ATHENS UNIVERSITY
OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

STATISTICAL MODELLING FOR FOOTBALL DATA: A ROBUST APPROACH BASED ON WEIGHTED MAXIMUM LIKELIHOOD

By

Georgios M. Kalamidas

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
2005

ATHENS UNIVERSITY
OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Statistical Modelling for football data: A Robust Approach based on Weighted Maximum Likelihood

By

Georgios M. Kalamidas

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
January 2005

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

Στατιστική Μοντελοποίηση για ποδοσφαιρικά δεδομένα: Μια Ανθεκτική Προσέγγιση βασισμένη στη Σταθμισμένη Μέγιστη Πιθανοφάνεια

Γεώργιος Μ. Καλαμίδας

ΔΙΑΤΡΙΒΗ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Ιανουάριος 2005

ATHENS UNIVERSITY
OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

A Thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science

STATISTICAL MODELLING FOR FOOTBALL DATA: A ROBUST APPROACH BASED ON WEIGHTED MAXIMUM LIKELIHOOD

Georgios M. Kalamidas

Approved by the Graduate Committee

D. Karlis	I. Ntzoufras	V. Vasdekis
Assistant Professor	Assistant Professor	Assistant Professor
Thesis Supervisor	Members of the Committee	

Athens, June 2005

Michael Zazanis, Professor
Director of the Graduate Program

DEDICATION

I dedicate this M.Sc. Thesis:

- ☺ Firstly, to my family; my father Michail, my mother Eleni and my brother Antonios (who also belongs to the statistical community). Without them, it is almost sure that the reader would not have the chance to read this report!
- ☺ And secondly, to the Greek National Football Team for the unique achievement in the Euro Cup 2004 (Portugal).

ACKNOWLEDGEMENTS

First of all, I am grateful to my supervisor, Dimitris Karlis, for his valuable and essential help. I would like also to thank my colleagues Antonia Korre, Eleni Costi, Evangelia Stamouli-Letta, Stamatia Nikolaou and Nikolaos Mitsopoulos; in a spirit of noble emulation and through our hard work and excellent collaboration for the past two years, we all became better statisticians.

Finally, I would like to thank all my close friends and my girlfriend for their moral support and patience during the time that we know each other.

VITA

Despite the fact that I was born in 2 June 1977 in Athens, I am very proud of my (and my parents') roots, the village Ramia. Ramia belongs in the Prefecture of Arta (Epirus, West Greece). It is located high up to the famous and exquisite Mt Tzoumerka (2429m. above sea level). Everyone should visit it!

My general interests are sports (and especially football) and the folk dances of Greece and other places in the rest of the world. I have great concern about the knowledge, the study, the preservation and the promotion of the Greek Folklore and Tradition. For this reason, I activate myself the late years in the Board of Cultural Association of my village and I participate in several folklore dancing groups.

I graduated from Department of Mathematics in the Faculty of Sciences of the National & Kapodistrian University Of Athens. I have degree in Mathematics with primary specialization in Applied Mathematics and secondary specialization in Mathematical Education. After I submit this Thesis, I will get the degree of Master of Science in Statistics from the Department of Statistics of the Athens University of Economics and Business.

I have taken part in many Statistical and Mathematical conferences and I am member of the Greek Mathematical Company and the Greek Statistical Institute. My scientific interest is Sports Statistics.

ABSTRACT

Georgios M. Kalamidas

Statistical Modelling for football data: A Robust Approach based on Weighted Maximum Likelihood

January 2005

Generalized linear models are an extension of classical linear models and can be used in a wide range of applications for the estimation of the unknown parameters. Maximum Likelihood Estimation (MLE) is explicitly model-dependent. Thus, a fine alternative and robust method is the Weighted Maximum Likelihood Estimation (WMLE). WMLE can be used in many areas of scientific research, but here we shall see an application to soccer data. Since different distributions could fit soccer data sets well, for many years there had been a long discussion on which one we should use and the most known debate is that of the Poisson versus the Negative Binomial distribution. For some reasons, we prefer to fit the Poisson distribution, although we have the suspicion that the underlying distribution is not exactly this one, but a (probably small) deviation of it. We shall apply robust methods, because Robust Theory can deal with both data contamination and model deviation. Our soccer data refer to the season 2003-2004 of the Greek National A Division (GNA). At the end, we give a small comparison between the bookmakers' and our results.

ΠΕΡΙΛΗΨΗ

Γεώργιος Μ. Καλαμίδας

Στατιστική Μοντελοποίηση για ποδοσφαιρικά δεδομένα: Μια Ανθεκτική Προσέγγιση βασισμένη στη Σταθμισμένη Μέγιστη Πιθανοφάνεια

Ιανουάριος 2005

Τα Γενικευμένα Γραμμικά Μοντέλα (ΓΓΜ) αποτελούν μια επέκταση των κλασικών γραμμικών μοντέλων και μπορούν να χρησιμοποιηθούν σε ένα ευρύ φάσμα εφαρμογών για την εκτίμηση των αγνώστων παραμέτρων. Η Εκτίμηση Μέγιστης Πιθανοφάνειας (ΕΜΕ) είναι απόλυτα εξαρτημένη από το μοντέλο. Έτσι, μια πολύ καλή εναλλακτική και ανθεκτική μέθοδος είναι η Εκτίμηση Σταθμισμένης Μέγιστης Πιθανοφάνειας (ΕΣΜΕ). Η ΕΣΜΕ μπορεί να χρησιμοποιηθεί σε πολλούς τομείς της επιστημονικής έρευνας, αλλά εδώ θα δούμε μια εφαρμογή σε ποδοσφαιρικά δεδομένα. Καθώς διαφορετικές κατανομές μπορούν να προσαρμοστούν σε σύνολα ποδοσφαιρικών δεδομένων, για πολλά χρόνια υπήρχε μια μακρά συζήτηση για το ποια θα έπρεπε να χρησιμοποιήσουμε και η πιο γνωστή διαμάχη είναι αυτή μεταξύ της Poisson κατανομής έναντι της Αρνητικής Διωνυμικής κατανομής. Για κάποιους λόγους προτιμούμε να προσαρμόσουμε την Poisson κατανομή, αν και έχουμε την υποψία, ότι η κατανομή που υπάρχει στο βάθος δεν είναι αυτή ακριβώς αλλά μια (πιθανώς μικρή) απόκλισή της. Θα εφαρμόσουμε ανθεκτικές μεθόδους, επειδή η Θεωρία Ανθεκτικότητας μπορεί να χειριστεί τόσο την παραποίηση των δεδομένων όσο και την απόκλιση των μοντέλων. Τα ποδοσφαιρικά μας δεδομένα αναφέρονται στην περίοδο 2003-2004 για την Ελληνική Α΄ Εθνική Κατηγορία (ΕΑΕΚ). Στο τέλος, δίνουμε μια μικρή σύγκριση ανάμεσα στα αποτελέσματα των γραφείων στοιχημάτων και των δικών μας.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Short biography of soccer

The name "soccer" appeared, when the Europeans emigrated in the United States in order to distinguish this sport with the football already being played there. Actually the birth name was "Association Football", but after while it was shortened to "Assoc. Football" and some people just called it "assoc." or "soc.". At that time, it was very common to add an "er" to words, especially from the students of the 1880s, so the name took its final form.

Various Egyptian tombs, as old as 2500 BC, provide evidence that football-like games existed in that region and time period. The Greeks, for whom ball games were an essential part of life, developed a kicking/throwing game, called *Phaininda* or *Episcyros* around 2000 BC (see also Appendix II). The Greek game of *Episcyros* was later adopted by the Romans and renamed it into *Harpastum* (meaning "the small ball game"). It is also believed that the Romans took *Harpastum* to the British Isles at the time of their expansion. Another one of the first soccer-like games was *Tsu Chu* (literally "kick ball") played in China. Its use was dual; firstly it was a part of the soldiers' training during the Ts' and Han Dynasties (255 BC-260 AD) and secondly it was played as a part of the emperor's birthday celebration. In Japan, records show that around the 5th century AD there existed a similar game called *Kemari*. The French medieval *Soule* and the Italian *Calcio* are also among the historical antecedents of soccer. In Mexico and Central America, between 600 and 1600 AD, the Mayans and Aztecs played a game akin to soccer. North American folklore tells of 17th century indigenous Americans playing "*Pasuckquakkohwog*" (which means "they gather to play ball with the foot"). Finally, the Eskimos in Canada and Alaska are reputed to have played *Aqsatuk* on ice.

After this travel in time, the game arrived in England with a fairly bad reputation among British royalty. The government passed laws against soccer and throughout the centuries, English monarchs tried to ban this version of the game. The game became so popular by 1800 that, in certain contests in northern and middle England, huge masses of people were gathered and demonstrated against such prohibitions. In 8 December 1863, the “Football Association” (FA) was created in London and a group representing various clubs adopted a code of rules. Marples (1954) quoted two decisive rules:

1. “*No player shall run with the ball*” and
2. “*Neither tripping nor hacking [kicking an opponent on the shin] shall be allowed and no player shall use his hands to hold or push his adversary*”.

Of course, the goalkeeper is the exception to that rule and the other players can use their hands only in the case when they have to bring the ball back to the field from the sidelong lines.

These rules regarding handling and running with the ball defined the essential difference between the soccer and the rugby or the American Football. In just over 140 years, the rules of the game have remained essentially unchanged. There have been minor changes, such as those allowing for substitutes, the determination of a result in some games by several forms of extra-time or penalty shoot-outs and developments in the offside rule. Nowadays, a soccer game lasts 90 minutes totally (two half times of 45 minutes), played between 2 teams of 11 players each and in a rectangular shaped playing ground with fixed dimensions. This football ground is separated into two mirror image semi-grounds, which are possessed by each team. In the second half time the two teams change sides, so as to ensure as far as possible equal terms for both of them. Finally, there is a referee with the help of two linesmen, who is the objective judge of the debatable phases.

There is an indication in a Chinese text at Munich Ethnological Museum in Germany that the first international game was played by Japan and China in the year of 50 BC. However, it is known for sure that a game was played in 611 in the ancient Japanese’s capital of Kyoto. After the foundation of FA, the first international game was played between England and Scotland in 1872 and on 1 May 1904 Belgium faced France in Brussels. The first official football (or soccer) club worldwide was an English one named Sheffield Club (1855) and in 1862 England’s oldest professional club, Notts County, formed. In 1871, the English Cup was established followed by the international championship in 1884. In July 1885, professional football was legalised by the F.A. in England in response to the increasing number of working class players in the game and the revenue gained from rising attendances. In 1888 the first professional Football League was created and a second division was added in 1892; modern football was well and truly born. The first steps of soccer in Greece peeped out around 1895 and the primacies belong to Thessalonica's "Omilos Filomouson" (meaning the "Club of Music Lovers"), which is the predecessor of Iraklis. Iraklis and Panathinaikos were the two first official Greek clubs founded in 1908.

By May of 1904 the "Fédération Internationale de Football Association" (FIFA) was established in Paris in order to govern and control the sport worldwide. The 7 founding members were France, Belgium, Denmark, Spain, the Netherlands, Sweden and Switzerland. Today, FIFA counts 204 member nations. The European part of the FIFA, the "Union des Associations Européennes de Football" (UEFA), celebrated its Jubilee on 15 June 1954 in Basel of Switzerland. UEFA became the guiding hand and governing body of European football, on and off the field. Initiatives to found continent-wide competitions were soon acted upon. Nowadays, there are 6 continental confederations, which (except for the national championships) take care of continental and international tournaments and of course, the big event of the World Cup (or the "Mundial") every 4 years. Uruguay was the first World Cup winner in 1930.

Men's Football was introduced as a demonstration sport in the 1896 Olympic games of Athens. It became an official Olympic sport in the 1908 Games in London. FIFA launched the Olympic Football title in 1924 and the World Cup in 1930. Nearly a century passed before the first ever FIFA World Cup for women was held in China in 1991 and women's Football was included in the Olympic competition program in the 1996 Games of Atlanta. Until the 1984 Olympic games, participation was restricted to amateur players and eastern European countries dominated Olympic Football. After the inclusion of professional athletes at the Olympic games, participation rules have been a subject of debate between the International Olympic Committee and FIFA. As a result, new regulations were established specifying the age of participant athletes. The new regulations have enabled many African countries to display a rich pool of Football talents. Among the web sites listed in the Appendix, more details can be found in the URL address **http://www.fifa.com** or alternatively in **http://www.soccerway.com**.

1.2 The appearance of statistics in soccer

Soccer is the most popular sport in the world, played by nearly 250 million people, including 40 million women. The numbers of spectators in the stadiums and of television viewers amounts to billions. Since the reputation and the acceptance for this sport was of universal interest, it was very logical for money to be implicated. On one hand, relative sport industries and companies made increasing investments on this area. On the other hand, the populace became impatient either just simply to predict or even to bet on the outcome of a match event.

Many sports in general and soccer in particular are a fertile land for applying statistical methodologies and developing methods for dealing with athletic data. Many times scientific search provide fundamental help in taking several crucial decisions. In the middle of the twenty first century, statisticians started creating statistical models in order to predict the outcome of soccer games. The United Kingdom and countries mainly from Central Europe, the United States of America and Canada have long tradition in betting on the outcome of soccer or other games. Lately, several variants of betting have occurred. So, someone can now bet for example on the half time outcome, the exact final scores, who will be the first player to score, if the first goal will be scored by a header, a penalty or a foul, which team will come first in specific (small) group of teams in a competition, which team will be the winner of the tournament and many more. The basic aim was to perform better predictions than the bookmakers.

The challenge in making bets was, and still is, to find those in which the considered probability of occurrence is higher than the corresponding probability determined by the bookmakers' odds and thus the expected gain is high. Statistical models can be very helpful tools for such purposes. Usually, the odds in soccer are fixed some days before the matches are played. This fact gives the right to the researcher estimate the probabilities under his theory and to compare them with those of various bookmakers. So, he can exploit any weakness in the bookmakers' specification and possibly take advantage of it.

The core of our work will be Maher's (1982) model (see Section 2). According to Dixon and Coles (1997) a statistical model for soccer games should first of all take into account the different abilities of both teams in a match and it should allow a "home effect", which means that generally most of the teams perform better when they play at home. Because of the nature of soccer, it is probably better to divide each team's ability into two parts; the ability to attack and score (i.e. the "offensive ability") and the ability to defend and not concede goals (i.e. the "defensive ability"). Also, it is reasonable to measure each team's ability based on more recent results and to take into account the ability of the teams that they have played against. Of course, many more requirements could be opposed, but it is far from clear that it is not practical to obtain empirical estimates of probabilities of matches' outcomes that account for all these constraints. Since we process count data, we shall use a generalized linear model under the Poisson family following the assumption made by Maher (1982) and other authors, that the number of goals scored by the two teams are independent Poisson variables, whose means are determined by the respective attack and defence qualities of each side. The novel idea is to apply a soccer modelling technique (after taking advantage of all the available literature) from another point of view, which is based on robust theory.

Since different distributions could fit soccer data sets well, for many years there had been a long discussion on which one we should use. The known debate of the Poisson versus the Negative Binomial distribution detained the statisticians long. At the end, Karlis and Ntzoufras (1998) and Baxter and Stevenson (1998) found that these two distributions were very close to each other and that in practice there was not much difference. For the above reason (described more analytically in Section 2), we prefer to fit the Poisson distribution for our soccer data, although we have the suspicion that the underlying distribution is not exactly this one, but a (probably small) deviation of it. We should not forget to refer to the remark of Douglas (1994), that short-tailed observed frequency distributions are often well fitted by a number of different theoretical discrete distributions, with little discriminatory power. Under this thought, we will try to apply robust methods.

The need for a different approach, such as robust theory and methods, issued from the fact that parametric models were only approximations to reality. In our case, as we just said, the choice of the Poisson distribution may not be the right one. In conformity to common belief, Statistics is the science of extracting useful information from empirical data. An effective way for conveying the information is to use parametric stochastic models. Nevertheless, there are some reasons described in Section 3 that do not allow in each case complete freedom of action and a pressing need of using nonparametric methods arises. Robust statistics combines the virtues of both approaches. Parametric models are used as vehicles of information, and procedures that do not depend critically on the assumptions made in these models are implemented.

We shall later see, that there are four types of deviations from strict parameters models and according to Grunert and Fieller (1995) there are two main types of distortion; data contamination and model deviation. Finally, the main aims of robust statistics are: (i) to describe the structure best fitting the majority of the data, (ii) to identify outliers or deviating substructures, (iii) to identify leverage points and (iv) to deal with unsuspected serial correlations and deviations from the assumed correlation structures. If the data are high quality or they do not have any outliers, robust methods are not absolutely necessary, but they can give a noticeable improvement over classical ones.

The remaining of this thesis is organized as follows:

Section 2 gives an extensive literature review in soccer research and several aspects of it and presents most of the statistical methods used for soccer data. Several models are presented for predicting the outcome of a specific game or even a whole tournament and there are basically three different approaches. The first one is trying to model only the outcome of a match and the models specified can be used also for ranking the teams. For the second approach, this is the point for the argument in the choice of the appropriate fitting distribution. The models here try to predict the number of goals scored by each team and they are divided into two basic categories; (i) the teams' performances are constant and (ii) these performances change across the time. The third approach deals with several aspects of other characteristics that can be found in this sport. This Section concludes with the implication of soccer statistics in economics.

Section 3 presents the main robust theory that we used. Again, there are two basic approaches; (i) the work of Huber (1964) and (ii) the "Infinitesimal approach" (Huber, 1972). This Section continues with the importance and the treatment of outliers and after some more theoretical issues, the weighted maximum likelihood method gradually develops. We focus on Lindsay's (1994) approach and we basically work on a new algorithm and its relevant equations. We also mention the importance of the generalized linear models and how robustness is involved. The Section concludes with the MM algorithms, which a rising area of research in statistical science.

At the beginning of Section 4, we detect the disadvantage of maximum likelihood estimation in soccer data and we stress our interest in weighted maximum likelihood estimation. We give an example, in which different estimation methods can actually produce different results. We use the Greek National A Division League for the season of 2003-2004 and explain the structure of these soccer data. In the following, we simulate 1000 leagues through these two estimating methods, while for the second one we define two types of weights different from those proposed by Lindsay (1994). There is a full description and comparison of the results and at the end we say some interesting things about the betting market.

Finally, Section 5 concludes and suggests refinements that, we believe, would lead to further improvement in return. In the end, after the two Appendices that contain a list of web links for soccer statistics and a proof of the soccer's birth, there is an extensive list of references.

CHAPTER 2

Review of soccer research

This section contains most of the methods used in soccer statistics. The aim of the research had mainly been to examine the nature of the data in order to find a distribution that fitted closely to the number of goals. There are also many other parameters that could affect the final result. So, several models were adopted for predicting the outcome of a specific game or even a whole tournament. A lot of papers have been published presenting statistical methods for soccer data. Someone could locate three basic approaches in the research of soccer statistics. The first approach focuses on the win or the loss of a game, so it models the outcome of a game. The models specified could be used also for ranking soccer teams. The second category investigates models for the prediction of the number of goals scored by each team or else, it concentrates on score modelling. The third category deals with several aspects of other characteristics that can be found in this sport. This Section quotes the main topics in literature for the three above approaches and concludes with how the economics in general could be implicated with soccer statistics.

2.1 Modelling the win or a loss of a game

Bradley and Terry (1952) introduced a general model where are represented the results of some experiments. The responses are pair-wise rankings of the n elements of a set $A=\{\alpha_1,\alpha_2,...,\alpha_n,\}$. This set can contain objects, items, persons, teams or treatments. Such a model and developments of it have been used during the last decades in several scientific areas such as psychology, economy and biometrics. It should be mentioned that much earlier Zermelo (1929) had designed the first model for paired comparisons in order to rate chess skills. One of the applications is the paired comparison of all the teams in a Round Robin Tournament. Let $y_{i,j}$ denote the result ($y_{i,j}=1$, if i team wins and $y_{i,j}=2$, if j team wins) of a match between the pair (i,j). Then the paired comparisons may be presented as:

$$P\left(y_{i,j}=1\right)=\frac{\exp\left(a_i-a_j\right)}{1+\exp\left(a_i-a_j\right)},$$

where α_k represents the strength of team k, for every k.

