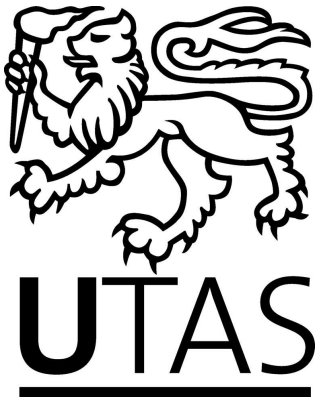

University of Tasmania

School of Mathematics and Physics

Testing for a Winning Mood Effect and Predicting Winners at the 2003 Australian Open

Lisa Miller

October 2005

Submitted in partial fulfilment of the requirements for the Degree of Bachelor of Science with Honours

Supervisor: Dr Simon Wotherspoon

Acknowledgements

I would like to thank my supervisor Dr Simon Wotherspoon for all the help he has given me throughout the year, without him this thesis would not have happened. I would also like to thank my family, friends and fellow honours students, especially Shari, for all their support.

Abstract

A 'winning mood effect' can be described as a positive effect that leads a player to perform well on a point after performing well on the previous point. This thesis investigates the 'winning mood effect' in males singles data from the 2003 Australian Open. It was found that after winning the penultimate set there was an increase in the probability of winning the match and that after serving an ace and also after breaking your opponents service there was an increase in the probability of winning the next service. Players were found to take more risk at game point but less risk after their service was broken. The second part of this thesis looked at predicting the probability of winning a match based on some previous measure of score. Using simulation, several models were able to predict the probability of winning based on a score from earlier in the match, or on a difference in player ability. However, the models were only really suitable for very large data sets. Isotonic regression was applied to data from the 2003 Australian Open and the first server of the match was found to have a higher chance of winning the match when the difference in points won, or the difference in games won, was small.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Four Grand Slam Tennis Tournaments are held every year and each tournament is played in a different location subject to differing weather conditions and court surfaces. Many tennis players have preferences to certain Grand Slams because the court surface suits their playing style. Magnus and Klaassen have performed several studies on data taken from Wimbledon mentioning the idea of what they termed a 'winning mood effect', where a player having done well on one point goes on to perform well on the next point. The first part of this thesis, which is addressed in Chapters 2-6, looks at determining if the same effect can be found at the 2003 Australian Open.

Chapter 2 provides an introduction to the Australian Open, giving a brief outline of its history and beginnings. The conditions of the Australian Open and how they may differ from other Gram Slams are discussed and it outlines how the statistics are collected at the Australian Open.

Chapter 3 reviews several papers on commonly espoused issues relating to tennis that were tested at Wimbledon by Magnus and Klaassen and provides an explanation of a 'winning mood effect'. Many of the hypotheses that were tested in this thesis were by-products of hypotheses tested by Magnus and Klaassen when they found evidence of a 'winning mood effect'.

Chapter 4 outlines some of the basic statistical methods used in later chapters. It discusses several basic statistical methods for testing for in-

dependence and then extends these to look at GLM's and random effects models. Gibbs sampling is also introduced in this chapter and is used later to develop a Bayesian approach to analysing data from the Australian Open.

Many of the variables in the dataset for the 2003 Australian Open could not be used for analysing the hypotheses. Chapter 5 explains the format in which the data were received and the construction of new variables used in the analysis. Information such as when game points, break points and set points occurred had to be derived as well as who was the winner of each point.

There were several different approaches used to test hypotheses relating to a 'winning mood effect'. Chapter 6 explains the problems that arose when comparing probabilities of a single player or within pairs of players due to the symmetry of a winning point. It also outlines what hypotheses were used to test for a 'winning mood effect' and the results of these tests.

The second part of this thesis, which is addressed in Chapters 7 and 8, looks at predictive modelling of win-loss scenarios based on a measure of score. The methods behind the models are explained as well as how the models performed. This thesis looks at modelling two different types of win-loss scenarios based on models that assume a monotonic relation between player skill, or score, and the probability of winning. The first scenario looks at predicting the match winner based on a fixed score during a match using isotonic regression, truncated beta deviates, monotone regression splines, penalty splines, and probit regression. The second scenario looks at predicting a tournament outcome based on differences in player ability, which were determined by the Bradley Terry model, commonly used for ranking. Isotonic regression and generalised additive models are used to generalise the Bradley Terry model in an attempt to predict tournament outcomes. Several of the models were tested on simulated data and isotonic regression was tested further on data from the 2003 Australian Open. Unfortunately some of the methods were too computationally complex to implement.

Chapter 2

Australian Open Tennis

The Australian Open is the first of four Gram Slam Tournaments played at various locations and times throughout the year. The Australian Open is held in mid-January on hard courts (Rebound Ace), the French Open in May-June on clay courts, Wimbledon in June-July on grass courts and the US open in August-September on hard courts (DecoTurf II). Wimbledon is the oldest of the Grand Slams, followed by the US Open, French Open and Australian Open. All the Grand Slam tournaments include a males and females singles competition, males, females, and mixed doubles, as well as junior and master competitions.

2.1 History

The Australian Open was first held in 1905 as the Australasian Championships at the Warehouseman's Cricket Ground in St Kilda Road, Melbourne. In 1927 the Tournament became the Australian Championships and was renamed in 1969 as the Australian Open. Since 1905 the Tournament has been held at six different venues, 56 times in Melbourne, 17 in Sydney, 14 in Adelaide, eight in Brisbane, three in Perth and twice in New Zealand (in 1906 and 1912). In 1972 it was decided that one city should permanently host the Australian Open and as Melbourne attracted the largest support

the Kooyong Lawn Tennis Club was chosen. Following a decrease in interest, the Australian Open was shifted to Melbourne Park (formerly Flinders Park) for the 1988 Australian Open. This proved to be an immediate success with attendance increasing 90 percent from 140,000 in Kooyong 1987 to 266,436 in Melbourne 1988.

2.2 Conditions

The Australian Open is held in January and players have to deal with the Australian summer, where temperatures on the court can reach more than 40 degrees Celsius. These extreme temperatures can cause the tennis balls to expand and shrink, creating problems for players as this changes the reaction of the ball.

The Australian Open is played on a hard court surface known as Rebound Ace, which is made of asphalt and sand. Originally, while at Kooyong, the Australian Open was played on grass courts but changed to a Rebound Ace when moved to Melbourne Park. Rebound Ace was designed to provide a surface that is suitable to all playing styles and provides a consistent ball bounce and speed. However, Rebound Ace is a slower surface compared to the hard court Deco Turf II used in the US Open and thus benefits the slower, perhaps taller players. Some players have complained that Rebound Ace increases ankle and knee injuries as the players feet tend to grip to the surface too well.

The Australian Open is the only Grand Slam tournament that can feature indoor play due to the movable roof on Rod Laver Arena, centre court and the first show court. Matches on these courts can be played with the roof open on fine days, or closed on wet or extremely hot days.

2.3 Collecting Statistics

Statistics for tennis events have grown substantially over the past few years. Statistics use to be limited and only viewed in magazines or newspapers the following day or even week. They are now collected at all courts for all events and available on live broadcasts or via the web. Tennis Statisticians are selected to record the statistics for each match and are required to undergo training in the lead up to the event. The term 'Tennis Statistician' is somewhat misleading, their role is really data collection. The majority are under the age of 25 and have no statistical training. The Tennis Statistician is seated on the court to record data using a notebook computer linked to the central IBM scoring network. This is used to update the scoreboards on the court and around the venue and also goes live to the media and the Internet.

Chapter 3

Literature Review

Magnus and Klaassen have published several papers on different areas of tennis. This thesis focuses on two of these papers, one of which looks at some frequently espoused ideas about tennis and the other about the final set in a tennis match. They analysed 88,883 points taken over four years at Wimbledon (1992 - 1995) for singles matches played on show courts finding different results for male and female matches and different combinations of seeded and non-seeded players. During their research they found evidence of what they termed a 'winning mood effect' - loosely defined as 'if you do well in one point (game) you will also do well in the next point (game)' (p. 26, 1996). This effect is contrary to the idea often promoted by commentators that the opponent will hit back after losing the previous point (or game). A person having won a point generally gains confidence to go on and perform better in the next point, or conversely a person having lost a point loses confidence and goes on to perform badly in the next point.

Magnus and Klaassen (1996) found players performance and strategy was affected by the outcome of the previous point, suggesting points were not independently distributed. After a double fault there was a decrease in the percentage of aces served, which suggests following a double fault the server increases his efforts to ensure the next first service is in. Conversely, after a double fault the percentage of first serves in decreases when a seed is serving

against a non-seeded player. After serving an ace, a player was found to take more risk with less first services in, more aces served and more points won on service, which suggests serving an ace also affects the next point.

Magnus and Klaassen (1996) found evidence that game points and break points were not played equally. At game point the server had more aces and more points won on service and at break point, which would be considered a more important point than a game point, the server made sure their next first service was in. Thus at game point the server took more risk and at break point the server took less risk.

Further evidence for a 'winning mood effect' was found when looking at break points. After breaking an opponents service there was an increased chance of winning a service game. Conversely, after missing break points in the previous game there was an increased chance of losing a service game, particularly for games between two non-seeds.

Looking at the relationship between the final game in one set and the first game in next set, Magnus and Klaassen (1996) found that a player who serves first in a set is more successful if they won the previous set. However, this was similar for all service games in that set, which is not surprising as the better player probably won the previous set and is likely to go on and have more success in the subsequent set. It appears that for males singles tennis if you are able to break once in the set it gives you a big enough edge to be able to go on and win the set. The stronger males singles players appear to be less affected by what happens in previous points and consequently seeded players are more likely to take risks when serving at break point.

Commentators often suggest that the player who wins the fourth set will win the match. Magnus and Klaassen (1999) found this is only true for matches between two non-seeds and when two seeds play each other the winner of the fourth set will probably lose the match. Magnus and Klaassen found a seeded player is the favourite, over a non-seed, to win the match at the start of the final set, but if a non-seed is to win it is more likely to occur in males rather than females singles tennis. This is because the difference in

quality between the two players in the final set is smaller for males than for females. In males tennis the final set is the fifth set and a non-seed must win two sets in order to reach the final, whereas, in females tennis the final set is the third and a non-seed need only win one.

Chapter 4

Statistical Methods

This chapter introduces the statistical methods that will be used for hypothesis testing and predictive modelling in this thesis. The methods in this chapter use maximum likelihood estimation techniques and tests for independence in 2x2 contingency tables. Pearson's Chi-Squared test, Cochran's Mantel Haenszel test, generalised linear models and random effects models are used to test the hypotheses in Chapter 6. Maximum likelihood estimation and gibb's sampling are also introduced and used for predictive modelling in Chapter 7.

4.1 Maximum Likelihood Estimation

Maximum likelihood estimation is a general process of obtaining estimated values for unknown parameters. The maximum likelihood estimate (MLE) is the parameter value that maximises the likelihood, the probability of the data as a function of the model parameters.

For example consider m observations $y_1, \ldots, y_m$ of a binomial distributed random variable

$$Y \sim \text{Bin}\,(n, \pi)$$

with n known and π to be estimated. The likelihood $l(\pi)$ is the probability of observing $y_1, \ldots, y_m$ as a function of π.

From the binomial probability mass function it follows immediately that

$$l(\pi) = \prod_{i=1}^{m} \binom{n}{y_i} \pi^{y_i} (1-\pi)^{n-y_i}.$$

We wish to maximise the likelihood function with respect to π, thus only terms involving this parameter are considered. As log is monotonic, the likelihood l attains its maximum where $\log l$ is a maximum, and typically it is more convenient. Ignoring additive constants the binomial log likelihood takes the form

$$\begin{aligned} l(\pi) &= \sum_{i=1}^{m} \log \left[\pi^{y_i} (1-\pi)^{n-y_i} \right] \\ &= \sum_{i=1}^{m} \left(y_i \log \pi + (n-y_i) \log (1-\pi) \right) \end{aligned}$$

Taking the partial derivative of the log likelihood with respect to π gives

$$\begin{aligned} \frac{\partial l}{\partial \pi} &= \sum_{i=1}^{m} \left(\frac{y_i}{\pi} - \frac{(n-y_i)}{(1-\pi)} \right) \\ &= \sum_{i=1}^{m} \frac{(y_i - n\pi)}{\pi (1-\pi)}. \end{aligned}$$

At the maximum $\pi = \hat{\pi}$, and

$$\left. \frac{\partial l}{\partial \pi} \right|_{\pi=\hat{\pi}} = \sum_{i=1}^{m} \frac{(y_i - n\hat{\pi})}{\hat{\pi}(1-\hat{\pi})} = 0,$$

from which it follows that

$$\hat{\pi} = \frac{1}{mn} \sum_{i=1}^{m} y_i.$$

4.2 Pearson's Chi-Squared Test

Pearson's Chi-Squared test is a large sample approximation of Fisher's Exact test for two-sided p-values. Fisher's Exact test is used for small samples to give an exact test of independence. Pearson's Chi-Squared test can be used

to test independence and perform goodness of fit tests to determine how well a model will predict the observed data.

Given a multinominal sample of size n the marginal distribution of the observed data, n_{ij}, is the binomial distribution

$$\text{Bin}\left(n_{ij}, \pi_{ij}\right).$$

The binomial distribution is a special case of the multinominal distribution with $c = 2$. The number of observations with explanatory variable i and response variable j is

$$n_{ij} = n_{11}, \ldots, n_{rc}$$

where $i = 1 \ldots, r$ and $j = 1 \ldots, c$.

