A Statistical Approach to Sports Betting

submitted by

Anton Altmann

for the degree of Doctor of Philosophy in Statistics at

City University, Department of Actuarial Science and Statistics

January 2004

Contents

List of Tables

List of Figures

Figures for chapter 6: Estimating NBA scoring rates: a question of quarters

Acknowledgements

Firstly I would like to express my gratitude towards Dr. Mark Dixon for his large supply of enthusiasm and expertise while supervising this thesis. Furthermore I would like to thank Prof. Richard Verrall, Dr. Julie Badcock and Dr. Matthew Dominey for reading earlier drafts of the thesis and providing many helpful suggestions, and Dr. Russell Gerrard for several valuable bits of advice for some of the technical problems.

In addition, I would like to thank the programmers of the statistical package *R* for producing such a useful piece of software and supplying it for free. In addition, Luigi Colombo, John Harrison and others made City University's research room a fun place to be, John Harrison in particular providing several useful ideas that have contributed towards the thesis. Finally, I'd like to thank my parents for their constant support, and occasional necessary nudge when the writing up process started to stall.

Declaration

I declare that powers of discretion are granted to the University Librarian to allow this thesis to be copied in whole or in part without further reference to the author. This permission covers only single copies make for study purposes, subject to normal conditions of acknowledgement.

Abstract

While gambling on sports fixtures is a popular activity, for the majority of gamblers it is not a profitable one. In order to make a consistent profit through gambling, one of the requirements is the ability to assess accurate probabilities for the outcomes of the events upon which one wishes to place bets. Through experience of betting, familiarity with certain sports and a natural aptitude for estimating probabilities, a small number of gamblers are able to do this. This thesis also attempts to achieve this but through purely scientific means. There are three main areas covered in this thesis. These are the market for red and yellow cards in Premier League soccer, the market for scores in American football (NFL) and the market for scores in US Basketball (NBA).

There are several issues that must be considered when attempting to fit a statistical model to any of these betting markets. These are introduced in the early stages of this thesis along with some previously suggested solutions. Among these, for example, is the importance of obtaining estimates of team characteristics that reflect the belief that these characteristics adjust over time. It is also important to devise measures of evaluating the success of any model and to be able to compare the predictive abilities of different models for the same market.

A general method is described which is suitable for modelling the sporting markets that are featured in this thesis. This method is adapted from a previous study on UK soccer results and involves the maximisation of a likelihood function. In order to make predictions that have any chance of competing with the odds supplied by professional bookmakers, this modelling process must be expanded to reflect the idiosyncrasies of each sport.

With the market for red and yellow cards in Premier League soccer matches, in addition to considering the characteristics of the two teams in the match, one must also consider the effect of the referee. It is also discovered that the average booking rate for Premier League soccer matches varies significantly throughout the course of a season.

The unusual scoring system used in the NFL means that a histogram of the final scores for match results does not resemble any standard statistical distribution. There is also a wealth of data available for every NFL match besides the final score. It is worth investigating whether by exploiting this additional past data, more accurate predictions for future matches can be obtained.

The analysis of basketball considers the busier schedule of games that NBA teams face, compared to NFL or Premier League soccer teams. The result of one match may plausibly be affected by the number of games that the team has had to play in the days immediately before the match. Furthermore, data is available giving the scores of the game at various stages throughout the match. By using this data, one can assess to what extent, and in which situations, the scoring rate varies during a match.

These issues, among many others, are addressed during this thesis. In each case a model is devised and a betting strategy is simulated by comparing model predictions with odds that were supplied by professional bookmakers prior to fixtures. The limitations of each model are discussed and possible extensions of the analysis are suggested throughout.

Chapter 1

Introduction

Among the many applications of probability and statistics, gambling is maybe one of the more widely known and appreciated by the general public. There are many popular forms of gambling in society today, such as state lotteries, casino games and sports betting, which is the focus of this thesis.

The aim of this study is to incorporate many of the ideas that are considered by professional bookmakers and successful gamblers within a formal statistical framework. By employing well-known and well-understood statistical procedures, the intention is to combine these ideas in such a way as to optimise the ability to predict future results. Ideally the probabilities produced will be more accurate than the bookmaker's odds[1].

In Section 1.1 of this introductory chapter some differences, in terms of specification of a probability model, between sports betting and various other types of gambling are considered. The strengths and weaknesses of both the statistical approach and the intuitive approach favoured by bookmakers and the majority of gamblers are discussed in Section 1.2. In Section 1.3 a brief explanation of the betting opportunities available today is given. Some details concerning the scope of this study are outlined in Section 1.4, since restrictions are placed on the type of sports analysed. In Section 1.5 the structure of the thesis is outlined.

1.1 Some key features of sports betting

One way in which sports betting is different from many other types of betting, from a statistical point of view, is that the probabilities cannot be fully specified. For

[1]Superior predictions are not in fact essential in order to win money against the bookmakers, as will be shown in Section 5.6.1.

many casino games, such as Roulette, the probabilities are entirely known by both the gambler and the bookmaker. As a result there is no way the gambler can make a profit on a long-term basis due to the small bias in the Casino's favour, assuming the roulette wheel to be fair[2]. Meanwhile, for some card games, such as Poker, the entire probability distribution for future events conditional on the information currently available to the player is *theoretically* possible to calculate but in practice it is impossible for a poker player to process the full set of calculations while the game is taking place. Nevertheless, with experience, the player can approximate the odds of various outcomes in order to assess which decisions should be made. Also, on balance, every player has access to an equal amount of information (although the importance of this information will vary from hand to hand). For sporting events however, the number of factors that determine the probabilities are far more numerous, and their importance towards forming the probabilities of various occurrences in the fixtures is generally impossible to quantify. Furthermore, not all gamblers/bookmakers have equal access to relevant information.

1.2 A statistical approach to gambling versus an intuitive approach

Despite the large number of sports that are popular with gamblers, and the large number of markets within these sports, a relatively small amount of formal literature concerning odds for sporting events has been published. In fact, the vast majority of odds that are available for sporting events are not derived through advanced statistical techniques. In general they are determined through the practical experience of those setting the odds combined with selective use of basic figures such as team/player averages, or in many cases it is the beliefs and behaviour of the market that determines the odds.

Most successful gamblers apply a thorough knowledge of the sport on which they are betting combined with a need-to-know knowledge of mathematics and probability in order to choose the most attractive bets. To some extent this thesis takes the opposite approach. Some of the advantages and disadvantages of either perspective are obvious. In particular, most prospective gamblers find the process of accumulating

[2]There have been instances of players keeping totals of the frequency of each roulette number's occurrence and discovering numbers on certain areas of the wheel to occur more frequently, due to small inclines on the surface on which the roulette table is placed.

knowledge about a sport more inviting than the process of acquiring the statistical techniques required to produce accurate sports models as well as the computing experience necessary to implement these models. On the other hand, while the initial process of producing a statistical model is labour intensive and frequently frustrating, should a successful model eventually be created, far less effort is involved in generating further predictions. Furthermore, by not using any quantitative methods it is hard for a gambler to combine their experience and knowledge in an optimal way. For example, Forrest and Simmons (2000a) studied the predictions of English and Scottish soccer results made by professional advisers working for various British newspapers. It was investigated whether widely available information, such as recent form of the teams or the difference in league positions between the two soccer teams participating in a match, were used by the advisers. It was concluded that some of the information seemed to correlate strongly with the forecasts made by the advisers. However excessive weight was attached to some parts of it while other parts were not exploited by the forecasters even though they were important predictors of soccer results. Furthermore, while a statistical model's effectiveness can be quantified by using projected profit curves or confidence intervals on projected returns, for example, the success of an intuitive approach is not easily measured and attempts to do so are frequently inaccurate, optimistic or both[3].

The statistical approach meanwhile is heavily dependent on assumptions and is thus inflexible to certain important factors. In horse or greyhound racing, for example, the odds change very quickly in the hours leading up to a race to accommodate new information such as paddock gossip, results of previous races that day or weather changes. Obtaining this information in a convenient format for computers, such as a spreadsheet, for past fixtures and attempting to fit a reliable model in order to update predictions hour-by-hour is impractical. Therefore, before choosing to take a statistical approach towards modelling a sport, with regard to producing a prediction system that is on average superior to the intuitive approach, one must decide carefully which sports are suitable.

[3] "In betting on races, there are two elements that are never lacking: hope as hope and an incomplete recollection of the past" E.V. Lucas, New York Times, 7 October 1951.

1.3 Betting opportunities

While the importance of producing accurate probabilities for sporting events is evident, it is also necessary to maximise the potential profits of any betting strategy using these probabilities by selecting the most attractive bets available. There are many different mediums through which bets can be made and many different types of bets are available for most fixtures.

Until the mid 1990s in the UK, visiting a high street bookmaker such as Ladbrokes or Coral was the most popular method of placing a bet. More recently, many find it more convenient to place bets with an online bookmaker, of which there are now many such as Sportingbet or Premierbet. Another recent opportunity for gamblers is the use of *betting exchange* websites such as Betfair where betting odds and maximum stakes are offered by gamblers to other gamblers, thus removing the bookmaker's role of specifying odds. For most websites of this type, a small proportion of the profits from winning bets goes to the website administrators.

The types of bets available for most sports fall into two basic categories. The more traditional format is *fixed odds*, which is the system used by high-street bookmakers as well as many online bookmakers. A more recent format that has become popular within the last decade is *spread betting* (otherwise known as *index* or *range* betting).

1.3.1 Fixed odds

For European sports, a fixed odds system generally offers odds on each outcome of an event. For example, these odds were available from Sportingbet for a Premier League soccer match between Liverpool and Middlesbrough on 8 February 2002:

Liverpool victory: 8:15

Draw: 12:5

Middlesbrough victory: 5:1

If £K is staked on an outcome with offered odds O1:O2, then £(O1/O2+1)*K is returned to the gambler, yielding a profit of £K*O1/O2. If an outcome has odds O1:O2, then this corresponds to a probability of O2/(O1+O2) of it occurring. Hence the probabilities suggested by the bookmaker for the match above of the outcomes (Home Win, Draw, Away Win) are (0.65, 0.29, 0.17). Note that the sum of these

probabilities is 1.11 rather than 1. The surplus of 0.11 is known as the bookmaker's *take* or *overround*. By scaling down all probabilities by $\frac{1}{1.11}$, implied probabilities of (0.59, 0.26, 0.15) are obtained. For every £1 that is placed, the bookmaker makes a profit of £0.11 assuming equal money is placed on all three outcomes. If money is not placed equally on all three outcomes, the bookmaker's expected profit increases as more money is staked on the outcomes with lower odds (so higher probability). The size of the bookmaker's overround varies across different sports and fixtures. In general, the more popular fixtures are frequently bet on by less discerning gamblers who do not seek the most favourable odds, which allows the bookmaker to offer less competitive odds yet still attract custom.

Fixed odds for many American sports differ slightly from those of European sports in that *handicap* bets are more common[4]. With these, one of the sides starts the game with a points handicap (known as a *line*) specified by the bookmaker, and the gambler may place bets on which side will win the match, after the scores are adjusted to include the handicap. The payouts for either bet are equal (this is known as *even odds*). For example, this line from *Sportingbet.com*'s website allows two bets:

Tennessee Titans: -7.0 -1.10

Pittsburgh Steelers: +7.0 -1.10

The gambler should back Tennessee Titans if they believe it is likely that the Titans will beat Pittsburgh Steelers by more than seven points in the match. Conversely, if they believe that Pittsburgh Steelers are likely either to win, or to lose by six points or fewer, they should back them. Should either bet be successful, on a £1 stake, the gambler will be returned $£1 + £1 * (1/1.10) = £1.91$. Should the bookmaker have equal amounts of money staked on either team, its overround is $1 - \frac{1}{1.1} = 0.091$.

Another type of fixed odds bet is an *asian handicap* bet which combines the two types of bet listed above. A handicap is offered but the odds are not always even. Also, in the event of the match ending in a draw, once the handicap is included, all stakes are returned (this is known as a *draw no bet* arrangement).

[4]Often referred to as *point spreads* in the US, not to be confused with UK spread betting.

1.3.2 Spread betting

Spread betting is a stock-market style of betting where sporting outcomes are traded like commodities. As an illustrative example, in the Arsenal - Manchester United match played on 22 August 1999, prior to kickoff one spread betting firm displayed a *spread* for the total number of bookings points for the match as $30 - 34$. The spread reflects the firm's expectation of the number of bookings points (10 points for every yellow card and 25 for every red card) given during the match (if a player receives two yellow cards in the match, resulting in a red card, then overall 35 points are included in the final total for this player).

The number of yellow and red cards that will be scored are being treated as a commodity and gamblers can *buy* this commodity at 34 or *sell* at 30. A typical bet on this match might be to SELL at 30 for £1 per point. If the final total is

- 4 yellows and 0 reds = 4*10+0=40 points then the return is $(30-40) \times 1 =$ -£10
- 0 yellows and 1 red (10*0+25*1)=25 points then the return is $(30-25) \times 1 =$ +£5.

A BUY at 34 of bookings points works in a similar way, with returns of +6 and -9 respectively. In addition, the spreads often fluctuate up and down in the lead-up to kick-off to reflect the betting behaviour of gamblers, for the same reasons and in the same way as stock market prices move.

One key advantage spread betting has over fixed odds betting is that when a spread bet is made, the profit or loss can adjust right up to the conclusion of the fixture. If a fixed odds bet is made, this is not always the case. For example, if one bets that the score of a soccer match will be 0-0, the bet (and possibly interest in the fixture) is lost as soon as a goal is scored. On the other hand, some gamblers are deterred from spread betting since the maximum loss is uncertain prior to a fixture. For example, if in the example above the number of bookings is sold, the maximum return is +30 times the selected stake whereas the maximum loss could be as high as around 130 times the stake. Some spread betting firms offer *stop loss* accounts, where the maximum gain or loss on each market is set to some agreed limit.

For the bookmaker the preferred situation is to have equal amounts of money staked on either side of the offered spreads. Assuming, for simplicity, that £K has been both bought and sold on a constant 36-40 spread for a fixture, this will result in a risk-free profit of £$4K$ for the bookmaker. Analogously, in the case of fixed odds betting such

as an NFL handicap bet as described above, the bookmakers will ideally have equal amounts of money staked on both outcomes of the match. Thus, contrary to popular belief, the points handicap is not specified with the aim of estimating the most likely difference in score in the match but with the aim of attracting an equal volume of bets on each outcome. As a result it seems feasible that the probabilities inferred from the odds available from bookmakers are not always unbiased. Besides the obvious difficulties in evaluating unbiased probabilities for each outcome of any sporting event it is sometimes in the interests of the bookmaker to offer inaccurate probabilities in order to minimise their financial exposure.

This can lead to market inefficiencies - that is, there may be opportunities for gamblers to have an expected gain by placing bets on some outcomes. There have been several academic studies into the market efficiency of various sporting markets. Gandar *et al* (2000) investigate the market efficiency of the betting markets for Major League Baseball (MLB) and NBA. The accuracy with which handicaps accommodate the effect of a side playing at their home ground is examined. It is concluded for both sports that win rates of sides playing at home do not differ significantly from the rates inferred by the odds. Vergin (2001) simulated a set of betting strategies designed to investigate the theory that handicaps reflected a tendency of the public to overestimate the ability of a team with recent good results and to underestimate the ability of a team with recent bad results. The theory appeared to be true in the case of a side with recent good results. Furthermore, Simmons, Forrest and Curran (2003) investigate the efficiency of the handicap and spread betting markets of Rugby League fixtures. It is concluded that the handicap market does not fully incorporate the home advantage but does correctly evaluate the relative strengths of the two competing teams (thus has no *favourite - underdog bias*). The spread betting market is unbiased in these two respects however. It is speculated that the handicap betting market may comprise gamblers who select bets largely on an emotional basis whereas due to the higher risks and higher returns associated with the spread betting market, any spread bet placed must seem financially sound to the gambler. In a study of NFL betting lines, Vergin (2001) supports this theory by reporting that "Line-makers report that bets from the unsophisticated general public far outweigh bets of expert handicappers. Therefore they give primary weight to the biases of the uninformed general public".

Note that bias in the probabilities inferred from a set of odds does not imply market inefficiency, which, as stated above, only arises if an expected gain is available by

placing a bet on at least one outcome. There may not be such a bet if the bookmaker's overround prevents the gambler from exploiting the inaccuracy in the probabilities in order to realise an expected gain.

1.4 Choosing appropriate sports for modelling

As mentioned in Section 1.2 some sports are more amenable to statistical analysis than others. There are several considerations in this respect:

- Are there significant betting opportunities for the sport? Ideally it would be possible to have a successful model for predictions of the results for a sport and to use these in order to develop a profit-making strategy by placing bets with bookmakers. The scope for a model to be profit-making is increased with a greater number of events to bet on, and with a larger selection of different bets available.

- Are there adequate data resources available? There are detailed match reports available on the internet for all fixtures since 1995 for all of the four major American sports (NFL, NFL, MLB and the National Hockey League). Some European websites have reasonably detailed and consistently formatted statistics available for UK soccer matches since 1996. More recently, detailed web pages for cricket and rugby matches have become available.

There are of course other criteria besides those mentioned above, not least the ease with which the sport can be modelled statistically. Certain sports, such as cricket and golf, do not have scoring systems which can be easily approximated by the well-known distributions, while another problem is that the covariates required in establishing a model for some markets, such as the number of corners in a soccer match, are not always clear.

Although quantitative analysis may be of interest for many sports, in the interests of achieving a reasonable level of depth to the analysis of the sports it studies, this thesis restricts itself to specific types of sports, so that various routine procedures can easily be reproduced for different applications. In particular this study is restricted to sports where

- individual matches take place between two opponents, which could be individual players or teams. This excludes horse racing, golf and motor racing for example.

- the outcome is described by an accumulated total, rather than a time or an event. This excludes sports such as boxing or athletics.

Among the sports which do fulfill the necessary criteria are UK soccer, NFL, and NBA. In addition, this thesis only considers the statistical methods required in order to predict final results of matches and does not attempt to model various other aspects of sports fixtures, such as the time until the first goal or the winner of a league. Also, since the final result is the variable of ultimate interest, within-game modelling is not generally considered. For a detailed analysis of within-game models for soccer data, the reader is referred to Hirotsu and Wright (2002).

1.5 Outline of thesis

So far some of the issues involved in producing a profit-making gambling strategy through the production of an accurate set of probabilities have been discussed at a very general level. The remainder of this thesis is structured as follows. In Chapter 2 the main issues that need to be addressed when modelling sporting events of the type listed in Section 1.4 are specified. Possible solutions are discussed by means of a literature review. Chapter 3 gives a detailed explanation of one of the stages of the modelling process, namely the estimation of the parameters of a specified statistical model. Three individual sports markets are treated in depth in Chapters 4, 5 and 6 using the procedure outlined in Chapter 3. These are respectively the markets for red and yellow cards for Premier League soccer matches, NFL scores and NBA scores. An alternative estimation procedure to that explained in Chapter 3 is implemented and evaluated in Chapter 7. Chapter 8 concludes.

Chapter 2

Overview of sports modelling techniques

As stated in Chapter 1, the motivation for this thesis is to produce statistical models for outcomes of certain sports events. Ideally these will be able to generate probabilities which are competitive with the odds provided by professional bookmakers. In order to approach this task, some of the techniques already applied to modelling sports will now be discussed via a literature review, with the aim of outlining some of the problems encountered, and attempts at solutions.

2.1 Sports modelling - the main issues

On consideration of the sports of interest, as detailed in Section 1.4, the general task is to produce the most accurate possible prediction for the final result of a fixture between two teams. This can be expressed as result (X, Y) between teams (t_1, t_2). Note that (X, Y) does not have to denote a final score, and could express the total number of red cards or the number of shots on goal for example. The following issues must all be considered:

2.1.1 Selecting a suitable distribution for responses

The scale, variance and range of X and Y vary from sport to sport. Some thought must be given as to which of the common statistical distributions are most suitable, or if a combination of these distributions is more appropriate, or indeed if a non-parametric distribution is required. For many betting markets $X + Y$ or $X - Y$ is frequently of interest, as highlighted in Section 1.3.

2.1.2 Accommodating dependency between the home and away scores

It is plausible for many sports that information about the value of X may affect one's belief about the density of Y. Hence the joint distribution of (X,Y) may need to accommodate the possible dependency between X and Y. The difficulty of fitting such a joint distribution depends on both the response distribution chosen and the characteristics of the relationship between X and Y. The Normal distribution has a well known bivariate form whereas the Poisson distribution, for example, has a considerably more complicated, and thus less flexible, bivariate form. In addition, the relationship between X and Y may be of such an intricate form that no existing probability distribution adequately represents the joint distribution (X, Y).

2.1.3 Representing the team abilities

The ability of sporting teams, in terms of their impact on X and Y, frequently is not adequately expressed by a single parameter. For example, if the expected value of $X - Y$ is large and positive, this could arise because t_1 has a tendency to play in such a way that high values of X are expected. Alternatively their style of play may generally prevent high values of Y arising. These situations are analogous to a team respectively being predominantly *attacking* or *defensive* if X and Y are soccer scores. In addition to studying the mean values of the team parameters, their variance may also be of interest, since some teams may be less consistent than others.

2.1.4 Including covariates other than the abilities of the teams

For several sports, totals for *in-game statistics* such as attempted shots, fouls committed and time spent in possession of the ball are available. In addition to these, factors such as the effect of playing at home, the length of time since the previous fixture, the key injured players and many other relevant factors could be included to improve the accuracy of predictions.

2.1.5 Allowing parameters to adjust over time

It seems reasonable that the values of the parameters of t_1 and t_2 should vary over time. However, the way in which past information, such as results of previous fixtures, should be used in order to determine these parameters is a complicated issue. How much importance should be attached to a result from a fixture that occurred one year

ago compared to the result of a fixture that took place the previous week?

2.1.6 Finding techniques in order to obtain estimates of the parameter values

While exploration of the data based on one's knowledge of the sport can lead to the specification of a model, the process of obtaining estimated values of the parameters employed in the model can be a considerable task. An overly ambitious model may even make the process mathematically intractable. In some cases compromises with regard to the model specification may be necessary.

2.1.7 Validating the model and assessing predictive capability

The task of producing statistical models for sports results differs from the task of producing statistical models for some other applications in the sense that it is the prediction of future events, as opposed to the interpretation of existing data, that is of interest (although these processes are of course linked). In particular, the danger of *over-fitting* must be considered. For example consider the situation that in the first five matches of the data set where a team played on the birthday of the wife of the team's manager, the team won. If the objective is to interpret existing data a common step would be to obtain the optimal fitted values for the data. To achieve this, an indicator vector signalling matches when teams played on the manager's wife's birthday should be included. Assuming that these results are entirely coincidental, as seems likely, the predictive power of the model will be greatly harmed by doing so. Hence standard measures for determining the accuracy of a fitted model, such as R^2 or C_p, are only considered provided the conclusions drawn from their use are also reflected by an improvement in the predictive capability. Methods of measuring this will be devised.

2.1.8 Comparing predictions obtained through a statistically-based procedure with bookmakers' odds

The stated ambition in Chapter 1 was to produce predictions that are superior to those of professional bookmakers. A variety of criteria are necessary in order to determine if this has been achieved.

2.1.9 Considering betting strategies based on model predictions

The development of a betting strategy once predictions have been obtained from a model is potentially an extremely involved task. A relatively small amount of attention is placed on the subject in this thesis and some of the more straightforward approaches are considered.

Note that in order to consider issues 1 to 4 in particular, some knowledge of the sports being modelled is required. In addition, much of the treatment of issues 1 to 4 is specific to each sport. For example, finding the optimal statistical distribution to represent NBA scores does not necessarily ease the task of finding the optimal statistical distribution for other sports results. As a result, discussion of issues 1 to 4 with reference to previous literature is deferred to the relevant sections of Chapters 4, 5 and 6, which include detailed explanations of the sports markets they cover. Issues 5 to 9 generally require less detailed knowledge of specific sports and their treatment mainly involves statistical methods that can be generalised to many other applications, including many sports markets. Hence these issues are discussed in the remainder of this chapter.

This chapter contains mainly technical material which is included in order to suggest some techniques that could be considered for the treatment of tasks that arise later in this thesis, and to introduce the reader to some of the thinking behind existing attempts to model sports. A thorough understanding of this material is *not* essential in order to follow the development in later chapters so if the reader is more interested in the techniques applied specifically in this thesis, the remainder of this chapter does not need to be read immediately and can be referred to, where directed by the text, if necessary at later stages of reading.

2.2 A simple example of a sports model

In order to link the ideas outlined very generally above with the formal analysis summarised in the remainder of this chapter, it is helpful to provide an example of a model specification that could be employed in any general attempt to model sports of the type that this thesis considers.

It is assumed the data set includes N matches. For each match k, $k \in [1, \ldots, N]$, the home and away scores, X_k and Y_k respectively, are available as are the identities

of the home team and away team, denoted by $i(k)$ and $j(k)$ respectively. The term 'score' is used here for convenience, however, as mentioned at the start of Section 2.1 X_k and Y_k could also represent match totals of figures other than the final score, such as shots on goal or the number of fouls.

For this example it is assumed that each team's ability (to affect X_k and Y_k) can best be described by two parameters. These parameters are assumed to alter over time hence the two ability parameters of team $i(k)$ at time t are denoted by $\alpha_{i(k),t}$ and $\beta_{i(k),t}$. If X_k and Y_k represent the final match scores of a fixture then the α and β terms represent the attacking and defensive abilities of the team respectively.

Two further parameters are included in this example model. The effect of playing at home is described by a single parameter δ and it is assumed this effect is the same for all teams. Finally, since it is desired that the α and β terms have a mean of zero to aid their interpretation, a term to represent the global mean is included, denoted by γ. Note that the inclusion of δ means that γ is effectively the mean for all away fixtures.

With these terms defined it is now possible to specify a model. The expected scoring rates of both sides for match k, which takes place at time t is as follows:

$$\begin{aligned} E[X_k] &= exp(\gamma + \alpha_{i(k),t} + \beta_{j(k),t} + \delta) \\ E[Y_k] &= exp(\gamma + \alpha_{j(k),t} + \beta_{i(k),t}) \end{aligned}$$

Some previous studies which employ model specifications similar to this one subtract the β terms in the right hand sides of the above equations. The above formulation is entirely equivalent, although the interpretation of the β parameter estimates has to be inverted. The right hand sides of the above equations are exponentiated since for almost all sporting markets (including those studied in this thesis) match totals are always greater than or equal to zero. This implies that their means must be strictly positive, hence the use of the exponential function.

Not all previous academic sports studies use the match scores as the outcome variable. For example Forrest and Simmons (2000b) when studying English and Scottish League soccer results represent outcomes using a vector that can take on three values, in order to represent the three outcomes of home win, draw and away win.

2.3 Allowing parameters to adjust over time

In most applications to sports the α and β parameters need to adjust over time. Generally their abilities change for many different reasons and the parameters need to reflect teams drifting in to and out of form, losing players or changing coach, for example. Large sections of the previous relevant literature are devoted towards modelling this process, and some of the ideas are summarised in this section.

The approach employed by Dixon and Coles (1997) when modelling English League and Cup soccer scores is to taper the log-likelihood, so at time-point t the following psuedo-loglikelihood is maximised:

$$\sum_{k=1}^{M_t} log(L(X_k, Y_k|\Theta_t)) * exp(-\varsigma(t - t_k)) \tag{2.3.1}$$

where M_t is the number of matches played prior to time t, and Θ_t is the set of parameter values at time t. In this framework the same values of α_i and β_i are employed for every match in which team i plays. However the importance of matches towards obtaining optimal estimates for α_i and β_i decreases the longer ago the match is, via the exponential term. Since the quantity in Equation 2.3.1 is maximised at many time-points throughout the data set, the estimates of the α and β terms evolve through time even though for each individual maximisation the same team parameter value is used for each match in which the team plays. ς is selected in order to maximise the predictive capability of the model. Due to the non-standard nature of the predictive likelihood, this can only be achieved by inspection through testing a range of values of ς and monitoring the resulting value of predictive likelihood.

Another logical approach is to define a distribution explicitly for α_i as a function of time for which there are several methods available. The starting point is to assume α_i follows a random walk, as implemented by Fahrmeir and Tutz (1994) in the development of a paired comparisons system, so that

$$\alpha_{i,t+1} = \alpha_{i,t} + u_{i,t+1} \tag{2.3.2}$$

where

$$u_{i,t+1} \sim \mathcal{N}(0, \sigma_i)$$

Note that this formulation allows team-specific movements, defined by σ_i. Glickman

and Stern (1998) modified this concept slightly for a model of NFL scores, so that

$$\alpha_{i,t+1} = \kappa(\alpha_{i,t} - \overline{\alpha_{.,t}}) + u_{i,t+1} \tag{2.3.3}$$

where $u_{i,t}$ is as in Equation 2.3.2, (but in fact was not set to be team specific), κ represents a shrinkage/expansion factor, to accommodate the possibility that there is a trend where the overall disparity between the ability of teams increases, or decreases, over time. Glickman and Stern obtained a posterior value of 0.99 for κ, suggesting that no such trend necessarily exists.

Knorr-Held (2000) argued that one possibly undesirable property of the Glickman and Stern formulation is that if team i does not play at time $t+1$, but other teams do, then

$$E[\alpha_{i,t+1}] = \alpha_{i,t} - \frac{1}{n}\sum_{j=1}^{N}\alpha_{j,t}$$

(where the κ term included in Equation 2.3.3 is assumed to be 1). Hence team i's parameter adjusts even though they didn't actually play, since the model has not been designed with the property that

$$\frac{1}{n}\sum_{j=1}^{N}\alpha_{j,t} = 0$$

In certain situations, there is some sense behind this phenomenon. For example, suppose Liverpool beat Southampton 3-0 on day t and on day $t+1$, Manchester United beat Southampton 7-0. Liverpool's 3-0 result looks a little less impressive after time $t+1$, so it seems appropriate that Liverpool's attacking parameter could drop.

Crowder *et al* (2000), who included two parameters for each team in the production of an English League soccer results prediction system, applied a modified extension of the Glickman and Stern setup, by employing an AR[1] process:

$$\begin{pmatrix} \alpha_{i,t+1} - \alpha_{i,0} \\ \beta_{i,t+1} - \beta_{i,0} \end{pmatrix} = \begin{pmatrix} \rho_{\alpha\alpha} & \rho_{\alpha\beta} \\ \rho_{\beta\alpha} & \rho_{\beta\beta} \end{pmatrix} \begin{pmatrix} \alpha_{i,t} - \alpha_{i,0} \\ \beta_{i,t} - \beta_{i,0} \end{pmatrix} + \begin{pmatrix} u_{i,t+1} \\ v_{i,t+1} \end{pmatrix}$$

where

$$\begin{pmatrix} u_{i,t+1} \\ v_{i,t+1} \end{pmatrix} \sim \mathcal{N}_2(0, \Sigma)$$

independently.

This structure allows dependence between a team's attacking and defensive param-

eters, firstly from the constant (with respect both to time and teams) autoregressive component, via $\rho_{\alpha\beta}$ and $\rho_{\beta\alpha}$, and also the time-specific, team-specific variations $u_{i,t}, v_{i,t}$ can be mutually dependent.

Another solution is to allow the amount of change in ability to be proportional to the time since the previous estimation. Note that the treatments above have indexed time, hence either assume that all points of estimation are equally far apart, or that a team's ability varies equally between each time-point when a game is played regardless of the time differences between these time-points. However, Rue and Salvesen (1997) use Brownian motion to model the evolution of parameters, so the attacking parameter at time $t+s$ is modelled thus:

$$\alpha_{i,t+s} = \alpha_{i,t} + (B_{\alpha_i}(\frac{t+s}{\tau}) - B_{\alpha_i}(\frac{t}{\tau}))\sigma_{\alpha_i} \tag{2.3.4}$$

where $B(t)$ is standard Brownian motion starting at level 0 and τ is the non-team-specific inverse loss of memory rate for the α parameter. The defensive parameters are similarly defined.

Harville's (1980) treatment of NFL models the team abilities as a random effect, within a mixed linear model framework, where team abilities are assumed to vary from season-to-season, but not from game-to-game within a season. Hence the score difference S_k for match k between sides $h(k)$ and $a(k)$ during season m can be represented as

$$S_k = T_{h(k),m} - T_{a(k),m} + H + R_k$$

where $T_{h(k),m}$ and $T_{a(k),m}$ represent the abilities of team $h(k)$ and $a(k)$ relative to the average ability during season m, H is the home effect and R_k is match k's random residual effect.

Finally, it is necessary to make suitable adjustments for season breaks, particularly when one considers Equation 2.3.4. The English League soccer season typically breaks for around 3 months over summer, while the NBA season has a six month breaks, and the NFL season breaks for 8 months. It seems unsatisfactory to treat these breaks as any other and one might expect team abilities to vary at a different rate during season breaks compared to the gaps between fixtures during the regular season. Glickman and Stern (1998) acknowledge this effect when specifying an NFL model and incorporate two further parameters into their model so that, if time $t+1$ is the time of the first

match of a new season,

$$\alpha_{i,t+1} = \kappa_s(\alpha_{i,t} - \overline{\alpha_{.,t}}) + u_{i,t+1}$$

where

$$u_{i,t} \sim \mathcal{N}(0, \sigma_s).$$

κ_s is a season-to-season shrinkage/expansion regression parameter and σ_s is the between-season evolution standard error. The posterior value for κ_s obtained was 0.82, with 95% posterior interval (0.52, 1.26), based on six season's worth of data. The fact that $\kappa_s < 1$ is plausible since the post-season drafting system is designed so that the most promising American football players from US colleges are allocated to the worst performing teams from the previous season, in an attempt to prevent the hierarchy becoming too ingrained. In English League soccer, with no such system in place, it is plausible that a similar study may conclude that $\kappa_s \geq 1$.

2.4 Finding techniques in order to obtain estimates of the parameter values

Dixon and Coles (1997) employ Newton-Raphson maximisation routines in order to find the MLEs of the parameters. With a modern computer this can be accomplished in a matter of seconds and has to be repeated for each time-point, or for however many estimations are required for the desired level of accuracy. However, as outlined in Section 2.3 the Dixon and Coles formulation does not feature 'true' dynamic team abilities since in every match in which a team participates, the same parameters are employed for that team. Furthermore, while it is possible to obtain both MLEs and their standard errors (by taking the diagonal elements of the inverse of the observed information matrix of the likelihood), the full posterior distribution of the parameters is not available. This makes it difficult to verify the validity of the parameter distributional assumptions, for example.

On the other hand, now consider the likelihood functions that apply to a true dynamic model. In general, if $\mathbf{G_t}$ denotes the set of results observed at time t, and θ_t denotes the set of all parameter values at time t, then the relevant likelihood function,

conditional on initial value θ_0 is

$$\begin{aligned} L_t(\underline{\theta}) &= p(\mathbf{G_1}, ..., \mathbf{G_t}|\theta_0) \\ &= \int_{\theta_1} p(\mathbf{G_1}, ..., \mathbf{G_t}, \theta_1|\theta_0)\mathbf{d}\theta_1 \\ &\vdots \\ &= \int_{\theta_t} \dots \int_{\theta_1} p(\mathbf{G_1}, ..., \mathbf{G_t}, \theta_1, \dots, \theta_t|\theta_0)\mathbf{d}\theta_1 \dots, \mathbf{d}\theta_t \end{aligned} \tag{2.4.1}$$

For a model of NBA scores, using Brownian motion for the drift of team ability as demonstrated in Equation 2.3.4, the parameters required include attack and defensive parameters for each team at each time-point, plus constant parameters for the global mean, home effect, memory loss, and between-season expansion/shrinkage. This gives a total of $t * 29 * 2 + 4$ parameters at time t. This is the dimension of the integral in 2.4.1, which means Newton-Raphson is not appropriate. Both Glickman and Stern (1998) and Rue and Salvesen (1997) apply a *Markov Chain Monte Carlo* (MCMC) technique, although doing so requires considerable thought in order to divide the complete posterior distribution into more convenient and computationally efficient conditional posterior distributions. Also, some inspection is required in order to find suitable prior distributions. MCMC will be considered in more detail in Chapter 7.

Crowder *et al* (2000) devised a fairly intricate approximation to the likelihood for their English League soccer model, which avoided the used of MCMC and was computationally more efficient. However, it is not readily adaptable to alternative, more complicated model formulations. This approximation is not considered in this thesis.

One other technique that has been considered in order to obtain time-dependent estimates is application of a *Kalman Filter*. In general, one considers applying a Kalman filter when one wants to represent a stochastic process x governed by the following linear stochastic differential equation:

$$x_k = Ax_{k-1} + w_{k-1}$$

via a measurement z such that

$$z_k = Hx_k + v_k$$

Random variables w_k and v_k respectively represent the process and measurement error. The Kalman filter approach produces predictions for x via a set of *prediction*

equations, which predict values for x and the error covariance matrix, and a set of *measurement update* equations, which act as feedback to the prediction process. The update equations are designed in order to improve future predictions. The *extended Kalman filter* relaxes the assumption that the process must be linear.

Fahrmeier and Tutz (1994) apply an extended Kalman filter to the set of parameters, which include response thresholds (since their response vector is categorical), team abilities and, optionally, parameters for any other covariates. The estimation of parameters is achieved via posterior modes, and the likelihood is considered as a combination of *filters* and *smoothers*. The filters are the loglikelihood of the data given parameter estimates at the latest time point and the smoothers are the loglikelihood of the transitions of the parameter values from one time-point to another. At each time-point it is assumed that

$$\beta_t \sim \mathcal{N}(T_t\beta_{t-1}, Q_{t-1})$$

where β_t is the set of parameters described above, T_t is the transition matrix at time t and Q_t is the error process.

Some other studies do not explicitly include team abilities as parameters which need to be estimated by the model. Forrest and Simmons (2000a), for example, approximate team abilities with a range of measures such as recent form, league positions and total scored/conceded goals in the current season. The match result is then regressed against this set of measures and the coefficients of this regression are the parameters to be estimated. As mentioned in Section 2.2, Forrest and Simmons classify a match result as either a home win, a draw or an away win. An ordered logit model is used to obtain parameter estimates and this could in principal be extended to a model which estimates team parameters. While many popular statistical packages supply routines for ordered logit analysis, ideally the estimation would be adapted to allow team abilities to vary over time. This could be an onerous task given the large computational requirements of a conventional ordered logit likelihood maximisation.

2.5 Validating the model and assessing predictive capability

2.5.1 Discrepancy measures

One suitable technique to see how closely the specified model mimics the observed data is to compare the predictive distribution to the data. This can be done by simulating a suitable number of samples from the predictive distribution and comparing these samples to the observed data. There are usually various aspects of the data that can be checked and it is therefore useful to devise one or more test quantities. If a model is being developed within a classical framework, this is a scalar summary of the data. If the problem is being considered within a Bayesian framework, this is a scalar summary of both the data and the parameters. This test quantity $T(\mathbf{y})$ is known as a *test statistic*, in the classical case, and as a *discrepancy measure* $T(\mathbf{y}, \theta)$ in the Bayesian case (Gelman *et al* 1995). Using these test quantities, tail area probabilities to quantify the scale of disagreement between model and data can be approximated.

In the classical case, suppose there are n observed values $\mathbf{y} = (y_1, \ldots, y_n)$. K copies of *replicated* data, $\mathbf{y_1^*}, \ldots, \mathbf{y_K^*}$ given the model and the estimated value of θ can be generated. Hence $\mathbf{y_i^*}$ is a vector of simulated values $(y_{i1}^*, \ldots, y_{in}^*)$. Then set $T(\mathbf{y_i^*}) = min(y_{i1}^*, \ldots, y_{in}^*)$, for example. Then the tail area probability could be defined as the length of the vector

$$\{i : T(\mathbf{y_i^*}) > T(\mathbf{y}), i \in [1, K]\}$$

divided by K. For large enough K, this is an approximation to $P(T(\mathbf{y_\cdot^*}) \geq T(\mathbf{y})|\theta)$. If this tail area is close to zero or one, it suggests a possible discrepancy.

The Bayesian method differs, since a posterior distribution of the parameters is considered, rather than their point estimates. As such, sample values $\theta_1^*, \ldots, \theta_K^*$ are generated from the posterior distribution of θ, then for each generated value, a single $\mathbf{y_i^*}|\theta_i$ is generated. Tail areas can be computed as above, to approximate $P(T(\mathbf{y_\cdot^*}, \theta) \geq T(\mathbf{y}, \theta)|\mathbf{y})$.

Glickman and Stern (1998) apply this technique to their NFL model, which is defined from a Bayesian perspective. One assumption which they test is that the variance of the score difference, conditional on its mean, is equal for all games. Two discrepancy measures, which are sensitive to this assumption, that they use to test

this are

- the difference between the largest annual average squared score-prediction residual, and the smallest (Glickman and Stern have six years of data available)
- the difference between the largest and smallest average squared score-prediction residual for each team.

In fact, Glickman and Stern do not obtain any significant evidence to suggest that the variance of a match score is a function of its mean. They also conclude using discrepancy measure techniques that it is necessary to include team-specific home effects, a feature not present in most other studies.

2.5.2 Predictive ability summary statistics

Predictive ability summary statistics can be used either to test the effect of model enhancements, such as adding new covariates into the mean function, or in order to find optimal values for parameters with respect to predictive ability.

Knorr-Held's (2000) paper proposes four measures that could be suitable for evaluating the predictive ability of a sports model. Noting that the only outcomes in Knorr-Held's model are win/draw/lose, then given a total of N matches, and R possible outcomes (R=3 in this case), let $\hat{p}_k^r$ denote the estimated probability that the result of match k will be r, where $r \in (1,\ldots,R)$ and $k \in (1,\ldots,N)$. Note that $\hat{p}_k^r$ is calculated only using data available prior to match k. Also, let the observed result be denoted by s for each match, hence $\hat{p}_k^s$ is the probability of each observed result, as estimated by the model prior to match k.

The four measures are defined as:

1. the number of correctly predicted results, where the predicted result is the outcome with the highest estimated probability.
2. $\frac{1}{N}\sum_{k=1}^{N} log(\hat{p}_k^s)$
3. $-\frac{1}{N}\sum_{k=1}^{N}((1-\hat{p}_k^s)^2 + \sum_{r\neq s}(\hat{p}_k^r)^2)$
4. $\frac{1}{N}\sum_{k=1}^{N}\hat{p}_k^s$

Measure 2 is similar to a measure employed by Dixon and Coles (1997) and Crowder *et al* (2000) and will be referred to as the *predictive likelihood*. Generalising it to be able to calculate the predictive abilities of models which provide probabilities for the entire

set of scores rather than just win/draw/lose, it can be defined as follows. If t denotes the time at which match k takes place and Θ_t represents the parameter estimates based on all data available up to, but not including, time t then the predictive likelihood is defined as

$$\sum_{k=m}^{N} log(P(X_k, Y_k|\Theta_{t(k)}))$$

where X_k, Y_k are the observed home and away scores and m is the first match for which predictions are made[1]. It must be used with some caution, however. Firstly, it is sensitive to outliers, although this is less of a problem if it is used only to compare nested models. Secondly, it is not robust to mis-specification of the response distribution.

Measure 1 is only suitable where the number of possible outcomes is small, although a similar measure to compare two models could be to count the number of occasions one set of predictions is closer to the observed score, and to verify if the proportion is significantly different from 0.5. This can be done via a straightforward binomial signs test. A consequence of using such a measure is that the magnitude of error is not considered. As a result, this measure may not pick up model deficiencies particularly well.

Measure 3 is a quadratic loss, while measure 4 is similar to measure 2. However, measure 4 has the disadvantage that if a result occurs to which the model had assigned an extremely low, or even zero, probability, the measure isn't greatly penalized.

2.6 Comparing predictions obtained through a statistically-based procedure with the bookmaker's odds

Some of the previous studies considered here attempt to compare the accuracy of the model predictions with the accuracy of the probabilities inferred from the bookmaker's odds. In order to do this, a suitable definition of 'accuracy' is needed. Stefani (1980) uses the absolute average difference between the predicted score and the observed score for both College Football and NFL games. Harville (1980) also uses this statistic in order to compare the accuracy of predictions from an NFL scores model with a bookmaker's line. Another measure considered is the squared difference between predicted and observed score, which penalises larger discrepancies more severely. Another measure used by Harville is the proportion of occasions the prediction system correctly

[1]Many models require a 'burn-in' period so that predictions are only evaluated once sufficient data has been observed to make reasonable estimates of parameter values.

predicts the winner of a fixture.

Stefani conducted a year-by-year comparison between the statistical model and the bookmaker's line and concluded that the bookmaker's line was consistently more accurate each year. The general conclusion to Harville's comparisons was that the bookmaker's predictions were more accurate at the start and end of an NFL season, while the model performed better during the middle of the season. It is suggested that at the start of the season the bookmaker takes account of factors such as roster changes, injuries and pre-season exhibition game results, while at the end of the season the importance of late-season matches differs from team to team (this is discussed in more detail in Chapter 5). The model implemented by Harville is based solely on match scores (excluding exhibition games) hence does not accommodate such information.

Glickman and Stern (1998) comment that their NFL predictions' Mean Square Error was smaller than that of the bookmakers, and also claim that for 65 out of the 110 validation matches the model predictions would have produced winning bets. They also comment that 'for this small sample, the model fit outperforms the point spread, though the difference is not large enough to generalise'.

2.7 Betting strategies based on model predictions

Harville (1980) suggests that bets could be made if the following ratio exceeds 0.5 by a sufficient amount:

$$\frac{P(S_k > B_k)}{P(S_k > B_k) + P(S_k < B_k)}$$

where S_k represents the score predicted by the model and B_k represents the bookmaker's line. The paper states that 'the proposed betting scheme would generally have shown a profit during the 1971-77 period', however it is not stated whether the bookmaker's overround (as explained in Section 1.3) is included.

Dixon and Coles (1997) use a betting strategy similar to Harville's although they also adjust for the bookmaker's overround. Hence, repeating the notation used to describe the four measures suggested by Knorr-Held in Section 2.5.2, if they estimate the probability of outcome r in match k to be p_k^r while the bookmaker's probability (converted from the 'odds' format described in Section 1.3.1) is b_k^r, then the expected gain by placing a unit stake on outcome r in this match is

$$\frac{p_k^r}{b_k^r} - 1 \tag{2.7.1}$$

(note that $\sum_{r=1}^{r=R} b_k^r > 1$ for any bookmaker, which reflects their overround). So bets should be placed provided the value in Equation 2.7.1 exceeds some cut-off value ξ. Using predictions generated by a statistical model, Dixon and Coles simulate such a strategy for different values of ξ during the 1995/96 English League soccer season and discover that overall profit can be made for $\xi > 0.15$. There is considerable variance in this profit and the 90% bootstrap confidence intervals of the realised profit when $\xi > 0.15$ generally include 0, and indeed also the loss that one would realise on average were bets placed randomly. Nevertheless there is some indication that their predictions, from a relatively simple model, can form the basis of a profit-making betting strategy.

Rue and Salvesen (1997) suggest that a betting strategy could take account not just the expected profit from making a bet but also the variance of that profit. Hence bets should be placed with regard both to maximising profit but also restricting the probability of ruin. Defining P as the profit on a bet, μ_r^k and σ_r^k as the expected profit and standard deviation for betting a unit amount on outcome r on match k, β_r^k to be the proportion of capital to be staked on outcome r of match k and $\mathcal{B}$ to be the set of matches which can be bet on, then the optimal values of β_r^k can be found by maximising

$$E(P) - Var(P) = \sum_{j \in \mathcal{B}} \beta_r^k (\mu_r^k - \beta_r^k (\sigma_r^k)^2)$$

The solution to this is $\beta_r^k = max(0, \frac{\mu_r^k}{2(\sigma_r^k)^2})$. By simulating betting throughout the 95/96 and 96/97 Premier League soccer season, profits of 47% and 22% were returned on original capital.

More complex strategies than this can be considered. In particular, one could take into account the amount of capital available and the utility of money. In addition, many recommended betting strategies assume 'correct' probabilities are available, whereas ideally, the distribution of the estimates of the parameters which form the estimated probabilities should be considered.

This chapter has discussed some important issues concerning the modelling process from a statistical perspective. There are other criteria to consider besides these. It is important that a method does not require excessive computational resources in order to be used. In practice, development of a model is generally performed in stages, where flaws in the model assumptions or methodological errors become apparent through trial and error. Thus the computation time required to implement any stage of the modelling process has to be short enough for the model development to take place on

a practical timescale.

An additional issue is the availability and cost of data required by a process. There may be situations where an alternative method is attractive from a statistical perspective in that it may, for example, produce estimates of quantities that have lower variance, or have lower expected bias, than an existing method. However, these improvements may only be observable given a suitably large amount of data. Consideration must be given towards how much data is likely to be necessary and whether such a quantity of data can be obtained at an affordable cost, in terms of money or time, before deciding to implement an alternative method in a situation such as this.

Chapter 3

A general method for obtaining parameter estimates

In the previous chapter some of the key issues involved in the modelling of sports results were described and a selection of previous treatments were summarised. Some of this material is helpful to raise awareness of the potential problems that arise, while some is of more direct importance since it can be applied, with minor modifications, to the sports markets that are to be modelled in this thesis.

To clarify the objective of this chapter, in terms of how it ties in with the other material in this thesis, it is helpful to outline the general procedure involved in modelling a sport. It can be considered as a three stage process:

1. Specification of a model for the sport, on consideration of issues listed in Sections 2.1.1 to 2.1.4. This is generally done using both one's existing knowledge about a sport, and by exploring the available data. This stage concludes with the specification of a statistical model which relates the parts of data that are deemed to be important to each other via a set of parameters (such as team abilities).

2. A procedure is implemented in order to estimate the values of the parameters included in the model specified in stage 1.

3. Using the estimates obtained in stage 2, the validity of the specified model is evaluated. If the model is considered to be satisfactory, the estimates have a number of possible uses. They may be informative in themselves since the ranking of teams by ability, or the average effect of a covariate on the outcome of fixtures may be of interest to the statistician. The estimates can also be used to generate

predictions for future sporting fixtures and these predictions can form the basis of a betting strategy.

Chapters 4, 5 and 6 mainly cover stages 1 and 3 of this process. Stage 2 is covered in this chapter. The material in this chapter is quite technical although an exhaustive understanding of it all is not necessary in order to follow the developments in the main sections of this thesis, which, in Chapters 4, 5 and 6, is the construction and application of sports specific models. This chapter can be read in its entirety if the reader is interested in the technical aspects of the parameter estimation process, otherwise the reader may find it more useful for occasional reference, where indicated in the text, while reading Chapters 4, 5 and 6.

The procedure outlined in this chapter is based on a procedure employed by Dixon and Coles (1997). The original application was the modelling of UK soccer scores and the procedure they used can be extended quite easily to model other sports. It is by no means the only technique that has been applied in studies of sports modelling but it is used by all models in Chapters 4, 5 and 6. Some other parameter estimation procedures are mentioned in Section 2.4, one of which (the Markov Chain Monte Carlo approach) is applied to a model for NFL scores in Chapter 7. For the models elsewhere in this thesis, the procedure outlined in this chapter is considered to be more suitable. The strengths and weaknesses of it compared to the Markov Chain Monte Carlo approach are discussed in Chapter 7.

3.1 The Dixon-Coles MLE procedure - an introduction

In this section the procedure employed by Dixon and Coles (1997) to obtain estimates of model parameters of English soccer teams in summarised. From here on this procedure is referred to as the *MLE method*. Initially the model specification that Dixon and Coles chose is described. It is assumed that home and away goals follow independent Poisson distributions. Given N matches in total, then for match k between teams $i(k)$ and $j(k)$, the probability associated with each match score is

$$L_k = P(X_k = x_k, Y_k = y_k) = (\frac{e^{-\lambda_k}\lambda_k^{x_k}}{x_k!})(\frac{e^{-\mu_k}\mu_k^{y_k}}{y_k!}) \tag{3.1.1}$$

where

- X_k and Y_k are the number of home and away goals,

- $\lambda_k = e^{\alpha_{i(k)}+\beta_{j(k)}+\delta}$ and $\mu_k = e^{\alpha_{j(k)}+\beta_{i(k)}}$
- $\alpha_{i(k)}, \alpha_{j(k)}$ represent the home and away sides' attacking capabilities,
- $\beta_{i(k)}, \beta_{j(k)}$ represent the home and away sides' defensive capabilities,
- δ represents the effect of playing at home.

In this way, probabilities $L_1, ..., L_N$ for each match are obtained. Standard likelihood maximisation procedure suggests maximising the sum of the logs of these N probabilities with respect to the (α, β, δ) parameters in order to obtain maximum likelihood estimates $(\hat{\alpha}, \hat{\beta}, \hat{\delta})$. However, to do so in this case assumes that all parameters, including team abilities, are fixed over time which in practice is not believed to be the case. Various treatments of this problem for other modelling frameworks are outlined in Section 2.3. The MLE method uses a 'weighting' factor, Υ_k, for each match. Hence the pseudo-loglikelihood to be maximised is

$$\sum_{k=1}^{N} log(L_K)\Upsilon_k \tag{3.1.2}$$

The parameter Υ_k should be larger the more recently the match took place. The form for Υ_k chosen by Dixon and Coles is

$$\Upsilon_k = exp(-\varsigma(t - t_k))$$

where t is the current time, t_k is the time match k took place and $\varsigma < \infty$ is a coefficient chosen in order to maximise the predictive ability of the model, rather than the loglikelihood specified in Equation 3.1.1. Note that in the interests of readability, this pseudo-loglikelihood is referred to as the 'likelihood' throughout this thesis.

From here on, Υ_k is referred to as an *external* parameter, while the team, global mean and home effect parameters, which are maximised at each time-point as part of the likelihood, are referred to as *internal* parameters. There is no algebraic solution to finding the maximum likelihood estimates of the internal parameters but Newton-Raphson maximisation techniques can be used without major difficulty.

In order to assess the predictive ability of the model, a scalar quantity referred to as the *predictive likelihood* (PL) can be used. It is defined as the sum of the loglikelihoods of the observed scores given the predicted scores based only on data available up until

the time of the match. Hence

$$PL = \sum_{k=m}^{N} log(P(X_k, Y_k)|\Theta_{t(k)}) \tag{3.1.3}$$

where m denotes the first match after which sufficient data has been observed in order to be able to make reliable estimates of the parameters, $t(k)$ is the time at which match k takes place and Θ_t is the set of (α, β, δ) estimates based on all matches up to but *not* including time t. This sum must be computed for a range of values of ς until an approximate maximum value of PL has been found.

Before the likelihood can be maximised based on the model suggested, one remaining problem is that the α, β parameters are unconstrained. Thus there is no unique solution to the likelihood maximisation, since a constant can be added to all the α's and subtracted from all the β's without affecting any of the score predictions. Dixon and Coles introduced the constraint that $\frac{1}{N}\sum_{i=1}^{N} \alpha_i = 0$ to achieve a unique maximum likelihood. The effect of this along with some alternative solutions, are now discussed.

3.2 Modifications and extensions

3.2.1 Constraints on team ability parameters

In order to generalise the MLE method described in the previous section so that parameter estimates for other sports can be obtained, some modifications are required. For example, Dixon and Coles have all English soccer results from all divisions, plus results from cup games, in their database and so it is fairly rare that a new team enters into their likelihood. In certain other situations, such as the yellow cards application in Chapter 4 where only Premier League data is employed, teams frequently enter and leave the data set. As a result, the sum-to-zero constraint applied to the teams' parameters to ensure a unique solution to the maximum likelihood could be problematic. The following simplified version of a league system illustrates this.

Suppose that data from two seasons of a league system is to be modelled. The league contains teams A,B,C in the first season and teams A,B,C and D in the second season. Each team's ability is constant throughout time and can be summarised by a single parameter. These abilities relative to team A's are (0, 0.3, -0.5, -1.0). The interpretation of these parameters is that team B on average beats team A by 0.3 goals, for example and similarly for teams C and D. Furthermore, an intercept term γ

is required so that if team A were to play a team of equal ability, on average a total of 3.0 goals would be scored. The score of a match is not affected by whether the match is played at the home ground of either side in the match.

