	43.768/14	52.571/17	43.971/18	35.762/16	35.767/10	24.75/12	15.542/12
5				32.643/9	41.479/16	40.403/17	31.163/16	27.537/9	27.182/11	14.923/13
4				25/4	31.396/18	34.163/17	26.6/12	25/10	23.531/13	15.35/10
3					22.778/10	26.68/16	19.148/13	16.524/8	19.435/14	12.068/13
2					19/4	14.073/12	18.212/16	12.686/8	9.375/8	7.941/17
1					11/2	5.5/5	6.25/15	5.571/10	7.4/5	4.308/13

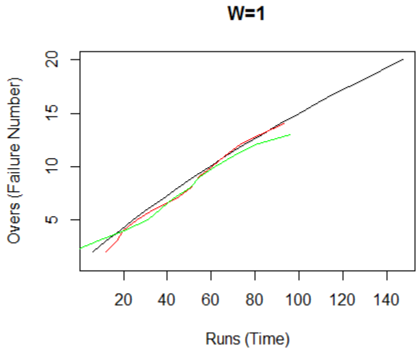
Appendix 4 - Repairable system graphs

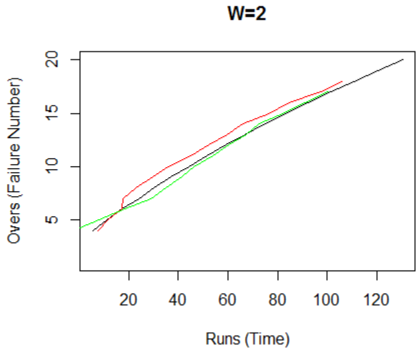
4.1 - Testing to see if the ROCOF was constant

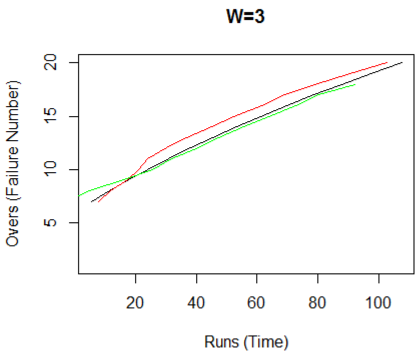
Graphs of runs vs overs to test to see if the ROCOF was constant or not.

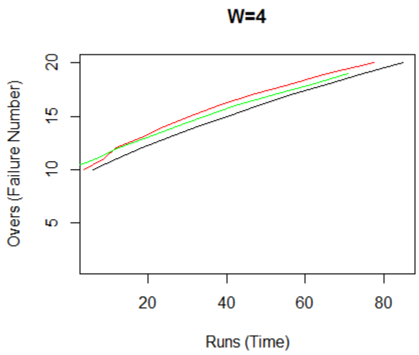
T20 graphs

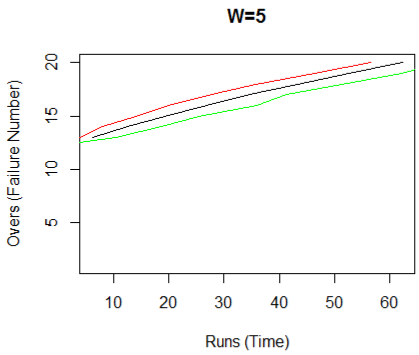


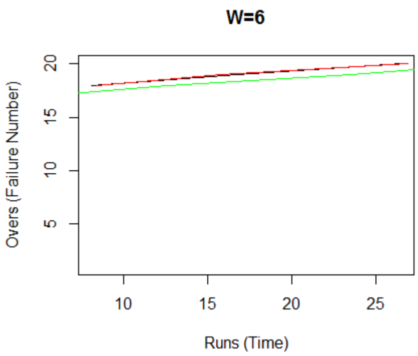


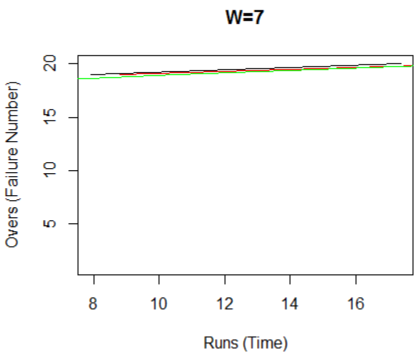




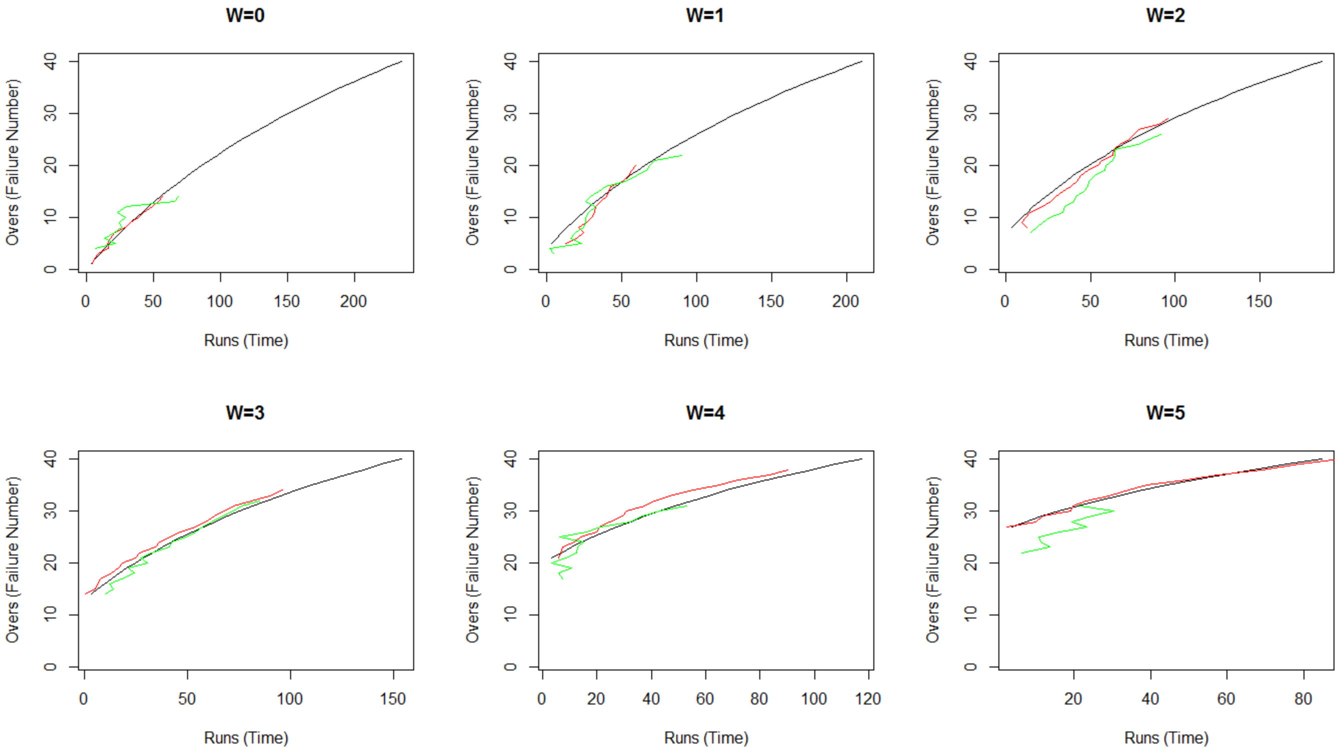








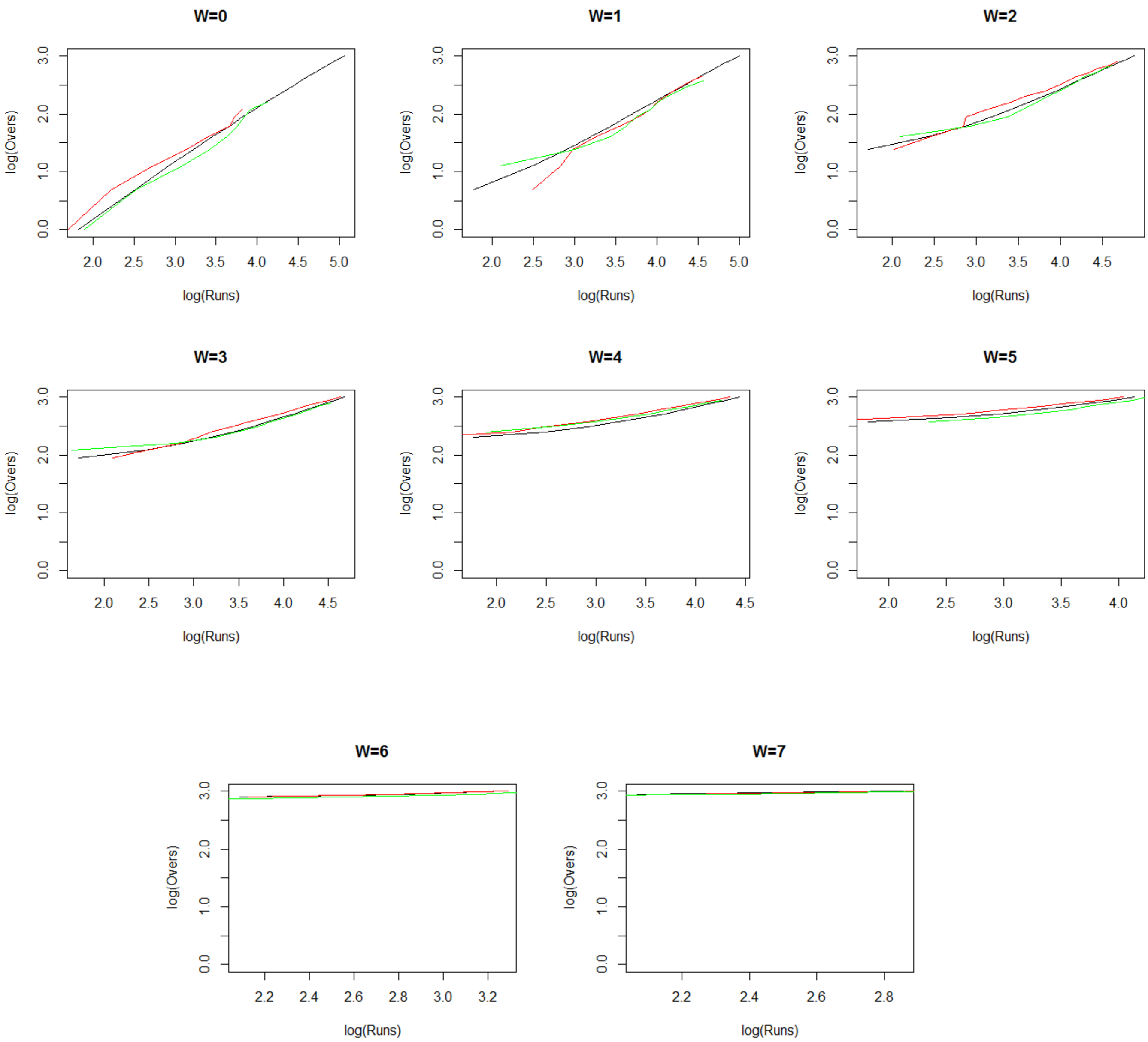
Pro40 graphs


W=0
W=1
W=2
W=3
W=4
W=5
Overs (Failure Number)
Runs (Time)
0
10
20
30
40
50
60
80
100
120
150
200