To test independence between the row and column marginals consider the null hypothesis

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}.$$

Under H_0 the expected value of n_{ij} is

$$\mu_{ij} = n\pi_{i+}\pi_{+j}.$$

These are called the expected frequencies when the null hypothesis is true. The column variable is j, the row variable is i and the symbol + denotes a summation where $i+$ denotes the row marginal totals and $+j$ denotes the column marginal totals.

The probabilities $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ are restricted to

$$\sum_i \pi_{i+} = 1 = \sum_j \pi_{+j}.$$

Since these are usually unknown, the MLE's must be obtained. The MLE's are the sample marginal proportions $\hat{\pi}_{i+} = n_{i+}/n$ and $\hat{\pi}_{+j} = n_{+j}/n$ and therefore the estimated expected frequencies become

$$\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n_{i+}n_{+j}/n^2.$$

The Pearson's Chi-Squared statistic measures the deviation between the observed and expected frequencies

$$\chi^2 = \sum_r \sum_c \frac{(n_{ij} - \hat{\mu}_{ij})}{\hat{\mu}_{ij}} \tag{4.1}$$

and in the absence of the association follows a Chi-Squared distribution with degrees of freedom

$$df = (I-1)(J-1).$$

4.3 Cochran's Mantel Haenszel Test

Cochran's Mantel Haenszel test is a non-model based test of the null hypothesis to test conditional independence in (IxJxK) tables, where $i = 1 \ldots, r$, $j = 1 \ldots, c$ and $k = 1 \ldots, s$, assuming that there is no three-way interaction. Mantel Haenszel tests estimate the strength of an association rather that testing hypotheses about the association. With several outcome categories, marginal homogeneity can be tested by applying the Cochran's Mantel Haenszel test using a single stratum for each subject with each row representing a particular outcome.

Considering (2x2xk) tables, $c = r = 2$, condition on both the predictor totals (n_{1+k}, n_{2+k}) and the response outcome totals (n_{+1k}, n_{+2k}). Assuming the distribution of the frequencies in stratum k is a product of two binomial probabilities, one for each condition, the Cochran's Mantel Haenszel statistic is calculated by taking one of the four cells, say n_{11k}, for each stratum and finding its expected value and variance. The same null hypothesis is tested as for Pearson's Chi-Squared test, no relationship between the explanatory variable and the response variable, but now independence within the stratum k is considered. Under H_0, the expected value of n_{11k} becomes

$$\mu_{11k} = E(n_{11k}) = n_{1+k}n_{+1k}/n_{++k}.$$

Similarly the variance of n_{11k} becomes

$$Var(n_{11k}) = n_{1+k}n_{2+k}n_{+1k}n_{+2k}/n_{++k}^2(n_{++k} - 1).$$

Cochran's Mantel Haenszel statistic (CMH) combines the information from the k strata by comparing $\sum_k n_{11k}$ to its null expected value and the test statistic for independence of row and column marginals across strata is

$$CMH = \frac{\left[\sum_k \left(n_{11k} - \mu_{11k}\right)\right]^2}{\sum_k Var\left(n_{11k}\right)}. \tag{4.2}$$

Under the null hypothesis the CMH statistic is asymptotically $\chi^2(1)$. When there is only one stratum being considered the CMH statistic reduces to the Pearson's Chi-Squared statistic for independence between (2x2) contingency tables.

Cochran's Mantel Haenszel test treats the rows in each (2x2) table as two independent binomial distributions rather than as a hypergeometric distribution, as in the Mantel Haenszel test. CMH statistic is an alternative to the Mantel Haenszel (MH) statistic, but summed over the strata s, the only difference is the variance of the MH statistic takes the form,

$$Var\left(n_{11k}\right) = n_{1+k}n_{2+k}n_{+1k}n_{+2k}/n_{++k}^3.$$

4.4 Generalised Linear Model

A Generalised Linear Model (GLM) is a generalisation of a linear regression, which extends an ordinary regression model to include response distributions that are non-normal and model functions of the mean. A GLM is made up of a random component, a systematic component and a link function. The random component identifies the response variable from a natural exponential family of distributions, the systematic component identifies the explanatory variables that enter the model in the form of a linear predictor, and the link function relates the expected value of the response variable, the mean, back to the linear model.

The response variable y may be distributed about its expected value μ in relation to any distribution F from the exponential family,

$$y_i \sim F\left(\mu_i\right).$$

The explanatory variables $x_1, \ldots, x_m$ enter the model through the linear predictor η, which is related to μ by a monotonic function $\eta_i = l(\mu_i)$ called the link function,

$$l(\mu_i) = \eta_i = X\beta.$$

For example consider m observations $y_1, \ldots, y_m$ of a binomial distributed random variable

$$y_i \sim \text{Bin}(n_i, \pi_i)$$

with n_i known and π_i to be estimated.

Let the linear predictor η_i of the dependent variable Y be a linear function of the explanatory variable X,

$$\eta_i = X\beta = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_m x_{mi}.$$

In the general linear model, the conditional mean of the dependent variable Y is estimated. For GLM an identity link function is used to relate μ_i to the linear predictor η_i,

$$l(\mu_i) = \eta_i. \tag{4.3}$$

Thus a nonlinear model can be transformed to a model that is linear in its parameters. The conditional mean function of Y is

$$\mu_i = n_i \pi_i$$

and the variance is

$$\nu_i = n_i \pi_i (1 - \pi_i).$$

Assuming Y is drawn from a standard logistic distribution, π_i takes the form,

$$\pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} = \frac{1}{1 + \exp(-\eta_i)} \tag{4.4}$$

where π_i is a nonlinear function of η_i. However, the linear predictor function given by a logit transformation (or link) makes the model linear in the logit scale

$$\eta_i = \log\left(\frac{\pi_i}{1 - \pi_i}\right). \tag{4.5}$$

4.5 Random Effects Model

Random Effects models are a class of Generalised Linear Mixed Models (GLMM) that extend the Generalised Linear Model by modelling a GLM as well as a random effect. The parameters of the random effects act as nuisance parameters that make estimation of the fixed effects difficult and measure the amount of variability not explained by the fixed effects.

The linear predictor for a random effects model is similar to the linear predictor for a GLM but with an additional vector of random effects values. It takes the form

$$l(\mu) = X\beta + U \tag{4.6}$$

where U is a vector of random effect values. This model assumes that the U_i's are independently Normally distributed random variables with zero mean and unknown variance.

Random Effects models can be fitted using a method known as Penalised Quasi Likelihood (PQL). Instead of maximising the likelihood function, PQL uses iterative fitting to maximise an approximation to the likelihood function. PQL is frequently used because it does not involve numerical integration or Monte Carlo approximation. For large data sets it is more computationally feasible and easier to program compared with exact maximum likelihood methods.

4.6 Gibbs Sampling

In order to make inferences about model parameters and to make predictions it is sometimes necessary to integrate over high dimensional probability distributions. Gibbs sampling, a special case of the single component Metropolis-Hastings algorithm, is a technique used to construct a Markov chain from a required distribution with the purpose of forming sample averages to approximate the expected value of some function of a vector of random effects. Gibbs sampling is simply a Metropolis-Hastings algorithm

in which the proposed distribution for each parameters is chosen to be the full conditional distribution for that parameter. As a consequence, the acceptance probability for the Metropolis-Hastings algorithm is always one, that is no proposals are ever rejected, making the scheme more efficient than simple Metropolis-Hastings.

Let X be a vector of k random variables comprising of random effects, $\{X_0, X_1, X_2, \ldots, X_k\}$ and $\pi(.)$, the distribution of X, be a likelihood.

We wish to evaluate the expectation

$$E[f(x)] = \frac{\Sigma f(x)\,\pi(x)\,dx}{\Sigma \pi(x)\,dx}.$$

Then at each time t, the next state X_{t+1} is chosen by first sampling a candidate point Y from a proposal distribution $q(.|X_t)$, which may depend on the current X_t. The candidate point Y is then accepted with probability

$$\alpha(X_t, Y)$$

and the next state becomes $X_{t+1} = Y$. If the candidate is rejected, the chain does not move, and $X_{t+1} = X_t$.

For gibbs sampling, the proposal for updating the i^{th} component of X is

$$q_i(Y_{.i}|X_{.i}X_{.-i}) = \pi(Y_{.i}|X_{.-i}) \tag{4.7}$$

where the full conditional distribution is

$$\pi(Y_{.i}|X_{.-i}).$$

In this case $\alpha(X_t, Y) \equiv 1$, and candidates are never rejected. The resulting gain in computational efficiency must be weighed against the complexity of drawing from the full conditional distribution.

Chapter 5

Description of Data

The data in this thesis was based on 118 matches played at the 2003 Australian Open in the males singles championship and was obtained courtesy of Swinburne University of Technology and the Australian Open. This chapter looks at describing the format of the data and how derived variables were constructed to facilitate the hypothesis tests to be discussed in Chapter 6.

5.1 The Data

There were originally 127 matches played at the 2003 Australian Open, with a total of 27,738 point by point data for these matches. However, in this thesis only complete matches were considered, matches that did not involve a retirement or a walkover. Retired matches could dramatically effect the probability of winning a match if the person that was winning the match was the retiree. Thus the person who won the match would have a low probability of winning. Walkover matches contain no relevant point data, except the person who won the match, so there was no need for them to be included in the analysis. This reduced the data set to 26,874 points from 118 matches. Basic summaries, such as means and counts, could be obtained from the data, but in order to test the hypotheses additional columns needed to be added to the data set to identify important points. The data set provided

information on who was serving, whether the first or second serve was in, whether the point was decided by a double fault, ace or winner, and whether the point was won at the net or not. To test the hypotheses for a 'winning mood effect' information of who won the point and whether the point was a game point or a break point was required.

Table 5.1 provides a summary of the data.

Number of ...	**Males Singles**
Matches	118
Sets	448
.. final	26
.. non final	422
Games	4250
.. final	269
.. non final	3981
Points	26874
Sets/Match	3.80
Games/Set	9.49
Points/Game	6.32

Table 5.1: *Summary of Males Singles Data from 2003 Australian Open.*

5.2 The Original Variables

The data was received from Swinburne University in the form of point by point data for every match, along with a paper by Clarke and Norton (2002) describing the process of collecting statistics at the Australian Open. The format of the data was a series of numerical symbols indicating different events in a tennis match represented in different columns for every point played in the match. The meaning of these numbers was different across columns and had to be interpreted in order to analyse the data. A description

of the original variables can be found in Table 5.2.

Variable	Description
Match	match number
Set	set number
ptforc	total points for first server in current game
ptagtc	total points against first server in current game
gmforc	total games for first server in current set
gmagtc	total games against first server in current set
server	server of current point
fserve	outcome of first serve
sserve	outcome of second serve
actor	last player to play the ball
stroke	type of last stroke played
effect	effect of last play
atnet	number of players at net
points	point number

Table 5.2: *Summary of the Original Variables.*

Match

Match identifies the round number, the match number and the type of match. For example $ms101$ means males singles match, round 1, game 1.

Set

Set is the number of the set at that point. This has a minimum of three and a maximum of five.

Ptforc

Ptforc is the number of points won by the first server in the game at that point. It can range from zero to arbitrarily large, depending on the number of deuces, but typically ranges from 0 to 3.

Ptagtc

Ptagtc is the number of points won against the first server in the game

at that point, i.e. the number of points won by the second server. It can range from zero to arbitrarily large, depending on the number of deuces, but typically ranges from 0 to 3.

Gmforc

Gmforc is the number of games won by the first server in the set at that point. This can range from 0 to 7.

Gmagtc

Gmagtc is the number of games won against the first server in the set at that point, i.e. the number of points won by the second server. This can range from 0 to 7.

Server

Server is the server of the point. This will be a 1 for the first server in the match and a 2 for the second server in the match.

Fserve

Fserve represents the first serve at that point. 1 - in play, 2 - fault, 3 - winner, 4 - ace.

Sserve

Sserve represents the second serve at that point. 0 - no second serve, 1 - in play, 2 - fault, 3 - winner, 4 - ace.

Actor

The actor is the last player to make a play on the ball at that point. 0 - no play on last ball, 1 - player who served first at the start of the match, 2 - player who served second at the start of the match.

Stroke

Stroke is the type of stroke used by the last player to make a play at that point. 0 - no stroke, 1 - forehand, 2 - backhand, 3 - overhead, 4 - volley.

Effect

The effect represents the last play at that point. 0 - no play on the ball, 1 - unforced error, 2 - forced error, 3 - winner.

Atnet

The exact meaning of atnet was unclear from the paper that was received

with the data set (Clarke and Norton, 2002) however, this was of no concern as atnet was not used in the analysis. It could mean the number of players at the net at the end of the play at that point. 0 - no players at net, 1 - either player at the net, 2 - Both players at the net. It could also mean the player at the net at the end of play at that point. 0 - No players at the net, 1 - first person to serve in the match at the net, 2 - second person to serve in the match at the net. However, the first is more plausible as the latter does not take into account the possibility of both players being at the net.

Point

Point is the number of points played in the match at that point. It can range from zero to arbitrarily large, however the average number of points per match was 228.

5.3 Scoring System

The scoring system for points is represented as a 0 standing for love, 1 standing for 15, 2 standing for 30 and 3 standing for 40. For players to be on deuce, both players must have won the same number of points. A deuce for both players was represented by an odd number greater than or equal to 5. An advantage to a player was represented by an even number great than or equal to 4. Thus for the first server in a match to win the game he must win the point and have a *'ptforc'* of an odd number greater than or equal to 3, providing the other player *'ptagtc'* is not equal to 4. In a tiebreak *gmforc=gmagtc=6*, to win the game the first server must win the point and *'ptforc'* must be greater than or equal to 6 with a lead of 1.