The model specification can be expressed as follows: if X, Y respectively represent the scores of teams i and j in a match then

$$\begin{aligned} E[X+Y] &= \gamma + \alpha_i + \alpha_j \\ E[X-Y] &= \alpha_i - \alpha_j \end{aligned}$$

The simultaneous equations to be solved in order to convert the listed team abilities so that they satisfy a sum-to-zero constraint are, in matrix form:

$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \gamma \\ \alpha_A \\ \alpha_B \\ \alpha_C \end{pmatrix} = \begin{pmatrix} 3.3 \\ 2.5 \\ 2.8 \\ -0.3 \\ 0.5 \\ 0.8 \\ 0.0 \end{pmatrix}$$

The solution to this is

$$(\gamma, \alpha_A, \alpha_B, \alpha_C) = (2.87, 0.067, 0.37, -0.43)$$

However, with the addition of team D to the data set, the simultaneous equations to be solved in order to satisfy a sum-to-zero constraint are:

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \gamma \\ \alpha_A \\ \alpha_B \\ \alpha_C \\ \alpha_D \end{pmatrix} = \begin{pmatrix} 3.3 \\ 2.5 \\ 2.0 \\ 2.8 \\ 2.3 \\ 1.5 \\ -0.3 \\ 0.5 \\ 1.0 \\ 0.8 \\ 1.3 \\ 0.5 \\ 0.0 \end{pmatrix}$$

to which the solution is

$$(\gamma, \alpha_A, \alpha_B, \alpha_C, \alpha_D) = (2.4, 0.3, 0.6, -0.2, -0.7)$$

The problem here is that while the abilities of teams A,B and C have not changed, α_A, α_B and α_C have changed in order to satisfy the sum-to-zero constraint. The predictions of match results are still valid but the parameter estimates become less interpretable. One could choose the alternative constraint that the ability of team A is always set to be zero but this means that the abilities of teams B,C and D can only be expressed with respect to the ability of team A which, for real life applications, would be likely to change over time.

Another approach is to include a prior on the team abilities when the likelihood is maximised. Referring back to the example described by Equation 3.1.1, the most natural prior assumption to place on the team abilities is to assume they are Normally distributed with mean zero. In addition, the overall mean scoring rate can be accommodated by the inclusion of a global mean parameter γ. Hence the optimal estimates of α, β, γ and δ maximise

$$\prod_{k=1}^{M} P(X_k, Y_k | \alpha_{i(k)}, \alpha_{j(k)}, \beta_{i(k)}, \beta_{j(k)}, \gamma, \delta) \prod_{i=1}^{N} \pi(\alpha_i) \prod_{i=1}^{N} \pi(\beta_i)$$

where

$$\pi \sim \mathcal{N}(0, \tau_{\alpha\beta})$$

$\tau_{\alpha\beta}$ is an external parameter, hence the optimal value of $\tau_{\alpha,\beta}$, like ς, must be found by inspection of the predictive likelihood defined in Equation 3.1.3.

It should be clarified at this stage that the prior distribution referred to here does not serve the purpose conventionally served by a prior term in the context of Bayesian statistics. The Bayesian interpretation of the prior term used here would be that before any data has been observed, it is believed that all teams have equal α's and β's and that this belief is modified upon observing data. This is not the reason for the inclusion of the prior term in this case. The prior term here serves an entirely different purpose, which is to act as a constraint on the estimates of the parameters so that the likelihood can be maximised.

3.2.2 Application of prior values to other parameters

One problem caused by the MLE method for parameter estimation is that by down-weighting matches that took place less recently in the likelihood, information about all parameters (rather than just the team abilities) is down-weighted. Thus when the parameters for the global mean, γ, and the effect of playing at home, δ, are estimated recent matches are given greater weight. Since these parameters are considered in practice to be constant throughout time, this is not desirable. This same issue applies to the estimation of the score standard deviation term if a Normal distribution is employed for the scores, for example, and the correlation coefficient if a bivariate Normal distribution is used. Hence, at the start of a season in particular, the estimates of these parameters are based largely on recent results. While in some cases it may be desirable that the global mean and other terms vary with time to some extent, there is a danger of excessive variation occurring. This problem can also be addressed by including more prior terms in the likelihood. This treatment fits more naturally into the Bayesian modelling philosophy. Hence, for example, by specifying that

$$(X_k, Y_k) \sim \mathcal{N}_2(\gamma + \alpha_{i(k)} + \beta_{j(k)} + \delta, \gamma + \alpha_{j(k)} + \beta_{i(k)}, \sigma_X, \sigma_Y, \rho)$$

then a suitable likelihood function could be

$$L(\mathbf{X},\mathbf{Y}|\alpha,\beta,\gamma,\delta,\sigma_X,\sigma_Y,\rho) = \prod_{k=1}^{M} P(X_k,Y_k|\alpha_{i(k)},\alpha_{j(k)},\beta_{i(k)},\beta_{j(k)},\gamma,\delta,\sigma_X,\sigma_Y,\rho)$$
$$* \quad \pi(\gamma)\pi(\delta)\pi(\sigma_X)\pi(\sigma_Y)\pi(\rho)\prod_{i=1}^{N}\pi(\alpha_i)\pi(\beta_i) \qquad (3.2.1)$$

where the $\pi(.)$ terms are such that

- $\gamma \sim \mathcal{N}(\gamma_0,\tau_\gamma)$
- $\delta \sim \mathcal{N}(\delta_0,\tau_\delta)$
- $\sigma_X \sim \mathcal{N}(\sigma_{X0},\tau_{\sigma X})$
- $\sigma_Y \sim \mathcal{N}(\sigma_{Y0},\tau_{\sigma Y})$
- $\rho \sim \mathcal{N}(\rho_0,\tau_\rho)$

It is then necessary to choose appropriate mean $\gamma_0,\delta_0,\sigma_{X0},\sigma_{Y0},\rho_0$ and variance $\tau_\gamma,\tau_\delta,\tau_{\sigma X},\tau_{\sigma Y},\tau_\rho$ values for the priors described above.

Selecting prior mean values

The prior values for γ_0 and δ_0 at time t could respectively be $\overline{\mathbf{Y_k}}|(t(k)<t)$ and $(\overline{\mathbf{X_k}}-\overline{\mathbf{Y_k}})|(t(k)<t)$. It is more difficult to select suitable initial values for the σ_{X0},σ_{Y0} and ρ_0 terms. For example, σ_X and σ_Y are *conditional* standard deviations, conditional on covariates including team parameters and the effect of playing at home. Thus in order to produce a reliable estimate of $(\sigma_X|\alpha,\beta,\gamma,\delta)$, suitable estimates for the α,β terms, for example, are necessary. Estimates for these parameters can only be obtained by maximising the quantity in Equation 3.2.1. Yet it is for this process that suitable values of σ_{X0} and σ_{Y0} are required. Similarly, ρ is a *conditional* correlation, so a similar argument applies. There are still various possible estimates for σ_{X0}, σ_{Y0} and ρ_0 that can be considered. For example, non-time-dependent estimates of the team parameters can be obtained using a more straightforward parameter estimation process such as fitting a generalized linear model (if the probability distribution chosen for scores is not Normal) or least squares regression (if Normally distributed scores are assumed). Estimates of the standard errors and correlation conditional on these estimates could then be calculated. It is rather time consuming to repeat this process on every occasion that the model parameters need to be estimated and so setting σ_{X0}, σ_{Y0} and ρ_0 to

be the unconditional standard deviation of home scores, the unconditional standard deviation of away scores and the unconditional (home score, away score) correlation is a straightforward alternative. This will normally give inflated estimates for the σ_X and σ_Y terms, since team abilities account for some of the variance in almost all of the situations which are investigated in this thesis. Techniques to scale down these figures could be considered, although these would be chosen in order to maximise the predictive ability of the model, along with the other external parameters.

Selecting prior variance values

Suitable variance quantities $\tau_\gamma, \tau_\delta, \tau_{\sigma X}, \tau_{\sigma Y}, \tau_\rho$ for the priors are also required. A quantity that allows sufficient movement from the initial estimate of the parameter, without allowing excessive fluctuation of the estimate (which could bias predictions of future results) is desirable. One obvious candidate is the standard error of the initial value described in the above paragraph. This can be obtained either through formulae if possible (see below), or alternatively a simple model with no team effects can be maximised. Standard errors of the terms of interest can be obtained by taking the diagonal terms of the inverse of the information matrix.

Hence it is natural to define

$$\tau_\gamma = \sqrt{\frac{Var(\mathbf{Y_k}|(t(k) < t))}{N}}$$

$$\tau_\delta = \sqrt{\frac{Var(\mathbf{X_k} - \mathbf{Y_k})|(t(k) < t)}{N}}$$

As previously discussed in this section, selecting appropriate values for σ_{X0}, σ_{Y0} and ρ_0 is problematic thus $\tau_{\sigma X}$, $\tau_{\sigma Y}$ and τ_ρ may need to be selected to create suitably weak priors.

3.2.3 Including additional covariates

The only covariates in the models specified so far in this chapter are two ability parameters for each team and the effect of playing at home. However, as discussed in Section 2.1.4, there are often additional covariates that may improve the accuracy of predictions. The starting point is to assume a linear, or loglinear, relationship between the covariate and the response. One straightforward model for soccer scores (X_k, Y_k) could involve using the attempted goals, or shots, (HS_k, AS_k) as covariates, with two

extra parameters κ_1 and κ_2:

$$\begin{aligned} E[X_k|HS_k, AS_k] &= exp(\gamma + \alpha_{i(k)} + \beta_{j(k)} + \delta + \kappa_1 * HS_k + \kappa_2 * AS_k) \\ E[Y_k|HS_k, AS_k] &= exp(\gamma + \alpha_{j(k)} + \beta_{i(k)} + \kappa_1 * AS_k + \kappa_2 * HS_k) \end{aligned} \quad (3.2.2)$$

A model for the prediction of (HS_k, AS_k) could be developed and combined with that described in Equation 3.2.2 to obtain a joint distribution for (X_k, Y_k, HS_k, AS_k), which may give more accurate marginal distributions for (X_k, Y_k).

The problem that affects the estimation of global parameters such as γ and δ, which is described in Section 3.2.2, also affects the estimation of the κ_1 and κ_2 terms in Equation 3.2.2. The true values of these parameters are not considered in practice to vary over time but the parameter estimation procedure, as described thus far, places greater emphasis on recent results when κ_1 and κ_2 are estimated. The problem is compounded if the covariate is an indicator variable representing a rare or seasonal event. An example of this could occur in soccer where a variable Z_k could be defined so that

$$Z_k = \begin{cases} 1 & \text{if both teams in match } k \text{ are threatened by relegation from the league should} \\ & \text{the match be lost} \\ 0 & \text{otherwise} \end{cases}$$

After the first match of a season where $Z_k = 1$, the estimate of its coefficient is heavily affected by the result of this match, rather than averaged out over that match and all others in previous seasons as desired. While this is also true for parameters such as the global intercept γ or home effect δ, it is less critical since the presence of γ and δ in the specification of the conditional mean of every match ensures that their estimates are based on reasonably large quantities of data.

There is no 'clean' way of solving this problem within the MLE method framework, but the following correction technique can be employed. Let

- Λ represent all time-dependent team ability parameters
- Υ represent all non-time-dependent covariate parameters such as κ_1 and κ_2, and global parameters such as γ and δ.
- $(\mathbf{X}, \mathbf{Y})$ represent all data

Next, the following functions are defined:

- $l_1(\Upsilon|\Lambda^*) = \prod_{k=0}^{M} P(x_k, y_k|\Lambda = \Lambda^*, \Upsilon)$
- $l_2(\Lambda|\Upsilon^*) = \prod_{k=0}^{M} P(x_k, y_k|\Lambda, \Upsilon = \Upsilon^*) exp(-\varsigma(t - t_k))$

Initially Λ^* is a vector of zeroes of length $2N$, hence all offensive and defensive parameters are set to zero. Next, $l_1(\Upsilon|\Lambda^*)$ is maximised in order to obtain non-time-dependent estimates Υ^*. Then, using this value, $l_2(\Lambda|\Upsilon^*)$ is maximised to obtain Λ^{**}. Hence Υ^* is obtained by giving equal weight to all matches, but assuming that all teams are of equal ability, while Λ^{**} is obtained by giving more importance to recent matches. Team abilities are estimated subject to a restricted value of Υ^*.

One could next consider maximising $l_1(\Upsilon|\Lambda^{**})$ to obtain Υ^{**} and repeating the process described above until some desired number of iterations has been implemented but this would be time-consuming and would also require a good understanding of the behaviour of both l_1 and l_2, which is rather difficult given their large dimensionality. For this reason only one implementation has been used. In certain situations it is more appropriate to maximise l_2 before maximising l_1. Where implemented in subsequent chapters in this thesis, the process described above has been modified so that the global parameters such as γ, δ and σ_X have been re-evaluated along with Λ, and a prior term for them has been included as outlined in Section 3.2.2. The justification is that their ubiquity in the likelihood function means they are less sensitive to the loss of information brought about by the inclusion of a down-weighting term in the likelihood maximisation process.

3.2.4 Season breaks

One further issue that arises as a result of the down-weighting system employed by the MLE method, although it also applies to any analysis which attempts to allow estimates of teams' abilities to adjust over time, is the need to accommodate the break that occurs between the seasons of most sports. The MLE method as outlined so far assumes that team abilities adjust at the same rate over the season break as they do during the season. This seems like an unrealistic assumption and so the solution used in this thesis is to add, for every season before the current one, a between-season-truncation adjustment w on to the time-points of matches when the likelihood is maximised. Hence the time-points of the matches during the season prior to the current one have w added to them while the time-points of the matches during the

season prior to that one have $2w$ added etc. This quantity is an external parameter and this also has to be deduced by inspection.

3.3 Effect of the extensions on the maximisation of the predictive likelihood

The external parameters, i.e. offensive and defensive tightnesses $\tau_{\alpha\beta}$, time down-weighting parameter ς and between-season-truncation adjustment w, all have to be chosen in order to maximise the predictive likelihood. Since the functional form of the predictive likelihood is too complicated to analyse algebraically, the only way to find the optimal values is by using the rather crude technique of trying out many sets of values, recording the predictive likelihood each time and choosing the set which corresponds to the highest value of predictive likelihood. This technique is only valid if the surface of the predictive likelihood is reasonably well-behaved, in particular it helps if it is unimodal. It should be noted that by using such a procedure, valid comparisons between predictions obtained and those offered by bookmakers can only be made if the optimal values for the external parameters are found using data which occurred prior to the sample which is to be compared to bookmakers' predictions. Hence, in order to have genuine comparisons between model and bookmaker predictions, one must divide the data set into two sections. The earlier section is used in order to find optimal values for $\tau_{\alpha\beta}$, ς and w. Then the updating of all non-external parameters is performed at each time-point on the latter section of the data using the optimal values of the external parameters. Predictions are then made using the most recent parameter estimates and these predictions can be compared with the bookmaker's. If sufficient time and computing resources are available, re-evaluation of the external parameters could be performed at every time point, and summary statistics on the comparison between bookmaker and model predictions could be computed at every stage.

3.4 Model comparison techniques

The most frequently used statistic in this thesis in order to assess the validity of models is the predictive likelihood, as defined by Equation 3.1.3. Strictly speaking it is a device used in stage 3 of the modelling process outlined in the introduction to this chapter, which deals with model evaluation. This chapter focuses on stage 2 of this process,

however, given the frequent use of the predictive likelihood throughout the next three chapters, it seems appropriate to include a short discussion of it here.

The statistic must be used with some caution as the following, rather extreme, example illustrates. Suppose the following scores are observed,

$$(1, 0, 8, 2, 2, 1, 2, 1, 1, 3)$$

and prior to each match it was believed that each score had an expected value of $\mu = 2.1$. In this case a predictive likelihood of -19.895 is obtained, assuming the scores follow a Poisson distribution. Alternatively, suppose it was believed that each score had an expected value of $\mu = 2.01$. Here the predictive likelihood becomes -19.915. Hence a superior predictive likelihood is obtained assuming $\mu = 2.1$ whereas if it is assumed that $\mu = 2.01$, closer predictions for eight of the ten scores are obtained. For spread betting, one should asymptotically make more profit using $\mu = 2.1$ as the prediction, since returns are proportional to the closeness of the predictions. For fixed odds betting one would lose rather a lot of money by assuming $\mu = 2.1$. In fact, by looking at the logs of the observed probabilities of the scores given $\mu = 2.1$ and Poisson distributed responses:

(-1.358,-2.100,-6.769,-1.309,-1.309,-1.358,-1.309,-1.358,-1.358,-1.666),

it can be seen that the third score is atypical (which one cannot always see by looking at the raw data of more complicated data sets than in this example). This suggests that either the third score is an outlier, an extreme event, or that an extra covariate to describe a characteristic feature of the third match is required. Some general understanding of the sport being modelled may be important in order to decide this.

The material covered in this chapter has been selected with the aim of describing methods that are common to all three markets that are covered in the next three chapters. It also suggests a suitable method for other studies of similar sporting markets. Each study must extend or modify these general methods in order to accommodate the specific features of the market. The next three chapters illustrate this.

Chapter 4

Harsh referees and dirty teams: estimating booking rates in soccer

This chapter investigates the rate of bookings in Premier League soccer. Motivated by the rapidly growing and financially lucrative sports spread betting markets, the aim is to estimate the distribution of the numbers of cautions and dismissals (yellow and red cards) given by the referee in a particular future match. This is achieved using a detailed statistical model to account for the characteristics of the two teams playing, the referee and several other factors. The aim is to obtain predictions that could be used as the basis for a profit making strategy on UK sports spread betting markets.

This chapter presents the first application of the likelihood maximising procedure outlined in Chapter 3 in order to obtain estimates of parameter values. This application is introduced in Section 4.1. The basis of the model development is past data in the form of numbers of home and away yellow and red cards observed in Premier League soccer from 1994-2001, which is explored in Section 4.2. In Section 4.3 the adjustments required to model the bookings process and fit the model to the data are discussed. In Section 4.4 the results are reviewed and the utility is examined of the model using approximately 1150 booking spread prices. In Section 4.5 some possible improvements to the model are suggested and Section 4.6 gives the conclusions to this chapter.

4.1 Bookings in soccer - an overview

Soccer, or Association Football, has been played with the same basic rules for over 100 years. The main change over this period has been the increase in its popularity and the financial consequences for good or bad performance. As soccer and its participants have become more professional, and success has become more important, players must play close to the boundaries of the rules, and inevitably sometimes break them. To avoid unfair advantages to rule-breakers, and to ensure the safety to participants, match officials (referees) are given the power to penalise a player who commits a serious breach of the rules, or who continually commits minor offenses. Penalties can range from a free-kick through to cautioning and ultimately dismissing (sending off) a player. Every caution or dismissal by a referee is indicated to the offending and other players by clearly displaying a yellow card (for a booking or caution) or a red card (for a dismissal). The likely number of red and yellow cards to be shown in the match differs from game to game depending on various factors: some players are more prone to committing punishable offenses, while some referees tend to caution and dismiss more readily than others. Estimating the distribution of the number of red and yellow cards in a given match and investigating the influence of such factors are the subjects of the chapter.

There have been a number of studies of both the statistics and the psychology of the bookings process, with a variety of motivations. For example, Ridder *et al* (1994) examine the effect of a red card on the outcome of the match, and even suggest situations in which it might be advantageous to commit deliberate (unethical) fouls. The aim here is rather different and is motivated by the opportunity of spread betting, which is described using an example from this particular market in Section 1.3.2. The volumes of bets on bookings markets can be huge: in fact for some firms it is the most popular form of betting, so as a consequence there are strong financial incentives to both bookmakers and gamblers for models that can accurately estimate the probabilities of various outcomes of bookings markets and this is the underlying motivation for this study.

Although spreads are quoted for hundreds of markets and for many sports, what makes the study of bookings markets appealing is the apparent difficulty in estimating the mean number in a given match. This difficulty is reflected in the prices offered by different spread betting companies who offer spreads independently of each other.

Although prices are generally driven by gambler behaviour, the opening, or initial spreads are quoted based on the spread companies' subjective probability of the mean bookings points for the match. While for most sports markets the opening spreads across bookmakers generally agree, for the booking markets, the opening prices are usually very different. In the Arsenal versus Manchester United example detailed in Section 1.3.2, the opening spread offered by one bookmaker was 30-34 but other firms opened at 20-24 and 36-40. This level of discrepancy in opening prices is not atypical.

By developing a detailed statistical model, the aim is to estimate the distribution of the numbers of yellow and red cards in a given match, and in addition to gain an understanding of the bookings process. The fact that there is no existing literature on the development of such a model that the writer is aware of is likely to be due to lack of motivation: before the spread markets became popular, there was little desire to know such probabilities.

4.2 UK Bookings data

The data available include Premier League soccer matches since the start of the 1994/1995 season. The available data vary in detail: for later matches the referee name is available, whereas for the early matches only the home and away red/yellow card numbers are recorded. The data is split into three parts.

- Aug 1994- May 1997. Home/away red/yellow card numbers. (1222 matches)
- Aug 1997-May 1999. Home/away red/yellow card numbers with referee names (760 matches).
- Aug 1999-May 2002. Home/away red/yellow card numbers with referee names, and most spread betting prices (1140 matches). Table 4.1 gives the first five matches in this data set.

Table 4.1: Five lines of the dataset

referee	date	home team	away team	home score	away score	hm. yels.	aw. yels.	hm. reds	aw. reds	spread
DGallagher	20010421	Arsenal	Everton	4	1	0	2	0	1	37
NBarry	20010421	Bradford	Derby	2	0	0	2	0	1	41
MDean	20010421	Chelsea	Charlton	0	1	3	0	0	0	37
GBarber	20010421	Ipswich	Coventry	2	0	1	4	0	0	35
GPoll	20010421	West Ham	Leeds	0	2	5	2	0	1	48

The raw data for the second and third part of the data set are displayed in Figure 4.1 which displays histograms of numbers of yellow and red cards. Figure 4.1 suggests

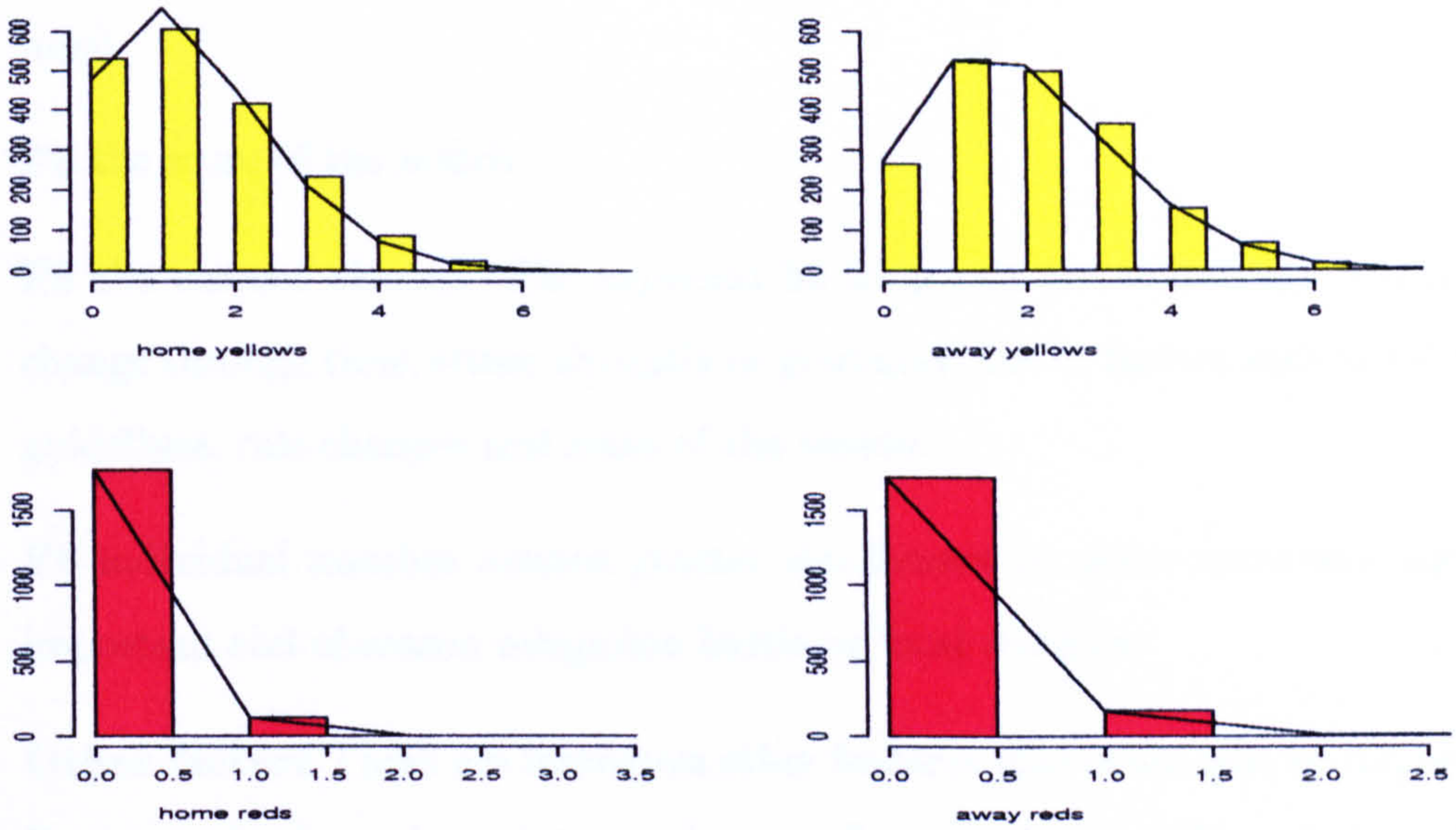


Figure 4.1: Histograms of yellow and red cards, with Poisson distribution lines overlaid.

sides tend to collect more bookings when playing away from home.

With the number of yellow cards being discrete and generally quite low, the Poisson distribution appears to be the most natural probability distribution to employ here. The overlaid lines on the plots in Figure 4.1 represent Poisson probabilities whose parameters are the overall means of the displayed data. While the overlaid lines do appear to depart slightly from the histogram, note that each match has a different expected number of yellows, so an exact fit of a Poisson model to the collated data over matches is not expected, even if the distribution of yellows in each match has a Poisson form. Hence it is assumed that yellow cards follow a Poisson distribution from now on.

Concentrating initially on the second section of the data set, there is information on 29 teams and 34 referees, with most referees officiating a match approximately once every fortnight. Based on some initial thought, and simple exploration of the data, the following factors are worth consideration in any given match:

- **F1** the two teams' propensities to pick up bookings (hereafter termed the teams' *dirtinesses*)

- **F2** the two teams' propensities to provoke the opposition into getting booked (hereafter termed the teams' *provocations*)

- **F3** the referee's propensity to give out cards (hereafter termed the referee's *harshness*)

- **F4** the score of the match

- **F5** the current *climate*. The expected booking rate for an average match will change through time, either abruptly or gradually, due to factors such as referees' guidelines, rule changes and state of the season.

- **F6** individual matches assume greater significance on some occasions, e.g. an important end-of-season relegation battle or local rivalries.

- **Other factors** There are numerous other features that could also be important. For example, dependence between home yellows and away yellows, in that if one side collects many yellows, the general match temperature will rise and may provoke fouls from the opposing side. Also the weather, longer-term consequences of a large number of bookings (player suspensions) on individual players, individual player rivalries or friction between certain individuals, crowd intimidation on players and referees, may, among other factors, all be influential on the bookings rate.

In Sections 4.2.1-4.2.4, what are considered to be the main effects, namely factors F1-F6, are explored using empirical summaries of the data.

4.2.1 F1-F3: team dirtiness, team provocation, and referee harshness

During Premier League soccer seasons 1997/98 until 2001/02, Derby collected an average of 2.242 bookings over 190 matches, with a bootstrap confidence interval of (2.050, 2.434). Manchester United collected an average of 1.432 (1.244, 1.620) in the same time period. It is well acknowledged that some teams have players who are more likely to collect bookings. What may be more surprising is that, for example, Leeds provoked on average 2.453 (2.222, 2.684) bookings from their opponents, while during the same period, Southampton provoked only 1.489 (1.295, 1.683) bookings. As for referees, G Barber booked on average 4.269 (3.881, 4.657) players in each match, whereas the equivalent statistic for P Durkin is 2.832 (2.383, 3.281). This suggests definite team and referee specific effects for factors F1-F3. It is interesting to note

that the dirty teams are not necessarily the most provocative as one might expect. For example, only two teams collected more bookings than Nottingham Forest during the 1998/99 season, yet only one team attracted fewer bookings. In fact, for every booking Nottingham Forest provoked, they collected 1.79 themselves. Figure 4.2, which plots the average number of bookings sides attracted against those they provoked in the 1998/1999 season, emphasises this lack of association.

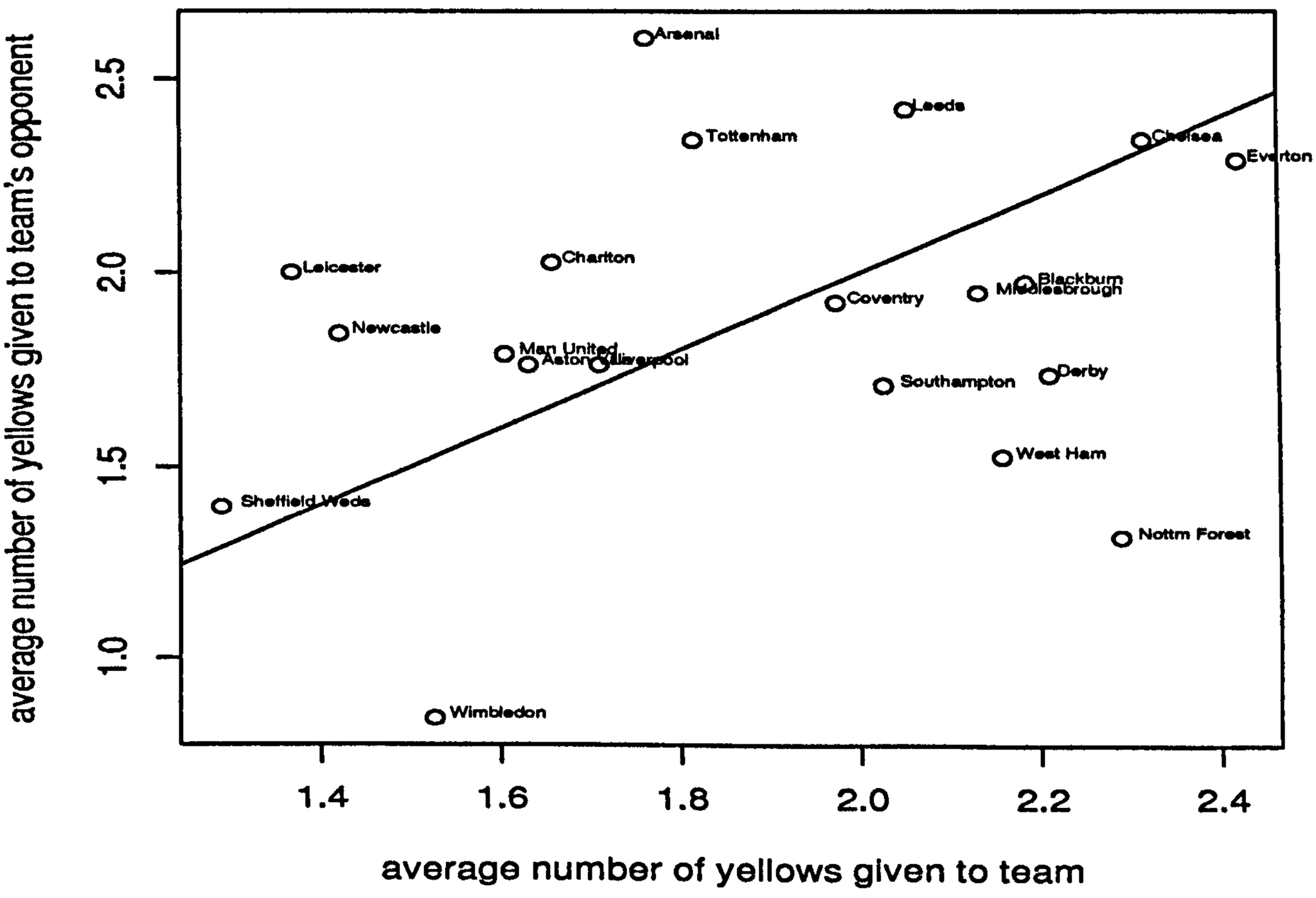


Figure 4.2: Cards collected versus cards provoked, season 1998/1999

4.2.2 F4: the score of the match

Table 4.2: Average cards collected in all matches versus goal difference

goal difference	≤ -5	-4	-3	-2	-1	0	1	2	3	4	≥ 5
cards collected	1.5652	1.987	1.9074	1.9974	1.9725	1.8522	1.6419	1.3714	1.2222	1.1688	0.913

Table 4.2 displays the average number of red or yellow cards collected by a side, compared to the difference in score of the match. There is little doubt that the worse

the result of the match is for a team, the more likely they are to collect bookings.

4.2.3 F5: the climate

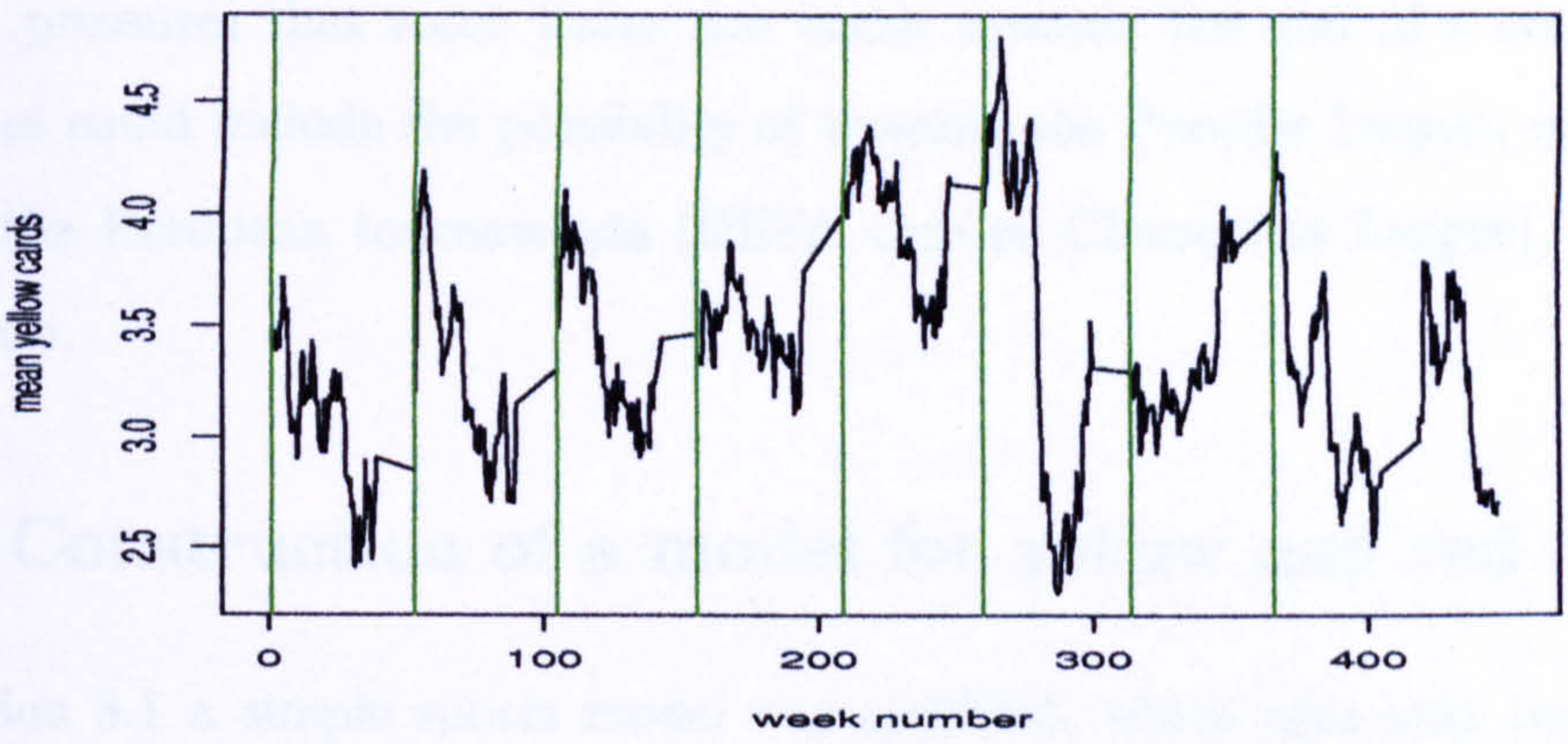


Figure 4.3: Moving average of number of yellow cards awarded in each match. The vertical green lines denote season breaks

Figure 4.3 displays the moving average (block size 50) of the total number of cards given out for all matches since August 1994. It suggests the bookings market has underlying non-stationarity. Particularly noticeable is the trough in the global booking rate that occurred just before week 300. In fact, at around that time (January 2000) the FA issued instructions to all Premier League referees advising them to exercise more caution when issuing yellow and red cards. However, looking at the entire graph, it appears that the awarding of bookings is more 'fashionable' at certain times than others.

4.2.4 F6: match-specific factors

Table 4.3: Bookings data for matches between known rivals

team 1	team 2	number of times played each other	average yellows against each other	average reds against each other	team 1 average cards in all matches	team 2 average cards in all matches
Sunderland	Middlesbrough	6	5.33	0.67	1.93	1.88
Coventry	Aston Villa	8	4.38	0.12	1.8	1.46
Leicester	Derby	10	5	0.2	1.45	2.18
Tottenham	Arsenal	10	4.9	0.3	1.66	1.73
Liverpool	Everton	10	4.2	0.5	1.44	1.94

Table 4.3 displays information concerning the numbers of bookings awarded during matches between various pairs of teams who are recognised as being strong rivals.

Although the information is based on quite a small number of matches, and there are other factors which determine the bookings rate in a match, it does appear that there may be a genuine effect from these rivalries.

On a similar theme, it could be considered whether there is any effect from the various pressures that some teams are under towards the end of a season. These pressures could include the possibility of winning the Premier League, qualifying for one of the European tournaments (UEFA Cup or Champions League), or avoiding relegation.

4.3 Construction of a model for yellow and red cards

In Section 3.1 a simple sports model was specified, which uses only the two teams involved and a home effect as relevant predictors. This model is the template upon which the model for yellow and red cards will be developed. For this application, the attacking and defensive capabilities can be substituted by teams' dirtiness and provocation levels to model booking rates. However, as discussed in Section 4.2, now there are other first-order effects that need to be included.

4.3.1 Basic Extensions

The referees can be treated in the same way that the teams' dirtiness and provocation factors are. So a harshness parameter is associated with each referee and the individual referee harshness coefficients are added to the parameter space. For this study referee data is only available since the start of the 1997/1998 season, hence likelihood maximisation, and prediction of scores, will only be performed over matches which have taken place since then. However, data from seasons 1994-1997 is included elsewhere in the modelling process.

Note that there are four, not two as in Equation 3.1.1, data points for each match k, those being home and away yellows (HY_k, AY_k) and home and away reds (HR_k, AR_k). Of ultimate interest in this study is the joint distribution of (HY_k, AY_k, HR_k, AR_k). However, it will be assumed that the home and away bookings rates are independent, so the task reduces to finding the joint distributions $(HY_k, HR_k) = (HR_k|HY_k)(HY_k)$ and $(AY_k, AR_k) = (AR_k|AY_k)(AY_k)$. This will be attempted in Section 4.3.7. For Sections 4.3.2 to 4.3.6 it is the expected number of *yellow* cards that is examined unless otherwise indicated. The validity of the assumption of independent home and

away booking rates will be discussed in Section 4.5.2.

4.3.2 Accounting for the result of the match

The obvious problem with trying to include the result of a match in the model is that the result of the match is not known at the time the prediction needs to be made. An attempt can be made however to predict which matches are more likely to result in a larger difference in score. This can be achieved quite easily by using the model specified by Equation 3.1.1 and implementing the MLE procedure described in Chapter 3 in order to obtain parameter estimates for teams' goal-scoring abilities. This is in fact the original application for which Dixon and Coles developed this procedure. Table 4.4 gives both the attacking and defensive parameters for all teams just after the matches played on 11/05/2002. Table 4.5 provides predictions for the matches which took place on 11/05/2002. No adjustments are made for home/away score dependence or match incentives, since only approximate estimates of abilities are required.

Table 4.4: Goal-scoring offensive ($\hat{\alpha}$) and defensive ($\hat{\beta}$) team ability estimates, May 2002

Team	$\hat{\alpha}$	rank	$\hat{\beta}$	rank
Arsenal	0.2875	2	-0.2853	1
Aston Villa	-0.0602	12	-0.1023	6
Barnsley	-0.1061	14	0.2382	26
Blackburn	-0.0058	7	0.0121	9
Bolton	-0.0804	13	0.1574	22
Bradford	-0.3022	29	0.2959	28
Charlton	-0.1424	18	0.1003	19
Chelsea	0.2058	4	-0.2225	3
Coventry	-0.1467	19	0.0742	16
Crystal Palace	-0.1081	15	0.1734	23
Derby	-0.1897	25	0.1029	20
Everton	-0.0587	11	0.0673	14
Fulham	-0.2336	26	-0.0465	7
Ipswich	-0.0483	10	0.0771	17
Leeds	0.1412	5	-0.1659	5
Leicester	-0.161	21	0.0653	13
Liverpool	0.2087	3	-0.2606	2
Man City	-0.1806	24	0.23	25
Man United	0.4852	1	-0.1888	4
Middlesbrough	-0.1629	22	-0.0019	8
Newcastle	0.1104	6	0.0202	10
Nottm Forest	-0.2342	27	0.2486	27
Sheffield Weds	-0.1108	16	0.1329	21
Southampton	-0.1177	17	0.0874	18
Sunderland	-0.167	23	0.022	11
Tottenham	-0.0066	8	0.04	12
Watford	-0.2482	28	0.3398	29
West Ham	-0.013	9	0.0733	15
Wimbledon	-0.1469	20	0.1832	24

Table 4.5: Score predictions for 11/5/2002

Home team	Away team	Home predicted goals	Away predicted goals
Arsenal	Everton	2.1375	0.7793
Blackburn	Fulham	1.4241	0.874
Chelsea	Aston Villa	1.6644	0.822
Leeds	Middlesbrough	1.7251	0.7851
Leicester	Tottenham	1.3297	1.1566
Liverpool	Ipswich	1.9971	0.801
Man United	Charlton	2.6952	0.7832
Southampton	Newcastle	1.3614	1.3293
Sunderland	Derby	1.4077	0.9223
West Ham	Bolton	1.7288	1.0973

The next step is to include the score predictions generated in this way in the predictions for yellow cards. Figure 4.4 plots a moving average of predicted score difference versus collected yellow cards, for both the home side and the away side. It appears that for the home side at least, if a side is expected to win, then their average number of yellow cards decreases. For the away side, the situation is less clear. The approach taken is to include separate home and away parameters to reflect the likelihood of a team experiencing a bad result in a match.

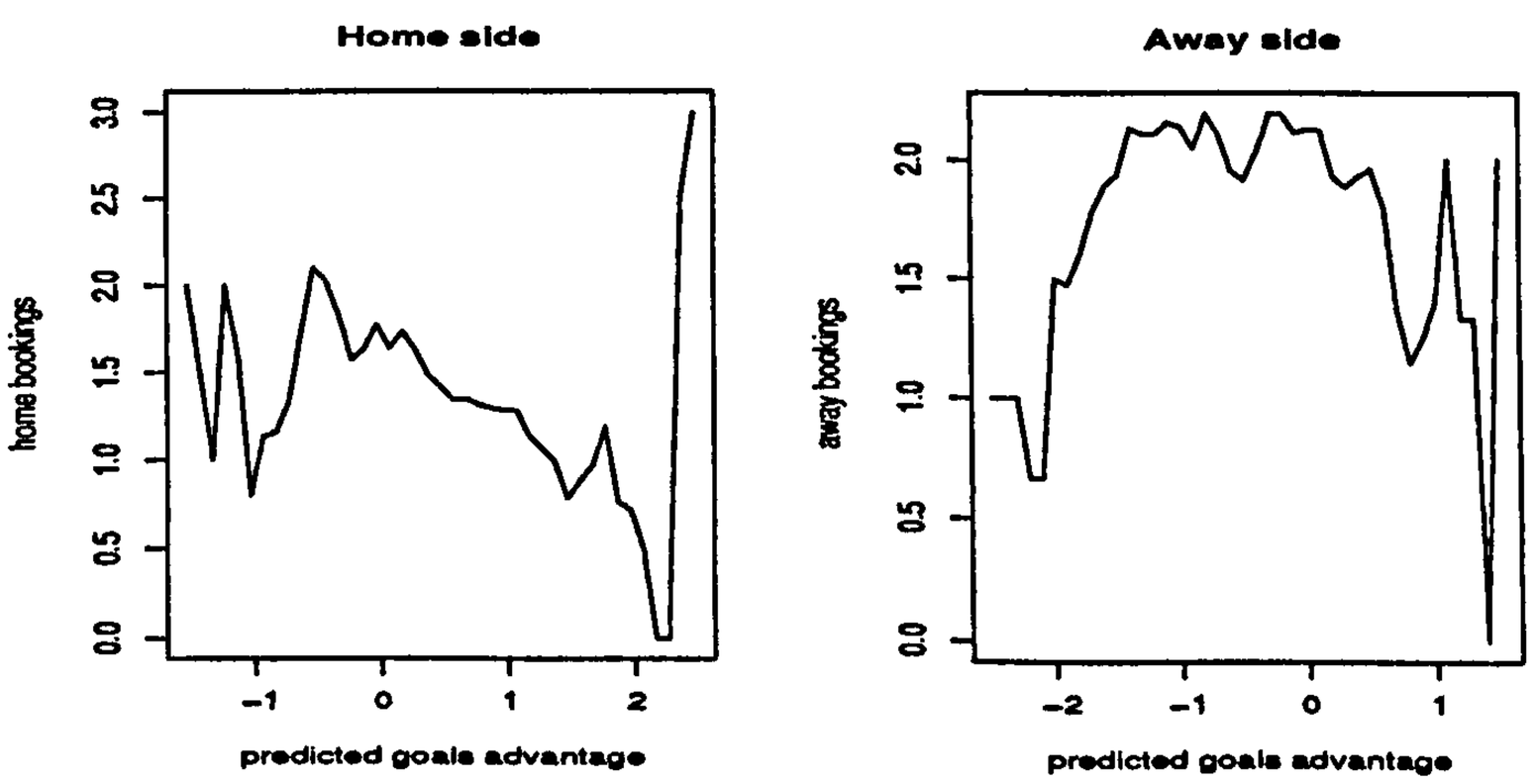


Figure 4.4: Predicted score advantage versus yellows collected

4.3.3 Modelling the climate

By examining Figure 4.3 it appears that generally yellow cards are awarded most frequently at the start of a season, then tail off gradually until the end of the season. It is important to acknowledge this non-stationarity. For example, if a side collects a large number of yellow cards in a match shortly after the start of the second season, the

model should acknowledge the generally high booking rate during that period when evaluating the parameters for that team in order to avoid unnecessary bias.

There are two issues which need to be resolved here:

- How is the climate estimated for matches which have taken place already, in order to minimise bias in the maximum likelihood estimation of the other model parameters?

- How is an estimate provided for the climate of a future match for which a prediction of the number of yellow cards is required?

To resolve the first issue a smooth curve is fitted which reflects the trends observed in Figure 4.3. To do this, Epanechnikov's kernel is used, with smoothing parameter set to 5 weeks (see Section 4.7.2 in the additional comments section for an explanation of this technique and the definition of Epanechnikov's kernel).

Figure 4.5 displays the curve obtained in this way at the final time-point in the data set, plotted over the moving average of observed yellow cards. Note that the procedure is modified directly after the trough in bookings rates around week 120, as explained in Section 4.2.3. In this case, kernel smoothing is applied only to data observed after the time-point when the trough took place.

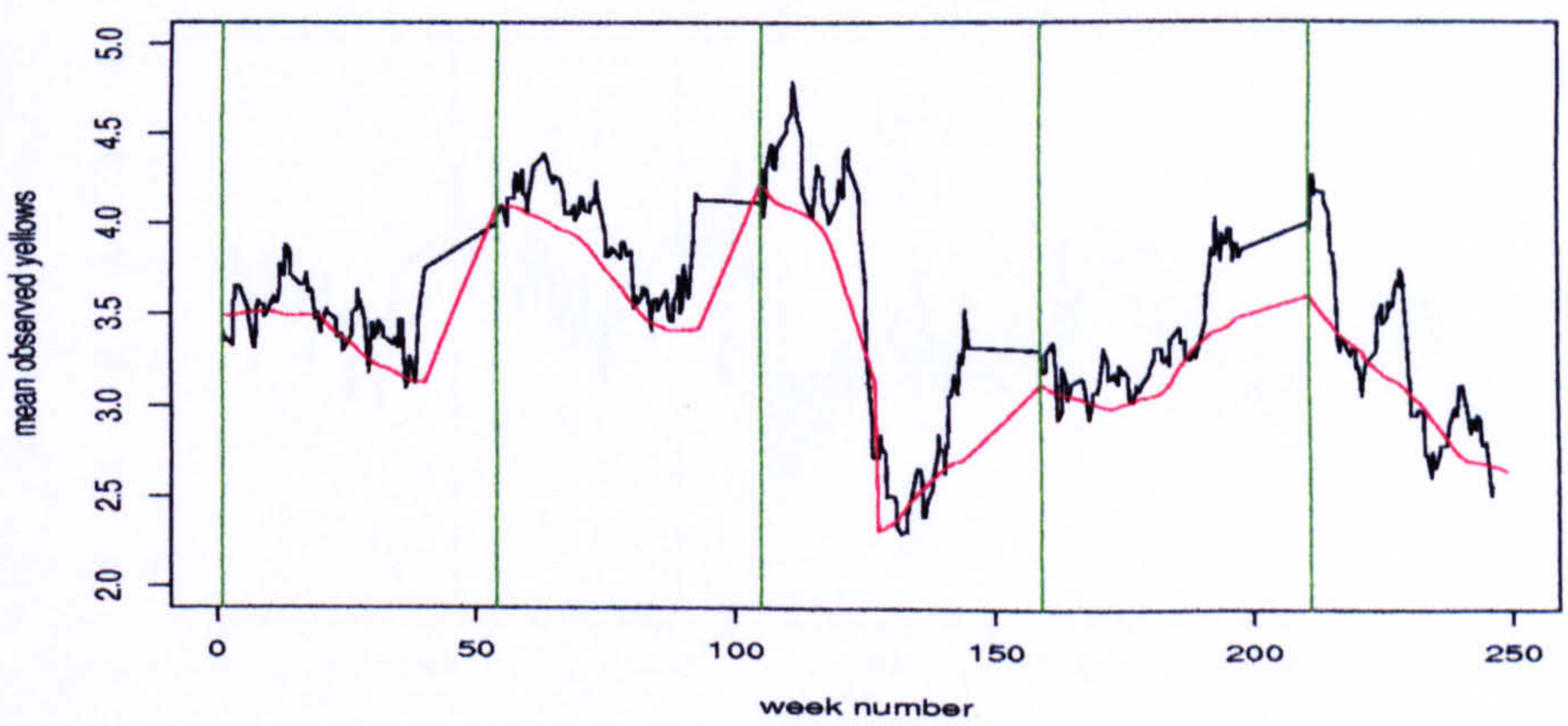


Figure 4.5: Moving average of observed yellows (___) and estimated climate (___). Vertical lines denote the start of a football season.

Issue 2 is resolved as follows: in normal circumstances the previous match's estimated climate appears to be a sensible prediction of the next fixture's climate. The exception to this is at the start of the season when, on inspection of Figure 4.5, a rise in the climate is likely to occur. The reasons for this are not entirely understood

to the writer (possibly new guidelines for certain offenses are issued at the start of most seasons). However, to accommodate this effect, the following simple procedure is employed:

Let S be the number of seasons in the data set. Let $IC_1, ..., IC_S$ be the initial climate of each season and $FC_1, ..., FC_S$ be the final climate of each season, as displayed in Figure 4.5. If a prediction $E[IC_i]$ of the climate at the start of season i is required

$$E[IC_i] = FC_{i-1} + \frac{\sum_{j=2}^{j=i-1}(IC_j - FC_{j-1})}{i-2}$$

So the expected climate at the start of a season is the climate at the end of the previous season, plus the mean change in climate from the end of one season to the start of the next, for all seasons observed until then. This value is carried through the first ten time-points in each season, to remove the instability that arises from having only a small number of matches over which to apply kernel smoothing.

Figure 4.6 plots the predicted climate, with the start-of-season adjustment described above. It incorrectly predicts a jump in the climate at the start of the fourth season, but generally seems to predict the climate adequately. Note that seasons 94-02 are employed to obtain the data for the season-jump, but only the climate for seasons 97-02 is plotted.

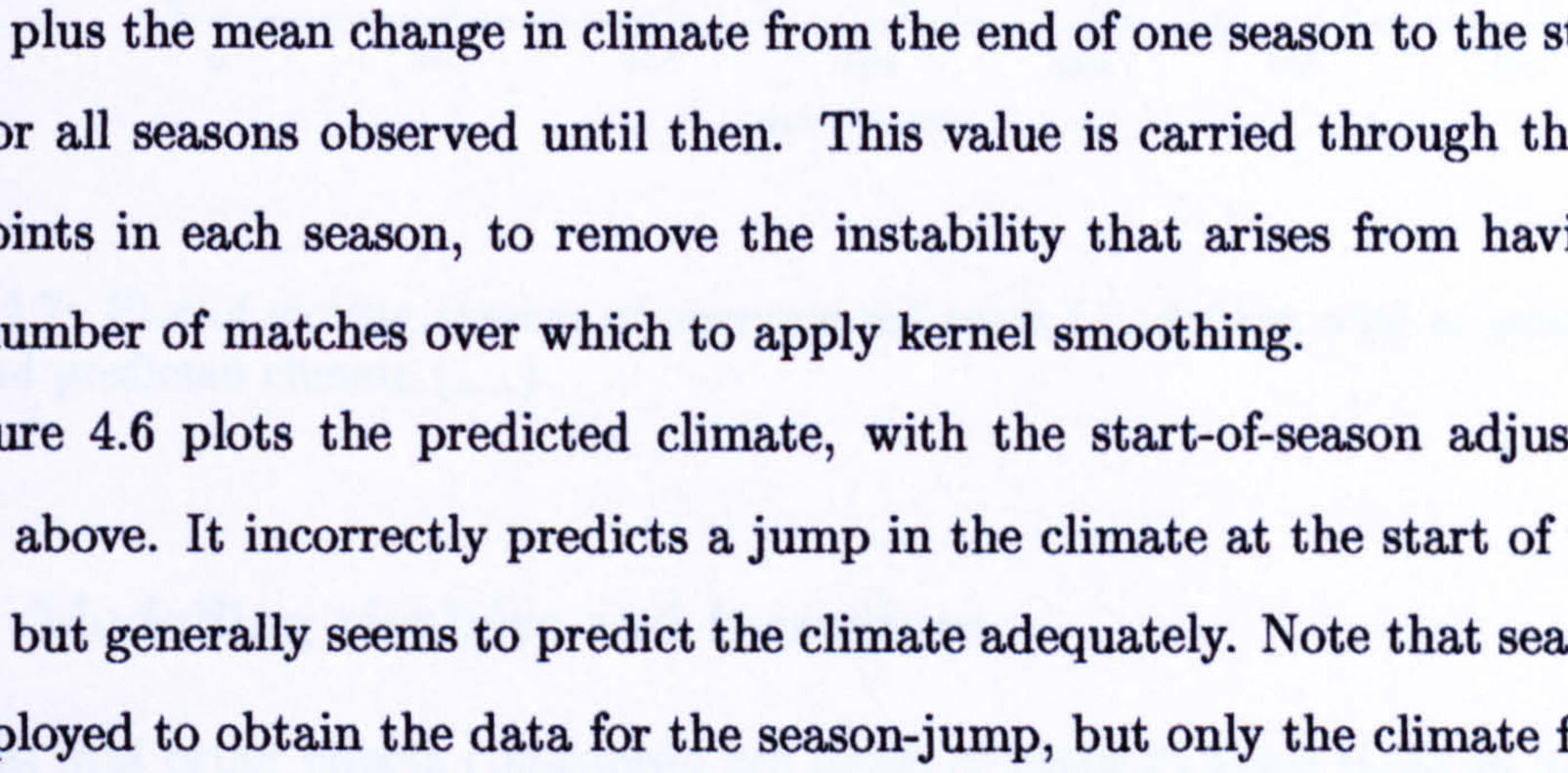


Figure 4.6: Plot of moving average of observed yellows (___) along with predicted climate (___)

Finally, the discussion above is concerned with the climate of yellow cards. It is also necessary to repeat the methodology in order to obtain an estimate for the climate of red cards, since the awarding of red cards is also subject to various external pressures.

Figure 4.7 displays the moving average, fitted climate and predicted climate for red cards. Again a bandwidth of 5 weeks appears to provide a satisfactory fit of the curve to the observed data.