Closely related to this model was Kuonen's (1997) logistic regression model. He tried to predict the probability of each team in a European soccer Cup to reach to a certain round and the probability to win the Cup. He proposed three different methods and the best was the one that assumed constancy of the team strength. His calculations were divided into two parts. First, he considered for the year x, C_x to be the ratio points achieved over games played for the 3 past years and then he took a weighted mean of the ratio points achieved over games played for each of these 3 years:

$$\begin{cases} C_x=\dfrac{po\operatorname{int}s\ \textit{achieved during the three year period}}{\textit{games played during the three year period}} \\ \textit{Coefficient}=\dfrac{3}{6}C_{x-1}+\dfrac{2}{6}C_{x-2}+\dfrac{1}{6}C_{x-3} \end{cases}$$

Secondly, he calculated the probabilities described above as:

$$\begin{cases} P\left(i\ \textit{wins in leg}\ k\right)=P\left(\boldsymbol{i\ wins\ in\ leg\ k}-1\right)\cdot\sum_{j\in J}P\left(\boldsymbol{j\ wins\ in\ leg\ k}-1\right)\cdot P_k\left(i,j\right) \\ P\left(i\ \textit{wins the tournament}\right)=P\left(\boldsymbol{i\ wins\ in\ leg\ k}=\log_2 n\right) \end{cases}$$

where n teams participate in a tournament of k legs and J is the set of all potential opponents of i ($k=2,\ldots,\log_2 n$). He succeeded to predict correctly about 64.49% of the 376 game outcomes of the European Cups from 1992 to 1996.

Later on, Kuonen and Roehrl (2000) took all the information from a ranking system for a simple probability model and tried to predict results for the World Cup of France '98. A ranking system is an attempt to represent quantitatively the strength of a team. They combined Stefani's (1980) model (see section 2.3.5) for the Round Robin part with Kuonen's (1997) model for the knockout part. They used only the results of the preliminary rounds and ignored the long-term past performance of the teams until the start of the World Cup in order to construct a rating scheme of their own possibly different from the already known, such as the FIFA/Coca-Cola World ranking or the World Soccer Elo. They pinpointed differences and indicated among others, that France deserved the World Cup and not Brazil, which was up to then top-ranked. Thus, they verified Kuonen *et al.* (1999); after using a simplified version of Kuonen's

(1997) model, they had predicted that France would have won the Cup with a probability of 57.4%, whereas Brazil was 3:1 favourite at the bookmakers' odds!

Sometimes, a game is included in football pools but it cannot be played due to bad weather. For gambling purposes, there are panels of experts who determine the result. In Britain, Forrest and Simmons (2000) used an ordered logit model proposed by Zavoina and McElvey (1975) and showed that panel results are more predictable than real results. They suggest that the pools panel should add some random noise to simulate real result more closely. An introduction of approximately 10% of "wild" predictions into their results would reflect better the unpredictable nature of real games.

2.2 Score modelling

2.2.1 Which distribution should we use?

Poisson distribution has a formal theoretical basis and is naturally used for events that occur randomly and at a constant rate over the observed time period. The Negative Binomial distribution (NBD) belongs to the family of mixed Poisson distributions and should be in our mind, if we assume that the scoring ability varies across time, teams and so on. It is derived from the simple Poisson distribution by assuming that its parameter varies according to a Gamma distribution. A reason why someone might not prefer it instead of the Poisson distribution, it is because it appears more complicated.

One basic property of the Poisson distribution is that the mean is equal to the variance and it is often examined by a measure called "index of dispersion", which is the variance to the mean ratio. In practice, for soccer data and for each team we can calculate through a set of matches the means of goals, the variances and these indexes. Anderson and Siddiqui (1994) noticed that, if the Poisson assumption is valid, then we should expect almost half of the teams to have index of dispersion greater than one and the others to have index of dispersion less than one. Karlis and Ntzoufras (1998) showed that the distribution of the number of the goals is over-dispersed relative to the simple Poisson distribution (i.e. a significantly larger percentage than 50% of the teams have index of dispersion greater than one).

The first one who presented a statistical model to predict the outcome of a soccer game was Moroney (1956). He declared that the Poisson and even better the NBD are the most appropriate to find the probability of winning a game. None the less, a sufficiently developed model came much later by Maher (1982), whose work comprised the basis for many subsequent writers. Specifically, if team i is playing at home against team j and the observed score is $\left(x_{ij}, y_{ij}\right)$, he assumed that:

$$X_{ij} \sim Poisson\left(a_i \beta_j \gamma\right) \text{ and } Y_{ij} \sim Poisson\left(a_j \beta_i\right)$$

where X_{ij} and Y_{ij} are independent variables, α_i and β_j measure their attack and defence qualities, while the team's i home ground advantage (HGA) is measured by γ. Each match has a different fitted Poisson distribution and this was the main differentiation from Moroney (1956) and Reep and Benjamin (1968) who fitted a single distribution to scores from all matches and concluded that this distribution should be the NBD and not the Poisson. After that, Maher (1982) examined the difference $Z_{ij} = X_{ij} - Y_{ij}$ between the teams' scores and actually improved the fit by using a bivariate Poisson model with correlation of about 0.2, since the independence assumption was not totally valid.

The choice of either the NBD or the Poisson distribution also concerned Baxter and Stevenson (1988). They fitted the two distributions on scores between 1946 and 1984 and found that, before the 1970, the NBD provided superior description of the data, but after 1970 both distributions are adequate. Furthermore, to discriminate between the different mechanisms requires large quantities of data. Also, Leroux and Puterman (1992) remarked that there is little statistical evidence to favor the choice of a finite mixture as opposed to a NBD, although the physical interpretation of the former appears more meaningful. Douglas (1994) added that many distributions may fit a data set well and the choice of which distribution to apply might be a difficult one. Generally, short-tailed observed frequency distributions are often well fitted by a number of different theoretical discrete distributions, with little discriminatory power.

Karlis and Ntzoufras (2003) proved, that under the assumption that the joint distribution of the number of goals scored by each team is a bivariate Poisson distribution, the outcome (win, draw or loss) does not depend on the correlation parameter of the bivariate Poisson distribution. Some models treat the number of goals using bivariate distribution (Maher (1982), Karlis and Ntzoufras (2003)). This approach assumes dependence between the number of goals scored by each team. In a recent work, Karlis and Ntzoufras (2003) proposed an inflated Bivariate Poisson model to account for the excess of draws found in certain championships. This model generalizes the idea of zero-inflated multivariate Poisson models of Li, *et al.* (1999).

Statistical methodologies developed for other sports are also applicable. Since soccer was not so popular in the United States, several researchers studied sports like basketball or rugby. Croucher (1995) faced descriptively the scoring patterns in Rugby League, while Lee's (1999) procedure for a bivariate NBD tried to model the negatively correlated scores taken from the rugby league. He modelled the half-scores in order to simulate the total scores by adding independent Poissons to twice the half scores. Gill (2000) assumed that basketball and American football scores are normally distributed (see Section 2.3.15) and hockey scores vary according to a Poisson distribution. He applied a probabilistic model for late-game reversals by using data from the 1997-98 regular seasons of the NBA, NFL and NHL sports, respectively, and suggested, that the leader of the final period wins the game about 80% of the time.

2.2.2 Some more basic models

The probability that a match between teams i and j will end as a draw was first proved by Keller (1994) to satisfy:

$$P(i\ ties\ j) = \frac{d}{d\lambda} P(i\ beats\ j),$$

where λ is the parameter of the goal scoring distribution of team i. This holds only if the goal scores are Poisson distributed. He used a maximum likelihood method to estimate the probabilities of the possible outcomes in a soccer match and had in mind only the recent results involving the two specific teams. He did not consider at all any home advantage and a lot of data were necessary in order to obtain good estimators.

Kuonen (1996), Lee (1997) and Karlis and Ntzoufras (1998) used the model proposed Maher's (1982) in order to estimate the goal scoring distribution parameters of each competing team but through less data than the Keller's (1994) model needed. Kuonen (1996) noticed that it was better for someone to bet during the end of the season in major European championships like the Italian, the French and the German and for the years 1993-1995. Also, Lee (1997) found for the English Premier League in 1995-1996 that Manchester United could easily have lost the championship or that

Liverpool should have finished at the second place instead of Newcastle United. Karlis and Ntzoufras (1998) observed in the Greek League table for the 1997-1998 season, that Panathinaikos had a higher expected value of points than Olympiakos (the winner), because they had a better attack and defence, but they lost important games against Olympiakos and AEK (the 3rd). Also, four specific teams had a probability of relegation higher than 30% and actually three of them were relegated.

Among several models Karlis and Ntzoufras (2000b) concluded in a basic one, also used by Kuonen (1996) and Lee (1997), which assumes that offensive (a_i) and defensive abilities (d_j) of each team change in home and away games. Let n_{kij} and λ_{kij} be the observed and the expected number, respectively, of the goals scored by team i against team j in the football ground k (1 for home/2 for away), h_k the home/away effect and μ a constant. It is easy to see that this model is equivalent to modeling two distinct models for home and away games:

$$\begin{array}{lcl} n_{ij}^H \sim Poisson\left(\lambda_{ij}^H\right),\ \log\left(\lambda_{ij}^H\right) = \mu^H + a_i^H + d_j^H & \text{where:} & n_{ij}^H = n_{2ij},\ \mu^H = \mu + h_2 \\ n_{ij}^A \sim Poisson\left(\lambda_{ij}^A\right),\ \log\left(\lambda_{ij}^A\right) = \mu^A + a_i^A + d_j^A & & n_{ij}^A = n_{1ij},\ \mu^A = \mu + h_1 \end{array}.$$

Karlis and Ntzoufras (2000b) after facilitating a backward method starting from the full model and removing terms, arrived at the conclusion, that models with interactions $h.\alpha_{ki}$ and $h.d_{kj}$ (indicating that offensive and defensive abilities of each team change in home and away games) are not significantly better than the one of independence. The final ranking and the number of goals scored and conceded by each team were calculated from 24 leagues and were correlated up to 0.85 showing that goal scores can be used to determine the performance of a team. They also found evidence that there is a rather small dependence in the number of goals scored by the two opponents and a small over-dispersion.

Koning *et al.* (2003) estimated the scoring intensities, i.e. the expected number of goals in a complete match. This information was then used as input for a simulation model that computed the probability for each team to win the tournament. The realization of $\left(\mathrm{N}_{\mathrm{ij}}, \mathrm{N}_{\mathrm{ji}}\right)$ gives the result of a group match, where the number of goals N_{ij} scored during 90 minutes of play by team i against team j follows a Poisson distribution $\left(\mathrm{N}_{\mathrm{ij}} \sim P\left(\lambda_{\mathrm{ij}}\right)\right)$. Let also T_{ij} be the waiting time until team i scores during the extra (maximum of) 30 minutes time and that both teams are equally skilled at taking penalties. Then:

$$\left|\begin{array}{l}\Pr(i\,beats\,j)=\Pr(N_{ij}>N_{ji})+\Pr(T_{ij}<T_{ji},N_{ij}=N_{ji},T_{ij}<30)+\frac{1}{2}\Pr(N_{ij}=N_{ji},T_{ij}>30,T_{ji}>30)\\ \Pr(i\,wins\,the\,final)=\frac{1}{S}\cdot\sum_{s}\Pr(i\,wins\,the\,final\,|R=R_s)\end{array}\right.$$

where R_S is a ranking obtained by simulation and S is the number of simulations. In most applications they used for the scoring intensities as estimator an average of goals scored by team i, weighted with the relative quality of team j's defense:

$$\hat{\lambda}^{H}_{ij}=\frac{1}{K^{H}_{i}+K^{A}_{i}}\left(\sum\nolimits_{k}N^{H}_{ik}\frac{\lambda^{A}_{\bullet j}}{\lambda^{A}_{\bullet k}}+\sum\nolimits_{k}N^{A}_{ik}\frac{\lambda^{A}_{\bullet j}}{\lambda^{A}_{\bullet k}}\right),$$

where K^{H}_{i} denotes the number of home matches played by team i, N^{H}_{ij} the goals scored by team i against team j in a home match and $\lambda^{A}_{\bullet j}$ the average number of goals conceded by team j in away games. Koning *et al.* (2003) answered to many questions and except for the European Cup of 1996 indicated the favourites.

2.2.3 Dynamic modelling

A significant extension of the method of paired comparisons appeared by Fahrmeir and Tutz (1994). The pairs (i,j) and (j,i) should not necessarily be the same and $y_{i,j}$ could take other values (such as 3 meaning a draw). Their approach was based on a response model that specified the connection between the observations and the underlying abilities and a transition model that specified the variation of abilities over time. The basic assumption was, that for each team i there exists a latent random utility $U_i=a_i+\varepsilon_i$, where a_i is constant and ε_i is a random variable:

$$y_{i,j}=r\Leftrightarrow\theta_{r-1}<U_j-U_i<\theta_r,$$

where $-\infty=\theta_0<\ldots<\theta_k=\infty$ are thresholds. Assuming a continuous distribution $F(\cdot)$ for the differences $\varepsilon_i-\varepsilon_j$, a general ordinal paired comparison model yields:

$$P(y_{i,j}=r)=F(\theta_r+a_i-a_j)-F(\theta_{r-1}+a_i-a_j)$$

and after introducing time dependence, the simplest response model was:

$$P(y^{(t)}_{i,j}=r)=F(\theta_{t,r}-a_{t,i}-a_{t,j})-F(\theta_{t,r-1}-a_{t,i}-a_{t,j}),\forall t,i.$$

For more complex cases they used a Kalman filter for paired comparisons. The results from an application to soccer data of the German Bundesliga were impressive showing a remarkable fit to the variation of abilities of each team through time.

Dixon and Coles (1997) noticed that almost any model up to that time was assuming constant performance rate through time. Their idea was to introduce a dependence parameter ρ, because they believed that team performances were varying through time. Furthermore, they observed the dependence between low scores such as 0-0, 1-0 or 0-1. They modified Maher's (1982) model to:

$$\Pr\left(X_{i,j}=x,Y_{i,j}=y\right)=\tau_{\lambda,\mu}(x,y)\frac{\lambda^{x}\exp(-x)}{x!}\frac{\mu^{y}\exp(-y)}{y!}$$

It was supposed that teams i and j scored x and y goals respectively, $\lambda=a_i\beta_j\gamma$, $\mu=a_j\beta_i$, $\max(-1/\lambda,-1/\mu)\le\rho\le\min(1/\lambda\mu,1)$ and they used the definition:

$$\tau_{\lambda,\mu}(x,y)=\begin{cases}1-\lambda\mu\rho & , \quad if\ x=y=0\\ 1+\lambda\rho & , \quad if\ x=0,y=1\\ 1+\mu\rho & , \quad if\ x=1,y=0\\ 1-\rho & , \quad if\ x=y=1\\ 1 & , \quad otherwise\end{cases}$$

They took all the scores (x_k,y_k) for the likelihood function under one crucial assumption; a team's performance is likely to be more closely related to their performance in recent matches than in earlier matches, so the parameters are locally constant through time. For every time point t they constructed a 'pseudolikelihood':

$$L_t(a_i,\beta_i,\rho,\gamma;i=1,\dots,n)=\prod_{k\in A_t}\left\{\tau_{\lambda_k,\mu_k}(x_k,y_k)\left[\exp(-\lambda_k)\lambda_k^{x_k}\right]\left[\exp(-\mu_k)\mu_k^{y_k}\right]\right\}^{\phi(t-t_k)},$$

where $\lambda=a_{i(k)}\beta_{j(k)}\gamma$, $\mu=a_{j(k)}\beta_{i(k)}$, $A_t=\{k:t_k<t\}$ and t_k corresponding to the time that match k was played. Among several choices of the non-decreasing function of time ϕ, they worked with $\phi(t)=\exp(-\xi t)$ down-weighting exponentially all previous results according to a parameter $\xi>0$. By the maximization of L_t, they calculated the score probabilities and they estimated the probability of a home win in match k as:

$$p_k^H=\sum_{l,m\in B_H}Pr(X_k=l,Y_k=m),$$

where $B_H = \{(l,m): l > m\}$. The final step was to take the probabilities p_k^A, p_k^D of an away win and a draw respectively, $\delta_k^H = 1$ for a home win and 0 otherwise and define:

$$S(\xi) = \sum_{k=1}^{N} \left(\delta_k^H \log p_k^H + \delta_k^A \log p_k^A + \delta_k^D \log p_k^D \right).$$

They concluded that the teams' performances are genuinely dynamic.

A further improvement of the models presented by Maher (1982) and later on by Dixon and Coles (1997) came by Dixon and Robinson (1998). They thought of the home-away scoring process as a two-dimensional birth process:

$$\begin{cases} \lambda_k(t) = \lambda_{xy}\lambda_k \\ \mu_k(t) = \mu_{xy}\mu_k \end{cases},$$

where λ_{xy} and μ_{xy} (for $x, y = 0,1,\dots$) were the parameters that determined the (homogeneous) scoring rates during which the score is (x,y). They had used λ_k and μ_k given by Dixon and Coles (1997) for match k at time t and score (x,y) and denoted the goal times in match k by:

$$(t_k, J_k) = \left\{ (t_{k,l}, J_{k,l}) : \ l = 1, \dots, m_k \right\},$$

where $m_k = x_k + y_k$ and $t_{k,l}$ are the total number of goals and the time of the l-th goal in match k respectively and $J_{k,l}$ is 0 for a home goal and 1 for an away goal. An extension was made to model injury time. Since there were not available data showing how much injury time was added, goal times over 45 and 90 minutes were considered as (possibly) censored observations. They introduced the parameters ρ_1 and ρ_2 for the multiplicative adjustment to the scoring intensities over the extra times. Thus, the home scoring rate (similarly for the away scoring rate $\mu_k(t)$) was taken to be:

$$\lambda_k(t) = \begin{cases} \rho_1 \lambda_{xy}\lambda_k, & for\, t \in (44/90, 45/90] \\ \rho_2 \lambda_{xy}\lambda_k, & for\, t \in (89/90, 90/90] \\ \lambda_{xy}\lambda_k, & otherwise \end{cases}$$

The best-fitted model was the one that added variation parameters ξ_1 and ξ_2 and allowed for intensities a linear change through time:

$$\begin{cases} \lambda_k^*(t) = \lambda_k(t) + \xi_1 t \\ \mu_k^*(t) = \mu_k(t) + \xi_2 t \end{cases},$$

and λ_{xy} (similarly for μ_{xy}) was defined to be:

$$\lambda_{xy} = \begin{cases} 1, & for\, x = y = 0 \\ \lambda_{10}, & for\, x - y \geq 1 \\ \lambda_{01}, & for\, x - y \leq -1 \\ \lambda_{21}, & for\, x - y \geq 1, x \geq 2 \\ \lambda_{12}, & for\, x - y \leq -1, y \geq 2 \end{cases} .$$

They found a continuously increasing rate for both teams, perhaps due to tiredness of players, which leads to mistakes in defending. The attack and defense parameters decreased and increased respectively from a higher to a lower division. The scoring rates of home and away teams depended on the current score, especially when the home team had a small lead. No evidence for the immediate strike back was found.

Another paper based on Lee (1997) and Dixon and Coles (1997) was published by Rue and Salvesen (2000), who used Markov Chain Monte-Carlo (MCMC) methods and took the modified model:

$$\begin{aligned} \log(\lambda) &= c^{(x)} + a_i + d_j - k\Delta_{i,j} \\ \log(\mu) &= c^{(y)} + a_j + d_i + k\Delta_{i,j} \end{aligned} ,$$

where $\Delta_{i,j} = \dfrac{a_i + d_i - (a_j + d_j)}{2}$ is the difference in strengths between the two teams and $k > 0$ is a small constant for the intensity of the psychological effect, in which the stronger team i underestimates the underdog j. They also used $\tau_{\lambda,\mu}(x, y)$ defined by Dixon and Coles (1997) with $\rho = 0.1$. The next step was to truncate the Poisson distribution after 5 goals and they proposed the following robust model:

$$\pi(x, y | \lambda, \mu) = (1 - \varepsilon) \cdot \pi^{*}(x, y | \lambda, \mu) + \varepsilon \cdot \pi^{*}\left(x, y \middle| \exp\left(c^{(x)}\right), \exp\left(c^{(y)}\right)\right),$$

where π^{*} was the resulting truncated law. Finally, they introduced time through a Brownian motion. They found many interesting things for the English Premier League 1997-1998 season; Manchester United should be the champions and not Arsenal and Aston Villa could easily have finished in the 15th position instead of the 7th.