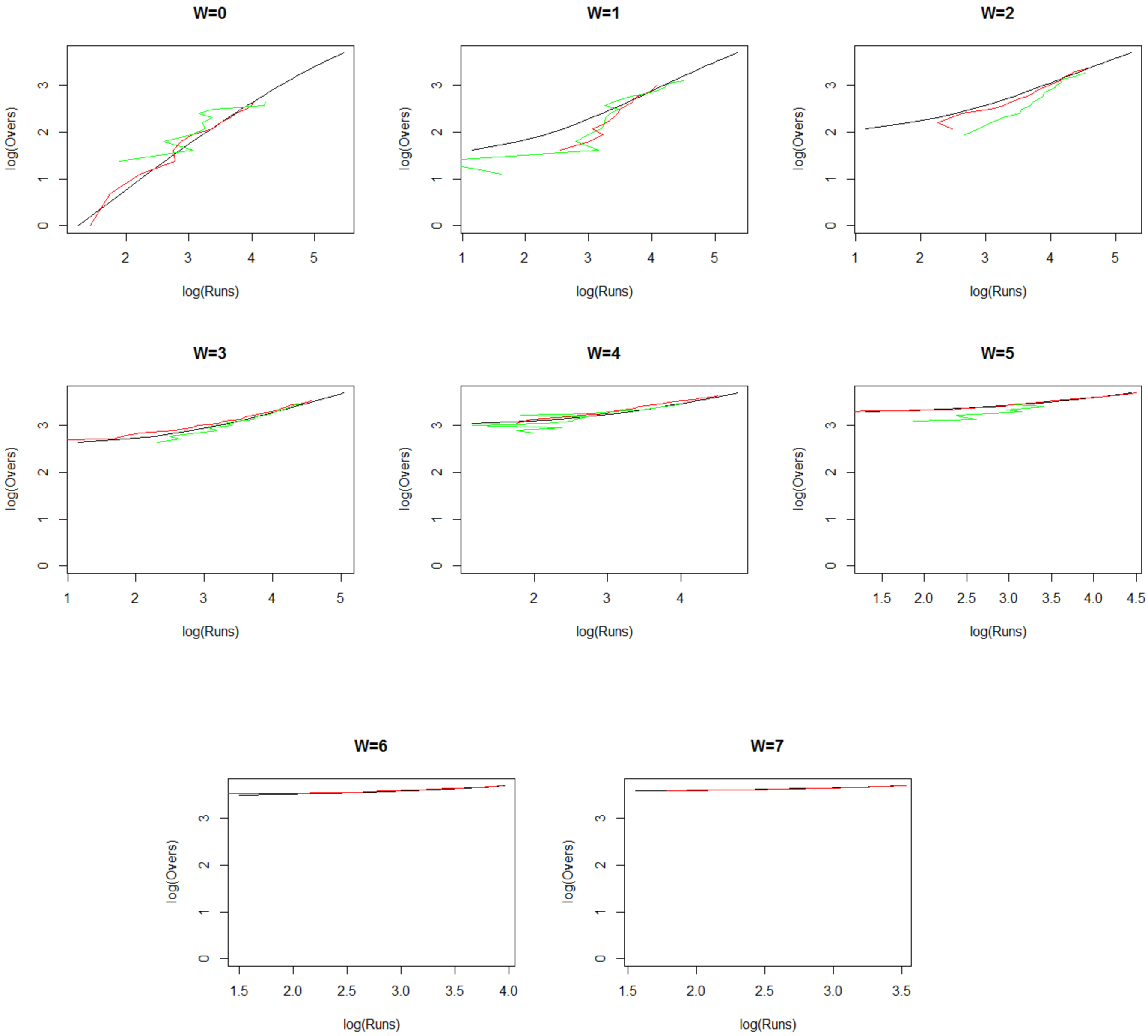
4.2 - Testing the suitability of $v_2(r)$

Graphs of log(runs) vs log(overs) to test to see if $v_2(r)$ was the suitable form for the ROCOF to take.

T20 graphs



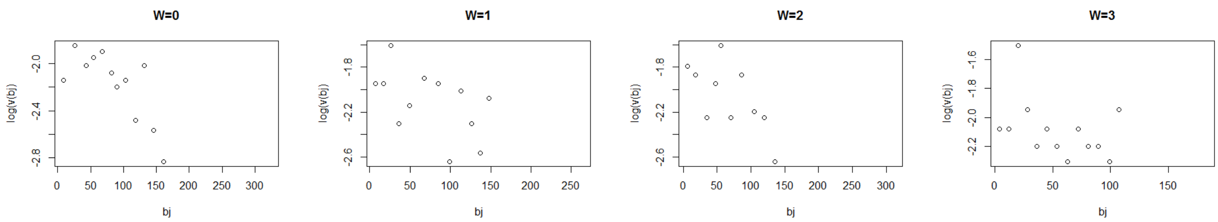
Pro40 graphs

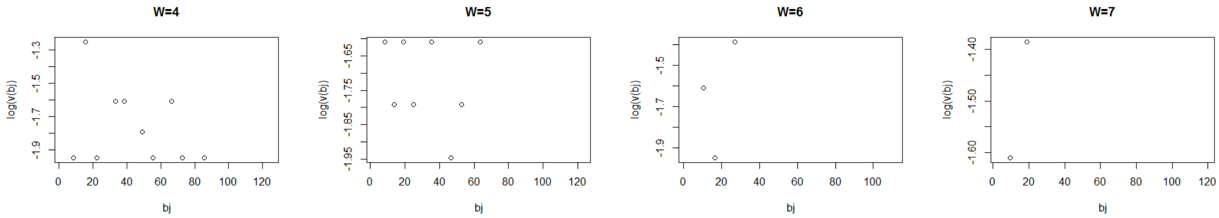


4.3 - Testing the suitability of $v_1(r)$

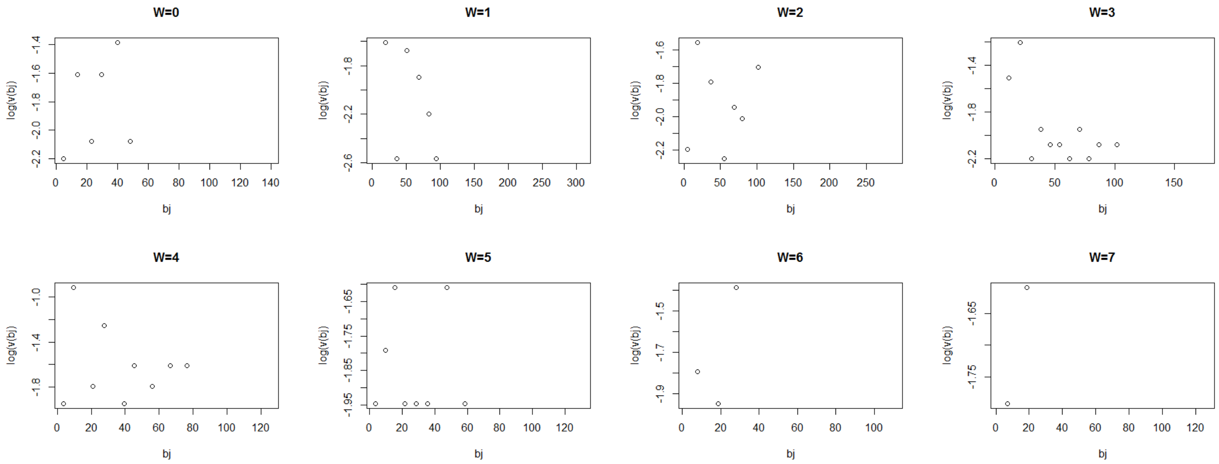
Graphs of b_j vs $log(\tilde{v}(b_j))$ to test to see if $v_1(r)$ was the suitable form for the ROCOF to take.