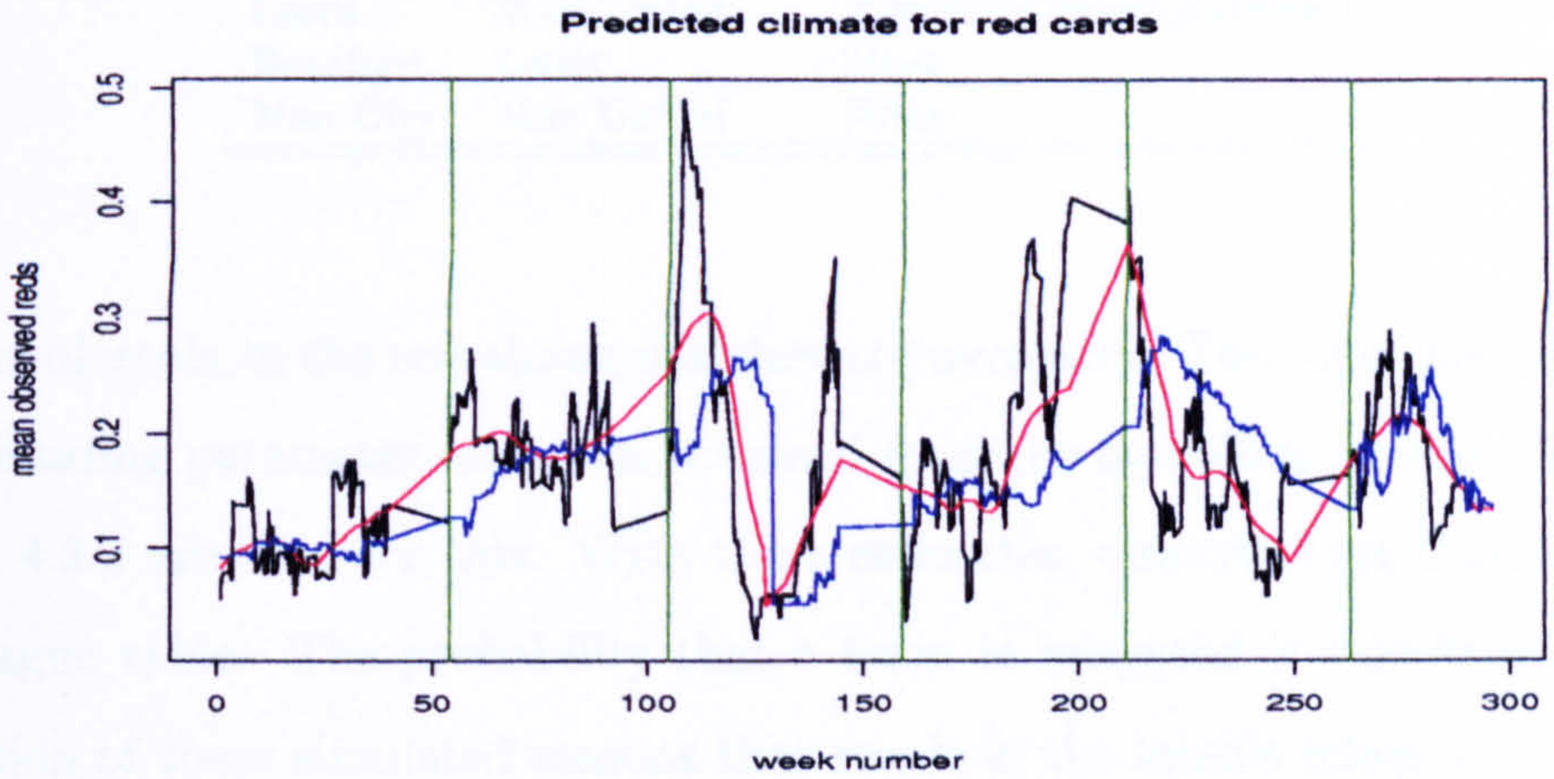


Figure 4.7: Plot of moving average of observed red cards (___) along with smoothed climate (___) and predicted climate (___)

4.3.4 Modelling rivalries and incentives

The final first order effects considered are those of team-to-team rivalries and specific match incentives. Once the levels of rivalry and incentives have been determined, their inclusion as factors in the model is straightforward, although it must be done at a later stage of the modelling process, for reasons which will be discussed shortly. In this case, the level of rivalries that exist between specific teams is determined empirically following consultation with Tony Bloom, who has researched thoroughly the soccer clubs' official magazines produced for supporters, collected prior to the first matches of the data set, in order to determine the traditional rivalries. Table 4.6 displays some of the levels of rivalry employed. Note that rivalries are not entirely symmetric: for example, Leeds are a strong rival of Bradford, but not vice versa. This reflects the fact that it is believed that a Leeds - Bradford game is of greater significance to Bradford than Leeds.

Meanwhile, a match is deemed to have a specific incentive attached to it if the result of the match may have an abnormally significant effect on the future of the club. Specifically, a team has an incentive if the match result may affect to a large extent the probability that the team wins the Premier League or is relegated from the Premier League. In order to calculate the probabilities of these two events, predictions for the

Table 4.6: Level of rivalry between teams

Team	Strong rivalries	Mild rivalries
Coventry	Aston Villa	Leicester,Derby
Everton	Liverpool	None
Leicester	None	Coventry,Aston Villa,Derby
Leeds	Man United	Barnsley,Chelsea,Bradford
Bradford	Leeds	None
Man City	Man United	None

numbers of goals in the remaining matches are necessary. The team goal-scoring and goal-conceding parameter estimates obtained from the modelling process described in Section 4.3.2 are used for this. With these estimates, outcomes are simulated of the final league table. The probability that a team is relegated is calculated to be the proportion of these simulated seasons that result in the team's relegation. The probability that a team wins the Premier League is similarly defined. Table 4.7 lists the final matches of the 2001/2002 season along with these probabilities before the matches take place. The probabilities of qualifying for two lucrative soccer tournaments, the UEFA Cup and the Champions League, are not considered due to the rather complicated rules which determine the chance of either event taking place, although this is a possible topic for further research.

Table 4.7: Title and relegation probabilities at end of 2001/2002 season. The probabilities apply before the listed match takes place.

date	home	P(win)	P(releg.)	away	P(win)	P(releg.)	score
20020427	Aston Villa	0	0	Southampton	0	0.002	2-1
20020427	Charlton	0	0.002	Sunderland	0	0.456	2-2
20020427	Derby	0	1	Leeds	0	0	0-1
20020427	Fulham	0	0	Leicester	0	1	0-0
20020427	Ipswich	0	0.395	Man United	0.124	0	0-1
20020427	Middlesbrough	0	0	Chelsea	0	0	0-2
20020427	Newcastle	0	0	West Ham	0	0	3-1
20020427	Tottenham	0	0	Liverpool	0.071	0	1-0
20020428	Everton	0	0	Blackburn	0	0	1-2
20020429	Bolton	0	0	Arsenal	0.776	0	0-2
20020508	Liverpool	0	0	Blackburn	0	0	4-3
20020508	Man United	0.122	0	Arsenal	0.878	0	0-1
20020511	Arsenal	1	0	Everton	0	0	4-3
20020511	Blackburn	0	0	Fulham	0	0	3-0
20020511	Chelsea	0	0	Aston Villa	0	0	1-3
20020511	Leeds	0	0	Middlesbrough	0	0	1-0
20020511	Leicester	0	1	Tottenham	0	0	2-1
20020511	Liverpool	0	0	Ipswich	0	0.937	5-0
20020511	Man United	0	0	Charlton	0	0	0-0
20020511	Southampton	0	0	Newcastle	0	0	3-1
20020511	Sunderland	0	0.063	Derby	0	1	1-1
20020511	West Ham	0	0	Bolton	0	0	2-1

In order to assess the effects of incentives and rivalries as accurately as possible, some realistic match home and away yellow card predictions are needed. This is because in order to detect if these factors affect booking rates, it is necessary to compare a set of predictions for yellow cards which take account of these factors with a set of reasonably reliable predictions that do not. To obtain a set of predictions of the second type, a model incorporating factors F1-F5 as outlined in Section 4.2 is fitted.

4.3.5 Model construction from factors F1-F5

By adapting the model specified in Section 3.1, at time t, the expected number of home and away yellow cards (HY_k and AY_k) for match k between teams $i(k)$ and $j(k)$, refereed by official $r(k)$ are:

$$\begin{aligned} E[HY_k] &= CY_k * \delta * exp(\mu_h + \alpha_{i(k)} + \beta_{j(k)} + \gamma_{r(k)} + s_h * \Delta_k) \\ E[AY_k] &= CY_k * (1-\delta) * exp(\mu_a + \alpha_{j(k)} + \beta_{i(k)} + \gamma_{r(k)} + s_a * (-\Delta_k)) \end{aligned} \quad (4.3.1)$$

where

- CY_k represents the estimated yellow cards climate at the time match k takes place, as displayed in Figure 4.5.
- $0 < \delta < 1$ represents the proportion of total yellow cards that are collected by home sides
- $\alpha_{i(k)}$, $\alpha_{j(k)}$ are team $i(k)$ and $j(k)$'s dirtiness parameters
- $\beta_{i(k)}$, $\beta_{j(k)}$ are team $i(k)$ and $j(k)$'s provocation parameters
- $\gamma_{r(k)}$ is referee $r(k)$'s harshness parameter
- $\Delta_k = E[HSC_k] - E[ASC_k]$ where $E[HSC_k], E[ASC_k]$ are the home and away predicted scores (estimated using the estimates from Section 4.3.2)
- s_h and s_a are the home and away coefficients for the effect of home and away predicted superiority.
- μ_h and μ_a are intercepts.

Note that CY_k is not a parameter to be estimated in the likelihood maximisation, since it has been separately determined in Section 4.3.3. It should also be noted

that the time down-weighting, prior tightness and seasonal truncation parameters, as defined in Sections 3.1, 3.2.1 and 3.2.4 and referred to as *external* parameters, are as yet undetermined. To determine these, several sets of their values are fixed and for each set, the entire set of *internal* parameters (the parameters included in Equation 4.3.1) are estimated at each time-point. They are then used to find predictions for the numbers of yellow cards given and the resulting predictive likelihood statistic is monitored. Table 4.8 displays the predictive likelihoods obtained in this way. The optimal value is highlighted in red and it appears that (0.02, 0.2, 20) is close to the optimal values for the time down-weighting, prior tightness and seasonal truncation parameters respectively.

Table 4.8: Predictive likelihood of yellow cards model obtained for different choices of external parameters

Truncation w= 5 weeks:

		Prior variance $\tau_{\alpha\beta}$ of offensive and defensive estimates			
		0.05	0.1	0.2	0.5
	0.001	-5908.616	-5866.486	-5858.54	-5894.354
	0.005	-5911.713	-5864.679	-5852.256	-5887.95
Weight ς	0.01	-5916.128	-5864.963	-5846.599	-5882.129
	0.02	-5924.553	-5871.035	-5841.957	-5877.462
	0.05	-5941.492	-5896.95	-5852.989	-5891.999

Truncation w= 10 weeks:

		Prior variance $\tau_{\alpha\beta}$ of offensive and defensive estimates			
		0.05	0.1	0.2	0.5
	0.001	-5908.656	-5866.419	-5858.394	-5894.212
	0.005	-5911.986	-5864.538	-5851.674	-5887.377
Weight ς	0.01	-5916.717	-5865.076	-5845.814	-5881.361
	0.02	-5925.566	-5872.059	-5841.579	-5877.161
	0.05	-5942.824	-5899.437	-5854.699	-5894.17

Truncation w= 20 weeks:

		Prior variance $\tau_{\alpha\beta}$ of offensive and defensive estimates			
		0.05	0.1	0.2	0.5
	0.001	-5908.761	-5866.232	-5857.992	-5893.817
	0.005	-5912.715	-5864.196	-5850.117	-5885.842
Weight ς	0.01	-5918.25	-5865.535	-5843.871	-5879.461
	0.02	-5928.067	-5874.945	-5841.084	-5876.981
	0.05	-5946.164	-5905.606	-5860.328	-5902.203

Truncation w= 30 weeks:

		Prior variance $\tau_{\alpha\beta}$ of offensive and defensive estimates			
		0.05	0.1	0.2	0.5
	0.001	-5908.725	-5866.309	-5858.153	-5893.975
	0.005	-5912.449	-5864.339	-5850.736	-5886.456
Weight ς	0.01	-5917.691	-5865.367	-5844.635	-5880.213
	0.02	-5927.174	-5873.874	-5841.271	-5877.038
	0.05	-5944.897	-5903.322	-5858.008	-5898.61

4.3.6 Modelling Incentives

By obtaining MLEs for the parameters in the model specified by Equation 4.3.1 (subject to near-optimal values for the time down-weighting, prior tightness and seasonal truncation parameters), predictions can be generated that are necessary for the final stage of the modelling process. This is to test the effect of specific match incentives and rivalries on the bookings rate. The effect of the incentives and rivalries is examined by fitting several generalised linear models.

Before constructing these models, the following variables are defined:

$$r_k^s = \begin{cases} 1 & \text{if match } k \text{ is between two teams who are strong rivals} \\ 0 & \text{otherwise} \end{cases}$$

$$r_k^m = \begin{cases} 1 & \text{if match } k \text{ is between two teams who are mild rivals} \\ 0 & \text{otherwise} \end{cases}$$

$$i_k^{rh} = \begin{cases} 1 & \text{if home side of match } k \text{ is in danger of relegation} \\ 0 & \text{otherwise} \end{cases}$$

$$i_k^{wh} = \begin{cases} 1 & \text{if the home side of match } k \text{ can win Premier League} \\ 0 & \text{otherwise} \end{cases}$$

(i_k^{ra}, i_k^{wa}) are defined similarly for the away side, while

$$i_k^r = \begin{cases} 1 & \text{if both sides of match k are in danger of relegation} \\ 0 & \text{otherwise} \end{cases}$$

$$i_k^w = \begin{cases} 1 & \text{if in match k both sides can win Premier League} \\ 0 & \text{otherwise} \end{cases}$$

Incidentally, a team is deemed to be facing relegation if their probability of being relegated lies between 0.05 and 0.95, to ensure that teams whose predicament is effectively sealed are not classified as having an incentive. The same principal is applied to the teams who can win the Premier League. Also, in situations when rivals are also, for example, both fighting against relegation, then the rivalry indicator is set to zero, since it is assumed that the threat of relegation is the more dominant effect in the match, and that the effects of these two factors are not additive (data are too sparse to test this belief). Table 4.9 displays the relevant results.

Table 4.9: Investigating effect of derbies and incentives. $\hat{h_k}$ and $\hat{a_k}$ represent the predictions for home and away yellow cards from the model constructed from factors 1-4

Model	Coefficients and p-values
$H_k - \hat{h_k} \sim r_k^s + r_k^m$	r_k^s: (0.2,0.02) r_k^m: (0.23,0.01)
$H_k - \hat{h_k} \sim r_k^s + r_k^m + i_k^{rh}$	r_k^s: (0.2,0.02) r_k^m: (0.23,0.01) i_k^{rh}: (0,0.97)
$H_k - \hat{h_k} \sim r_k^s + r_k^m + i_k^{wh}$	r_k^s: (0.2,0.02) r_k^m: (0.23,0.01) i_k^{wh}: (-0.01,0.88)
$H_k - \hat{h_k} \sim r_k^s + r_k^m + i_k^r$	r_k^s: (0.2,0.02) r_k^m: (0.23,0.01) i_k^r: (-0.01,0.87)
$H_k - \hat{h_k} \sim r_k^s + r_k^m + i_k^w$	r_k^s: (0.21,0.01) r_k^m: (0.24,0.01) i_k^w: (0.2,0.03)
$A_k - \hat{a_k} \sim r_k^s + r_k^m$	r_k^s: (0.25,0) r_k^m: (0.05,0.52)
$A_k - \hat{a_k} \sim r_k^s + i_k^{ra}$	r_k^s: (0.25,0) i_k^{ra}: (0.01,0.84)
$A_k - \hat{a_k} \sim r_k^s + i_k^{wa}$	r_k^s: (0.25,0) i_k^{wa}: (0.02,0.67)
$A_k - \hat{a_k} \sim r_k^s + i_k^r$	r_k^s: (0.25,0) i_k^r: (-0.01,0.82)
$A_k - \hat{a_k} \sim r_k^s + i_k^w$	r_k^s: (0.26,0) i_k^w: (0.16,0.04)

Some of the results in Table 4.9 are a little surprising. For example, it appears that the threat of relegation has no effect on booking rates, even if both teams in the match are relegation rivals. Similarly the booking rate for a match involving a side in contention for winning the Premier League only rises if both sides participating are title contenders. The effect of inter-team rivalries is generally as expected, although it is interesting that the mild rivalries affect home, but not away, booking rates. It is for this reason that the term for the mild relegation indicator is not included in the final four models tested in table 4.9. Thus the only alterations needed in the model are the additions of parameters that allow the expected number of yellow cards to increase in matches between sides who are both in contention to win the Premier League and matches where the two sides are traditional rivals. If this rivalry is mild, only the home side's expected number of yellow cards is adjusted.

Having displayed the necessary extensions of the Dixon-Coles model in Sections 4.3.1 to 4.3.4 it is now possible to state the specification of the final model for the mean yellow card rates. For match k between home team $i(k)$, away team $j(k)$ and refereed by official $r(k)$ the expected number of home and away yellow cards are:

$$\begin{aligned} E[HY_k] &= CY_k * \delta * exp(\mu_h + \alpha_{i(k)} + \beta_{j(k)} + \gamma_{r(k)} \\ &\quad + s_h * \Delta_k + r_k^s * \lambda_s + r_k^m * \lambda_m + i_k^w * \nu) \\ E[AY_k] &= CY_k * (1-\delta) * exp(\mu_a + \alpha_{j(k)} + \beta_{i(k)} + \gamma_{r(k)} \\ &\quad + s_a * (-\Delta_k) + r_k^s * \lambda_s + i_k^w * \nu) \end{aligned} \tag{4.3.2}$$

where, in addition to the parameters described in Section 4.3.5

- λ_s is the parameter for the effect of playing against a strong rival

- λ_m is the parameter for the effect of playing against a mild rival

- ν is the parameter for the effect of both teams being rivals for overall victory in the Premier League

4.3.7 Model for red cards

Finally, a model for red cards conditional on the number of yellow cards is required. Denoting the number of home and away red cards by HR_k and AR_k and the fitted climate for red cards displayed in Figure 4.7 by CR_k, a straightforward model, assuming a Poisson distribution in the likelihood, is:

$$\begin{aligned} E[HR_k|HY_k = hy_k] &= CR_k * \delta_r * exp(\mu_h^r + \varrho_h * hy_k) \\ E[AR_k|AY_k = ay_k] &= CR_k * (1 - \delta_r) * exp(\mu_a^r + \varrho_a * ay_k) \end{aligned} \tag{4.3.3}$$

where home effect, intercept and slope parameters $(\delta_r, \mu_h^r, \mu_a^r, \varrho_h, \varrho_a)$ are to be estimated.

Note that the parameters included in Equation 4.3.2 are not all estimated within a single likelihood maximisation. This is due to a feature of the parameter estimation process that is described in detail in Section 3.2.3. Essentially, it is desirable that λ_s, λ_m and ν are treated as parameters that are constant throughout time. However, the parameter estimation procedure for allowing team parameters to be based on more recent results also bases its estimates of the λ_s, λ_m and ν parameters on more recent matches unless modifications to the parameter estimation process are made. So in practice parameter estimates are obtained using a procedure similar to that outlined in Section 3.2.3. Applying it to this example the procedure is as follows:

1. Find maximum likelihood estimates of the parameters contained in Equation 4.3.1.

2. Fit the generalised linear model described by Equation 4.3.2 via Poisson regression, with the α, β, γ parameters treated as constants.

3. Again perform maximum likelihood estimation of the model described in Equation 4.3.2 but where the $\mu_h, \mu_a, \delta, s_h, s_a, \lambda_s, \lambda_m$ and ν parameters are treated as constants, and the α, β, γ parameters are re-evaluated.

4. Perform maximum likelihood estimation of the red cards model described in Equation 4.3.3.

By repeating this procedure at each time-point, estimates for each parameter are obtained.

4.4 Results from the models

4.4.1 Parameter estimates

Tables 4.10 and 4.11 display the estimates for team and referee parameters, obtained at time-point 256 (by which time 124 weeks have elapsed in the data set) and at time-point 512 (when 249 weeks have elapsed), the final time-point in the data set at the time of writing. Note that Ipswich, Manchester City and Fulham had not played in the Premier League by time-point 256, hence do not have any estimates here. Figure 4.8 plots the team and referee estimates for selected teams and referees over time. The period where Blackburn's estimate is almost flat corresponds to the two year period when Blackburn were not playing in the Premier League due to being relegated at the end of the 1998/99 soccer season. The curve is not *totally* flat though, because although Blackburn do not participate in any matches during this period, their opponents and referees do. As a result, the parameter estimates for Blackburn are slightly re-evaluated based on data about opponents and referees that the parameter estimation procedure subsequently incorporates.

4.4.2 Model evaluation

The predictive ability of the model can be assessed via its *predictive likelihood* statistic as defined in Section 3.1. Table 4.12 displays this statistic, plus predictive likelihood statistics for some simpler models, in order to gain a clearer picture of the model's accuracy. Note that the joint likelihood of the number of (home yellow, away yellow, home red, away red) cards is calculated rather than the points make-up, which has a rather less tractable distribution. Model 1 predicts that total bookings in any match will be the mean total bookings observed in all matches prior to the game, in

Table 4.10: Team dirtiness ($\hat{\alpha}$) and provocation ($\hat{\beta}$) parameter estimates, with ranking displayed in brackets

Team	$\hat{\alpha}$, t=256	$\hat{\alpha}$, t=512	$\hat{\beta}$, t=256	$\hat{\beta}$, t=512
Arsenal	0.056 (11)	0.124 (7)	0.175 (3)	0.145 (4)
Aston Villa	-0.008 (16)	-0.078 (22)	-0.025 (16)	0.021 (13)
Barnsley	0.039 (12)	0.016 (16)	0.005 (15)	0.001 (17)
Blackburn	0.099 (6)	0.029 (14)	0.065 (10)	0.086 (8)
Bolton	0.031 (13)	-0.012 (18)	0.048 (13)	-0.077 (23)
Bradford	-0.15 (23)	-0.084 (23)	-0.071 (22)	-0.119 (26)
Charlton	-0.093 (20)	-0.046 (20)	0.087 (8)	0.063 (11)
Chelsea	0.192 (2)	0.162 (4)	0.073 (9)	-0.056 (22)
Coventry	-0.001 (15)	0.075 (8)	-0.026 (17)	0.007 (15)
Crystal Palace	-0.034 (19)	-0.008 (17)	0.052 (11)	0.02 (14)
Derby	0.197 (1)	0.229 (2)	-0.049 (18)	-0.043 (20)
Everton	0.12 (5)	0.134 (6)	0.175 (4)	0.09 (6)
Fulham	-	-0.028 (19)	-	0.233 (2)
Ipswich	-	-0.286 (29)	-	-0.091 (25)
Leeds	0.121 (4)	0.236 (1)	0.22 (1)	0.298 (1)
Leicester	-0.254 (25)	0.02 (15)	0.05 (12)	0.092 (5)
Liverpool	-0.013 (18)	-0.093 (24)	-0.062 (20)	-0.027 (19)
Man City	-	0.04 (12)	-	0.06 (12)
Man United	-0.144 (22)	-0.072 (21)	-0.155 (24)	-0.164 (27)
Middlesbrough	0.08 (9)	0.055 (10)	0.098 (7)	0.082 (9)
Newcastle	-0.131 (21)	-0.178 (28)	0.119 (6)	0.068 (10)
Nottm Forest	0.093 (7)	0.071 (9)	-0.147 (23)	-0.078 (24)
Sheffield Weds	-0.205 (24)	-0.17 (27)	-0.26 (25)	-0.205 (28)
Southampton	0.01 (14)	-0.157 (26)	0.008 (14)	-0.045 (21)
Sunderland	0.185 (3)	0.163 (3)	0.123 (5)	0.162 (3)
Tottenham	0.084 (8)	0.034 (13)	0.207 (2)	0.006 (16)
Watford	-0.01 (17)	0.046 (11)	-0.052 (19)	-0.026 (18)
West Ham	0.078 (10)	0.142 (5)	-0.063 (21)	0.088 (7)
Wimbledon	-0.261 (26)	-0.134 (25)	-0.515 (26)	-0.36 (29)

other words that the booking rate in a match is not dependent on the referee, the teams playing or the climate and can best be predicted by the overall mean number of bookings for all matches. Model 2 is more sophisticated, where for each match the home prediction is a combination of the mean number of yellows the home team has collected, the mean number of yellows the away team has provoked and the mean number of cards the referee has awarded in previous matches (all weighted according to how recently these matches occurred). The away prediction is calculated similarly. Also, the prevailing climate for bookings is accommodated. The exact method used is outlined in Section 4.7.1 of the additional comments. This model has been devised since it does not employ any advanced statistical methods, and might well be an approach a non-statistician, with access to the relevant data, would use. Model 3 is the model incorporating factors **F1** to **F5** described in Section 4.3, hence does not consider rivalries or incentives. Model 4 is similar to Model 3 but with rivalries and incentives included, hence is the most advanced model constructed in this chapter and

Table 4.11: Referee parameter estimates at timepoints 256 and 512. The number in brackets is their ranking out of all the referees who had officiated at that time-point

Referee	$\hat{\gamma}$, t=256	$\hat{\gamma}$, t=512
P.Alcock	-0.063 (20)	-0.037 (25)
G.Ashby	0.007 (13)	0.006 (19)
G.Barber	0.148 (2)	0.089 (9)
N.Barry	0.037 (9)	0.003 (20)
S.Bennett	0.092 (6)	0.057 (12)
M.Bodenham	0.004 (14)	0.001 (21)
K.Burge	-0.217 (25)	-0.095 (29)
M.Dean	-	0.096 (8)
P.Dowd	-	0.106 (4)
S.Dunn	-0.022 (17)	-0.05 (26)
P.Durkin	-0.197 (24)	-0.254 (34)
A.Durso	0.031 (10)	0.049 (13)
D.Elleray	-0.122 (22)	-0.148 (32)
C.Foy	-	0.128 (3)
D.Gallagher	-0.176 (23)	-0.008 (23)
M.Halsey	-0.024 (18)	-0.174 (33)
R.Harris	0.119 (4)	0.078 (10)
P.Jones	-0.054 (19)	0.001 (22)
B.Knight	0.116 (5)	0.105 (5)
S.Lodge	0.031 (11)	0.01 (18)
M.Messias	-	0.015 (17)
G.Poll	0.086 (7)	-0.069 (27)
D.Pugh	-	0.022 (16)
M.Reed	0.151 (1)	0.101 (7)
U.Rennie	0.048 (8)	-0.126 (31)
M.Riley	0.004 (15)	0.166 (1)
R.Styles	-	0.139 (2)
P.Taylor	-	-0.015 (24)
A.Wiley	-0.021 (16)	-0.075 (28)
C.Wilkes	-	0.103 (6)
A.Wilkie	0.024 (12)	0.027 (15)
G.Willard	0.143 (3)	0.065 (11)
J.Winter	-0.064 (21)	-0.124 (30)
E.Wolstenholme	-	0.048 (14)

Table 4.12: Predictive likelihood for different models

Model	Likelihood statistic
1	-5888.927
2	-5720.314
3	-5680.083
4	-5670.205

is described by Equation 4.3.2.

Since a higher predictive likelihood statistic is desirable, it is reassuring to note that the most advanced model is the one with overall the most accurate predictions. It is unfortunately not possible to produce equivalent figures for the bookmaker's predictions, since they provide only a prediction for the total number of points accumulated in the match, where 10 points are awarded for each yellow card and 25 points are awarded

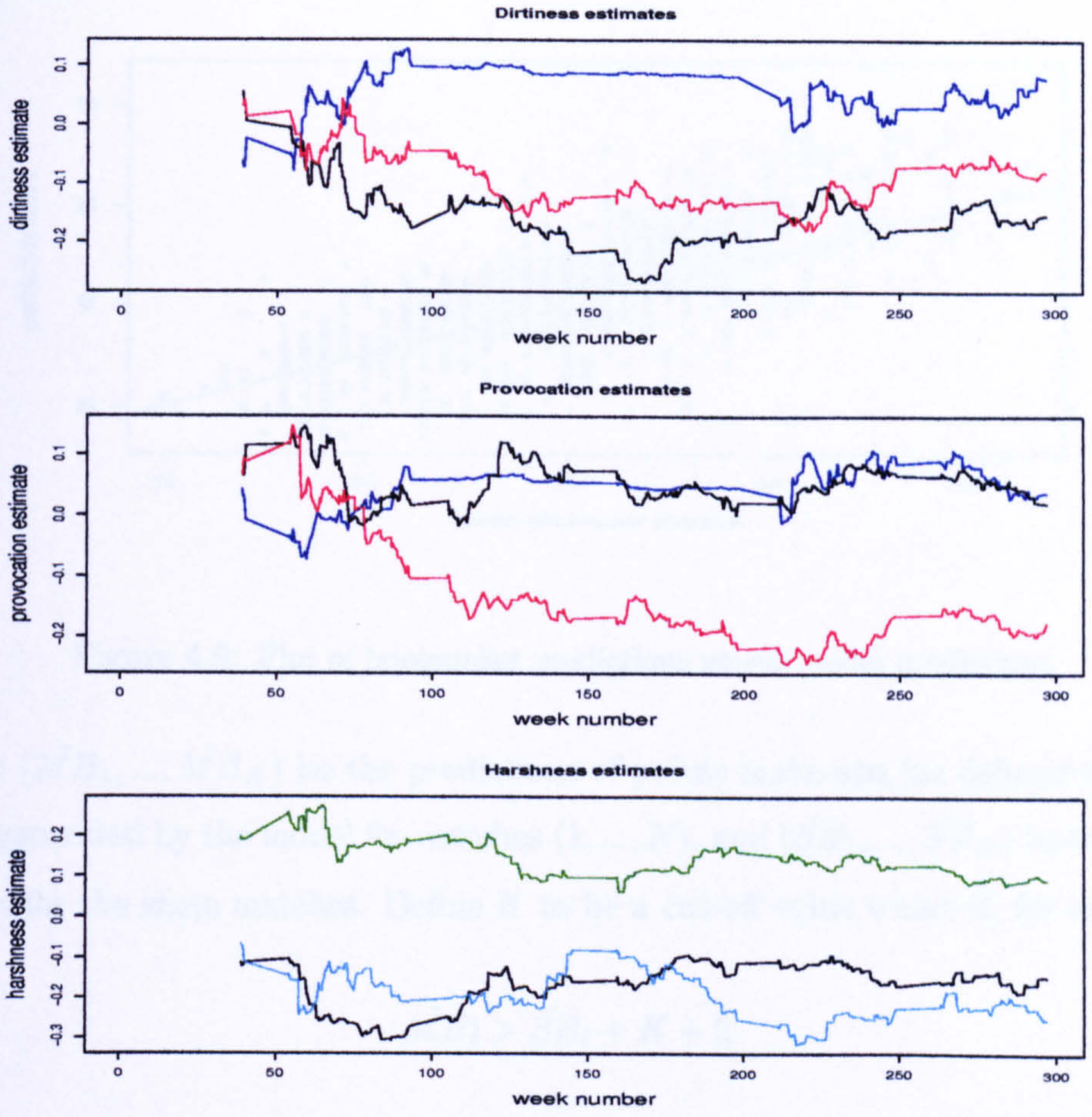


Figure 4.8: Plots of team dirtiness and provocation estimates over time, for Blackburn (___), Newcastle (___) and Man United (___). Also plots of referee harshness estimate for D. Elleray (___), P. Durkin (___) and G. Barber (___)

for each red card (as outlined in Section 1.3.2). These values cannot be converted into Poisson-distributed predictions for home and away yellow and red cards.

4.4.3 Betting strategy and success

Using the predictions produced from the most advanced model it is of interest to formulate a betting strategy and observe the returns it would generate. Bets should be placed when a discrepancy arises between the model predictions and the spread provided by a bookmaker. The model predictions for individual yellow and red cards can easily be converted into predictions for points make-ups by summing the probabilities of all the permutations of cards which result in each possible make-up. Figure 4.9 plots the quoted spread prices against the model predictions of points make-ups. While there is broad agreement, it is the points away from the diagonal which represent matches on which we should be most inclined to bet.

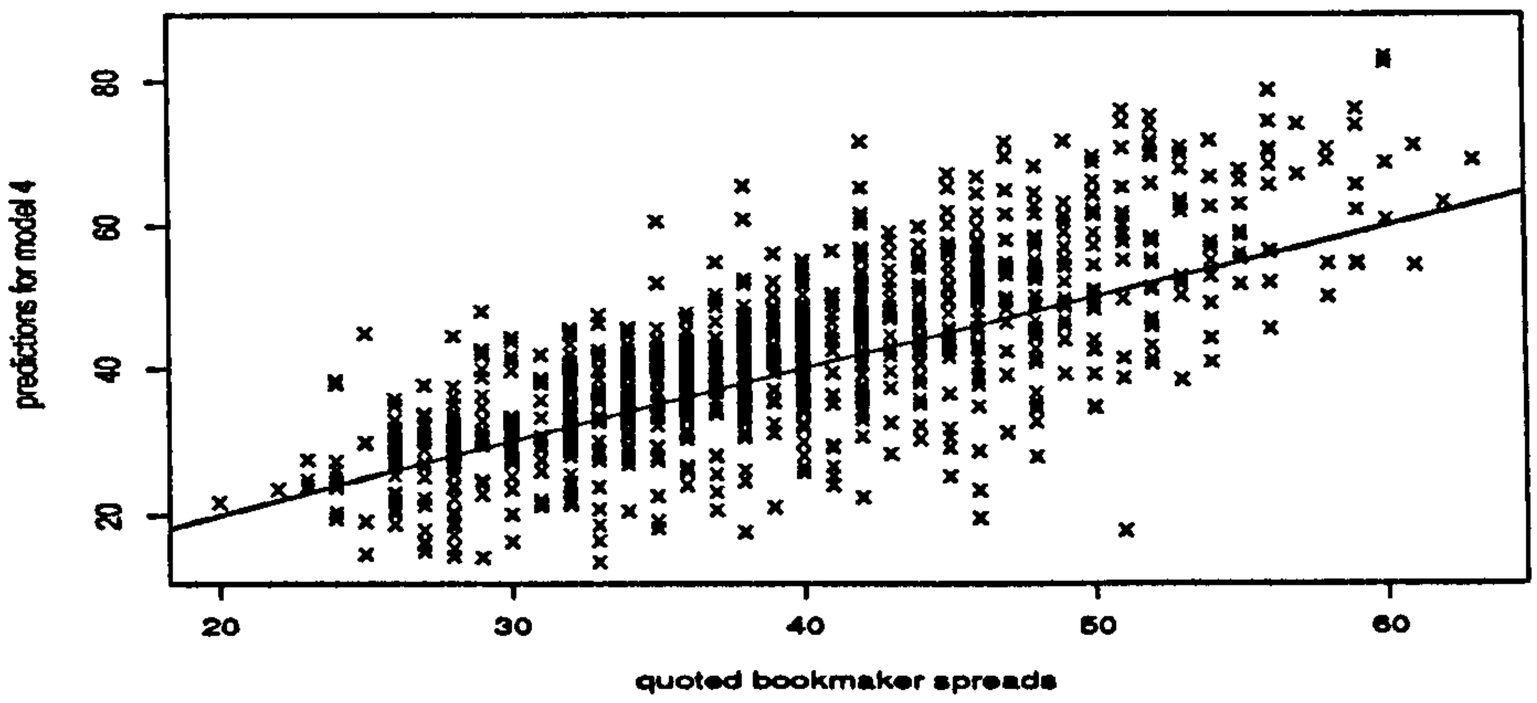


Figure 4.9: Plot of bookmaker predictions versus model predictions

Let $(\hat{MB}_1, ..., \hat{MB}_N)$ be the predictions of points make-ups (as defined in Section 1.3.2) generated by the model for matches $(1, ..., N)$, and $(\hat{SB}_1, ..., \hat{SB}_N)$ be the quoted spreads for the same matches. Define K to be a cut-off value where if, for match i,

$$\hat{MB}_i > \hat{SB}_i + K + 2$$

then a bet is placed on high bookings and if

$$\hat{MB}_i < \hat{SB}_i - K - 2$$

then a bet is placed on low bookings. The 2 point addition or subtraction appears because the bookmaker offers 4-point spread intervals, rather than a single number, in order to make its profit. The profit or loss made by following this betting strategy, for different values of K, is considered. Figure 4.10 plots annual returns, in points, for various values of K, for the 99/00, 00/01 and 01/02 seasons.

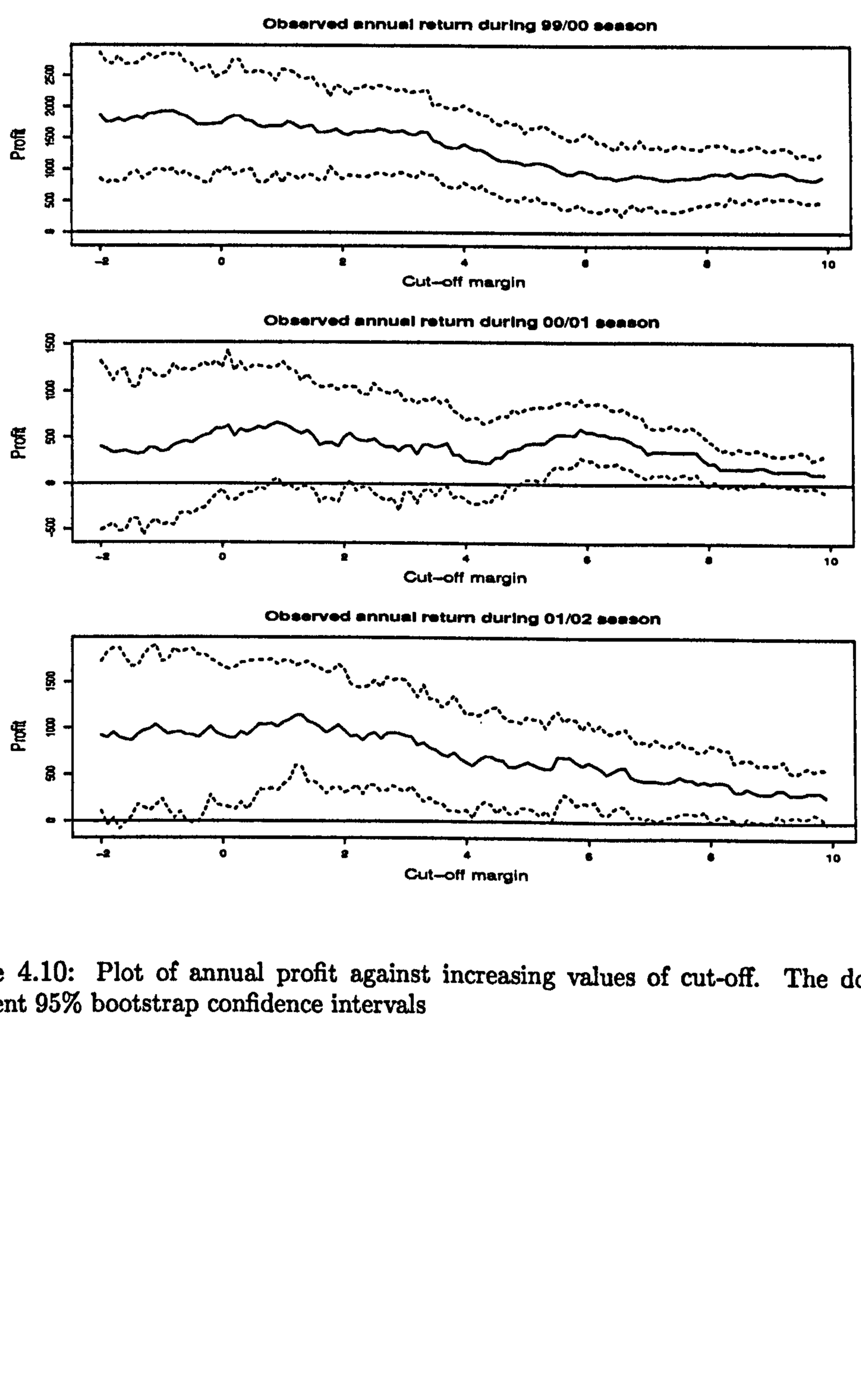


Figure 4.10: Plot of annual profit against increasing values of cut-off. The dotted lines represent 95% bootstrap confidence intervals

The observed return curves are somewhat bizarre, since it seems that even by betting on matches with a negative cut-off, hence negative expected return, one still makes a profit in all seasons. The 99/00 return curve must be regarded with some skepticism since it was during the middle of this season that the dramatic drop in booking rates highlighted in Section 4.2.3 occurred. In reality there was considerable uncertainty concerning the behaviour of referees for many weeks subsequent to this drop. Therefore many of the winning bets theoretically placed during the 99/00 season and included in the 99/00 return curve could not have been placed with any confidence.

Note that the returns generated when the expected return is negative do not correspond to "random betting", since this strategy still excludes what the model considers to be especially unattractive bets even if the cut-off value $K < 0$. A random betting strategy does not do this. Interestingly, the sum of the spread sell points for the 00/01 and 01/02 seasons was 33844, while the total points make-ups for the same matches was 33320, meaning one would have achieved a profit of 524 points by selling every match.

The strategy employed in Figure 4.10 is rather naive since bets with equal expected return but different variances are treated equally whereas the bets with lower variance are more attractive to many gamblers. For example, consider the two matches detailed in Table 4.13. According to the model predictions, the bookings total should be sold in both matches and both matches have similar expected return. The difference in the variance of the return according to the model is displayed in Figure 4.11.

Table 4.13: Data for two matches in data set with equal mean returns

Date	*Home team*	*Away team*	*Spread*	*Model prediction*	*Expected return*	*Variance of return*
20000514	Sheffield Weds	Leicester	22-26	17.88	4.12	236.54
20000826	Everton	Derby	52-56	47.92	4.08	639.66

- For the first match, the maximum possible win is 22 points and there is a 19% chance of this occurring. The probability of losing 50 points or more is 0.5%.

- For the second match, there is a 32% chance of winning 22 points or more and a 1% chance of winning a maximum of 52 points. The probability of losing 50 points or more is 3%.

Many gamblers would consider minimising the probability of financial ruin to be a key criteria when selecting a staking plan so would place more of their assets on the

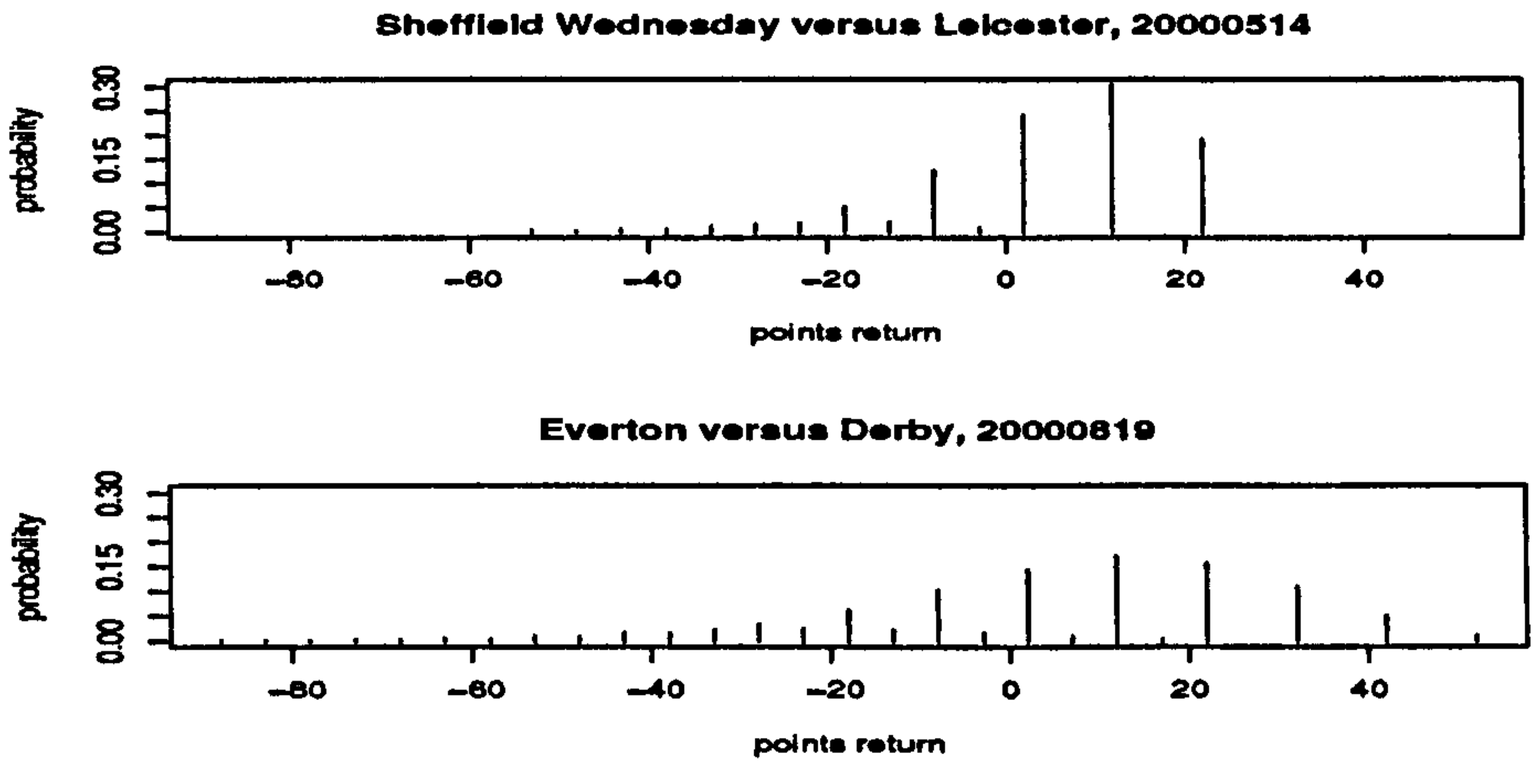


Figure 4.11: Density functions of returns on two bets with equal expected return but different variances

first match.

There is another more subtle consideration concerning the variance of the expected profit of each bet. So far, the variance of the parameter estimates has been used only in Section 4.3.6 in order to evaluate the significance of extra covariates (including indicator variables concerning the effect of historical rivalries or match incentives), with regard to assessing if they are worthwhile inclusions in a model. However, if a betting strategy should consider the total variance of the expected make-up, then in addition to the variance of the data given the conditional mean specified by the model, it may be worth considering the model within a Bayesian context and including also the variance of the parameter estimates that are present in the conditional mean for a given match. For example, the parameters employed in the prediction for a match involving a newly promoted team or a new referee are subject to more uncertainty than the parameters for a match involving teams and a referee that have been observed in many matches. Calculating the total variance of estimates of all parameters involved in the prediction of a match score is computationally awkward, but may be useful for more sophisticated betting strategies.

4.5 Possible improvements to the model

4.5.1 Hierarchical modelling using foul rates

One match statistic to which the bookings rate might well be related is the fouling rate. In particular, it is plausible that if the number of fouls in a particular match

was known, the prediction of the number of bookings would be influenced. Define HF_k, AF_k to be the number of fouls by the home and away sides in match k and HY_k, AY_k to be the number of yellow cards. A multivariate model can be formulated that suggests a distribution for the number of fouls and from that a distribution for the number of yellows can be deduced.

One possible model could be

$$\begin{aligned}
HF_k|\Theta^F_{k_h} &\sim Pois(g_1(\Theta^F_{k_h})) \qquad g_1 : \Re :\rightarrow \Re_+ \\
AF_k|\Theta^F_{k_a} &\sim Pois(g_1(\Theta^F_{k_a})) \\
HY_k|\Theta^Y_{k_h}, hf_k &\sim Bin(hf_k, g_2(\Theta^Y_{k_h})) \quad g_2 : \Re :\rightarrow \Re_+ \\
AY_k|\Theta^Y_{k_a}, af_k &\sim Bin(af_k, g_2(\Theta^Y_{k_a}))
\end{aligned}$$

where Θ^F represents a set of parameters which may determine the foul rate, such as the teams involved, and Θ^Y represents a set of parameters which determine the proportion of fouls which convert to yellow cards. These may also be team-specific. An approach similar to this is carried out on NFL match scores in the next chapter.

4.5.2 Dependence of home and away bookings

The assumption made throughout this chapter that the booking rates of the home and away sides are independent of each other simplifies the model but it does seem dubious. For example, if a side collects five bookings against a side which collects none, that appears to be a 'dirtier' performance than if the opposition had also collected five bookings, since in the latter case, the high bookings rate can be put down to the generally high match 'temperature'. Table 4.14 displays

$$\frac{\tilde{f}(i,j)}{\tilde{f}_H(i)\tilde{f}_A(j)}$$

for each joint home and away bookings rate (i,j) $i = 0,\ldots,9$ and $i = 0,\ldots,8$, where $\tilde{f}, \tilde{f}_H, \tilde{f}_A$ are the joint and marginal empirical probability functions for home and away bookings.

A pattern to Table 4.14 is observed. Entries on or close to the *(home bookings=away bookings)* diagonal generally occur more frequently, and entries away from the diagonal occur less frequently than would be expected under an independence assumption. Hence ideally a bivariate distribution which can model the surface of Table 4.14 would be found.

Table 4.14: Frequency of observed joint scores divided by expected frequency given independence assumption

		Away bookings								
		0	1	2	3	4	5	6	7	8
Home bookings	0	1.76	1.28	0.85	0.66	0.61	0.4	0.3	0	0
	1	0.93	1.01	1.09	0.89	1.11	1.08	0.76	0.45	0
	2	0.77	0.97	0.95	1.23	0.85	1.27	1.47	1.96	1.53
	3	0.44	0.8	1.11	1.24	1.34	1.48	0.96	1.15	2.67
	4	0.15	0.42	1.02	1.52	1.91	1.79	3.1	5.54	3.23
	5	0.44	0.22	1.06	1.97	1.94	0	2.24	0	9.31
	6	0	0.53	2.85	0	0	0	10.86	0	0
	7	0	0	0	5.15	0	0	0	0	0
	8	-	-	-	-	-	-	-	-	-
	9	0	0	0	5.15	0	0	0	0	0

4.6 Conclusion

Overall, the results obtained from the model implemented are quite encouraging since consistent profits are made for each year that a relatively naive betting strategy is simulated. In fact, the profit curves displayed in Figure 4.10 are likely to be conservative estimates since it is the average spread available from four bookmakers rather than the most favourable price offered that has been used to calculate hypothetical profit curves. Therefore many of the winning bets in practice would have resulted in slightly greater wins than recorded here and many of the losing bets would have resulted in slightly smaller losses. Also, more bets would have been placed if a larger range of spreads were available for each match. It seems reasonable to assume that these would also overall have been profitable.

One problem with betting on this market in practice is that the intrinsic high variability of booking rates in soccer means that all bets are relatively high risk. While it is true that with any gambling system stakes must be decided in such a way that the probability of financial ruin is kept to an acceptably small level, the non-negligible probability of very large make-ups (2.0% of matches result in a total points make-up of 100 or more) in booking rates means that any Sell bet is potentially risky. Approximately 65% of bets are Sells if a cut-off value of 4 is used when placing bets. In order to make large amounts of money by betting on this market one must be able to sustain occasional large losses. The next two chapters, which concentrate on modelling NFL and NBA scores for fixed odds betting (for which the maximum possible loss on any bet is restricted by the gambler), attempt to realise similarly profitable strategies but with a more stable return curve.

4.7 Additional comments and information

4.7.1 Generating model 2 predictions

Model 2, as employed in Section 4.4.2, creates predictions for yellow and red cards without recourse to formal statistical modelling. The prediction for the home number of yellows in a match is calculated as follows. First, the climate is estimated by calculating the mean of the total number of yellows collected in the fifty matches prior to the time when the match of interest takes place. The likely increase in booking rates at the start of the season is estimated by a similar method to that described in Section 4.3.3, by using the mean jump in the climate at the start of previous seasons. This number is added to the climate at the end of the previous season to obtain the climate for the first fifty matches of any season. Also, after the sudden drop in bookings observed in January 1999 (week number 127), the mean number of yellows in all matches since week number 127 is used, until week number 134 (which corresponds to approximately 50 matches). Figure 4.12 plots this climate.

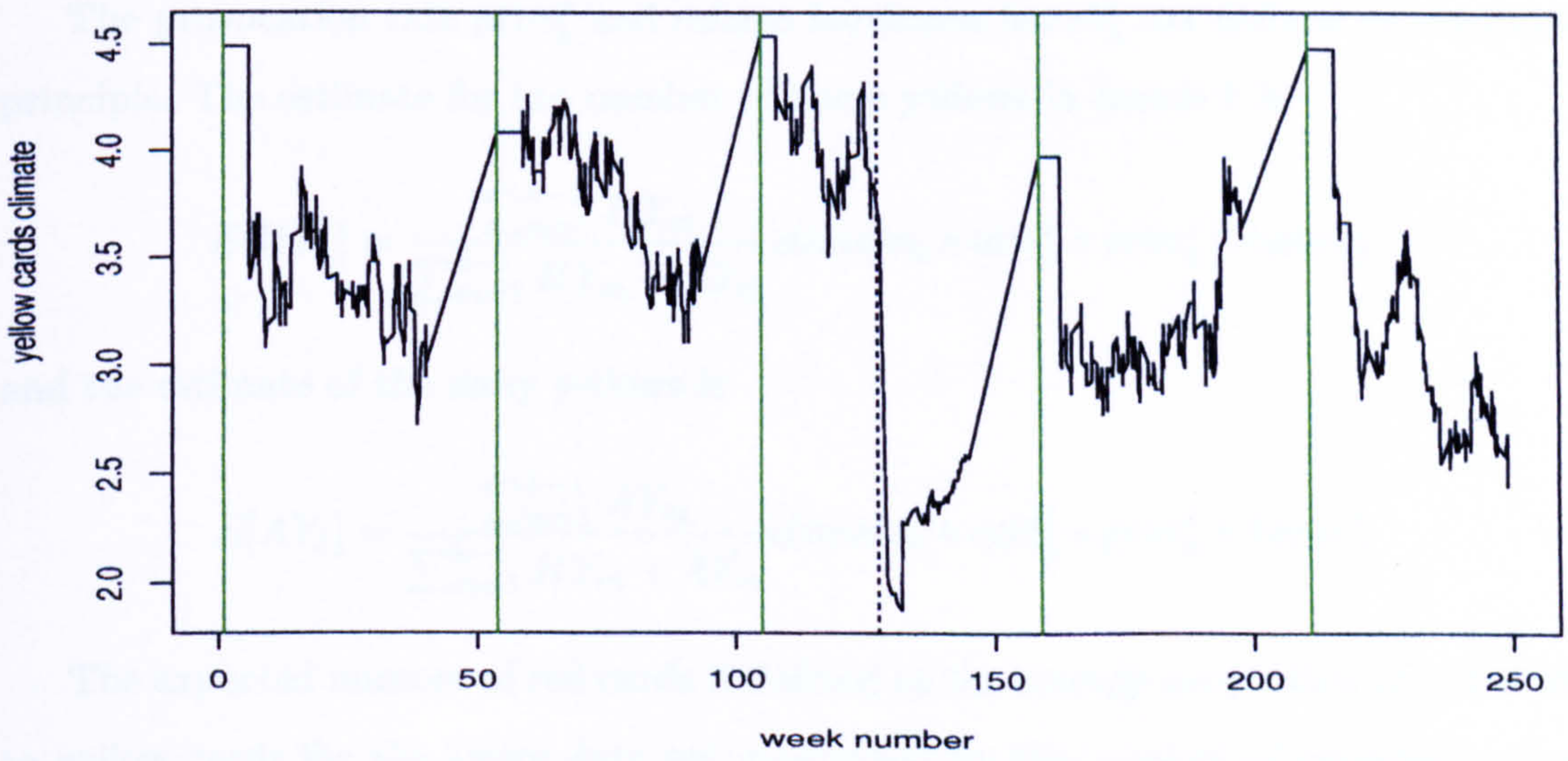


Figure 4.12: Predicted climate curve. The solid green lines denote the start of a new season, the dotted line denotes the time-point where referees were advised to be more cautious with regards to issuing cards

Next, estimates for teams' attacking and provoking parameters and the referees' harshnesses are needed. This is done using a weighted mean, weighted according to time. Let HY_k and AY_k represent the number of home and away yellows observed in match k between sides i and j. Suppose team i has played at home in

matches $i_H(1),\ldots,i_H(N_{iH})$ and away from home in matches $i_A(1),\ldots,i_A(N_{iA})$ prior to match k. The yellow yards they have collected in these matches are therefore $HY_{i_H(1)},\ldots,HY_{i_H(N_{iH})}$ and $AY_{i_A(1)},\ldots,AY_{i_A(N_{iA})}$. Also, let $t(k)$ be the time match k takes place.