5.4 Processing

In order to perform more in-depth analysis of the data, pre-established columns were used to construct derived variables corresponding to more convenient events. This included identifying when a match, set, game and point

had been won and who was the winner and also when there was a game point, a set point and a break point. A description of the derived variables can be found in Table 5.3.

Variable	**Description**
Gwin	game winning point
Swin	set winning point
Mwin	match winning point
Winner	winner of point
Mwinner	match winner
Swinner	set winner
Gwinner	game winner
Gamept	game point
Breakpt	break point
Setpt	set point

Table 5.3: *Summary of the Derived Variables.*

Gwin

Gwin is a game winning point that occurs when the game number changes and at the very last point played. TRUE - game winning point, FALSE - not a game winning point.

Swin

Swin is a set winning point that occurs when the set number changes and at the very last point played. This can only occur on a game winning point. TRUE - set winning point, FALSE - not a set winning point.

Mwin

Mwin is a match winning point that occurs when the match number changes and at the very last point played. This can only occur on a game winning point and a set winning point. TRUE - match winning point, FALSE - not a match winning point.

Winner

To obtain the winner of each point, you need to look at the effect of the last play in the point. If there is no effect, then it is either a double fault where the non server wins the point, or an ace or winner where the server wins the points.

Effect	**Serve**	**Winner**
0 - no effect	double fault	non server
0 - no effect	ace or winner	server

If the effect of the last play is an unforced error or a forced error then the person who is not the actor is the winner of the point. The actor wins the point if the effect of the last play is a winner.

Effect	**Winner**
1 - unforced error	non actor
2 - forced error	non actor
3 - winner	actor

Gwinner

Gwinner is the winner of the game. The winner of each game was obtained by selecting all the game winning points and finding the winner for each of these points.

Swinner

Swinner is the winner of the set. The winner of each set was obtained by selecting all the set winning points and finding the winner for each of these points.

Mwinner

Mwinner is the winner of the match. The winner of each match was obtained by selecting all the match winning points and finding the winner for each of these points.

Gamept

Gamept is a game point. A point is a game point if winning the point could potentially win the game, and is used to determine a set point and a break point. TRUE - game point, FALSE - not a game point.

To find a game point, firstly identify the set number, because in set five if the players have each won six games then the set is not decided by a tiebreak, they continue play till one player has a two game lead on his opponent. However, for sets one to four, if the players are tied at six games each then a tiebreak is played, where the game point would be the point where a player has won at least six points, with a lead of one point. If not in a tiebreak, a point is a game point if one of the players has won at least three points and both players have not won the same amount of points.

Breakpt

Breakpt represents a break point, where winning the point could potentially allow the non-server to win the game. This can only occur on a game point and the server must have at least one less point than the non server. TRUE - break point, FALSE - not a break point. For a player to be on break point, the server, i.e. his opponent, must have one of the following scores,

$$0-40, 15-40, 30-40, \text{advantage non-server.}$$

However in a tiebreak, for a player to be on break point the server must have one of the following scores,

$$0-6, 1-6, 2-6, 3-6, 4-6, 5-6, 6-7, 7-8, \ldots.$$

Set Point

Setpt is a set point, where winning the point could potentially win you the set. This can only occur on a game point. TRUE - set point, FALSE - not a set point. For a fifth set, a point is a set point if it is a game point for the player who has won at least six games and both players have not won the same amount of games. For sets one to four, a point is a set point if it is a

game point for the player who has won five games and both players have not won the same amount of games, or at least one player has won six games. For a player to be on set point in either of the first four sets he must be on game point and have one of the following scores,

$$5-1, 5-2, 5-3, 5-4, 6-5, 6-6.$$

However in the fifth set the player must be on game point and have one of the following scores,

$$5-1, 5-2, 5-3, 5-4, 6-5, 7-6, 8-7, \ldots.$$

Chapter 6

Winning Mood Effect

There has been some speculation as to the effect of certain points on performance in males tennis matches. A study by Magnus and Klaassen found evidence toward what they termed a 'winning mood effect', where a player having done well in a previous point goes on to perform well in the next, or subsequent, points. In this chapter eight hypotheses are introduced and later tested for a 'winning mood effect' using Chi-Squared, Mantel Haenszel, GLM and Random Effects models. If such an effect is true then the idea of testing tennis hypotheses assuming points are independently distributed is incorrect. This chapter also looks at the method behind Magnus and Klaassen's analysis and the problems that were faced when fitting models for each of the hypotheses.

6.1 Hypotheses

This section outlines the eight hypotheses that were tested in this chapter. Each hypothesis was tested for a 'winning mood effect' and could be classified according to three categories; effects associated with break points, effects associated with winning a service and effects associated with winning a set.

The hypotheses tested were:

1. At break point does the server take less risk?
2. After breaking his opponents service is there an increased chance that a player will win his own service?
3. After missing a break point in the previous game is there an increased chance that a player will lose his own service?
4. At game point does the server take less risk?
5. After a double fault do most players make sure their next first service is in?
6. Is an ace worth more than just one point?
7. After a player wins a set is there an increased chance that he will win the first game of the next set?
8. Does winning the penultimate set provide an advantage in the final set?

6.2 Magnus and Klaassen

The process that Magnus and Klaassen used for their analysis was not entirely clear. It appeared they averaged over the entire data set by constructing the service characteristics based on a single match. This found, for example the average percent of aces served in a match by a male singles player. The percentage of aces was constructed as a ratio of the number of aces in the first and second service to the number of points served. The limitation to this approach was that they averaged over players with different abilities. To overcome this they classified the matches according to seeds and computed separate averages for each of the four categories.

- Sd-Sd - seeded against a seeded player
- Sd-NSd - seeded against a non-seeded player
- NSd-Sd - non-seeded against a seeded player
- NSd-NSd - non-seeded against a non-seeded player

By grouping matches by seeding the variability between matches should decrease because matches involving players of the same seeding combination would be expected to have similar differences in playing ability.

Magnus and Klaassen may have used a test such as Cochran's Mantel Haenszel test to stratify by seeding and then perform 2x2 contingency tests for independence between the marginal rows and columns. However, their analysis did not state any p-values when they confirmed a hypothesis was statistically significant, only percentages and standard errors. This could mean that they constructed confidence intervals for the probability of, say, serving an ace following a double fault and then compared them with the constant probability of serving an ace.

6.3 Fitting Models

The approach taken in this thesis is different from what we believe Magnus and Klaassen adopted. The probability of winning a point, game or set was assumed to be constant for a pair of players in each match. By adopting models that allow for match to match variability in this probability, we allow for the variation in skill of the players.

In essence, we are adjusting for the comparative skill of each pairing of players, and all comparisons are made within pairs or matches rather than between pairs or matches.

A binomial GLM was fitted using two different approaches. Both approaches takes into account match to match, or player to player, variability, just in different ways. Firstly, a GLM was fitted assuming the effects due to

matches, or players were fixed. Secondly, a random effects model was fitted assuming the effects due to matches were random. There was no real justification for assuming the matches were random, and normally distributed, because after the first round the players were deciding who they played by winning the match in the previous round. Also the first round was established so that the first and second seeded players, that is the best two players in the tournament, did not meet before the final. This means it is possible the results found in this thesis may not hold on a different data set even if the same analysis was performed.

Player to player variability was accounted for when making comparisons within a single player, i.e. a players probability of serving an ace at game point. Match to match variability was accounted for when making comparisons within a pair of players, i.e. a players probability of winning a match.

6.4 Player Probabilities

When testing whether there was an increased chance of winning a point based on an event in a tennis match it was important to only compare probabilities of the same person or pair. For example, if the test was to determine the chance of serving an ace at break point, only the probability for the person serving needed to be considered. If the test was to determine the chance of winning the match having won the previous set, the probabilities within a pair of players needed to be compared. The comparison would be the probability of a player winning the match after winning the penultimate set, compared with the same persons probability of winning a match after losing the penultimate set. However, this assumed each player had a fixed probability of winning the match, which was the compliment of his opponent losing the match. This created problems with symmetry when comparing probabilities within a pairs of players, which did not effect the probability of a single player. The next section attempts to deal with these problems.

6.5 Symmetry

It is important to recognise that a match involves a pair of players, and that a win for one player means a loss for his opponent. The probability P of one player winning the match is the compliment of the probability his opponent will win. If all the players were considered the same probabilities would be compared twice. Thus the basic unit of analysis is the match, or pair of players, not the individual.

Consider the probability of winning a point. Due to the effect of symmetry within a pair of players, only one player's probability of winning needs to be compared, the other player's probability would be the compliment of this. Which player of the pair is analysed is largely arbitrary, provided the choice is consistent. Typically, either the first server in the match, or the match winner is chosen and the choice will depend on the hypothesis being tested. If the hypothesis being tested was the probability of winning the match having won the penultimate set, then the probability should be modelled in terms of the first or second server. Modelling the probability of winning the match based on the match winner, results in non sensible comparisons, as this player must always win the match. For hypotheses not testing the probability of winning the match, the match winner should be used as they are considered the better player and consequently the variation in the players probabilities would be less. More variation would be expected when modelling in terms of the first server because the first server may or may not be the better player. This introduces a random selection of good and bad players and consequently high and low probabilities of a win.

Consider testing the probability of winning the match having won the penultimate set. The model is based on the first server of the match. In order to fit the model an indicator variable is used to assign

0 not a match winning point

1 match winning point for the first server

-1 match winning point against the first server

Here a particular player in each match has been chosen for the comparison. In this example the model had to be in terms of the first server because the hypothesis being tested was the probability of winning the match.

Another way to control for variability, but still holding the condition of symmetry, is to model the explanatory variable in terms of a factor. This would model the probability of winning for the match winner separately from the probability of winning against the match winner.

In this thesis, the probability of a first serve in and the probability of serving an ace was modelled in terms of the the player because symmetry did not apply to these cases. For example, if player one served an ace it did not decrease the probability of player two serving an ace. This focused on the player to player variability per match so each game was split into the first and second server for each match and their probabilities were analysed separately. First service in and aces served were tested using Chi-Squared, Mantel Haenszel, GLM (fixed effects) and Random Effects models. Services won were tested using GLM and Random Effects models because of the symmetry of winning.

6.6 Break Points

There were three hypotheses tested in this section concerning break points. If there was evidence of a 'winning mood effect' more risk would be taken in the service following a break point, that is, an increase in aces served and services won and a decrease in first services in. If a player were to break his opponents service he would be expected to gain in confidence and win his service. At the same time, the player having lost his own service may feel discouraged and lose confidence in his own ability. The ability to win your service is a big factor in winning tennis matches. If a player serves first and breaks his opponents service then he leads the set by two games. If the player then wins his service he will have a lead of three games on his opponent.

Hypothesis 1: At break point does the server take less risk?

Hypothesis 1 looks at first serves in, aces served and services won at break point. A player may be inclined to take less risk on a break point, as they do not want to lose their service. The results for hypothesis 1 can be found in Table 6.1.

Probability of...	**Model**	**Lower Limit**	**Upper Limit**
First service in	Fixed Effects	42.5	51.5
	- Change in Odds	0.87	1.02
	Random Effects	42.9	51.2
	- Change in Odds	0.87	1.02
Ace	Fixed Effects	23.7	41.9
	- Change in Odds	0.55	0.85
	Random Effects	21.9	40.1
	- Change in Odds	0.53	0.83
Winning service	Random Effects	28.1	35.1
	- Change in Odds	0.64	0.75
.. breakpt for		65.3	76.7
	- Change in Odds	1.38	1.82
.. breakpt against		30.1	39.4
	- Change in Odds	0.66	0.80

Table 6.1: *Probability of the server playing certain points at break point and the change in odds (odds below 1 indicate a decrease in the probability, odds above 1 indicate an increase in the probability).*

At break point the probability of serving an ace was between 24 and 42% using a fixed effects model. The decrease in percentage of aces served suggested that less risk was taken by the server at break point. There was evidence of a slight decrease in the chance of a player getting his first service in (ranging from 43 to 52%), but this was not significant as the player still had a 50% chance of getting his first service in.

Similar results were found when accounting for the player to player variability in the random effects model and there was still not enough evidence to conclude a decrease in the percentage of first services in at break point. The reduction in the probability of a player serving an ace and winning his service suggested, at break point, players served less aces and won less of their services, hence taking less risk for an unsuccessful result. Looking at the match winners probability of winning the service at break point for his opponent, it was still very much in favour of the non-server.

Hypothesis 2: After breaking his opponents service is there an increased chance that a player will win his own service?

Hypothesis 2 compared the effect of a player winning his service after breaking his opponent's service, thus only game points were consider. It would be expected that a player would gain confidence from breaking his opponent's service, and hence be more likely to take risks in the following game.

Probability of...	**Model**	**Lower Limit**	**Upper Limit**
Winning service	Fixed Effects	83.5	90.2
	- Change in Odds	2.27	3.00
	Random Effects	85.2	91.2
	- Change in Odds	2.41	3.19

Table 6.2: *Probability of a player winning his service after breaking his opponents service and the change in odds (odds above 1 indicate an increase in the probability).*

Using a fixed effects model the probability of a player winning his own service after breaking his opponents service was between 84 and 90%. A slightly higher probability of services won was obtained when accounting for match to match variability in the random effects model. If a player was to break his opponents service then he had an increased chance of winning his own service, which reflected an increase in the level of risk taking. This

supported evidence of a 'winning mood effect' and the increase in the players performance could have been due to gaining confidence from breaking his opponent service.