Knorr-Held's (2000) main concern was rating and had in mind that recent results have more influence in estimating current abilities than earlier results. In his cumulative link model for ordered responses, the latent parameters represented the team strength. These were allowed to evolve through time according to specific constrained random walks with independent normal increments. This model treated all teams symmetrically. Posterior mode estimators of the abilities were calculated with an extended Kalman filter together with an *ad hoc* method for a variance parameter. For the German Bundesliga in the period of 1996-1997, he found different patterns in the estimated abilities for the various teams and interesting temporal trends.

Crowder *et al.* (2002) presented an autoregressive model $AR(1)$:

$$\gamma_{it} - \gamma_{i0} = R\left(\gamma_{i,t-1} - \gamma_{i0}\right) + u_{it}$$

for the underlying parameters $\gamma_{it} = \left(\gamma_{a_{it}}, \gamma_{\beta_{it}}\right)^T$. The attack α_{it} and defence β_{it} abilities of each team i at time t were expressed in terms of a basic set of unconstrained parameters. Also, $R = \begin{pmatrix} \rho_{\alpha\alpha} & \rho_{\alpha\beta} \\ \rho_{\beta\alpha} & \rho_{\beta\beta} \end{pmatrix}$ was a 2×2 matrix of auto-regression parameters, u_{it} were independent $N_2(0, \Sigma)$ innovations and γ_{i0} was the base-line value towards which γ_{it} was drawn, if R was small. The two approximation were:

$$\alpha_{it} - \alpha_{i0} = R\left(\alpha_{i,t-1} - \alpha_{i0}\right) + u_{it} \quad \text{and} \quad \begin{cases} \alpha_{it} = m_{it} + \sum r_{ijt} \\ \alpha_{jt} = m_{jt} + \sum r_{jit} \end{cases}$$

or if teams i and j did not play at time t:

$$\begin{cases} \alpha_{it} = m_{it} \\ \alpha_{jt} = m_{jt} \end{cases},$$

where $m_{it} = \alpha_{i0} + R\left(\alpha_{i,t-1} - \alpha_{i0}\right)$ and r_{ijt} represented the discrepancies between the home and away goals and their expected values. Hence, the second approximation along with the reduced formula constituted the formal model for $\left(a_{it}, \beta_{it}, a_{jt}, \beta_{jt}\right)$ without losing too much predictive power in comparison with the $AR(1)$ model. This new method for prediction was computationally fast. They focused on modelling the 92 soccer teams in the English Football Association League and compared the results with those of Dixon and Coles' (1997) model. They could both predict about the same home wins, but the model of Crowder, *et al.* (2002) was slightly better for away wins. Dixon and Coles' (1997) method started poorly but improved very quickly and became better than the approximation method around week 210.

2.3 Other aspects of modelling

2.3.1 Is chance more important than skill?

Reep and Benjamin (1968) came up to a ‘historical’ conclusion using the famous in literature ‘r-pass movement’. This was nothing more than the success of exact r passes among players of the same team during a match. A “0-movement” pass was defined in situations like penalties or when the first attempted pass was intercepted. The probability of an r-pass movement was given by:

$$P(r)=\left[p_1 \cdot p_2 \cdot \ldots \cdot p_j \cdot \ldots \cdot p_r \cdot \left(1-p_{r+1}\right)\right],$$

where p_r is some function with $p_1 > p_2 > p_3 > \ldots > p_r > p_{r+1}$. In other words, one would expect p_1 to be fairly high (but less than unity) and p_r to fall rapidly to some low value beyond which there is little further decrease; a form like this function’s behaviour is the exponential one. They suggested that, if the probability of a pass succeeding varies from pass to pass throughout an attempted chain, from chain to chain and from game to game, then the distribution of length of pass-chains would be approximately the NBD or a compound Poisson. The most known result was that ‘*chance does dominate the game*’.

In a next paper Reep, *et al.* (1971) modified that suggestion. They said that the probability of pass-success not only varies from player (chain initiator) to player (another chain initiator), but also from throughout the chain, from one period in a game to another and from game to game. Therefore the underlying distribution was not an *exact* compound Poisson. They excluded the goals shots arisen from penalties or interceptions, which were erroneously included in the count of 0-pass move. The probability of “success” is relatively invariant over movements and between the players. The new result was that skill supplants chance, when the player is in the shooting area, where he has two choices: passing or shooting.

Hill (1974) confirmed, in a sense, the latter. He compared expert forecasts’ tables before the beginning of a specific season to the tables of the final season’s results and proved statistically, that throughout a whole season and not in each game, skill supplants chance. This statement sparked off tries to discover ways of predicting the outcome of soccer matches in long runs (rather than in short ones).

2.3.2 First team players

It is reasonable for a coach to start a game with his best players, or to be more precise, with the players that are high skilled and well exercised at the same time. The physical characteristics that the starters have, is significant to be discovered. Thus, the players of the bench who lack on such characteristics should work on these, the trainers could improve their work and generally a team would be built more carefully.

Snyder *et al.* (1998) were interested in finding a way to classify individuals' features based upon some physical variables, for each of the different field positions. They presented several approaches, where efficient and realistic imputation algorithms were sought. Any information that measured the examined physical characteristics could be used for imputation. Cluster analysis gave bad results, projections and linear DA were not so good, but Logistic DA or CART showed that the variables were indeed divided into subsets that represented these attributes.

2.3.3 Playing strategies: the model of Pollard and Reep (1997)

Any coach would like to possess all possible information about his players' abilities. Many data considering a variety of moves during matches had been collected from Church and Hughes (1987), Franks (1988), Paukku (1994) and some others. These data could consist a unit of measurement. Ali (1988) and Pollard *et al.* (1988) were the first who analysed such collected data. A typical example for studying the different strategies is the ball possession. A 'good' ball possession has more chance to produce a goal. Olsen (1988) and Pollard (1995) studied the case of weighting, since each shot has different probability to score.

Ball possession had been in the work of Pollard and Reep (1997) the basic unit of measurement. They assigned each shot a weight according to its estimated probability p of scoring. For example shots from central locations in the penalty area were on average 15 times more likely to end up in goals than shots from outside. The pitch was divided in 6 zones starting from the defence towards the attacking line (see Figure 2.3.1). The zone, in which the ball possession started, was recorded. There were two different types of possession; the 'set play' (such as a free kick or a corner) and the 'open play'. For possession of type j starting in zone i, the probability of scoring a goal was estimated as:

$$p_{ij} = \sum_{k=1}^{n_{ij}} p_{ijk} / n_{ij}, \; i = 1,2,\ldots,6, \; j = 1(open\,play) \; or \; 2(set\,play),$$

where the k^{th} ball possession p_{ijk} was basically the weighted value shot and n_{ij} the total number of team possessions of type j originated in zone i.

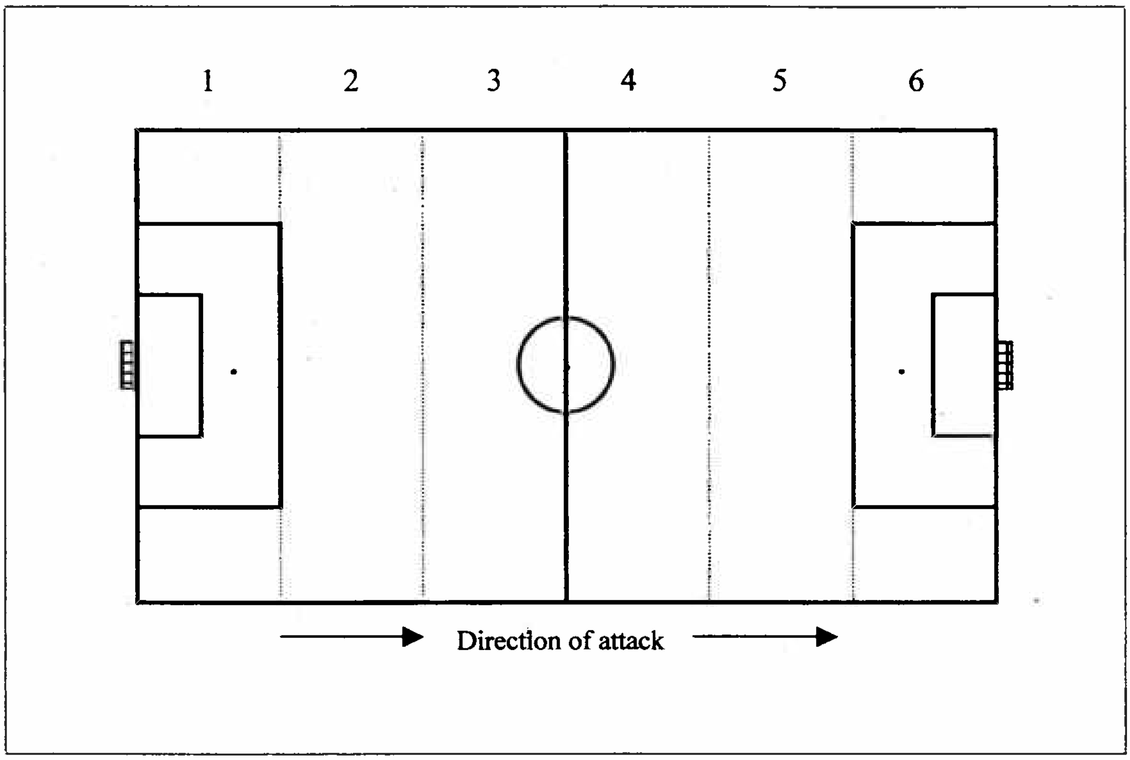


Figure 2.3.1: Division of the field of play into six zones
(taken from Pollard and Reep (1993))

In Tables 2.3.1 and 2.3.2 there are the results from the World Cup of 1986 in Mexico. From such tables it was feasible to compare or even to implement different strategies. They defined the 'yield' of a ball possession using estimated probabilities based on the scoring of a goal through the logistic regression analysis as the estimated probability of scoring a goal minus the estimated probability of conceding a goal, based on the outcome of the possession. The yield is easier to interpret when given as rate per 1000 ball possessions and the values give then the net yield in goals per 1000 ball possessions. For example, a yield of 3.5 (see Table 2.3.2) means that for every 1000 ball possessions a team expects to score 3.5 more goals than it concedes. Negative yield values mean that, on average, more goals were conceded than scored.

Zone of origin	**No. of team possessions for the following types of play:**		**Yield for the following types of play:**	
	Open play	Set play	Open play	Set play
1	865	651	5.9	2.2
2	822	244	8.5	0.5
3	837	321	6.2	2.2
4	473	450	10.9	8.5
5	318	336	24.8	12.6
6	111	416	78.3	18.0

Table 2.3.1: Yield per 1000 team possessions, classified by zone of origin and type of play (taken from Pollard and Reep (1993))

Situation	**Strategy**	**n**	**Yield**
Goal kick	Long	99	-2.7
Throw-in in own half	Short	276	-0.2
Possession in zone 4	Short passing only	1372	11.1
	Running with the ball	288	16.3
	Long forward pass	148	23.1
Free kick in zone 5	Direct shot	60	12.5
	Other	143	16.8
Throw-in in zone 6	Short	98	3.5
	Long towards goalmouth	32	21.7
Centres from zone 6	Above waist height	240	33.3
	Below waist height	103	96.6

Table 2.3.2: Yield per 1000 team possessions from playing strategies in different situations (taken from Pollard and Reep (1993))

2.3.4 Strategies and levels of measurements

In all soccer tournaments worldwide, when two or more teams finish in the same position, then there are some rules that determine the final ranking. In the Premier League of English Football the number of wins and draws firstly determines the positions. If the teams have the same number of points, then the goals scored and concede are examined. The team with the highest goal difference gets the highest place. If there is still problem, then the best team is the one that scored more goals. Otherwise, if the teams cannot be ranked under these rules, they play each other. The question is, if this is the right way to determine the final League positions or instead there exists another better measure.

Croucher (1984) noticed that such a change could affect team strategy, because there would be different motivations to raise or to lower the total number of goals in a game. Wright (1997) mentioned that this particular tie-breaking mechanism assumes an interval scale, since a 3-1 victory is the same as a 6-4 victory. Many fans support that a ratio scale might be better; then a 3-1 score would be of the same value as a 6-2 score. According to the ratio tie-breaking mechanism, a team loosing 3-1 would try to score, because the ratio would be doubled with only one goal scored. So, the loosing team would be more offensive than the winning.

2.3.5 Feeling stronger at home

The first two approaches that tried to quantify the home ground advantage (HGA) came by Stefani (1980) and Pollard (1986). The first one used the formula:

$$w_{ij} = u_i - u_j + h_i + \varepsilon_{ij}, \forall i \neq j,$$

where w_{ij} is the goal margin in a match with team i playing at home against team j, u_i and h_i are measures of team's i ability and HGA respectively and ε_{ij} is a zero-mean random error. This model was often used later on by many researchers, such as Clarke and Norman (1995) and Kuk (1995). On the other hand, Pollard (1986) just counted the number of matches won by home teams as a percentage of all games played and did not take in mind the different skills. This was acceptable only in the case that each team had about the same abilities. Both agreed in the existence of HGA.

Someone might wonder what causes a HGA. Courneya and Carron (1992) could observe it but not answer to the question. Clarke and Norman (1995) used least squares for soccer data in English football and produced a HGA effect for each team in addition to a team rating. They showed that a team's HGA varies from year to year. Some teams have negative HGA, which it may have greater effect on winning than on goal difference. On average it is worth just over half a goal. There is some evidence for club effect, but this is not indisputable. A possible reason of its existence might be the geographical distance between the two teams' origins. None-the-less, it is almost the same in all divisions agreeing with Dowie (1982). So, because the top teams have more fans and belong to the high divisions, the audience's size does not affect HGA.

Bland (1995) calculated the correlation between home points and attendance, away points and attendance, home goal difference and attendance, and away goal difference and attendance. He found that the difference between home points and away points was positively related to average attendance. Therefore, the audience's size is a cause of HGA. Bland and Bland (1996) disagreed openly with the earlier argument of Clarke and Norman (1995). They believed that the situations are not the same; for a player in lower division playing in front of a crowd of 10,000 is might be the same as for a player of major division playing in front of 30,000 fans.

2.3.6 Empty and crammed tiers

It is very common the late years that people go more infrequently to stadiums to watch soccer games. Generally, there is variability in the crowd's size from match to match. When a match is indifferent form point view or a big favourite confronts an underdog, less fans watch it. On the contrary, if there is big uncertainty about the outcome or need for points, then the interest increases and is maximized in a total uncertain future result. Peel and Thomas (1988, 1992) thought that bookmakers' odds might contain information about the connection between attendance demand and uncertainty of the outcome. They found a U-shape relationship between attendance and the home win probability odds, because more fans want to watch a game when either the home or away team stands high in the League table.

Forrest and Simmons (2002) interpreted this result by saying that as the teams' chances of winning grow less equal, attendance generally falls away. More spectators are attracted to matches where the prospects of the competing teams are evenly balanced. They used a two-stage model and data from 1997-1998 bookmakers' odds in England in order to measure this uncertainty, but they allowed for the possibility that these are biased predictors of the outcomes of the matches. In the first stage a latent regression generates the probability of a win for either side. The second stage is just a simulation model. They found that it would be inappropriate to assume efficiency when modelling attendance demand, because that particular year the soccer betting market was not fully efficient. Also, the paradox was that, even though the fans prefer well-balanced games, this might run the risk of lowering attendances.

2.3.7 Artificial pitch surface

There are some few teams worldwide that play their home games in an artificial pitch, but it is observed that the number of such teams increases in recent years. Barnett and Hilditch (1993) examined the possible presence of any HGA due to this in the four divisions of the English Football League for a period of 10 years. In Table 2.3.3, it seems that teams with an artificial pitch won much more games playing at home than the others and succeed to reduce a lot the percentage of losses.

Pitch	Percentages for the following venues:					
	Home			Away		
	Win	Draw	Lose	Win	Draw	Lose
Natural	47.7	27.1	25.2	24.8	27.0	48.2
Artificial	57.7	25.1	17.1	25.1	26.2	48.6

Table 2.3.3: Percentages of game outcomes
(taken from Barnett and Hilditch (1993))

In away games there are not any significant differences in the patterns of the three possible outcomes for the two groups. After analysing different measures of (relative) performance, they concluded that there exists a statistically significant advantage for teams employing an artificial pitch, when playing at home. A large variety of alternative explanations for the differences in performance were rejected.

2.3.8 Dismissal of a player

The biggest punishment in a soccer match by a referee is to show a player the red card and to disqualify him for the rest of the game and thus, the team that loses a player, finds it difficult to win. Ridder *et al.* (1994) stated that the sooner a team loses a player the more probable is to lose the game too, because the according probability increases considerably (see Table 2.3.4). Three assumptions were made:

1. The two teams scored according to two independent Poisson processes.
2. The scoring intensities' ratio of the two full teams is a constant for each game.
3. After the red card the scoring intensities vary (and actually increases as shown by Morris (1981)) over time.

Minute of red card	Probability		
	Team of 11 wins	Draw	Team of 10 wins
0	0.65	0.17	0.18
15	0.62	0.18	0.20
30	0.58	0.20	0.22
45	0.54	0.21	0.25
60	0.49	0.23	0.28
75	0.44	0.24	0.32
90	0.375	0.25	0.375

Table 2.3.4: Probabilities of the Outcome of the Match by Minute of the Red Card (taken from Ridder *et al.* (1994))

Ridder *et al.* (1994) estimated the effect of the red card by linear regression. Their estimator was based on a comparison of the number of goals scored by the same team before and after the red card. They noticed that the goals after the red card increased and usually this was given to the already weaker team. Also, if a player had instantly to make decision to risk a red card, there is a unique time moment in the game at which the optimal action of the defender changes; after that moment, he should trip up the opposing player (see Table 2.3.5).

Relative strength of teams	Probability of score		
	0.30	0.60	1.00
0.5	70	42	0
1	71	48	16
2	72	52	30

Table 2.3.5: Time (Minute of Game) after which a defender should stop a breaking-away player by probability of score and relative strength of the defender's team by minute of the red card (taken from Ridder *et al.* (1994))

2.3.9 When the coach foots the bill

Most of the times, when a team has a run of bad results, the board decides to fire the coach expecting to reverse the situation with a new one. It is common belief that the new coach will motivate the players better, and therefore improve the results. One could compare the new with the old coach, but there are several difficulties. For example, how the performance should be measured? Does the moment of dismissal play any role? And of course, it is natural that the two coaches do not face the same conditions or equal teams in strength.

Brown (1982) used as measure of performance the percentage of wins. The change of a coach cost about 11% in the percentage of games won; in a season of 14 games, it cost a little bit more than a one game won during the season. The board prefers a tactic like this in order to appease fans and press media, because it is more difficult to hire new players during the season. The findings of Van Dalen's (1994) model were that all coach changes have a positive effect on the goal difference and the effect is significantly positive in the three of the five cases examined. The decision whether or not to fire a coach is significantly positive related to the ranking for almost all teams in baseball and basketball according to Scully (1995).

On the other hand, Koning (2003) noticed that there are times that a team performs worse with the new coach. His model included the non-constant quality of a team and the also non-constant HGA and could separate the defensive from the offensive skills. He did this separation, because he felt that a new coach would try not to lose at first and not necessarily to play well. An extra defensive effort might reduce the offensive efforts and therefore we may not see any change at the goal difference. Hence, he found that the defensive performance did improve, but in 11 out of 28 coach changes the quality of a team and HGA decreased in general. Not even a temporary improvement exists.

2.3.10 Balance

Koning (2000) used a probit model to assess whether the balance in competition in Dutch professional soccer had changed over time. He defined a soccer league to be in perfect balance for a certain year, if the probability that any team won a home game did not vary with the opponent or with the team. The conclusions were born out by three different measures of balance. Contrary to popular belief, the balance has not been changed much for about the past 30 years.

2.3.11 The structure of a tournament

Let us consider a tournament with groups of teams in the first Round Robin round, where the best two teams qualify to the next round in a knock-out procedure up to the final. There are two ways of continuing; either after a lottery-pick or based on standard fixed structure. McGarry and Schutz (1994) thought the case, where in the first part the teams are separated into 6 groups according to the FIFA/Coca-Cola World ranking system, but the first two seeds are reserved from the actual holder of the Cup and the host team. It was shown, that the (almost sure) promotion of the host team to this high seed does not affect the final winner. In the second part, the seeding depends on the first round's results. Because the tournament was not balanced, they gave a rating score to each team and used a paired comparison model with the help of a Monte Carlo procedure for the simulation. They found that this structure was not very fair. Due to the seeding of the knockout part the last two groups were underprivileged. The first and the third groups were preferable from the rest.