Then when team i plays in match k the estimate of their attacking rate is defined as follows:

$$attr_k^i = \frac{\sum_{m=1}^{N_{iH}} HY_{i_H(m)} * exp(-w * (t(k-1) - t(i_H(m))))}{\sum_{m=1}^{k-1} HY_m} + \frac{\sum_{m=1}^{N_{iA}} AY_{i_A(m)} * exp(-w * (t(k-1) - t(i_A(m))))}{\sum_{m=1}^{k-1} AY_m} \quad (4.7.1)$$

The rate defined by Equation 4.7.1 is a time-weighted mean of all of team i's home and away yellows, divided by the mean home and away yellows in all matches before match k. Equation 4.7.1 is equal to 1 if team i has an average booking rate, compared to all teams. The weighting factor w is set to be the same value (0.02) as that selected in Section 4.3.5, where the values for the external parameters for the yellow cards model were determined.

The provocation rate $prov_k^j$ and referee harshness $harsh_k^r$ are defined on a similar principle. The estimate for the number of home yellows in match k is

$$E[HY_k] = \frac{\sum_{m=1}^{k-1} HY_m}{\sum_{m=1}^{k-1} HY_m + AY_m} climate_k * attr_k^i * prov_k^j * harsh_k^r$$

and the estimate of the away yellows is

$$E[AY_k] = \frac{\sum_{m=1}^{k-1} AY_m}{\sum_{m=1}^{k-1} HY_m + AY_m} climate_k * attr_k^j * prov_k^i * harsh_k^r$$

The expected number of red cards is defined as the average proportion of red cards to yellow cards for the entire data set, multiplied by the number of expected yellow cards for that match:

$$E[HR_k] = \frac{\sum_{m=1}^{k-1} HR_m + AR_m}{\sum_{m=1}^{k-1} HY_m + AY_m} * E[HY_k]$$
$$E[AR_k] = \frac{\sum_{m=1}^{k-1} HR_m + AR_m}{\sum_{m=1}^{k-1} HY_m + AY_m} * E[AY_k]$$

Finally, in the event of a new team or referee entering, the values obtained through equations of the type observed in Equation 4.7.1 are replaced by the home/away climate

until the team or referee has participated in five matches.

4.7.2 Kernel Regression

Kernel regression is a non-parametric regression technique. In general, non-parametric regression is attractive when there is no obvious appropriate structure (e.g. linear, trigonometric, polynomial) for the curve that best fits the relevant data. One non-parametric regression technique, which has its roots in density estimation, is the kernel regression technique. There are several possible implementations of kernel regression (see Silverman (1986) for examples).

Given i.i.d. data $(X_1, Y_1), ..., (X_N, Y_N)$, a suitable form that represent Y_i as a function of the X_i is required. A kernel function $K(t)$ can be thought of as a generalisation of a weight function, which satisfies the condition that $\int_{-\infty}^{\infty} K(t)dt = 1$. There are various estimators that make use of kernel functions, one of the more popular choices being the *Nadayara-Watson* estimator, as outlined in Wand and Jones (1995):

$$\hat{Y(x)} = \frac{\sum_{k=1}^{N} K_h(x - x_k) Y_k w_k}{\sum_{k=1}^{N} K_h(x - x_k) w_k}$$

where w_k is the square root of the number of observations with value x_k and K_h is the kernel function with bandwidth h. The next decision is the choice of kernel function. There are several, which have different properties. Most are conceived with the aim of minimising the *mean integrated square error* (MISE), defined by

$$E \int \hat{f}(t) - f(t)^2 dt = \int E\hat{f}(t) - f(t)^2 dt + \int var\hat{f}(t)dt \qquad (4.7.2)$$

where $f(t)$ is the function to be estimated with $\hat{f}(t)$, which in this case is the Nadayar-Watson estimator. Equation 4.7.2 gives the MISE as the sum of the integrated square bias and the integrated variance. Silverman (1986) details various methods that can be employed to find kernel functions which result in small values of MISE. One of these is the Epanechnikov kernel which is given by

$$K_h(x) = \begin{cases} \frac{3}{4}(1 - \frac{x^2}{5h^2})/(h\sqrt{5}) & \text{for } |x| < h\sqrt{5} \\ 0 & \text{otherwise} \end{cases}$$

and it is this kernel that has been applied in this chapter. It also requires relatively

little computational effort, which is another important criterion.

It is also necessary to choose a suitable bandwidth h. There are various ways of doing this, although it is to a large extent dependent on the intended application of the regression. In some cases it may be necessary to have an automated process that chooses h by some objective process. In this case, since suitably powerful software is available, it is possible to try out various values, look at the resulting curves, and make a decision based on existing knowledge of the climate. Figure 4.13 displays curves resulting from various choices of bandwidth. The curve arising from bandwidth set to

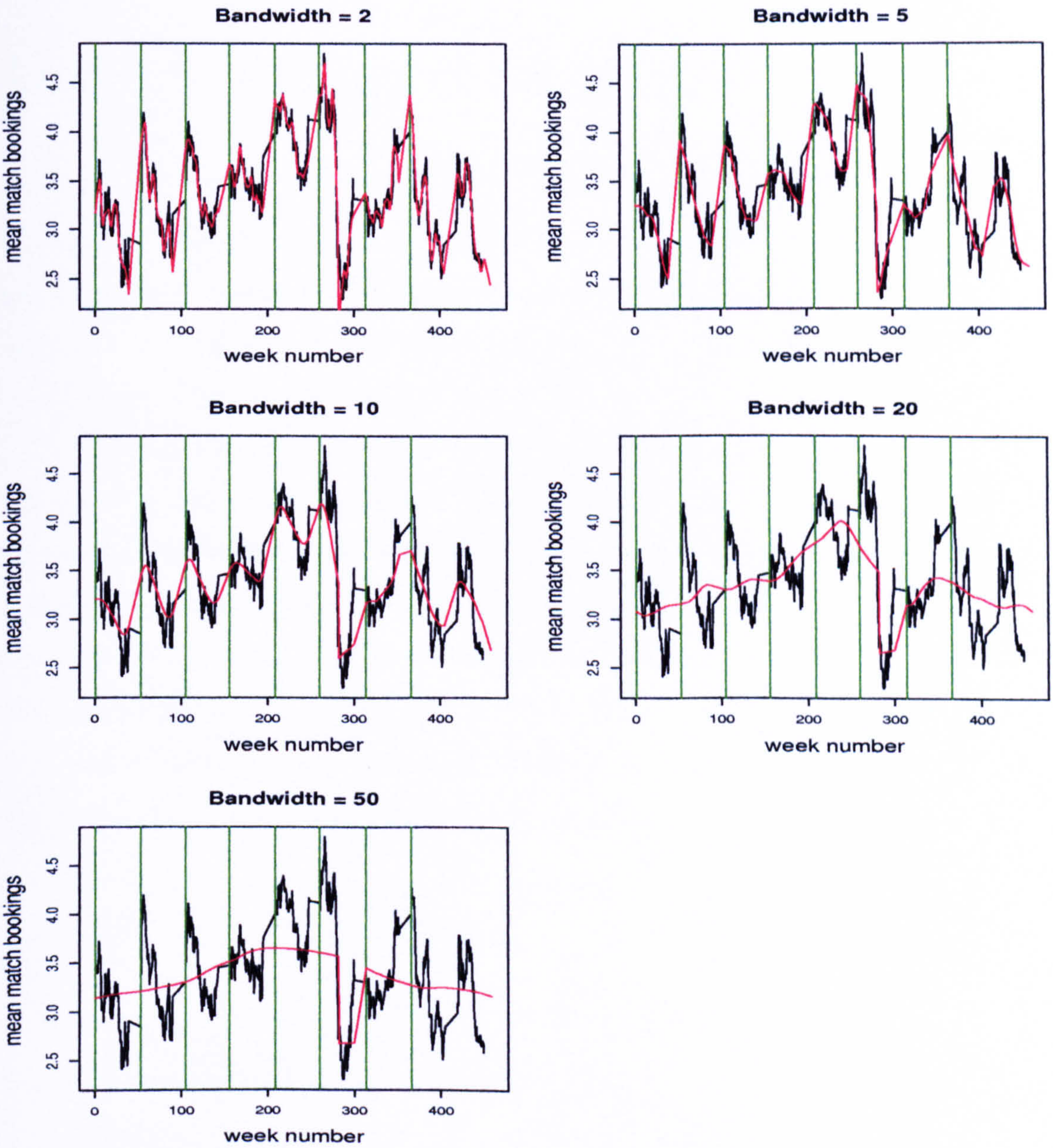


Figure 4.13: Kernel regression estimates for different choices of bandwidth. The solid black line represents observed moving average, the red line represents estimated climate using that bandwidth

be 5 seems to provide the best fit.

Chapter 5

Estimating NFL scores: the threes and sevens distribution

NFL has traditionally been one of the most popular sports in the United States, among gamblers and the general public alike. Betting on NFL is extremely popular and there are numerous casinos that offer bets on the sport, usually in the form of a fixed odds handicap bet (discussed in Section 1.3.1).

In this chapter firstly a brief overview of NFL is given, explaining the structure of the season and the game regulations. Section 5.2 describes the data available while in Section 5.3 a basic model, assuming two independent Normal distributions for the home and away scores, is specified and fitted. Section 5.4 attempts to find an alternative to the Normal distribution in order to represent the home and away scores. More match data is incorporated into a large multivariate structure in Section 5.5 in an attempt to find more accurate score predictions while Section 5.6 presents a more straightforward use of this extra data. Section 5.7 presents the conclusions to the chapter.

5.1 NFL - a brief summary

5.1.1 NFL season structure

The NFL season is divided into two stages:

- *the regular season* This involves six leagues containing five or six teams each[1]. Normally half of the matches take place between teams within the same league,

[1]This is true for the data set being analysed which contains matches until January 28 2001. At the start of the 2002/2003 season teams were reallocated into eight divisions each containing four teams

with the remaining matches being against selected teams from other leagues. Opponents for matches outside a team's league are selected by the NFL administration so that successful teams from the previous season play other successful teams, similarly for unsuccessful teams, in an attempt to handicap the better teams. Approximately 30 teams play 16 games each season, hence approximately 240 games are played during each regular season.

- *the play-offs* The twelve most successful teams from the regular season filter into a knockout tournament. The final match of this tournament is the *Superbowl* and is one of the most popular sporting events worldwide.

5.1.2 NFL game structure

The matches consist of four fifteen minute periods. Each team consists of two separate squads of players, one being the offensive squad, one being the defensive squad. Each squad contains 11 players. At the start of the first quarter one side is designated to be in possession of the ball and this side fields its offensive squad while the side not in possession of the ball fields its defensive squad. Upon a change of possession of the ball, which can take place in several ways, the offensive players of the side that has just lost possession are substituted by the defensive players in their side, while the side that has won possession replaces its defensive squad with its offensive players. A detailed explanation of many details of the match regulations and the important aspects of NFL matches, such as the ways in which possession of the ball can be lost, is deferred to Section 5.5.1 (in order to understand the intervening sections of this chapter, a thorough knowledge of such details is not required). Points are scored either through

- **Field Goals:** these are scored when a team kicks the ball through a set of raised posts at the opponent's end of the field and are worth 3 points.

- **Touch Downs:** these are scored when a team places the ball over a line at the opposing team's end of the pitch and are worth 6 points.

- **1-Point Conversions:** after a Touch Down is scored, a team is given one extra play. Should they successfully kick the ball between the raised set of posts at the opposing end of the field using this play, they score one extra point.

- **2-Point Conversions:** if, after a Touch Down, the team succeeds in placing the ball over the line at the opponent's end of the field with the extra play, they

score two extra points.

- **Defensive Conversions:** these occur when a team scores a Touch Down at their own end of the field. Their opponents score 2 points in this situation. Teams would only do this if their opponents are likely to score a Touch Down otherwise.

Table 5.1 displays the frequency of the events listed above.

Table 5.1: Average frequency of scoring opportunities in each match

	home	away
Field Goals	1.524	1.405
Touch Downs	2.539	2.132
1-Point Conversion	2.333	1.893
2-Point Conversions	0.069	0.089
Defensive Conversions	0.042	0.029

If the two teams have an equal number of points after the four periods, an extra period, known as *overtime*, is played. This period ends as soon as one side scores either a Touch Down or Field Goal, with this side being declared the winner[2].

5.2 NFL data

Two data sets are available for this analysis and they are described below.

1. NFL final scores for the home and away side for seasons 1983/84 - 2000/01, along with a bookmaker's line for score differences.

2. For seasons 1997/98 - 2000/01 the following figures are available for both the home and away side,

   - the final match score
   - the points scored in each quarter of the match, including any overtime periods
   - the number of Touch Downs, Field Goals, 1-Point Conversions, 2-Point Conversions and Defensive Conversions scored in each match
   - the match totals for yards passed, yards rushed, number of attempted passes, number of completed passes, number of rushes, number of inter-

[2]Strictly speaking this period would also end if one side scored a defensive conversion and thus yielded two points to the opposing side. This would be a bizarre tactic however, since it would result in the side immediately losing the match.

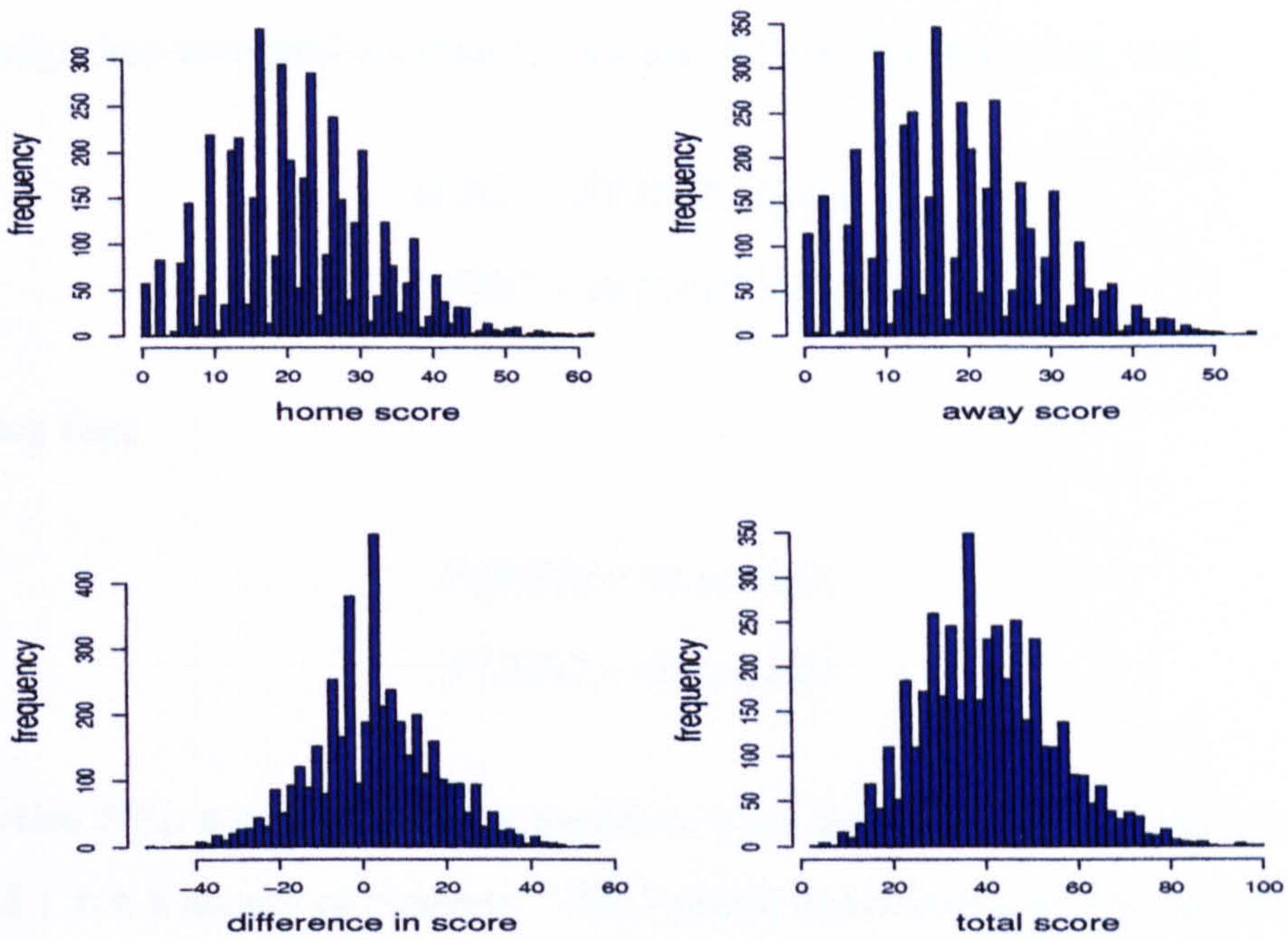


Figure 5.1: NFL score histograms 1983-2001

ceptions and time in possession of the ball (these terms are explained when used in Section 5.5)

- a bookmaker's line for score differences and total scores

Sections 5.3 and 5.4 uses the first data set, while Section 5.5 uses the second.

5.3 A basic model for NFL scores

Figure 5.1 displays histograms for home scores (HSC), away scores (ASC), score differences and score totals. The home mean, away mean, home standard deviation, away standard deviation and home and away correlation for scores are 22.15, 19.02, 10.41, 9.97 and -0.03 respectively. Two independent univariate Normal distributions seems to be the most obvious distribution to employ in order to model the home and away scores and this was the distribution chosen in several previous studies of NFL, including Stern (1991), Harville (1980) and Glickman and Stern (1998). Stern conducts a Kolmogorov-Smirnov test which rejects Normality at the 5% significance level. Harville comments that 'approximations to the posterior probabilities may be somewhat crude, however it is not clear how to improve on them by other than ad hoc procedures'. Due to the way in which points are scored in NFL, scores which are combinations of 3s and

7s are more likely to occur. Furthermore, by applying the Normal distribution one may assign non-zero probabilities to negative scores. By assuming that

$$HSC \sim \mathcal{N}(22.15, 10.41)$$
$$ASC \sim \mathcal{N}(19.02, 9.97)$$

it follows that

$$P(HSC < 0) \approx 0.015$$
$$P(ASC < 0) \approx 0.025 \tag{5.3.1}$$

In practice NFL scores cannot be negative, thus the non-zero probabilities in Equation 5.3.1 are a source of concern. The Normal distribution will be employed in the first model attempt but concerns about its suitability, along with some alternative approaches, are discussed in Section 5.3.1.

The small correlation coefficient between the home and away scores suggests that the dependence between them is not straightforward enough to be modelled by a bivariate Normal distribution. However, this does not suggest that the home and away scores are independent. In NFL, possession is crucial and any possession of the ball by one side implies lack of ball possession by the other, which restricts their scoring opportunities. The linear models summarised in Table 5.2 reveal some curious trends.

Table 5.2: *Coefficients and significance levels, modelling NFL Home Score (HSC) against Away Score (ASC), Home Rushed Yards (HRY) and Away Rushed Yards (ARY)*

Model	Coefficients and p-values
HSC~ASC	ASC: (0.00211,0.94798)
HSC~ASC+HRY	ASC: (0.10526,0.00056), HRY: (0.08326,0)
HSC~ASC+HRY+ARY	ASC: (0.18828,0), HRY: (0.07449,0), ARY: (-0.04668,0)

While there is clearly enormous dependence between the play of the two teams, the structure of this dependence is not immediately obvious.

For now a straightforward model is specified which can later be modified where necessary. For match k that takes place between team $i(k)$ and team $j(k)$, at team $i(k)$'s ground,

$$HSC_k \sim \mathcal{N}(\mu_k, \sigma)$$
$$ASC_k \sim \mathcal{N}(\lambda_k, \sigma) \tag{5.3.2}$$

where

-

$$\mu_k = \gamma + \alpha_{i(k)} + \beta_{j(k)} + \delta$$
$$\lambda_k = \gamma + \alpha_{j(k)} + \beta_{i(k)} \tag{5.3.3}$$

- σ is the standard deviation for home and away scores
- γ is the global mean
- $\alpha_{i(k)}, \alpha_{j(k)}$ are offensive parameters for respectively the home and away teams
- $\beta_{i(k)}, \beta_{j(k)}$ are defensive parameters for respectively the home and away teams
- δ is the home effect

This model will be referred to as the *basic model.*

In order to obtain MLEs for the parameters included in the model specified by Equation 5.3.2, values for the external parameters, as defined in Chapter 3, must be fixed. The process used in Section 4.3.5 concerning the analysis of bookings rates is repeated here, by trying a range of values for these parameters and monitoring the predictive likelihood. Table 5.3 displays the predictive likelihood for different sets of values of the external parameters, and it appears that the near-optimal (time-down-weighting (ς), offensive/defensive prior tightnesses ($\tau_{\alpha\beta}$), seasonal truncation (w)) values are (0.05,5,20), which are highlighted in red. Table 5.4 displays the estimates of the team parameters for this model at the final time-point in the data set. In contrast to the estimates presented for Premier League soccer team abilities in Chapter 4, it is rare that NFL teams have both a strong offense and a strong defense. The drafting system used by the NFL that is described in section 2.3 puts a ceiling on the number of highly rated players that any squad can contain. As a result teams are forced to make compromises concerning the quality of some sections of their squad. In the case

Table 5.3: Predictive likelihood obtained for different choices of external parameters for final scores

Truncation w= 5 weeks:

		Prior variance $\tau_{\alpha\beta}$ of offensive and defensive estimates			
		2	5	10	20
	0.005	-5696.8346	-5693.0002	-5698.3407	-5701.1115
	0.01	-5694.2798	-5687.5982	-5693.1642	-5696.1395
Weight ς	0.02	-5692.254	-5678.6826	-5684.7334	-5688.226
	0.05	-5696.3219	-5665.0395	-5673.7137	-5680.1769
	0.1	-5708.4818	-5663.6557	-5680.9272	-5698.2981

Truncation of w= 10 weeks:

		Prior variance $\tau_{\alpha\beta}$ of offensive and defensive estimates			
		2	5	10	20
	0.005	-5696.4878	-5691.8315	-5697.2038	-5700.0071
	0.01	-5693.5382	-5685.5082	-5691.1315	-5694.1826
Weight ς	0.02	-5691.844	-5675.5469	-5681.7373	-5685.4487
	0.05	-5697.8174	-5662.9761	-5672.9169	-5680.8305
	0.1	-5710.4193	-5665.455	-5687.4899	-5710.9848

Truncation of w= 20 weeks:

		Prior variance $\tau_{\alpha\beta}$ of offensive and defensive estimates			
		2	5	10	20
	0.005	-5695.9376	-5689.5608	-5694.9919	-5697.8628
	0.01	-5692.3865	-5681.6207	-5687.3551	-5690.5718
Weight ς	0.02	-5691.7564	-5670.4412	-5677.0025	-5681.249
	0.05	-5700.7032	-5662.4368	-5676.0887	-5687.9537
	0.1	-5712.6495	-5670.3222	-5703.2531	-5741.8719

Truncation of w= 30 weeks:

		Prior variance $\tau_{\alpha\beta}$ of offensive and defensive estimates			
		2	5	10	20
	0.005	-5695.6078	-5686.8531	-5692.3737	-5695.3268
	0.01	-5690.8531	-5677.4607	-5683.3596	-5686.7768
Weight ς	0.02	-5691.3705	-5665.9569	-5673.183	-5678.1523
	0.05	-5706.0655	-5667.5716	-5687.5513	-5705.6403
	0.1	-5719.4015	-5679.3656	-5727.9586	-5789.005

of St Louis and Miami in particular, it is clear which aspect of the game they have chosen to specialise in.

Figure 5.2 plots a moving average of predicted scores versus observed scores, for the home scores, away scores, scores differences and total scores. It reveals that the predictions appear to be broadly accurate. Some instability is observed towards the left hand and right hand edges of the plots. This is due to the number of observations upon which the values are calculated decreasing in these areas.

Note that while Figure 5.2 suggests that the model in general makes sensible predictions it does not prove that there are no systematic biases within the model. For example if a team plays a fixture without one or more of their most highly valued players, their expected score supremacy is usually lower. Since on average teams benefit

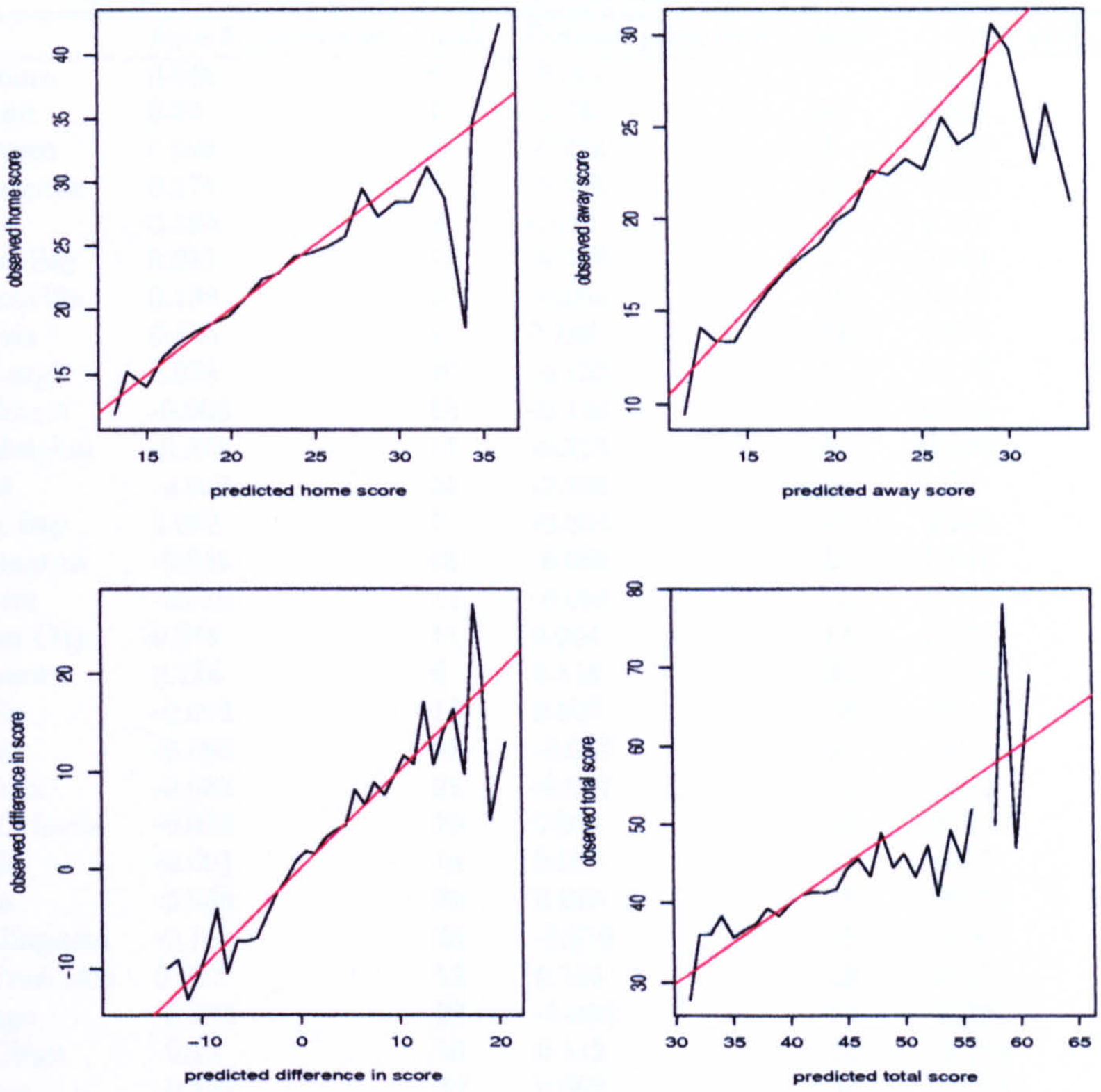


Figure 5.2: Plot of moving average of predicted scores versus moving average of observed scores

from injuries to their opponents as often as they suffer from their own injuries, the net effect of injuries for both sides across all matches is approximately zero. However, there are biases in the predictions for matches where one, or both, of the squads is significantly weaker than usual.

Not all previous attempts to model NFL have involved two parameters to represent team abilities, with many employing only a single parameter. However, Section 5.8.2 in the additional comments section of this chapter outlines and implements a technique which compares the predictive power of models using differing numbers of parameters to represent team abilities. The results suggest that using two parameters seems suitable.

Only one parameter, δ, is used to represent the effect of playing at home although Glickman and Stern (1998) employed a separate home effect parameter for each team. The method they used to test the need for such a specification is outlined briefly in Section 2.5.1. It is plausible that with games being played in such a variety of climates,

Table 5.4: Rankings of all NFL teams after January 28, 2001

Team	Attack parameter	rank	Defense parameter	rank	Overall ability	rank
Baltimore	0.048	9	-0.391	1	0.439	1
Oakland	0.22	2	-0.095	9	0.315	2
Tennessee	0.048	8	-0.224	2	0.271	3
Indianapolis	0.176	4	-0.006	14	0.182	4
Denver	0.184	3	0.012	20	0.172	5
Tampa Bay	0.011	13	-0.138	4	0.149	6
Jacksonville	0.138	5	-0.004	16	0.141	7
St Louis	0.325	1	0.188	29	0.137	8
Pittsburgh	0.024	10	-0.102	7	0.126	9
NY Giants	-0.003	15	-0.126	5	0.123	10
Philadelphia	-0.025	18	-0.121	6	0.096	11
Miami	-0.069	24	-0.139	3	0.07	12
Green Bay	0.062	7	-0.004	15	0.066	13
Washington	-0.051	20	-0.099	8	0.048	14
NY Jets	-0.018	17	-0.054	10	0.036	15
Kansas City	0.019	11	0.004	18	0.015	16
Minnesota	0.128	6	0.119	27	0.008	17
Buffalo	-0.005	16	0.007	19	-0.012	18
Detroit	-0.065	23	-0.047	11	-0.018	19
Carolina	-0.053	21	-0.027	12	-0.026	20
New Orleans	-0.033	19	0.024	22	-0.057	21
Seattle	-0.001	14	0.069	23	-0.07	22
Dallas	-0.056	22	0.019	21	-0.075	23
New England	-0.112	25	-0.018	13	-0.094	24
San Francisco	0.017	12	0.155	28	-0.137	25
Chicago	-0.223	29	-0.003	17	-0.22	26
San Diego	-0.13	26	0.112	26	-0.242	27
Atlanta	-0.151	27	0.093	25	-0.244	28
Cincinnati	-0.208	28	0.075	24	-0.283	29
Arizona	-0.233	30	0.205	31	-0.438	30
Cleveland	-0.281	31	0.197	30	-0.478	31

and with journeys to some games being particularly long, the disadvantage of playing at other grounds is not homogeneous. In all the models employed in this chapter only one parameter is used to represent the effect of playing at home although further research may cast doubt on the validity of this assumption.

5.3.1 Discussion: suitability of Normal distribution

Figure 5.3 displays the histogram of actual scores, versus the density of scores predicted using the basic model, given three different predicted score intervals. It can be seen that scores are not Normally distributed and indeed they do not follow any standard statistical distribution. To understand the distributions observed in Figure 5.3 the way points are collected in NFL needs to be considered.

Referring to Table 5.1, it is noted that almost all points are obtained via Field Goals (3 points), Touch Downs (6 points) and subsequent 1-Point Conversions after scoring a Touch Down. Figure 5.4 displays the histogram for the entire set of final

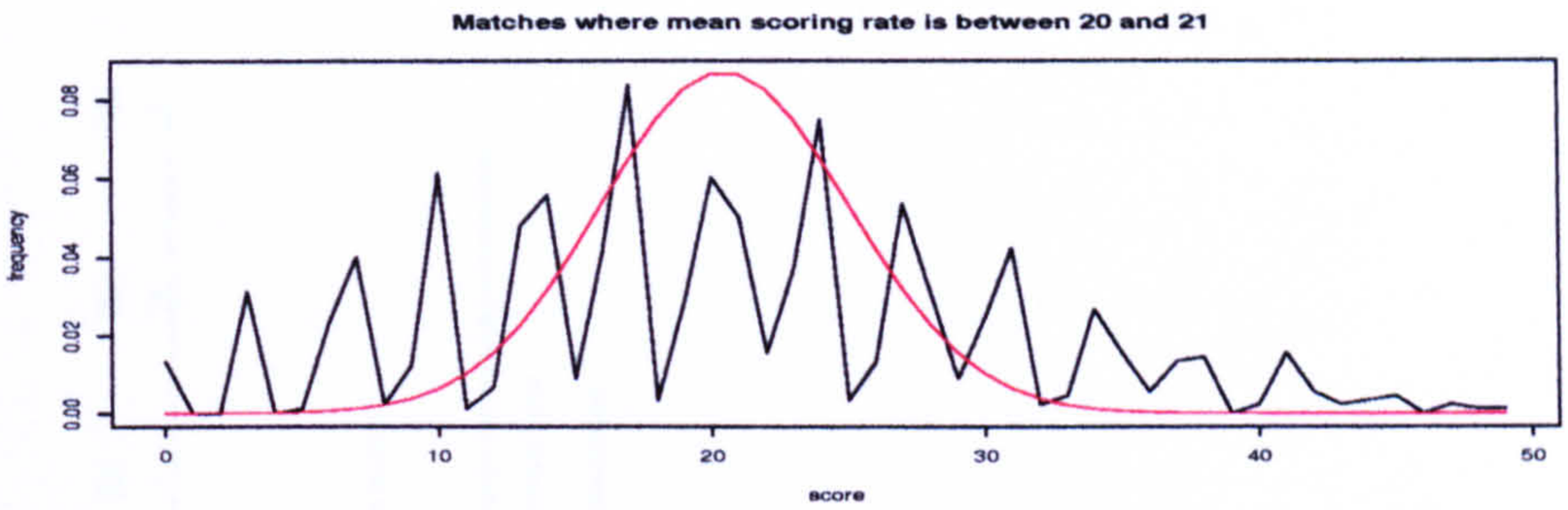


Figure 5.3: Plots of observed histograms of score frequencies (___) along with theoretical frequencies obtained assuming normal distribution applies (___) given three different match means

scores, and peaks are observed at all numbers which are combinations of a low number of 7s or 3s.

It is important to have a reasonably accurately specified distribution function for scores when betting. As discussed in Section 1.3 one of the most widely available betting markets for NFL is handicap betting. To illustrate the problem that arises by using the Normal distribution to predict scores, two possible betting situations are considered. For the purposes of these examples, the term 'score difference' is used to signify the home score minus the away score of a match. It is frequently of interest to know if P(score difference > handicap). Suppose the basic model gives a predicted score difference, $E(X-Y)$, of 2.5 points, while the handicap offered by the bookmaker is -2.5 points (i.e. it believes the median score supremacy of the home team over the away team is 2.5 points). A bet on the home side is won providing $X - Y \geq 3$. According to the basic model and applying a continuity correction, the probability of

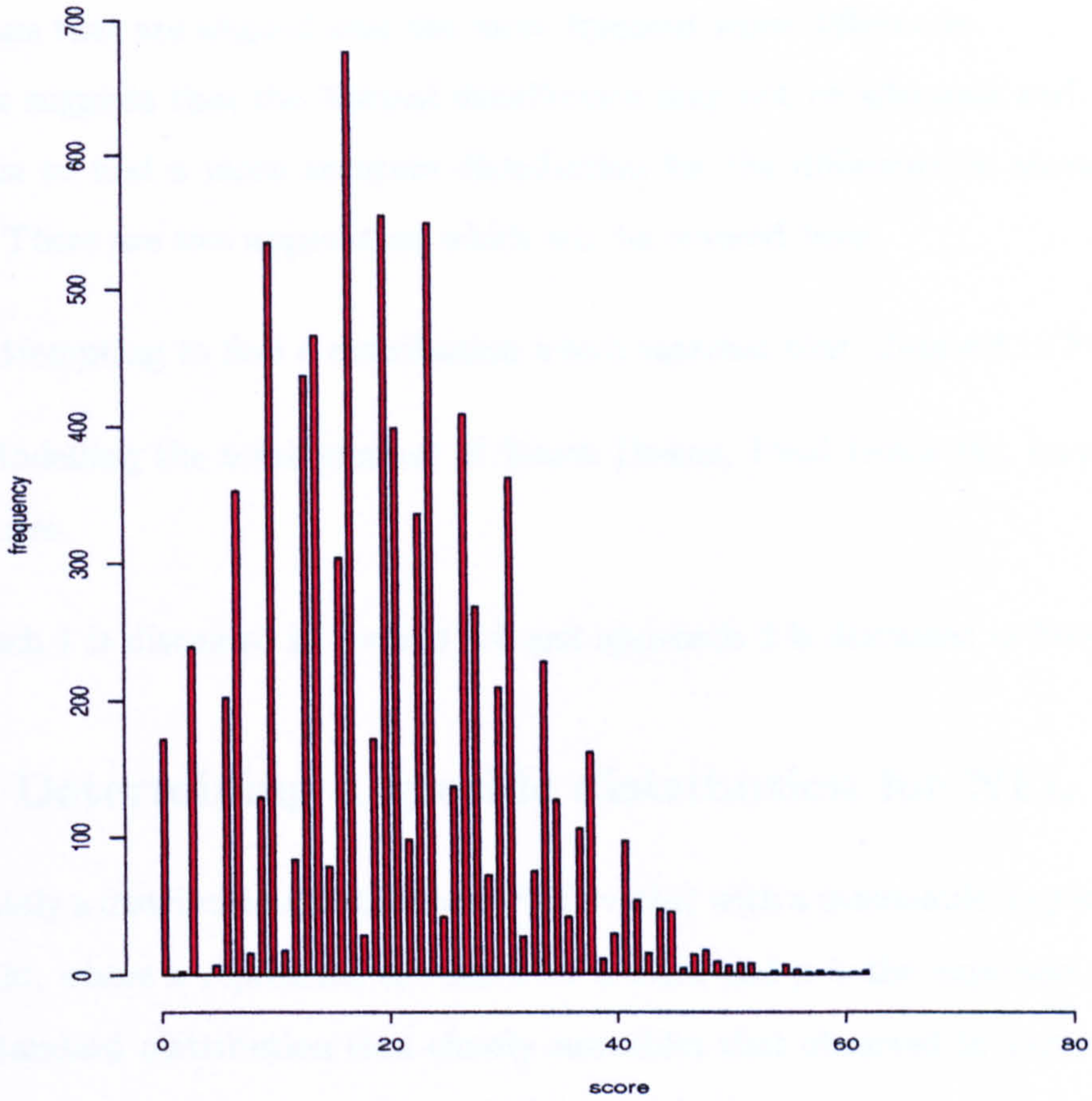


Figure 5.4: Histogram of all scores, either side, 1983-2000

winning the bet is $P(X - Y > 2.5) = 0.5$ where $X - Y \sim \mathcal{N}(2.5, \sqrt{2} * 9.14)so^3$ it does not appear to be worth making such a bet. However, of the matches where the basic model predicts a score difference between 1.5 and 3.5, 54.9% have a final score difference ≥ 3. Hence the basic model estimates this bet to be less attractive than it is.

Meanwhile, suppose for another match that the basic model predicts that $E(X - Y) = 4$ and the bookmaker offers a handicap of -3. In this case it is tempting to back the home team and such a bet is won providing $X - Y \geq 4$. According the basic model, the probability of winning this bet is $P(X - Y > 3.5) = 0.515$, where $X - Y \sim \mathcal{N}(4, \sqrt{2} * 9.14)$. Of the matches where the basic model predicts a score difference between 3 and 5, only 49.8% have a final score difference ≥ 4. In this case, the basic model thinks this bet is more attractive than it is, since it is unaware that only a small number of matches (134) have a final score difference of 4 but many

[3] the MLE obtained for σ with the basic model is 9.14

more matches (340) have a score difference of 3. A similar trend is observed for other handicaps that are aligned near the more frequent score differences.

This suggests that the Normal distribution may not be adequate and it would be desirable to find a more accurate distribution for the difference in scores and total scores. There are two approaches which will be covered here.

1. Attempting to find a distribution which matches that observed in Figure 5.3.
2. Modelling the total number of Touch Downs, Field Goals etc, instead of total score.

Approach 1 is discussed in Section 5.4 and approach 2 is discussed in Section 5.5.

5.4 Determining a specific distribution for NFL

Ultimately a distribution that reflects $P(X = x|\mu)$ with a reasonable degree of accuracy is sought, where x represents the score for a team and μ is the expected score. There is no standard distribution that closely resembles that observed in Figure 5.4 so it is necessary to develop a nonparametric density of some kind.

It may seem attractive to use the values of score frequencies plotted in Figure 5.3 as the probabilities. So, for example, $P(X = 0|\mu = 20.5) = 20/925$, since of the 925 matches where either team's mean scoring rate was between 20 and 21, 20 resulted in a score of zero. However, the problem with this solution is seen by considering Figure 5.5.

Here the proportion of occasions in which the score was 21, given differing values of the predicted score, is plotted. The predicted scores are obtained using the model described in 5.3. To adjust for the continuity of the predicted score, $P(X = 21|\mu)$ is defined as the proportion of matches for which $\mu \in (\mu - \epsilon, \mu + \epsilon)$, for some chosen value of ϵ (0.1 in this case). A smoother version of Figure 5.5 is preferable, and by obtaining smooth versions for all scores, a full density for $P(X = x|\mu)$ for all values of x and all values of μ is obtained. To clarify this process, consider Table 5.5.

For example, 0.05714 of the matches where $\mu \in (25.65, 25.75)$ resulted in a score of 21. The NAs signify that no match actually had that predicted mean in the dataset. The rows in Table 5.5 sum to 1. The accuracy of the probabilities in the rows of Table 5.5, which represent the density of interest, is improved by smoothing down the columns. Kernel regression (described in Section 4.7.2) is applied in order to achieve

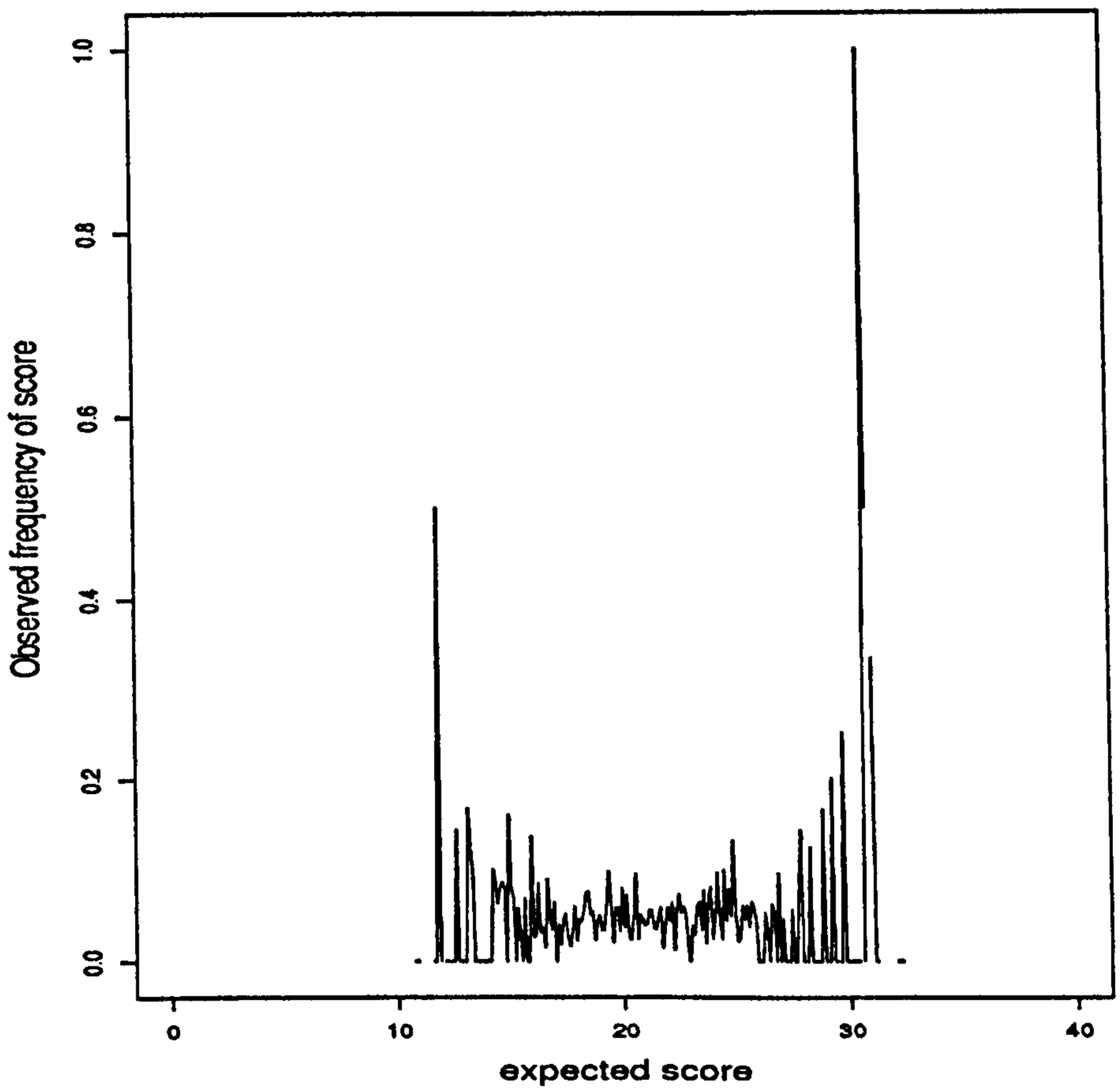


Figure 5.5: Frequency plot of P(score=21|μ), for different values of μ

Table 5.5: Observed proportions of scores, for given means

		scores frequency given μ									
		0	1	2	...	21	22	23	...	98	99	100
	0.0	NA	NA	NA		NA	NA	NA		NA	NA	NA
	0.1	NA	NA	NA		NA	NA	NA		NA	NA	NA
	0.2	NA	NA	NA		NA	NA	NA		NA	NA	NA
	...											
μ	25.7	0	0	0		0.05714	0.02857	0.14286		0	0	0
	25.8	0.04545	0	0		0.04545	0	0.04545		0	0	0
	25.9	0	0	0		0	0	0		0	0	0
	...											
	39.8	NA	NA	NA		NA	NA	NA		NA	NA	NA
	39.9	NA	NA	NA		NA	NA	NA		NA	NA	NA
	40.0	NA	NA	NA		NA	NA	NA		NA	NA	NA

this. In effect smooth versions for function $f(\mu) = P(X = x|\mu)$ are obtained for all observed values of X.

This density will be referred to as the *NFL distribution*. Figure 5.6 contrasts the density obtained for the scores 0, 7 and 21, once smoothing has been applied.

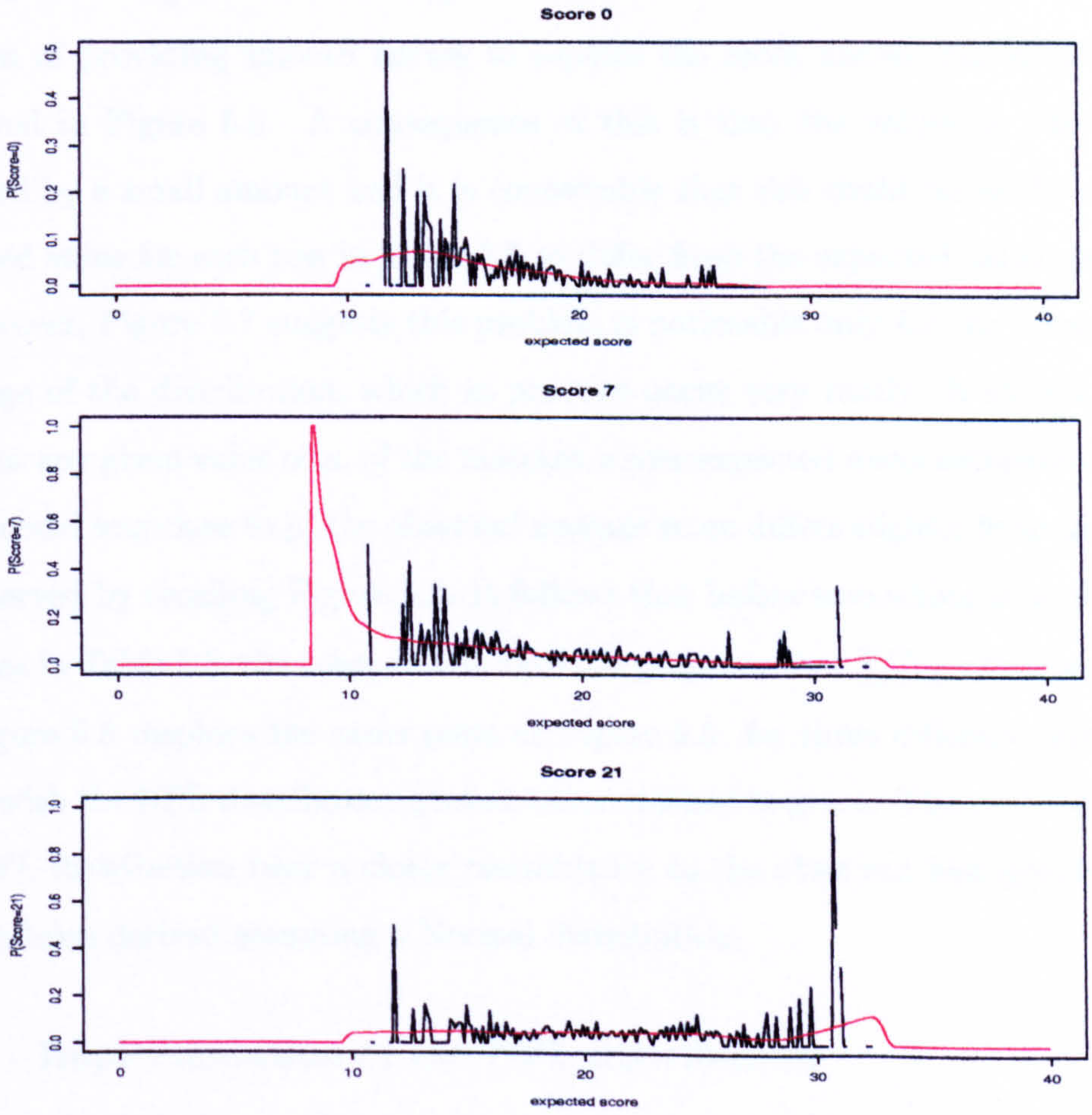


Figure 5.6: Plot of $f(\mu) = P(X = x|\mu)$ (___) with kernel-smoothed curve overlaid (___), for $x = (0, 7, 21)$

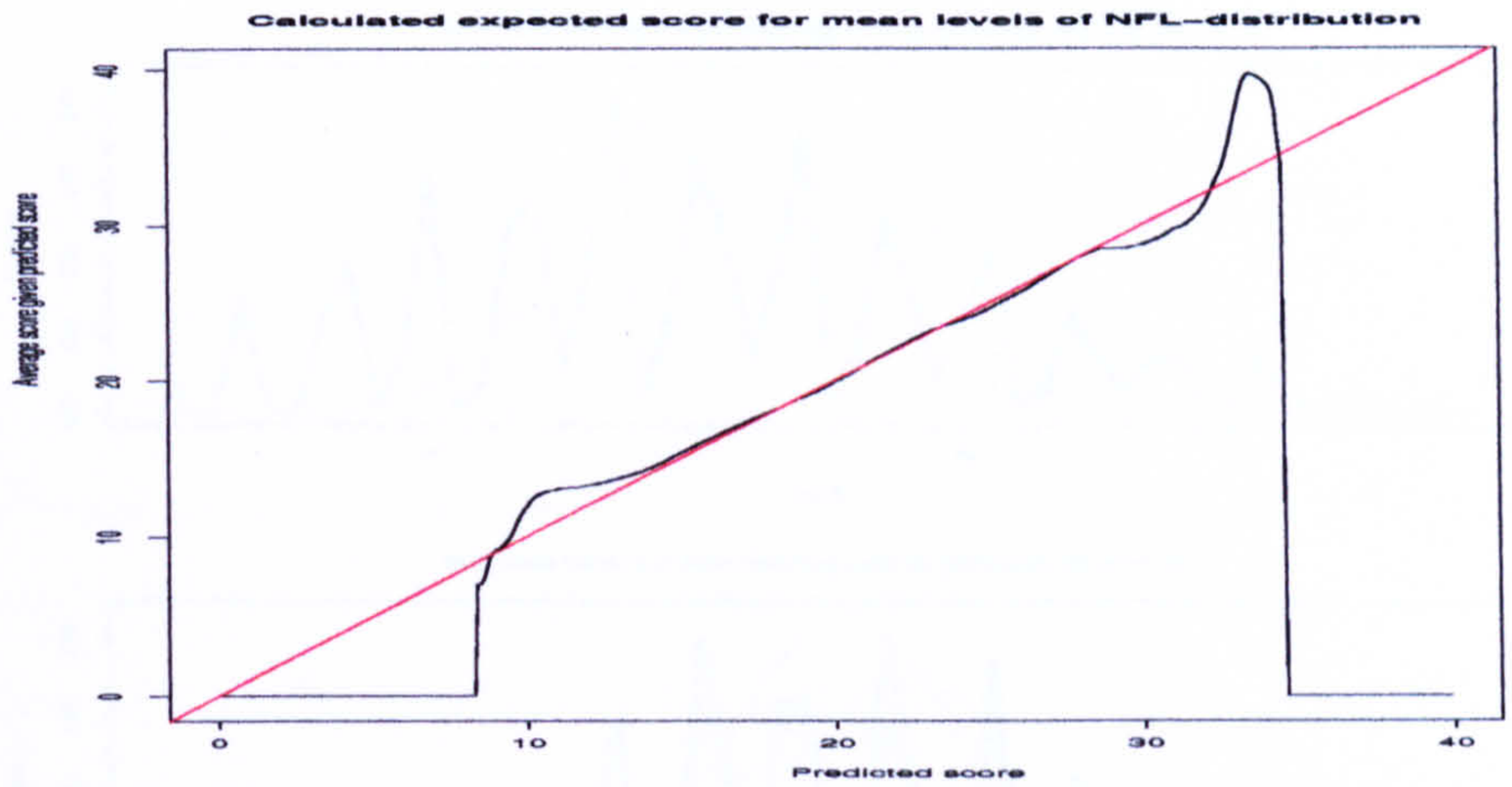


Figure 5.7: Plot of $\sum_{x=0}^{99} xP(x|\mu)$, for each value of μ, where the probabilities are those of the NFL distribution.

One possible concern by applying this technique is that, after the smoothing is applied to $P(X = x|\mu)$ for fixed values of x, there may be values of μ for which $\sum_{x=0}^{\infty} xP(X = x|\mu) \neq \mu$. This may arise since the kernel smoothing is applied with the aim of providing smooth curves to replace the more uneven curves of the type displayed in Figure 5.5. A consequence of this is that the values in Table 5.5 are adjusted by a small amount and it is conceivable that this could cause the calculated expected value for each row in Table 5.5 to differ from the expected value specified by μ. However, Figure 5.7 suggests this problem is noticeable only for the values towards the edge of the distribution, which in practice occur very rarely. It should be noted that, for any given value of μ, of the matches whose expected mean estimated using the basic model was close to μ, the observed average score differs slightly from μ. This can be observed by recalling Figure 5.2. It follows that before smoothing is applied to the columns in Table 5.5, the rows do not have the property that $\sum_{x=0}^{\infty} xP(X = x|\mu) \neq \mu$.

Figure 5.8 displays the same plots as Figure 5.3, for three different mean scoring rates, with the NFL distribution probabilities overlaid in green. The probabilities from the NFL distribution bear a closer resemblance to the observed histograms than the probabilities derived assuming a Normal distribution.