T20 graphs - DL Resources

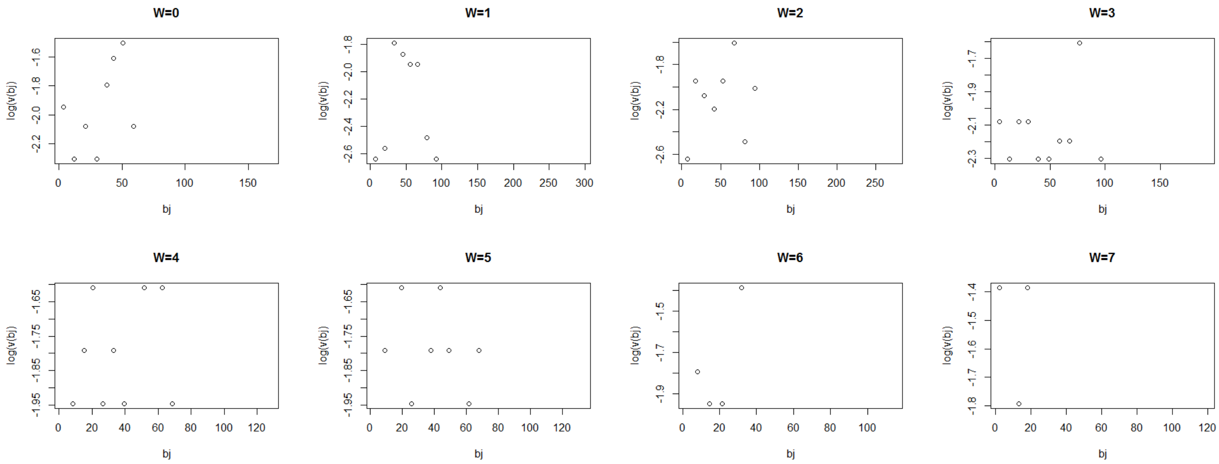




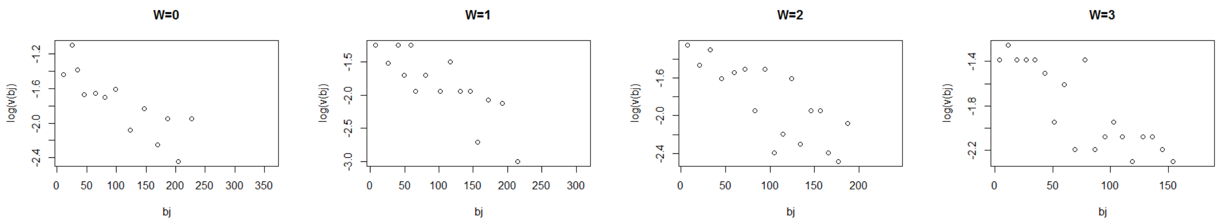
T20 graphs - First Innings

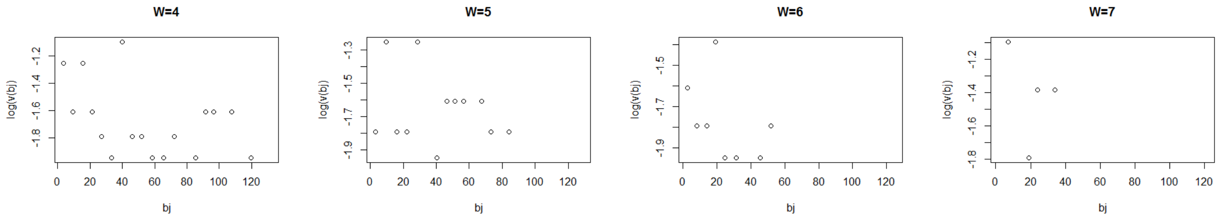


T20 graphs - Second Innings

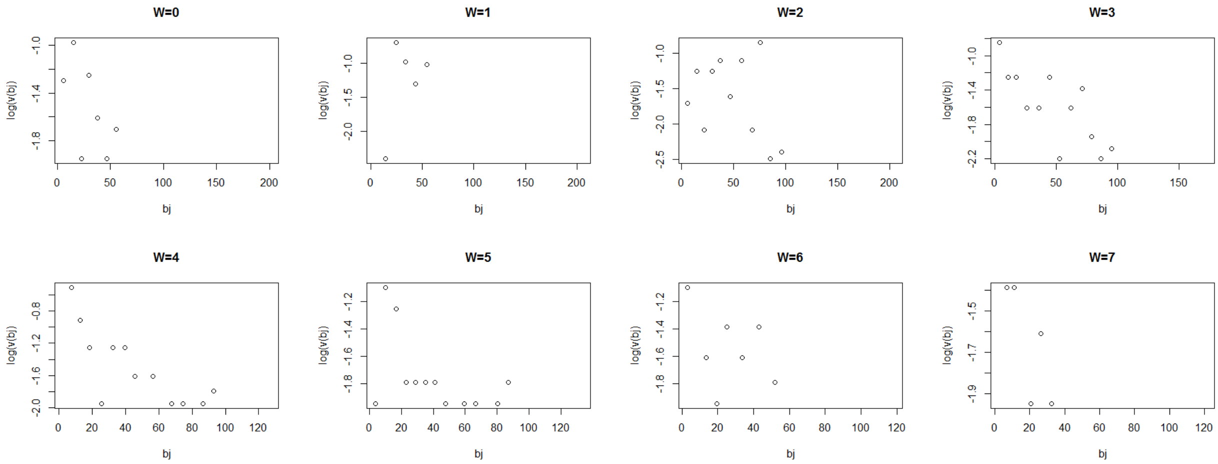


Pro40 graphs - DL Resources

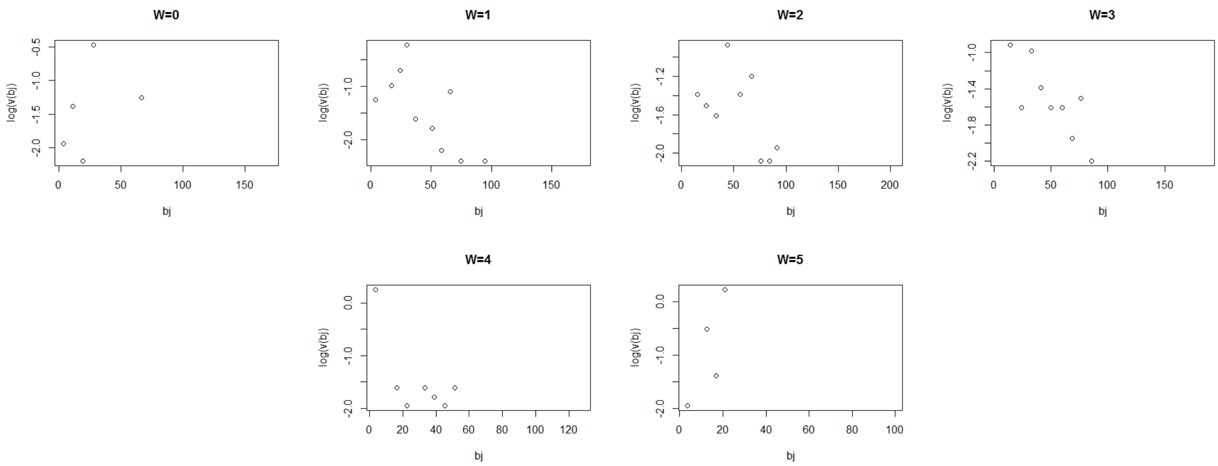




Pro40 graphs - First Innings

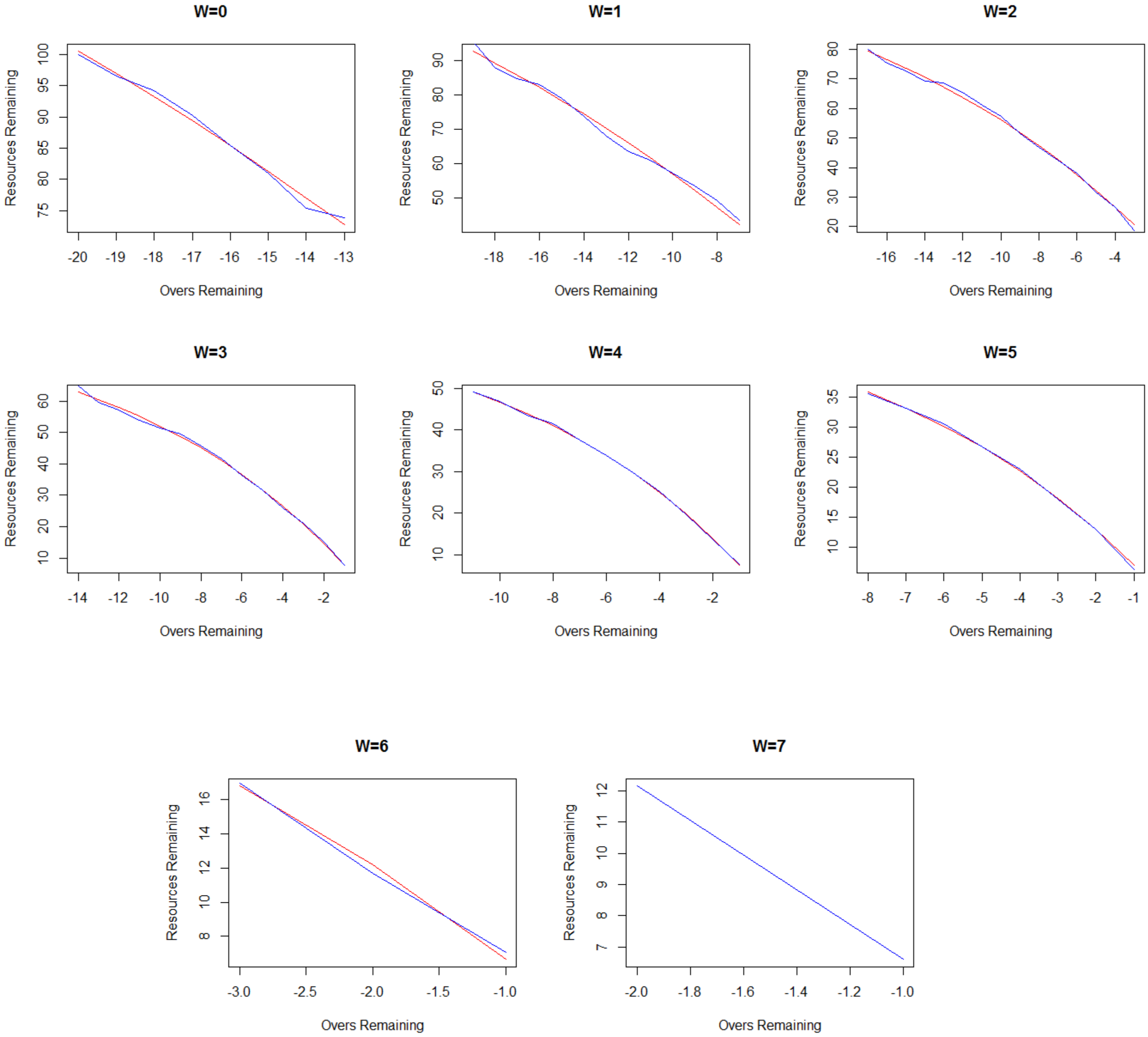


Pro40 graphs - Second Innings

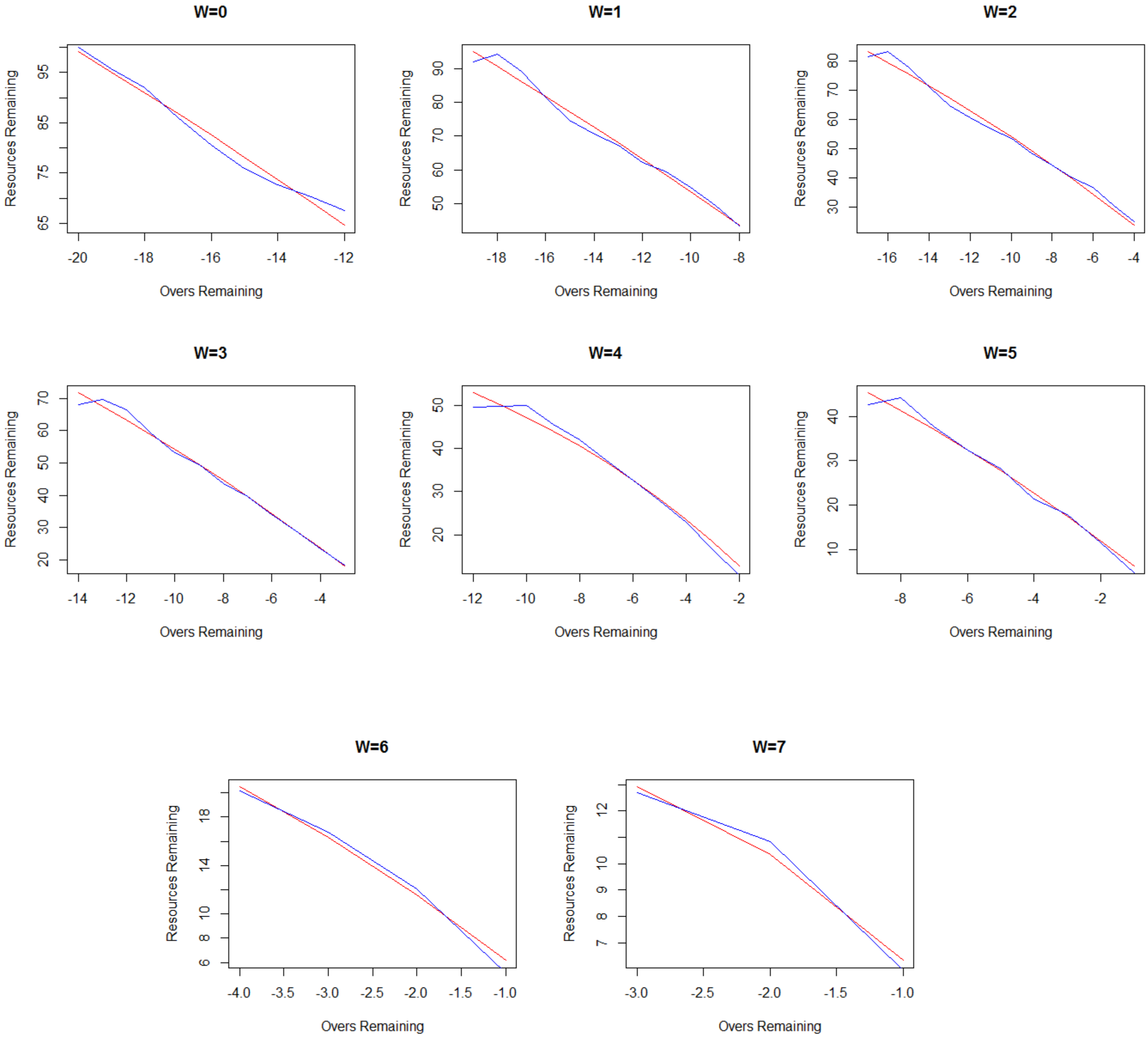


Appendix 5 - Nonlinear Least Squares, Fits and Diagnostics

5.1a - T20 first innings fitted (red) vs observed (blue) values



5.1b - T20 Second innings fitted (red) vs observed (blue) values

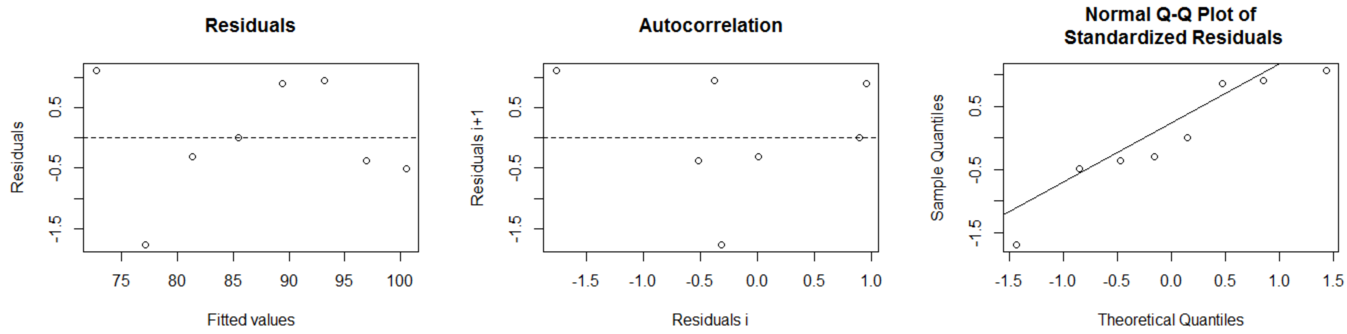


5.2 T20 Diagnostics

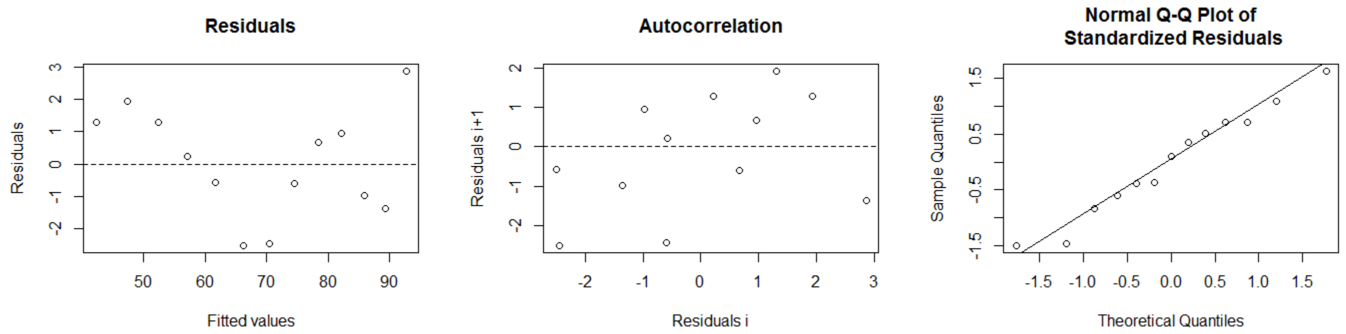
Below are Residual vs Fitted value plots, Autocorrelation plots and Normal QQ plots for the T20 fitted models.