Marchand (2002) considered a classic 16-team knockout tournament. The standard method usually gives big advantage to the higher seeded teams. In the random method the initial structure is totally at random and the higher seeded teams may lose this advantage. He focused on the probability that a top ranked team finally wins and how this probability may vary between the two cases. He indicated that the outcomes of the standard knockout tournament and the random knockout tournament might not differ as much as one expects. The advantage of the standard draw for the top seeded team may be generally overestimated.

2.3.12 Long spells of same football results

Dobson and Goddard (2003) investigated the issue of persistence in sequences of consecutive match results. They followed Koning's (2000) modelling approach and used a Monte Carlo analysis to test for short-term persistence effects in the presence of team heterogeneity. The conclusion was that the hypothesis of non-persistence could not be rejected in cases of sequences of consecutive losses and sequences of consecutive matches without a loss. On the contrary, in sequences of consecutive wins or consecutive matches without a win it seems to exist a negative persistence.

2.3.13 The betting market

Many of the already mentioned papers have applications in the betting market. Each bookmaker employs a panel of selectors to create predictions, which are viewed as unbiased by an average bettor. Usually a fixed percentage of the betting money returns to the bettors and the rest remain to the bookmaking agency (perhaps the government) for various expenses and sport programs. The amount of money spent is fixed, but the possible profit can be fixed or variable. Also it seems to exist a significant negative correlation between the cost per bet and the yearly amount spent per bettor. If the cost of a bet rises, then the bettor spends less money and tries to find fewer but more accurate bets. A simple and not expensive betting scheme appears to provide increased gross revenue.

Stefani (1980) observed some tendencies in the outcomes of the games from 6 major National Championships in Europe and 3 European Cups; on average the possible outcomes were separated as 49% home wins, 27.8% ties and 23.2% home losses, 2.73 goals were scored per game and playing at home gave an advantage of 0.47 goals per game. Certain nations deviated a lot from those averages and this must be taken in mind, if a betting is to be designed. Stefani (1980) considered the betting strategies from two approaches and tried to optimize them. In the first one, patterns of selections are chosen independently from the teams that are playing. The best random pattern consists of all home wins alternated one at a time with ties. The bettor could wait for whole years to gain some money, which also depend on payoffs. In the second one, a least square method is used to make predictions and the selections are made based upon the previous performances of the teams. It is necessary to permutate selected home wins with home losses when a tie is predicted. The latter approach provides better short and long term return than the former one.

In a subsequent paper, Stefani (1987) explained the gambling on American football games. He compared four different estimators; James-Stein's, Harville's, the Least Square and the Weighted Least Squares. The first two are biased, while the other two are unbiased. No statistically significant differences were found. Each one lagged about 1% behind a typical sports book in selecting the winning team and about 0.25 points per game in average absolute error. He supported that little or no profit is possible, if a bet is placed on every game. The bettor should be selective in order to make some profit, because he would be fortunate to be 0.524 accurate. Only few, highly skilled and selective bettors can reach a level of 0.55 or maximum 0.60 accuracy against the sports book.

Index betting is a recent way for betting on sports. Jackson (1994) dealt with this new area and tried to make a link with finance (see also Section 2.4). He pointed out the similarities and the two main differences between index betting on sports and gambling on the future price of stocks. The first is that in sports the event takes place and this determines a final value for the index, often with much bigger variability than stocks. The second is the modelling; in sport there is a wide range of applications due to the quantity of data and models, while in the stock market the models that describe the underlying process are extremely few and complex.

Boulier and Stekler (2003) were interested to find whether the expert's predictions of National Football League games were more accurate than those that would have been solely on the rankings. They compared some methods of forecasting and their analysis derived solely from the rankings based on 'power scores'. They predicted the probability that a higher ranked team will win by estimating a probit model. A team that is ranked 1 position higher than its opponent has only slightly better than even chance of winning, while a team that is 30 positions above its opponent should win more than three-quarters of the time. After that, they used the technique of recursive regression, where the estimations were made from the data available from the first 6 weeks. From week 7 and then, the probits were weekly updated so as to predict the outcomes of every next week. According to 'Brief score' the betting market was the best predictor, the recursive probit model was the second best and the sports editor was the worst. They concluded that the statistical model of the sports editor's forecasting procedure yields slightly superior forecasts to the actual forecasts in a 'real world' situation for a large sample of predictions and that the information contained in the betting market is the best predictor of the outcomes.

2.3.14 Genetic and Neuro Tuning

Prediction of the outcome of a match has been a hot theme the late years for everyone who watches sports. Many models and PC-programs have been developed for this cause and most of them use stochastic methods of uncertainty description. Recently, some models appeared that use neural networks for the results of football game predictions and deal with non-linear dependencies.

Rotshtein *et al.* (2003) proposed a model with two phases. In the first phase, they define the fuzzy model structure, which basically uses information about both teams' previous games results. The second phase consists of fuzzy model tuning. It is based on the method of identification of the non-linear dependencies trained by experimental data 'past-future'. They use fuzzy IF-THEN rules and for tuning they combine a genetic algorithm with neural network.

2.3.15 Brownian motion

A last special case, which also responds to the question of Section 2.2.1, contains sport scores (such as basketball scores) that can reasonably be approximated by a continuous distribution. Stern (1994) wanted to estimate the probability that the home team wins the game given that they lead in score in specific time moment.

At the beginning, he transformed the time scale to the unit interval. Then he represented by $X(t)$ the (positive or negative) lead of the home team at time t and assumed that it could be modeled as a Brownian motion process with drift μ and variance σ^2 per unit time:

$$
\begin{aligned}
&X(t) \sim N(\mu t, \sigma^2 t) \\
&X(s) - X(t), s > t, \textit{ is independent of } X(t). \\
&X(s) - X(t) \sim N(\mu(s-t), \sigma^2 (s-t))
\end{aligned}
$$

The probability that the home team wins a game is $\Pr(X(1) > 0) = \Phi(\mu/\sigma)$, where Φ is the cumulative density function of the standard Normal distribution. If the home team is leading (or losing) by l points at time t, then under the random walk this probability is:

$$\Pr_{\mu,\sigma}(l,t) = \Pr(X(1) > 0 | X(t) = l) = \Pr(X(1) - X(t) > -l) = \Phi\left(\frac{l + (1-t)\mu}{\sigma\sqrt{(1-t)}}\right).$$

His application to 493 games from the 1991-1992 National Basketball Association (NBA) season indicated that the Normal distribution appears to be a satisfactory approximation to the distribution of score differences in each of the four quarters. Surprisingly, the Brownian motion model is well applicable to baseball too.

2.4 Economic implications

Professional soccer is often used as a test for the validity of some econometric models. Success might be seen in both ways; from the team's ranking in the final League table and from the annual report of the 'team company' at the end of the season. According to Dobson and Goddard (1995), teams that have been in the League a long time are found generally to enjoy higher attendances. However, teams from towns with a high proportion of professional and managerial employees are not much affected by price changes, or indeed by the team's form. In addition, there exist a 'loyalty' factor, which seems to affect all clubs, regardless of position or status.

An econometric model appeared by Szymanski and Smith (1997). A competitive market is supposed to exist, where football skills can be bought. A team's position in the League is determined by the skill they managed to buy and fixes in advance the yield of this investment. Each team's aim is to reach the highest possible ranking, because improvements in League position trade profit. This model performs reasonably well and it shows indeed that some of the top teams make profits, while the rest suffer losses. It can also estimate how much money is needed by a team to move up the League.

Szymanski and Smith's (1997) work could be extended to several fields of soccer industry. One of them is the player's transfer fee. A transfer fee is defined to be the total amount of money one team must pay to another, so as to obtain the services of a player. Dobson and Gerrard's (1997) model confirmed that the selling team is able to make profit through the difference between the value placed on the transferred player and the reserve price for the player.

Many financiers study the effectiveness and consequences of incentives (we took a first glimpse of motivations in section 2.3.4). A key for strategic behavior is the pay-offs, especially when these change. Dewenter (2003) believed that sports are appropriate area of such a study, since there are competitive situations, strategic behaviors and, of course, available data. He focused his analysis on the effects of changing the pay-off system by the FIFA. He applied panel count data techniques so as to examine, if the outcomes of the matches had changed after introducing a three-point system in the Portuguese first division. He found that the transition to this system resulted in less attractive matches, because the underdogs defended even more than before. None the less, only the scores had been affected and not the outcomes.

Actually, this had negative effect on the home advantage; the home goals reduced more than the away goals and the goal difference decreased.

The efficiency of financial markets is recently tested through the efficiency of sports betting market and the predictability of match results. Gandar *et al.* (1988) proposed several economic tests, which select bets purely based on the teams' past performances and attempt to exploit certain hypothesized patterns of the public. Any biases in the bookmakers' odds are assumed to last long enough to be detected and exploited by bettors. Goddard and Asimakopoulos (2004) proposed a forecasting model based on ordered probit regression, whose main advantage derives from its ability to predict matches played in the closing (and perhaps at the starting) stages of the season. It contains additional information that is not impounded into the bookmakers' odds, and that the latter are weak-form inefficient. A fine strategy for exploiting inefficiencies in the bookmakers' prices is to place those bets, which the model assesses to be good value at specific times in the season. Such a strategy appears capable of generating a positive return.

CHAPTER 3

Weighted maximum likelihood estimation

Our approach will be based on modeling the scores between the competent teams and not the final outcome (win, draw or loss). The basic problem is that an "unusual" score could produce invalid results. For example, if we examine the strength (i.e. the defensive and the offensive ability) of a specific team and we observe a win with the "unusual" score of 7-0, it is obvious that this team will appear stronger than it really is. We know that the number of goals scored by a team is a sufficient indicator for the strength of a team, since it must score in order to win. Also a high number of goals scored leads to a high final position and this is not necessarily true. Karlis and Ntzoufras (2000a) proved that there is a high correlation between the final ranking and the number of goals scored and conceded by each team.

The need for another way of score modeling, which will give more robust estimates, is clearly obvious. In Section 3.1.1 we shall see that Grunert and Fieller (1995) noticed that there are two main types of distortion from the model in use and its assumptions; data contamination and model deviation. Our main concern will be the location of any possible outliers and then to treat them with a special manner. There are several suggestions in the literature, such as rejection. We shall attempt to conclude to more robust estimates by finding and giving the observations the appropriate weights, i.e. unexpected scores will be down-weighted.

This section presents some of the basic robust theory including the several approaches made by Huber and another basic one called "the Infinitesimal approach". This theory is used in order to introduce the weighted maximum likelihood method and among several researches we focus on Lindsay's (1994) approach. We also mention the importance of the generalized linear models and how robustness is involved. The section highlights a new algorithm and the relevant equation known as "iterative reweighting least squares" algorithm and "iteratively reweighted estimating equations" respectively, and concludes with a recent developed area in scientific research, the MM algorithms.

3.1 Robust statistics

3.1.1 General

Hampel *et al.* (1986) gathered and presented most of the hitherto known theory in robust statistics. Of course, this area of statistics is huge of its own, so we shall try to restrict as far as possible only to the essential parts of it, which will be used in the following for our work. In this Section and mainly in Section 3.1.4 we shall point out some fundamental aspects about robust theory already used in their book, “Robust Statistics” (1986).

It is very common in statistics to exist several assumptions that are approximations to reality. The problem with the theories of classical parametric statistics is that they derive optimal procedures under exact parametric models, but say nothing about their behaviour when the models are only approximately valid. A possible approach would be to replace a given parametric model by another one and to enlarge it to a “supermodel” by adding more parameters. This basic idea made Hampel *et al.* (1986) declare that robust statistics, as a collection of related theories, is the statistics of approximate parametric models. Robust statistics deal with several known areas, such as the rejection of outliers or even the violation of the independence assumption. It should not be confused with nonparametric statistics, because neighborhoods of parametric models are considered.

The need for a different approach ensued from the fact, that the theories of parametric models were only approximations to reality. Also, certain central limit theorems gave information about an imaginary limit under certain assumptions and could not clarify how far we are still away from that limit or whether the assumptions are fulfilled. There are four types of deviations from strict parameters models:

1. The occurrence of gross errors.
2. Rounding and grouping.
3. The model may have been conceived as an approximation anyway.
4. Apart from the distributional assumptions, the assumption of independence may only be approximately fulfilled.

We should mention here, that gross errors mainly occur as copying or keypunch errors and they are the most frequent reasons for outliers.

According to Grunert and Fieller (1995) there are basically two types of distortion; data contamination and model deviation. The first one is given where the model corresponding to the data differs from the model of the sample representation of "reality", while the second one occurs when the assumptions do not correctly describe the "reality". "Reality" stands for a probability distribution of a random variable X and the random sample of variables $X_1,...,X_n$, independently and identically distributed (i.i.d.) like X, constitute the sample representation of "reality" (Barnett (1982)). Let X_i^D, $i \in \{1,...,n\}$, denote the random variable corresponding to the i-th observation of the sample data and F_X, $F_{X(D)}$ be the distribution functions of X and X^D respectively. Data contamination occurs either when $X_1^D,...,X_n^D$ are *not* identically and/or independently distributed or when $X_1^D,...,X_n^D$ are i.i.d. like X^D, but $F_{X(D)} \neq F_X$. The differences are:

i. It is *not* always true that the same kind of distortion holds for both data and model assumptions. For example, for a nonparametric model for F_X with no other assumption apart from continuity, a distortion like a mixture model could affect the data but cannot cause wrong model assumptions, since nothing other than continuity is assumed.
ii. The consequences of data contamination and model deviation are not necessarily equivalent under the same kind of distortion. For example, they could differ with respect to bias.

Finally, the main aims of robust statistics are:

- To describe the structure best fitting the bulk of the data
- To identify outliers or deviating substructures
- To identify leverage points (i.e. points that influence the estimating parameters a lot)
- To deal with unsuspected serial correlations and deviations from the assumed correlation structures

If the data are of high quality or they do not have any outliers, robust methods are not absolutely necessary, but they can still give a noticeable improvement over classical ones.

3.1.2 Huber's approaches

As we know, the *likelihood function* expresses the probability of the observed data as a function of the unknown parameters. The values that maximize this function are called *maximum likelihood estimators (MLE)* of these parameters. The disadvantage of MLE is that they are explicitly model-dependent and as such, they are criticized as being non-robust. Let $X_1,...,X_n$ be i.i.d. observations, which belong to some sample space $\mathcal{X}$ with a density f. For the location estimate T of an unknown parameter θ, instead of solving the ML equations $\sum f'/f(x_i-T)=0$ or minimizing a relationship of the form $-\sum \log f(x_i-T)$, Huber (1964) solved the equations $\sum \psi(x_i-T)=0$ or minimized a relationship of the form $\sum \rho(x_i-T)$. So, the two approaches optimise a different function. Any estimator defined by either of these two last equations is called an *M-estimator*. Huber (1967) called them the "*maximum likelihood estimates under non-standard conditions*". The next step was to withdraw from a strict parametric model of the form $G(x-\theta)$ for known G and to assume that a (known and fixed) fraction ε ($0 \leq \varepsilon < 1$) of the data might be consisted of gross errors with an arbitrary (unknown) distribution $H(x-\theta)$. Thus, he introduced the "gross-error model":

$$F(x-\theta)=(1-\varepsilon)G(x-\theta)+\varepsilon H(x-\theta).$$

As we have just seen, the *M*-estimators were just a slight generalization of MLE in the sense that M-estimators are based on different estimating equation that give more or less weight in a certain observation. The introduction of this flexible class of estimators gave a very useful tool and properties like consistency and asymptotic normality were derived. Huber's (1964) main aim was to optimise the worst that could happen over the neighbourhood of the model, as measured by the asymptotic variance of the estimator. He had to make some restrictions in order to be able to ignore or at least control the asymptotic bias, which in real life is unavoidable. Then, he used the formalism of a two-person zero-sum game: Nature chooses an F from the neighbourhood of the model, the statistician chooses an *M*-estimator via its ψ, but in reality an F is chosen from the neighbourhood of the model, and the gain for Nature the loss for the statistician is the asymptotic variance $V(\psi,F)$, which under mild regularity conditions turn out to be $\int \psi^2 dF \big/ \left(\psi' dF\right)^2$ (see also Definition 2(b) in Section 3.1.5). It was shown that under very general conditions there exists a saddle point of the game.

In the gross-error model case, it consists of what has been called "Huber's least favourable distribution", which is normal in the middle and exponential in the tails. The very known *Huber-estimator* is given in the form:

$$\psi_b(x) = \min\{b, \max(x, -b)\} = x \cdot \min\left(1, \frac{b}{|x|}\right),$$

for $0 < b < \infty$ (see Figure 3.1). It is the MLE for the distribution with density:

$$f(x) = \left[\exp\left(-\int_0^x \psi_b(t)\,dt\right)\right] \Big/ \int \left[\exp\left(-\int_0^u \psi_b(t)\,dt\right)\right] du,$$

which is least favourable in the minimax sense.

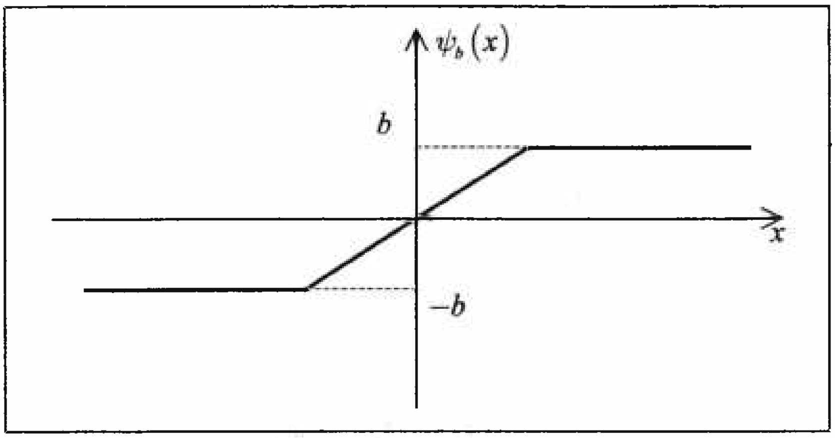


Figure 3.1.1: ψ-function defining the Huber-estimator with cut-off point b

As we can see, the ψ-function gives as weights the observations themselves inside a specific space of the real line, while outside that the weights become equal to the two edges of the space ($-b$ and b respectively). Hence, the observations are censored, since they are not allowed to take very large or very small values.

After the Minimax approach for robust estimation, Huber (1965) evolved another one using Robustified Likelihood Ratio Tests. He censored the classical likelihood ratio tests by putting a bound (possibly asymmetric) from above and below on the log likelihood ratio of each observation. So, a single observation could not carry any more the test statistic to $+\infty$, even if the likelihood ratio was unbounded. This method could also give robust confidence intervals and point estimates of location. In particular, Huber (1968) followed the next procedure:

> "*Given the length* $2a > 0$ *of the confidence interval, look for the estimate* T *which minimizes the maximum probability (over the neighborhoods of the parametric model distributions) of overshooting or undershooting the true* θ *by more than* a. *The estimate can be derived via a maximum test between* $\theta = -\alpha$ *and* $\theta = +\alpha$."

In this sense, Huber-estimators form the optimal robust estimators for the normal location model.

3.1.3 The Infinitesimal approach

The fact that many statistics depend only on the empirical cumulative distribution function of the data is useful for this approach, where three central robustness concepts are studied. Huber (1972) explained their use by linking them to the stability of aspects of a bridge, for example.

First of all, Qualitative Robustness is defined as the equicontinuity of the distributions of the statistic as n changes. It is very closely related to continuity of the statistic viewed as functional in the weak topology and it can be considered as a necessary but rather weak robustness condition. According to Huber (1972), a small perturbation to the bridge should have small effects. On the other hand, more informative is the Influence Function (IF), which measures the effects of infinitesimal perturbations (see also Definition 2(a) in Section 3.1.5). So, $IF(x;T,F)$ describes the effect of an additional observation in any point x on a statistic T, given a (large) sample with distribution F. In the end, the Breakdown Point measures the distance from the model distribution beyond which the statistic becomes totally unreliable and uninformative. It is guidance up to what distance from the model the local linearization provided by the IF can be used or it tells us how big the perturbation can be before the bridge breaks down.