Hypothesis 3: After missing a break point in the previous game is there an increased chance that a player will lose his own service?

Hypothesis 3 only analysed games that had break points to determine the relationship between missing a break point in the previous game and the probability of winning the next service. Missing a break point is where a player who was in the position of breaking his opponent's service, i.e. was on break point, ended up losing the game. If this was weighing on the player's mind when he came out to serve in the next game it could affect his performance and cause him to lose confidence and take less risk.

Probability of...	**Model**	**Lower Limit**	**Upper Limit**
Winning service	Fixed Effects	35.1	48.5
	- Change in Odds	0.73	0.97
	Random Effects	35.9	49.2
	- Change in Odds	0.75	0.99

Table 6.3: *Probability of a player winning his service after missing a break point in the previous game and the change in odds (odds below 1 indicate a decrease in the probability).*

The results from Table 6.3, show a decrease in the probability of winning a service after missing a break point in the previous game. Therefore the probability of a break was higher after the player missed the opportunity to break his opponent's service. This found in favour of both a 'winning mood effect' and conversely a 'discouragement effect'. A 'discouragement effect' is where the player is negatively influenced by what happened in the previous game. They feel discouraged by missing the opportunity to break and are unable to concentrate enough to perform at the level required to win their

next service. Their probability of winning their service was between 35 and 49%.

6.7 Winning your service

The aim of any tennis player is to win their own service. Does this mean that at game point the server is likely to take less risk? Does a player gain the confidence from serving an ace to perform better in the proceeding points? After serving a double fault is a player so concerned about winning his own service that he takes less risk on the next service? This section was concerned with answering such questions to determine if winning a service could provide evidence of a 'winning mood effect'.

Hypothesis 4: At game point does the server take less risk?

The previous section showed that being broken or failing to take the opportunity to break decreased the chance of the player winning his next service game. If a player cannot break his opponent's service then he cannot win the match. Therefore being able to hold your own service is a big step in winning a match. Does a server then decide to be conservative in order to win his service game or does he decide to play for the big shots?

Table 6.4 suggested Hypothesis 4 was not correct for game points. The probability of a player getting his first service in at game point showed a decrease to between 44 and 50%, indicating the player took more risk in order to win the game. In taking more risk, the players probability of serving an ace increased along with his probability of winning the game. At game point the player had extra confidence in his own ability, thus taking more risk in order to win the game and consequently achieved a more successful result. If the match winner is on game point then he has a 57 to 64% chance of winning the service, however if it is game point against the match winner, then the server only has a 36 to 43% chance of winning the service.

Probability of...	**Model**	**Lower Limit**	**Upper Limit**
First service in	Fixed Effects	44.1	49.7
	- Change in Odds	0.89	1.00
	Random Effects	44.1	49.7
	- Change in Odds	0.89	1.00
Ace	Fixed Effects	50.2	60.8
	- Change in Odds	0.99	1.260
	Random Effects	50.0	60.8
	- Change in Odds	1.01	1.23
Winning service	Random Effects	58.7	63.1
	- Change in Odds	1.20	1.30
.. gamept against		35.5	42.5
	- Change in Odds	0.74	0.87
.. gamept for		57.4	63.9
	- Change in Odds	1.17	1.32

Table 6.4: *Probability of the server playing certain points at game point and the change in odds (odds below 1 indicate a decrease in the probability, odds above 1 indicate an increase in the probability).*

Hypothesis 5: After a double fault do most players make sure their next first service is in?

Generally, a player serves a double fault through taking more risk and trying to place his service out of the reach of his opponent. After a player has served a double fault does he then decide that he has to be cautious and make sure his next first service is in?

Table 6.5 suggested Hypothesis 5 was not true. In fact there appeared to be some evidence that a player took more risk after serving a double fault. There was evidence that the probability of a player getting his first serve in decreased following a double fault, but not significantly. After serving a double fault the player had between 44 and 51% chance of getting his first

Probability of...	Model	Lower Limit	Upper Limit
First service in	Fixed Effects	44.4	51.9
	- Change in Odds	0.90	1.05
	Random Effects	43.5	50.5
	- Change in Odds	0.87	1.02

Table 6.5: *Probability of a player getting his first service in after serving a double fault and the change in odds (odds below 1 indicate a decrease in the probability, odds above 1 indicate an increase in the probability).*

service in.

Hypothesis 6: Is an ace worth more than just one point?

Serving an ace could potentially have a big effect on a player's confidence because his opponent was not able to make a play at the ball. Does this mean that an ace could be worth more than just one point?

Probability of...	Model	Lower Limit	Upper Limit
Winning service	Fixed Effects	62.5	71.3
	- Change in Odds	1.30	1.58
	Random Effects	63.2	71.9
	- Change in Odds	1.31	1.60

Table 6.6: *Probability of a player winning his service after serving an ace and the change in odds (odds above 1 indicate an increase in the probability).*

After serving an ace the probability of winning the next service increased significantly, see Table 6.6. If a player served an ace then he had a probability of between 63 and 71% of winning his next service. This was true even after accounting for player to player variability. After serving an ace a player gained in confidence giving him an increased chance of winning his next service, supporting evidence of a 'winning mood effect'.

6.8 Winning a Set

Winning a set potentially gets a player closer to winning the match. In males singles matches a player must win three sets to win the match. Does winning a set give the player confidence to perform better in the next set? Can winning the final set be determined by the winner of the penultimate set?

Hypothesis 7: After a player wins a set is there an increased chance that he will win the first game of the next set?

The results for hypothesis 7, assuming the probability of winning the first game of the set was dependent on the match and who won the previous set, can be found in Table 6.7. Only game points needed to be analysed as the hypothesis was testing who won the first game of the set.

The fixed effects model showed evidence to suggest that winning the previous set actually decreased a players chance of winning the first game of the next set. This went against the idea of a 'winning mood effect', suggesting that if a player won the previous set then his chance of winning the first game of the next set was between 24 and 43%. However, having lost the previous set there was insufficient evidence to suggest that a player would win the first game of the next set.

After accounting for variability due to matches in the random effects model, there was still evidence of a decrease in the probability of winning the first game in a set having won the previous set. If a player won the set then his probability of winning the first game of the next set was between 28 and 49%. If a player lost the previous set there was evidence of an increased chance of winning the match but not significantly (36 to 76%). At first glance this appeared to provide evidence against a 'winning mood effect', however winning the previous set could still give the player a 50% chance of winning the match and losing the previous set did not guarantee the player won the match. Also most players would probably win a set on their own service,

Probability of...	Model	Lower Limit	Upper Limit
Winning game	Fixed Effects	23.7	43.5
	- Change in Odds	0.55	0.87
	Random Effects	28.1	49.0
	- Change in Odds	0.63	0.98
.. lost previous	Fixed Effects	40.8	82.0
	- Change in Odds	0.88	2.14
	Random Effects	36.3	76.4
	- Change in Odds	0.75	1.80
.. won previous	Fixed Effects	21.3	44.8
	- Change in Odds	0.52	0.91
	Random Effects	24.8	49.5
	- Change in Odds	0.58	0.98

Table 6.7: *Probability of a player winning the first game in the set after winning the previous set and the change in odds (odds below 1 indicate a decrease in the probability, odds above 1 indicate an increase in the probability).*

which meant they would have to break their opponents service in order to win the first game of the next set.

Hypothesis 8: Does winning the penultimate set provide an advantage in the final set?

Hypothesis 8 looked at whether winning the penultimate set gave the player enough confidence to win the match.

The results in Table 6.8 suggested the probability of winning the match after winning the penultimate set was between 55 and 94% for a GLM. This provided evidence that winning the penultimate set gave the player an advantage in the final set. There was evidence of a decreased chance of winning the match having lost the penultimate set. This was not significant as the player still had between 8 and 72% chance of winning the match.

Probability of...	**Model**	**Lower Limit**	**Upper Limit**
Winning Match	Fixed Effects	54.6	94.0
	- Change in Odds	0.14	3.78
	Random Effects	86.3	98.0
	- Change in Odds	2.29	7.61
.. lost previous	Fixed Effects	8.3	71.8
	- Change in Odds	0.31	1.54
	Random Effects	2.9	29.1
	- Change in Odds	0.18	0.60
.. won previous	Fixed Effects	58.9	99.0
	- Change in Odds	1.25	9.30
	Random Effects	88.1	99.6
	- Change in Odds	2.46	18.17

Table 6.8: *Probability of a player winning the match after winning the penultimate set and the change in odds (odds below 1 indicate a decrease in the probability, odds above 1 indicate an increase in the probability).*

Taking into account match to match variability in the random effects model produced very different results. The chance of winning the match significantly increased, with a probability between 88 and 100% if the first server won the penultimate set. However, if the first server lost the penultimate set, there was a significant decrease to between 3 and 29% of winning the match.

These findings reflect support toward a 'winning mood effect' after winning the penultimate set. Winning the penultimate set provided the player with enough confidence to be able to go on and win the match, or the player having lost the penultimate set was so discouraged that he played poorly in the final set.

Chapter 7

Predicting Win-Loss Scenarios based on Scores

This chapter is focused on using win-loss scenarios for predicting the winner of a tennis match based on some measure of 'score'. There are two main problems that are considered in this thesis. Firstly, given some initial measure of performance, can the outcome of the match be predicted? For example, given the number of games a player wins in the first two sets, can the winner of the match be determined? The second problem examines models using tournament outcomes of previous matches to infer some comparative measure of player skill, and hence predict the outcome of future matches. For example, given the outcome of matches leading to the finals, can the outcome of the finals be predicted? Several methods are applied to these two problems using simulated data to construct fictitious matches and games. The probability of winning is binomial and the predictions can take any arbitrary form for the link function, instead of one of the standard links. Thus some of the models considered have proved to be rather complex and although the idea is appropriate their implementation was not attempted at this time.

7.1 Generalised Binomial Models

We are interested in using a measure of 'score', such as past performance in a match or tournament, to predict future performance. For example, in a tennis tournament we may be interested in using the initial sets to predict the outcome of a match. If a player has won the majority of games in the first few sets of the match he is most likely the better player and consequently should have a higher chance of winning the match. Similarly early matches in a tournament could be used to predict the outcome of the tournament. If a player is winning the majority of matches easily, he should be the better player and have a better chance of winning matches in the finals.

The generalised binomial model for predicting win-loss scenarios takes the form,

$$y_i \sim \text{Bin}\left(n_i, \pi_i\right)$$

$$l\left(\pi_i\right) = \eta_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m$$

where y_i is the number of wins from n_i attempts in the i^{th} match, or game, and π_i is the probability of a win. Here l is the link function and X is the predictor score.

7.2 Bradley Terry Models

The Bradley Terry model takes a similar approach to the generalised binomial model provided in the previous section but instead of basing the probability of winning on a single predictor score, it bases it on unknown values of the player's ability.

The Bradley Terry model for predicting win-loss scenarios takes the form

$$y_{ij} \sim \text{Bin}\left(n_i, \pi_{ij}\right)$$

$$l\left(\pi_{ij}\right) = \lambda_i - \lambda_j = q_{ij}.$$

Here λ_i and λ_j represent the ability of player i and player j respectively, and the probability of a win is related to the difference in abilities q_{ij}.

The only constraint on the link function l is that it be monotonic, thus for any two additional players s and t that play against each other,

$$q_{ij} \geq q_{st}$$

if and only if

$$\pi_{ij} \geq \pi_{st}.$$

Bradley Terry models are used in situations that involve two players being compared at the same time to determine a preference over a rating period such as a tennis tournament. Player i is preferred to player j with probability $\lambda_i/(\lambda_i + \lambda_j)$.

7.3 Arbitrary Link Function

For a binomial GLM there are three standard link functions

Logit

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)$$

Probit

$$\Phi^{-1}(\pi_i)$$

where Φ^{-1} is the standard normal distribution

Complementary log log

$$\log\left\{-\log\left(1-\pi_i\right)\right\}.$$

The logit link is particularly popular because parameters can be interpreted as log odds. The probit and logit link functions are very similar in that they both assume the probability distribution is symmetric,

$$l(\pi_i) = -l(1-\pi_i)$$

and take the shape similar to Figure 7.1. The complementary log log link function is a more flexible function as it is asymmetric.

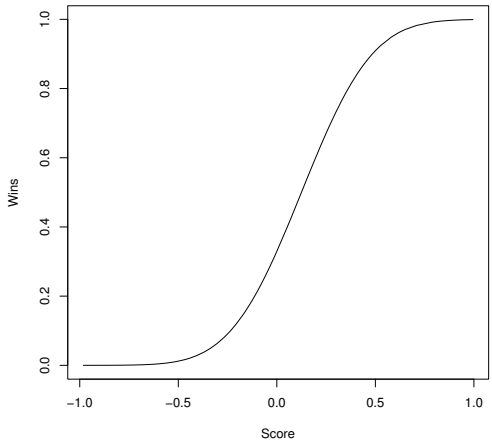

Figure 7.1: *Probability of winning against score assuming a Probit link.*

The link function provides the relationship between the score and the final probability of winning, however this relationship may not be well represented by any of the three standard links. The choice of the link function is essentially arbitrary and an improvement on the fit can be made by allowing for some flexibility in its choice. The only constraint that is placed on the link function in this thesis is the same constraint that applied to the Bradley Terry model in the previous section, it must be a monotone function, so that increasing the score implies an increase in the probability of winning.