First Innings

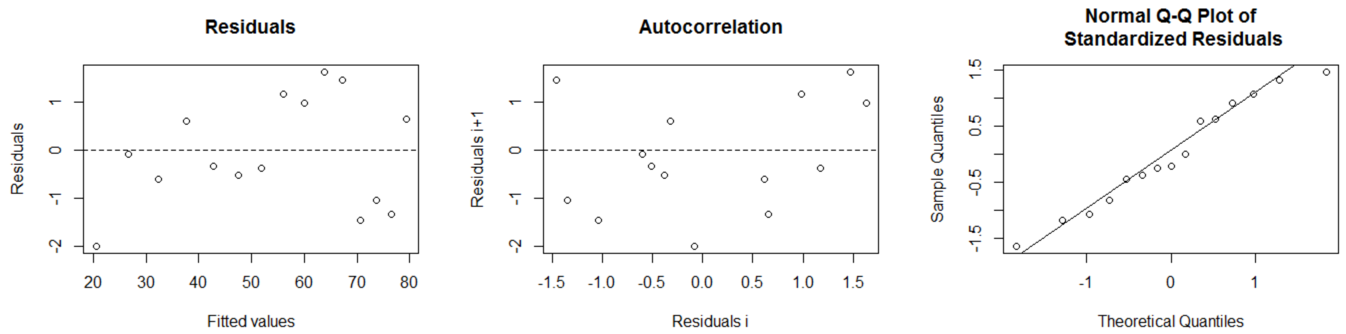
W=0



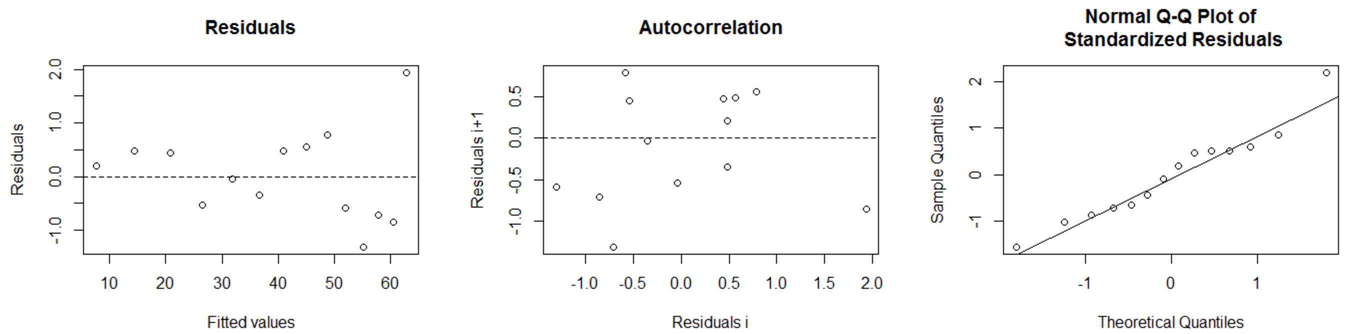
W=1



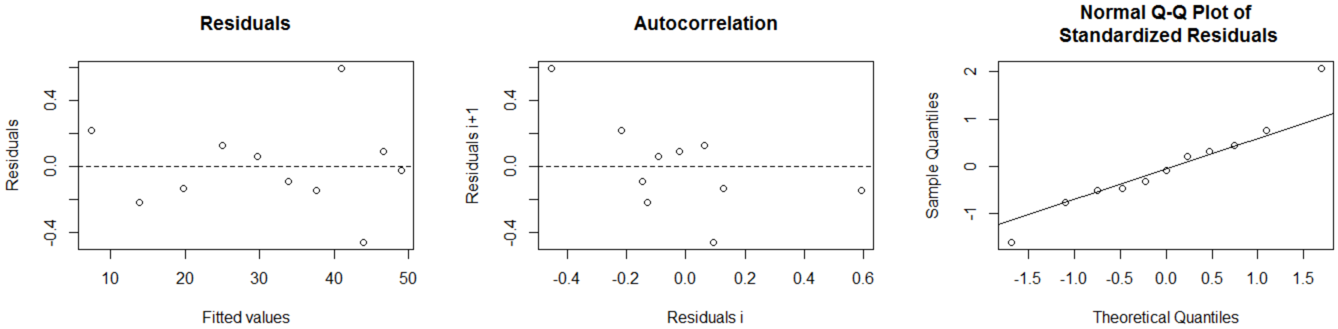
W=2



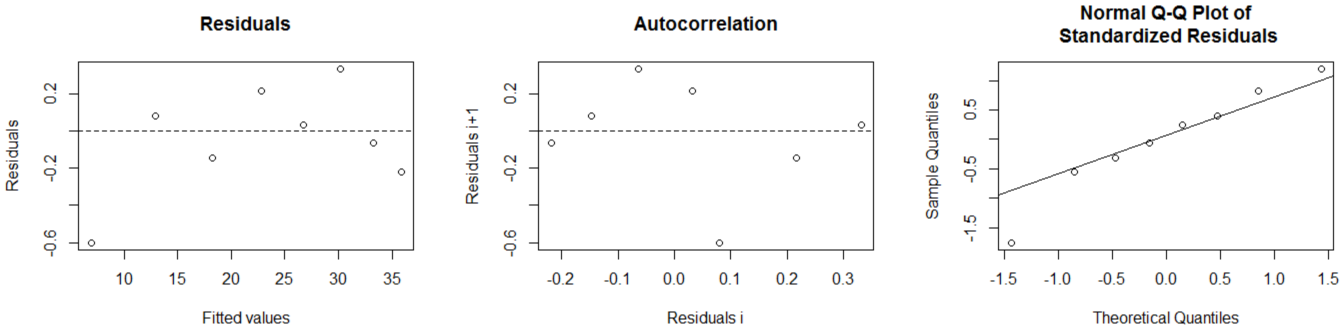
W=3



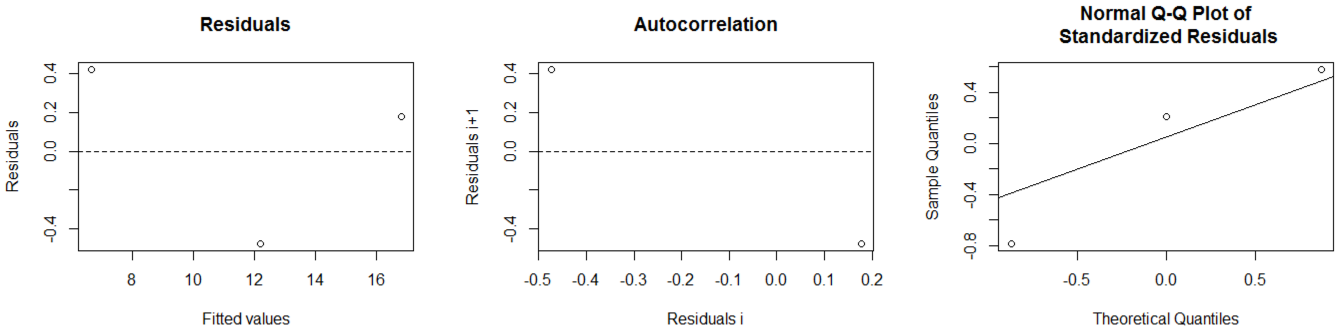
W=4



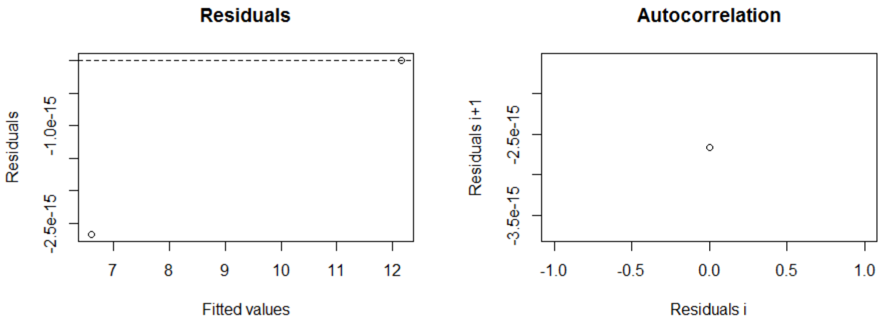
W=5



W=6



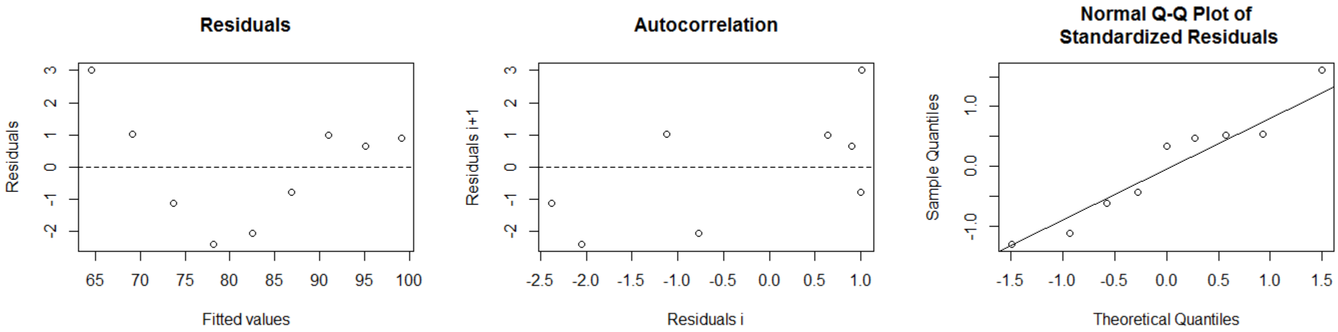
W=7



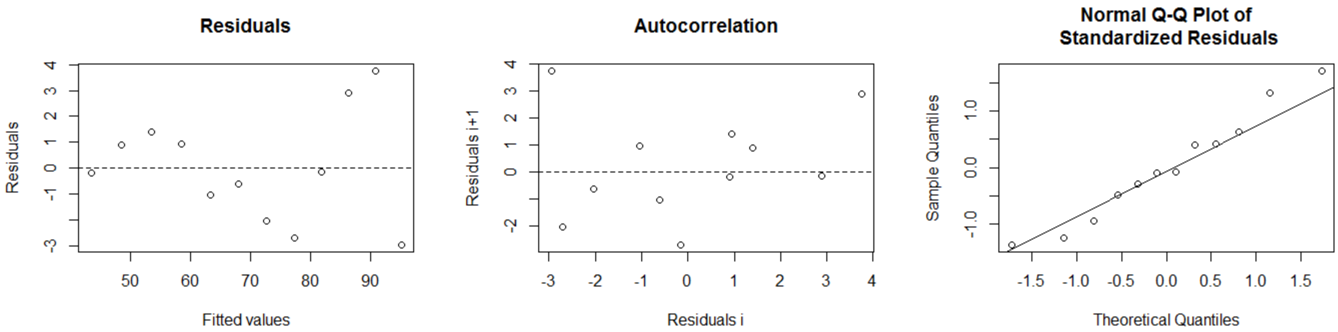
There were not enough data points to produce a Normal QQ plot when W=7