Two important norms of IF are the *sup*-norm over x, known as "gross-error sensitivity" γ^*, which measures the maximum bias caused by infinitesimal contamination (and the stability of T under small changes of F) and the L_2-norm with respect to F, namely $\int IF^2 dF$. Both norms depend on F, so they can be seen as two new functionals measured by the "change-of-bias function" CBF and the "change-of-variance function" CVF. The *sup*-norms of the two latter functions are the "change-of-bias sensitivity" and the "change-of-variance sensitivity".

It should be mentioned, that robustness theory requires high breakdown point and low gross error sensitivity. The concept of "breakdown point" belongs to the maximum permitted percentage of the minority of the data, such that it has only limited influence. A value of 0.5 is the best possible for the breakdown point, which indeed indicates that any majority can overrule any minority. The problem with the gross-error sensitivity is that a low value contradicts the efficiency requirement of low asymptotic variance under the parametric model. Both of them have positive lower bounds and as a rule these bounds cannot be reached at the same time. Hampel (1968) came up to the famous result:

"*The most robust, the less efficient*"

Actually, there is an optimal class of compromise statistics or "admissible robust statistics", such that while the one bound increases, the other must be decreased.

3.1.4 Outliers

Most of the times many people usually regard an outlier solely as gross error. None the less, there are situations where the "outlier" is a proper observation and it is the most informative of all, so that the rest of the sample may be forgotten. Classical examples are the exams, where an outlier is the only correct value and the bulk of results are false. Also, in other cases the outlier indicates a different model for all data. When someone rejects an outlier, he should be very careful and know the aims that a rejection like this serves. Generally, if an outlier is too far "away", it is deemed too unlikely under the parametric model used and should be rejected, because the aim is the safety of the main statistical analysis. The danger of keeping the outlier could be disastrous and probably much bigger than the efficiency loss caused by rejection, if that was a proper observation. Another aim is the identification of interesting values for special treatment. The underlying model or some kind of effect may be of great interest. Even the gross error could be corrected or studied in more detail.

The two quantitative robustness tools that we have already seen can be used, so as to treat any possible outliers. On one hand, IF gives much information about the local behavior near the parametric model. It is characterized by its high jumps at the rejection points, which cause relatively large efficiency losses and whose height depends on the local density of the underlying distribution at the rejection points. On the other hand, the breakdown point of the combined rejection-estimation procedures tells us about their global reliability and how many distant outliers can be safely rejected. We must not forget that the ability to reject depends on the proportion of outliers and the best we can succeed is a value of 0.5 (50%). It also explains the "masking effect" of the outliers; an outlier masks a second one close by if the latter can be rejected alone, but not any more "in company".

Hampel *et al.* (1986) suggested two major ways of treating outliers; the first is to "move them in" close to the good data; the second is to reject them "smoothly", with continuously decreasing weight and influence. They arrived at some conclusions regarding the behavior of rejection rules for robust estimation. First of all, any way of treating outliers, which is not totally inappropriate, prevents the worst. Totally inappropriate are considered a non-robust computer program (e.g. least squares!) without any built-in checks and without a careful follow-up residual analysis, and furthermore, some objective rejection rules, like Studentized range. Most methods still lose unnecessarily at least 5-20% efficiency in some realistic situations. In general, one should not only detect and accommodate outliers, but also interpret and correct them. Identification of outliers and suspect outliers can be done much safer and better by looking at residuals from a robust fit, rather than a non-robust fit. Finally, rejection rules with subsequent estimation are nothing but special robust estimators.

3.1.5 Some more theory

Let $X_1,...,X_n$ be i.i.d. observations, which belong to some sample space $\mathcal{X}$. A *parametric model* consists of a family of probability distributions F_θ on $\mathcal{X}$, where the unknown parameter θ belongs to some parameter space Θ. In robust theory, the model $\{F_\theta;\theta\in\Theta\}$ is a mathematical abstraction, which is only an idealized approximation of reality. Our aim is to find or to construct statistical procedures, which still behave fairly well under deviations from this assumed model. So, we do not only consider the distribution of estimators under this specific model, but also under other probability distributions. We consider estimators, which are functionals

$\left[i.e.\ T_n(G_n)=T(G_n)\ \ for\ all\ n\ and\ G_n\right]$ or can asymptotically be replaced by functionals. This means that we can assume that there exists a functional $T : domain(T) \rightarrow \mathbb{R}$ such that $T_n(X_1,\dots,X_n)_{n\rightarrow\infty} \rightarrow T(G)$ in probability, when the observations are i.i.d. according to the true distribution G in $domain(T)$. The latter is the set of all distributions in $\mathcal{F}(\mathcal{X})$ for which T is defined and $G_n = \frac{1}{n}\sum_{i=1}^{n}\Delta_{x_i}$ is the *empirical distribution* of the sample, where Δ_x represents a degenerate distribution that gives probability 1 at point x and 0 elsewhere. We say that $T(G)$ is the *asymptotic value* of $\{T_n;\ n \geq 1\}$ at G.

Definition 1.

A functional T will be called *Fisher consistent*, if $T(F_\theta)=\theta\ for\ all\ \theta\ in\ \Theta$. So, *at* the model, the estimator $\{T_n; n \geq 1\}$ asymptotically measures the right quantity.

Definition 2.

(a) The *influence function (IF)* of T at F is:

$$IF(x;T,F)=\lim_{t\rightarrow 0}\frac{T\left((1-t)F+t\Delta_x\right)-T(F)}{t},$$

in those $x \in \mathcal{X}$ that the limit exists.

(b) The *asymptotic variance* is closely related to IF. Actually, it is given by:

$$V(T,F)=\int IF(x;T,F)^2\, dF(x).$$

(c) On the other hand, the *asymptotic relative efficiency* of a pair of estimators $\{T_n; n \geq 1\}$ and $\{S_n; n \geq 1\}$ is given by:

$$ARE_{T,S}=\frac{V(S,F)}{V(T,F)}.$$

Except for the expected square of IF, there are at least other three important summary values. The first one is the supremum of the absolute value, which defines the *gross-error sensitivity* of T at F as $\gamma^* = \gamma^*(T,F)=\sup_x\left|IF(x;T,F)\right|$. This measures the worst (approximate) influence, which a small amount of contamination of fixed size can have on the value of the estimator and for that, it can be thought as an upper bound on the (standardized) asymptotic bias of the estimator. The second one deals with small fluctuations in the observations. The worst (approximate and standardized) effect of adding an observation at y and removing another one at x is examined via the *local-shift sensitivity* $\lambda^{*} = \sup_{x \neq y} \frac{\left|IF(y;T,F) - IF(x;T,F)\right|}{y-x}$. On the contrary, the third one refers to the complete rejection of the extreme outliers. Extreme outliers will be considered (and then are entirely rejected) all observations further away than the *rejection point* $\rho^{*} = \inf_{x}\left\{r > 0;\ IF(x;T,F) = 0\ \ when\left|x\right| > r\right\}$.

Let us now assume a countable set $\mathcal{X} = \{0,1,\dots,K\}$, $K \leq \infty$, as sample space. Also, $m_{\beta}(x)$ will be called the *model vector* and is a family of probability densities on $\mathcal{X}$, with $m_{\beta}(x) > 0$ (*for all* $x \in \mathcal{X}$). The i.i.d. observations $X_1,\dots,X_n$ are made from $m_{\beta}(x)$. The *data vector* $d(x)$ expresses the proportion of the n observations that have value x and the function $t(x)$ denotes some nominal *true density*.

Definition 3.

The *Pearson residual function* is defined as:

$$\delta(x) = \frac{\left[d(x) - m_{\beta}(x)\right]}{m_{\beta}(x)}.$$

Remark 1.

These residuals are not standardized to have identical variances. Also, it is important to notice, that they have range $[-1,\infty)$.

Definitions 4.

(a) The (squared) *Hellinger distance (HD)* between the data vector and the model vector is defined by $\sum\left[\sqrt{d(x)} - \sqrt{m_{\beta}(x)}\right]^{2}$

(b) and the *minimum Hellinger distance estimator (MHDE)* is that value of β, which minimizes the above distance.

Remark 2.

The pioneering work by Beran (1977) showed that the MHDE could achieve first-order efficiency and robustness properties. Also, Titterington *et al.* (1985) proved that MHDE is an important alternative to the MLE among various minimum-distance estimates, while Simpson (1987) found that in the Poisson model the Hellinger distance has an asymptotic breakdown point of 0.5. For a more recent comparison of the MHDE with the MLE, as well as the balance between robustness and efficiency, the reader may refer to Lindsay (1994) and Karlis and Xekalaki (1998, 2001).

Remark 3.

Let $\xi \in \mathcal{X}$, $\chi_\xi(x)$ be a degenerate distribution at ξ and:

$$t_\varepsilon(x)=(1-\varepsilon)t(x)+\varepsilon\chi_\xi(x)$$

an ε-contaminated version of density $t(x)$. A different expression of the IF is:

$$T'(\xi)=\left.\frac{\partial T\left((1-\varepsilon)t+\varepsilon\chi_\xi\right)-T(F)}{\partial\varepsilon}\right|_{\varepsilon=0}=\frac{\partial T(t_\varepsilon)}{\partial\varepsilon},$$

where t_ε is an ε-contaminated version of density $t(x)$ defined right above. When the functional is the MLE $T_{ML}(t)$ or the MHDE $T_{HD}(t)$, then $T'(\xi)=i(\beta)^{-1}u(\xi;\beta)$, where $i(\beta)$ is the Fisher information and $u(\xi;\beta)=\nabla\log(m_\beta(\xi))$ is the *score function* (the symbol ∇ denotes the differentiation with respect to β).

Remark 4.

When the model is correctly specified, any estimator with the same IF as the MLE (given by the latter expression) has the same efficiency and so is optimal.

Remark 5.

All the first-order efficient estimators could be considered non-robust, because they have the same sensitivity to contamination as the MLE. However, Taylor series approximations like $\Delta T(\varepsilon):=T(t_\varepsilon)-T(t)\cong\varepsilon T'(\xi)$ can be very misleading.

3.2 Weighted maximum likelihood method

The idea of robustness depends largely on stability of the parameter estimates under slight departures from the model. We would like to "correct" surprising observations; "surprising" in the sense that they occur in locations ξ with small probabilities $m_\beta(\xi)$. Thus, it seems very natural to downweight data points with large Pearson residuals or generally of dubious authenticity.

3.2.1 Lindsay's (1994) approach

Lindsay (1994) worked with another function, the "residual adjustment function" in order to find the key structural element that links ML and minimum HD and to measure the robustness properties of MHDE, since the results from the IF were poor. The central point of his research was estimating equations for β of the form:

$$\sum A(\delta(x))\nabla m_\beta(x)=0,$$

where $A(\delta)$ satisfies the next assumption.

Assumption 1.

The *residual adjustment function (RAF)* is assumed to be an increasing twice-differentiable function $A(\delta)$ on $[-1,\infty)$, with $A(0)=0$ and $A'(0)=1$.

Examples

- Using the linear RAF $A_{LD}(\delta)=\delta$, we take the important case of the MLE:

$$0=\sum\delta(x)\nabla m_\beta(x)=\sum(d(x)-m_\beta(x))u(x;b)=\sum d(x)u(x;b).$$

- For the minimum HD, we also take the form:

$$\sum A(\delta(x))\nabla m_\beta(x)=0,$$

if we use as RAF the $A_{HD}(\delta)=2\left[\sqrt{\delta+1}-1\right]$.

A very important class of RAFs has the form:

$$A_\lambda(\delta)=\frac{(1+\delta)^{\lambda+1}-1}{\lambda+1}.$$

For several values of λ, we obtain known corresponding results such as:

→ ML for $\lambda=0$

→ HD for $\lambda = -\frac{1}{2}$

→ Minimum Pearson's chi-squared for $\lambda = 1$

→ Minimum Neyman's chi-squared for $\lambda = -2$

→ Minimum Kullback-Leibler divergence for $\lambda \to -1$

The most important property of RAF is that solving the estimating equations $\sum A(\delta(x))\nabla m_\beta(x) = 0$ corresponds to the minimization of a measure of "distance" between the data d and the model m_β. Also, the degree of robustness relative to ML depends on how much $A(\delta)$ deviates from linearity. Note that in fact $A(\delta(x))$ is a weight for the observation x. So, observations with large Pearson residuals are given smaller weights.

Lindsay (1994) developed a large subclass of density based minimum distance estimation, called "minimum disparity estimation", of which minimum HD estimation is a part. This estimation is an efficient and robust estimation method in parametric models and succumbs the disadvantages of Huber's (1981) minimax approach, which despite its fine theoretical properties, it is very difficult to apply in problems other than location. For discrete models Lindsay's (1994) method extends easily the ideas from MLE, since it compares the observed probability function to the expected under the assumed model probability function, via a suitable chosen disparity function. For continuous models this is not straightforward, because the observed measure is a discrete one, while the assumed density is continuous. There are several solutions proposed in the literature mainly based on "kernel density estimators", but our work will not extend to this area.

Definition 5.

(a) Suppose that $G(\cdot)$ is a real-valued thrice-differentiable function on $[-1,\infty)$ with $G(0) = 0$ and $\delta(x)$ is the Pearson residual. For any pair of densities $m_\beta(x)$ and $d(x)$, the *disparity measure* determined by G is defined as:

$$\rho(\delta, m_\beta) = \sum m_\beta(x) G(\delta(x)).$$

(b) The *minimum disparity estimator (MDE)* is that value of β -call it $T(d)$- which minimizes ρ.

Cressie and Read (1984 and 1988) introduced an important class of such measures, known as Cressie-Read family of *power divergence measures*:

$$PWD\left(d,m_\beta\right)=\sum d\left(x\right)\frac{\left\{\left[d\left(x\right)/m_\beta\left(x\right)\right]^{\lambda}-1\right\}}{\lambda\left(\lambda+1\right)}=\sum m_\beta\left(x\right)\frac{\left\{\left[1+\delta\left(x\right)\right]^{\lambda+1}-1\right\}}{\lambda\left(\lambda+1\right)}.$$

For several values of λ, we take known measures such as:

→ Likelihood disparity for $\lambda=0$:

$$LD\left(d,m_\beta\right)=\sum d\left(x\right)\left[\log\left(d\left(x\right)\right)-\log\left(m_\beta\left(x\right)\right)\right]$$

→ Twice-squared HD for $\lambda=-\frac{1}{2}$:

$$HD\left(d,m_\beta\right)=2\sum\left[\sqrt{d\left(x\right)}-\sqrt{m_\beta\left(x\right)}\right]^2$$

→ Pearson's chi-squared divided by 2, for $\lambda=1$:

$$PCS\left(d,m_\beta\right)=\sum\frac{\left[d\left(x\right)-m_\beta\left(x\right)\right]^2}{2m_\beta\left(x\right)}$$

→ Neyman's chi-squared divided by 2, for $\lambda=-2$:

$$NCS\left(d,m_\beta\right)=\sum\frac{\left[d\left(x\right)-m_\beta\left(x\right)\right]^2}{2d\left(x\right)}$$

→ Kullback-Leibler divergence for $\lambda=-1$:

$$KL\left(d,m_\beta\right)=\sum m_\beta\left(x\right)\left[\log\left(m_\beta\left(x\right)\right)-\log\left(d\left(x\right)\right)\right].$$

Lindsay (1994) used all the former information and proved that on one hand the RAF determines various second-order measures of efficiency and robustness through a scalar measure called the "estimation curvature" and on the other hand the breakdown properties of the estimators through its tail behavior (a breakdown point of 50% is given). Furthermore, a second-order approximation of the RAF, $A\left(\delta\right)\cong\delta+A_2\delta^2/2$, shows how the curvature parameter A_2 becomes a measure of the trade-off between the efficiency and robustness in a second-order sense. In the following, we shall present only some basic of his findings. For further explanation and proof, the reader can refer to the paper of Lindsay (1994).

✍ The first "suspicion" is that the deficiency of an MDE is a simple function of the estimating curvature A_2 and a nonnegative quantity D depending on the model but not on $A(\delta)$.

Proposition 1.

Suppose that the sample space is finite $(K < \infty)$. The *second-order efficiency* of a MDE with RAF $A(\delta)$ is:

$$E_2(MDE) = E_2(MLE) + A_2^2 D.$$

✍ Suppose that the true density is a contaminated model. Generally, the objective is to create a test procedure, more robust in the sense of preserving size and hence confidence intervals coverage, but equivalent to the likelihood ratio test, when the model is right. If we want to test the hypothesis

$$H_0 : \beta = \beta_0,$$

we can express the likelihood ratio test in terms of the likelihood disparity function:

$$LRT = 2n\left[LD\left(d, m_{\beta_0}\right) - LD\left(d, m_T\right)\right], \quad with\ T = T_{ML}(d).$$

Simpson (1989) used the *disparity difference test statistic*:

$$DDT = 2n\left[\rho\left(d, m_{\beta_0}\right) - \rho\left(d, m_T\right)\right], \quad with\ T = T(d),$$

in the case that ρ equals to squared HD. We consider its behaviour as a test statistic for $H : T(t) = \beta_0$, where t may or may not be in the model. Under certain conditions, it holds the next:

Theorem 1.

(i) If $t(x)$ is in the model, then, under the null hypothesis:

$$DDT \rightarrow X^2_{\dim(\beta)}.$$

(ii) Under the null hypothesis, if $t(x)$ is the true density and $\dim(\beta) = 1$, then:

$$DDT \rightarrow c(t) X_1^2,$$

where: $c(t) = Var_t\left[T'(t, X)\right] \cdot \nabla\rho\left(t, m_\beta\right)\Big|_{\beta=\beta_0}$.

Lindsay (1994) noticed that the likelihood disparity was not bounded over all d and m_β. Hence, there was obvious the need for some new disparities, so as moments of all orders to exist. For any fixed number $\alpha \in [0,1]$ and $\bar{\alpha} = 1 - \alpha$, such disparities are:

- the *blended weight chi-squared disparity*:

$$BWCS(\alpha) = \sum \frac{[d(x) - m_\beta(x)]^2}{2[\alpha\, d(x) + \bar{\alpha}\, m_\beta(x)]},$$

where the corresponding RAF is:

$$A_\alpha(\delta) = \frac{\delta}{1+\alpha\delta} + \frac{\bar{\alpha}}{2}\left[\frac{\delta}{1+\alpha\delta}\right]^2$$

- and the *blended weight Hellinger chi-squared disparity*:

$$BWHD_\alpha = \sum \frac{[d(x) - m_\beta(x)]^2}{2\left[\alpha\sqrt{d(x)} + \bar{\alpha}\sqrt{m_\beta(x)}\right]^2},$$

where the corresponding RAF is:

$$A_\alpha(\delta) = \delta[w(\delta)]^{-2} + \frac{\bar{\alpha}}{2}\delta^2[w(\delta)]^{-3},$$

with $w(\delta) = \alpha\sqrt{\delta+1} + \bar{\alpha}$.

Finally, another desirable fixture is the robustness against "inliers". Although minimum Hellinger distance has the right tail behavior to deal with large outliers, it has some deflects with respect to inliers. For this, we should try to find adjustment functions that down-weight both positive and negative residuals relative to ML, in the sense that $|A(\delta)| \le |\delta|$. If we start with the convex function $G(\delta) = e^{-\delta} - 1$, then we obtain the *negative exponential (NE) disparity measure* with RAF:

$$A_{NE}(\delta) = 2 - (2+\delta)e^{-\delta}.$$

This disparity measure has the advantage, that it generates a bounded RAF. Indeed, the minimum NE disparity estimator generates a second-order efficient estimator that shrinks both positive and negative residuals, because:

$$\left|\begin{array}{l} A'(\delta) = (1+\delta)e^{-\delta} \\ A''(\delta) = -\delta\, e^{-\delta} \\ A'''(\delta) = (\delta - 1)e^{-\delta} \end{array}\right|.$$

The key element in the WMLE approach is to define the weights in a "clever" way, so as to down-weight appropriately the ML score function (see also Section 3.3.1) and thus to obtain robust estimates with high efficiency. By suitably defining the weights, we may obtain weighted likelihood estimates that reflect disparities evaluated at the observed data points and not at the whole real line. By such an approach we overcome problems with any numerical integration needed (in continuous cases), which can lead to stability problems.