5.4.1 Implementation of the NFL distribution

Although a correctly specified distribution is necessary in order to optimise the accuracy of predicted probabilities for future matches, it is not necessary in order to obtain consistent estimates for the parameters. In this case these are the teams' offensive

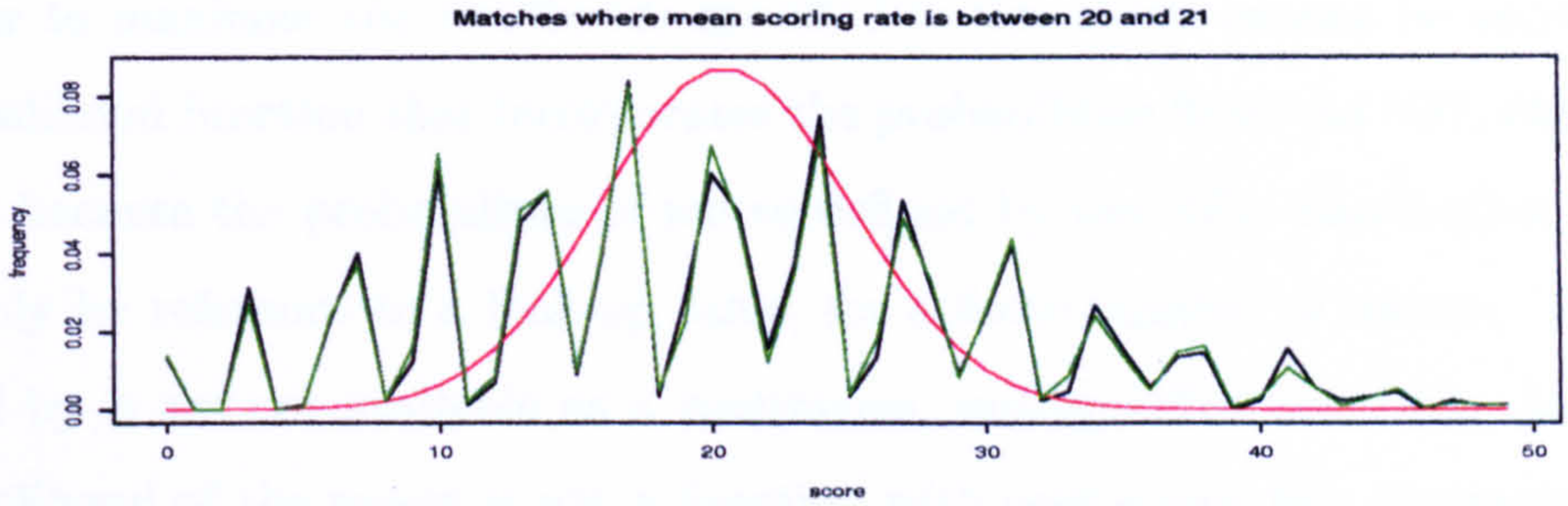


Figure 5.8: Plots of observed histograms of score frequencies (___), theoretical frequencies obtained assuming normal distribution applies (___), and also the computed NFL distribution (___), for three sets of means

and defensive abilities, as well as the global mean, home effect and score variance. In general, if the form specified for the density is incorrect but the conditional mean of the data generating process is specified correctly (that is, the functional form and explanatory variables are the same as those of the true data generating process), in certain situations it is possible to obtain asymptotically consistent estimates. In particular this is true if the assumed density is a member of the exponential family, which the Normal distribution is. However, the fact that asymptotically consistent estimates can be obtained only guarantees that as the amount of data available becomes infinite, the estimates of the parameters converge to the 'true' values. However, the rate at which they converge to them increases the closer the assumed density is to the true underlying data generating process.

Hence ideally MLEs for the parameters would be obtained using the NFL distribution. Unfortunately the numerical routines such as Newton-Raphson that are employed

in order to maximise the likelihoods specified in this thesis cannot be applied easily to a likelihood function that incorporates the probabilities from the NFL distribution. This is because the probabilities of scores defined by the NFL distribution are available only by reference to a look-up table, for a finite number of means. The values defined by it are not available as a continuous, well-specified equation. As a result, the likelihood of the scores is not a function with continuous first derivatives, which is an essential criterion in order to use the numerical maximisation routines employed throughout this thesis. For this reason the MLEs estimated from the basic model are used in the remainder of this section. The NFL distribution is employed in order to predict the probabilities of match outcomes.

While having the conditional mean correctly specified leads to asymptotically consistent estimates (despite an incorrectly specified probability density for the data), the fact that the variance is mis-specified means that the standard errors obtained for MLEs are invalid. As a result, valid t-statistics or confidence intervals for the parameters cannot be obtained using the basic model. That would be a problem if a selection of different models using a range of different factors were being fitted and the significance level of these factors were being investigated. In this application, the parameter estimates are used only to produce an expected mean as in Equation 5.3.3. Therefore, for this application the standard errors for the parameters are not required. A more detailed discussion on this topic can be found in Cameron and Trivedi (1998) pp27-31.

5.4.2 Model evaluation

It is of interest to see how the two models above perform relative to the bookmaker's line. One betting strategy is to place a bet on a match provided that, according to the model, the probability of winning is greater than a cut-off value k, for example 0.55. The success rate of such a strategy, for varying values of k, is displayed in Figure 5.9 for both the basic model and the NFL distribution model. Also included is a y=x line, which represents the curve that would be realised with a theoretically optimal model, where bets are placed knowing the true probability that the bet is successful. Overall, the plot is not conclusive, but it appears that for the majority of sensible candidate values for k, the bets made using the NFL distribution slightly out-perform those made using the basic model. Both models seem to perform quite respectably compared to the bookmaker's line. Note that only the proportion of bets won, rather than profit,

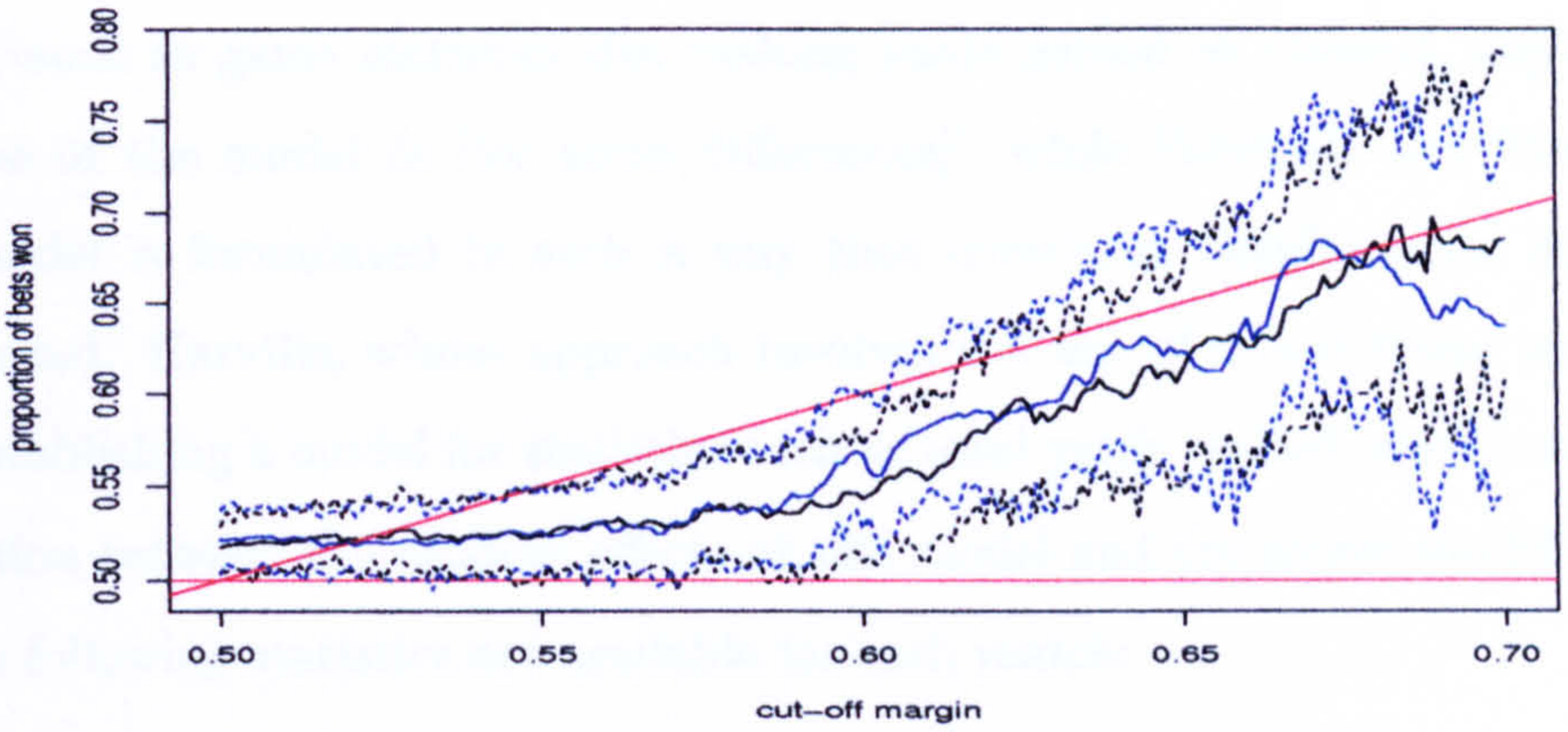


Figure 5.9: Proportions of bets won, where a bet is made provided P(Win)≥cut-off, according to both the basic model (___) and the NFL distribution (___)

is plotted in Figure 5.9, hence the bookmaker's overround is not considered.

Generally most bookmakers return $\frac{1}{1.1}$ times the original stake on a winning bet of the type described above, as explained in Section 1.3.1. So the expected profit for the gambler by betting a unit stake on a result with outcome q is

$$\frac{q}{1.1} - (1 - q) \tag{5.4.1}$$

This is positive if $q > 0.524$, although the rate at which profit is made is too slow for most gamblers unless the success rate is considerably higher than this. The profit curve of Figure 5.9 appears to win approximately 55% of the time if the estimated probability of success is above 60%.

5.5 Inclusion of more covariates

As mentioned in Section 5.2, the second data set available includes only four years of data, but there is more data available for each match. While ultimately it is only the distribution of the home and away score that is of primary interest, it is conceivable that the marginal distribution for the home and away score, derived from a joint distribution of many match variables, may be more accurate than the basic model which includes no information beyond the score of the match.

While none of the previous studies of NFL scoring rates that the writer is aware of include any information besides the scores and identities of the teams in the model specification, several papers suggest that some benefit may be derived by including

other match statistics. Glickman and Stern (1998) state that 'use of covariate information, such as game statistics like rushing yards gained or allowed, might improve precision of the model fit [for score differences]', while Fahrmeir and Tutz's (1994) NFL model is formulated in such a way that covariates besides team abilities can be included. Harville, whose approach involves the use of mixed linear models, suggests establishing a model for statistics such as total yards rushed, and monitoring the correlation between the random effects of this model and the scores model.

The following statistics are available for each match:

- the number of Touch Downs (HTD, ATD), Field Goals (HFG, AFG), 1-Point Conversions (H1C, A2C), 2-Point Conversions (H2C, A2C) and Defensive Conversions (HDC, ADC) scored in each match by the home and away side.

- Yards rushed (HRYD, ARYD) and yards passed (HPYD, APYD). It is crucial that a team moves the ball towards the opponent's end, firstly in order to increase their chances of scoring points, and secondly because they are forced by the regulations to surrender possession of the ball if they do not advance the ball more than 10 yards every 4 plays (this is explained in more detail in Section 5.5.1). The ball can be advanced towards the opponent's goal either by running with the ball or by successfully passing to another player.

- The number of rushes (HR, AR), the number of attempted passes (HPA, APA) and the number of completed passes (HPC, APC).

- The number of pass interceptions (HPINT, APINT). Should a player from a defensive team catch the ball while the offensive team is in possession of the ball, his side assumes possession of the ball.

The mean values of these figures are displayed in Table 5.6 in order to demonstrate the approximate scale of each figure.

Table 5.6: Mean values for figures in data set, 1997-2001

	Home	*Away*
Rush/Pass Attempts	61.13	60.21
Rushes	28.51	27.07
Yards Rushed	114.58	106.66
Pass Attempts	32.62	33.14
Completed Passes	18.76	18.66
Yards Passed	209.68	202.91
Had Intercepted	1	1.14

To create a joint distribution involving the covariates listed above, a set of marginal and conditional distributions of the covariates must be established. There are a large number of configurations for this, but the approach taken here reflects the approximate pattern by which NFL play proceeds.

5.5.1 NFL pattern of play

Play effectively starts with a *scrimmage*, which is similar to the *scrum* in rugby and involves a set of players from either side forming two lines standing opposite each other. In NFL it is the offensive team that always begins with possession of the ball, and the ball is almost always passed by the offensive players in the scrimmage to the quarterback, who stands behind the scrimmage, protected from the opposing team's defensive players by his own offensive players. The quarterback most often attempts to pass the ball on to another player. This action is counted as a Pass Attempt. If this pass is successful, the player receiving the ball either tries to run with the ball, which is recorded as a Rush, or very occasionally pass it once more to another player (only backwards passes are permitted in this case), which is recorded as a Pass Attempt. This initial activity, which represents the start of any attack, is summarised in the data set by the number of Rushes or Pass Attempts.

The first dependent variable is the decision the team makes concerning whether to Rush or make a Pass Attempt. Now the procedure that follows a Rush or Pass Attempt is considered.

A Rush almost always concludes with the player with the ball being impeded by the opposition either by being thrown to the ground, or forced to run out of the field of play. The action in the game stops and another scrimmage takes place from the place where the rushing player was halted, provided play from the last four scrimmages have resulted in the offensive team advancing at least ten yards towards the opponent's end of the field. If this is not the case, the offensive team loses possession of the ball to the defensive team, and all players on the field are substituted appropriately, as explained in Section 5.1.2. In the case of a Pass Attempt, three things can occur. Firstly, the player may catch the ball and continue to attack. Hence the number of Passes Completed as a proportion of the number of Passes Attempted is the next dependent variable. The other two situations occur if the Pass is unsuccessful. Normally the ball is not caught completely by either side in which case a scrimmage takes place from the point where the Pass Attempt was started, again provided that play from the

previous four scrimmage has resulted in a gain of at least ten yards by the offensive team. However occasionally (average 1.00 by the home side and 1.14 times by the away side in each match) a player from the defensive side catches the ball. In this case this player's side gains possession of the ball and becomes the offensive side. The number of Pass Interceptions is therefore the next dependent variable.

The next two variables incorporate two of the previous variables, namely the number of Yards Rushed as a result of the number of Rushes made, and the number of Yards Passed as a result of the number of Passes Completed.

Finally, the total number of Touch Downs and Field Goals that result from all the action described above is modelled.

Figure 5.5.1 is a diagrammatic representation of the conditional structure outlined above. The conditioning progresses from left to right then from top to bottom. Hence the maximal model for Away Rush/Pass Attempts (ARPA) can only have Home Rush/Pass Attempts (HRPA) as a covariate, while the model for Away Field Goals (AFG) features potentially all of the other variables.

Finally points can also be scored through 1- or 2-Point Conversions following a Touch Down, and through Defensive Conversions. The rates at which these are achieved are approximately equal for all teams in all matches so, unlike the distributions described above, do not require extensive treatment. The procedure adopted for these variables is discussed later.

5.5.2 Response distributions

The individual probability distributions involved are now discussed.

Combined Rush and Pass attempts

Histograms of home and away rush and pass attempts are displayed in Figure 5.11. They appear to be Normally distributed and their (home,away) correlation coefficient is -0.527. The bivariate Normal distribution seems to be an appropriate distribution.

Pass Attempts and Completed Passes

The most obvious way to model these is using binary logistic regression, using respectively HRPA, ARPA, HPA and APA as the group size, and the covariates being selected in accordance with Figure 5.5.1.

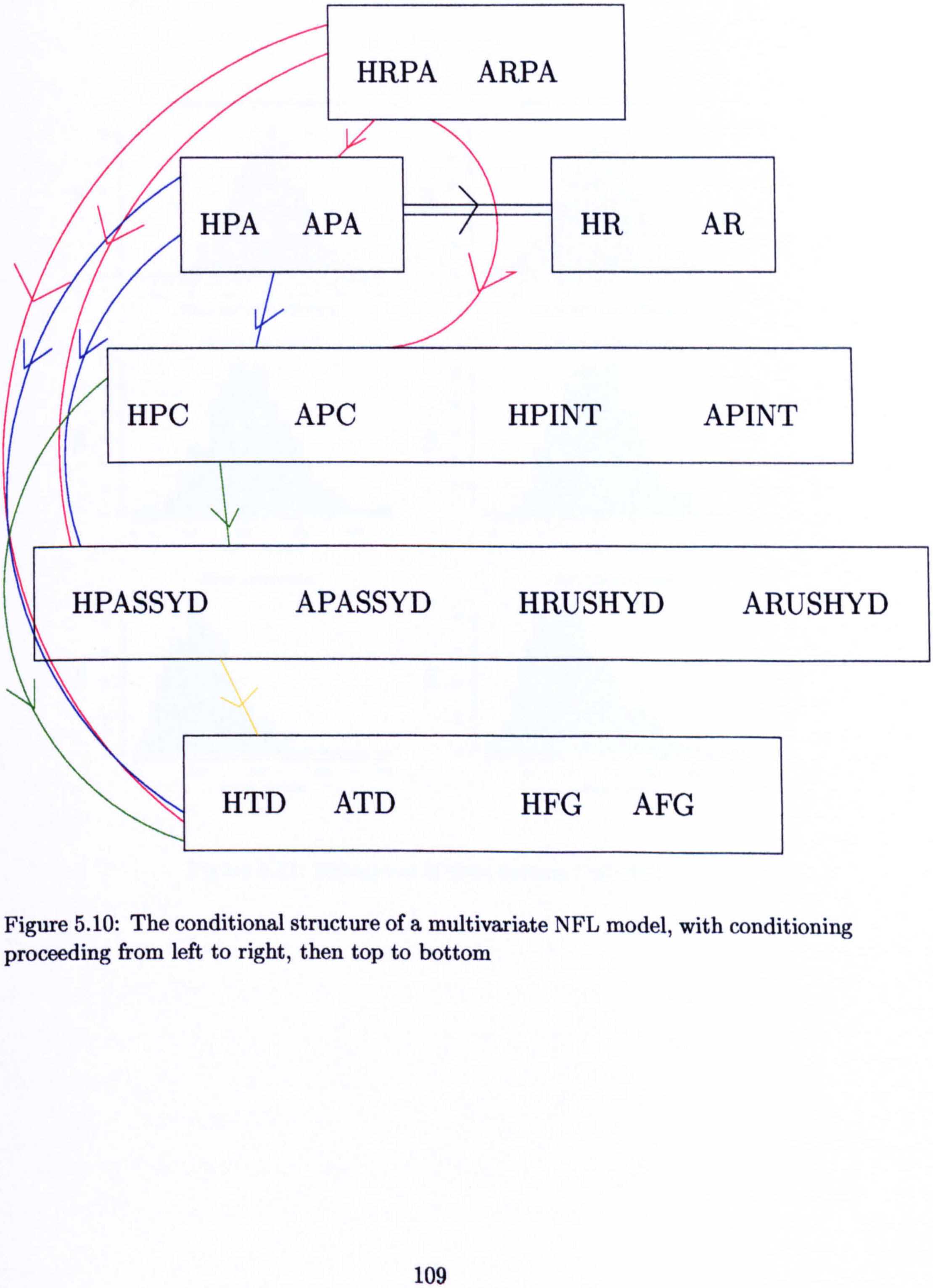


Figure 5.10: The conditional structure of a multivariate NFL model, with conditioning proceeding from left to right, then top to bottom

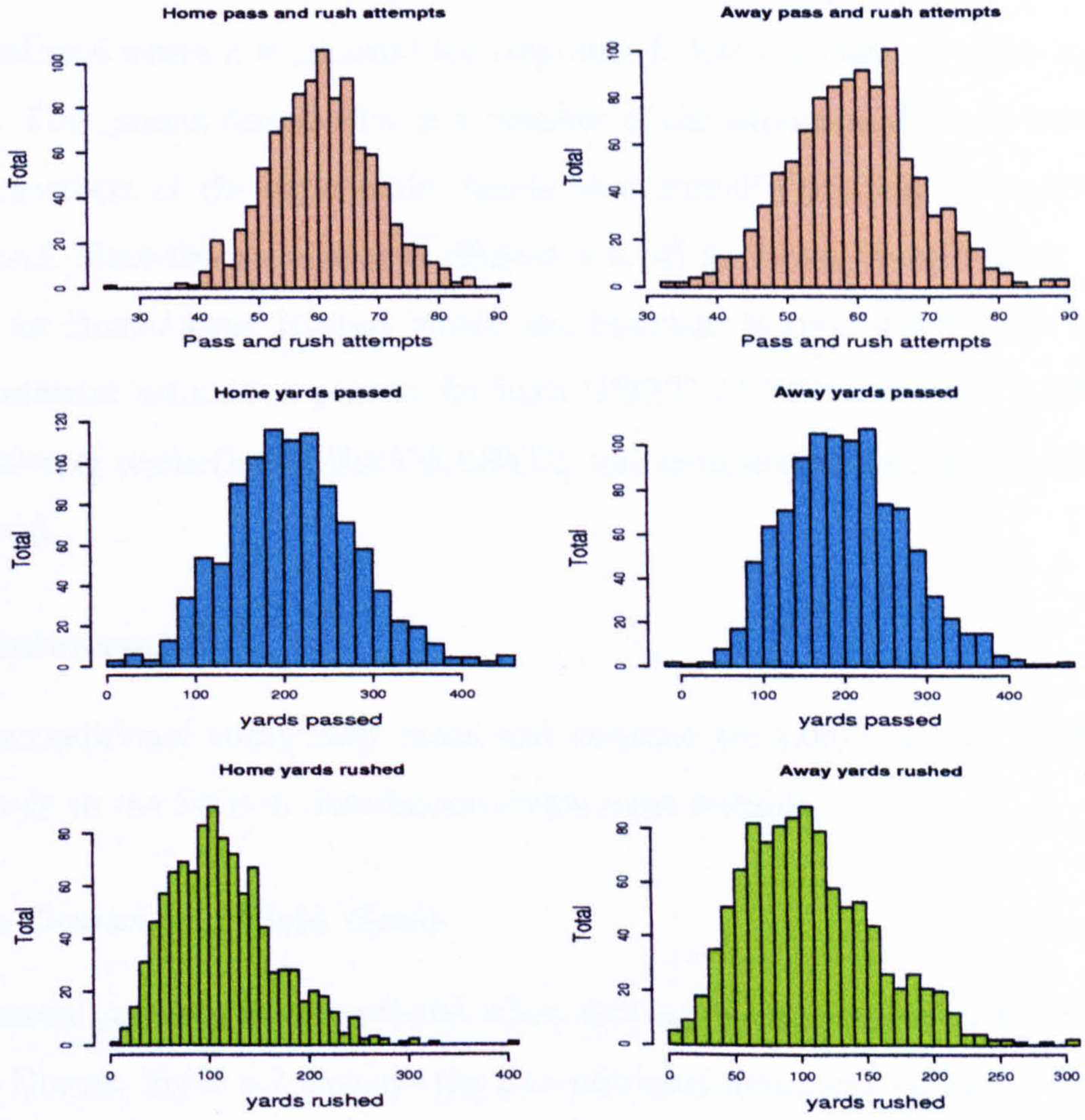


Figure 5.11: Histograms of data, seasons 1997-2001

Total Passed Yards and Rushed Yards

Figure 5.11 displays histograms for the home and away total passed and rushed yards. While the passed yards seem to follow Normal distributions, the rushed yards are significantly positively skewed. The most obvious alternative distribution is the gamma distribution. Unfortunately attempting to maximise the likelihood of data that is assumed to be gamma distributed is very time consuming, However, as stated in Section 5.4, asymptotically consistent estimates of parameters can be obtained, provided that the conditional mean of a data generating process is correctly specified, by maximising the likelihood where it is assumed the responses follow any member of the exponential family. The gamma distribution is a member of the exponential family but there are other members of the exponential family that simplify the task of maximising the likelihood. Since the correlation coefficient is 0.147 for Home/Away Passed Yards and -0.290 for Home/Away Rushed Yards, the bivariate Normal distribution is used for the parameter estimation process for both (HRYD,ARYD) and (HPYD,APYD). For simulation or prediction of (HRYD,ARYD), two independent gamma distributions are employed.

Pass Interceptions

The unconditional home/away mean and variance are 1.00/1.14 and 1.152/1.120 respectively so the Poisson distribution seems most suitable.

Touch Downs and Field Goals

An unusual problem is encountered when looking at the number of Field Goals and Touch Downs. Table 5.7 displays the unconditional mean and variances of these variables, which suggests that the assumption of equality of the mean and variance required when using the Poisson distribution is violated. Note that the Poisson condition is that the mean is equal to the *conditional* variance, conditional on any relevant covariates. Thus the under-dispersion of these variables is more severe than recorded in Table 5.7 since the conditional variance, given the team parameters and the other covariates, is less than or equal to the unconditional variance. For the ultimate application of this problem, it is necessary to calculate probabilities such as $P(HSC > k)$. Since $HSC = 3 * HFG + 6 * HTD$ (suppressing 1- and 2-Point Conversions and Defensive Conversions for now), if the Poisson distribution is employed to model Touch Downs and Field Goals, scores further from the mean have their probability of occurrence

Table 5.7: Touch Down and Field Goals means and variances 1997-2001

	Touch Downs	*Field Goals*
Home mean	2.539	1.524
Home variance	2.184	1.365
Away mean	2.132	1.405
Away variance	1.942	1.35

over-estimated, and scores closer to the mean have their probability underestimated. While none of the well-known statistical distributions is suitable for modelling under-dispersed count data such as this[4], Efron's Double Poisson distribution (1986) can be used in this situation.

Efron's Double Poisson distribution arises as an exponential combination of two Poisson distributions, $Pois(\mu)$ and $Pois(y)$, hence

$$f(y,\mu,\phi) = K(\mu,\phi)Pois(\mu)^{\phi}Pois(y)^{1-\phi}$$

where ϕ represents a dispersion parameter and $K(\mu,\phi)$ is a normalising constant. The expansion of this expression is:

$$f(y,\mu,\phi) = K(\mu,\phi)\phi^{1/2}e^{-\phi\mu-y}\frac{y^y}{y!}(\frac{e\mu}{y})^{\phi y} \qquad (5.5.1)$$

where

$$\frac{1}{K(\mu,\phi)} \doteq 1+\frac{1-\phi}{12\phi\mu}(1+\frac{1}{\phi\mu})$$

The mean and variance of the distribution are approximately μ and $\frac{\mu}{\phi}$. Although originally proposed as a means of modelling over-dispersed count data, it is also suitable for under-dispersed data. The normalising constant is approximately 1, and can be suppressed for maximum likelihood estimation. Since the first order maximisation conditions are the same as those for maximum likelihood estimation of Poisson distributed data, the MLEs obtained using either approach are the same. However, the standard errors decrease, in the case of under-dispersed data, which has two effects on the application in question. Firstly, inferences obtained via p-values are affected and secondly, the variances of the parameters change which, if the parameters are considered from a Bayesian point of view, affects the variance of the predictive distributions.

In figure 5.12, the density of 3*HFG+6*HTD is plotted in blue assuming a Poisson

[4]the negative binomial distribution is suitable for modelling overdispersed count data

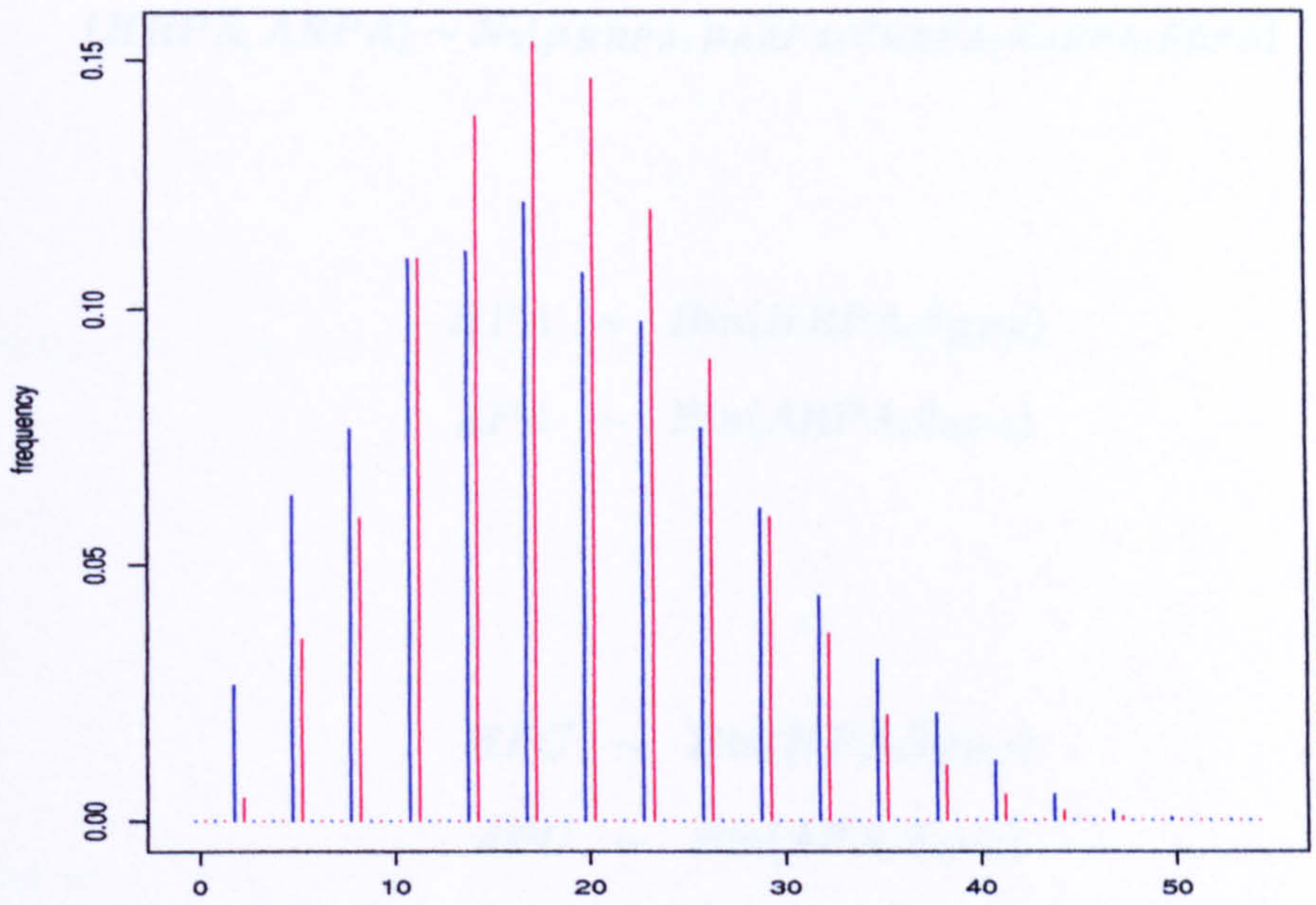


Figure 5.12: Density of 3*FG+6*TD, assuming FG and TD are Poisson Distributed (___) and Efron distributed (___)

distribution for both HFG and HTD with the rate parameters set respectively to be the overall means of the home Field Goals and Touch Downs. In red, the density is plotted assuming Efron's Double Poisson distribution, with rate parameters as before, but with dispersion parameter defined as the respective mean/variance ratios. It can be observed that the probability of many low scores is far lower assuming Efron's Double Poisson distribution.

Since the MLEs obtained by using Efron's Double Poisson distribution to model the response variable are the same as those obtained using a standard Poisson distribution, for parameter estimation purposes it is more convenient to employ the Poisson distribution in the likelihood function. To generate predictions or to simulate outcomes, Efron's Double Poisson distribution is employed using the parameter estimates obtained with the Poisson distribution. In order to do this, an estimate for the dispersion parameter ϕ is required and the ratio of the mean to the conditional variance of the variable is a suitable choice. After the final time-point, this was found to be 1.280 for Touch Downs and 1.185 for Field Goals[5].

To summarise, the following distributions are employed:

[5]Note that when employing the Double Poisson distribution, a value for the dispersion parameter lower than 1 corresponds to over-dispersion, a value greater than 1 corresponds to under-dispersion.

- $$(HRPA, ARPA) \sim \mathcal{N}_2(\mu_{HRPA}, \mu_{ARPA}, \sigma_{HRPA}, \sigma_{ARPA}, \rho_{RPA}) \qquad (5.5.2)$$

- $$\begin{aligned} HPA &\sim Bin(HRPA, \theta_{HPA}) \\ APA &\sim Bin(ARPA, \theta_{APA}) \end{aligned} \qquad (5.5.3)$$

- $$\begin{aligned} HPC &\sim Bin(HPA, \theta_{HPC}) \\ APC &\sim Bin(APA, \theta_{APC}) \end{aligned} \qquad (5.5.4)$$

- $$(HPYD, APYD) \sim \mathcal{N}_2(\mu_{HPYD}, \mu_{APYD}, \sigma_{HPYD}, \sigma_{APYD}, \rho_{PYD}) \qquad (5.5.5)$$

- $$(HRYD, ARYD) \sim \mathcal{N}_2(\mu_{HRYD}, \mu_{ARYD}, \sigma_{HRYD}, \sigma_{ARYD}, \rho_{RYD}) \qquad (5.5.6)$$

For prediction,

- $$\begin{aligned} HRYD &\sim Gamma(\alpha_{HRYD}, \lambda_{HRYD}) \\ ARYD &\sim Gamma(\alpha_{ARYD}, \lambda_{ARYD}) \end{aligned}$$

- $$\begin{aligned} HPINT &\sim Pois(\lambda_{HPINT}) \\ APINT &\sim Pois(\lambda_{APINT}) \end{aligned} \qquad (5.5.7)$$

- $$\begin{aligned} HTD &\sim Pois(\lambda_{HTD}) \\ ATD &\sim Pois(\lambda_{ATD}) \end{aligned} \qquad (5.5.8)$$

For prediction,

$$\begin{aligned} HTD &\sim Pois^2(\lambda_{HTD}, \phi_{HTD}) \\ ATD &\sim Pois^2(\lambda_{ATD}, \phi_{ATD}) \end{aligned}$$

where Pois^2 denotes Efron's Double Poisson distribution,

•

$$\begin{aligned} HFG &\sim Pois(\lambda_{HFG}) \\ AFG &\sim Pois(\lambda_{AFG}) \end{aligned} \tag{5.5.9}$$

For prediction,

$$\begin{aligned} HFG &\sim Pois^2(\lambda_{HFG}, \phi_{HFG}) \\ AFG &\sim Pois^2(\lambda_{AFG}, \phi_{AFG}) \end{aligned}$$

The form of the mean terms such as λ_{HFG} or μ_{ARPA} has not yet been specified. This is the topic of the next section.

5.5.3 Selection of covariates for specific models

Ideally, by studying the past relationships between the covariates, a reliable framework which produces predictions for future events could be obtained. With the large number of covariates, there is a danger of over-fitting. That is, having a set of variables that explain past data very precisely, but by modelling the random error rather than the underlying relationships. Hence the predictive power may well be disappointing. This problem is illustrated in Tables 5.8 and 5.9.

Table 5.8 displays the results of fitting a model for the number of Home Yards Rushed using various covariates, however the model is fit separately for each season of the data. The coefficient for the number of Home Rush/Pass Attempts (HRPA) is highly significant for seasons 2 and 3, but the size of the coefficient, and hence statistical significance is far lower in seasons 1 and 4. The coefficients for Away Home Rush/Pass Attempts (ARPA) and Away Pass Interceptions (APINT) display a similar problem for different seasons. Table 5.9 displays the results of performing binary logistic regression on the proportion of Home Pass Attempts (HPA) that result in

Table 5.8: Coefficients and values for Home Rushed Yards model, using various covariates, regressed over each season individually

Covariate	*Season*			
(coef,p-value)	*1*	*2*	*3*	*4*
HR	**(5.511,0)**	**(5.852,0)**	**(5.189,0)**	**(5.533,0)**
AR	(-0.615,0.18)	(-0.272,0.505)	(-0.059,0.873)	(0.345,0.38)
HRPA	(0.406,0.185)	**(1.408,0)**	**(0.846,0.009)**	(0.551,0.112)
ARPA	**(-0.979,0.004)**	(-0.22,0.484)	(0.071,0.826)	(-0.226,0.527)
HPINT	(4.434,0.07)	(1.226,0.588)	(1.863,0.351)	(0.954,0.691)
APINT	**(-6.802,0.002)**	**(-8.789,0)**	(-3.387,0.087)	(-3.875,0.096)

Table 5.9: Coefficients and values for Home Pass Conversion ratio model, using various covariates, regressed over each season individually

Covariate	*Season*			
(coef,p-value)	*1*	*2*	*3*	*4*
HRPA	(-0.004,0.328)	(-0.008,0.118)	**(-0.013,0.004)**	**(-0.013,0.006)**
ARPA	**(-0.02,0)**	**(-0.023,0)**	**(-0.028,0)**	**(-0.027,0)**
HPA	(0.002,0.735)	(-0.002,0.649)	**(0.009,0.033)**	**(0.012,0.005)**
APA	(0.006,0.169)	**(0.017,0)**	**(0.022,0)**	**(0.019,0)**
HPINT	**(-0.068,0.005)**	(-0.022,0.318)	**(-0.074,0)**	**(-0.085,0)**
APINT	(-0.006,0.814)	**(-0.051,0.018)**	(-0.003,0.903)	(-0.039,0.084)

a Completed Home Pass (HPC) using a number of covariates, where the regression is again performed separately for each season of the data. As in Table 5.8, some of the covariates' significance varies drastically from season to season. One possible cause is the correlation between some of the covariates. Techniques such as Principal Component Analysis could be considered in this situation. The approach taken here is to select only the most essential set of covariates for each model, although with further research, more suitable models may be discovered.

A procedure to discover the essential covariates

To choose the most appropriate set of covariates for each model, the following procedure is used:

1. For each of the models specified by Equations 5.5.3 to 5.5.9, obtain team parameter estimates for the models, with no covariates involved in the conditional

mean. For example, the conditional mean for the models specified by Equations 5.5.3 and 5.5.4 are defined to be, for match k between sides $i(k)$ and $j(k)$,

$$\begin{aligned} logit(E[X_k]) &= \mu + \alpha_{i(k)} + \beta_{j(k)} + \delta \\ logit(E[Y_k]) &= \mu + \alpha_{j(k)} + \beta_{i(k)} \end{aligned} \tag{5.5.10}$$

where

- X_k and Y_k are set to be the home and away response variables in Equations 5.5.3 to 5.5.9. The distributions for X_k and Y_k are as stated in Equations 5.5.3 to 5.5.9.
- the α and β parameters are the teams' offensive and defensive capabilities with respect to the relevant response variable.
- μ and δ are the global mean and effect of playing at home, with respect to the response variable.

Note that for the other models, the specification of the conditional means such as those listed in Equation 5.5.10 need to modified appropriately depending on whether the data is considered to binomially distributed, Normally distributed or Poisson distributed in Equations 5.5.3 to 5.5.9.

The loglikelihoods are maximised using the MLE procedure that has been applied to other models of this type in this thesis and is explained in Chapter 3. Note that it is appropriate to exponentiate the right hand side of the means when Equations 5.5.3, 5.5.4, 5.5.7, 5.5.8 and 5.5.9 are being fitted. This is because the proportion parameter in the binomial distribution, and the rate parameter in the Poisson distribution, are necessarily greater than zero.

2. Using the results of step 1, create team effects for each of the models featured in Equations 5.5.3 to 5.5.9. For each match this is $\hat{\alpha}_{i(k)} + \hat{\beta}_{j(k)}$ for the model for the home response and $\hat{\alpha}_{j(k)} + \hat{\beta}_{i(k)}$ for the model for the away response.

3. For each of the models specified by Equations 5.5.3 to 5.5.9, fit a generalised linear model including, in the conditional mean, both the team effects from stage 2 and the available covariates (obtained by reference to Figure 5.5.1). These models are all fitted separately for the data of each season. The consistency of the estimates from one season to the next is examined. Upon consideration of the results of

this, the covariates are chosen with the aim of including all necessary information while minimising the risk of over-fitting by modelling random error.

The logic behind this procedure is that it is conceivable that the problems involved in selecting which covariates are important that is observed in Tables 5.8 and 5.9 may be reduced by modelling more of the variance where possible. This can be achieved by including team parameters, in addition to the covariates, in the specifications of the means of the models listed in Equations 5.5.3 to 5.5.9.

When a variable is significant for some seasons but not others, the decision whether to include it is a subjective one based on various factors. Firstly, how close it is to statistical significance in each season is considered. If it is the model for a home response about which there is uncertainty, then the corresponding model for the away response is examined (and vice versa).

Results of covariate selection procedure

Based on this analysis, Table 5.10 displays which covariates have been selected for the final models.

- There is a general symmetry between the home and away categories, which is logical, although the partly subjective selection of covariates was made using this as a criterion. Total symmetry is not expected since NFL teams, as in many other sports, vary tactics according to whether the match is being played at home or away.

- The sufficiency of Completed Passes towards predicting Passed Yards, and the the sufficiency of Rushes towards predicting Rushed Yards, is unsurprising.

- The importance of information concerning Touch Downs towards the prediction of Field Goals is logical, since usually when a team is faced with a potential scoring opportunity, it has to decide between trying to obtain a Touch Down for 6 points or settling for a Field Goal for 3 points. Thus a larger number of Touch Downs than expected is likely to lead to a smaller number of Field Goals than expected and vice versa.

- One parameter to represent both the effect of a home covariate, such as HRPA or HPA, on the home result and the effect of an away covariate, such as ARPA or APA, on the away result is generally considered appropriate.

Table 5.10: Final model for each covariate

rush and pass attempts

$$HRPA_k \sim \gamma + \delta + \alpha_{i(k)} + \beta_{j(k)}$$
$$ARPA_k \sim \gamma + \alpha_{j(k)} + \beta_{i(k)} + \lambda_{11a} HRPA_k$$

proportion of rush and pass attempts that result in a rush

$$\frac{HPA_k}{HRPA_k} \sim \gamma + \delta + \alpha_{i(k)} + \beta_{j(k)} + \lambda_{21} HRPA_k + \lambda_{22} ARPA_k$$
$$\frac{APA_k}{ARPA_k} \sim \gamma + \alpha_{j(k)} + \beta_{i(k)} + \lambda_{21} ARPA_k + \lambda_{22} HRPA_k + \lambda_{23a} HPA_k$$

pass interceptions

$$HPINT_k \sim \gamma + \delta + \alpha_{i(k)} + \beta_{j(k)} + \lambda_{31} HPA_k + \lambda_{32} APA_k + \lambda_{33h} HRPA_k$$
$$APINT_k \sim \gamma + \alpha_{j(k)} + \beta_{i(k)} + \lambda_{31} APA_k + \lambda_{32} HPA_k$$

pass completion ratio

$$\frac{HPC_k}{HPA_k} \sim \gamma + \delta + \alpha_{i(k)} + \beta_{j(k)} + \lambda_{41} ARPA_k + \lambda_{42} APA_k + \lambda_{43h} HRPA_k$$
$$\frac{APC_k}{APA_k} \sim \gamma + \alpha_{j(k)} + \beta_{i(k)} + \lambda_{41} HRPA_k + \lambda_{42} HPA_k + \lambda_{43a} APINT_k$$

total yards passed

$$HPYD_k \sim \gamma + \delta + \alpha_{i(k)} + \beta_{j(k)} + \lambda_{51} HPC_k$$
$$APYD_k \sim \gamma + \alpha_{j(k)} + \beta_{i(k)} + \lambda_{51} APC_k$$

total yards rushed

$$HRYD_k \sim \gamma + \delta + \alpha_{i(k)} + \beta_{j(k)} + \lambda_{61} HR_k$$
$$ARYD_k \sim \gamma + \alpha_{j(k)} + \beta_{i(k)} + \lambda_{61} AR_k$$

touch downs

$$HTD_k \sim \gamma + \delta + \alpha_{i(k)} + \beta_{j(k)} + \lambda_{71} HRYD_k + \lambda_{72} APYD_k + \lambda_{73} HPA_k + \lambda_{74} APA_k$$
$$ATD_k \sim \gamma + \alpha_{j(k)} + \beta_{i(k)} + \lambda_{71} ARYD_k + \lambda_{72} HPYD_k + \lambda_{73} APA_k + \lambda_{74} HPA_k + \lambda_{75a} HPC_k + \lambda_{76a} HTD_k$$

field goals

$$HFG_k \sim \gamma + \delta + \alpha_{i(k)} + \beta_{j(k)} + \lambda_{81} HPINT_k + \lambda_{82} APINT_k + \lambda_{83} HTD_k + \lambda_{84} HPYD_k$$
$$AFG_k \sim \gamma + \alpha_{j(k)} + \beta_{i(k)} + \lambda_{81} APINT_k + \lambda_{82} HPINT_k + \lambda_{83} ATD_k + \lambda_{84} APYD_k + \lambda_{85a} HPA_k$$

Model fitting concerns and estimated coefficients

Having specified a model for each covariate, it is important to recall a key problem associated with the MLE method of obtaining parameter estimates of models that feature covariates beside team abilities. By placing less weight on the information from matches that took place longer ago, teams' more recent performances contribute more to the estimates of their parameters. As a result, the information from less recent matches which is helpful towards estimating the effect of factors besides team parameters is also down-weighted. These effects are not considered time dependent

so all information concerning these factors is of interest, hence this is not a desirable property, as explained in more detail in Section 3.2.3. Hence the procedure outlined in that section is employed, namely

- obtain estimates for the non-time-dependent covariates assuming no team effects,
- treating these estimated values as constants, estimate the team effects using the MLE procedure outlined in Chapter 3.

Table 5.11 displays results from the first stage of this process. The coefficients for the covariates obtained at a time-point halfway through the data set, and at the final time-point are displayed.

Table 5.11: Final model coefficients at final time-point

Away Rush and Pass Attempts coefficients					
HRPA: -0.626					
Home Decision to Attempt Pass coefficients					
HRPA: 0.009	ARPA: 0.021				
Away Decision to Attempt Pass coefficients					
ARPA: 0.039	HRPA: 0.016	HPA: -0.034			
Home Pass Interceptions coefficients					
HPA: 0.04	APA: -0.014	HRPA: -0.02			
Away Pass Interceptions coefficients					
APA: 0.04	HPA: -0.014				
Home Pass Completion Ratio coefficients					
ARPA: -0.006	HRPA: -0.02	APA: 0.012			
Away Pass Completion Ratio coefficients					
HRPA: -0.02	HPA: 0.012	APINT: -0.086			
Home Yards Passed coefficients					
HPC: 10.072					
Away Yards Passed coefficients					
APC: 10.072					
Home Yards Rushed coefficients					
HR: 4.838					
Away Yards Rushed coefficients					
AR: 4.838					
Home Touch Downs coefficients					
HRUSHYD: 0.004	HPASSYD: 0.005	HPA: -0.022	APA: 0.017		
Away Touch Downs coefficients					
ARUSHYD: 0.004	APASSYD: 0.005	APA: -0.022	HPA: 0.017	HPC: -0.003	HTD: 0.069
Home Field Goals coefficients					
HPINT: -0.182	APINT: 0.151	HTD: -0.19	HPASSYD: 0.003		
Away Field Goals coefficients					
APINT: -0.182	HPINT: 0.151	ATD: -0.19	APASSYD: 0.003	HPA: 0.007	

For the second stage, it is necessary to repeat the process of finding the values for the down-weighting, team prior tightness and seasonal truncation which maximise the predictive power of the model. Table 5.12 displays the optimal values for the external parameters for each model.

Table 5.12: Optimized values of external parameters for each model involved in creation of joint distribution for NFL final scores

	ς_w	$\tau_{\alpha\beta}$	ζ_s
DPA	0.1	0.1	5
PINT	0.01	0.1	10
PCR	0.05	0.1	5
PASSYD	0.01	20	2
RUSHYD	0.01	20	2
TD	0.05	0.2	5
FG	0.05	0.2	10

Now that team parameters and the coefficients are available, predictions for each of the variables being modelled can be generated. The next stage is to examine how closely the joint distribution implied by the predictions obtained throughout the modelling process resembles the joint distribution of the observed data.

5.5.4 Posterior analysis of the modelling process

In Section 2.5.1 several possible techniques in order to evaluate the suitability of a model are suggested. One of these, *the discrepancy measure* technique, is used here. To summarise the technique, samples are generated using the specified distribution and the parameter estimates. A scalar summary statistic of the data is calculated for the observed data set and for each of the simulated samples. The value of these statistics for the simulated samples is compared to the value of the statistic for the observed data. Recall from Section 2.5.1 that the discrepancy measure technique can be applied within either a Bayesian framework or a classical framework. The Bayesian framework is more appropriate if the entire distribution of the model parameters is of interest. In this application, the maximum likelihood estimates of each parameter are employed in order to produce predictions for the covariates in the conditional mean of each model, but the distribution of these parameters is not considered. It is the distribution of the data that is of primary interest, hence the classical form of discrepancy measure test is employed. The discrepancy measures used in order to diagnose any problems with the predictive distribution of the statistics are the mean and variance of each of the variables featured in the model.

Table 5.13 displays the observed values of the measures, along with the (2.5%, 97.5%) quantiles obtained through 1000 simulations of the same statistics using the final models obtained in Section 5.5.3. Only data from the 1999/00 and 2000/01 seasons, which constitutes the second half of the data set, is included so that sufficient

data has been observed in order to make valid predictions. If the observed value for any of the statisticsis not contained in the (2.5%, 97.5%) quantiles of the simulated data, that suggests a model deficiency. The results are now considered.

Table 5.13: Statistics for observed values of NFL variables, with confidence intervals of simulated values in brackets

Variable	Mean	Variance
HRPA	60.82 (60.39,61.89)	73.76 (69.79,88.74)
ARPA	60.06 (59.42,60.99)	84.41 (74.16,95.4)
HPA	32.96 (31.88,32.94)	65.89 (38.92,49.26)
APA	33.34 (33.23,34.32)	72.59 (41.59,52.61)
HR	27.86 (28.22,29.29)	71.17 (39.04,49.34)
AR	26.71 (25.91,26.98)	72.2 (39.86,50.59)
HPC	19.07 (18.3,19.06)	29.62 (19.21,24.61)
APC	19.04 (18.69,19.54)	31.24 (22.48,28.91)
HPINT	1.07 (0.86,1.03)	1.23 (0.84,1.15)
APINT	1.14 (1.06,1.26)	1.22 (1.08,1.49)
HPYD	213.44 (202.46,213.35)	5600.37 (4050.25,5190.09)
APYD	203.74 (202.23,213.75)	6014.05 (4538.7,5745.8)
HRYD	111.59 (111.03,118.93)	2778.79 (1972.28,2467.92)
ARYD	105.3 (100.24,109.98)	2541.86 (1760.42,2230.69)
HTD	2.5 (2.39,2.65)	2.31 (2.17,2.9)
ATD	2.12 (2.03,2.27)	2.01 (1.85,2.48)
HFG	1.54 (1.5,1.74)	1.43 (1.74,2.58)
AFG	1.4 (1.38,1.61)	1.35 (1.54,2.17)

The variance of the simulated HPA is far lower than that of the observed data. Recall that HPA is modelled by estimating $\frac{HPA}{HRPA}$ within a binary logistic regression framework, which assumes a binomial distribution where the group size is $HRPA$. The binomial distribution only has one parameter and may not be flexible enough to represent the process by which the data is generated in practice. One distribution that may be more suitable is the beta-binomial distribution, which has two shape parameters. Its density is:

$$P(X = x|N, a, b) = \binom{N}{x} \frac{\Gamma(a+b)\Gamma(a+x)\Gamma(b+N-x)}{\Gamma(a)\Gamma(b)\Gamma(a+b+N)}$$

The beta-binomial distribution is often used to model count data for which the variance is greater than the mean (and is thus *overdispersed*). By employing the beta-binomial density, both parameters a and b could be estimated in such a way that both the mean and variance of the simulated samples match that of the observed data more closely. The complicated nature of its density suggests that maximisation of the likelihood of such a model, including time-downweighted team effects, would be complicated and could be considered as an extension to this work.

These comments also apply to the number of Away Pass Attempts and the number of Home and Away Passes Completed, which were assumed to be binomially distributed with the number of Home and Away Pass Attempts as the group sizes respectively.

According to Table 5.13, the predictions for Pass Interceptions seem satisfactory for away teams, but not for home teams. The simulated values of these statistics are however computed using the previously simulated values of the number of Pass/Rush Attempts and Pass Attempts, which are known to be inaccurate. An adjustment to the simulation procedure that may test more accurately the reliability of the Pass Interceptions model is to reproduce the adjustment applied to the process of simulating values of Pass Attempts. That is, to generate 1000 simulations based upon the observed, rather than simulated, values of Pass Attempts and Rush/Pass Attempts. This update is repeated for the remaining variables in the distribution. Table 5.14 displays the results.

Table 5.14: Observed statistics for variables along with simulated values in brackets. Values are simulated using observed values of explanatory variables

Variable	Mean	Variance
HPINT	1.07 (0.94,1.06)	1.23 (1.04,1.31)
APINT	1.14 (1.08,1.22)	1.21 (1.18,1.62)
HPYD	213.48 (207.86,215.88)	5576.57 (5042.6,6033.64)
APYD	204.48 (203.5,211.35)	6077.13 (5234.15,6377.64)
HRYD	111.73 (108.1,113.91)	2772.12 (2610,3162.87)
ARYD	105.14 (106.75,112.27)	2543.75 (2311.98,2776.07)
HTD	2.5 (2.47,2.67)	2.3 (2.13,2.77)
ATD	2.12 (2.1,2.28)	2.01 (1.68,2.24)
HFG	1.54 (1.45,1.64)	1.42 (1.22,1.63)
HFG	1.4 (1.38,1.56)	1.34 (1.17,1.55)

The majority of the observed values are within the confidence intervals obtained using simulated values of the relevant quantities. A small number of the observed statistics lie outside the the confidence intervals, however this is not unexpected given the large number of comparisons made. Overall the discrepancy measures applied here have not identified any clear deficiencies in the models for these variables. However it is possible that alternative summary measures would discover some discrepancies between the observed data and the simulated samples.

Ideally, to conclude this section of analysis, simulated values of home and away scores would be generated by using the simulated values of Touch Downs and Field Goals. Unfortunately, due to the unsatisfactory variance of simulated samples of some of the covariates involved in this modelling process it is clear that the variance of these simulated scores would not be accurate. Even if the variances of the simulated

quantities were closer to that of the observed data it would still be necessary to compare the covariance structure of the simulated samples of data with that of the observed data. In this case there is little interest in doing so since it has already been established that some of the response distributions applied are not suitable. Therefore a more straightforward approach than the one described in this section is now considered.

5.6 A quasi-multivariate model

In order to exploit some of the data from the in-match totals available, a less ambitious version of the model attempted in Section 5.5 can be considered. The following models are fitted:

$$\begin{aligned}
E[HTD_k] &= \gamma_{td} + \delta_{td} + \alpha^{td}_{i(k)} + \beta^{td}_{j(k)} \\
E[ATD_k] &= \gamma_{td} + \alpha^{td}_{j(k)} + \beta^{td}_{i(k)} + \lambda_{htd|atd} HTD_k \qquad (5.6.1) \\
E[HFG_k] &= \gamma_{fg} + \delta_{fg} + \alpha^{fg}_{i(k)} + \beta^{fg}_{j(k)} + \lambda_{hfg|htd} HTD_k + \lambda_{hfg|atd} ATD_k \qquad (5.6.2) \\
E[AFG_k] &= \gamma_{fg} + \alpha^{fg}_{j(k)} + \beta^{fg}_{i(k)} + \lambda_{afg|htd} HTD_k + \lambda_{afg|atd} ATD_k \\
&+ \lambda_{afg|hfg} HFG_k \qquad (5.6.3)
\end{aligned}$$

where

- γ_{td}, γ_{fg} are intercepts for the Touch Down and Field Goal scoring rates,
- δ_{td}, δ_{fg} are parameters representing the effect of playing at home on the Touch Down and Field Goal scoring rates,
- the $\alpha^{td}_{\cdot}$, $\alpha^{fg}_{\cdot}$ parameters are teams' abilities to score Touch Downs and Field Goals,
- the $\beta^{td}_{\cdot}$, $\beta^{fg}_{\cdot}$ parameters are the teams' abilities to prevent opponents from scoring Touch Downs or Field Goals,
- the λ are the effect of observed Touch Downs or Field Goals in a match on scoring rates.

The λ terms, which are coefficients of the HTD, ATD and AFG terms, in all the above models are all highly significant. This formulation is simple to implement and also

generates predictions for Touch Downs and Field Goals, which can then be combined to generate a probability distribution that resembles the distribution for NFL scores that was observed in Figure 5.4. Unlike the methods used in Section 5.4 there is no need to use computer-intensive techniques such as kernel smoothing.

It is necessary to select an appropriate probability distribution to model the responses in Equations 5.6.3. The approach described in Section 5.5.2 to model Touch Downs and Field Goals in the more complex multivariate model is also used here. To recap, the Poisson distribution is implemented when the likelihood is maximised and asymptotically consistent estimates are produced. However, the Touch Downs data is under-dispersed, meaning the variance of the data is lower than the mean. The same is true of the Field Goals data. As a result, a distribution that can simulate this feature of the data, such as Efron's Double Poisson (defined by Equation 5.5.1) is implemented in order to provide probabilities for future events or to simulate outcomes. The estimates of teams' abilities to score and prevent Touch Downs and Field Goals after the final time-point is displayed in Table 5.15. The lack of similarity between the rankings of teams across the four categories further supports the suggestion that the better teams do not consist of players of equal calibre throughout the squad.