Consider the case of predicting the winner of a match based on earlier games in the match. Assuming the distribution of winning the match was not symmetric and flattens out when two players have about the same chance of winning the match, then the function of probability would not adopt a standard probit or logit distribution. Two players may have a 50% chance of winning the match when they have both won a certain amount of games and then once a player reaches a certain threshold of games won his probability of winning may increase dramatically.

The idea of choosing an arbitrary link function is similar to that of Generalised Additive Models, which are discussed in the next section. In the

generalised binomial model case we are estimating an arbitrary link function, whereas the generalised additive model estimates an arbitrary transformation of each prediction, or in our case 'score'.

7.4 Generalised Additive Models

Generalised additive models (GAM) are another application of monotone function estimation. GAM allows for additive binomial regression as it recognises the form of the response and will not give predicted values below zero or above one. GAM is fitted by backfitting, subtracting from the log-likelihood function a penalty function that increases as the fitted function becomes less smooth. The standard GAM for predicting win-loss scenarios takes the form,

$$y_i \sim \text{Bin}\left(n_i, \pi_i\right)$$

$$l\left(\pi_i\right) = F_1\left(x_1\right) + \ldots + F_m\left(x_m\right)$$

where the link function l is fixed and the F_i's are arbitrary smooth functions to be estimated.

The backfitting algorithm is a general algorithm for fitting an additive model using any regression fitting mechanisms. The backfitting algorithm starts by fitting a basic model, extracting the partial residuals and smoothing over them. These partial residuals are then used to fit the new model.

In general the backfitting algorithm takes the form,

1. Initialise Z estimates corresponding to the predictor variables and fit the model using the Z estimates.

2. Smooth the covariates one at a time by smoothing the partial residuals

3. Refit the model using the new Z estimates

4. Repeat steps 2- 4 until convergence

In this thesis we smooth assuming known link function l,

$$l(\pi) = F(\eta)$$

to estimate the smooth function F. This is equivalent to a GLM with link

$$F^{-1}l(\pi) = \tilde{l}(\pi) = \eta$$

where $\tilde{l}$ is a generic link function, not necessarily monotonic and

$$\pi = \left(l^{-1}F\right)(\eta).$$

Backfitting algorithm

Firstly, the $Z_i^0 = \eta_i$ are initialised and then the backfitting algorithm for a binomial GLM takes the form,

1. Fit a Binomial GLM with the current Z estimates and extract the coefficients, $\hat{\beta}_i$, residuals, r_i, and weights, w_i, of the fit.
2. The covariates are smoothed one at a time by
   (a) forming the partial residuals $pres_i$ for each covariate
   $$pres_i = r_i + \hat{\beta}_i * Z_i^k$$
   (b) form $Z_i^{(k+1)}$ by smoothing the partial residuals.
   $$Z_i^{(k+1)} = sm(pres_i)$$
3. The new Z estimates are used to refit the model
4. Steps 2 - 4 are repeated until the model converges

Later in this chapter a GAM is applied to the Bradley Terry Model by smoothing over the link function instead of the normal predictor function, the difference in player's ability. GAM and the backfitting algorithm are explained in more detail by Hastie & Tibshirani (1990).

7.5 Prediction based on Score

The models in this section are evaluated using simulated data. Fictitious scores are simulated for a specified number of games in order to test the models. The models discussed in this section are isotonic regression, truncated beta deviates, monotone regression splines, penalty splines and probit regression. Monotone regression splines and penalty splines are associated with smoothing the function using a penalty for roughness. Isotonic regression and truncated beta deviates assume a generalised link function to be monotonic under an order constraint.

Given observations $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$, such that

$$y_i \sim \text{Bin}\,(n_i, \pi_i)$$

$$l\,(\pi_i) = x_i.$$

Without loss of generality we assume

$$x_i \leq x_{i+1} \qquad \text{for all } i.$$

Assuming l is a monotonic increasing function, this implies

$$\pi_i \leq \pi_{i+1}$$

and the problem reduces to estimating the π_i's subject to this ordering constraint.

7.5.1 Isotonic Regression

Isotonic regression is one approach that can be used to predict the winner of a match based on a score during the match. Consider two players, i and j. The winner of player i vs player j can be modelled as binomially distributed with a probability of π_{ij}. Then the probability that player i beats player j takes the form,

$$F\,(\pi_{ij}) = S_{ij} \tag{7.1}$$

where F is constrained to a monotone increasing function and S_{ij} is the score between player i and player j at a fixed time during the match.

In isotonic regression, a probit or logit function of the probability is not assumed, but any arbitrary function that is constrained to being monotone increasing. Isotonic regression plots the scores from a match between two players and fits a line using the lower convex hull. This line is then used to estimate the winner of a match, set or game. As can be seen in Figure 7.2 it is not symmetric, as in Figure 7.1, and is therefore able to capture the different probabilities of winning the match as the differences in score vary.

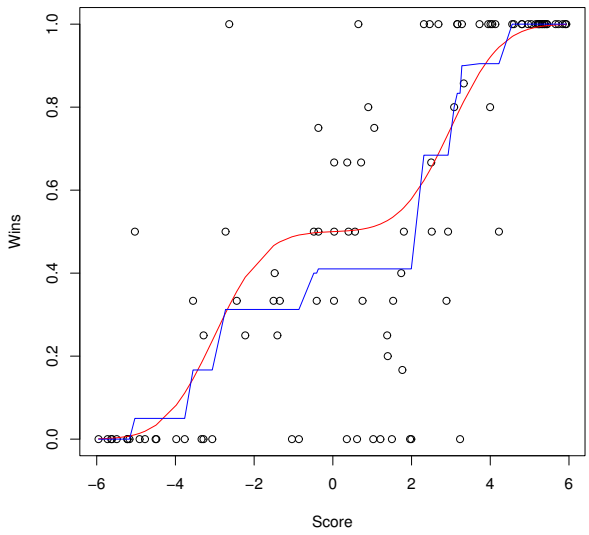

Figure 7.2: *Probability of winning against score using Isotonic Regression.*

The line in Figure 7.2 was fitted to 100 simulated data points and appears to have captured the rough shape but there are many discontinuities. By increasing the number of data points the line becomes a closer fit to the model, suggesting this method works best for very large data sets. Although isotonic regression provides optimal flexibility and computational ease in estimating monotone functions, it can be seen from Figure 7.2 that it produces a non-differentiable step function that may not be visually appealing.

Where variable values are naturally ordered monotonically, isotonic regression can be viewed as a particular case of monotone splines. The piecewise linear monotone function that is produced by this model corresponds to monotone regression splines defined with degree one and a knot positioned at every data point.

7.5.2 Convex Hull

Isotonic regression requires the minimisation of the negative log likelihood

$$J = -\sum_{i=1}^{n} \left(y_i \log \pi_i + (n_i - y_i) \log \left(1 - \pi_i\right) \right) \tag{7.2}$$

with respect to π_i, subject to the order constraints

$$\pi_i < \pi_{i+1}.$$

In matrix form we may write the constraint as

$$A\pi \leq 0$$

and A takes the form

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 & ... & 0 \\ 0 & 1 & -1 & 0 & ... & 0 \\ 0 & 0 & 1 & -1 & ... & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & ... & 1 & -1 \end{pmatrix}$$

Introducing a vector s of slack variables removes the inequality constraint,

$$A\pi + s = 0$$

where $s_i \geq 0$. If $s_i = 0$ the i^{th} constraint is said to be active.

The Kuhn-Tucker conditions (Nash & Sofer, 1996) imply that at the minimum

$$\nabla J = -A^T \alpha \tag{7.3}$$

where for all i, $\alpha_i \geq 0$ and $\alpha_i s_i = 0$. The α's are called the dual variables and are the coefficients in a convex linear combination of normals defined by the active constraints. Applying the Kuhn-Tucker conditions to equation 7.2 yields the system of equations

$$\begin{aligned}
\frac{n_1\pi_1 - y_1}{\pi_1(1-\pi_1)} &= -\alpha_1 \\
\frac{n_2\pi_2 - y_2}{\pi_2(1-\pi_2)} &= -\alpha_2 + \alpha_1 \\
&\vdots \\
\frac{n_n\pi_n - y_n}{\pi_n(1-\pi_n)} &= -\alpha_n + \alpha_{n-1}.
\end{aligned}$$

If these equations are summed, each successive α_i cancels to yield

$$\sum_{i=1}^{j} \left(\frac{n_i\pi_i - y_i}{\pi_i(1-\pi_i)} \right) = -\alpha_j.$$

Since $\alpha_j \geq 0$, this implies that for all j

$$\sum_{i=1}^{j} \frac{n_i\pi_i}{\pi_i(1-\pi_i)} \leq \sum_{i=1}^{j} \frac{y_i}{\pi_i(1-\pi_i)}.$$

Moreover, $\alpha_j > 0$ implies $s_j = 0$ and $\pi_{j-1} = \pi_j$. Combined, these conditions imply that the π_i's are the slopes of the segments of the lower convex hull to the polygonal path with vertices

$$\left(\sum_{i=1}^{j} n_i, \sum_{i=1}^{j} y_i \right).$$

7.5.3 Truncated Beta Deviates

Consider a Bayesian approach to the isotonic regression problem. We wish to estimate $\pi_1, \ldots, \pi_n$ given observations $y_1, \ldots, y_n$ such that

$$y_i \sim \text{Bin}(n_i, \pi_i)$$

$$\pi_i \leq \pi_{i+1}.$$

The likelihood for this problem takes the form

$$p\left(y|\pi\right) = \left[\prod_{i=1}^{n} \binom{n_i}{y_i} \pi^{y_i} \left(1-\pi\right)^{(n_i - y_i)}\right] I\left(\pi_1 < \pi_2 < \cdots < \pi_n\right)$$

where I is the characteristic function

$$I\left(A\right) = \begin{cases} 1 & \text{if A,} \\ 0 & \text{if not A.} \end{cases}$$

Adopting conjugate Beta priors

$$\pi_i \sim \text{Beta}\left(\alpha_i, \beta_i\right)$$

and the full posterior is proportional to

$$p\left(\pi|y\right) \propto \left[\prod_{i=1}^{n} \pi^{(y_i + \alpha_i - 1)} \left(1-\pi\right)^{(n_i - y_i + \beta_i - 1)}\right] I\left(\pi_1 < \pi_2 < \cdots < \pi_n\right).$$

So the full conditional distribution for the π_i's are

$$\begin{aligned} \pi_1|\pi_{j\neq 1}, y &\sim \text{Beta}\left(y_1 + \alpha_1, n_1 - y_1 + \beta_1\right) I\left(\pi_1 \leq \pi_2\right) \\ \pi_i|\pi_{j\neq i}, y &\sim \text{Beta}\left(y_i + \alpha_i, n_i - y_i + \beta_i\right) I\left(\pi_{i-1} \leq \pi_i \leq \pi_{i+1}\right) \\ \pi_n|\pi_{j\neq n}, y &\sim \text{Beta}\left(y_n + \alpha_n, n_n - y_n + \beta_n\right) I\left(\pi_{n-1} \leq \pi_n\right) \end{aligned} \tag{7.4}$$

where here I indicates a truncation of the distribution in the obvious manner. These lead immediately to a Gibb's sampling scheme for the π_i's, which are updated by successively drawing from equation 7.5.

Figure 7.3 plots the median probability as well as the 2.5 and 97.5 % percentiles against the difference in score. The truncated beta deviate is unable to capture the curvature of the probability distribution, even after providing curvature prior to sampling. By increasing the mean number of games per score the truncated beta deviate appears to capture the curvature of the distribution better, see Figure 7.4.

7.5.4 Monotone Regression Splines

Like isotonic regression, monotone regression splines can be used to define the binomial regression function that models the probability of success in a

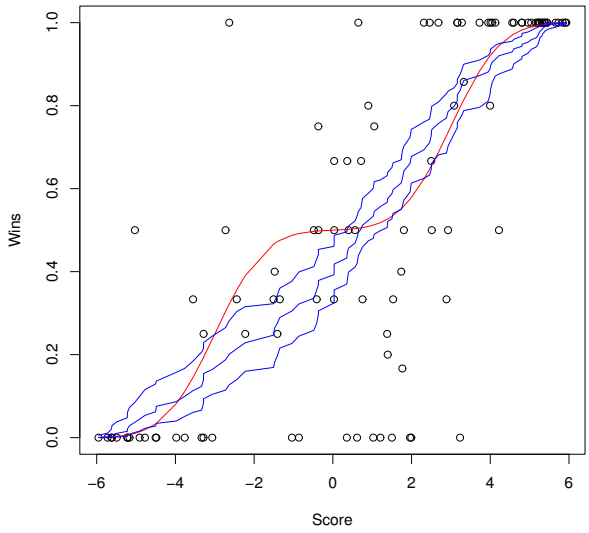

Figure 7.3: *Probability of winning against score using Truncated Beta Deviates.*

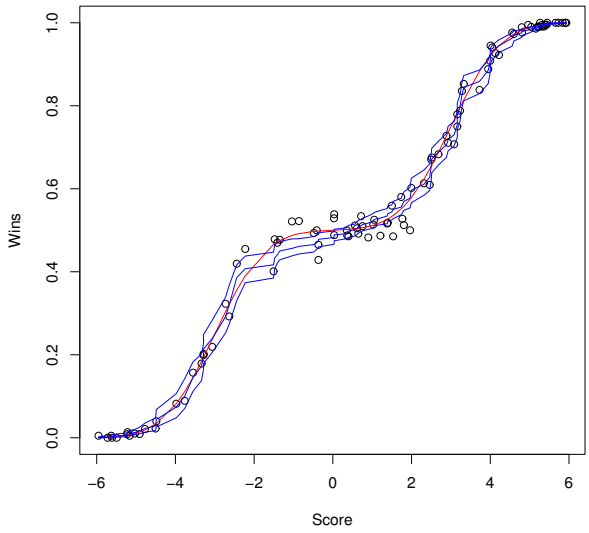

Figure 7.4: *Probability of winning against score using Truncated Beta Deviates allowing for more games per score.*

certain number of trials. The flexibility of monotone regression splines, as discussed by Ramsay (1988), allows the detection of characteristics that may be associated with the score of a game, or match, that are not observable using probit or logit based approaches. Ramsay (1988) showed that the flexibility is retained by imposing non-negativity on the parameters which define the spline function, thus constraining it to be non-decreasing.