Remark 6.

It is important to note that the two approaches (the approach of Huber and the one defined above) share some common elements. The most important is that the RAF operates in the Pearson's residuals as Huber's ψ-function operates in simple residuals. Thus both methods down-weight unexpected observation, the former is based on simple residuals while the later in Pearson's residuals (see Agostinelli, 2002 for a discussion on this issue).

The Algorithm

Basu and Lindsay (1993b) proposed to rewrite $\sum A(\delta(x))\nabla m_\beta(x)=0$ in a weighted form of the likelihood equation, with weights defined by:

$$\begin{aligned}\sum A(\delta(x))\nabla m_\beta(x)&=\sum\left[A(\delta(x))-A(-1)\right]\nabla m_\beta(x)=\\&=\sum w^*(x)(1+\delta(x))\nabla m_\beta(x)=\\&=\sum w^*(x)d(x)u(x;\beta)\end{aligned}$$

The term $A(-1)$ in w^* forces the weights to be nonnegative (since A is increasing) and it can be replaced by any other constant without changing the above equalities. The algorithm is as follows:

"Given current estimate b, create weights $w^*(x)$

and solve for the above equations equal to zero

with these weights fixed."

For example, if m_β is the Poisson model (mean parameter β), then the algorithm gives a re-weighted mean as the next value of the parameter:

$$b_{new} = \frac{\sum w^{*}(x)d(x)x}{\sum w^{*}(x)d(x)}.$$

When the algorithm has converged, the final set of weights $w^{*}(x)$ reflects the relative influence that the observed cells had in the final solutions. Finally, we just mention that in the continuous case, in the place of the two summations, there are respectively two integrals. The main issue upon the updating equation is that one needs to evaluate the two integrals numerically. Usually this can be made using standard algorithms which are based on the feature that evaluate the function to be integrated at suitably chosen points.

3.2.2 Related material

Basu and Sarkar (1994) made an extensive empirical study to compare the estimators and the disparity tests generated by the NE disparity at the normal model to those generated by the BWHD family. Generally the efficiency of an estimator T relative to the MLE is estimated by the ratio of the mean square errors (MSEs):

$$efficiency = \frac{MSE(MLE)}{MSE(T)}.$$

It is shown that the curvature parameter of the RAF is not always an adequate global measure of the trade-off between robustness and efficiency of the MDE. The estimator obtained by minimizing the NE disparity is an attractive robust estimator with good efficiency properties.

Harris and Basu (1994) also studied the HD and expressed it in the form of a penalized log-likelihood. They considered a parametric family of distributions with countable support and their aim was to make inference about the parameter based on a random sample of size n from m_β. For simplicity, they examined cases where θ was scalar, but the methods could be generalized to multiparameter θ. Let f_β be a data driven modification of the model, such that the minimization of the Kullback-Leibler (KL) divergence between d and f_β generates the MHDE of the parameter based on m_β. They showed that minimizing the HD corresponds to minimizing:

$$\sum_{N_x} d(x)\left[\log\left(d(x)/f_\beta(x)\right)\right] + \sum_{N'_x} m_\beta(x),$$

where $N_x=\{x:d(x)\neq 0\}$ and N_x' is the complementary of N_x. The term $\sum_{d(x)=0} m_\beta(x)$ can be thought of as a penalty applied to the KL divergence of a modified function $f_\beta(x)$ and minimizing the HD can be thought as equivalent to maximizing a penalized log-likelihood. According to Good and Gaskins (1971), the usual method of penalized KL divergence minimizes:

$$\sum_x d(x)\left[\log\left(d(x)/f_\beta(x)\right)\right]+hJ(\beta),$$

where $J(\beta)$ is a penalty function and h is the weight put on the penalty. The minimum HD corresponds to using the penalty $\sum_{d(x)=0} m_\beta(x)$ with the penalty weight h being equal to 1. Harris and Basu (1994) used another class of estimators, the *minimum penalized Hellinger distance estimators (MPHDE's)*, which are constructed by minimizing:

$$\sum_x d(x)\left[\log\left(d(x)/f_\beta(x)\right)\right]+h\sum_{d(x)=0} m_\beta(x).$$

They found that most of the MHDEs' robustness was derived from the modification in the density rather than the use of the penalty. This modification produces a function that essentially "ignores" the outlying observations. If the weight on the penalty is changed, this will not alter the robustness properties of the parameter estimate.

Field and Smith (1994) also worked on the uncertainty of the data source. They assumed a parametric model $f(x;\theta)$ and modified the usual likelihood equations in order to achieve robust estimates with high breakdown properties. The main tools were the score function $f'/f(x,\theta)$ and two weight functions $w(x,\theta)$. The first one is similar to a Huber-style estimate, since it truncates the score function but the truncation is carried out on the probability scale and not in a Euclidean scale. The idea is to consider the supremum of each score function over the central $(1-2p)\%$ of the distribution as determined by the current value of θ. The j-th component is:

$$w_j(x,\theta)=\min\left[\sup_{y\in A(\theta,\rho)}\left\{\left|\frac{\left|\frac{f'}{f}(y,\theta)_j\right|}{\left|\frac{f'}{f}(x,\theta)_j\right|}\right|\right\},1\right],$$

where $A(\theta,\rho)=\{x \mid p \leq F(x,\theta) \leq 1-p\}$. On the contrary, the second one uses the same weight for each component of the score function:

$$w(x,\theta)=\begin{cases}\dfrac{F(x,\theta)}{p} & \text{if } F(x,\theta)<p \\ 1 & \text{if } p \leq F(x,\theta) \leq 1-p . \\ \dfrac{1-F(x,\theta)}{p} & \text{if } F(x,\theta)>1-p\end{cases}$$

Both weight functions have the important property being invariant under monotone transformations of the data. The latter additionally down-weights smoothly any points, which do not lie within the central $(1-2p)\%$. They used an iterative procedure, starting with the ML score function. By truncating its norm and adjusting the score function to have Fisher consistency and the pre-assigned bound on the supremum of the IF, the optimal score function is obtained. The starting point should be a high breakdown estimate. After some examples, they calculated the efficiencies and a simulation study compared their estimator's performance to some others.

Markatou (1996) pointed out the dilemma between weighted likelihoods and usual *M*-estimation for random variables, which follow a continuous distribution. She studied a contaminated Normal model $(1-\varepsilon)N(0,1)+\varepsilon N(\mu,\sigma^2)$ for various values of ε, μ and σ^2 and presented a Monte Carlo comparison between the methods of Basu *et al.* (1995) and the classical Huber's (1981) robust methods. Robustness is defined by the existence of a root at or near the parameters of the component with the larger mass and by the existence of a root at or near the parameters of either component when $\varepsilon=0.50$. She used the *negative exponential (NE) RAF*, $A(\delta)=2-(2+\delta)\exp(-\delta)$ and the RAF based on HD. The algorithm requires appropriate starting values and in her simulations she only uses the pairs (μ,σ^2) of starting values $\left(med(X_i), 1.48 \cdot med\left|X_j-med(X_i)\right|\right)$ and the corresponding ML estimates of location and scale. Convergence is achieved, when the difference in the estimators between two consecutive steps is less than or equal to 10^{-6}. One result is that the biases of the location and scale estimates do not differ much, when different starting values are used. Generally speaking, when ε increases, the power of p in the *weighted score estimation equations*:

$$\sum_{i=1}^{n}\left[w\left(x_i;M_\beta,\hat{F}\right)\right]^{p}u\left(x_i;\beta\right)=0$$

should increase so as to guarantee a smaller weight for the aberrant data. For certain combinations of p and ε, if we use the appropriate RAF, estimators with very small bias are generated. Basu *et al.* (1995) recommend the use of a grid of starting values, mainly with high percentages of contamination.

Basu *et al.* (1997) worked on the same basis and established that the *Minimum NE disparity estimator (MNEDE)* is asymptotically as efficient as the MLE at the model and robust under data contamination. The MNEDE is obtained by minimizing the NE disparity $D(\hat{g}_n,\theta)$ between $\hat{g}_n$ and f_θ over Θ, where $f_\theta \in \mathcal{F}_\theta$, $g \in \mathcal{G}$,

$$D(g,\theta)\equiv\int\left\{\exp\left[-\delta(g,\theta,x)\right]-1\right\}f_\theta(x)dx \quad \text{and} \quad \delta(g,\theta,x)\equiv\left(\frac{g(x)}{f_\theta(x)}-1\right).$$

The MNEDE, like the MHDE, is a very attractive robust estimator since it attains its robustness properties without sacrificing first-order efficiency at the model. A very nice feature of the MNEDE is the robustness it provides against inliers, a property that the MHDE does not share. On the whole, the MNEDE appears to be a promising estimator and a major competitor of the MHDE within the class of robust first-order efficient estimators.

Finally, we should make two last important remarks. The first of the two is that Böhning and Hoffmann (1982), in their attempt to find the MLE for a certain class of discrete sampling models, pointed out that one of the main restrictive assumptions someone has to make is the concavity of the log-likelihood function. Secondly, Markatou (1999a) took a closer look at the performance of weighted likelihood in the context of mixture models. It is shown that the weighted likelihood methodology produces robust and first-order efficient estimators for the model parameters. When the number of components in the true model is higher than the number of components specified in the hypothesized model, the weighted likelihood equations have multiple roots.

3.3 Generalized Linear Models (GLMs)

Up to this point, we have discussed procedures for simple data sets, having a single variable. Generalized linear models (GLMs) are an extension of classical linear models. A vector of observations y having n components is assumed to be a realization of a random variable Y whose components are independently distributed with means μ. The systematic part of the model is a specification for the vector μ in terms of a small number of unknown parameters $\beta_1,\dots,\beta_p$. In the case of ordinary linear models, this specification takes the form:

$$E(Y_i) = \mu_i = \sum_{j=1}^{p} x_{ij}\beta_j,\ i = 1,\dots,n,$$

where the βs are parameters whose values are usually unknown and have to be estimated from the data and x_{ij} is the value of the j-th covariate for observation i. We may rewrite it in matrix notation:

$$\mu_{(n\times 1)} = X_{(n\times p)}\beta_{(p\times 1)},$$

where X is the model matrix and β is the parameter vector. We also make the assumption, that we know the covariates that influence the mean and can measure them effectively without error. For the random error, we assume independence and constant variance of errors.

The classical linear model can be separated in three parts. The *random component* consists of the components of Y being independently Normal with $E(Y) = \mu$ and constant variance σ^2. The *systematic component* consists of the covariates $x_1,\dots,x_p$, which produce a *linear predictor*

$$\eta = \sum_{j=1}^{p} x_j\beta_j .$$

Finally, the *link* between the random and the systematic part is: $\mu = \eta$. If we write $\eta_i = g(\mu_i)$, then $g(\cdot)$ will be called the *link function*. GLMs allow two extensions;

(i) the distribution may come from an exponential family (and not explicitly from the Normal case).

(ii) $g(\cdot)$ may become any monotonic differentiable function.

Definitions 6.

A probability distribution is said to be a member of the *exponential family*, if its probability density function (or probability function, if discrete) can be written in the form:

$$f_Y(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right).$$

The parameter θ is called the *natural* or *canonical* parameter and the parameter ϕ is usually assumed known. When it is assumed unknown, it is often called the *nuisance* parameter. The basic properties are:

$$E[Y] = b'(\theta)$$

$$Var[Y] = a(\phi)b''(\theta).$$

Robust inference about GLMs is very limited. Sections 3.3.1. and 3.3.2. present some important papers and research on GLMs or robustness in GLMs.

3.3.1. Iterative Reweighting Least Squares (IRLS)

In order to find the MLE $\hat{\beta}$ of β, we must maximize the log-likelihood function. If we use the exponential family, the joint density for $Y = (Y_1, Y_2, \ldots, Y_n)^T$ is:

$$f_Y(y;\theta,\phi) = \prod_{i=1}^{n} f_{Y_i}(y_i;\theta_i,\phi_i) = \exp\left\{\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + \sum_{i=1}^{n} c(y_i,\phi_i)\right\},$$

where $\theta = (\theta_1, \ldots, \theta_n)^T$ stands for canonical parameters and $\phi = (\phi_1, \ldots, \phi_n)^T$ for nuisance parameters. Hence, the log-likelihood function can be written in the form:

$$\log(f_Y(y;\beta,\phi)) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + \sum_{i=1}^{n} c(y_i,\phi_i)$$

and depends on β through: $\begin{cases} \mu_i = b'(\theta_i) \\ g(\mu_i) = \eta_i \\ \eta_i = x_i^T\beta = \sum_{i=1}^{n} x_{ij}\beta_j \end{cases}$ $\quad i = 1, \ldots, n$.

A relatively detailed proof for the MLE of $\hat{\beta}$ starts by considering the *scores*:

$$u_k(\beta)=\frac{\partial}{\partial\beta_k}\log\left(f_Y(y;\beta,\phi)\right),\quad k=1,2,\ldots,p$$

and then $u_k(\beta)=0,\quad k=1,2,\ldots,p\ \Rightarrow\ \mathbf{u}\left(\hat{\beta}\right)=\mathbf{0}$. From the exponential family:

$$\begin{aligned}
u_k(\beta)=\frac{\partial}{\partial\beta_k}\log\left(f_Y(y;\beta,\phi)\right)&=\frac{\partial}{\partial\beta_k}\sum_{i=1}^{n}\frac{y_i\theta_i-b(\theta_i)}{a(\phi_i)}+\frac{\partial}{\partial\beta_k}\sum_{i=1}^{n}c(y_i,\phi_i)=\\
=\sum_{i=1}^{n}\frac{\partial}{\partial\beta_k}\left[\frac{y_i\theta_i-b(\theta_i)}{a(\phi_i)}\right]+0&=\sum_{i=1}^{n}\frac{\partial}{\partial\theta_i}\left[\frac{y_i\theta_i-b(\theta_i)}{a(\phi_i)}\right]\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_k}=\\
&=\sum_{i=1}^{n}\frac{y_i-b'(\theta_i)}{a(\phi_i)}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_k}=\\
&=\sum_{i=1}^{n}\frac{y_i-b'(\theta_i)}{a(\phi_i)}\left[\frac{\partial\mu_i}{\partial\theta_i}\right]^{-1}\left[\frac{\partial\eta_i}{\partial\mu_i}\right]^{-1}\frac{\partial}{\partial\beta_k}\sum_{j=1}^{p}x_{ij}\beta_j=\\
&=\sum_{i=1}^{n}\frac{y_i-b'(\theta_i)}{a(\phi_i)}\frac{1}{b''(\theta_i)}\frac{1}{g'(\mu_i)}x_{ik}=\\
&=\sum_{i=1}^{n}\frac{y_i-b'(\theta_i)}{a(\phi_i)}\frac{x_{ik}}{b''(\theta_i)g'(\mu_i)}=\\
&=\sum_{i=1}^{n}\frac{y_i-\mu_i}{Var(Y_i)}\frac{x_{ik}}{g'(\mu_i)},\qquad k=1,\ldots,p,
\end{aligned}$$

which depend on β through $\mu_i=E(Y_i)$ and $Var(y_i),\ i=1,\ldots,n$. In theory, we solve the p simultaneous equations $u_k\left(\hat{\beta}\right)=0,\quad k=1,2,\ldots,p$ to evaluate $\hat{\beta}$. In practice, these equations are usually non-linear and have no analytic solution. Therefore, we rely on numerical methods to solve them.

First, we note that the Hessian $\boldsymbol{H}$ and the Fisher information I matrices can be derived directly from the last expression. Recall that these two matrices are closely connected, because *I is minus the expected value of the Hessian matrix*. So, we take:

$$\begin{aligned}
\left[\boldsymbol{H}(\beta)\right]_{jk} &= \frac{\partial^2}{\partial\beta_j\partial\beta_k}\log\left(f_Y(y;\beta,\phi)\right)= \\
&= \frac{\partial}{\partial\beta_j}u_k(\beta)= \\
&= \frac{\partial}{\partial\beta_j}\sum_{i=1}^{n}\frac{y_i-\mu_i}{Var(Y_i)}\frac{x_{ik}}{g'(\mu_i)}= \\
&= \sum_{i=1}^{n}\frac{-\frac{\partial\mu_i}{\partial\beta_j}}{Var(Y_i)}\frac{x_{ik}}{g'(\mu_i)}+\sum_{i=1}^{n}(y_i-\mu_i)\frac{\partial}{\partial\beta_j}\left[\frac{x_{ik}}{Var(Y_i)g'(\mu_i)}\right]
\end{aligned}$$

and:

$$\begin{aligned}
\left[I(\beta)\right]_{jk} &= E\left[-\left[\boldsymbol{H}(\beta)\right]_{jk}\right]= \\
&= \sum_{i=1}^{n}\frac{\frac{\partial\mu_i}{\partial\beta_j}}{Var(Y_i)}\frac{x_{ik}}{g'(\mu_i)}-\sum_{i=1}^{n}\left(E[Y_i]-\mu_i\right)\frac{\partial}{\partial\beta_j}\left[\frac{x_{ik}}{Var(Y_i)g'(\mu_i)}\right]= \\
&= \sum_{i=1}^{n}\frac{\frac{\partial\mu_i}{\partial\beta_j}}{Var(Y_i)}\frac{x_{ik}}{g'(\mu_i)}-\sum_{i=1}^{n}(\mu_i-\mu_i)\frac{\partial}{\partial\beta_j}\left[\frac{x_{ik}}{Var(Y_i)g'(\mu_i)}\right]= \\
&= \sum_{i=1}^{n}\frac{\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j}}{Var(Y_i)}\frac{x_{ik}}{g'(\mu_i)}-0= \\
&= \sum_{i=1}^{n}\frac{\frac{1}{g'(\mu_i)}x_{ij}}{Var(Y_i)}\frac{x_{ik}}{g'(\mu_i)}= \\
&= \sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{Var(Y_i)g'(\mu_i)^2}
\end{aligned}$$

or equivalently in the very important relationship:

$$I(\beta)=X^TWX,$$

where:

$$X=\begin{bmatrix} x_1^T \\ \cdot \\ \cdot \\ \cdot \\ x_n^T \end{bmatrix}=\begin{bmatrix} x_{11} & \dots & x_{1p} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ x_{n1} & \dots & x_{np} \end{bmatrix},$$

$$W = diag(w) = \begin{pmatrix} w_1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & w_2 & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & \cdot & & \cdot \\ \cdot & & & & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & 0 & w_n \end{pmatrix}$$

and:

$$w_i = \frac{1}{Var(Y_i)g'(\mu_i)^2}, \quad i = 1,\ldots,n.$$

The Fisher information matrix depends on β through μ and $Var(Y_i)$, $i = 1,\ldots,n$.

Now, the scores can be written as:

$$u_k(\beta) = \sum_{i=1}^{n}(y_i - \mu_i)x_{ik}w_i g'(\mu_i) = \sum_{i=1}^{n} x_{ik}w_i z_i, \quad k = 1,2,\ldots,p,$$

where $z_i = (y_i - \mu_i)g'(\mu_i)$, $i = 1,\ldots,n$. Therefore: $u(\beta) = X^T Wz$. One method to solve the p simultaneous equations $u_k(\hat{\beta}) = 0$, $k = 1,2,\ldots,p$ that give $\hat{\beta}$ could be the (multivariate) Newton-Raphson method. If β^i is the current estimate of $\hat{\beta}$, then the next estimate is:

$$\beta^{i+1} = \beta^i - H(\beta^i)^{-1}u(\beta^i)$$
$$\Leftrightarrow \quad .$$
$$\beta^{i+1} = \beta^i + I(\beta^i)^{-1}u(\beta^i)$$

This iterative algorithm is called *Fisher scoring*. By using former relationships and substituting in the latter one, we get:

$$\begin{aligned} \beta^{i+1} &= \beta^i + \left[X^T W^i X\right]^{-1} X^T W^i z^i = \\ &= \left[X^T W^i X\right]^{-1}\left[X^T W^i X\beta^i + X^T W^i z^i\right] = \\ &= \left[X^T W^i X\right]^{-1} X^T W^i\left[X\beta^i + z^i\right] = \\ &= \left[X^T W^i X\right]^{-1} X^T W^i\left[\eta^i + z^i\right] \end{aligned}$$

where η^i, W^i and z^i are all functions of β^i. This is a weighted least squares equation, that is β^{i+1} minimizes the weighted sum of squares:

$$(\eta + z - X\beta)^T W(\eta + z - X\beta) = \sum_{i=1}^{n} w_i\left(\eta_i + z_i - x_i^T\beta\right)^2$$

as a function of β, where $w_1,\ldots,w_n$ are the weights and $\eta + z$ is called the *adjusted dependent variable*. The Fisher scoring algorithm (also known as *IRLS*, because it involves iteratively minimising a weighted sum of squares) proceeds as follows:

1. *Choose an initial estimate* β^i *for* β *at i=0.*
2. *Evaluate* η^i, W^i *and* z^i *at* β^i.
3. *Calculate* $\beta^{i+1} = \left[X^T W^i X\right]^{-1} X^T W^i \left[\eta^i + z^i\right]$.
4. *If* $\left\|\beta^{i+1} - \beta^i\right\|$ *>some pre-specified (small) tolerance, then set* $i \to i+1$ *and go to 2.*
5. *Use* β^{i+1} *as the solution for* $\hat{\beta}$.