Once probability distributions for HTD, ATD, HFG and AFG are obtained in this way, in order to obtain a distribution for final scores it is also necessary to consider the distribution of 1-Point, 2-Point and Defensive Conversions (discussed in Section 5.1.2). The most obvious formulation is

$$\begin{aligned}
(H1C_k + H2C_k) &\sim Binom(HTD_k, \theta_1) \\
(A1C_k + A2C_k) &\sim Binom(ATD_k, \theta_1) \\
H1C_k &\sim Binom(H1C_k + H2C_k, \theta_2) \\
A1C_k &\sim Binom(A1C_k + A2C_k, \theta_2) \\
HDC_k &\sim Pois(\theta_3) \\
ADC_k &\sim Pois(\theta_3)
\end{aligned}$$

where θ_1 represents the average proportion of Touch Downs which result in either a 1- or 2-Point Conversion, θ_2 represents the average proportion of 1- or 2-Point conversions that result in a 1-Point Conversion and θ_3 represents the average number of Defensive

Table 5.15: Offensive and defensive ability estimates of NFL teams in terms of Touch Down and Field Goal conceding rates after 28 January, 2001

Team	Touch Down offensive estimate	rank	Touch Down defensive estimate	rank	Field Goal offensive estimate	rank	Field Goal defensive estimate	rank
Arizona	-0.0926	28	0.0093	18	-0.337	29	0.3156	31
Atlanta	-0.0198	19	-0.0277	11	-0.2865	28	0.0771	22
Baltimore	0.0389	11	-0.0987	1	0.0436	12	-0.6111	1
Buffalo	0.0487	8	0.0541	26	6e-04	16	-0.0311	13
Carolina	0.0256	14	-0.0214	13	-0.1472	24	-0.014	14
Chicago	-0.0698	26	-0.0341	9	-0.3615	30	0.0757	21
Cincinnati	-0.1393	30	0.0432	22	-0.2717	27	0.1273	25
Cleveland	-0.156	31	0.0828	30	-0.4617	31	0.2446	29
Dallas	0.0375	12	-0.0178	14	-0.1356	23	0.0097	16
Denver	0.0616	6	-0.0873	2	0.2409	3	0.0572	20
Detroit	0.0394	10	0.0309	21	-0.0893	21	-0.0814	11
Green Bay	0.1009	2	0.0063	16	0.0126	15	0.037	19
Indianapolis	0.1534	1	0.0714	28	0.2242	4	-0.032	12
Jacksonville	0.1007	3	-0.0536	7	0.1606	6	0.0219	17
Kansas City	-0.0225	20	-0.0717	5	0.0951	8	0.0258	18
Miami	0.0644	4	4e-04	15	-0.1279	22	-0.225	4
Minnesota	0.0526	7	0.0517	25	0.1713	5	0.1421	27
New England	0.0464	9	-0.0311	10	-0.2006	26	-0.0035	15
New Orleans	-0.0304	22	-0.0365	8	-0.0292	19	0.0927	24
NY Giants	-0.041	23	-0.0788	4	0.0703	9	-0.1162	9
NY Jets	0.0204	16	-0.0221	12	0.0313	13	-0.1017	10
Oakland	-0.062	25	0.0759	29	0.2962	2	-0.1853	8
Philadelphia	-0.0759	27	0.059	27	-0.0492	20	-0.264	3
Pittsburgh	-0.0107	18	0.0435	23	0.0619	11	-0.1935	6
San Diego	0.021	15	0.0281	20	-0.1847	25	0.1404	26
San Francisco	-0.0263	21	0.0081	17	0.0695	10	0.1896	28
Seattle	0.026	13	0.0882	31	0.0185	14	0.0819	23
St. Louis	0.0053	17	0.0499	24	0.4339	1	0.2466	30
Tampa Bay	-0.0618	24	-0.0633	6	-0.0172	17	-0.204	5
Tennessee	0.0624	5	-0.0859	3	0.1036	7	-0.324	2
Washington	-0.1093	29	0.0151	19	-0.0229	18	-0.1866	7

Conversions by either team in a match.

All three of these distributions use a single parameter to determine the probabilities of the outcomes in all games, for both home and away teams. Almost 95% of Touch Down Conversions result in either a 1- or 2-Point Conversion and of these approximately 96% are 1-Point Conversions, while on average only 0.04 Defensive Conversions are scored in each match. Hence there is very little variance in the data to verify if any more flexible specification is appropriate.

Once all the parameters from the models described in this section have been estimated it is possible to produce probability densities for the scores of the games. Firstly it is necessary to simulate, for example, 10000 samples of each of the distributions defined above. So for each match k, values HTD_k^{*i}, ATD_k^{*i}, HFG_k^{*i}, AFG_k^{*i}, $H1C_k^{*i}$, $A1C_k^{*i}$, $H2C_k^{*i}$, $A2C_k^{*i}$, HDC_k^{*i}, ADC_k^{*i}, $i \in (1, 10000)$ are obtained. Then for each

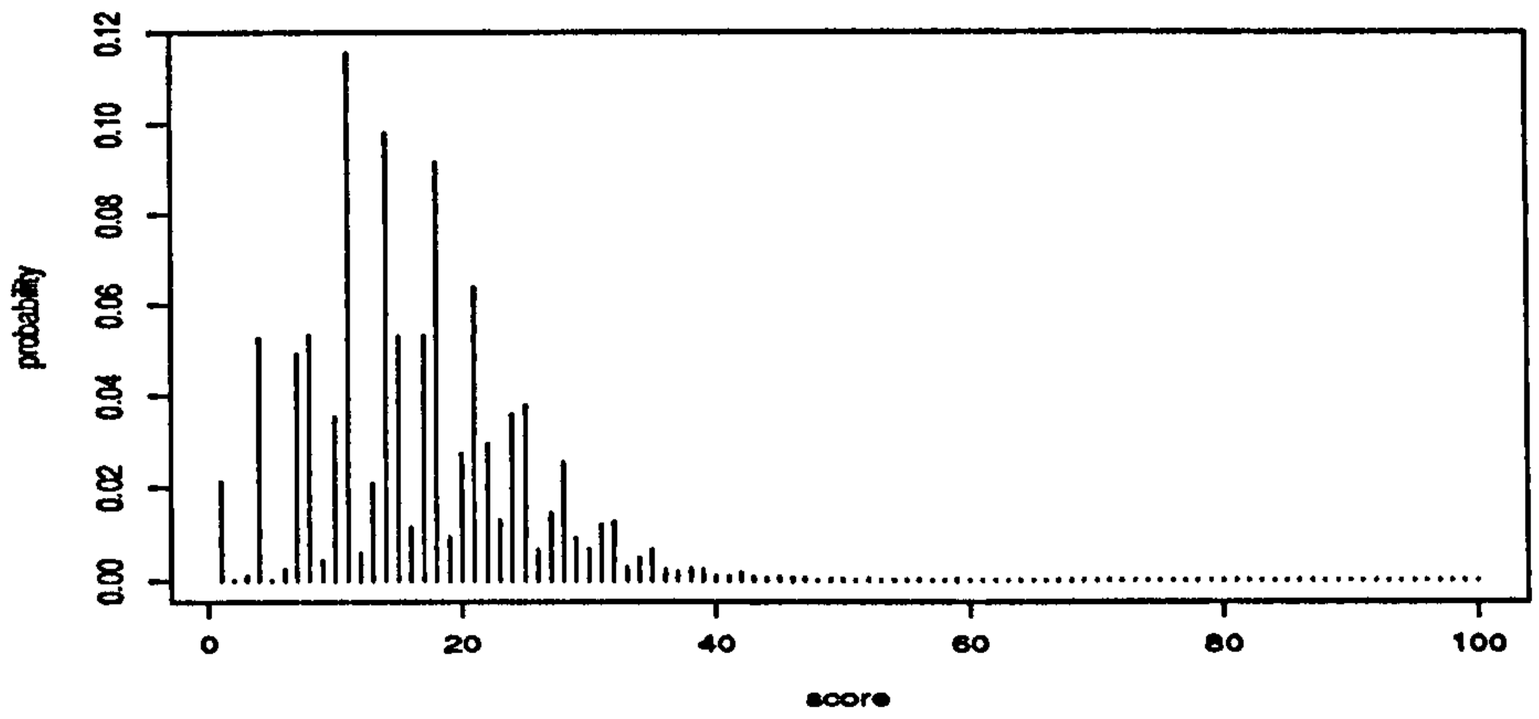


Figure 5.13: Probability density obtained from quasi-multivariate model for New York Giants' final score, SuperBowl 2000/01

match, 10000 simulated values of the home and away scores can be obtained via

$$\begin{aligned} HSC_k^{*i} &= 6 * HTD_k^{*i} + 3 * HFG_k^{*i} + H1C_k^{*i} + 2 * H2C_k^{*i} + 2 * HDC_k^{*i} \\ ASC_k^{*i} &= 6 * ATD_k^{*i} + 3 * AFG_k^{*i} + A1C_k^{*i} + 2 * A2C_k^{*i} + 2 * ADC_k^{*i} \end{aligned}$$

5.6.1 Model evaluation and betting success with quasi-multivariate model

Figure 5.13 plots the density of scores for New York Giants' score in the final match in the data set, which was the 2000/01 SuperBowl. This density is obtained by simulating 10000 outcomes using the distributions obtained in Section 5.6. The uneven density that was treated in Section 5.4 is mimicked. Figure 5.14 displays a moving average plot of predictions against observed values, which reveals that the predictions are broadly reliable.

In Figure 5.15 the betting strategy attempted in Section 5.4.2 is repeated, where bets are placed on the difference in scores and the total score of matches. The results are interesting in that the bets on differences in scores generally win just about frequently enough to ensure a positive expected gain (as discussed in Section 5.4.2, this requires a win rate greater than 52.4%), provided bets are placed when the probability of success is estimated to be greater than around 57%. Nevertheless the profit curve is not close to the red line which represents the proportion of winning bets one would realise if the model probabilities were the 'true' probabilities. The return curve for bets on total scores is very disappointing, even though the lower graph in Figure 5.14 suggests that

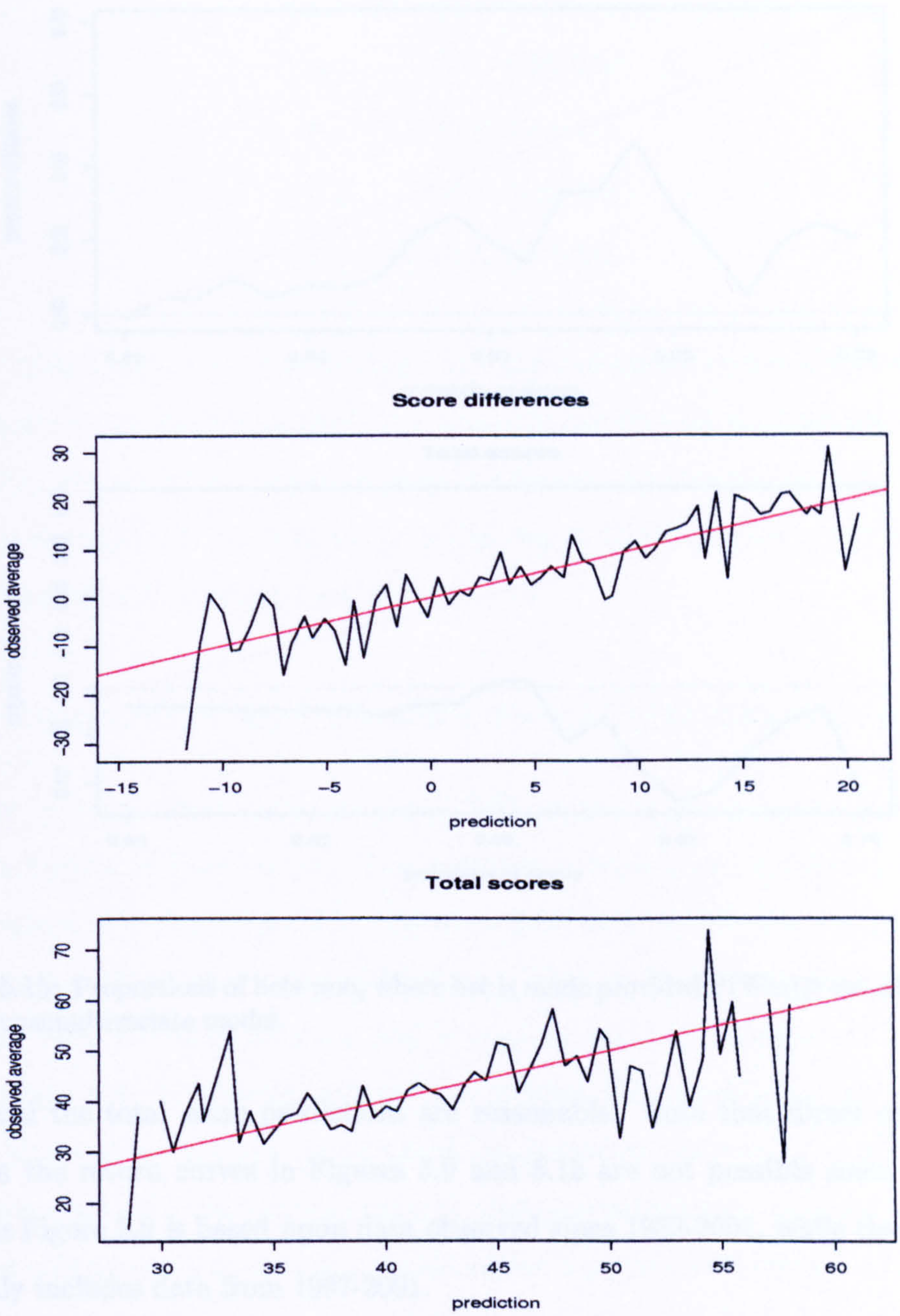


Figure 5.14: Moving average plots of predicted score difference versus observed score difference, and predicted total score versus observed total score for quasi-multivariate model

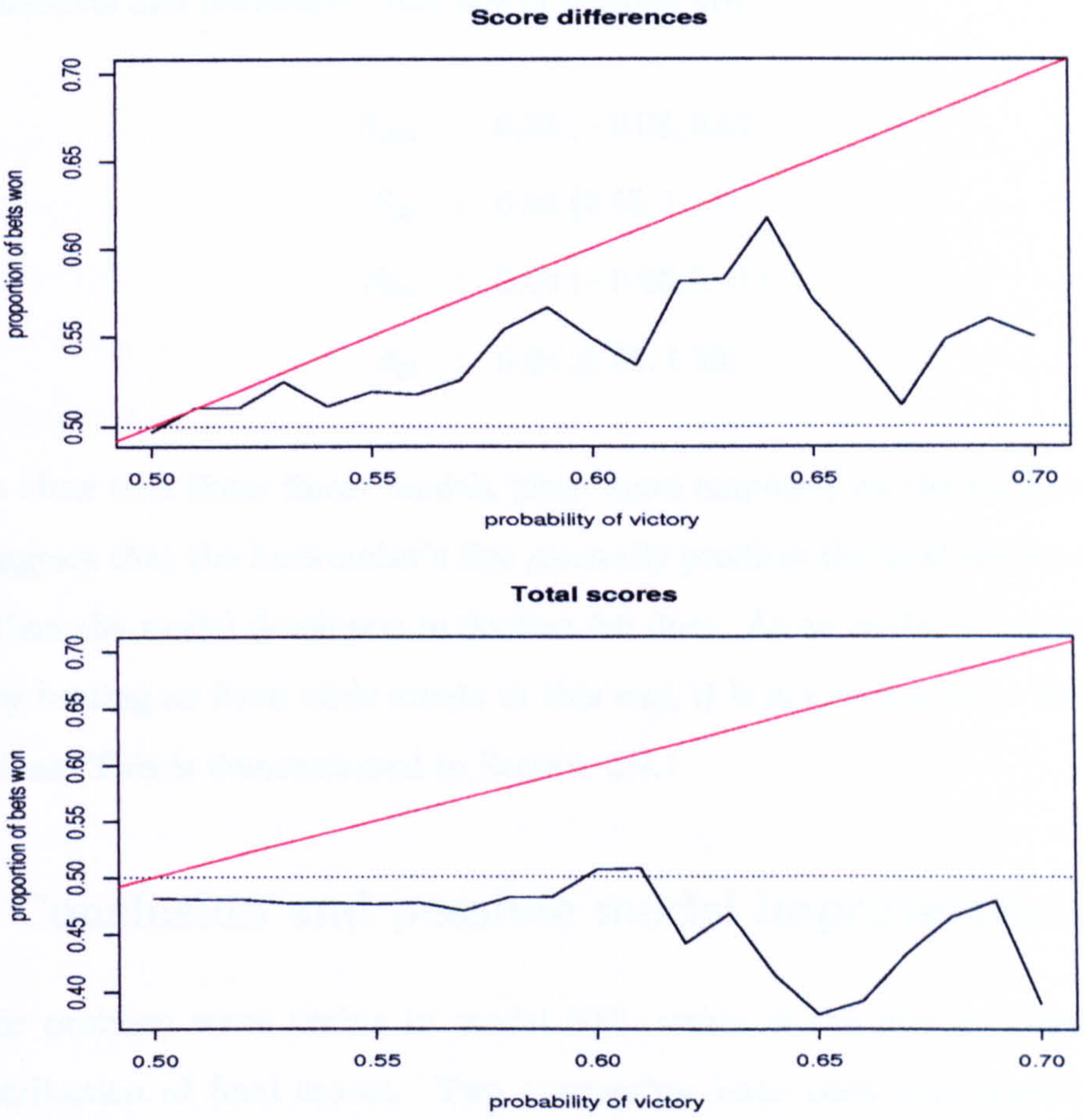


Figure 5.15: Proportions of bets won, where bet is made provided P(Win)$\geq$ cut-off, according to the quasimultivariate model.

in general the total score predictions are reasonable. Note that direct comparisons between the return curves in Figures 5.9 and 5.15 are not possible since the return curve in Figure 5.9 is based upon data observed since 1983-2001, while that of Figure 5.15 only includes data from 1997-2001.

A revealing comparison of the accuracy of the two model predictions against the bookmaker's line is observed by fitting two linear models. Denoting the model means for score differences and totals by ED_m and ET_m and the bookmaker's equivalents by ED_b and ET_b,

$$\begin{aligned} E(HSC - ASC) &\sim \alpha_d + \beta_{dm} ED_m + \beta_{db} ED_b \\ E(HSC + ASC) &\sim \alpha_t + \beta_{tm} ET_m + \beta_{tb} ET_b \end{aligned}$$

the coefficients and confidence intervals of interest are

$$\begin{aligned}
\beta_{dm} &: 0.23\ (-0.08, 0.55) \\
\beta_{db} &: 0.84\ (0.53, 1.15) \\
\beta_{tm} &: 0.08\ (-0.26, 0.41) \\
\beta_{tb} &: 0.94\ (0.55, 1.33)
\end{aligned}$$

It is clear that these linear models place more emphasis on the bookmaker's line. This suggests that the bookmaker's line generally predicts the final results more accurately than the model developed in Section 5.6 does. As an aside, in order to make a profit by betting on fixed odds events in this way, it is not necessary to have superior predictions. This is demonstrated in Section 5.8.1.

5.7 Conclusion and possible model improvements

A major problem when trying to model NFL scores is the non-standard nature of the distribution of final scores. Two approaches have been considered in tackling this problem. The first approach, covered in Section 5.4 constructs a non-parametric density using kernel smoothing techniques. While a distribution that reflects that of actual NFL scores is obtained and is used in order to calculate probabilities of winning bets with the bookmaker, the distribution cannot easily be used in order to obtain MLEs of parameters via standard maximisation routines. The main problem with such a model is its impracticality, in that probabilities can only be obtained by constant reference to a large look-up table, rather than by straightforward calculations. Hence simple tests which make use of this model's predictions and probabilities require continual use of uninviting and error-prone matrix manipulation.

The second approach, developed in Section 5.5, involves forming a model that predicts the events that form the scores, rather than the final score itself. Hence the number of Touch Downs and Field Goals are modelled. Initially, this is attempted by using many other statistics available for each match. Unfortunately the excessively complicated relationship between these variables, and the fact that the statistical distributions of these variables are frequently quite complicated, prevents an accurate marginal distribution for scores from being obtained. A simpler version of this model is implemented in Section 5.6 and while reasonable predictions for scores are obtained,

the predictions inferred from the bookmaker's line are superior.

The focus of this chapter has been more on general statistical methods and little consideration has been given to the nature of NFL itself. This is in contrast to Chapter 4 on yellow and red cards, where the effect of the prevailing climate, inter-team rivalries and the pressure of matches are all taken into account in the model building process. There is plenty of scope for improving the models outlined here in a similar way. In particular, data is available which identifies which players actually participated in each NFL match and the extent of their participation. Certain players, such as the quarterback, are central to the passage of play and many teams do not have two quarterbacks of a similar level of quality or experience. Thus an injury to the first-choice quarterback or other key players, which are not uncommon, are likely to impact upon the expected final score.

Another feature of NFL games is that teams adjust their tactics frequently throughout a match depending on the score of the game. This is true of all sports, where teams frequently become more defensive if they are ahead on goals. In NFL this tactic is used far more regularly, since the stop-start nature of the game permits constant re-organisation and re-evaluation of game strategy. However, all models in this chapter assume a constant scoring rate throughout the match. An alternative is to use quarterly scores for each match (an approach used in Chapter 6 for NBA scores) or even using the time of goals and analysing matches by treating the scoring rates as birth processes.

5.8 Additional comments and information

5.8.1 How a gambler can make a profit off a bookmaker with equally accurate probabilities

With equally accurate probabilities, a gambler can make a profit by betting with a bookmaker which offers odds for all events. To illustrate this, Table 5.16 displays a set of bets on events which each have two possible outcomes and each (unknown to both bookmaker and gambler) has a 50% probability of occurring. While the bookmaker and the gambler disagree on the probability of many outcomes, they are overall equally accurate. While Table 5.16 presents a simple example, the corollary of it can nevertheless be generalised to any situation where bookmakers and gamblers have differing but equally accurate predictions.

Table 5.16: A gambler's decisions and expected returns if a gambler has equally good predictions to bookmaker

Bookmaker's odds	Inferred bookmaker's probability	Gambler's probability	Gambler's decision to bet (Y/N)	Expected return for gambler
11:9	0.45	0.55	Y	$\frac{11}{9}\frac{1}{2} - \frac{1}{2} = \frac{1}{9}$
9:11	0.55	0.45	N	0
9:11	0.55	0.6	Y	$\frac{9}{11}\frac{1}{2} - \frac{1}{2} = -\frac{1}{11}$
4:6	0.6	0.55	N	0
6:4	0.4	0.45	Y	$\frac{6}{4}\frac{1}{2} - \frac{1}{2} = \frac{1}{4}$
11:9	0.45	0.4	N	0

Hence the only occasion when the gambler has an expected loss is when both the bookmaker and the gambler overestimate the probability of a certain outcome but the gambler overestimates it by more. However this expected loss is more than offset by the expected gain on the occasions when both the gambler and the bookmaker underestimate the probability of this outcome but the bookmaker underestimates it more drastically. Both of these situations should occur equally often in the long run since the model and bookmaker are assumed to be equally accurate overall. This leaves the occasions when one of the gambler or bookmaker overestimates the probability of the outcome but the other underestimates it. These correspond to the first two rows of Table 5.16. The gambler either does not bet, or has a positive expected gain. However the situations when the gambler's estimate of the probability of an outcome is higher than that of the bookmaker do not occur frequently, since the bookmaker's probabilities are always inflated to include their own overround. Hence although the gambler makes a long term profit it may accumulate rather slowly.

5.8.2 Procedure to determine the level of parameterisation of team abilities

As mentioned in Section 2.1, there is considerable debate as to what level of detail is required to represent the ability of each team in any given sport. There are various levels of parameterisation that could be considered, such as

- allowing attack and defense parameters for both sides, and a single home effect parameter that applies for all teams
- allowing attack and defense parameters for both sides, and separate home advantage parameter for each side
- allowing home attack, home defense, away attack and away defense parameters

for each side. The team-specific home effect is subsumed by this parameterisation.

Table 5.17 lists the different specifications that could be considered.

Table 5.17: *List of models with different levels of parameterisation*

Level	Model
1	$E[X_k] = E[Y_k] = \gamma$
2	$E[X_k] = \gamma + \delta$ $E[Y_k] = \gamma$
3a	$E[X_k] = \gamma + \delta + \alpha_{i(k)}$ $E[Y_k] = \gamma + \alpha_{j(k)}$
3b	$E[X_k] = \gamma + \delta + \beta_{j(k)}$ $E[Y_k] = \gamma + \beta_{i(k)}$
4	$E[X_k] = \gamma + \delta + \alpha_{i(k)} + \beta_{j(k)}$ $E[Y_k] = \gamma + \alpha_{j(k)} + \beta_{i(k)}$
5a	$E[X_k] = \gamma + \delta + \alpha_{i(k)} + \beta_{j(k)} + \lambda_{i(k)}$ $E[Y_k] = \gamma + \alpha_{j(k)} + \beta_{i(k)}$
5b	$E[X_k] = \gamma + \delta + \alpha_{i(k)} + \beta_{j(k)}$ $E[Y_k] = \gamma + \alpha_{j(k)} + \beta_{i(k)} + \mu_{i(k)}$
6	$E[X_k] = \gamma + \delta + \alpha_{i(k)} + \beta_{j(k)} + \lambda_{i(k)}$ $E[Y_k] = \gamma + \alpha_{j(k)} + \beta_{i(k)} + \mu_{i(k)}$

- In model 1 all teams are assumed to be of equal ability regardless of whether they play at their home ground.
- In model 2 all teams are assumed to be of equal ability except there is an effect from playing at the home ground.
- In model 3a the identity of the opponent is irrelevant when predicting the the score of either side.
- In model 3b, in order to predict the score of either side, the identity of that side is irrelevant and the predicted score is determined by the identity of the opponent.
- Model 4 has the level of parameterisation used in each model so far in this chapter.
- In model 5a, the expected goals scored by the home side, relative to the number they would be expected to score against the same opponents at the opponent's ground or on a neutral ground, varies for each team.
- Model 5b is similar to model 5a except the number of goals the home side concedes, rather than the number of goals they score, is modelled.
- In model 6, both the effect on scoring and conceding rates for the home side of playing a match on their home ground, relative to a match played away from their home ground, varies for each team.

Ideally, one would implement the likelihood maximisation process described in Chapter 3 for each level of parameterisation for every time-point in the whole data set and compare predictive likelihood statistics. However, a less labour and computer intensive method to gain a rough idea as to how many parameters need to be included, the following procedure can be implemented for each level of parameterisation

1. fit a separate generalised linear model as specified in the 'Model' column of Table 5.17 for each season of data available. In this way, a separate vector of estimates for team parameters is obtained for each season. Thus team abilities are assumed to remain constant throughout each season.
2. monitor how well the estimates for a set of team parameters for one season can be predicted from the previous season's.

The models have been fitted for NFL scores and before stage 2 of the procedure is implemented, it is checked that improvements in fit are observed with the addition of the extra team parameters. In general, if the probability distribution used to model data is a member of the exponential family, then if n extra parameters are included in a model which do not improve predictions significantly, the difference in the loglikelihood between the two models, multiplied by two, is asymptotically χ^2_{n-1} distributed. This was a procedure employed by Maher (1982) to determine how many parameters are required to represent soccer teams' abilities. The results of this check are displayed in Table 5.18. The best fit can be obtained with the maximal level of parameterisation.

Table 5.18: *Decrease, and significance of decrease, of deviance when additional team parameters are added into NFL model, season 1997/98 to 2000/01.*

Model number	Parameters in model	Comparison model	Year	Deviance reduction (df, p-value)
2	γ,δ	1	97/98	1038.41 (1,0)
			98/99	1700.75 (1,0)
			99/00	1424.48 (1,0)
			00/01	1177.53 (1,0)
3a	γ,δ,α_i	2	97/98	4861.07 (29,0)
			98/99	12686.35 (29,0)
			99/00	8890.41 (30,0)
			00/01	12151.14 (30,0)
3b	γ,δ,β_i	2	97/98	5206.37 (29,0)
			98/99	3959.65 (29,0)
			99/00	7035.53 (30,0)
			00/01	11418.87 (30,0)
4	$\gamma,\delta,\alpha_i,\beta_i$	3a, 3b	97/98	5132.11 (29,0) , 4786.81 (29,0)
			98/99	4021.2 (29,0) , 12747.9 (29,0)
			99/00	6382.31 (30,0) , 8237.19 (30,0)
			00/01	9130.24 (30,0) , 9862.5 (30,0)
5a	$\gamma,\delta,\alpha_i,\beta_i,\lambda_i$	4	97/98	1490.51 (29,0)
			98/99	1619.7 (29,0)
			99/00	1636.04 (30,0)
			00/01	2284.35 (30,0)
5b	$\gamma,\delta,\alpha_i,\beta_i,\mu_i$	4	97/98	3102.28 (29,0)
			98/99	2835.03 (29,0)
			99/00	3684.95 (30,0)
			00/01	2191.31 (30,0)
6	$\gamma,\delta,\alpha_i,\beta_i,\lambda_i,\mu_i$	5a, 5b	97/98	3042.22 (29,0) , 1430.46 (29,0)
			98/99	2556.85 (29,0) , 1341.53 (29,0)
			99/00	3550.3 (30,0) , 1501.39 (30,0)
			00/01	2211.74 (30,0) , 2304.78 (30,0)

While closer fitted values to past observations can be obtained by increasing the number of team parameters, this does not guarantee that superior predictions can be made. One possible reason could be that the improvements in fit observed by increasing the number of team parameters are caused by modelling the correlations within the random error of the data. Some measures of goodness of fit penalise the addition of extra parameters into a model in an attempt to prevent this. The method used here to detect if extra predictive power can be obtained by using more team parameters is by checking to see if the team ability estimates evaluated one year are informative about the team's ability the next season by applying the following simple least squares regression

$$W_s \sim \tau_0 + \tau_1 * W_{s-1}$$

where W_s is a vector of team coefficients $(\alpha_1, \ldots, \alpha_n)$ or $(\beta_1, \ldots, \beta_n)$ during season s. Examination of the τ_1 term suggests whether W_{s-1} can be used to predict W_s. Coefficients and p-values of the τ_1 terms for each season, for each level of parameterisation are listed in Table 5.19.

Table 5.19: *Coefficients and p-values obtained using previous year's parameters to predict next year's, for NFL, 1997/98 to 2000/01*

Parameters in model	Regression applied	Coefficient and p-value of previous year's parameter α_i	β_i	γ_i	δ_i
γ,δ,α_i	2~1	**0.86,0**			
	3~2	0.39,0.06			
	4~3	**0.9,0**			
γ,δ,β_i	2~1		**0.98,0**		
	3~2		**0.38,0.01**		
	4~3		**0.62,0**		
$\gamma,\delta,\alpha_i,\beta_i$	2~1	**0.89,0**	**0.42,0.03**		
	3~2	0.4,0.06	0.4,0.08		
	4~3	**0.85,0**	0.34,0.09		
$\gamma,\delta,\alpha_i,\beta_i,\lambda_i$	2~1	**0.88,0**	**0.42,0.04**	-0.08,0.66	
	3~2	**0.42,0.04**	0.34,0.15	-0.28,0.13	
	4~3	**0.76,0**	0.3,0.09	0.1,0.71	
$\gamma,\delta,\alpha_i,\beta_i,\mu_i$	2~1	**0.88,0**	0.34,0.06		0.95,0.13
	3~2	0.37,0.07	0.34,0.1		-0.08,0.82
	4~3	0.86,0	0.16,0.41		0.76,0.23
$\gamma,\delta,\alpha_i,\beta_i,\lambda_i,\mu_i$	2~1	**0.88,0**	**0.38,0.05**	0,0.99	0.22,0.33
	3~2	**0.39,0.05**	0.28,0.18	**-0.46,0.03**	0.28,0.08
	4~3	**0.79,0**	0.2,0.28	-0.17,0.46	**0.47,0**

It appears that overall there is a benefit in terms of predictive capability only up to parameterisation level 4. So including two team ability parameters appears to be the most suitable specification.

Chapter 6

Estimating NBA scoring rates: a question of quarters

Basketball is a sport with worldwide appeal and has been popular since the 1950s. This is particularly the case in the United States and since 1949 has been governed by the NBA. The NBA league forms the focus of the research carried out in this chapter.

Initially a model is constructed for NBA scores similar to the model specified by Equation 3.1.1, with parameters being estimated using the MLE procedure described in Chapter 3. As part of the process of building a more advanced model, techniques applied in Chapters 4 and 5 will be considered. Recall from Chapter 4 that details of specific Premier League soccer matches beyond the abilities of the two teams playing were examined, such as the importance of the match result or any particular rivalry between the two soccer teams. Similarly, in Chapter 5, match statistics besides the final scores of the two teams were considered, such as the number of yards either NFL side gained in the match. Approaches similar to these are taken in this chapter along with some new methods such as studying whether the scoring rate adjusts during the course of an NBA match. The data set for NBA is larger than that available for the studies in Chapters 4 and 5. This enables greater statistical significance to be observed when examining various aspects of the data, thus aiding the model enhancement process.

The structure of this chapter is as follows: initially the rules of NBA are summarised. In Section 6.2 the available data is introduced and a basic model for NBA scores is created in Section 6.3. The limitations of this model are discussed via an exploration of the data in Section 6.4. The information gained here is used to specify a more advanced model in Section 6.5, the accuracy of which is compared to both the

basic model and the lines offered by a professional bookmaker in Section 6.6. Section 6.7 concludes the chapter.

6.1 A brief introduction to the NBA League

The model construction process in this chapter considers both the rules of an individual NBA game as well as the regulations that govern the league structure so a brief summary of both is now presented.

6.1.1 League structure

The structure of the NBA league is as follows. 29 teams participate and they are grouped into four different divisions. These are the Atlantic and Central Divisions, which combine to form the *Eastern League*, and the MidWest and Pacific Divisions, which form the *Western League*. The NBA season is divided into a *regular season* and a *post-season*. During each regular season, each side plays 82 games between November and April. This means that teams play a fixture almost every two days throughout this period. Roughly two thirds of these regular season games are played against teams within a team's own league. The 16 teams who perform best in these divisions are allowed to participate in the subsequent post-season tournament. The post-season comprises a knock-out competition known as the *play-offs*. The 16 participating teams are grouped into eight pairs and in the first round, the teams in these pairs play against each other until one side has beaten the other three times. The eight victorious teams progress to another knockout stage where they are again placed into pairs and the victorious team from each pair is the first team to beat the other team in the pairing four times. This leaves four teams progressing to the subsequent round where a similar procedure is followed so that only two teams remain. These two teams qualify for the final round, which takes place in June. Again, the two teams play each other until one side has beaten the other four times. The team that achieves this is the League Champion.

6.1.2 Game regulations

While the rules of NBA are themselves quite complicated, an exhaustive knowledge of all of them is not necessary in order to understand the proceeding analysis. Typically each NBA side has a roster of around fifteen players. Only 12 are allowed to participate

in each game, with only five allowed to play on the court at any one time. Each game is split into four quarters, each lasting 12 minutes of playing time. Points are scored by placing the basketball into the net at the opposing team's end of the playing court. If a successful shot is taken within 25 meters of the net, 2 points are scored. If the shot is taken outside 25 meters, 3 points are awarded. If an infringement is committed against a player while they are in the act of shooting, that player's team is awarded one or two shots (known as *free-throws*) at a distance of four meters from the net. The opponents are not allowed to defend these shots. For each free-throw scored, one point is awarded. Should the scores be level at the end of the fourth quarter, the game then goes into overtime, where another 12 minutes of play are undertaken in order to decide the match winner. If the scores are still level at the end of this quarter, another period of overtime is played, and so on until a winner can be declared.

The divisional rankings, which at the conclusion of the regular season determine which teams qualify for the play-offs tournament, are decided according to the percentage of games the teams have won. Should this percentage be equal for two or more teams, a rather complicated set of rules determine the order in which these teams are ranked, such as individual results between these teams, or the percentage of games won against other teams in the league. It is only whether a team wins or loses that is recorded when teams are ranked in their division and the margin of victory in any games is not relevant at any stage. So, whether a team wins a match by 1 point or 25 points, their league position remains unchanged. This could be an important consideration when an attempt to predict the scoring rates is made.

Given its status as one of the premier US sports, it is not surprising that there is a great deal of interest in betting on NBA. While spread betting is possible, the majority of bets offered are fixed odds handicaps bets, which are described in Section 1.3.1.

6.2 NBA data

The data set over which the models of this chapter are developed includes all regular and post-season matches from the 1997/98 season until the 2000/01 season. The 1997/98, 1999/2000 and 2000/01 regular seasons all consist of 1189 matches while the respective post-seasons each consist of approximately 70 matches[1]. There was a

[1]Since the post-seasons are a set of mini-tournaments, with each being won by whichever team wins for the third time, in the case of the first round, and the fourth time in subsequent rounds, these mini-tournaments can consist of a variable number of matches. So the total number of post-season matches changes from year to year.

players' strike in the 1998/99 season so the regular season of that year only contained 725 matches, which took place between February and May of 1999. The total number of matches in the data set is 4568. This compares to 1020 matches in the main NFL data set used in Chapter 5 and the 1900 Premier League soccer matches used to predict booking rates in Chapter 4. The data available for each match includes, for both the home and away sides:

- the date of the match
- the final scores (including points scored in overtime periods)
- the number of points scored in each quarter
- the bookmaker's line for both the difference in score and the match total. The purpose of these lines is described in Section 1.3.1. They can be considered as a prediction of the median difference in score and total score of the match.
- the number of attempted and successful 2-point shots
- the number of attempted and successful 3-point shots
- the number of attempted and successful free throws

In order to provide an idea of the scale of these figures, Table 6.1 displays this information for the first five matches in the data set.

Table 6.1: First five matches in data set. The figures for the home team are listed above the figures for the away team

Date	*Teams*	*Score*	*Total Attempts*	*2-Point Attempts*	*3-Point Attempts*	*Free Throw Attempts*	*Bookmaker's Line*
19971031	Boston	92	105	72	18	15	-9
	Chicago	85	108	71	8	29	
19971031	Vancouver	88	109	76	13	20	2
	Dallas	90	106	63	11	32	
19971031	Miami	114	110	62	25	23	7
	Toronto	101	124	83	9	32	
19971031	Charlotte	85	114	57	13	44	2
	New York	97	103	59	10	34	
19971031	LA Lakers	104	115	49	29	37	2.5
	Utah	87	117	77	7	33	

6.3 A basic NBA scores model

For the proceeding analysis total scores at the end of the fourth quarter are studied. While extra information is available by using the final score (that includes points scored

in overtime periods), in order to make valid comparisons between the points scored in each game it is necessary to compare points scored in equal periods of time. It is also important that scores at the conclusion of the fourth quarter are used rather than the final score including overtime periods, because if points scored in overtime are included in the final score, the score variance estimate is inflated. So for this chapter, the term 'score' is defined as the score at the end of the fourth quarter rather than the final score of the match including overtime periods, unless otherwise indicated.

The (home mean, away mean, home standard deviation, away standard deviation, home and away correlation) for scores are respectively (95.88, 92.69, 11.83, 11.02, 0.37). Figure 6.1 displays a histogram of combined home and away scores for this data set. Its symmetry, combined with the relatively high correlation coefficient considering the number of observations (4568 matches) suggests that a bivariate Normal distribution seems suitable for the scores. Thus, the formulation is:

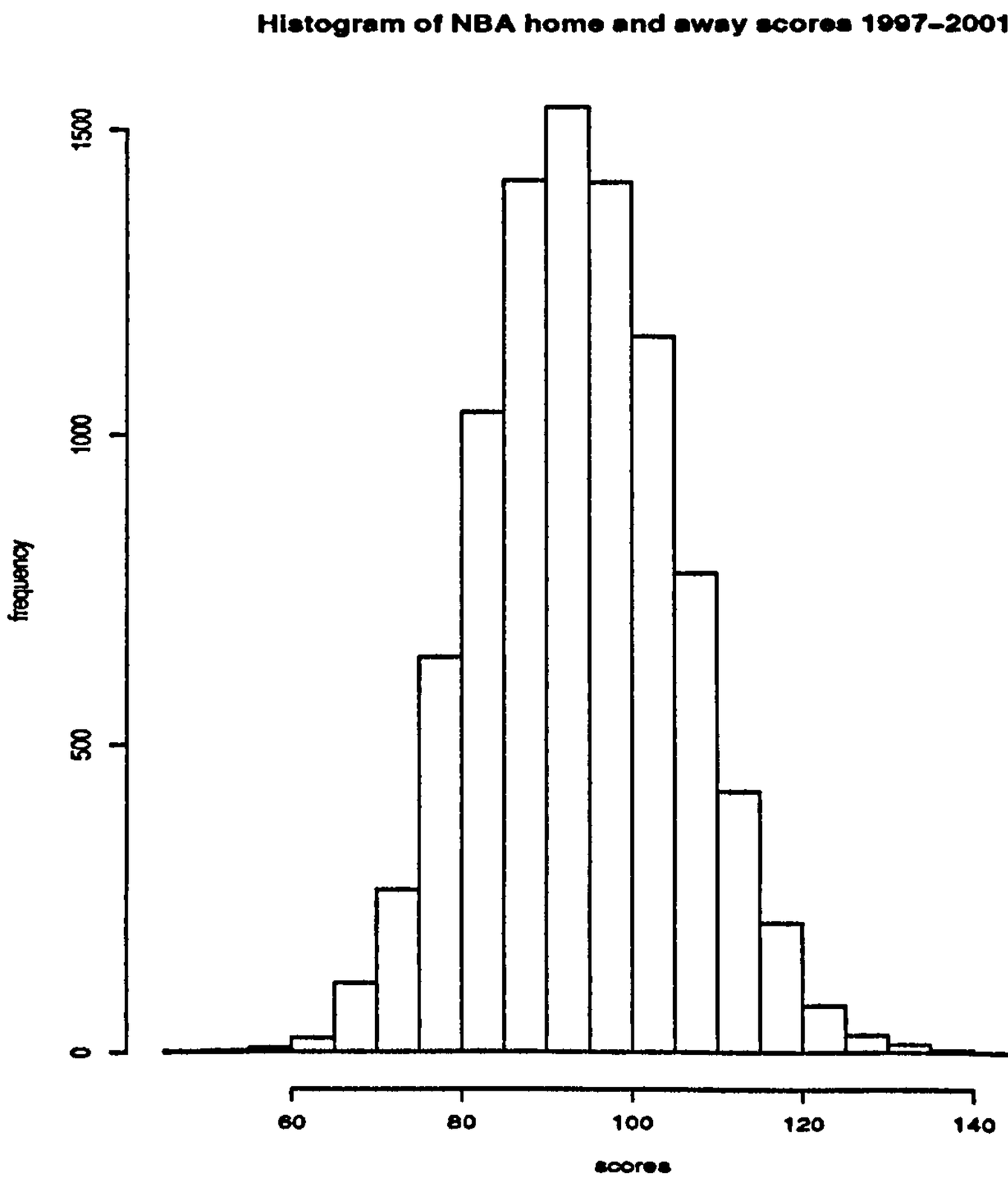


Figure 6.1: Histogram of NBA scores, 1997-2001

$$
\begin{aligned}
HSC_k &\sim \mathcal{N}(\mu_k, \sigma_h) \\
ASC_k | HSC_k = x &\sim \mathcal{N}(\lambda_k + \rho(x - \mu_k)\frac{\sigma_a}{\sigma_h}, \sigma_a\sqrt{1-\rho^2})
\end{aligned} \tag{6.3.1}
$$

and

$$
\mu_k = \gamma + \alpha_{i(k)} + \beta_{j(k)} + \delta
$$

$$
\lambda_k = \gamma + \alpha_{j(k)} + \beta_{i(k)}
$$

where

- HSC_k, ASC_k are the home and away scores in match k
- $\alpha_{i(k)}, \alpha_{j(k)}$ are offensive parameters for respectively the home and away teams
- $\beta_{i(k)}, \beta_{j(k)}$ are defensive parameters for respectively the home and away teams
- γ is the global mean
- δ is the home effect.
- σ_h, σ_a are the home and away score standard deviations
- ρ is the (home score, away score) correlation coefficient

In order to obtain parameter estimates for the parameters in the model specified by Equation 6.3.1 it is necessary to have near-optimal values for the external parameters, as described in Chapter 3. The procedure for obtaining them is applied in Sections 4.3.5, to obtain parameter estimates for a yellow card model, and 5.3 to obtain parameter estimates for a model of NFL scores and is repeated at this stage. Table 6.2 displays the predictive likelihood obtained for a range of values of the external parameters, suggesting that (0.1,5,20) are close to the optimal values of respectively the time down-weighting (ς), offensive/defensive prior tightnesses ($\tau_{\alpha\beta}$) and between-season truncation gap (w)[2]. Notice the difference between the time down-weighting

[2]Note that the predictive likelihood diverges for certain values of the external parameters. The reason for this is that at the start of one season, the high down-weighting factor combined with the large seasonal truncation values means that only data since the start of the season is 'remembered' by the likelihood function. Hence the number of available parameters is greater than the number of data points and the weak prior on the team parameters allows the MLE of the correlation coefficient to be almost 1. As a result, many subsequent observations are calculated to have probability of 0, hence the sum of the logs of these probabilities is $-\infty$.

parameter between the NFL model and the NBA model. For NFL the near-optimal value of ς is 0.05 whereas here it is 0.1. In the NBA application, a 10% weight is placed on a match 23 weeks ago when the likelihood is maximised, whereas for the NFL model, a match 46 weeks ago has a 10% weight. NBA teams play approximately three times a week for up to seven months of the year, whereas NFL teams play once a week for up to five months .It follows that the likelihood maximisation procedure for NFL scores includes a larger number of less recent matches in order to increase the number of observations with which parameters are estimated.

Table 6.2: Predictive likelihood obtained for different choices of external parameters for final scores

Truncation w= 5 weeks:

		Prior variance $\tau_{\alpha\beta}$ of offensive and defensive estimates			
		2	5	10	20
	0.01	-31596.3077	-31580.6296	-31590.3172	-31593.8397
	0.05	-31521.145	-31420.2025	-31428.6001	-31433.8074
Weight ς	0.1	-31580.2486	-31366.7204	-31370.8393	-31380.7614
	0.2	-31678.2108	-31364.5173	-31365.1443	-31394.373

Truncation w= 10 weeks:

		Prior variance $\tau_{\alpha\beta}$ of offensive and defensive estimates			
		2	5	10	20
	0.01	-31589.7093	-31567.5079	-31577.0747	-31580.5907
	0.05	-31514.5357	-31398.5752	-31406.7237	-31412.4433
Weight ς	0.1	-31591.1065	-31359.3255	-31360.6186	-31371.5109
	0.2	-31664.3299	-31361.3923	-31372.5807	-31412.9664

Truncation w= 20 weeks:

		Prior variance $\tau_{\alpha\beta}$ of offensive and defensive estimates			
		2	5	10	20
	0.01	-31598.5711	-31544.7656	-31554.0958	-31557.6064
	0.05	-31510.5953	-31370.856	-31378.1846	-31454.8696
Weight ς	0.1	-31588.9692	-31349.4321	-31352.4165	-31367.8302
	0.2	-31653.1635	-31897.5825	-Inf	-Inf

Truncation w= 30 weeks:

		Prior variance $\tau_{\alpha\beta}$ of offensive and defensive estimates			
		2	5	10	20
	0.01	-31601.3265	-31553.884	-31563.7258	-31566.3006
	0.05	-31519.1974	-31381.4687	-31389.139	-31464.7527
Weight ς	0.1	-31596.4297	-31353.7981	-31357.4613	-31378.109
	0.2	-31667.769	-31913.4894	-Inf	-Inf

Using the estimates obtained by maximising the likelihood at each time-point, predictions for each match can be made. Each team's offensive and defensive parameter estimate after the final match in the data set on 3 June 2001 is displayed in Table 6.3. The frequent large differences in the values of the offensive and defensive parameters of a team that was observed in the NFL study is also seen here. Like the NFL, the NBA also includes a 'draft' system whereby the least successful teams each season have first

choice of the graduating college Basketball players for the next season. These players are obliged to play for the team for a minimum of five years. This limits the number of 'star' players that play for any team and it follows that it is difficult for a team to be one of the best both offensively and defensively. Furthermore within a match a team's ability to score points is to some extent proportional to how willing they are to risk conceding points to their opponent. As a result each team's individual underlying strategy throughout a season is a trade-off between attacking and defensive play. This second consideration applies far less to NFL since the the offensive players and defensive players do not play at the same time in the match. Thus the defensive players are not normally expected to switch to an attacking mode of play and similarly for offensive players.

Table 6.3: Offensive ($\hat{\alpha}$) and defensive ($\hat{\beta}$) NBA team ability estimates according to basic model, June 2001

team	$\hat{\alpha}$	rank	$\hat{\beta}$	rank
Portland	0.896	12	-1.4147	9
Boston	1.4723	9	2.3218	20
Vancouver	-3.4054	25	2.4879	22
Miami	-6.9943	29	-5.0345	4
Charlotte	0.5661	13	-4.3919	5
LA Lakers	8.8196	1	-3.0014	7
Orlando	3.3811	5	3.1527	25
New Jersey	-1.2154	19	4.8569	26
Denver	0.2832	14	2.7809	24
Detroit	-0.1271	17	-0.6943	13
Houston	2.8479	6	0.1362	15
Philadelphia	-2.5215	24	-5.6697	3
Phoenix	-1.7862	23	-2.9609	8
Minnesota	1.3942	10	0.5105	16
Milwaukee	5.2354	4	-1.0619	12
Chicago	-4.6988	26	2.0572	19
LA Clippers	0.1763	16	-0.2841	14
Atlanta	0.233	15	6.9112	27
Utah	-1.2823	20	-3.0231	6
Indianapolis	-1.535	22	-1.1342	11
Seattle	1.5205	8	-1.1994	10
San Antonio	-1.2118	18	-7.0554	1
Washington	1.0371	11	8.6027	29
Sacremento	7.537	2	1.5947	18
New York	-5.9292	28	-6.7165	2
Cleveland	-1.4035	21	2.4694	21
Dallas	5.249	3	2.5838	23
Toronto	1.8681	7	1.5085	17
Golden State	-4.7234	27	7.3505	28

The moving average prediction of difference in score is compared to the moving average observed difference in score in Figure 6.2. Figure 6.3 plots the model predictions against the line offered by the bookmaker. From Figure 6.2 the predictions appear to

be generally sensible so it is not surprising that in Figure 6.3 a broad similarity between the two sets of predictions is observed. The matches furthest from the diagonal represent the matches where the bookmaker and the model disagree most strongly and these matches are examined in Section 6.4.6.

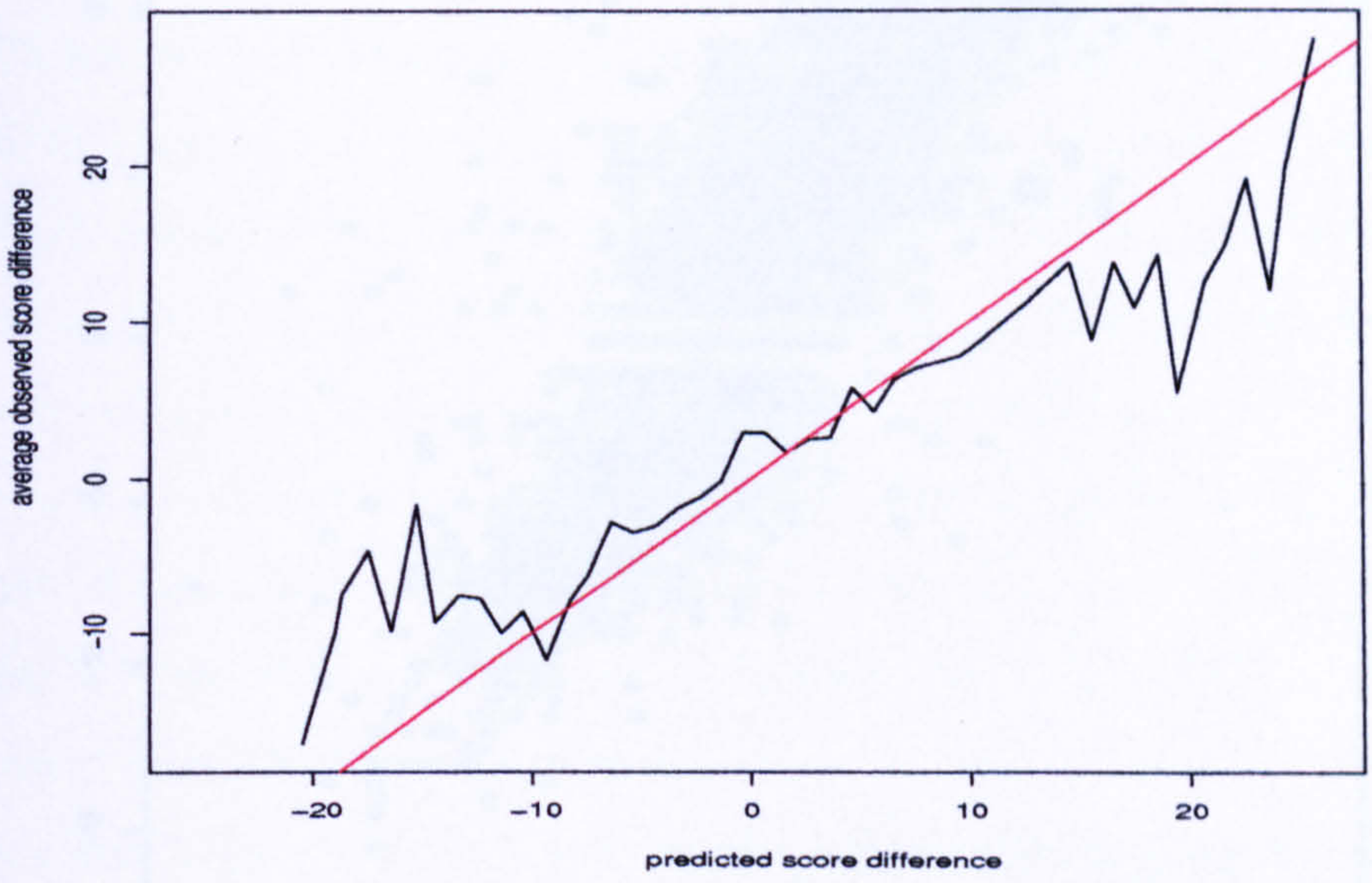


Figure 6.2: Moving average of expected (home-away) scores plotted against moving average of observed (home-away) scores

6.4 Possible improvements to the basic model

While the basic model can be used to produce sensible predictions for matches there are several restrictions implicit within this model that could be relaxed. Among the issues to be considered are that

- it is assumed that scoring rates are constant throughout a match, regardless of the size of the difference in score at any point. In practice the tactics of teams alter during the course of a match, depending on the score of the match, the condition of the players, the tactics being used by opposing teams and other factors.

- NBA teams play a large number of games within a short time-period, as described in Section 6.1. Players may become tired but the basic model does not adjust the scoring rates to reflect this.