Second Innings

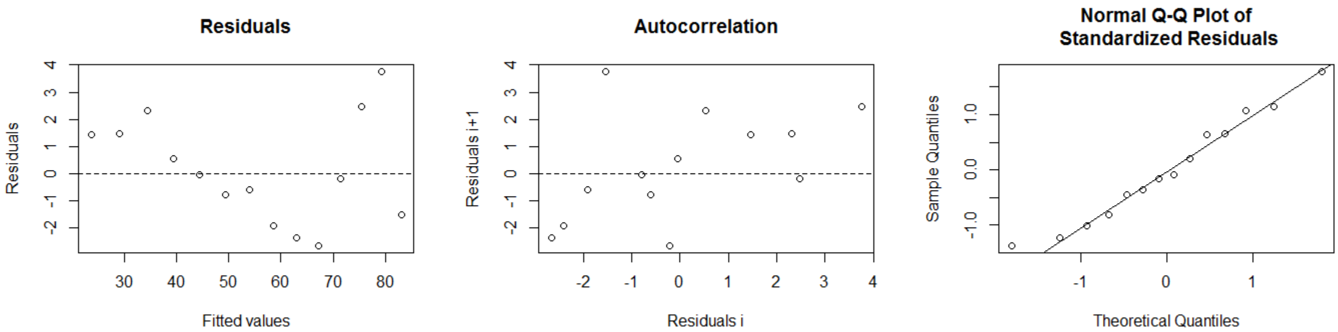
W=0



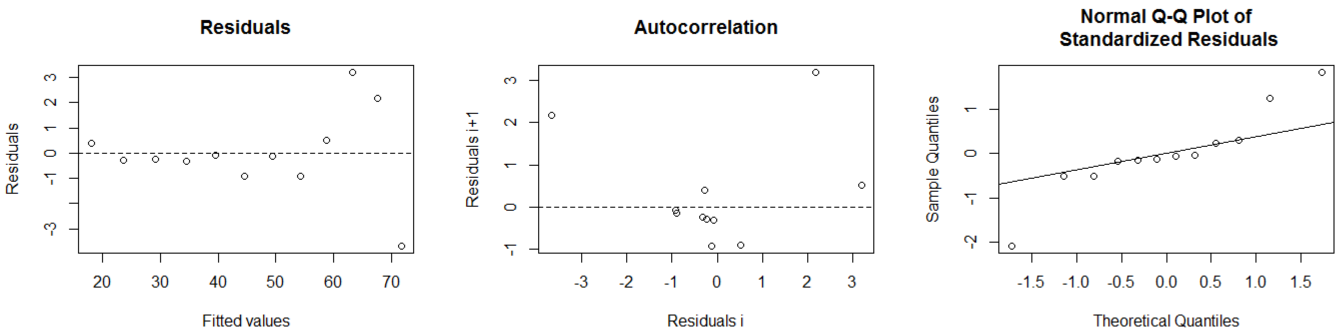
W=1



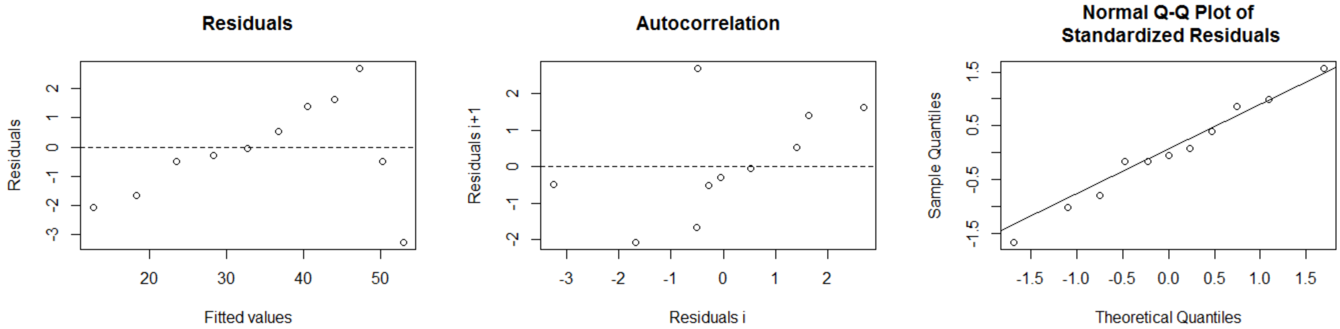
W=2



W=3

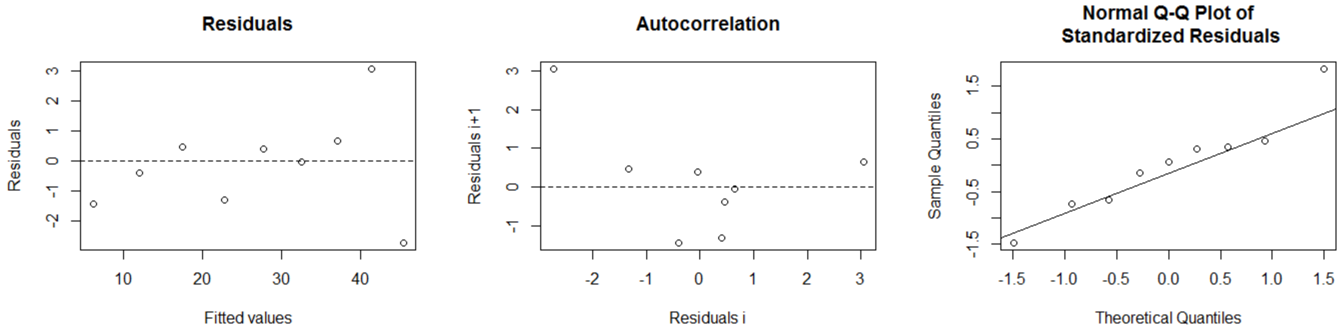


W=4



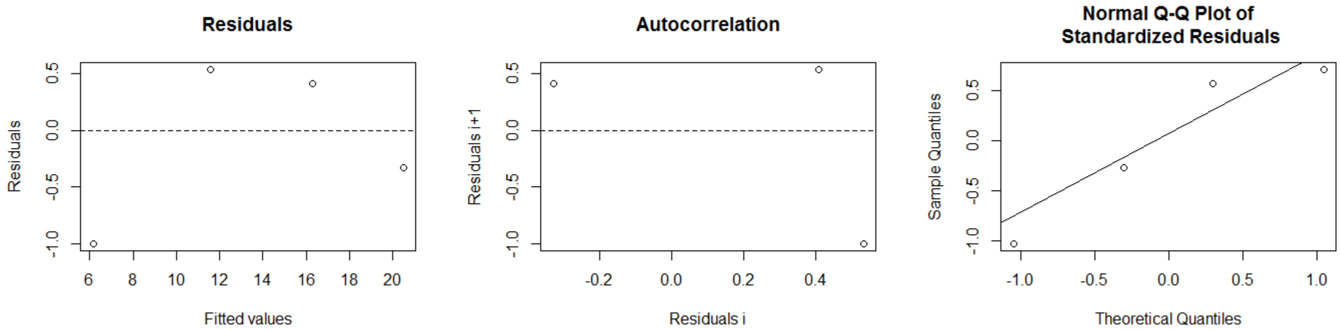
Residuals
Residuals
Fitted values
Autocorrelation
Residuals i+1
Residuals i
Normal Q-Q Plot of
Standardized Residuals
Sample Quantiles
Theoretical Quantiles