Remark 7.

The canonical link function is $g(\mu) = b'^{-1}(\mu)$. Because with this link $\eta_i = g(\mu_i) = \theta_i$, it holds:

$$\frac{1}{g'(\mu_i)} = \frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i), \quad i = 1,\ldots,n.$$

So, $Var(Y_i)g'(\mu_i) = \alpha(\phi_i)$ which does not depend on β and that is why:

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial}{\partial \beta_j}\left[\frac{x_{ik}}{Var(Y_i)g'(\mu_i)}\right] = 0, \quad \textit{for all } j = 1,\ldots,p.$$

It follows that $H(\beta) = -I(\beta)$ and, for the canonical link, Newton-Raphson and Fisher scoring are equivalent.

Remark 8.

The linear model is a GLM with identity link $\eta_i = g(\mu_i) = \theta_i$ and $Var(Y_i) = \sigma^2$, *for all* $i = 1,\ldots,n$. Therefore:

$$\begin{Bmatrix} w_i = \left[Var(Y_i)g'(\mu_i)^2\right]^{-1} = \sigma^{-2} \\ z_i = (y_i - \mu_i)g'(\mu_i) = y_i - \eta_i \end{Bmatrix}, \quad i = 1,\ldots,n.$$

Hence, neither $z + \eta = y$ nor $W = \sigma^{-2} I$ depend on β and the Fisher scoring algorithm converges in a single iteration to the usual least squares estimate.

Remark 9.

The standard errors (estimated standard deviations) are given by:

$$s.e.\left(\hat{\beta}_i\right)=\left[I\left(\hat{\beta}\right)^{-1}\right]_{ii}^{\frac{1}{2}}=\left[\left(X^T\hat{W}X\right)^{-1}\right]_{ii}^{\frac{1}{2}},\quad i=1,\ldots,p\,.$$

The asymptotic distribution of the MLE can be used to provide approximate large sample confidence intervals. We can find h such that:

$$P\left(-h\leq\frac{\hat{\beta}_i-\beta_i}{\left[I\left(\beta\right)^{-1}\right]_{ii}^{\frac{1}{2}}}\leq h\right)=\alpha\ \Leftrightarrow\ P\left(\hat{\beta}_i-h\left[I\left(\beta\right)^{-1}\right]_{ii}^{\frac{1}{2}}\leq\beta_i\leq\hat{\beta}_i+h\left[I\left(\beta\right)^{-1}\right]_{ii}^{\frac{1}{2}}\right)=\alpha\,.$$

The endpoints of this interval cannot be evaluated, because they also depend on the unknown parameter β. However, if we replace $I(\beta)$ by its MLE $I\left(\hat{\beta}\right)$, we obtain the approximate large sample $100\alpha\%$ confidence interval:

$$\left[\hat{\beta}_i-s.e.\left(\hat{\beta}_i\right)h\ ,\ \hat{\beta}_i+s.e.\left(\hat{\beta}_i\right)h\right].$$

As it is well known, for $\alpha=0.9,0.95,0.99$, we take $h=1.64,1.96,2.58$ respectively.

McCullagh and Nelder (1989) issued an important report on IRLS, where they applied it to the MLE of the parameters β in the linear predictor $\eta=\sum_{j=1}^{p}x_j\beta_j$ of a classic GLM. They did not use the dependent variable y but an *adjusted dependent variable* z, a linearized form of the link function applied to y. Also, the weights were functions of the fitted values $\hat{\mu}=\sum_{j=1}^{p}x_j\hat{\beta}_j$. The process is *iterative*, because both z and the weight W depend on the fitted values, for which only current estimates are available. The procedure is:

> *"Let $\hat{\eta}_0$ be the current estimate of the linear predictor, with corresponding fitted value $\hat{\mu}_0$ derived from the link function $\eta=g(\mu)$. Form the adjusted dependent variate with typical value:*
>
> $$z_0=\hat{\eta}_0+\left(y-\hat{\mu}_0\right)\left(\frac{d\eta}{d\mu}\right)_0,$$
>
> *where the derivative of the link is evaluated at $\hat{\mu}_0$. The quadratic weight is defined by:*

$$W_0^{-1} = \left(\frac{d\eta}{d\mu}\right)_0^2 V_0,$$

where V_0 *is the variance function evaluated at* $\hat{\mu}_0$. *Now regress* z_0 *on the covariates* $x_1, \ldots, x_p$ *with weight* W_0 *to give new estimates* $\hat{\beta}_1$ *of the parameters; from these, form a new estimate* $\hat{\eta}_1$ *of the linear predictor. Repeat until changes are sufficiently small."*

This algorithm has the advantage that it uses the data themselves as the first estimate of $\hat{\mu}_0$ and from this we get $\hat{\eta}_0$, $(d\eta/d\mu)_0$ and V_0. It is shown that the MLE for β_j are given by:

$$\sum W(y-\mu)\frac{d\eta}{d\mu}x_j = 0,$$

for each covariate x_j, where the summation is over the n units. Also, if I denotes the Fisher information matrix, then the new estimate $\hat{\beta}_{new}$ satisfies:

$$\left(I\hat{\beta}_{new}\right)_r = \sum_i W x_r \left\{\eta + (y-\mu)\frac{d\eta}{d\mu}\right\},$$

where the sum is over the n units. These equations have the form of linear weighted least-square equations with weight $W = V^{-1}\left(\frac{d\eta}{d\mu}\right)^2$ and dependent variate $z = \eta + (y-\mu)\frac{d\eta}{d\mu}$.

The iterative reweighting algorithm used by Basu and Lindsay (1993b) can be generally attributed to Beaton and Tukey (1974) and it is much simpler to apply than the Newton-Raphson method. Good references are also Holland and Welsch (1977) and Birch (1980), while Byrd and Pyne (1979) and Green (1984) discuss convergence results and Del Pino (1989) gives an extensive bibliography. Let:

$$Y_{n+1} = X_{nxp}\beta_{px1} + \varepsilon_{nx1}$$

be the standard regression model. A robust estimate $\hat{\beta}$ of β is found by minimizing:

$$\sum_{i=1}^{n} \rho\left(\frac{Y_i - X_i\beta}{\sigma}\right),$$

where σ is a known or previously estimated scale parameter. Let ψ represent the first derivative of ρ. Then $\hat{\beta}$ satisfies the estimating equation:

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{Y_i - X_i\hat{\beta}}{\sigma}\right) = 0, \quad for\ j = 1,2,\ldots,p\,,$$

where x_{ij} is the j-th component of the i-th row X_i of X. In order to avoid numerical methods, we use the weight function:

$$w(r) = \frac{\psi(r)}{r}$$

and the last equation becomes:

$$\sum_{i=1}^{n}\left(\frac{Y_i - X_i\hat{\beta}}{\sigma}\right) w\left(\frac{Y_i - X_i\hat{\beta}}{\sigma}\right) x_{ij} = 0, \quad for\ j = 1,2,\ldots,p\,.$$

This is a weighted version of the ordinary least squares. Now, this equation can be solved iteratively using a weighted least squares algorithm. Let W_β be the $n \times n$ diagonal matrix whose i-th diagonal element is:

$$w\left(\frac{Y_i - X_i\beta}{\sigma}\right).$$

Then for a given starting value β_0, the first iteration yields:

$$\beta_1 = \left[X^T W_{\beta_0} X\right]^{-1} X^T W_{\beta_0} Y\,.$$

This iteration scheme is continued till convergence to a specific level of tolerance is achieved.

3.3.2. Robustness in GLMs

Some of the well-known robust methods for estimating regression coefficients produce robust but usually inefficient estimators (Rousseeuw and Leroy (1987)). One of the first attempts to apply the minimum MHDE method to regression models was made by Pak and Basu (1994). They dealt with the MDE in linear regression models and showed that the estimators of the regression parameters are asymptotic normally distributed and efficient at the model, if the weights of the density estimators are appropriately chosen.

Markatou *et al.* (1997) discussed a method of weighting the likelihood equations and focused on the *weighted likelihood estimating equations*:

$$\sum_{i=1}^{n} w\left(x_i; M_\beta, \hat{F}\right) u\left(X_i; \beta\right) = 0,$$

where the weight function $w\left(x_i; M_\beta, \hat{F}\right)$ is selected such that it down-weights points that are inconsistent with the assumed model. They identified the outliers by the Pearson residual (also given by Definition 3):

$$\delta(t) = \frac{d(t)}{m_\beta(t)} - 1.$$

If the observed proportion of values at t is the same as the probability of observing t under the assumed model, then $\delta(t) = 0$; when the model is correctly specified, $\delta(t)$ converges to 0 almost surely. If there is no data observed at t, then $\delta(t) = -1$. For the model $m_\beta(t)$ the goodness-of-fit is examined by the *Pearson's chi-squared statistic*:

$$P^2 = n \sum m_\beta(t) \delta^2(t).$$

The most important thing is that the weights are functions of the Pearson's residuals and are defined via the RAF:

$$w\left(t, M_\beta, \hat{F}\right) = w\left(\delta(t)\right) = \frac{A\left(\delta(t)\right) + 1}{\delta(t) + 1}.$$

Defining the weights in this form, it is guaranteed that the weighted likelihood estimator is a root of the MDE equation:

$$\sum_{t} A\left(\delta(t)\right) \nabla_\beta m_\beta(t) = 0.$$

Markatou *et al.* (1997) noticed that the difference between two competing nested models in a logistic regression could be used as a chi-squared test of the smaller model against the larger. Also, we could expect the sum of weights to be roughly equal in magnitude to:

$$\left(n - n^*\right),$$

with: $n^* = \frac{1}{2} W_2^{-1}\left(k - 1 - \dim(\beta)\right)$ and: $W_2 = -A''(0)$. Thus, n^* reflects the loss of sample size necessary to achieve the improved robustness properties. Finally, they came up to some conclusions and made a comparison between MDE and weighted likelihood estimation:

i. if the weight functions generate increasing RAFs, then their method provides a link between the ML score equation and the MDE equations; otherwise, a selection criterion is needed to obtain the robust root

ii. their method provides furthermore a set of diagnostic sets

iii. the set of weights are extremely useful in testing goodness-of-fit and estimation.

In a following paper, Markatou (1999b) gave attention to linear regression and proposed a weighting scheme, where the components of the vector x_i, $i = 1,2,\ldots,n$ do not all take the same weight w_i, but each component x_{ij}, $i = 1,2,\ldots,n$, $j = 1,2,\ldots,p$ is down-weighted differentially. This is achieved by using a matrix of weights:

$$W(x_i) = diag\left(w_{i1}, w_{i2}, \ldots, w_{ip}\right)$$

There exists of a bound on a selected sensitivity and the efficiency of the sub-vector of parameters of interest is increased. The new estimators are asymptotically normal and have desirable robustness properties, but they are not invariant to non-singular reparametrization. This is not necessarily a disadvantage of the estimates, since there are many practical situations in which invariance might not be a desirable property.

Cantoni and Ronchetti (2001) proposed a class of robust testing procedures for GLMs. The main idea is that the robust estimation is made via another estimator, called the *Mallows quasi-likelihood estimator*. This is the solution of a special case of the estimating equations: $\sum_{i=1}^{n} \psi(y_i, \mu_i) = 0$, with ψ being the Huber function:

$$\psi_c(r) = \begin{cases} r, & if\ |r| \le c \\ c\, sign(r), & if\ |r| > c \end{cases}.$$

These procedures are very reliable in the presence of outlying points and other deviations from the assumed model. Further research includes the extension of these procedures to generalized estimating equations and to nonparametric models, like generalized additive models. However, there are computational problems.

By concluding this subsection, we refer to the paper of Lu *et al.* (2003), in which they proposed an estimation approach for finite mixtures of Poisson regression models based on MHDE. The methodology was also applicable to the standard Poisson regression model and the MHDE procedure could also be modified for a zero-inflated Poisson regression, which is a mixture of Poisson regression and a degenerate component whereby all of its mass is zero (see also Böhning, 1998). They developed a computational algorithm in order to extend the HELMIX algorithm of Karlis and Xekalaki (1998) to the finite mixtures of Poisson regression models setting. Through Monte Carlo procedures they showed that MHDE is a viable alternative to the MLE for both continuous and discrete random regressors. The MHDE outperformed the MLE, when the parameters were not sufficiently separated or near zero and for the regression coefficients of the component with small mixing probability, while the MLE appeared to be worse than the MHDE. Finally, a large sample size generally improved the performance of the estimators, but this might not be necessarily so, when the parameters were not well separated.

3.3.3. Iteratively Reweighted Estimating Equations (IREE)

As we have already seen for the simple case, where we did not use any covariates, the equations defined by the MDE are usually non-linear and we have to apply numerical methods in order to solve them. It is obvious that as the number of the parameters increase, so does the numerical difficulty. A similar technique to the IRLS was introduced by Basu and Lindsay (2004) with the major advantage of being vastly simpler to program. In addition, this new method does not require any matrix inversion per step and for example in a multidimensional normal model in d dimensions with $p=d+d(d+1)/2=d(d+3)/2$ unknown parameters, for estimating (μ,Σ), it requires $(p+2)$ numerical integrations. On the contrary, each step of the Newton-Raphson method requires $(p+1)(p+2)/2$ numerical integrations and the inversion of a p dimensional Hessian matrix.

First of all, for the estimating equation $\sum A(\delta(x))\nabla m_\beta(x)=0$, assuming that $\sum m_\beta(x)$ can be differentiated under the integral sign, we can write:

$$\sum A(\delta(x)-\lambda)m_\beta(x)\frac{\nabla m_\beta(x)}{m_\beta(x)}=0,$$

for any constant λ, or:

$$\sum w(x)\frac{\nabla m_\beta(x)}{m_\beta(x)}=0,$$

where:

$$w(x)=A(\delta(x)-\lambda)m_\beta(x).$$

This is a weighted version of the estimating equation of the likelihood disparity:

$$\sum d(x)\frac{\nabla m_{\beta}(x)}{m_{\beta}(x)}=\sum \delta(x)\nabla m_{\beta}(x)=0.$$

If $m_{\beta}(x)$ is in the exponential family and $\beta=(\beta_1,\beta_2,\ldots,\beta_p)$ represents the set of the natural parameters, then a relationship of the following form holds:

$$\frac{\nabla_{\beta_i} m_{\beta}(x)}{m_{\beta}(x)}=K(\beta)\left[S_i(x,\beta)-\beta_i\right]$$

for some functions K and S_i, which may depend on β. Hence, the i-th equation of the above weighted expression can be written as $\sum w(x)K(\beta)\left[S_i(x,\beta)-\beta_i\right]=0$ and thus we arrive at the *fixed-point equation* $\beta=F(\beta)$ with:

$$\beta_i=\frac{\sum w(x)S_i(x,\beta)}{\sum w(x)}.$$

The iteration is continued till convergence is achieved. Basu and Lindsay (2004) referred to this algorithm as the *IREE* algorithm. When they used $\lambda=A(-1)$, the weights $w(x)$ were non-negative, since the RAF $A(\delta)$ is increasing on $[-1,\infty]$, and they referred to this case as the *standard IREE* algorithm (or the IREE with standard weights).

Example 1

For the one-parameter exponential family, let $\mu=\beta$ be the mean and V the variance for the model m_{β}. Then:

$$\frac{\nabla_{\beta_i} m_{\beta}(x)}{m_{\beta}(x)}=\frac{(x-\mu)}{V}$$

and for μ the IREE will solve the equation:

$$\sum w(x)\left\{\frac{(x-\mu)}{V}\right\}=0.$$

This gives the fixed-point equation for μ as:

$$\mu=F(\mu)=\frac{\sum x\,w(x)}{\sum w(x)}.$$

In general, for a univariate parameter β:

$$F(\beta)=\frac{\sum w(x)S(x,\beta)}{\sum w(x)},$$

if $\frac{\nabla m_\beta(x)}{m_\beta(x)}=K(\beta)\left[S(x,\beta)-\beta\right]$ for some functions K and S.

The convergence of the fixed-point algorithm applied to the fixed-point formulation we have already seen, depends on the derivative of $F(\beta)$ at the solution and the rate of the convergence is quadratic, if this derivative is zero. If $S_i(x,\beta)=S_i(x)$ is independent of β, direct differentiation of:

$$F(\beta)=\frac{\sum w(x)S(x,\beta)}{\sum w(x)},$$

combined with the result that at the solution $\beta=F(\beta)=\frac{\sum w(x)S(x)}{\sum w(x)}$, gives:

$$F'(\beta)=\frac{\sum w'(x)\left(S(x)-\beta\right)}{\sum w(x)}$$

at the solution, where:

$$w'(x)=\frac{\partial w(x)}{\partial \beta}.$$

A nice improvement of the standard IREE is possible, if we allow negative weights. Basu and Lindsay (2004) referred to the case of $\lambda=-1$ as the *optimal IREE* algorithm (or the IREE with optimal weights).

Through some examples Basu and Lindsay (2004) tried to compare the several methods. As expected, the Newton-Raphson method converges substantially faster then the standard IREE. However, it is fair to say that the convergence of the standard IREE is moderately quick. For the Beran (1977) data, while the standard IREE requires about 2.5-3.5 times the number of steps needed for the Newton-Raphson method to converge, overall it only requires just about double the number of numerical integrations or less compared to what is necessary for the Newton-Raphson. This is because at each step the standard IREE requires only 4 numerical integrations whereas the Newton-Raphson method requires 6 integrations involving much more complex functions. Also, the optimal IREE is far superior to the standard IREE and comparable to the Newton-Raphson method.

Remark 10.

Sometimes a small decrease in efficiency of the optimal IREE is observed, which depends on outliers. The optimal IREE is a quadratically convergent algorithm only *at the model*. In terms of real data examples this means that the algorithm will perform best, when the data roughly follow the pattern dictated by the model. As an observation "goes" far away from the bulk of the data and the hypothesized model, the optimal IREE will need more steps to converge. In fact, the actual MHDE also are affected most by mid sized outliers. However, when the outlier becomes unacceptable large, most robust MDE would be able to clearly distinguish it as such and downweight it almost entirely. The majority of the data (excluding the outlier), which follow the model closely, would now govern primarily the performance of the estimator (as well as the IREE algorithm). For large outliers and robust initial estimates, the weights for values of X around the outlier are practically equal to zero (either for standard or optimal IREE), so that in extreme cases the algorithm works as if the outlier was simply not there. Consequently, the algorithm converges quickly.

3.4 MM algorithms

These kinds of algorithms can be thought as a part of the MM algorithms, which are an extension of the known EM (Expectation-Maximization) algorithm. A lot of things can be said about EM algorithm, but this is beyond the scope of our work here. Some central aspects for the reader to refer to EM algorithm can be found in Dempster *et al.* (1977), Little and Rubin (1987) and McLachlan and Krishnan (1997). The main idea is that the EM algorithm is an optimization transfer algorithm that depends on the notion of incomplete or missing data. On the other hand, MM algorithms do not involve missing data and Ortega and Rheinboldt (1970) first enunciated the general principle behind them. Lange *et al.* (2000) illustrated in their article a number of specific examples drawn from the statistical literature. Many important results emerged from this work and the discussion article that followed by Leeuw and Michailidis (2000), Wu (2000), Meng (2000), Groenen and Heiser (2000), Gelman (2000) and Hunter and Lange (2000). One of the topics discussed was the name itself and probably the most appropriate seems to be "MM algorithms". In minimization problems, the first M of MM stands for *majorize* and the second M for *minimize*. The opposite holds in maximization problems. Generally, MM algorithms seem to be easier to understand and sometimes easier to apply than EM algorithms.