- the basic model allows teams' abilities to adjust over time at a single rate, via the

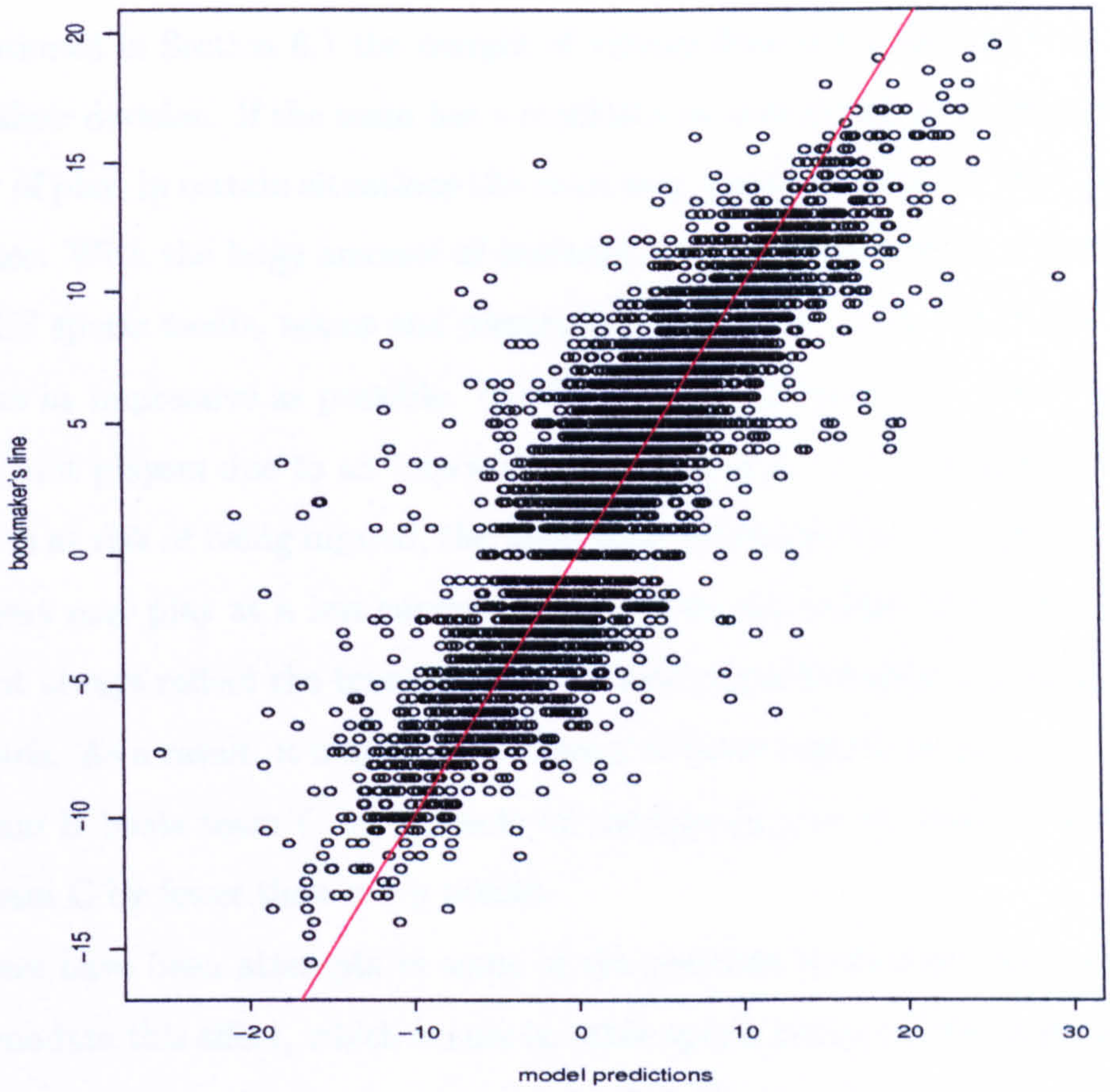


Figure 6.3: Plot of expected difference in scores according to model against bookmaker's line

parameter (ς). While the underlying ability of the players and management may adjust in the long term, it is conceivable that the standard of their performances fluctuate in the short term due to factors which vary more quickly, such as confidence or team morale.

- match scores are the only data values included in the model. Extra accuracy may be achieved using some of the match totals available for other aspects of the game, as listed in Section 6.2.

- two parameters are used to express the abilities of teams whereas other parameterisations might be more appropriate.

- the identity of the players participating in a match is likely to affect the scoring rates whereas the basic model does not consider the squad details of any fixtures.

Throughout the rest of this section, each of these restrictions will be evaluated by examining the available data.

6.4.1 Truncation of winning margins

As mentioned in Section 6.1 the margin of victory does not affect the team's ranking within their division. If the team has a comfortable lead at the conclusion of the third quarter of play, in certain situations the team may continue to try to increase the score difference. With the large amount of attention devoted to statistics in NBA coverage in the US sports media, teams and players may attempt to make their own individual statistics as impressive as possible. In other situations, such as if the team wants to rest its best players due to an important future fixture, or if it senses that one of its players is at risk of being injured, the team may withdraw their first choice players, or its players may play at a less energetic pace. Thus the margin of victory in a fixture does not always reflect the true disparity in level of performance during the course of the match. As a result, it may be that if team A beats team B by x points on average, and team B beats team C by y points on average $(x, y > 0)$, team A is expected to beat team C by fewer than $x + y$ points.

There have been attempts in some of the previous studies of sports modelling to accommodate this effect, which occurs in other sports besides NBA. Rue and Salvesen (1997) included an additional parameter in their Premier League soccer scores model that reflects their belief that a soccer team tends to underestimate its opponent if the opponent is weaker. Hence, defining X as the score of the home team, μ and λ as the abilities of the home and away teams respectively

$$E[X] = exp(\mu - \gamma(\mu - \lambda))$$

where γ is a term to allow $E[X]$ to vary according to the difference in ability between the two sides. If this effect is indeed present as Rue and Salvesen surmise, γ should be small and positive.

Stefani (1980) uses a similar idea to reflect this effect in a study of both NFL scores and College Football scores. The equation for a predicted winning margin w_k in match k between sides $i(k)$ and $j(k)$ is

$$E[w_k] = h_k + \lambda(r(i(k)) - r(j(k)))$$

where h_k represents the home advantage assuming match k is played on team $i(k)$'s field) and $r(i(k))$ and $r(j(k))$ are the ratings of team $i(k)$ and $j(k)$. $r(i(k))$ and

$r(j(k))$ can be considered as the average number of points advantage a team has over a reference team. The λ term is included to prevent over-predicting the margin of victory as the difference in team abilities increases. The estimated values for λ were 0.75 for College Football and 0.67 for NFL.

One strategy that can be readily attempted which exploits the available data is to develop a model for home and away scores at the end of the 3rd quarter, then model the 4th quarter scores conditional on the scores at the end of the 3rd quarter. Combining these models may obtain more accurate marginal distributions for the scores at the end of the 4th quarter than the basic model.

Figure 6.4 shows that there is, if anything, a negative correlation between the score difference at the start of the fourth quarter and the difference in points scored by the two teams in the final quarter. This suggests that teams' level of performance is not constant throughout an entire game and that the earlier stages of a match are frequently used to establish a score supremacy over opponents. The later stages of a match can be used to rest players while preventing opponents from scoring sufficient points to overturn the team's score advantage.

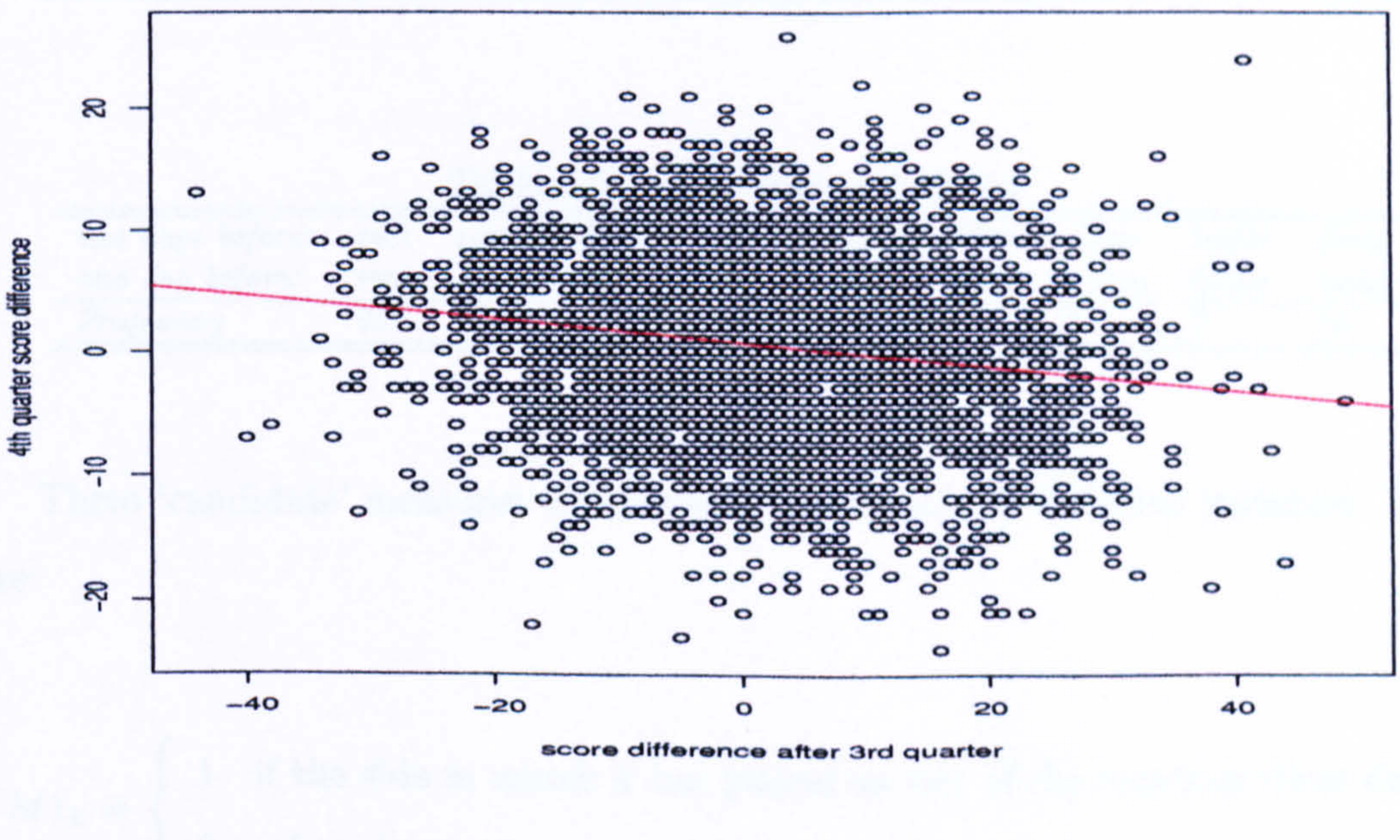


Figure 6.4: Plot of 4th quarter score differences against score difference at the end of the 3rd quarter for basketball, seasons 1997-2001

The line of best fit that is included in Figure 6.4 is obtained by regressing the difference in points scored in the final quarter between the two sides against the difference in score at the end of the third quarter. The coefficient and 95% confidence interval is -0.098 (-0.115, -0.081), leaving little doubt that this is an essential feature to be added

to the model.

6.4.2 Effect of schedule

With each NBA team playing 82 games within approximately seven months, teams are often required to play several matches in a short space of time, and their performance in some matches may be affected in some way due to this. Furthermore, the necessity to travel a long distance shortly before a match may be detrimental to the team's performance.

To study the suggestion that a team that is tired due to playing a busy schedule will perform less well, the level of tiredness needs to be quantified in some way. A measure of tiredness should be found that is specific enough to distinguish between enough situations, and general enough to include sufficient data to obtain significant results where appropriate. Tables 6.4 and 6.5 display counts concerning schedules of the teams in the two days prior to a match:

Table 6.4: Home team schedules

Two days before:	rest	rest	rest	home	away	home	away	home	away
One day before:	rest	home	away	rest	rest	away	home	home	away
Frequency	1242	95	589	1591	1018	22	4	0	7

Table 6.5: Away team schedules

two days before:	rest	rest	rest	home	away	home	away	home	away
one day before:	rest	home	away	rest	rest	away	home	home	away
Frequency	957	696	768	979	1136	9	6	7	10

Three 'candidate' measures are devised, in the form of indicator variables. These are:

$$M1_k = \begin{cases} 1 & \text{if the side in match } k \text{ has played on two of the previous three days} \\ 0 & \text{otherwise} \end{cases}$$

$$M2_k = \begin{cases} 1 & \text{if the side has had to travel in the last 24 hours to match } k \\ 0 & \text{otherwise} \end{cases}$$

$$M3_k = \begin{cases} 1 & \text{if the side in match } k \text{ has played a match anywhere on the previous day} \\ 0 & \text{otherwise} \end{cases}$$

These vectors are constructed for the home and away sides in each match. To verify if any of the measures listed above signify that a team is tired to the point where the match score may be affected, the observed average scores are compared to the predictions obtained using the basic model via simple linear regression. Tables 6.6 and 6.7 display the results for the home and away scores.

Table 6.6: For home teams, effect of schedule on average score difference

	$M1_k$: *Played 2 of 3 previous games*	$M2_k$: *Travelled to current fixture*	$M3_k$: *Played previous day*	*Number of games*	*Mean score difference minus prediction*
1	Y	Y	Y	345	-1.364 (-2.557,-0.172)
2	Y	Y	N	0	–
3	Y	N	Y	68	-0.369 (-2.658,1.92)
4	N	Y	Y	208	-1.552 (-2.906,-0.197)
5	Y	N	N	665	-0.285 (-1.114,0.544)
6	N	Y	N	0	–
7	N	N	Y	23	-0.989 (-5.259,3.281)
8	N	N	N	2763	0.39 (0.006,0.774)

By comparing rows 1 and 5 of Table 6.6, it seems that teams are tired by having to play the day previous to a fixture, while playing two games on successive days then having a rest before a match does not appear to tire teams. A comparison of row 1 with rows 3 and 7 would ideally clarify whether it is playing the previous day, or the travelling over the last 24 hours, that tires teams, but unfortunately rows 3 and 7 do not contain enough observations to draw any conclusions with confidence. Hence the overall message from Table 6.6, by comparing rows 1,3,4 and 7 with row 8, seems to be that playing the previous day on average reduces the score difference by between 1.5 and 2 points.

Table 6.7: For away teams, effect of schedule on average score difference

	$M1_k$: *Played 2 of 3 previous games*	$M2_k$: *Travelled to current fixture*	$M3_k$: *Played previous day*	*Number of games*	*Mean score difference minus prediction*
1	Y	Y	Y	946	-0.796 (-1.536,-0.055)
2	Y	Y	N	0	–
3	Y	N	Y	0	–
4	N	Y	Y	368	-0.262 (-1.35,0.825)
5	Y	N	N	574	-0.25 (-1.11,0.609)
6	N	Y	N	0	–
7	N	N	Y	0	–
8	N	N	N	2184	0.433 (-0.029,0.895)

Concerning teams who played away from home, it seems from row 5 of Table 6.7 that again, provided a team has rested the previous day, their score is not affected on average, even if they did play games on the two previous days. However, comparison of rows 1 and 4 suggests that playing two games in the last three days, one of which

took place the previous day, is more tiring than only playing a fixture the day before. However, the confidence intervals in rows 4 and 5 suggest that more observations are required to draw any strong conclusions with respect to this. Also, further discoveries may be made if the distance teams had to travel to a fixture is calculated and included within the framework above.

Overall, there is little doubt that a team's schedule affects their average result and measure $M3_k$ defined above seems to be the most appropriate one to include in the model. By comparing the parameter estimates for the following two linear models it appears that the specification by which the variables are included is quite crucial. Denoting the expected home and away score means from the basic model by μ_k and λ_k,

$$E[HSC_k - \mu_k] = \beta_0 + \beta_1 * HTIRED_k + \beta_2 * ATIRED_k \qquad (6.4.1)$$

$$E[ASC_k - \lambda_k] = \beta_0 + \beta_1 * ATIRED_k + \beta_2 * HTIRED_k \qquad (6.4.2)$$

For the first model, the coefficients and confidence of interval for β_1 and β_2 are -0.791 (0.159, -1.741) and 1.071 (1.767, 0.375), while for the second model they are -0.089 (0.566, -0.744) and 1.119 (2.013, 0.225). It appears that teams concede more points as a result of playing the day before, but their own scoring rate is not significantly affected.

6.4.3 Short-term form

The underlying ability of a team rarely changes drastically within a short period of time, since it is largely determined by the abilities of the players in the squad. These do not change on a regular basis. But while the team's ability may change slowly, its short-term form might fluctuate. For example a set of recent bad results may affect the team's confidence briefly. To detect this short-term form effect, a method to determine if the team is in a spell of particularly good or bad form is needed. Hence a *form measure* is needed for each team entering each match. Several methods are tried here.

Using recent results for prediction

It may be worth incorporating a team's most recent results, as well as their long term ability, in order to predict their scores. Two methods to measure a team's recent form are attempted.

Method 1 averages out the team's margin of victory or defeat in the previous k matches to create a form vector. The difference between the observed and predicted difference in score from the basic model is regressed against this form vector. The significance of the form vector is monitored. Thus if a team has performed badly compared to their opponents in many recent matches it is possible that their confidence or team morale has dropped to a point where their expected score in a future match is affected. The basic model may not adequately adjust sufficiently for the team's bad results since teams' abilities are assumed to adjust only over a longer period of time. A similar argument can be applied to suggest that the score of a team with recent good results may be significantly different to that predicted by the basic model.

Method 2 is similar to Method 1, but the difference between the margin of victory or defeat and the bookmaker's line for the previous k matches is calculated to create a form vector. So, instead of observed performance, it is the extent to which they have exceeded expectation that is considered. This may more accurately measure their confidence entering the next match.

As an illustrative example, the model fitted for Method 1, k=1 for the score difference in match m is

$$E[(HSC_m - ASC_m) - (\mu_m - \lambda_m)] = \beta_0 + \beta_1(HSC_n - ASC_n) \qquad (6.4.3)$$

where match n is the previous match in which the home side of match m participated, μ_m and λ_m are the expected home and away scores implied by the basic model and β_0 and β_1 are the coefficients to be obtained. A large significant value of β_1 would suggest that a team's performance relative to its opponent, compared to that predicted by the basic model, is significantly improved given a good result in a previous match.

The model fitted for Method 2, k=1 for the score difference in match m is

$$E[(HSC_m - ASC_m) - (\mu_m - \lambda_m)] = \beta_0 + \beta_1(HSC_n - ASC_n - B_n) \qquad (6.4.4)$$

where the μ_m, λ_m, β_0 and β_1 terms are defined similarly and B_n denotes the book-

maker's line for the home side's previous match. The bookmaker's line is used as an alternative to the prediction made by the basic model, $\mu_n - \lambda_n$, since the the dependency between $\mu_n - \lambda_n$ and $\mu_m - \lambda_m$ could produce misleading results if a model were fitted containing both of these terms.

Both methods are implemented on both home scores and away scores. The methods are tested for values of k between 1 and 10 and Table 6.8 displays the estimated coefficients for the β_1 values.

Table 6.8: Coefficients and significance levels for different values of k

k	method 1 at home (coef, p-val)	method 1 away (coef, p-val)	method 2 at home (coef, p-val)	method 2 away (coef, p-val)
1	(0.004, 0.738)	(-0.01, 0.38)	(-0.015, 0.298)	(-0.015, 0.263)
2	(0.004, 0.827)	(-0.016, 0.301)	(-0.023, 0.274)	(-0.026, 0.187)
3	(0.019, 0.338)	(-0.006, 0.759)	(-0.012, 0.643)	(-0.03, 0.233)
4	(0.016, 0.464)	(-0.017, 0.418)	(-0.015, 0.621)	(-0.061, 0.035)
5	(0.014, 0.549)	(-0.019, 0.384)	(-0.026, 0.461)	(-0.066, 0.049)
6	(0.012, 0.619)	(-0.031, 0.183)	(-0.063, 0.099)	(-0.103, 0.006)
7	(0.002, 0.924)	(-0.027, 0.265)	(-0.069, 0.101)	(-0.109, 0.008)
8	(0.015, 0.589)	(-0.03, 0.235)	(-0.047, 0.302)	(-0.119, 0.006)
9	(0.016, 0.553)	(-0.023, 0.364)	(-0.045, 0.365)	(-0.102, 0.027)
10	(0.01, 0.735)	(-0.021, 0.428)	(-0.059, 0.253)	(-0.12, 0.015)

The value of 0.004 obtained when Method 1 is applied to home team's score differences, for k=1, means that for every point by which the home side beat their opponent in their previous match, in the current match they beat their opponent by average 0.004 points more than the basic model predicts. The only significant results obtained for the models fitted are for Method 2, away games, values of $k > 3$ but in fact it seems that on average teams do worse playing away from home if their form prior to the game was good!

Additionally, it is of interest to see if the bookmaker's line is sensitive to recent runs of form by the teams involved. If this is so, then consideration of the results from Table 6.8 suggests that this reaction would be misplaced, thus presenting an area of the betting market to exploit. The first graph in Figure 6.5 plots, in blue, the average observed difference in score for each difference in form level between the two teams in a match, where form is defined by Method 1 above. Plotted over this, in black, is the average of the bookmaker's line for each match and, in red, the average predictions obtained from the basic model. The lower graph in Figure 6.5 is similar to the upper graph except form is defined by Method 2.

Note that the number of observations decreases towards both the right and left

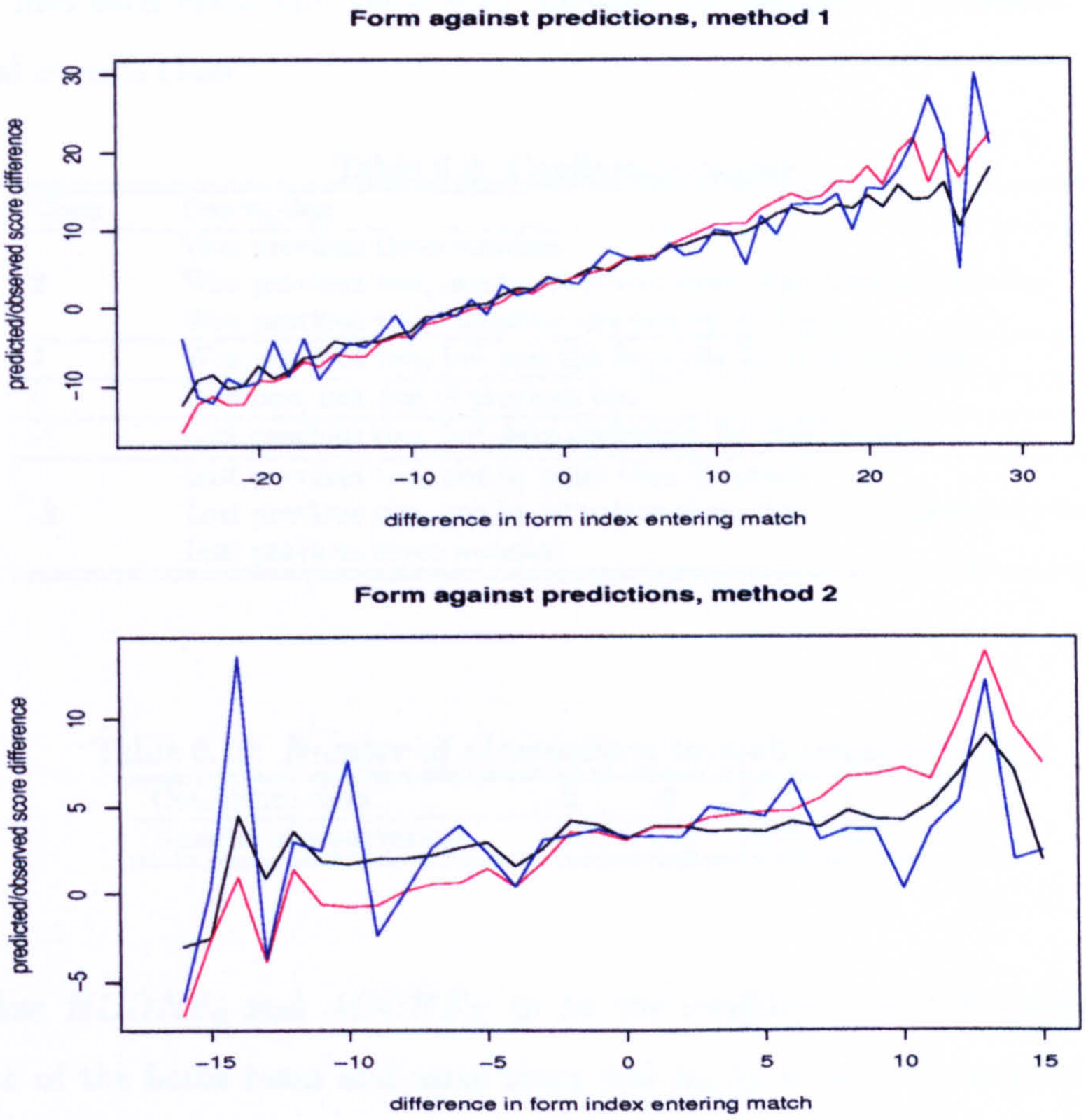


Figure 6.5: Plotting scores (___), model predictions (___) and bookmakers spreads (___) against recent runs of form

hand edges of the plots if Figure 6.5. This explains the large variance in these areas of the plots. If the bookmaker's line is sensitive to recent runs of forms, the black lines should increase above the blue ones on the right hand side, representing an over-reaction to good form by a team, and decrease below the blue line on the left hand side, representing an over-reaction to a bad run of form. In fact, the bookmaker's line ties in closely with the average observed scores regardless of the run of form suggesting that market over-reaction is not taking place.

Construction of a confidence score

Quantifying a team's level of confidence at any one time objectively is difficult. However, an attempt is made here via intuition to construct a vector which approximately represents each team's confidence entering the match.

Suppose that the confidence of a team can assume one of five levels, ranging from maximum confidence to minimum confidence. Table 6.9 displays how teams are al-

located into each class and Table 6.10 displays the number of occasions teams are classified in each class.

Table 6.9: Confidence classes

Class	Description
2	Won previous three matches Won previous two, one by 10 points more than bookmaker's line Won previous two, including one win by $\geq$15 points
1	Won previous two, but was the favourite for both matches
0	Won one, lost one of previous two
-1	Lost previous two, but were underdogs for both games
-2	Lost previous two, one by 10 points more than the bookmaker's line Lost previous three matches

Table 6.10: *Number of observations in each confidence class*

Confidence class	-2	-1	0	1	2
Number of observations	1939	819	3348	801	1283

Define $HCONF_k$ and $ACONF_K$ to be the confidence, as calculated above, in match k of the home team and away team and μ_k, λ_k to be the expected home and away score according to the basic model. The following linear model is fit.

$$\begin{aligned} E[HSC_k - \mu_k] &= \alpha_h + \gamma_h * HCONF_k \\ E[ASC_k - \lambda_k] &= \alpha_a + \gamma_a * ACONF_k \end{aligned}$$

For the coefficient γ_h, the estimate and confidence interval are -0.036 (-0.287, 0.215) and for γ_a they are -0.296 (-0.531, -0.061). Recent good results do not affect the home side at all, while again it seems the away side may, if anything, be at a disadvantage as their confidence increases.

Winning streaks

One frequently quoted statistic in media coverage of NBA in the build-up to a fixture is the length of winning streaks of the teams as they enter a fixture, a streak being defined as the number of consecutive victories immediately prior to the fixture. Again, it is conceivable that the market over-reacts to the importance of this short term run of form. To examine this, the length of winning streaks prior to every match in the data set is recorded, and model predictions are compared to the bookmaker's line. Figure

6.6 contains the relevant plot.

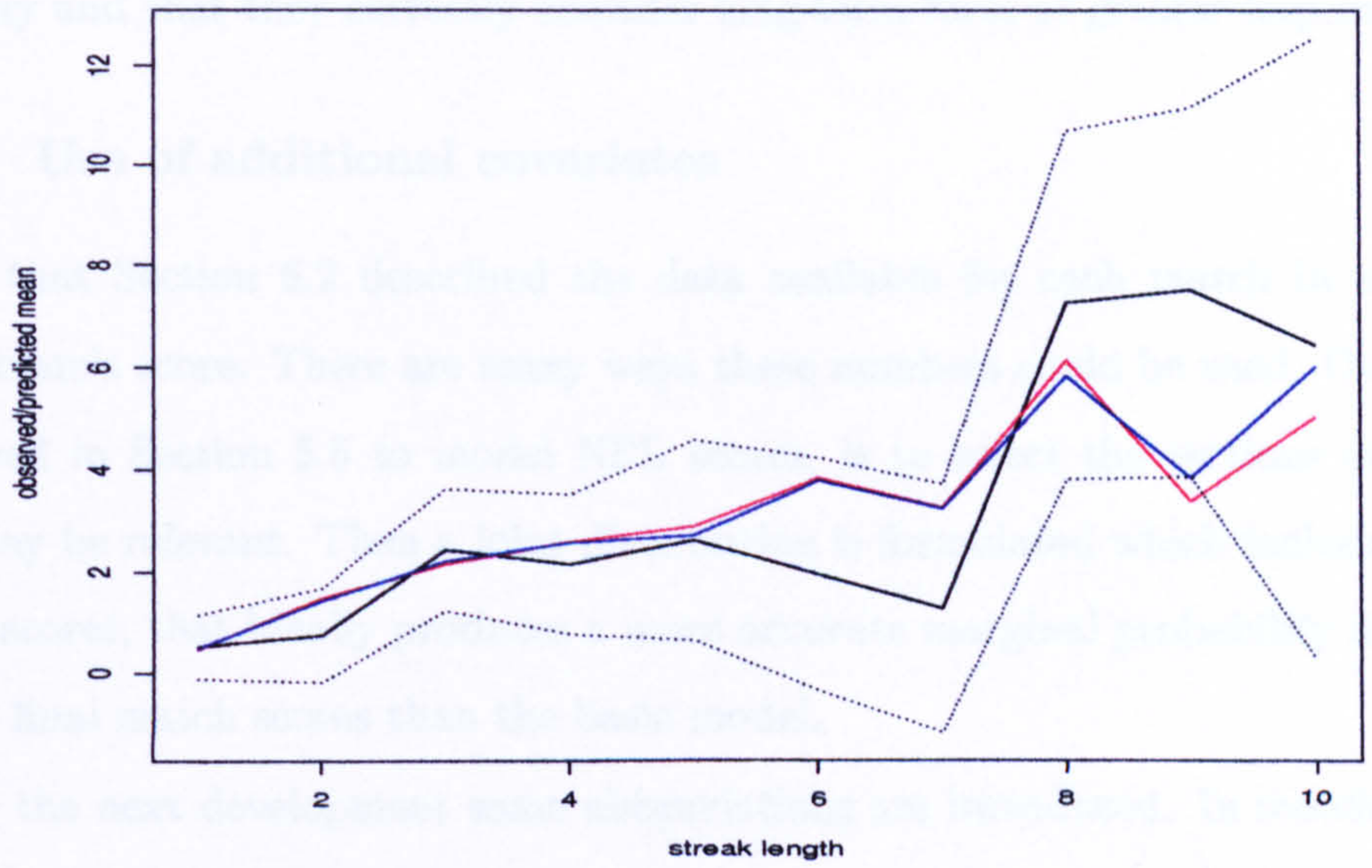


Figure 6.6: Plot of average observed score difference (___) plus confidence intervals (...), model predictions (___) and bookmaker's line (___) against length of winning streak prior to match

In fact, for streak lengths up to around 7, the bookmaker's line is very similar to the predictions from the basic model, which do not explicitly adjust for winning streaks. The differences observed for streak lengths above 7 are based on a small number of matches thus no firm conclusions can be made (only 103 matches from the sample of 4244 matches included in Figure 6.6 featured a side that was on a streak of 8 or more prior to the fixture). This suggests again that the bookmaker's line does not over-emphasise the significance of recent team form.

Short term form: conclusions

Two messages emerge from the investigation of teams' recent form. First, team abilities can be summarised by their long-term ability estimate obtained via the basic model. Factors such as confidence and motivation acquired through very recent runs of form on average do not significantly affect a team's results. There is mild evidence suggesting that a team's results are on average worse when playing away from home given recent good results. However, no adjustments are made to the model to accommodate short-term form.

Secondly the bookmaker's line does not appear to over-estimate the importance of recent good or bad results. The bookmaker's line is largely determined by the

behaviour of the betting market. However, despite the large emphasis that is placed on recent results in media reporting, it seems that gamblers are not susceptible to such publicity and that they correctly consider long-term form of greater importance.

6.4.4 Use of additional covariates

Recall that Section 6.2 described the data available for each match in addition to either team's score. There are many ways these numbers could be used. One method, employed in Section 5.5 to model NFL scores, is to select the sections of this data that may be relevant. Then a joint distribution is formulated which includes the final match scores, that ideally produces a more accurate marginal probability distribution for the final match scores than the basic model.

For the next development some abbreviations are introduced. In match k, for the home and away side respectively,

- HA_k, AA_k are the total shot attempts of any sort,
- $HFTA_k$, $AFTA_k$ are the number of free throw attempts,
- $HFTM_k$, $AFTM_k$ are the number of successful free throw attempts,
- $HF2A_k$, $AF2A_k$ are the number of 2-point attempts,
- $HF2M_k$, $AF2M_k$ are the number of successful 2-point attempts,
- $HF3A_k$, $AF3A_k$ are the number of 3-point attempts,
- $HF3M_k$, $AF3M_k$ are the number of successful 3-point attempts.

For each match k, the following procedure could be implemented.

1. Predict the number of shots, or attempts, of any kind that the home team and away team have, to obtain the distribution of (HA_k, AA_k).

2. Conditional on (HA_k, AA_k), predict how the attempts filter into 3-point attempts, 2-point attempts and free throw attempts. Hence the following two distributions are obtained:

$$(HFTA_k, AFTA_k | HA_k, AA_k) \tag{6.4.5}$$

$$(HF2A_k, AF2A_k | HA_k, AA_k) \tag{6.4.6}$$

from which this is inferred:

$$(HF3A_k, AF3A_k | HA_k, AA_k) \tag{6.4.7}$$

since

$$HA_k = HF2A_k + HF3A_k + HFTA_k$$

$$AA_k = AF2A_k + AF3A_k + AFTA_k$$

3. Conditional on the distributions listed in Equations 6.4.5, 6.4.6 and 6.4.7 a distribution for the number of each type of shot that is converted or *made* is sought.

$$(HFTM_k, AFTM_k | HFTA_k, AFTA_k, HA_k, AA_k)$$

$$(HF2M_k, AF2M_k | HF2A_k, AF2A_k, HA_k, AA_k)$$

$$(HF3M_k, AF3M_k | HF2A_k, AF2A_k, HA_k, AA_k)$$

Using the results of stage 3, the distribution of the final score for each team in match k can be obtained via the identities

$$HSC_k = 2 * HF2M_k + 3 * HF3M_k + HFTM_k$$

$$ASC_k = 2 * AF2M_k + 3 * AF3M_k + AFTM_k$$

An approach like this was applied to the NFL data in Section 5.5, although the multivariate distribution obtained did not match the observed distribution to a satisfactory extent. One reason for this was the unsuitability of the binomial distribution in the situations where a proportion was modelled. In stages 2 and 3 mentioned above, the most obvious way to model teams' decisions concerning which type of shot they have, or their shot conversion rates, is to employ the binomial distribution to model the proportions involved. However, it is worth checking the suitability of this method before implementing any modelling.

To test whether the $HF2M$ are binomially distributed with group size $HF2A$, samples of simulated values of $HF2M$ are generated assuming a binomial distribution. An estimate for the proportion of successful 2-point shots is obtained using a procedure explained below. These simulated samples are then compared with the observed data and the similarity is compared. To clarify the procedure:

1. for each match k between side $i(k)$ and $j(k)$, generate an estimate of the ratio $E[\frac{HF2M_k}{HF2A_k}]$ which is the mean of the average 2-point shot conversion rate for team $i(k)$ and the average conversion rate that $j(k)$ allow their opponents, during the last 500 NBA matches. This accounts for approximately half an NBA season and includes approximately 40 data points for each team.

2. for each match k, $k \in [1, N]$ where N is the total number of matches, simulate Z values, $HF2M^*_{k1}, \ldots, HF2M^*_{kZ}$ of home 2-point made shots, using a binomial distribution, given the observed number $HF2A_k$ as group size and the approximate conversion rate from stage 1 as the probability parameter. This yields Z samples of N simulated values of $HF2M$.

3. compare quantiles of the variance of samples $\mathbf{HF2M^*_1, \ldots, HF2M^*_Z}$ with the observed $Var(HF2M)$.

This same method can be applied to obtain simulated values of $AF2M$ as a proportion of $AF2A$ as well as

- $(HF2A + HF3A)$ as a proportion of HA and $(AF2A + AF3A)$ as a proportion of AA,
- $HF2A$ as a proportion of $(HF2A + HF3A)$ and $AF2A$ as a proportion of $(AF2A + AF3A)$,
- $HF3M$ as a proportion of $HF3A$ and $AF3M$ as a proportion of $AF3A$,
- $HFTM$ as a proportion of $HFTA$ and $AFTM$ as a proportion of $AFTA$.

Table 6.11 displays the results.

Table 6.11: Comparison of observed variance for variables, with simulated values assuming binomial distribution

Variable	*Observed Variance*	*Simulated 2.5% quantile*	*Simulated 97.5% quantile*
HF2A+HF3A	59.376	53.046	55.718
AF2A+AF3A	58.536	53.739	56.568
HF2A	70.098	59.055	61.967
AF2A	72.187	59.695	62.754
HF2M	28.867	30.563	32.974
AF2M	25.088	30.878	33.229
HF3M	6.847	6.343	6.894
AF3M	6.449	6.261	6.804
HFTM	42.268	40.838	42.598
AFTM	38.249	37.666	39.386

The observed values of $HF2A + HF3A$ and $AF2A + AF3A$ are over-dispersed compared to the samples simulated assuming a binomial response, as are the observed values of $HF2A$ and $AF2A$. Meanwhile the observed values of $HF2M$ and $AF2M$ are under-dispersed compared to the simulated samples. The beta-binomial distribution, described in Section 5.5.4 may be a more suitable response distribution than the binomial distribution where the simulations are under-dispersed. Given the computational difficulties involved in implementing a likelihood based procedure using a negative binomial distribution, this approach has not been considered in this investigation. Hence a full multivariate analysis of the type carried out in Section 5.5 for the modelling of NFL scores is a topic for possible further research.

6.4.5 Increasing levels of team parameterisation

Section 5.8.2 described a simple method that can offer some guidance on the appropriate number of parameters to include for each team in a model of NFL scores. For that application it appeared that two parameters was sufficient. Recall that firstly it is tested whether a better fit of the past data can be obtained by the inclusion of extra parameters to represent the teams' abilities. Secondly, in order to verify that extra predictive power can be gained from the extra team parameters (and that the better fit of past data is not obtained by modelling random error) it is investigated whether the prediction of team abilities from one season can be made using the fitted team abilities of the previous season (which are modelled on an entirely separate data set). The results of these tests, applied to NBA data, are displayed in Tables 6.12 and 6.13. It appears from 6.13 that the generally significant p-values of the most highly parameterised model that genuine effects rather than random error are being modelled with the extra parameters. Hence extra predictive power may be available by including four rather than two parameters for each team in the model.

Alternatively, as a possible area for further research, rather than modelling team abilities by allocating a set number of parameters for each team, other approaches could be taken to examine in what circumstances teams' strategies alter for different fixtures. One could examine in detail the effect on scoring and conceding rates of a big difference in ability between the two teams or the importance of the match, for example. It is possible that there are more efficient systems of summarising team abilities and how they vary their tactics from match to match than including extra parameters for each team.

Table 6.12: *Decrease, and significance of decrease, of deviance when additional team parameters are added into NBA model, season 1997/98 to 2000/01.*

Model number	Parameters in model	Comparison model	Year	Deviance reduction (df, p-value)
1	γ	-	97/98	-
			98/99	-
			99/00	-
			00/01	-
				-
2	γ,δ	1	97/98	4996.56 (1,0)
			98/99	4197.3 (1,0)
			99/00	7273.88 (1,0)
			00/01	5142.56 (1,0)
3a	γ,δ,α_i	2	97/98	33592.26 (28,0)
			98/99	19538.1 (28,0)
			99/00	35699.3 (28,0)
			00/01	29867.96 (28,0)
3b	γ,δ,β_i	2	97/98	47238.94 (28,0)
			98/99	30955.86 (28,0)
			99/00	40911.45 (28,0)
			00/01	34350.11 (28,0)
4	$\gamma,\delta,\alpha_i,\beta_i$	3a, 3b	97/98	46807.91 (28,0) , 33161.22 (28,0)
			98/99	27311.86 (28,0) , 15894.1 (28,0)
			99/00	42625.87 (28,0) , 37413.72 (28,0)
			00/01	34261.32 (28,0) , 29779.17 (28,0)
5a	$\gamma,\delta,\alpha_i,\beta_i,\lambda_i$	4	97/98	3148.84 (28,0)
			98/99	5311.63 (28,0)
			99/00	2908.18 (28,0)
			00/01	3379.12 (28,0)
5b	$\gamma,\delta,\alpha_i,\beta_i,\mu_i$	4	97/98	4173.95 (28,0)
			98/99	4816.92 (28,0)
			99/00	3333.23 (28,0)
			00/01	4411.47 (28,0)
6	$\gamma,\delta,\alpha_i,\beta_i,\lambda_i,\mu_i$	5a, 5b	97/98	4010.02 (28,0) , 2984.91 (28,0)
			98/99	4176.44 (28,0) , 4671.14 (28,0)
			99/00	3247.5 (28,0) , 2822.44 (28,0)
			00/01	4155.02 (28,0) , 3122.67 (28,0)

6.4.6 Inclusion of player information

There is little doubt that the identity of players participating in a match influences the performance of the team. Figure 6.3, which plots the model predictions against the bookmaker's line, reveals several points where the model and the bookmaker disagree. Table 6.14 displays all matches where the model and bookmaker's line for a match differ by more than 8 points.

Many disagreements occur at the start of seasons, for example, at the start of the

Table 6.13: *Coefficients and p-values obtained using previous year's parameters to predict next year's, for NBA, 1997/98 to 2000/01*

Parameters in model	Regression applied	Coefficient and p-value of previous year's parameter α_i	β_i	γ_i	δ_i
γ	2~1				
	3~2				
	4~3				
γ,δ	2~1				
	3~2				
	4~3				
γ,δ,α_i	2~1	**0.9,0**			
	3~2	**1.03,0**			
	4~3	**0.94,0**			
γ,δ,β_i	2~1		**1,0**		
	3~2		**1,0**		
	4~3		**0.97,0**		
$\gamma,\delta,\alpha_i,\beta_i$	2~1	**0.9,0**	**0.53,0**		
	3~2	**1.02,0**	**0.49,0**		
	4~3	**0.94,0**	**0.62,0**		
$\gamma,\delta,\alpha_i,\beta_i,\lambda_i$	2~1	**0.91,0**	**0.54,0**	**0.42,0.02**	
	3~2	**1.02,0**	**0.44,0.01**	**0.34,0.01**	
	4~3	**0.96,0**	**0.66,0**	**0.37,0.01**	
$\gamma,\delta,\alpha_i,\beta_i,\mu_i$	2~1	**0.91,0**	**0.42,0**		0.24,0.25
	3~2	**1.02,0**	**0.43,0.03**		0.05,0.77
	4~3	**0.94,0**	**0.59,0**		**0.58,0**
$\gamma,\delta,\alpha_i,\beta_i,\lambda_i,\mu_i$	2~1	**0.91,0**	**0.44,0.01**	0.31,0.08	**0.67,0.01**
	3~2	**1.02,0**	0.35,0.06	**0.41,0.01**	0.14,0.21
	4~3	**0.96,0**	**0.63,0**	**0.45,0**	**0.42,0.04**

Table 6.14: List of matches with big differences between bookmaker's line and model predictions

Date	*Home team*	*Away team*	*Model prediction*	*Bookmaker's prediction*	*Observed score difference*
19980124	Toronto	Minnesota	-4.29	5	0
19980215	Sacremento	Washington	-2.02	6	2
19980224	Washington	Houston	4.52	-5.5	12
19990205	Utah	Chicago	0.02	15	8
19990206	Charlotte	Milwaukee	2.64	-5.5	0
19990206	Golden State	Houston	1.74	-6.5	-2
19990208	Charlotte	Miami	1.49	-8.5	3
19990209	Chicago	Atlanta	4.94	-4.5	-16
19990210	Charlotte	Cleveland	3.85	-5.5	-10
19990211	Chicago	New York	4.8	-6	-5
19990216	LA Lakers	Charlotte	6.46	16	28
19990218	Indianapolis	Philadelphia	1.17	10	4
19991107	LA Lakers	Dallas	3.49	11.5	8
20000319	Golden State	Phoenix	-8.95	8	-17
20001104	Vancouver	LA Lakers	1.27	-9	-9
20001106	Sacremento	Portland	6.53	-2.5	4
20001112	Detroit	Seattle	6.57	-2	9
20001113	New Jersey	Portland	2.74	-6.5	-12
20001116	Sacremento	LA Lakers	6.85	-2	0
20001116	Toronto	Portland	5.74	-2.5	-6
20001127	LA Clippers	LA Lakers	-3.26	-11.5	-15
20001205	LA Lakers	Philadelphia	-0.15	8	11
20001221	Houston	LA Lakers	1.71	-8	-5

98/99 season (which started 5 February 1999 due to a players' strike), and the 00/01 season, which started on 31 October 2000. Between seasons clubs buy players, sell players, or players retire. Hence the roster of a team can change significantly between the end of one season and the start of the next. The bookmaker's lines generally consider such information. The procedure used in order to obtain parameter estimates for the basic model places more weight on recent results, hence information from previous seasons for all clubs is down-weighted. However it does not make adjustments for specific changes to a squad such as this. Unfortunately data concerning which players have participated in each match was not available during this study. It is a possible, and almost certainly worthwhile, area of further research.

6.5 Construction of more advanced model

On consideration of Sections 6.4.1, 6.4.2 and 6.4.5, the following construction will be implemented in order to seek a more effective NBA scores model:

$$(HSC3_k, ASC3_k) \sim \mathcal{N}_2(\mu_{h3k}, \mu_{a3k}, \sigma_{h3}, \sigma_{a3}, \rho_3) \qquad (6.5.1)$$

where

- $HSC3_k, ASC3_k$ are the scores at the end of the third quarter
- $\mu_{h3k} = \gamma + \delta + \alpha^h_{i(k)} + \beta^a_{j(k)} + \lambda_h ATIRED_k$,
  $\mu_{a3k} = \gamma + \alpha^a_{j(k)} + \beta^h_{i(k)} + \lambda_a HTIRED_k$
- γ is the global intercept
- δ is the home effect
- $\alpha^h_{i(k)}$, $\beta^h_{i(k)}$ are team $i(k)$'s offensive and defensive parameters while playing at home
- $\alpha^a_{j(k)}$, $\beta^a_{j(k)}$ are team $j(k)$'s offensive and defensive parameters while playing away from home
- $ATIRED_k$ is an indicator variable set to 1 if team $j(k)$ played a fixture the previous day, while $HTIRED_k$ is similarly defined for team $i(k)$. λ_h is the coefficient that expresses the average increase in points for the home side if the $ATIRED_k$ is equal to 1 and λ_a is similarly defined for the away side if the $HTIRED_k$ is equal to 1
- $\sigma_{h3}, \sigma_{a3}, \rho_3$ are the home standard deviation, away standard deviation and correlation coefficient of all third quarter final scores.

Then, the team abilities obtained from the above formulation are incorporated into a simple linear model, and treated as constants, to produce a model for final quarter scores $Q4HSC_k$ and $Q4ASC_k$.

$$(Q4HSC_k, Q4ASC_k) \sim \mathcal{N}_2(\mu_{hq4k}, \mu_{aq4k}, \sigma_{hq4}, \sigma_{aq4}, \rho_{q4}) \qquad (6.5.2)$$

where

- $\mu_{hq4k} = \gamma_4 + \delta_4 + \nu(\alpha^h_{i(k)} + \beta^a_{j(k)}) + \kappa(HSC3 - ASC3)$
  $\mu_{aq4k} = \gamma_4 + \nu(\alpha^a_{j(k)} + \beta^h_{i(k)}) + \kappa(ASC3 - HSC3)$
- γ_4 is the global intercept for fourth quarter scores
- δ_4 is the home effect for fourth quarter scores

- $\sigma_{hq4}, \sigma_{aq4}, \rho_{q4}$ are the home standard deviation, away standard deviation and correlation coefficient of fourth quarter scores

The following linear model is implemented to verify if the tiredness of the two teams affects the difference in score in the final quarter:

$$Q4HSC - Q4ASC \sim \beta_0 + \beta_1(HSC3 - ASC3) + \beta_2 HTIRED + \beta_3 ATIRED$$

where the $HSC3 - ASC3$ term is considered as a nuisance parameter since it is known that it is a strong predictor of $Q4HSC - Q4ASC$. The estimates and confidence intervals for the β_2 and β_3 terms are -0.456 (-1.055,0.143) and -0.118 (-0.556,0.319). There is no conclusive evidence that if a team plays on the day prior to a match, their score in the final quarter is affected, so a tiredness indicator vector is not included in the second model.

Ideally, a four-degree multivariate Normal distribution would be used in order to estimate all the parameters from the above models simultaneously since multiple regression of these four variables on each other reveals a strong dependence between them. However, due to the computational complexity involved in doing so, two independent bivariate Normal distributions are used and team parameters are estimated only through the first model. The estimates for the four team parameters are displayed in Table 6.15.

The similarity between the estimates for teams' parameters for their home games and their away games is not surprising given that it is the same players who participate in these games. The tactics may vary according to whether the team plays at home or away so an exact agreement is unlikely. The two sets of parameters are plotted both for the offensive parameters and the defensive parameters in Figure 6.7.

The estimated probability for the distribution of final scores can then be calculated using these two models.

6.5.1 Adjustment for overtime periods

The model development so far has focused on generating probabilities for the events $P(HSC - ASC)$ and $P(HSC + ASC)$ where HSC and ASC are the home and away score of a match at the conclusion of the fourth quarter of play. For betting purposes, it is the final score after possible overtimes that is of interest. Define

- $DSC4$ and $TSC4$ to be the difference in score and total score at the conclusion

Table 6.15: NBA team ability estimates for home offense ($\hat{\alpha^h}$), away offense ($\hat{\alpha^a}$), home defense ($\hat{\beta^h}$) and away defense ($\hat{\beta^a}$), June 2001

team	$\hat{\alpha^h}$	rank	$\hat{\alpha^a}$	rank	$\hat{\beta^h}$	rank	$\hat{\beta^a}$	rank
Portland	0.4427	13	0.1952	17	-3.2744	4	-3.5704	6
Boston	0.2986	16	1.3537	9	1.542	20	2.3514	23
Vancouver	-2.9833	26	-0.333	19	2.1558	24	2.8062	26
Miami	-2.4187	25	-3.3622	26	-3.8497	3	-5.8169	1
Charlotte	-0.6289	21	0.6117	13	-1.8768	6	-1.9093	8
LA Lakers	5.2324	2	3.2522	3	-0.1579	12	-1.4951	9
Orlando	2.0383	6	0.673	12	1.0169	15	-0.0434	12
New Jersey	1.2115	9	0.3824	15	3.0275	27	2.6821	24
Denver	1.081	10	0.3537	16	2.8064	26	4.6767	29
Detroit	0.9941	11	2.146	5	1.6314	21	1.6099	18
Houston	1.4774	8	1.0948	11	1.1891	16	0.1263	13
Philadelphia	-0.5255	19	-1.3798	22	-1.4715	8	-3.8216	5
Phoenix	0.1989	17	-0.6932	20	-1.5813	7	-2.8021	7
Minnesota	0.4542	12	1.7545	6	-1.3118	9	0.3479	14
Milwaukee	2.9362	3	4.161	2	1.1893	17	2.3237	22
Chicago	-6.739	29	-3.4667	27	-0.7823	11	0.758	15
LA Clippers	-3.4005	28	-1.452	23	1.9811	22	2.2405	21
Atlanta	-0.6606	22	-3.6421	28	1.4873	19	-0.0856	10
Utah	1.4809	7	-1.2072	21	-1.9445	5	-3.9773	4
Indianapolis	0.3686	14	1.0955	10	-1.0011	10	-0.0698	11
Seattle	2.6886	4	2.6519	4	0.3365	14	2.2309	20
San Antonio	-0.2353	18	0.4682	14	-6.0048	1	-4.5643	2
Washington	0.3143	15	-0.1212	18	3.9535	28	2.7331	25
Sacremento	5.8617	1	4.4384	1	2.194	25	4.1775	27
New York	-3.1018	27	-4.2615	29	-5.0601	2	-4.5078	3
Cleveland	-2.1034	24	-1.8492	24	0.1477	13	1.8894	19
Dallas	2.1973	5	1.6196	8	2.1116	23	0.8484	16
Toronto	-0.5286	20	1.6933	7	1.298	18	0.8781	17
Golden State	-1.3266	23	-2.0003	25	4.4248	29	4.6099	28

of the fourth quarter,

- DSC and TSC to be the difference in score and total score after overtimes are completed and
- $OTDSC_n, OTTSC_n$ to be the difference in score and total score in the nth overtime period,

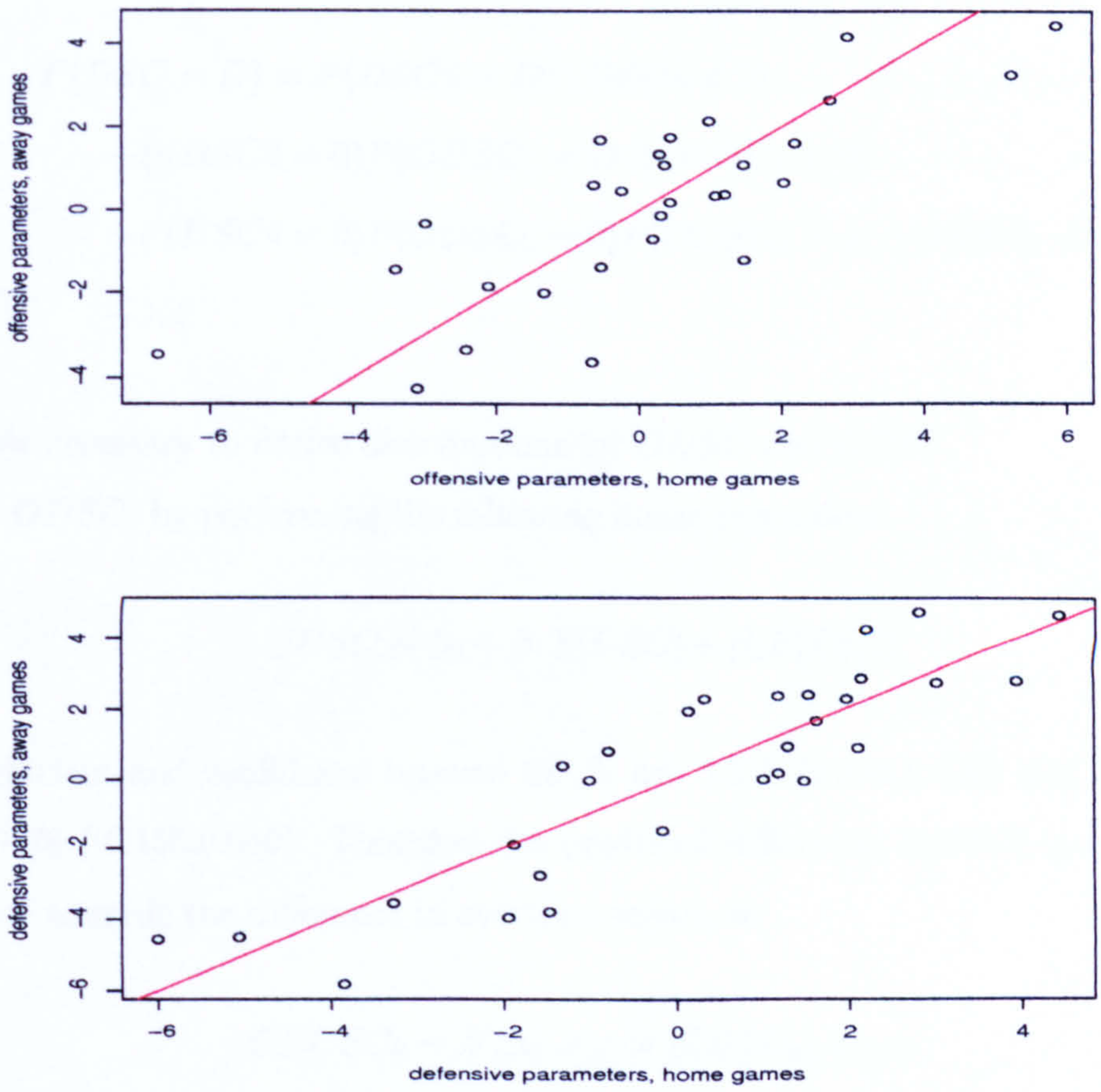


Figure 6.7: Plot of offensive parameters for home games of NBA teams at final time-point of data set against offensive parameters for away games at final time-point

the following formulae give the distributions for DSC and TSC:

$$
\begin{aligned}
P(TSC = T) &= P(TSC4 = T \cap DSC4 \neq 0) \\
&+ \sum_{y_1=0}^{T-1} P(TSC4 = y_1 \cap DSC_4 = 0) P(OTSC_1 = T - y_1 \cap ODSC_1 \neq 0) \\
&+ \sum_{y_1=0}^{T-1} \sum_{y_2=0}^{T-y_1-1} P(TSC4 = y_1 \cap DSC4 = 0) P(OTSC_1 = y_2 \cap ODSC_1 = 0) \\
&\quad * P(OTSC_2 = T - y_1 - y_2 \cap ODSC_2 \neq 0) \\
&+ \ldots
\end{aligned}
$$

and

$$P(DSC = D) = P(DSC4 = D \cap DSC4 \neq 0)$$
$$+P(DSC4 = 0)P(ODSC_1 = D \cap ODSC_1 \neq 0)$$
$$+P(DSC4 = 0)P(ODSC_1 = 0)P(ODSC_2 = D \cap ODSC_2 \neq 0)$$
$$+\ldots$$

It is now necessary to derive distributions for $ODSC$ and $OTSC$.