W=5



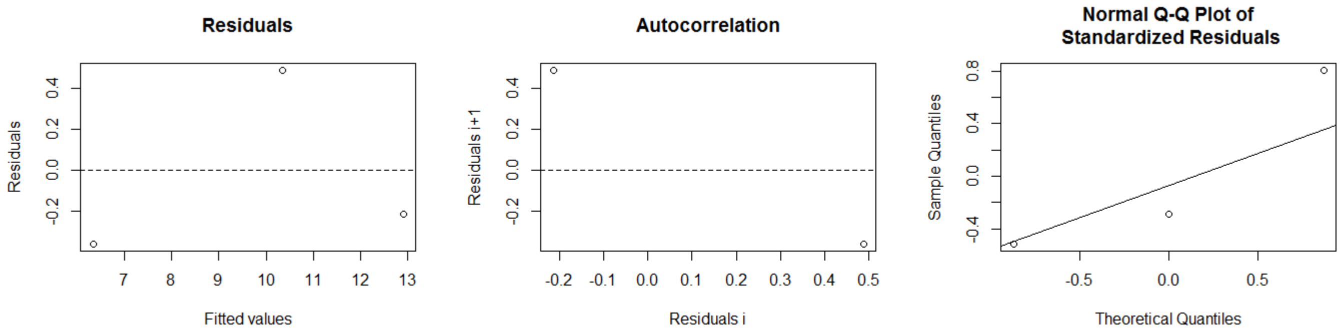
Residuals
Residuals
Fitted values
Autocorrelation
Residuals i+1
Residuals i
Normal Q-Q Plot of
Standardized Residuals
Sample Quantiles
Theoretical Quantiles

W=6



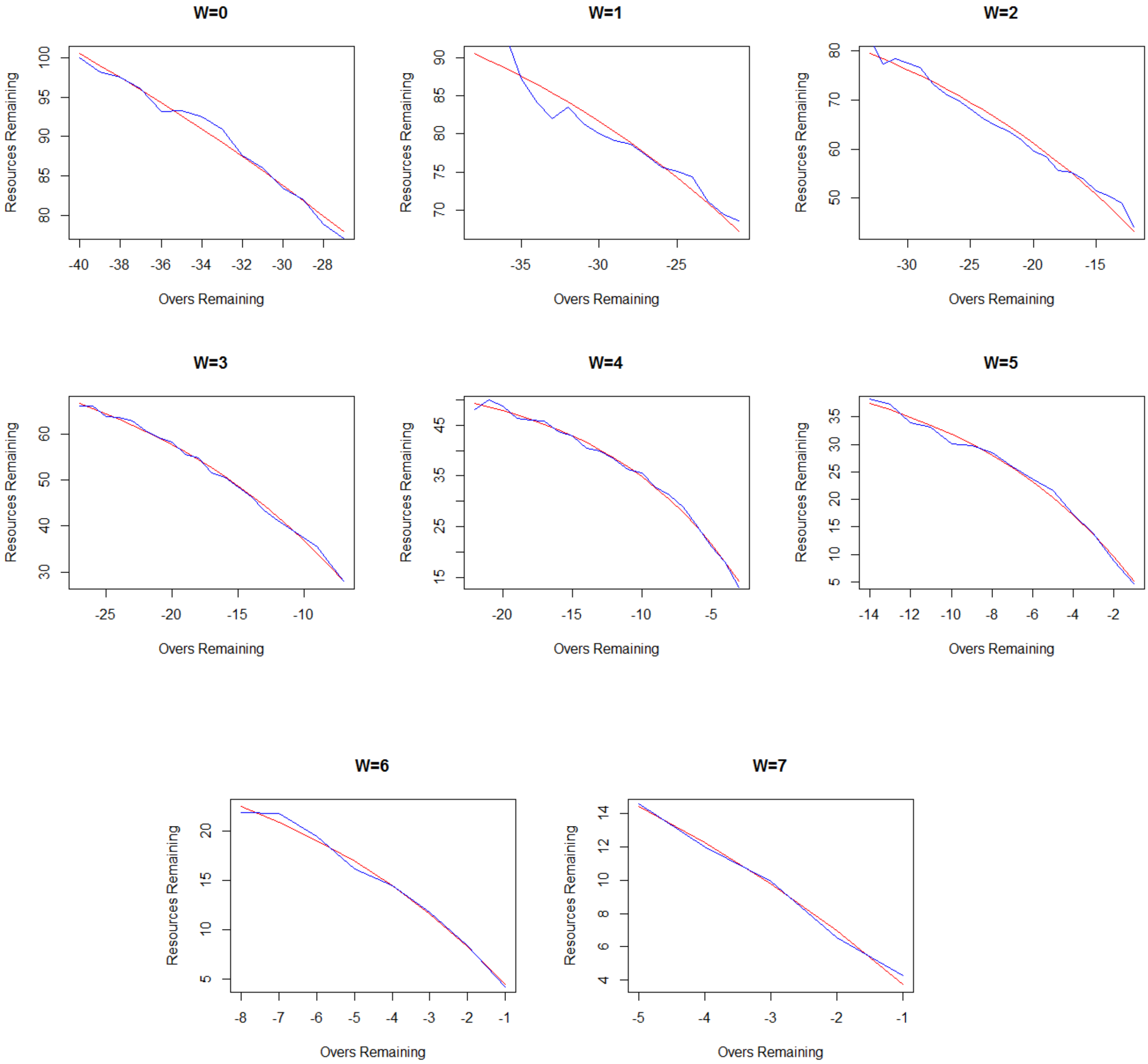
Residuals
Residuals
Fitted values
Autocorrelation
Residuals i+1
Residuals i
Normal Q-Q Plot of
Standardized Residuals
Sample Quantiles
Theoretical Quantiles

W=7

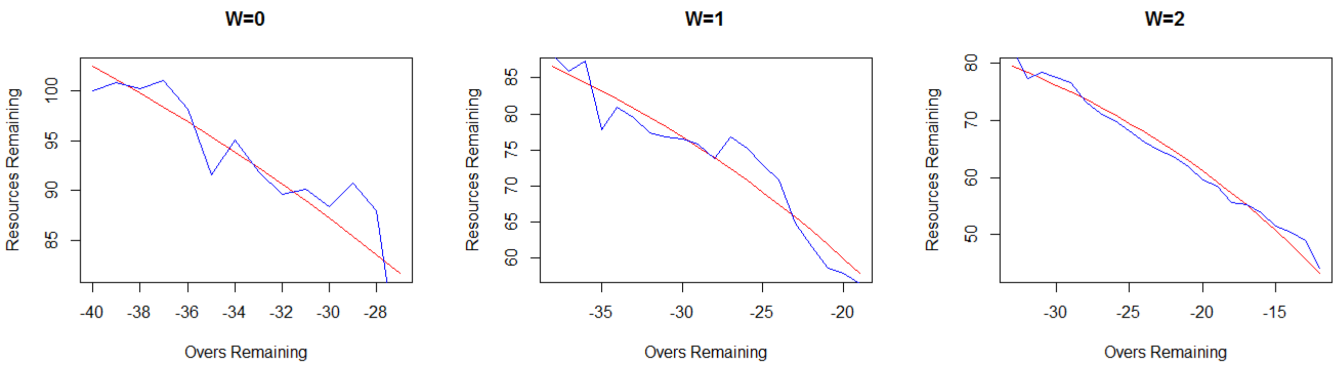


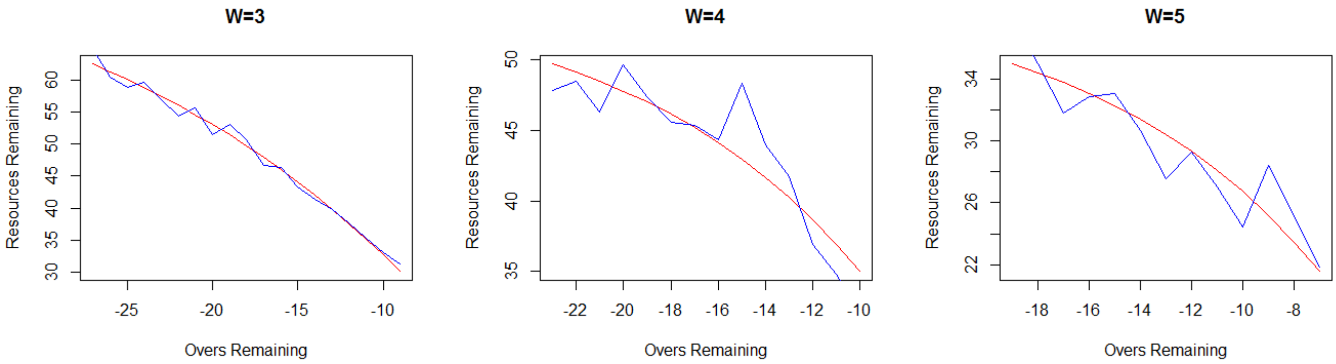
Residuals
Residuals
Fitted values
Autocorrelation
Residuals i+1
Residuals i
Normal Q-Q Plot of
Standardized Residuals
Sample Quantiles
Theoretical Quantiles