Monotone regression splines are constructed by the partial integration of M-splines of order k using non-negative linear combinations. These partial integrals are called I-Splines and are particularly useful in binomial models as they map a function to $[0, 1]$. An M-spline is a non-negative piecewise polynomial of degree $k - 1$ and the associated I-Spline is a monotone non-decreasing piecewise polynomial of degree k. The derivation of M-splines and I-splines is not discussed in this thesis, but Ramsay (1988) explains in great detail how monotone regression splines can be constructed using M-splines and I-splines.

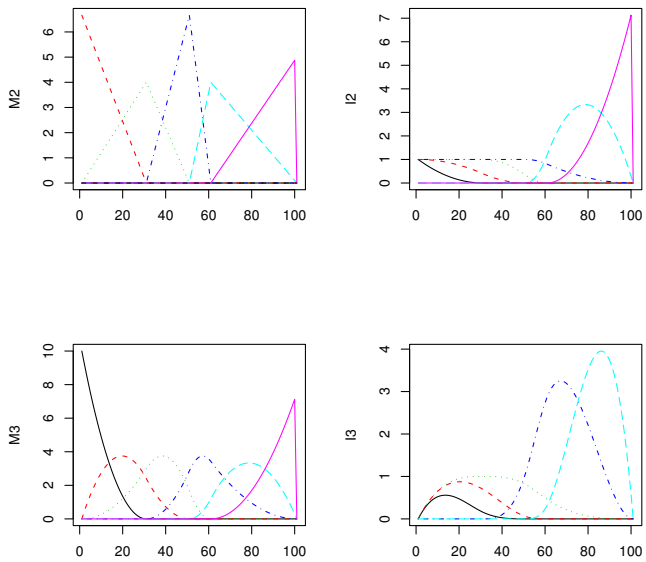

Figure 7.5: *M-Splines and associated I-Splines of order 2 and 3 respectively.*

As can be seen in Figure 7.5, a spline with order two consists of straight-line segments required to match at boundaries and a spline with order three is a piecewise quadratic with matching first derivatives. Associated with this set of basis splines is a knot sequence, t, where the linear combination of any basis spline will produce any other monotone regression spline associated with t.

When using splines the number of knots when fitting the spline can be reduced to allow for smoothing. Fitting one knot per observation will fit the data perfectly, but by increasing the number of observations per knot to ten the data will fit smoother. Another method, shown in Section 7.5.5, is penalty splines which applies a form of shrinkage, or smoothing penalty, to the sums of squares.

Monotone regression splines were not implemented in this thesis because of the difficulty of constraining the function F to be monotonic. For the sum

$$F = \sum \alpha_i f_i(x)$$

to be monotonic we must constrain

$$\alpha_i \geq 0.$$

In general, this is a difficult problem due to the potential for local maxima, and has not been considered further.

7.5.5 Penalty Splines

In the penalty spline approach, the link function is represented by a spline, and a penalty term is added to the log likelihood to penalise the link function for being too rough. That is, P is chosen to minimise the functional equation

$$\begin{aligned} J(P) = &- \sum_i \left(y_i \log P(x_i) + (n_i - y_i) \log\left(1 - P(x_i)\right)\right) \\ &+ \lambda \int_{-\infty}^{\infty} S(P)(x)dx \end{aligned}$$

where P is constrained to be monotonic. Here P represents an arbitrary monotonic link and the first term is the negative log likelihood and the second

term is a roughness penalty on P. For large λ the roughness penalty will dominate and P will be smooth at the expense of the likelihood. If λ is small the likelihood term will dominate and less penalty will be applied to the roughness function.

To constrain P to be monotonic, we represent P in the form

$$P\left(x_i\right) = \int_{-\infty}^{x_i} f\left(t\right)^2 dt$$

For the roughness penalty we choose

$$S\left(P(x)\right) = f\left(t\right)''^2$$

so that

$$\begin{aligned} J(f) = &- \sum_i \left(y_i \log \int_{-\infty}^{x_i} f\left(t\right)^2 dt + \left(n_i - y_i\right) \log \left(1 - \int_{-\infty}^{x_i} f\left(t\right)^2 dt \right) \right) \\ &+ \lambda \int_{-\infty}^{\infty} f''\left(t\right)^2 dt \end{aligned} \tag{7.5}$$

We can determine the f that minimises J through the calculus of variations.

Let f_0 be the function that minimises J, and write $f(t) = f_0(t) + \epsilon f_1(t)$ where $\epsilon \in R$ and f_1 is an arbitrary function. Since f_0 minimises J,

$$\left. \frac{dJ\left(f_0(t) + \epsilon f_1(t)\right)}{d\epsilon} \right|_{\epsilon=0} = 0.$$

Substituting from equation 7.6, we find

$$\begin{aligned} &\frac{dJ\left(f_0 + \epsilon f_1\right)}{d\epsilon} = \\ &\quad - \sum_i \left(\frac{y_i \int_{-\infty}^{x_i} 2 f_1 \left(f_0 + \epsilon f_1\right) dt}{\int_{-\infty}^{x_i} \left(f_0 + \epsilon f_1\right)^2 dt} - \frac{\left(n_i - y_i\right) \int_{-\infty}^{x_i} 2 f_1 \left(f_0 + \epsilon f_1\right) dt}{1 - \int_{-\infty}^{x_i} \left(f_0 + \epsilon f_1\right)^2 dt} \right) \\ &\qquad + \lambda \int_{-\infty}^{\infty} 2 f_1'' \left(f_0'' + \epsilon f_1''\right) dt \end{aligned}$$

Setting $\epsilon = 0$

$$\begin{aligned}\left.\frac{dJ(f)}{d\epsilon}\right|_{\epsilon=0} &= -\sum_i \left(\frac{y_i \int_{-\infty}^{x_i} 2f_1 f_0 dt}{\int_{-\infty}^{x_i} (f_0)^2 dt} - \frac{(n_i - y_i) \int_{-\infty}^{x_i} 2f_1 f_0 dt}{1 - \int_{-\infty}^{x_i} (f_0)^2 dt} \right) \\ &\quad + 2\lambda \int_{-\infty}^{\infty} f_1'' f_0'' dt \\ &= -2 \sum_i \left(\int_{-\infty}^{x_i} f_1 f_0 dt \, \frac{y_i - n_i \int_{-\infty}^{x_i} (f_0)^2 dt}{\int_{-\infty}^{x_i} (f_0)^2 dt \left(1 - \int_{-\infty}^{x_i} (f_0)^2 dt\right)} \right) \\ &\quad + 2\lambda \int_{-\infty}^{\infty} f_1'' f_0'' dt.\end{aligned}$$

Integrating by parts twice in the second term to eliminate f_1'' yields

$$\begin{aligned}\left.\frac{dJ(f)}{d\epsilon}\right|_{\epsilon=0} &= -2 \sum_i \left(\int_{-\infty}^{x_i} f_1 f_0 dt \, \frac{y_i - n_i \int_{-\infty}^{x_i} (f_0)^2 dt}{\int_{-\infty}^{x_i} (f_0)^2 dt \left(1 - \int_{-\infty}^{x_i} (f_0)^2 dt\right)} \right) \\ &\quad + 2\lambda \left([f_1' f_0'']_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f_1' f_0''' dt \right) \\ &= -2 \sum_i \left(\int_{-\infty}^{x_i} f_1 f_0 dt \, \frac{y_i - n_i \int_{-\infty}^{x_i} (f_0)^2 dt}{\int_{-\infty}^{x_i} (f_0)^2 dt \left(1 - \int_{-\infty}^{x_i} (f_0)^2 dt\right)} \right) \\ &\quad + 2\lambda \left([f_1' f_0'' - f_1 f_0''']_{-\infty}^{\infty} + \int_{-\infty}^{\infty} f_1 f_0^{(4)} dt \right).\end{aligned}$$

Outside the support of the data we assume $f_0'' = 0$ and $f_0''' = 0$. Introducing the Dirac delta function $\delta(x)$,

$$\left.\frac{dJ(f)}{d\epsilon}\right|_{\epsilon=0} =$$
$$2 \int_{-\infty}^{\infty} \left(\lambda f_0^{(4)} - \sum_i f_0 \left[\frac{y_i - n_i \int_{-\infty}^{x} (f_0)^2 dt}{\int_{-\infty}^{x} (f_0)^2 dt \left(1 - \int_{-\infty}^{x} (f_0)^2 dt\right)} \right] \delta(x - x_i) \right) f_1 dx.$$

Since this is true for arbitrary f_1, we must have

$$\lambda f_0^{(4)} = \sum_i f_0 \left[\frac{y_i - n_i \int_{-\infty}^{x} (f_0)^2 dt}{\int_{-\infty}^{x} (f_0)^2 dt \left(1 - \int_{-\infty}^{x} (f_0)^2 dt\right)} \right] \delta(x - x_i). \tag{7.6}$$

This expression can be immediately integrated once to yield

$$\lambda f_0''' = \sum_i f_0 \left[\frac{y_i - n_i \int_{-\infty}^{x} (f_0)^2 \, dt}{\int_{-\infty}^{x} (f_0)^2 \, dt \left(1 - \int_{-\infty}^{x} (f_0)^2 \, dt\right)} \right] H(x - x_i)$$

where $H(x)$ is the Heavy side-step function.

The presence of the Dirac delta function allows us to deduce that f_0 is a spline with knots at the observations. In principle f could be calculated from the differential equation 7.8. In practice this problem proves to be highly numerically unstable and has not been followed through.

Alternatively P could be represented as

$$p = \int_{-\infty}^{x} e^{f(t)} dt.$$

This choice leads to a similar calculation for f_0, but again, the practical problem is found to be numerically unstable.

7.5.6 Probit Regression

Generally probit regression is constrained to being symmetric, however, by modelling a mixture of probits a non-symmetric approach to estimation can occur. Probit regression takes a Bayesian approach to density estimation where a Markov Chain Monte Carlo (MCMC) algorithm, such as the Gibbs Sampler (mentioned earlier), can be used to simulate the posterior distribution for the beta's from a prior measure. This method gives a sample approximately drawn from the posterior density.

The method of probit regression used in this thesis looked at a mixture maximum likelihood estimation that fitted a mixture of normal distributions using a GLM with a probit link.

$$y_i \sim \text{Bin}(n_i, \pi_i)$$

$$\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m$$

which is a binomial regression model with a probit link. Let

$$y_i = \begin{cases} 1 & \text{if } Z \leq Z_0 \qquad \Rightarrow \text{win}, \\ 0 & \text{if } Z > Z_0 \qquad \Rightarrow \text{loss}. \end{cases}$$

In order to gibb's sample from this a latent variable is created that does not depend on the y's.

$$Z_i' \sim N(0,1)$$

and

$$\begin{aligned} Z_i' &\leq Z_{0i} \qquad \text{if } y_i = 1 \\ Z_i' &> Z_{0i} \qquad \text{if } y_i = 0 \end{aligned}$$

Equivalently

$$Z_i \sim N(Z_{0i}, 1)$$

or

$$Z_i \sim N\left((XB)_i, 1\right)$$

and

$$\begin{aligned} Z_i &\geq 0 \qquad \text{if } y_i = 1 \\ Z_i &< 0 \qquad \text{if } y_i = 0 \end{aligned}$$

This expression no longer depends on y, but on a threshold, Z_0. If the probability lies above the threshold Z_0 the player wins, and if the probability falls below the threshold Z_0 the player loses.

Therefore all the information that is required to gibb's sample is,

$$\begin{aligned} &Z_i' \sim N(0,1) \\ &Z_i = Z_{0i} - Z_i' \\ &Z_i \sim N\left((XB)_i, 1\right) \\ &Z_i' \leq Z_{0i} \Rightarrow Z_i \geq 0 \\ &Z_i' > Z_{0i} \Rightarrow Z_i < 0 \end{aligned}$$

The posterior does not need to be written out as the full conditional distributions can be immediately determined without refactoring the posterior. The problem with gibb's sampling is the full conditional distributions for every parameter has to be calculated so it is difficult to implement and if the parameters are correlated it can mix slowly. The mixture MLE was implemented using 100 simulated data points and proved to be quite slow with convergence occurring after 77 iterations.