For $ODSC$, by performing the following linear regression

$$ODSC \sim \beta_0 + \beta_1 E[DSC] + \beta_2 E[TSC]$$

the coefficient and confidence interval for β_1 are 0.315 (0.190,0.440) and for β_2 they are -0.0478 (-0.155,0.060). Therefore the predicted difference $E[DSC]$ is a significant predictor towards the difference in overtime scores, so

$$ODCSC_k \sim \mathcal{N}(\mu_0 + \mu_1 * E[DSC_k], \sigma_{OD}) \tag{6.5.3}$$

where $\mu_0, \mu_1, \sigma_{OD}$ are all evaluated using only data observed prior to match k. For the final match, $(\mu_0, \mu_1, \sigma_{OD}) = (-0.198, 0.268, 4.749)$.

By implementing the following linear regression:

$$OTSC \sim \beta_0 + \beta_1 E[DSC] + \beta_2 E[TSC]$$

the coefficients and confidence intervals for β_1 and β_2 are 0.082 (-0.158,0.322) and 0.188 (-0.0181,0.395). Since neither $E[DSC]$ nor $E[TSC]$ are significant predictors for $OTSC$

$$OTCSC_k \sim \mathcal{N}(\overline{OTSC}, \sigma_{OT}) \tag{6.5.4}$$

where again only data observed prior to match k are used. For the final match, $(\overline{OTSC}, \sigma_{OT}) = (21.008, 6.995)$.

6.6 Comparison of basic model and advanced model

Figure 6.8 plots the predicted mean of the difference in score from the basic model of Section 6.3 against the predicted mean of the difference in score using the more

advanced model constructed in Section 6.5 and similarly for the prediction of the total scores. While there is a broad agreement between the matches, there are also many matches whose predictions have changed greatly. Figure 6.9 displays a plot of moving average of predictions against observed values for the predicted values of the difference in score and the total score according to both models. It reveals that both model's predictions are generally reliable.

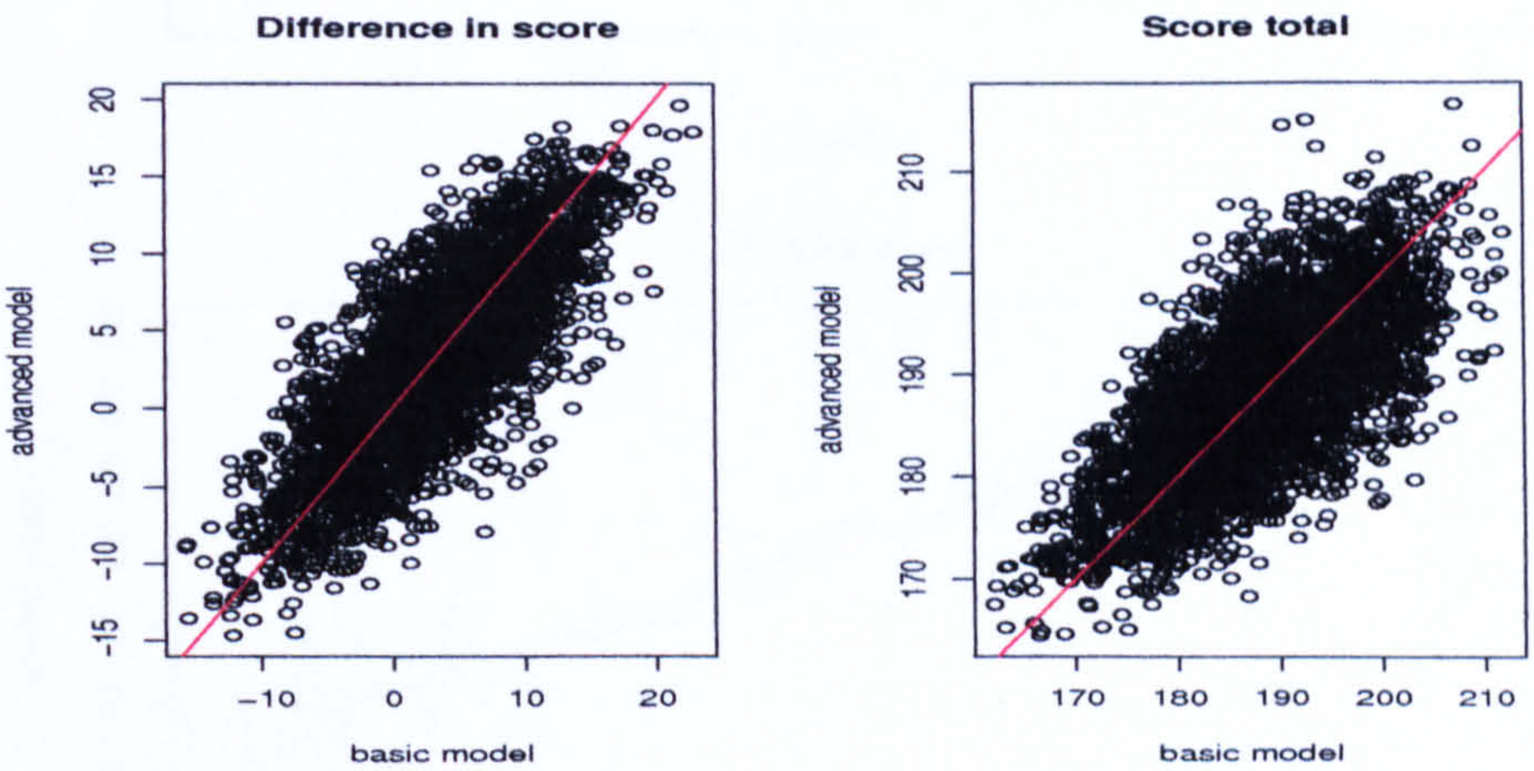


Figure 6.8: Plot of basic model score predictions against advanced model

6.6.1 Summary statistics

Two measures are used in order to compare the predictive ability of the basic model with the more advanced model of Section 6.5. The first measure, which uses scores at the end of the final quarter, counts the number of times that either model's mean prediction is closer to the final score. The second measure calculates the average predictive log-likelihood for each match. Table 6.16 displays the results of applying these measures.

Table 6.16: Comparison of basic model and quasimultivariate model via summary statistics

	Score difference		*Score total*	
	Basic model	*Advanced model*	*Model 1*	*Model 2*
Proportion closer	0.526	0.474	0.54	0.46
Mean loglikelihood	-3.923	-3.887	-4.246	-4.308

The first measure, which does not penalise the magnitude of difference between prediction and result, reveals that the basic model is closer to the observed result slightly more often.

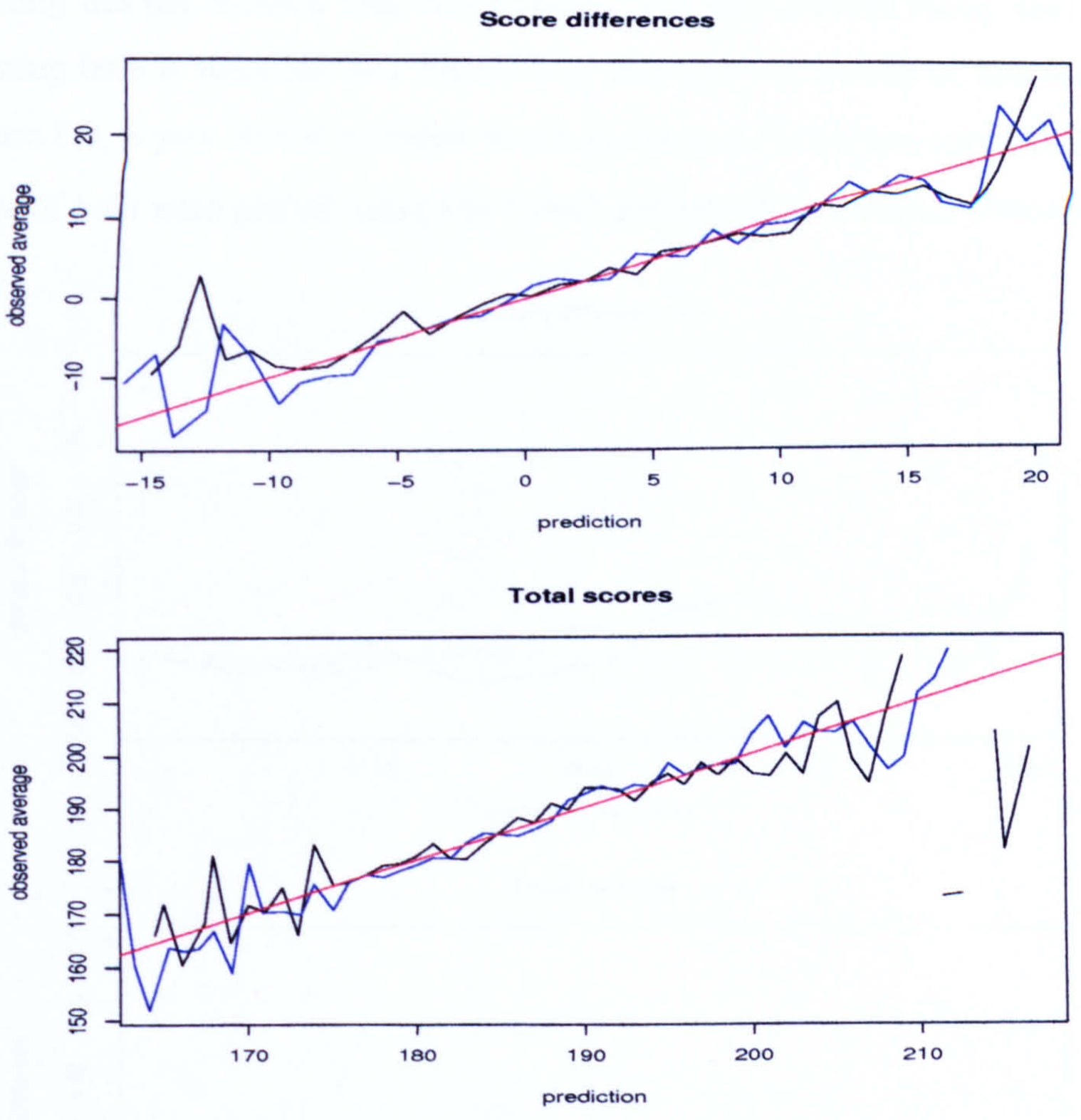


Figure 6.9: Moving average plots of predicted score differences and totals, for basic model (___) and advanced model(___)

The second comparison, which does penalise the magnitude of discrepancy, suggests the advanced model produces slightly better predictions for score differences but slightly inferior predictions for totals.

It should be noted that the score differences and totals have been modelled as a bivariate distribution. However, due to the complex procedure required to obtain predictions for final scores described in Section 6.5.1 it is not straightforward to calculate the loglikelihood of the joint (DSC, TSC) distribution for the advanced model. Hence Table 6.16 displays only marginal loglikelihoods.

6.6.2 Betting success

The betting strategy adopted here for NBA matches is similar to that outlined in Section 5.4.2 for betting on NFL matches. The probability of winning each bet offered by the bookmaker, either on the score difference or total score, is calculated. Then various cut-off values are chosen such that bets are only placed provided the probability

of winning the bet exceeds this cut-off value. For each cut-off value, the proportion of winning bets is recorded and Figure 6.10 displays the results of this strategy. As in Figure 5.9, a y=x line is included which represents the return curve that would be obtained if bets were placed using the "true" probabilities of match outcomes.

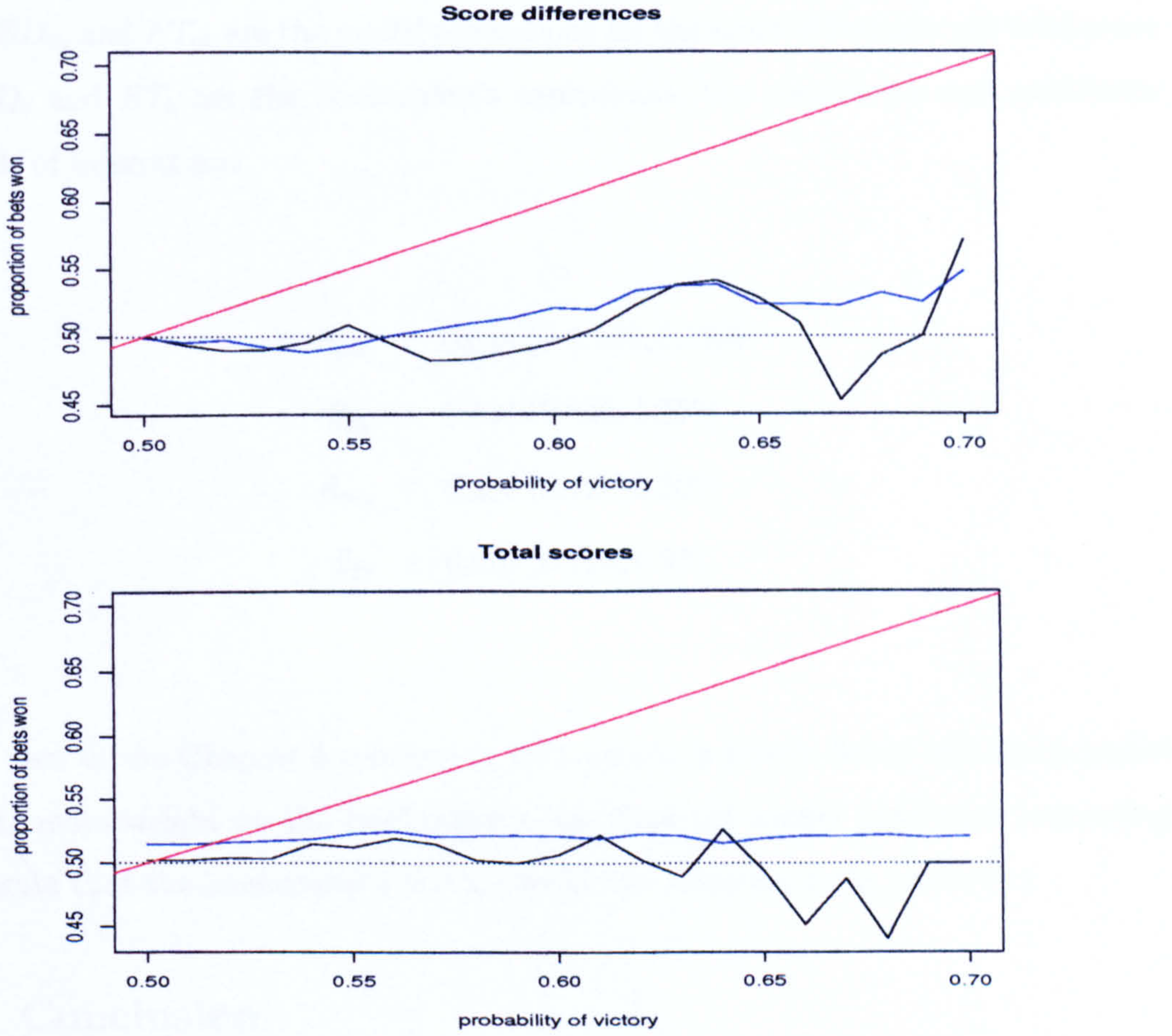


Figure 6.10: Proportions of bets won, where bet is made provided P(Win)$\geq$ cut-off, according to both the basic model (___) and the advanced (___).

It is clear that neither model produces predictions that can win bets against the bookmaker on a consistent basis. The horizontal dotted line is the 50% level, which represents the proportion of victories one would achieve by betting randomly. The advanced model is generally superior to the basic model, however, in order to make money by betting with the bookmaker in the long term, it is necessary to have bets that win often enough to exceed the bookmaker's overround. As mentioned in Section 5.4.2, this requires winning significantly more often than 52.5% of all bets. Unfortunately, even the advanced model does not produce predictions that win often enough.

A similar conclusion is reached by repeating the general linear model test previously performed on NFL scores in Section 5.6.1. After fitting the following two models

$$E(DSC) \sim \alpha_d + \beta_{dm} ED_m + \beta_{db} ED_b \quad (6.6.1)$$

$$E(TSC) \sim \alpha_t + \beta_{tm} ET_m + \beta_{tb} ET_b \quad (6.6.2)$$

where ED_m and ET_m are the model predictions for the score difference and total score and ED_b and ET_b are the bookmaker's equivalents, the coefficients and confidence intervals of interest are:

$$\beta_{dm} : 0.014\ (-0.101, 0.129)$$

$$\beta_{db} : 0.937\ (0.834, 1.039)$$

$$\beta_{tm} : 0.218\ (0.127, 0.309)$$

$$\beta_{tb} : 0.857\ (0.784, 0.93)$$

As seen in the Chapter 5 concerning NFL scores, a simple linear regression model puts far more weight on the bookmaker's line than the model prediction suggesting once again that the bookmaker's line is overall the more accurate prediction.

6.7 Conclusion

A basic model has been presented that produces predictions that are reasonably similar to those offered by the bookmaker. This model makes some assumptions that need to be relaxed so a more advanced model is introduced to reflect several effects. These include the tiredness teams feel as a result of playing a game on the previous day, the tendency for teams to relax should they be leading towards the end of a game and teams' different strategies for home and away fixtures. A new model is formulated to account for these effects and the success of a theoretical betting procedure based on odds offered by a professional bookmaker is evaluated. However, neither model produces predictions that are good enough to make a profit on a consistent basis.

It is likely that some of the assumptions that are still made by the advanced model, in particular that the players being fielded for a fixture does not affect the average result, are not being made by the bookmaker. While for many matches the

advanced model may, for example, correctly identify a 55% probability of winning a handicap bet, in other matches where the bookmaker has more accurate odds, then the probability of winning a bet drops to approximately 50% since the model is effectively placing a random bet. Thus the 55% success of the accurate bets is being averaged out with the 50% success rate of the inaccurate ones, putting a ceiling on the success rate of the betting strategies that have been considered.

Chapter 7

An alternative estimation method - Markov Chain Monte Carlo

While the MLE procedure used throughout this thesis is relatively simple to implement, from a statistical point of view it is a little unattractive. In particular, using the same team ability parameters for every match in which a team plays when parameters are estimated is not ideal even if the importance of older matches is down-weighted. It is more desirable to allow the team parameters to adjust over time by specifying some dynamic distribution for them. As emphasised by Equation 2.4.1, this does complicate the likelihood function considerably. The increase in parameters caused by this extension to the modelling procedure means that the numerical routines used in order to obtain parameters estimates used so far in this thesis are no longer appropriate.

There is an alternative method that can be used to obtain estimates of the parameters in a situation such as this. Instead of performing analytical evaluation of the likelihood function, inference on the parameters can be made by simulating a large number of samples from a posterior distribution of the parameter values given observed data. In this way, a large number of realisations of the values of parameters can be obtained and from these, estimates of the means, modes and variances of the parameter in the model can be made.

The main difficulty associated with this method is that of obtaining the simulated

samples. The posterior distribution of a set of parameters θ given observed data X is

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)} = \frac{P(\theta)P(X|\theta)}{\int_\theta P(\theta)P(X|\theta)d\theta} \tag{7.0.1}$$

where $P(\theta)$ is a prior distribution of θ and $P(X|\theta)$ is the likelihood of the data given θ. A likelihood function that includes separate parameters of each team at each time-point would be of very large dimension which would make evaluation of the integral in Equation 7.0.1 unfeasible. However, although the posterior distribution of the parameters cannot be specified in a closed form, a technique known as *Markov Chain Monte Carlo* (MCMC) can be used in order to simulate from it. A thorough understanding of the MCMC technique for simulation is not necessary in order to understand this chapter, however a brief summary is given in Section 7.6. There are several implementations of the MCMC technique, one of the most popular of which is the Gibbs Sampler. A Gibbs Sampling based MCMC program known as WinBUGS is used in order to perform the proceeding analysis.

By employing MCMC, it is possible to obtain parameter estimates for a sports model including genuinely dynamic team parameters. The advantages and disadvantages of this approach are demonstrated by an example. The market that analysed is NFL scores for seasons 1997/98 until 2000/01. The model specification is similar to that used in the basic model in Chapter 5 concerning NFL, although some modifications are made on consideration of Glickman and Stern's (1998) NFL model and Rue and Salvesen's (1997) soccer model. As outlined by Gilks *et al* (1995), the task of specifying a full probability model can be divided into the three stages:

1. Specification of model quantities and the dependency structure between them
2. Specification of the parametric form of direct relationships
3. Prior specifications

Each of these steps is now applied.

7.1 Specification of model quantities and the dependency structure between them

These are the relevant quantities used in the model:

- X_k and Y_k represent the home and away score in match k.

- μ_k and λ_k represent the mean home and away scoring level of match k
- κ represents the precision of X_k and Y_k conditional on μ_k and λ_k
- $\alpha_{i,t}$ and $\beta_{i,t}$ represent the offensive and defensive abilities of team i at time-point t
- γ represents the global mean for all games
- δ represents the benefit of playing at home
- The $\alpha_.$ and $\beta_.$ terms follow a Brownian motion, with drift precision τ_w between each time-point during a season and τ_s between the final time-point of one season and the first time-point of the following season.

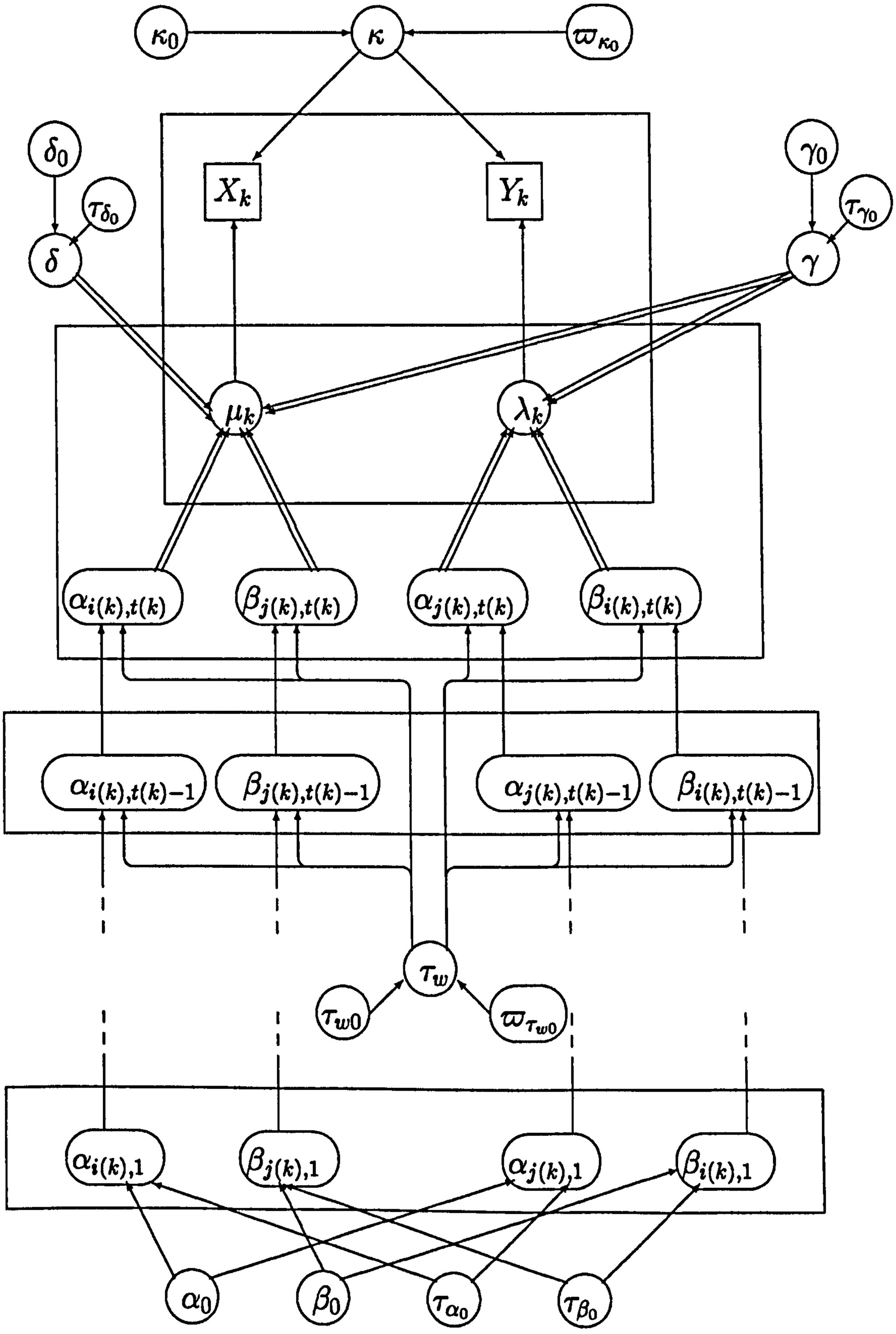
The form chosen for μ_k and λ_k for match k between teams $i(k)$ and $j(k)$ taking place at time $t(k)$ are:

$$\begin{aligned} \mu_k &= \gamma + \alpha_{i(k),t(k)} + \beta_{j(k),t(k)} + \delta \\ \lambda_k &= \gamma + \alpha_{j(k),t(k)} + \beta_{i(k),t(k)} \end{aligned} \tag{7.1.1}$$

Figure 7.1 is a cut-down *Directed Acyclic Graph* (DAG) representing these relationships, including only the parameters relevant to match k.

The DAG is a useful method for displaying the specification of Bayesian models, where parameters are considered to be random quantities, and are thus specified using a probability distribution. A single arrow that points from one quantity to another indicates that the probability distribution of the second quantity is some function of the first quantity. For example, the distribution for the δ term in Figure 7.1 depends on both δ_0 and τ_{δ_0}. Double arrows are used where several quantities combine to be re-expressed in a single quantity, in order to aid the presentation of the model. This is the case for the μ_k and λ_k terms, which are defined by Equation 7.1.1. The rectangular boxes in the DAG do not define anything with regard to the model specification but they are included in order to clarify to the user that the groups of quantities within the rectangle are considered to have a similar role in the model. In this case, it is helpful to group all team parameters that refer to the same time-point within one rectangle.

Figure 7.1: Cut-down *Directed Acyclic Graph* representing relationship between parameters of NFL model



7.2 Specification of the parametric form of direct relationships

The distribution of each variable and the relationship between each variable in the cut-down DAG is explained briefly.

- The scores X_k and Y_k are assumed to be independent and normally distributed with precision κ. Note that the *precision* as an alternative to the variance is used throughout this chapter, in order to be consistent with the WinBUGS nomenclature.

$$\begin{aligned} X_k &\sim \mathcal{N}(\mu_k, \kappa^{-\frac{1}{2}}) \\ Y_k &\sim \mathcal{N}(\lambda_k, \kappa^{-\frac{1}{2}}) \end{aligned}$$

- As mentioned above, standard Brownian motion is used to model the variation of a team's offensive ability. If t and $t+x$ are time-points in the same NFL season then

$$\alpha_{i,t+x} \sim \mathcal{N}(\alpha_{i,t}, (x/\tau_w)^{\frac{1}{2}})$$

  If $t+x$ is the time-point of the first fixture of one season and t is the time-point of the final fixture of the previous season then

$$\alpha_{i,t+x} \sim \mathcal{N}(\alpha_{i,t}, \tau_s^{-\frac{1}{2}})$$

  The teams' defensive abilities are modelled similarly.

- It is necessary to determine a prior mean and precision for the values of the α and β terms before any data is observed thus

$$\begin{aligned} \alpha_{.,1} &\sim \mathcal{N}(\alpha_0, \tau_{\alpha_0}) \\ \beta_{.,1} &\sim \mathcal{N}(\beta_0, \tau_{\beta_0}) \end{aligned}$$

- It is also necessary to determine a distributional form and relevant prior values for the other parameters in the model. It is assumed the global mean and home

effect parameters are normally distributed so

$$\begin{aligned} \gamma &\sim \mathcal{N}(\gamma_0, \tau_{\gamma_0}) \\ \delta &\sim \mathcal{N}(\delta_0, \tau_{\delta_0}) \end{aligned}$$

For precision parameters a conjugate prior is a gamma distribution with appropriate shape and scale parameters hence

$$\begin{aligned} \kappa &\sim \Gamma(\kappa_0, \varpi_\kappa) \\ \tau_w &\sim \Gamma(\tau_{w0}, \varpi_{w0}) \\ \tau_s &\sim \Gamma(\tau_{s0}, \varpi_{s0}) \end{aligned}$$

Note that there is insufficient space in Figure 7.1 to include the τ_s term. If the time difference between $t(k) - n + 1$ and $t(k) - n$ corresponds to a season break then τ_s replaces τ_w as the precision quantity that applies to $\alpha_{i(k),t(k)-n+1}$, $\alpha_{j(k),t(k)-n+1}$, $\beta_{i(k),t(k)-n+1}$ and $\beta_{j(k),t(k)-n+1}$. Similarly τ_{s0} and ϖ_{s0} replace τ_{w0} and ϖ_{w0}.

7.3 Prior specifications

Given the ubiquity of the γ,δ,κ and τ_w terms in the model, weak conjugate priors are employed. Since the mean away score during seasons 1997-2000 is 19.14

$$\gamma \sim \mathcal{N}(19, 0.01)$$

The mean (home-away) score during this period is 3.22 so

$$\delta \sim \mathcal{N}(3, 0.01)$$

The (unconditional) score variance is 106.11 hence the unconditional score precision is 0.00942. To have a flat prior on the score precision, a mean of 0.01, variance 10 can be used for which the relevant conjugate prior is

$$\Gamma(1.0 * 10^{-05}, 1.0 * 10^{-03})$$

While some detailed methods for setting prior values for the α and β terms could be

considered, such as using some function of the previous season's scored and conceded averages, for simplicity weak conjugate priors are again selected.

$$\alpha_{.,0} \sim \mathcal{N}(0, 0.01)$$
$$\beta_{.,0} \sim \mathcal{N}(0, 0.01)$$

A weekly deviation of 0.25 points in mean scoring or conceding level for a team seems plausible, or equivalently a precision of 16. The conjugate prior with mean 16, variance 100 is $\Gamma(2.56, 0.16)$. Therefore

$$\tau_w \sim \Gamma(2.56, 0.16)$$

Given the relatively small amount of data available in relation to τ_s, a stronger, informative prior is used. In order to set a prior distribution, first the mean scoring and conceding rate for each team in each year is calculated. Denote these values by $S_{i,j}$, $C_{i,j}$, $i = 1, \ldots, 4, j = 1, \ldots, 31$[1]. Defining $S_{21,j} = S_{2,j} - S_{1,j}$, it is calculated that

$Var[S_{21,.}] = 19.55 \quad Var[S_{32,.}] = 34.83 \quad Var[S_{43,.}] = 16.50$

$Var[C_{21,.}] = 12.16 \quad Var[C_{32,.}] = 14.06 \quad Var[C_{43,.}] = 29.92$

The reciprocals of these are respectively (0.0511, 0.0287, 0.0606, 0.0822, 0.0712, 0.0334). The mean and standard deviation of this vector are 0.05453 and 0.02100, so a reasonable conjugate prior is $\Gamma(6.75, 120)$, which has mean 0.05625 and standard deviation 0.02165.

7.4 Model implementation

Figure 7.2 displays selected portions of the output obtained by running 5 parallel chains of 5,000 iterations of the model described above in WinBUGS. The reason parallel chains are run is in order to check that convergence has been achieved. Should the disparity between the chains be similar to the variance within each chain, this suggests that the Markov chain has converged. It is expected that the different variables in the model will converge at different rates, hence traces of the parameter estimates for the offensive and defensive abilities of one team (the Denver Broncos) as well as the global parameters γ and δ are monitored in the right hand plots of Figure 7.2.

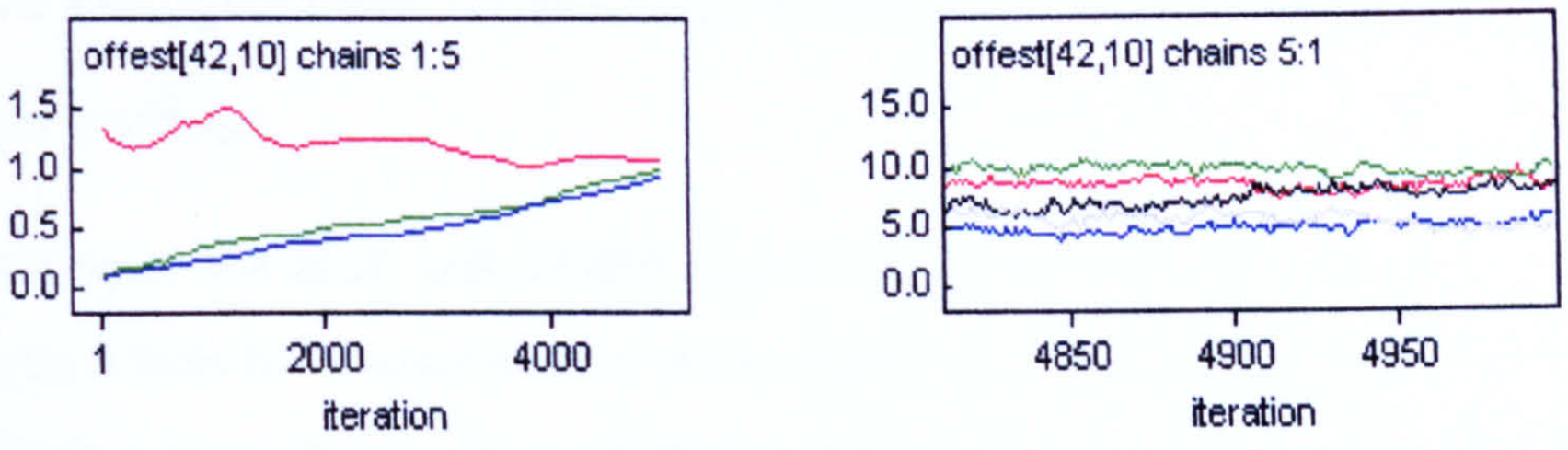
[1]Four seasons of data are used and 31 different NFL teams appear in this data set

While convergence appears to have been reached by the two global parameters, the traces of the chains for the team parameters that have been monitored suggest that overall convergence is still some way off. The traces of Denver Bronco's offensive and defensive parameter on 31 January 1999 reveal several non-intersecting chains, suggesting the stationary distribution has not been achieved. In fact, all these chains were started from the same initial values. By starting these chains at different values, as recommended in several texts, convergence may well seem even further away.

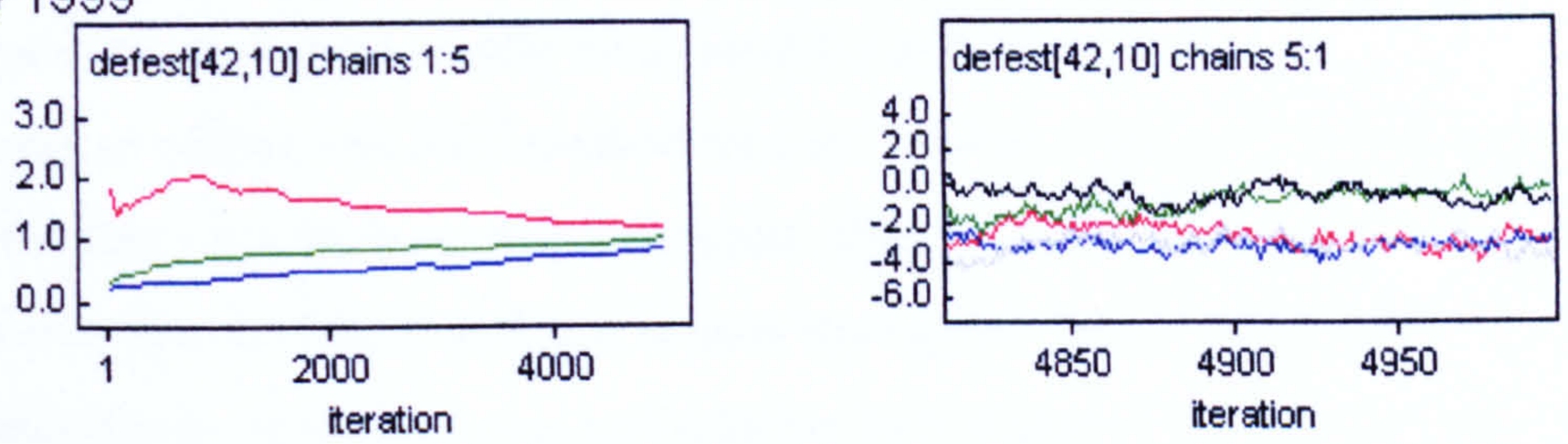
A more formal check for convergence, known as the *Gelman-Rubin diagnostic test*, is outlined in Section 7.6. Essentially, it produces a value known as the *potential scale reduction estimate*, with confidence intervals, which approach the value 1 as convergence of the Markov chain is achieved. The plots of the potential scale reduction estimate are displayed in the plots on the left hand side of Figure 7.2. They confirm what the traces suggest, namely that the global parameters have reached convergence but the team parameters have not.

To consider why this is so, note that the distribution of the Denver Bronco's offensive parameter on 31 Jan 1999 is determined primarily by a single data point, which is the Denver Bronco's score in the match that occurred close to 31 Jan 1999, and its relationship with the Denver Bronco's offensive parameters in the time-points immediately before and after 31 Jan 1999. It is also determined less directly through the complex dependence structure that exists between all the parameters featured in the model, which can partly be observed by recalling the DAG in Figure 7.1. Contrast this with the parameters for the global mean and home effect, which are determined using every match score, as well as the complex dependence structure. Given that far more data directly determines the global mean and home effect, it follows that the Gibbs Sampler converges more quickly towards suitable estimated values for them.

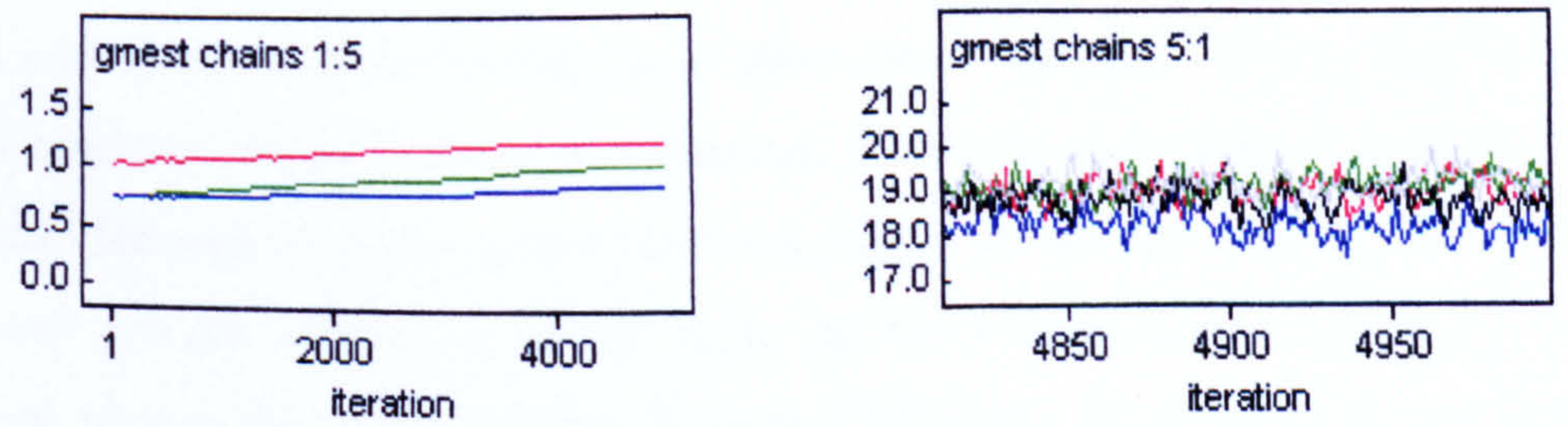
That some parameters have not converged after 5,000 iterations is not surprising given that Glickman-Stern's considerably simpler model was run for 18,000 iterations, by which stage one parameter had still not completely converged. Rue-Salvesen's model, with a similar level of complexity to the one employed in this example was run for 25,000 iterations, although certain parameters were not in fact evaluated via the MCMC routine. Unfortunately, WinBUGS was only able to perform approximately 8,000 iterations before encountering memory problems on a 700MhZ Pentium 3 PC with 384 MB RAM. Hence this approach shall not be pursued any further due to the computational limitations encountered. Incidentally, running the five chains of 5,000



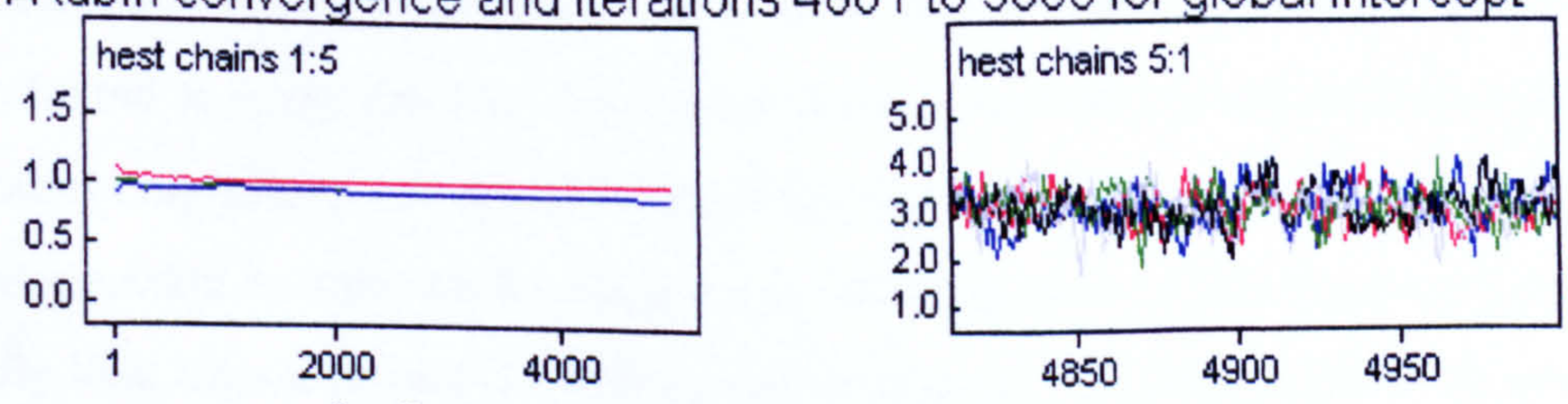
Gelman-Rubin convergence, and iterations 4801-5000 for Denver Broncos offense, 31 January 1999



Gelman-Rubin convergence, and iterations 4801-5000 for Denver Broncos defense, 31 January 1999



Gelman-Rubin convergence and iterations 4801 to 5000 for global intercept



Gelman-Rubin convergence and iterations 4801 to 5000 for home effect

Figure 7.2: Convergence-related output from MCMC treatment of NFL, season 1997/98-2000/01 continued

iterations took 27 minutes.

7.5 A comparison of the MLE and MCMC modelling approaches

Consideration of the MLE and MCMC approaches emphasises that there are several criteria which have to be considered in order to develop a model-building system, many of which are not related to the statistical qualities of the model or the estimation process. For example, it is important that the parameter estimation process is quick, easy to program and gives readily interpretable results.

In terms of output, the MLE method returns 2 parameters (representing the attack and defense) for each team at each time-point, plus an updated estimate for the global mean, home effect and score variance at each time-point. In the NFL data set used for this application there are 31 teams and matches take place on 182 different days. That leads to a total of (31*2+3)*182=11,830 parameters overall. The MCMC technique, on the other hand, returns t*31*2+5 parameter distributions when run at time-point t, since each team's offensive and defensive parameter is modelled as a dynamic process. In addition, the global mean, home effect, score variance, weekly precision and seasonal precision parameters are re-estimated. Note that these effects are assumed to be constant throughout time, unlike the team ability parameters. Hence, if each team is allocated two parameters for all 182 days, and the MCMC is run over all 182 days, the output to the MCMC simulation features $\sum_{t=1}^{182}(62t+5)$ =1,033,396 *distributions* (as opposed to to point estimates). It may seem excessive to allow teams' abilities to be re-evaluated at every day that any match takes place and instead abilities could be re-evaluated only after each occasion when the team in question has played a game. An approximation to this can be achieved by dividing each of the three seasons into 21 equally long time intervals (17 regular and 4 play-off, including Superbowl, weeks), giving 84 time-points overall. However, that still requires estimation of 221,660 distributions overall.

More information can undoubtedly be gained about teams using the MCMC technique but as far as applying the information towards a betting strategy is concerned, it is only the maximum value of the parameters' estimated distributions at the time-point immediately prior to a fixture that are essential for a simple betting strategy. The MLE method provides them. In addition, by using the MLE method, parameter

estimates for the entire four years can be obtained in approximately 35 minutes. This compares favourably to the 27 minutes required to run the 25,000 iterations of the earlier MCMC routine for just the final time-point. Crowder *et al* (2002) report similar findings. While greater efficiency and reliability could be achieved using a more customised MCMC sampler than WinBUGS, programming one would be an immense task. It is unlikely that it would be fast enough to make investigations into model enhancements practicable, especially when one considers that to investigate the effect of model adjustments on the predictive ability of a model, it is necessary that the process is run at many time-points throughout the data set. The MCMC techniques could be pursued further with regard to this application should considerable advances be made in computing power and MCMC software. Hence, on balance, while MCMC is certainly the more attractive approach from a statistical point of view, from a practical point of view, the MLE method is much easier to implement and is also far more suitable for the process of model development. Thus the MLE method has been the most suitable parameter estimation process to employ throughout this thesis.

7.6 Additional comments and information - Markov Chain Monte Carlo methods: a brief summary

In statistical analysis it is often necessary to study a data set via a multivariate interdependent set of variables Θ. If the analysis is being carried out within a Bayesian framework, Θ is a set of parameters and if the analysis is carried out within a frequentist framework, Θ is a set of observable data values. If there are various characteristics of Θ that are of interest, such as modes, higher posterior densities or quantiles of individual components, or relationships between different components of Θ, then algebraic manipulation of the joint distribution, π_Θ, of Θ is necessary. This can be a daunting, or even impossible, task.

MCMC techniques, subject to certain assumptions about relationships between subsets of members of Θ, can produce many samples from a Markov chain whose stationary distribution is π_Θ, hence inferences can be drawn about Θ using relatively straightforward analysis of these samples. There are various techniques employed in order to produce such a Markov chain. All are explained in more detail by Gilks *et al* (1996).

The most general specification of the MCMC technique employs the *Metropolis-*

Hastings algorithm. It produces a sequence of generated values $\Theta_1^*, \ldots, \Theta_n^*$ which, for a suitably large value of n, represent a sample from π_Θ. Firstly a *proposal distribution* $q(.|\Theta_t)$ is specified along with an initial value Θ_0^*. Next, a 'candidate' value $\Theta^\dagger$ is generated from $q(.|\Theta_0^*)$. $\Theta^\dagger$ is accepted with probability

$$\alpha(\Theta_0^*, \Theta^\dagger) = min(1, \frac{\pi(\Theta^\dagger)q(\Theta_0^*|\Theta^\dagger)}{\pi(\Theta_0^*)q(\Theta^\dagger|\Theta_0^*)} \tag{7.6.1}$$

If $\Theta^\dagger$ is accepted, $\Theta_1^* = \Theta^\dagger$ otherwise $\Theta_1^* = \Theta_0^*$. In fact $q(.|.)$ can be any distribution and the stationary distribution of $\Theta_1^*, \ldots, \Theta_n^*$ is always π_Θ, however the choice of distribution affects how quickly the chain converges. Also, it is necessary to have a 'burn-in' period of m iterations so that only samples $\Theta_m^*, \ldots, \Theta_n^*$ are considered to be representative samples from π_Θ. In this way the samples are not affected by the choice of starting value Θ_0^*.

There are various common implementations of this algorithm, in particular there are various ways of approaching the task of finding a suitable choice of $q(.|.)$. For example, the *Metropolis algorithm* involves choosing only symmetric forms, hence Equation 7.6.1 reduces to

$$\alpha(\Theta_0^*, \Theta^\dagger) = min(1, \frac{\pi_\Theta(\Theta^\dagger)}{\pi_\Theta(\Theta_0^*)}$$

Another form of this algorithm is known as *single-component Metropolis-Hastings* and involves decomposing Θ into a set of smaller subsets $\Theta_1, \ldots, \Theta_r$. Then these individual components are updated sequentially during each iteration, subject to an acceptance criteria similar to that of Equation 7.6.1. The most common example of single-component Metropolis-Hastings, and in fact the most widely used MCMC routine at the time of writing, is the *Gibbs Sampler*. When applying a Gibbs Sampler, the proposal distribution for updating the i^{th} component of Θ_t^* is the full conditional distribution

$$\alpha(\Theta_i^\dagger|\Theta_t^*) = \pi_\Theta(\Theta_i^\dagger|\Theta_{t,1}^*, \ldots, \Theta_{t,i-1}^*, \Theta_{t,i+1}^*, \ldots, \Theta_{t,K}^*)$$

where Θ is of dimension K.

One key task when using MCMC methods is monitoring convergence of the chain of values to ensure that the chain has settled into the required distribution. One way to do this, which the WinBUGS software implements, was outlined by Gelman-Rubin (1992). It involves running several chains in parallel and checking that they overlap to a satisfactory extent. To do this, define Ψ as a scalar summary figure of all the

simulated values of one of the parameters. Then two estimates of $var(\Psi)$ can be made. Defining ψ_{ij} to be the jth realisation of the summary figure of the values from chain i, these estimates are

- the *within-chains* variance

$$W = \sum_{i=1}^{m} \frac{s_i^2}{m}$$

where $s_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (\psi_{ij} - \overline{\psi_{i.}})^2$

- a weighted average of the *between-chains* variance

$$B = \frac{n}{m-1} \sum_{i=1}^{m} (\overline{\psi_{i.}} - \overline{\psi_{..}})^2$$

and W, thus

$$\widehat{var}(\Psi) = \frac{n-1}{n} W + \frac{1}{n} B$$

is a second estimate of $var(\Psi)$

$\widehat{var}(\Psi)$ is initially larger than $var(\Psi)$, while W is initially lower. However they both converge towards $var(\Psi)$ as the number of iterations continues. Hence it makes sense to define the following ratio, known as the *estimated potential scale reduction*,

$$\sqrt{\hat{R}} = \sqrt{\frac{\widehat{var}(\Psi)}{W}}$$

Once $\hat{R}$ and its confidence intervals ($\hat{R}$ follows a t-distribution) are suitably near 1, convergence has been achieved. This technique for analysing variance is known as the *Gelman-Rubin diagnosis.* Values of $\hat{R}$ and confidence intervals, choosing various parameter realisations as the scalar summary are included in the output provided for the application of MCMC in this chapter.

Chapter 8

Conclusion

As stated in Chapter 1, the aim of this thesis was to develop models for sporting events that produced probabilities that were at least as accurate as those inferred from odds offered by professional bookmakers. Three attempts have been made to achieve this, but only one, concerning the rate at which yellow and red cards are collected by soccer teams, appears to be successful.

It is not clear whether the greater success rate of the bookings model is because the model is more accurate, or because the market for bookings bets is less effective at forcing the mean towards the "true" probability than the market for NFL and NBA scores. It is not possible to produce figures such as the predictive likelihood for bookmakers' odds since only their expected mean is provided in the case of spread betting and only their expected prediction for the median is provided in the case of fixed odds handicap betting. The entire probability density for all outcomes is required for most summary statistics of predictive capability. Other commonly used goodness-of-fit summary statistics, such as C_p, R^2 or reduction in χ^2 used in regression, can only be used to compare nested models on the same data set, and cannot be used to compare the accuracy of different predictions for different sports. Hence a comparison of the accuracy of the central spreads for bookings with the accuracy of the NFL lines, for example, is not possible.

Nevertheless, it is not entirely surprising that the bookings model produces better returns when one considers the amount of match-specific information that is incorporated into each model prediction. As well as both teams' individual tendency to attract and provoke bookings, the referee, the difference in ability of the two teams, historical rivalries and match-specific incentives are accommodated into each prediction. Enhancing the NFL and NBA models in this way is necessary and entirely feasible since

the data concerning which players are injured, or whether a team has any unusual extra incentive entering a match, is available.

As far as further research is concerned, there are many other possible ways to enhance all models covered so far. For example, data is also available concerning the time that all points are scored, or bookings are collected. Using these, it is possible to improve not only the predictions generated for the match totals before the fixture takes place, as has been attempted throughout this thesis, but it is also possible to generate predictions while the match is in progress, given information already observed.

The fact that there are so many obvious options available for model enhancement is encouraging. Furthermore, the effect of any attempts to improve models can be analysed and developed further if necessary. This does not apply to the intuitive approach since once the knowledge concerning the sport has been acquired, and the skills in converting this knowledge into probabilities in a reliable way have been developed, it is not clear how any further improvements to the system can be made. While not all of the models in this thesis have the desired predictive capability, the approaches covered have much scope for improvement.

Bibliography

[1] A Colin Cameron and Pravin K Trivedi (1998) *Regression analysis of count data* Cambridge University Press

[2] Martin Crowder, Mark Dixon, Anthony Ledford and Mike Robinson (2002) "Dynamic modelling and prediction of English Football League matches for betting" *The Statistician* Vol.51, Part 2, 157-168

[3] M Dixon and S Coles (1997), "Modelling association football scores and inefficiencies in the football betting market", *Applied Statistics* Vol.46, 265-280

[4] Ludwig Fahrmeir and Gerhard Tutz (1994) "Dynamic stochastic models for time-dependent ordered paired comparison systems" *Journal of the American Statistical Association* Vol.89, 1438-1449

[5] David Forrest and Robert Simmons (2000a) "Forecasting sport: the behaviour and performance of football tipsters" *International Journal of Forecasting* Vol.16, 317-331

[6] David Forrest and Robert Simmons (2000b) "Making up the results: the work of the Football Pools Panel, 1963-1997" *The Statistician* Vol.49, Part 2, 253-260

[7] John M Gandar, Richard A Zuber, Reinhold P Lamb (2001) "The home field advantage revisited: a search for the bias in other sports betting markets" *Journal of Economics and Business* Vol.53, 439-453

[8] Andrew Gelman and Donald B Rubin (1992) "Inference from iterative simulation using multiple sequences" *Statistical Science* Vol.7, No.4, 457-472

[9] Andrew Gelman, John B Carlin, Hal S Stern and Donald B Rubin (1995) *Bayesian Data Analysis* London:Chapman & Hall

[10] W R Gilks, S Richardson and D J Spiegelhalter (1995) *Markov Chain Monte Carlo in practice"* CRC Press

[11] Mark E Glickman and Hal S Stern (1998) "A state-space model for National Football Scores" *Journal of the American Statistical Association* Vol.93, 25-35

[12] David Harville (1980) "Predictions for National Football League games via linear-model methodology" *Journal of the American Statistical Association* Vol.75, No.371, 516-524

[13] N Hirotsu and M Wright (2002) "Using a Markov process model of an association football match to determine the optimal timing of substitution and tactical decisions" *Journal of Operational Research Society* Vol.53, 88-96

[14] Leonard Knorr-Held (2000) "Dynamic rating of sports teams" *The Statistician(2000)* Vol.49, Part 2, 261-276

[15] M J Maher (1982) "Modelling association football scores" *Statistica Neerlandica* Vol.36, No.3, 109-118

[16] M Moroney (1951) *"Facts from figures"*, London, Pelican.

[17] G Ridder, JS Cramer and P Hopstaken (1994), "Down to Ten: Estimating the Effect of a Red Card in Soccer", *Journal of the American Statistical Association* Vol.89, No.427, 1124-1127

[18] Haavard Rue and Oyvind Salvesen (1997) "Predicting soccer matches in a league" *Technical Report, Statistics No.10, October 1997*, Department of Mathematical Sciences, Norwegian University of Science and Technology

[19] Robert Simmons, David Forrest and Anthony Curran (2003) "Efficiency in the handicap and index betting markets for English rugby league" extracted from *The Economy of Gambling* by Leighton Vaughan-Williams (Routledge)

[20] B W Silverman (1986) *Density Estimation* Chapman and Hall, London

[21] Raymond Stefani (1977) "Football and basketball predictions using least squares" *IEEE Transactions on Systems, Man and Cybernetics* February 1977

[22] Raymond Stefani (1980) "Improved least squares football, basketball and soccer predictions" *IEEE Transactions on Systems, Man and Cybernetics* Vol.SMC-10, No.2, February 1980

[23] Hal Stern (1991) "On the probability of winning a football game" *American Statistical Association* August 1991, Vol.45, No.3, 179-183

[24] Roger C Vergin (2001) "Overreaction in the NFL point spread market" *Applied Financial Economics* Vol.11, 497-509 Chapman and Hall, London

[25] M P Wand and M C Jones (1995) *Kernel Smoothing* Chapman and Hall, London