5.3a - Pro40 first innings fitted (red) vs observed (blue) values



5.3b - Pro40 second innings fitted (red) vs observed (blue) values



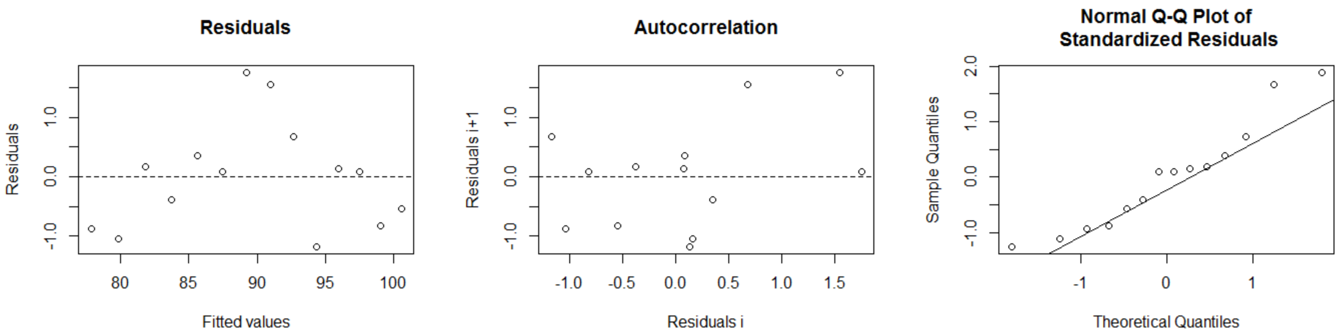


5.4 - Pro40 Diagnostics

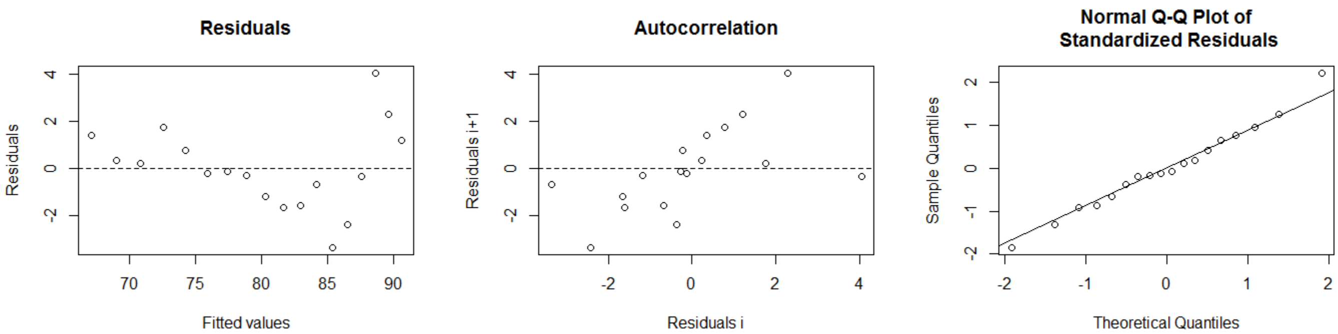
Below are Residual vs Fitted value plots, Autocorrelation plots and Normal QQ plots for the T20 fitted models.

First Innings

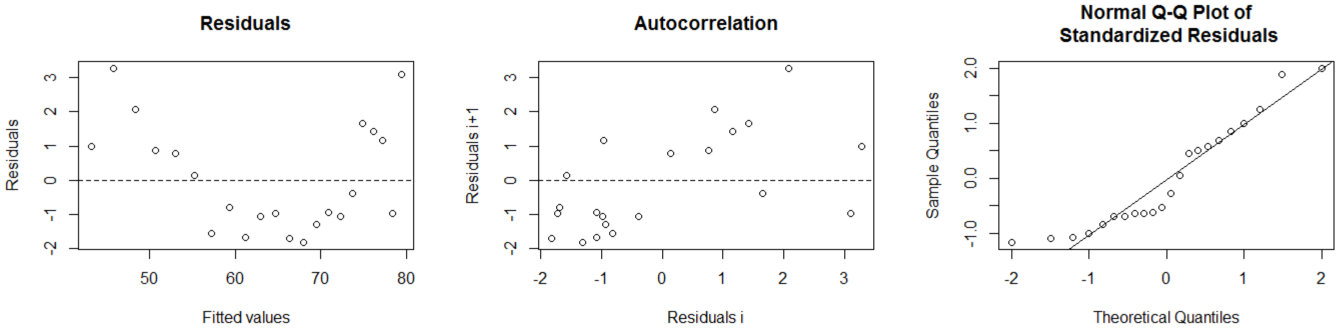
W=0



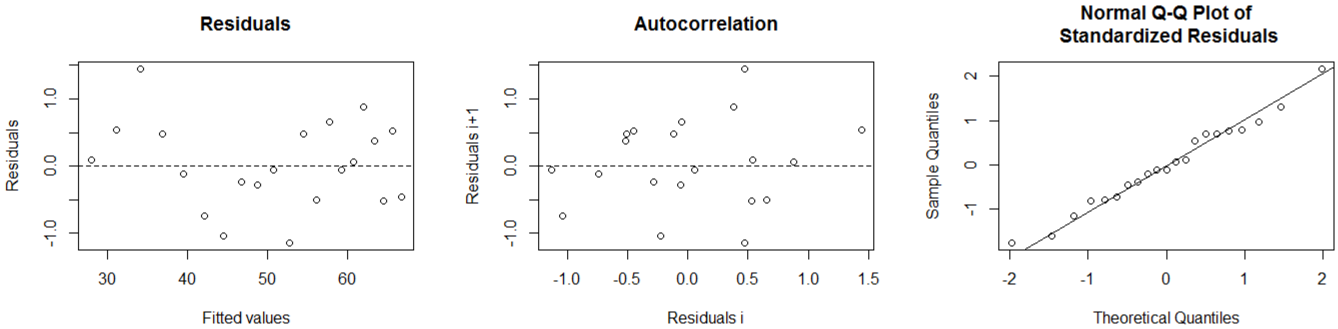
W=1



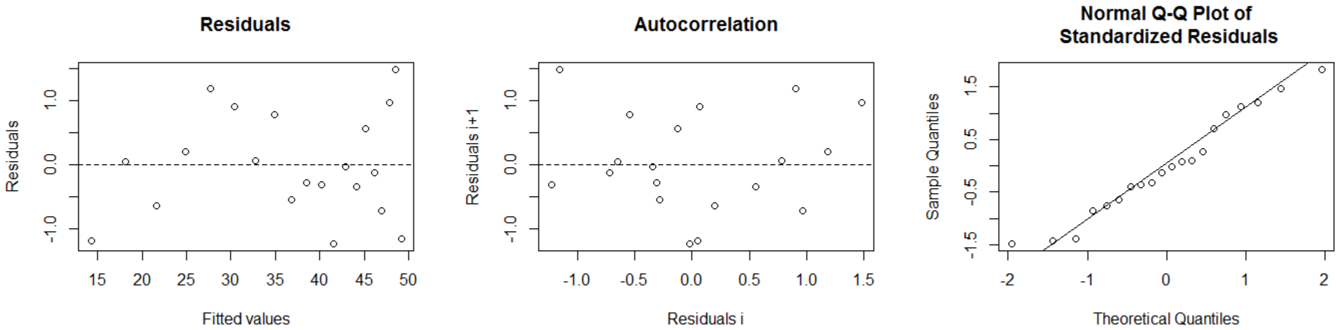
W=2


Residuals
Residuals
Fitted values
Autocorrelation
Residuals i+1
Residuals i
Normal Q-Q Plot of Standardized Residuals
Sample Quantiles
Theoretical Quantiles

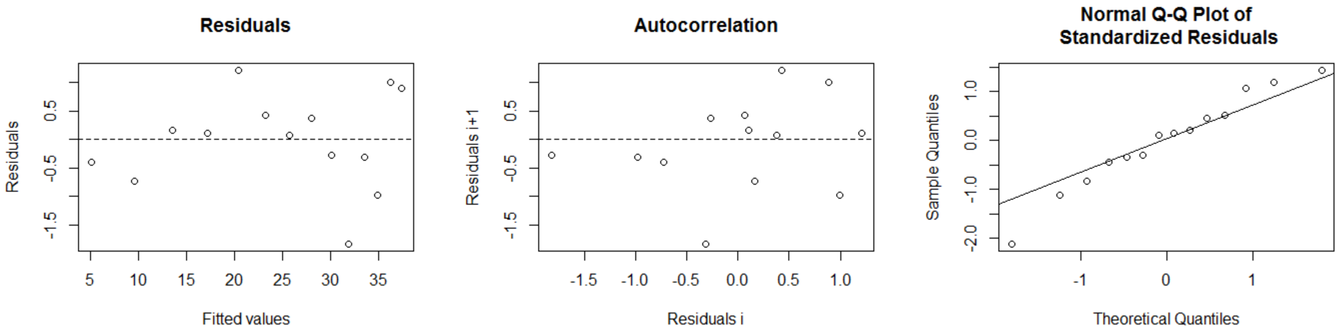
W=3


Residuals
Residuals
Fitted values
Autocorrelation
Residuals i+1
Residuals i
Normal Q-Q Plot of Standardized Residuals
Sample Quantiles
Theoretical Quantiles

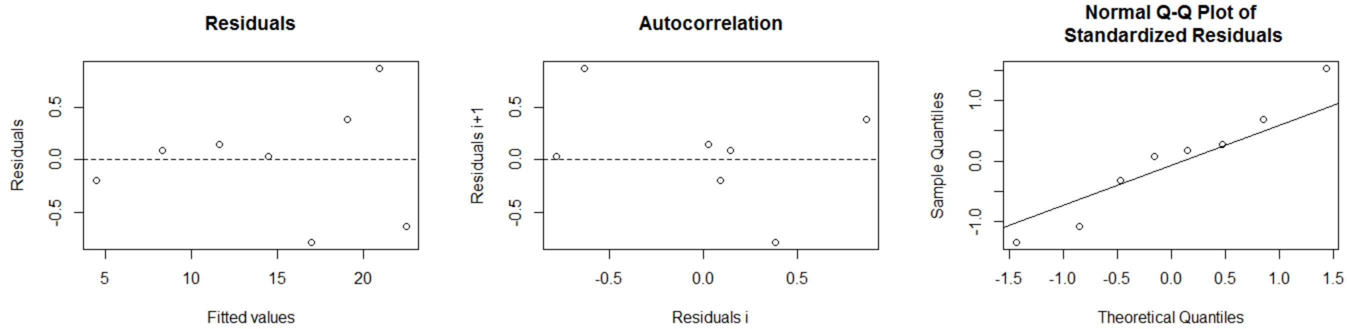
W=4


Residuals
Residuals
Fitted values
Autocorrelation
Residuals i+1
Residuals i
Normal Q-Q Plot of Standardized Residuals
Sample Quantiles
Theoretical Quantiles