7.6 Prediction based on Ability

The models in this section are based on the Bradley Terry model stated earlier in section 7.2 and are evaluated on simulated data, where for a fictitious competition each player competes with each other at random for a specified number of matches over a rating period. For this thesis we are assuming that a match results in only two possible cases, a win or a loss, and instead of adopting ties into the likelihood we assume a tie is halfway between a win and a loss. Firstly, a straight Bradley Terry model is applied to determine whether the difference in players ability can be used to rank players. Secondly, Bradley Terry with isotonic regression is used to determine whether differences in players ability can be used to predict the winner of a tournament. Lastly, a Bradley Terry Model with GAM is applied by smoothing over the link function instead of the difference in players ability.

7.6.1 Bradley Terry Model

The Bradley Terry model is fitted by maximising the negative log likelihood. Firstly, the player ability, λ, is simulated for each player from a normal distribution then the wins are predicted based on these λ's. Playing ability starts at zero and is updated after every game in order to predict the winner of a match.

7.6.2 Bradley Terry with Isotonic Regression

The Bradley Terry model can be extended to incorporate isotonic regression, where the winner of a tennis tournament is predicted based on a current score, i.e. the percentage of games won, after a preference between the players has been estimated from a difference in players ability using a Bradley Terry model. A Bradley Terry model with isotonic regression is applied using two different methods for selecting the players ability, Simulated Annealing and Newtons Method. The Bradley Terry with isotonic regression model takes the form,

$$\mathrm{l}\left(\pi_{ij}\right) = \lambda_i - \lambda_j$$

where l is an arbitrary link function.

For each method, abilities are simulated from a Normal distribution for 20 players, where every player plays every other player but themselves and wins are simulated using the Bradley Terry model based on a difference in ability between pairs of players. Isotonic regression then uses games for each pair of players, assumed to take a Poisson distribution, to predict the winner of the tournament. In the first method a discontinuous monotone link function is chosen to maxmise the likelihood using simulated annealing to estimate the players ability, λ, see Figure 7.6. For discontinuous data sets, simulated annealing maximises the likelihood by eliminating the process of continually moving in a downhill direction. The process of simulated annealing is explained further by Press, Flannery and Teukolsky (1992).

The second method applies Newtons Method to maximise the likelihood for selecting a players ability, this is similar to simulated annealing but is less likely to work as it assumes the link function is smooth but not continuous. For this reason the distribution can take any form, see Figure 7.7. Newtons Method is by far the worst at predicting the probability of a win and the distribution takes a discontinuous form, stretched out over the middle section. However, simulated annealing appears to give a reasonably fit to the data, but still displays the step-like characteristic of isotonic regression.

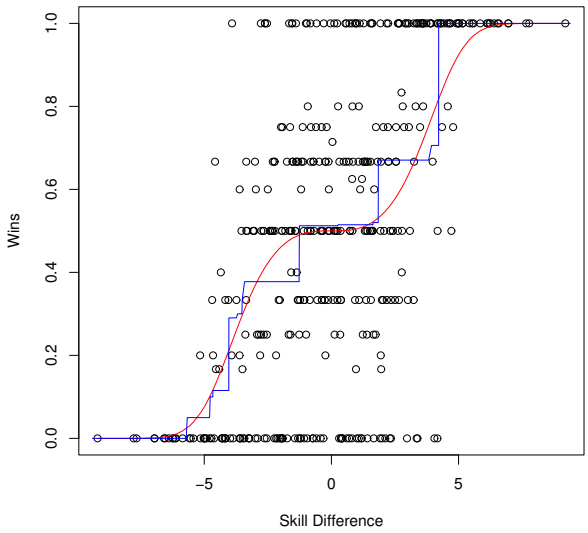

Figure 7.6: *Probability of winning a tournament using Bradley Terry with isotonic regression applying Simulated Annealing.*

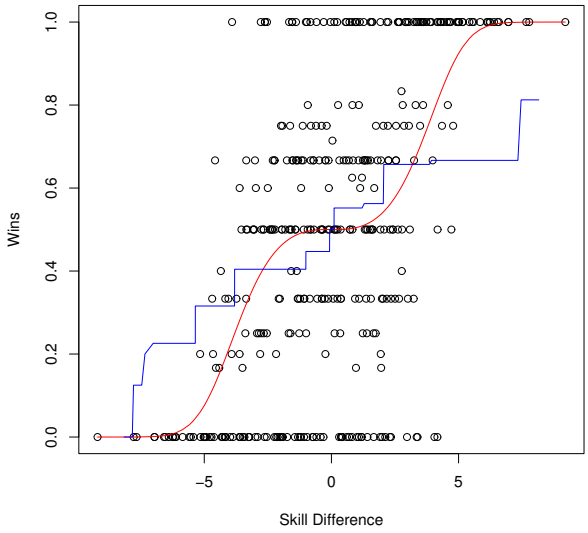

Figure 7.7: *Probability of winning a tournament using Bradley Terry with isotonic regression applying Newtons Method.*

7.6.3 Bradley Terry with GAM

The Bradley Terry model with GAM smooths over the link function rather than smoothing over each predictor variable as in the normal case. The Bradley Terry model with GAM takes the form,

$$\mathrm{l}\left(\pi_{ij}\right) = f\left(\lambda_i - \lambda_j\right)$$

where l is assumed to be a logit link function and takes the form

$$f^{-1}(\mathrm{logit}(\pi_{ij}))$$

Firstly a Bradley Terry model is fitted to the current f. Then a new f is generated using the GAM function to smooth over differences in player ability. Lastly, the new players ability is extracted as well as f and the process is repeated until convergence. To test the method, abilities were randomly generated from a Normal distribution for 20 players, where every player plays every other player but themselves and wins are simulated using the Bradley Terry model based on a difference in ability between pairs of players. GAM is then applied to the link function smoothing over the difference in player ability.

The GAM smoothing function estimates f, a function of the difference in player ability, thus is not constrained to being monotonic as can be seen in Figures 7.8 and 7.9. Figure 7.8 shows the implementation of a Bradley Terry model using the GAM function in R. This method had a convergence rate of 19 iterations. This fit is clearly not monotonic and is predicting two separate peaks when difference in player ability is closer to zero. Figure 7.9 shows the implementation of a Bradley Terry model using a self-constructed GAM, where convergence occurred after 14 iterations but is extremely slow. This method provides a relatively good fit to the data and is a lot smoother than using the GAM function in R but behaves oddly at the far ends of the distribution.

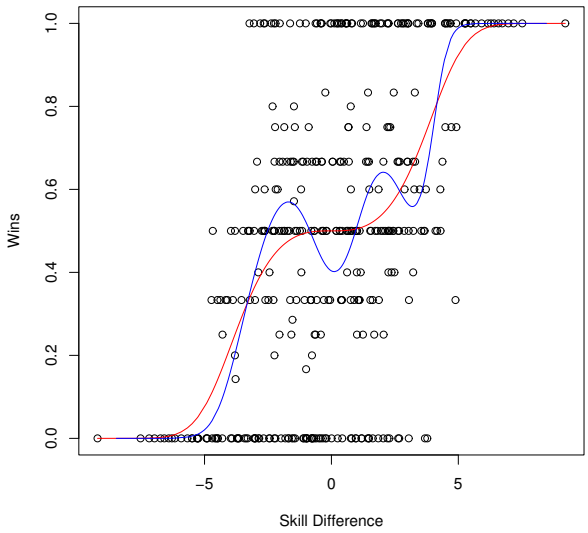

Figure 7.8: *Probability of winning a tournament using Bradley Terry and the GAM function in R.*

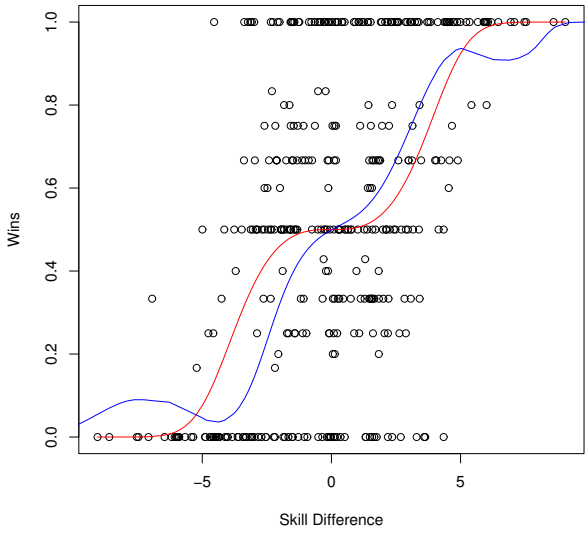

Figure 7.9: *Probability of winning a tournament using Bradley Terry and a self-constructed GAM function.*

Chapter 8

Predicting Winners at the Australian Open

This chapter looks at testing isotonic regression, which was explained in the previous chapter, on data from the 2003 Australian Open. The probability of winning the match is based on both the number of points won and the number of games won between the two players in a match for all sets except the last set.

8.1 Predicting the probability of winning

Consider predicting the probability of winning based on the number of points each player has won. Points that did not occur in the last set were used for the prediction and the sum of each players points were calculated. To compute the difference in points two methods could be used. Firstly, the difference in points between two players is compared to the winner of the match. The difference takes the form,

$$\text{Difference} = p_1 - p_2$$

where p_1 is the total number of points won by the first server prior to the final set and p_2 is the total number of points won by the second server prior to the final set.

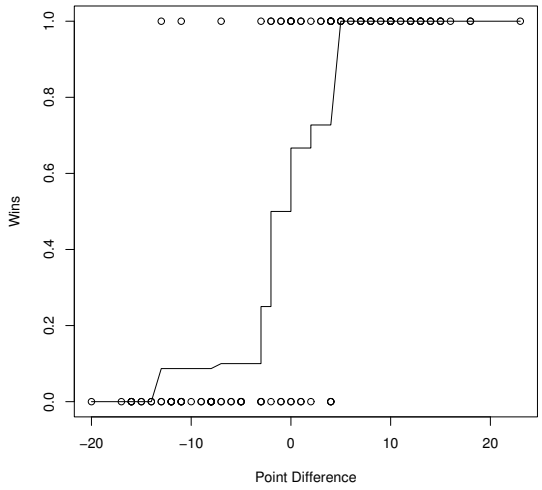

Figure 8.1: *Probability of the first server winning the match based on the difference in points won between players using isotonic regression.*

If the difference in points was positive then the person that served first at the start of the match had won the most points prior to the final set. Isotonic regression is applied using the difference in points for each player prior to the final set to predict the probability of winning the match. The second method averages the difference of points between players over the total points played and takes the form,

$$\text{Difference} = \frac{p_1 - p_2}{p_1 + p_2}$$

where p_1 is the total number of points won by the first server prior to the final set and p_2 is the total number of points won by the second server prior to the final set. The application of this method is then the same as the previous method.

Figure 8.1 shows the probability of winning the match was quite high for the player who served first and had won slightly more points in the match. The person who served second needed to win about 15 more points than the first server in order to win the match, whereas the first server only needed to win about five more. When both players had won the same amount of points

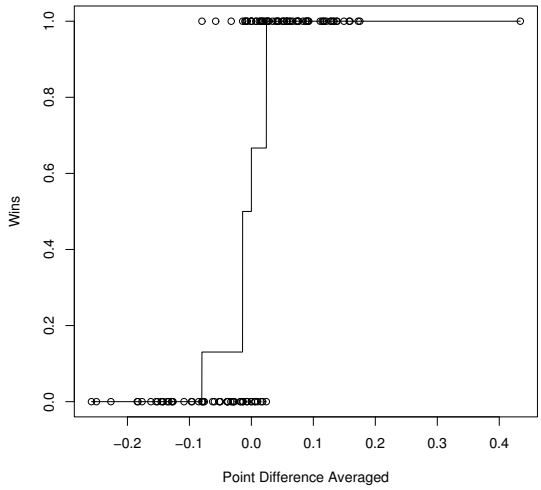

Figure 8.2: *Probability of the first server winning the match based on the averaged difference of points won between players using isotonic regression*

going into the last set the first server had over a 60% chance of winning the match.

Figure 8.2 shows the averaged difference of points between players over the total number of points played. This gives a much steeper curve and predicts that if the first server had won about 3% more of the points going into the last set then they would win the match, this was about 8% for the second server. The steepness of this curve is possibly due to the range in the difference in points won between the players. In one of the matches the first server won 20 points more than his opponent, this difference was 40% of the number of points played in the entire match. The sparcity in the range of the dataset could explain the steepness of the curve, particularly in Figure 8.2.

Table 8.1 shows the log likelihood and Akaike Information Criteria (AIC) for the probability of the first server winning the match based on both the difference in points won and the difference in games won. The averaged difference in points gives a much better fit for the model when based on

Model	**Method**	Log likelihood	AIC
Points	Diff	-30.015	296.03
	Avg Diff	-27.295	290.59
Game	Diff	-30.307	296.61
	Avg Diff	-29.33	294.67

Table 8.1: *AIC and Log Likelihood of points and games won.*

points, but only slightly better when the model is based on games.

Predicting the probability of winning the match based on games is the same procedure as for points but using the number of games won in the match, once again excluding the final set, instead of the number of points.

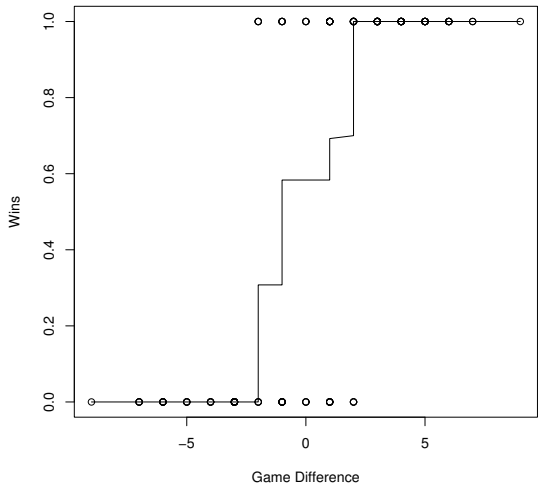

Figure 8.3: *Probability of the first server winning the match based on the difference in games won between players using isotonic regression.*

Figure 8.3 shows that if the first server has won one more game than the second server going into the final set then he has a 60% chance of winning the match, this is also true if he has won the same amount of games or even one less game than the second server. For either player to win the match

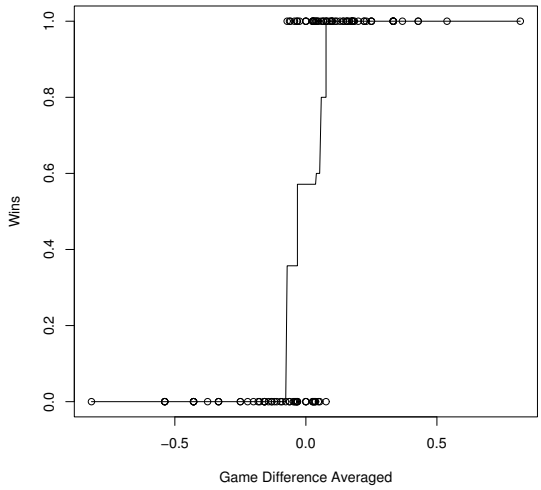

Figure 8.4: *Probability of the first server winning the match based on the averaged difference in games won between players using isotonic regression.*

they needed to win at least 3 games going into the final set. Figure 8.4, the averaged difference in games won, shows that to win the match the first server only needs to have won 0.5% more of the games prior to the final set, whereas, the second server needs to win 1% more of the games. Once again this curve is very steep and having a bigger range in data points may improve the fit.

The isotonic regression procedure was particularly fast, and so it was possible to estimate the variability in the fitted profile by bootstrapping over matches. Figures 8.5 and 8.6 plot the probability of winning the match based on the number of points won using 1000 bootstrapped samples from the original data. Similarly, Figures 8.7 and 8.8 plot the probability of winning the match based on the number of games won using 1000 bootstrapped samples.

Bootstrapping from the original data set appears to yield a similar distribution for the probability of winning a match based on both points won and games won. The first server appears to have the advantage over the second server even when they have won less points or games going into the final set.

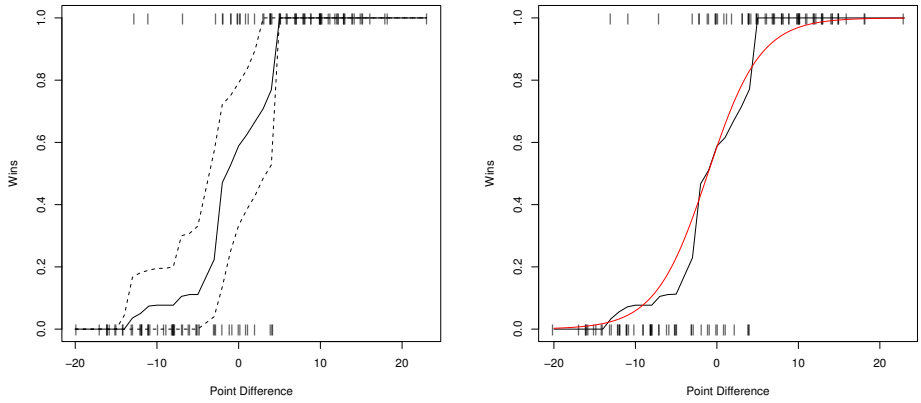

Figure 8.5: *Probability of the first server winning the match based on the difference in points won between players. The plot on the left shows the bootstrap median and pointwise 95% confidence interval for the profile fitted by isotonic regression. The plot on the right shows the profile fitted by isotonic regression and Binomial GLM with logit link.*

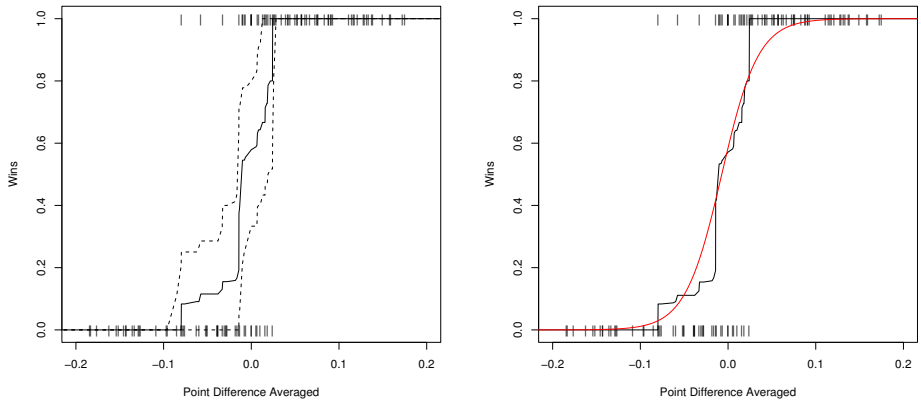

Figure 8.6: *Probability of the first server winning the match based on the averaged difference of points won between players. The plot on the left shows the bootstrap median and pointwise 95% confidence interval for the profile fitted by isotonic regression. The plot on the right shows the profile fitted by isotonic regression and Binomial GLM with logit link.*

It is important to remember that the model only predicts the probability of winning the match and the actual match winner is not always the player most probable to win. On three occasions, the first server won with a 10% probability of winning the match going into the final set. There was also one occasion when the second server had won three more points than the first server going into the final set, but the first server won the match with a probability of only 25%. There were several occasions when the second server won the match, with a 25% probability, going into the final set, with the first server having won four more points than them. This adds further evidence to the advantage that the first server has over the second server.

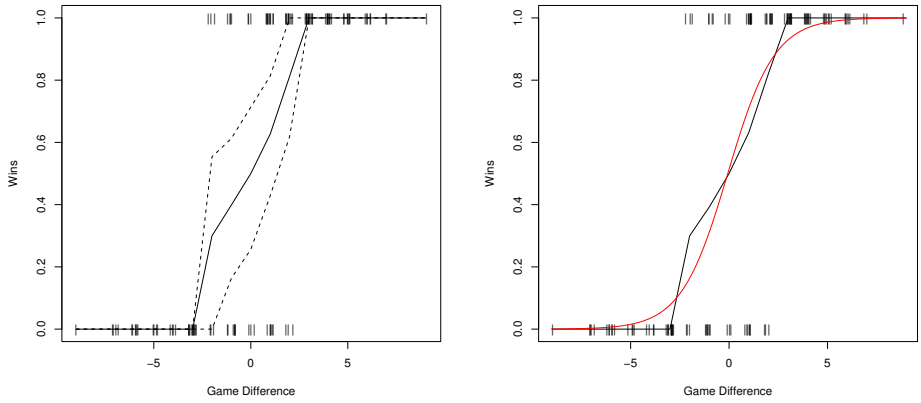

Figure 8.7: *Probability of the first server winning the match based on the difference in games won between players. The plot on the left shows the bootstrap median and pointwise 95% confidence interval for the profile fitted by isotonic regression. The plot on the right shows the profile fitted by isotonic regression and Binomial GLM with logit link.*

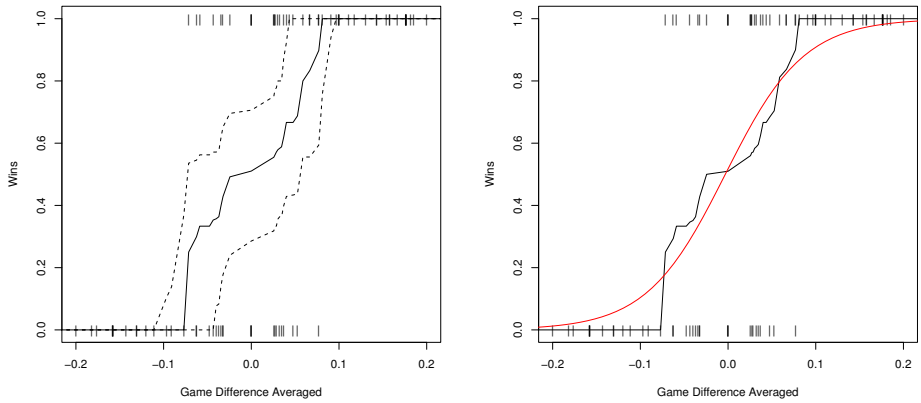

Figure 8.8: *Probability of the first server winning the match based on the averaged difference of games won between players. The plot on the left shows the bootstrap median and pointwise 95% confidence interval for the profile fitted by isotonic regression. The plot on the right shows the profile fitted by isotonic regression and Binomial GLM with logit link.*

Chapter 9

Conclusion

The first part of this thesis was based on looking for a 'winning mood effect' at the 2003 Australian Open. A 'winning mood effect' was found by Magnus and Klaassen when they analysed four years of data at Wimbledon. This thesis took a similar approach with the aim of determining whether such an effect could be found at the 2003 Australian Open as well. Evidence of a 'winning mood effect' was found after a player had served an ace or broken his opponents service, where there was an increase in the probability of winning the next service. At break point, players were inclined to take less risk on service which led to a decrease in the percentage of services won, this was also seen after missing a break point in the previous game. Although there was a slight decrease in the probability of first services in after serving a double fault there was not enough evidence to suggest players took more or less risk. After winning a set there was a decrease in the probability of winning the next game. This appeared to provide evidence against a 'winning mood effect', however, could be due to the player winning the set on their service. If this was the case the player would need to break his opponent to win the first game of the next set. If a player won the penultimate set then his probability of winning the match increased, whereas if the player had of lost the penultimate set then his probability of winning the match decreased.

The second part of this thesis looked at predictive modelling of win-

loss scenarios. Unfortunately, several of the models were too numerically complex to attempt to test on simulated data. The models that were tested on simulated data were isotonic regression, truncated beta deviates and the Bradley Terry models. Although the simulation showed these models were able to capture the relation between player ability and the probability of winning, our results suggested these methods are really only suitable for very large data sets. The Bradley Terry model used player ability for predictive modelling. This is just one indicator of the strength and skills of a player and if additional information, such as ability on different court surfaces or current injuries, was available then adjustments could be made to the model to allow better predictions for the probability of a win.

The isotonic regression model was also applied to data from the 2003 Australian Open. Going into the last set, the probability of winning the match was found to be in favour of the first server, even when they had won fewer points or games going into the final set of the match.

Bibliography

Agresti, A. (2002) *Categorical Data Analysis*, 2nd Ed. John Wiley and Sons, Inc., USA.

Agresti, A., Booth, J.G., Hobert, J.P., & Caffo, B. (2000) Random-Effects Modelling of Categorical Response Data. *Sociological Methodology*, Vol. 30, pp 27–80.

Clarke, S.R., & Norton, P. (2002) Collecting Statistics at the Australian Open Tennis Championship. *Proceedings of the Sixth Australian Conference on Mathematics and Computers in Sport*. Bond University, Queensland, Australia: UTS, pp 105–111.

Gilks, W.R., Richardson, S., & Spiegelhalter, D.J. (1996) *Markov Chain Monte Carlo in Practice*. Chapman and Hall.

Glickman, M.E. (1999) Parameter Estimation in Large Dynamic Paired Comparison Experiments. *Applied Statistician*, Vol. 48, No. 3, pp 377–394.

Hastie, T.J., & Tibshirani, R.J. (1990) *Generalized additive models*. Chapman and Hall, London.

Izenman, A.J. (1991) Recent Development in Nonparametric Density Estimation. *Journal of the American Statistical Association*, Vol. 86, No. 413, pp 205–224.

Magnus, J.R., & Klaassen, F.J.G.M. (1996) Testing some common tennis hypotheses: four years at Wimbledon. *Tilburg University, Center for Economic Research*, Discussion Paper 73, Tilburg University.

Magnus, J.R., & Klaassen, F.J.G.M. (1999) The final set in a tennis match: four years at Wimbledon. *Journal of Applied Statistics*, Vol. 26, No. 4, pp 461–468.

Nash, S.G., & Sofer, A. (1996) *Linear and Nonlinear Programming*. McGraw-Hill, Singapore.

Newton, M.A., Czado, C., & Chappell, R. (1996) Bayesian Inference for Semiparametric Binary Regression. *Journal of the American Statistical Association*, Vol. 91, No. 4338, pp 142–153.

Powers, D.A., & Xie. Y. (2000) *Statistical Methods for Data Analysis*. Academic Press, USA.

Press, W.H., Flannery, B.P., & Teukolsky, S.A. (1992)*Numerical Recipes in C*. Cambridge University Press, p 444–455.

Ramsay, J.O. (1988) Monotone Regression Splines in Action. *Statistical Science*, Vol. 3, No. 4, pp 425–441.

Ramsay, J.O. (1998) Estimating Smooth Monotone Functions. *Journal of the Royal Statistical Society. Series B*, Vol. 60, No. 2, pp 365–375.

Ramsay, J.O., & Abrahamowicz, M. (1989) Binomial Regression with Monotone Splines: A Psychometric Application. *Journal of the American Statistical Association*, Vol. 84, No. 408, pp 906–915.

Somes, G.W. (1986). The Generalized Mantel-Haenszel Statistic. *The American Statistician*, Vol. 40, No. 2, pp 106–108.