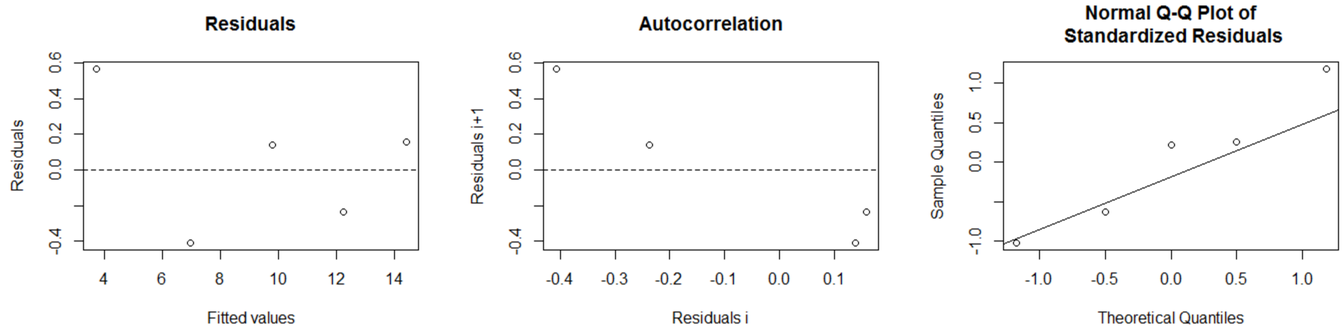
W=5


Residuals
Residuals
Fitted values
Autocorrelation
Residuals i+1
Residuals i
Normal Q-Q Plot of Standardized Residuals
Sample Quantiles
Theoretical Quantiles

W=6

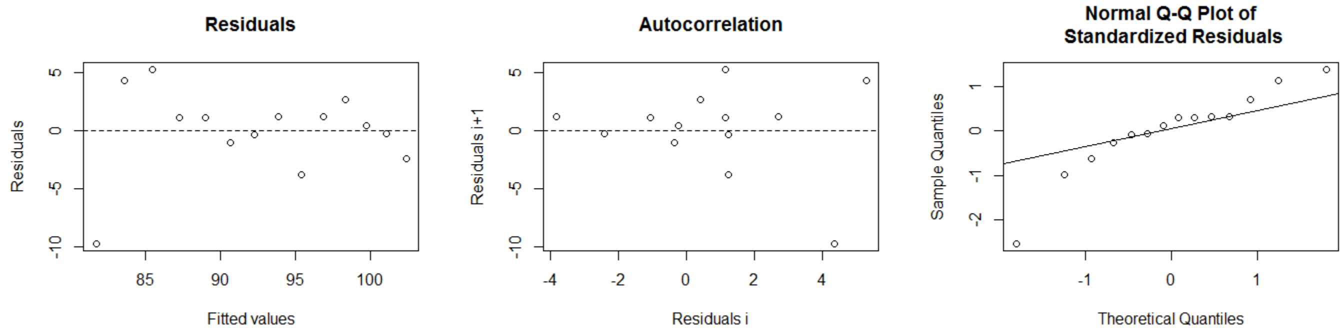


W=7

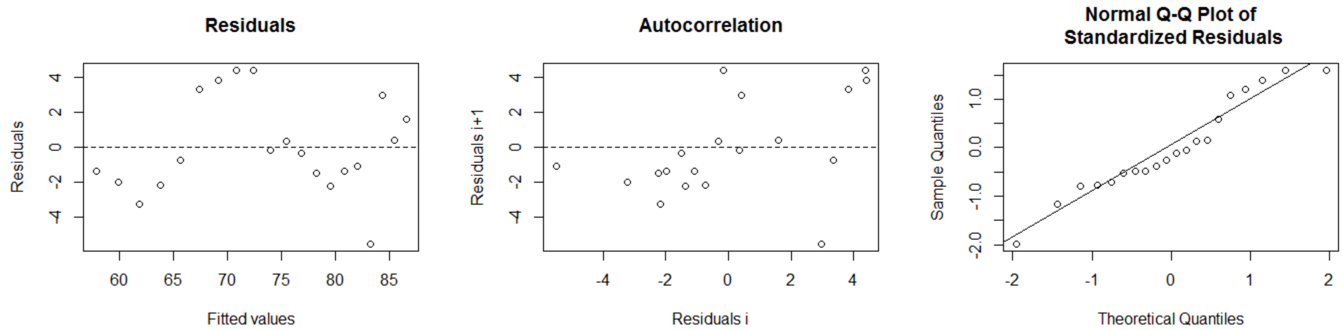


Second Innings

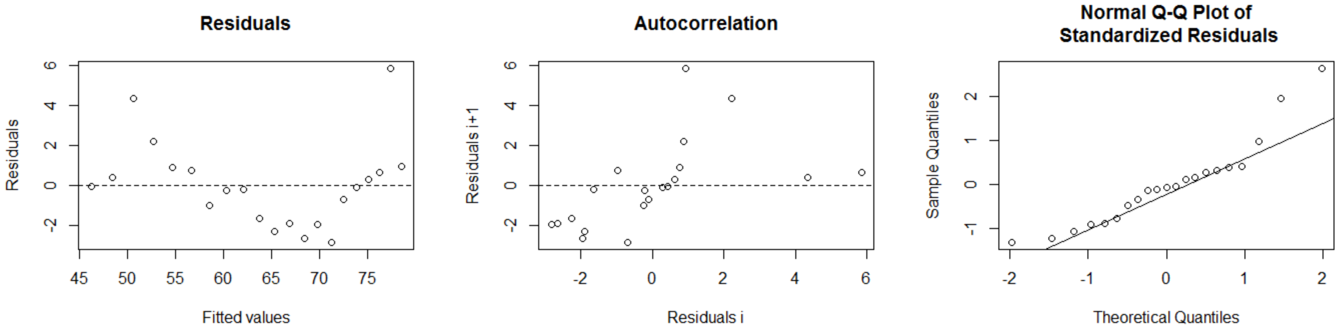
W=0



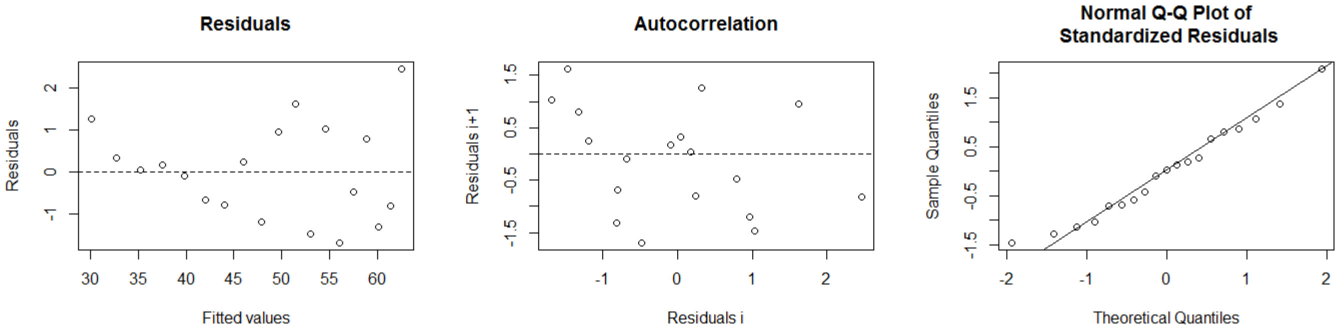
W=1



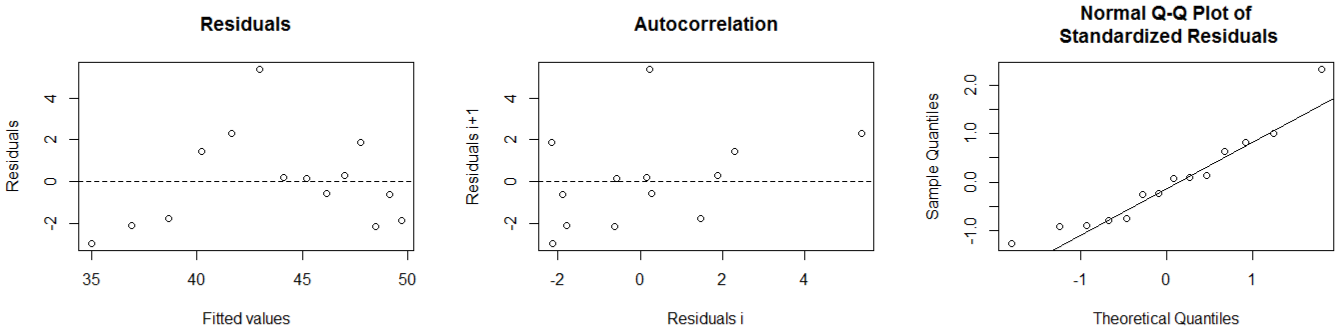
W=2


Residuals
Residuals
Fitted values
Autocorrelation
Residuals i+1
Residuals i
Normal Q-Q Plot of
Standardized Residuals
Sample Quantiles
Theoretical Quantiles

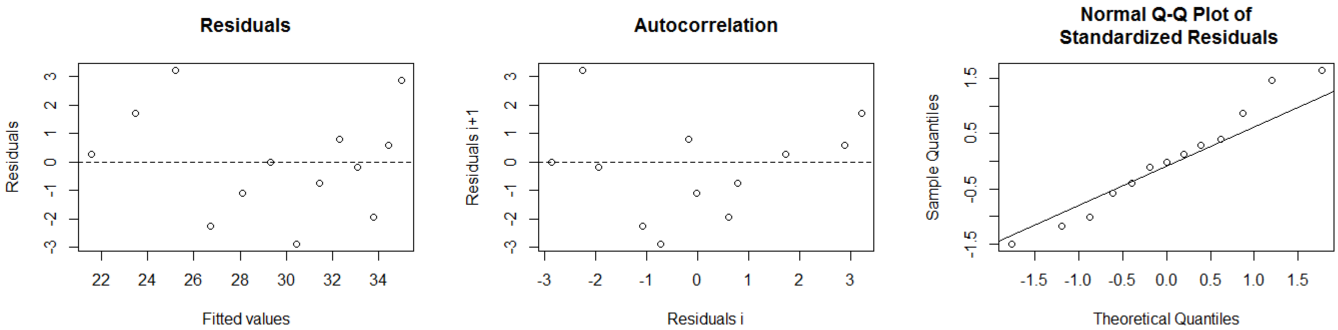
W=3


Residuals
Residuals
Fitted values
Autocorrelation
Residuals i+1
Residuals i
Normal Q-Q Plot of
Standardized Residuals
Sample Quantiles
Theoretical Quantiles

W=4


Residuals
Residuals
Fitted values
Autocorrelation
Residuals i+1
Residuals i
Normal Q-Q Plot of
Standardized Residuals
Sample Quantiles
Theoretical Quantiles

W=5


Residuals
Residuals
Fitted values
Autocorrelation
Residuals i+1
Residuals i
Normal Q-Q Plot of
Standardized Residuals
Sample Quantiles
Theoretical Quantiles