Since our work is focused on the Poisson regression model, it is worth mentioning the basic relevant result. If we have the observation y_i for case i, it is convenient to write the mean $d_i e^{x_i'\theta}$ as a function of a fixed offset $d_i > 0$ and a covariate vector x_i. If α_i are nonnegative coefficients and sum to 1, one method of constructing a majorizing function depends directly on the inequality:

$$f\left(\sum_i \alpha_i v_i\right) \leq \sum_i \alpha_i f(v_i)$$

defining a convex function $f(u)$. It is helpful to extend this inequality to:

$$f(c'v) \leq \sum_i \frac{c_i w_i}{c'w} f\left(\frac{c'w}{w} v_i\right),$$

when all components c_i and w_i of the vectors c and w are positive. In a medical imaging context, De Pierro (1995) introduced another method of optimization transfer. If $f(u)$ is convex, then he invoked the inequality:

$$f(c'v) \leq \sum_i \alpha_i f\left(\frac{c_i}{\alpha_i}(v_i - w_i) + c'w\right),$$

where:

$$\alpha_i \geq 0,\ \sum_i \alpha_i = 0 \text{ and } \alpha_i > 0 \text{ whenever } c_i \neq 0.$$

Furthermore, there are no positivity restrictions on the components c_i or w_i. Since the function:

$$f_i(u) = -d_i e^u + y_i u$$

is concave, this last inequality applies to the log-likelihood:

$$L(\theta) = \sum_{i=1}^{m} \left(-d_i e^{x_i'\theta} + y_i \ln d_i + y_i x_i'\theta - \ln y_i!\right).$$

In maximizing the corresponding surrogate function, one step of Newton's method yields the update:

$$\theta_j^{n+1} = \theta_j^n + \frac{\sum_{i=1}^{m} x_{ij}\left(y_i - d_i e^{x_i'\theta^n}\right)}{\sum_{i=1}^{m} d_i e^{x_i'\theta^n} x_{ij}^2 / \alpha_{ij}}.$$

The reader can consult Becker *et al.* (1997) for more details and other results of how De Pierro's method operates in GLMs. It is noteworthy that minorization by a quadratic function fails for Poisson regression, because the functions $f_i(u)$ do not have bounded curvature.

In this last paragraph, we shall see how Hunter and Lange (2004) applied the MM algorithms on a Poisson sports model. If we address back to section 2 and recall a simplified version of Maher's (1982) model, then the number of goals scored by team i against team j follows a Poisson process with intensity $e^{o_i - d_j}$, where o_i is the offensive ability for team i and d_j is the defensive ability for team j. If $\theta = (o, d)$ is the parameter vector, then the corresponding Poisson log-likelihood function is:

$$l_{ij}(\theta) = p_{ij}\left(o_i - d_j\right) - e^{o_i - d_j} - \ln p_{ij}!,$$

where the parameters should satisfy a linear constraint, such as:

$$\sum_i o_i + \sum_j d_j = 0.$$

Also, under the 2 assumptions that:

- *Different games are independent of each other and*
- *Each team's goals in a single game are independent of its opponent's,*

the full data log-likelihood is obtained by summing $l_{ij}(\theta)$ over all pairs (i, j). Setting the partial derivatives of the log-likelihood equal to zero leads to the equations:

$$e^{-\hat{d}_j} = \frac{\sum_i p_{ij}}{\sum_i e^{\hat{o}_i}} \quad \text{and} \quad e^{\hat{o}_i} = \frac{\sum_j p_{ij}}{\sum_j e^{-\hat{d}_j}}.$$

Of course, these equations do not admit a closed form solution, so we turn to an MM algorithm. Hunter and Lange (2004) via a procedure that uses some well-known inequalities, proved that the updates take actually the form:

$$o_i^{(m+1)} = \frac{1}{2}\ln\left\{\frac{\sum_j p_{ij}}{\sum_j e^{-o_i^{(m)} - d_j^{(m)}}}\right\} \quad \text{and} \quad d_j^{(m+1)} = -\frac{1}{2}\ln\left\{\frac{\sum_i p_{ij}}{\sum_i e^{o_i^{(m)} + d_j^{(m)}}}\right\}.$$

A good modification of this algorithm is to update the o vector before the d vector in each iteration, in order to use the updated subsets of the parameters as soon as they become available. Hence, we could replace the formula for $d_j^{(m+1)}$ above by:

$$d_j^{(m+1)} = -\frac{1}{2}\ln\left[\frac{\sum_i p_{ij}}{\sum_i e^{o_i^{(m+1)}+d_j^{(m)}}}\right].$$

In practice, an MM algorithm often takes less number of iterations, when we cycle through the parameters updating one at a time than when we update the whole vector at once. Such versions of MM algorithms are called *cyclic MM algorithms* and they generalize the ECM algorithms of Meng and Rubin (1993).

CHAPTER 4

Application

As we have discussed earlier in Sections 1 and 2, for some reasons we prefer to fit the Poisson distribution for our soccer data, although we have the suspicion that the underlying distribution is a (probably small) deviation of it. For soccer data, there are some references (see Karlis and Ntzoufras (2000, 2003)), which indicate that the goals scored by each team are slightly over-dispersed. The use of MLE in soccer data has several disadvantages, so we will try to apply robust methods via the WMLE, using of course the Poisson distribution. This comes in the wake of Douglas (1994)'s remark, that short-tailed observed frequency distributions are often well fitted by a number of different theoretical discrete distributions, with little discriminatory power. The theory of robustness deals not only with model deviation but also with data contamination. One of the solutions proposed in the literature is the "correction" of surprising observations by downweighting data points with large Pearson residuals.

In this Section, we will try to model soccer matches through a GLM taking into account the actual scores and not the outcomes. Generally, this could be very dangerous, since "unexpected" scores could influence heavily the results. In the beginning of Section 3, we said that a win with the "unusual" score of 7-0 is not the same with a win of a more "usual" score such as 1-0, 2-0 or 2-1. Of course, we are not talking about the points won from that game; we are talking about how strong this team appears to be and this is measured via its offensive and defensive ability. This has direct connection to the rating of the teams. Bassett (1997) gave a very good relative example, where different methods of estimating the parameters of a standard linear model can produce very different results. Hence, comparing least squares with least absolute value (or L_1), he found such differences in ratings and relative rankings. The main cause is, that the least squares can be strongly influenced by a single observation, since a team's rating depends on the relative strength of its opponent. This interdependence combined with the sensitivity of least squares is why only few "unusual" scores greatly influence the rating estimates.

4.1 An example

We shall now present a virtual example to show exactly what we mean. In the season 2004-2005, the schedule of the UEFA Cup changed. After the preliminary rounds, the best 40 teams are separated in 8 groups of 5 teams. Each team has to face every other of the 4 opponents, but the innovation is that do not play each other twice. All the teams have to play against the others only once. The 2 games will be played in home ground and the other 2 away. The three teams of each group that collect most points, can continue to the next round.

Let us suppose that one of the eight groups consists of teams A, B, C, D and E. In Columns 1 and 2 of Table 4.1.1 the matches and the hypothesized results are listed. Columns 3 and 4 refer to two kinds of weights produced by the WMLE for each score separately (details will be given later in Section 4.2.2), while Columns 5-7 show the expected scores from the three cases. We observe that all the weights take values from 0 to 1 and small weights indicate that the scores are "unexpected". We shall show that these two different weighting schemes produce different results one from the other and from the MLE. Especially in this case where the size of the data set is small, only one or very few "unusual" scores may influence a lot the ratings and the estimates.

Matches	**Scores**	$w_{ij(1)}$	$w_{ij(2)}$	**Expected Scores**		
				MLE	**WMLE (I)**	**WMLE (II)**
A-B	1-5	0.838-0.601	1.000-0.983	1.30-3.58	1.27-3.57	1.29-3.51
C-D	3-3	**0.508**-0.718	0.972-1.000	1.63-2.65	1.49-2.68	1.75-2.66
C-A	4-1	0.653-0.824	0.998-0.998	3.13-1.55	3.02-1.48	3.19-1.48
B-E	1-3	0.823-0.708	0.996-1.000	1.60-2.49	1.49-2.45	1.61-2.52
A-E	2-3	0.656-0.719	0.969-1.000	0.99-3.32	0.93-3.31	1.00-3.45
D-B	3-2	***0.698-0.766***	***0.998-1.000***	2.23-1.86	2.29-1.76	2.32-1.93
D-A	1-1	0.607-0.844	0.842-1.000	2.97-1.15	3.10-1.10	3.17-1.19
E-C	3-1	0.622-0.752	0.985-0.965	1.83-2.26	1.69-2.14	1.91-2.21
B-C	1-2	0.791-0.730	0.983-0.995	1.97-2.98	1.82-2.91	1.95-2.87
E-D	0-3	0.783-0.690	**0.712**-0.996	1.36-2.15	1.25-2.20	1.54-2.20

Table 4.1.1: Final scores and weights and expected scores from the three approaches

For example, for Method I team C is not expected to score 3 goals against team D, but for Method II the less logical score is the one of team's E, which does not succeed to score any goal against team D. This probably means that according to Method I and from team's C offensive ability in combination with team's D defensive ability, we should expect team C to score fewer goals in the specific match. Similar thoughts can be made for Method II in the match between teams E and D. We can also observe many more differences, like for example the game between teams D and B. For Method II the final result is absolutely expected, since both weights are equal or very close to 1, but for Method I we cannot draw safe conclusions. The weighs of 0.698 and 0.766 are mid-high values in the space $[0,1]$. This could be translated as logical outcome for the match, but we should probably expect fewer goals to be scored. The two methods will be discussed extensively in paragraph 4.2.2.

Table 4.1.2 shows the full data of the group. Except for the final ranking, the total points and the number of wins, draws and losses respectively (Columns 2-4), it is very interesting to take a look at Column 5.

Teams	Ranking	Points	W-D-L	Score Differences	
E	1	9	3-0-1	+2	(9-7)
D	2	8	2-2-0	+4	(10-6)
C	3	7	2-1-1	+2	(10-8)
B	4	3	1-0-3	0	(9-9)
A	5	1	0-1-3	-8	(5-13)

Table 4.1.2: Full data of the group

From the total points won, we know which is the true ranking (here, the best team is E and the worst is A). However, if we rely on the score differences (i.e. the total goals scored minus the total goals conceded), there seems to be some kind of confusion. In this case, team D appears as the best one, since it has the biggest (positive) difference, while the best attacks belong to teams D and C.

Table 4.1.3 shows the model's (see paragraph 4.2) expected final ranking and total points gathered. The ranking is of great interest, because it determines the next round's opponents for the three teams that will continue to the competition.

Teams	Observed		MLE		WMLE (w1)		WMLE (w2)	
	Points	Rank	Points	Rank	Points	Rank	Points	Rank
E	9	1	6.644	3	6.532	2	6.909	2
D	8	2	7.780	1	8.143	1	7.779	1
C	7	3	6.679	2	6.530	3	6.448	3
B	3	4	5.418	4	5.308	4	5.573	4
A	1	5	1.809	5	1.815	5	1.763	5

Table 4.1.3: Expected ranking and points

MLE suggests that the best three teams (ordered) are D, C and E. On the other hand, both Methods I and II of WMLE find as best teams D, E and C. The problem with MLE is that there is strong dependence on the goals scored. Hence, it chooses as best team one of D and C, because these two have the best attacks (each one scored 10 goals) and it founds D as the best team, because it has better score difference (+4). The same procedure is followed for positions 3 and 4 between teams E and B. This does not happen in WMLE, which clearly find E as the second best team (D is still expected to finish in the 1st place). Nonetheless, from the (expected) points, teams E and C are very close for Method I, but for Method II the difference is slightly more obvious.

4.2 A close look at the Greek League

Data refer to the season 2003-2004 of the Greek National A Division (GNA). According to Karlis and Ntzoufras (1998), soccer data form a kind of three-way "contingency table" with counts of the goals scored by team A, against team B, playing in ground C. The model can estimate the offensive parameters (by the factor A), the defensive parameters (by the factor B) and the home effect (C).

GNA consists of 16 teams playing with each opponent twice, once at home and once in away football grounds; so, each team plays 30 total games, 15 in home and 15 away. The final league consists of 240 soccer games (or 480 observed scores). Every win gives three (3) points to the winner, every draw one (1) point to each opponent and if a team looses, gets zero (0) points from that match. The team, which collects the highest number of points at the end of the season, is the winner of the league and becomes champion. Positions 2 and 3 are of crucial interest, since they give the right (including the champion) of playing in the prestigious and profitable European cup "Champions League" for the following year. Also, positions 4-6 give the corresponding right for the "UEFA" cup. Finally, for the bottom of the League table, the two teams that collect the fewest points are automatically relegated to the lowest division and are substituted in the next season by the two best teams of Greek National B Division (GNB). There is the rule that the third weakest team of GNA plays against the third strongest team of GNB. This is a double "fight" in both football grounds. If it is needed, extra time of 30 minutes is played and if it is not yet clear which is the winner, there is the process of penalties. The winning team will play in the GNA for the following year, while the loosing one will struggle in the GNB.

Ranking	Teams	Points	Goals	Goal Diff.	W-D-L
01	Panathinaikos	77	62-18	+44	24-05-01
02	Olympiakos	75	70-19	+51	24-03-03
03	PAOK	60	47-27	+20	18-06-06
04	AEK	55	57-32	+25	16-07-07
05	Aigaleo	52	37-26	+11	15-07-08
06	Panionios	47	40-29	+11	12-11-07
07	Chalkidona	45	40-39	+01	13-06-11
08	Iraklis	42	40-39	+01	12-06-12
09	Ionikos	33	33-43	-10	09-06-15
10	Xanthi	30	28-42	-14	08-06-16
11	OFI	29	27-44	-17	07-08-15
12	Kallithea	27	37-42	-05	05-12-13
13	Aris	27	24-46	-22	07-06-17
14	Akratitos	23	31-69	-38	05-08-17
15	Panileiakos	21	28-56	-28	04-09-17
16	Proodeftiki	20	26-56	-30	04-08-18

Table 4.2.1: GNA's full data

Greek league data were taken by an International Soccer Server web site available in the URL address **http://www.bettinggenius.com**. In the Appendix there is a list of some relative web links. Some basic data are concentrated in Table 4.2.1. Columns 1-5 show the final ranking of the teams, the points they gathered, they goals they scored and suffered and the goal difference respectively, while in Column 6 there are the numbers of wins (W), draws (D) and losses (L) for each team.

4.2.1 Estimating the parameters using MLE

In our application we use the Poisson log-linear model with the form:

$$\begin{cases} y_{ij} \sim Poisson\left(\lambda_{ij}\right) \\ \log\left(\lambda_{ij}\right) = \mu + h + o_i + d_j \end{cases}, \quad i,j = 1,2,\ldots,p$$

where p is the number of the teams in the league, y_{ij} and λ_{ij} are the observed and the expected number of the goals, respectively, scored by the home team i against the away team j; μ is a constant parameter, h is the all teams' common home effect parameter, o_i stands for the offensive ability of team i and d_j encapsulates the parameter for the defensive performance of team j.

Table 4.2.2 gives a first insight about the estimating parameters taken from the model including a constant parameter (intercept μ) and this model can be the basis for predicting future outcomes. Columns 2-4 show the estimated parameters, when the model uses all the data (240 games). On the other hand, Columns 5-7 show the estimated parameters, when the model uses the data only from the 1st round (first 120 games). We also fitted the same model but without the constant μ. The conclusions taken from both models are almost identical. The small differences in the coefficients were in their values themselves and not comparing each other. So, from now on, we shall not give results for both models but only for the one, which contains μ.

Teams	**Using ALL the data**			**Data only from the 1st round**		
		Offensive (o_i)	**Defensive** (d_i)		**Offensive** (o_i)	**Defensive** (d_i)
Intercept	*-0.040*			*0.007*		
Home effect	*0.370*			*0.338*		
Panathinaikos		0.471	-0.682		0.378	-0.528
Olympiakos		0.594	-0.613		0.501	-0.720
PAOK		0.208	-0.301		0.232	-0.372
AEK		0.411	-0.113		0.487	-0.182
Aigaleo		-0.033	-0.355		-0.042	-0.617
Panionios		0.050	-0.241		0.151	-0.045
Chalkidona		0.067	0.057		0.028	0.034
Iraklis		0.067	0.057		-0.003	-0.105
Ionikos		-0.120	0.143		-0.059	0.263
Xanthi		-0.286	0.111		-0.266	0.118
OFI		-0.319	0.156		-0.210	0.285
Kallithea		-0.003	0.173		0.017	0.200
Aris		-0.434	0.195		-0.374	0.235
Akratitos		-0.074	0.618		-0.268	0.793
Panileiakos		-0.262	0.400		0.211	0.409
Proodeftiki		-0.337	0.396		-0.783	0.235

Table 4.2.2: Model details

We take the constraint that all the offensive coefficients add up to zero and so do the defensive ones (that is, $\sum_{i=1}^{16} o_i = \sum_{i=1}^{16} d_i = 0$) and we can say, that each team is compared to a hypothesised "average" team, whose coefficients are expected to take values equal or very close to 0. All 16 teams can also easily be compared each other, since large values of o_i correspond to teams that have good attacks. One the other hand, small values of d_i belong to the teams with good defences.

According to the model parameters, at the end of the 1st round, Olympiakos had clearly a better attack and defence than Panathinaikos. It seems though, that in the 2nd round Panathinaikos improved its performance. Thus, at the end Olympiakos had better attack but slightly worst defence than Panathinaikos, which finished first in the league. Conversely, during the 1st round, the defence produced by Proodeftiki, Aris and Akratitos (and only one "step" away Xanthi) were the worst, while the worst attack came from Akratitos, Panileiakos and OFI. Nevertheless, what it really counts is the end of the championship. There, it was revealed that the three teams with the worst defence were Akratitos, Panileiakos and Proodeftiki (all three of them were relegated to GNB), while Aris, Proodeftiki and OFI had the worst attack. Maybe the entire above stand up for the fact that the teams should give more attention to their defences (either by buying good defending players or by putting in practice better defending tactics on the football grounds). One more final remark is that Chalkidona and Iraklis seem to be equal in strength. This appears first in Columns 4 and 5 of Table 4.2.1 and more intense in Columns 2 and 3 of Table 4.2.2. Just to reveal to the reader this equivalence, we give the corresponding coefficients with bigger accuracy:

$$\begin{aligned} &\mathit{Chalkidona}: && (o_7, d_7) = (0.066855433,\ 0.05673122) \\ &\mathit{Iraklis}: && (o_8, d_8) = (0.066857054,\ 0.05673065). \end{aligned}$$

For practical purposes, when statistical software (supporting GLMs) is not available, someone can estimate such coefficients as the mean number of goals scored and conceded, respectively. Norman and Clarke (1995) described a way of calculating the home effect and the probability of a win via simple packages supporting probability function calculation. All these calculations do not need special statistical knowledge and any non-statisticians can easily perform them. Also, the number of goals scored by a team is a sufficient indicator for the strength of a team, since a team must score in order to win. Karlis and Ntzoufras (2000) proved statistically this statement and they actually found high correlations between the final ranking and the number of goals scored and conceded by each team; hence, the goals scored can be used in order to determine the performance of a team. Furthermore, it was found that the distribution of the number of the goals is slightly over-dispersed relative to the simple Poisson distribution (see also Karlis and Ntzoufras (2003)).

From the assumed model, we used the above estimated parameters in order to generate replications of leagues. The total team points and the ranking of each replicated league were used to assess the distribution of the final league under the assumption that the model used is a sufficient summary of reality and the teams have the same performance as in the observed league. The analysis accounts for corrections of games that were surprisingly unfair or won by luck. For each data set we simulated 10000 leagues and the predicted results are shown in Table 4.2.3. The model predicted correctly 106 out of 240 or 44.17% of the total games played.

Teams		**Model using all the data**		**Model using data only from the 1st round**	
	Observed Points	**Predicted Ranking**	**Predicted Points**	**Predicted Ranking**	**Predicted Points**
01. Panathinaikos	77	**02**	**67.540**	**02**	**67.330**
02. Olympiakos	75	**01**	**69.861**	**01**	**69.475**
03. PAOK	60	**04**	**53.702**	**04**	**54.013**
04. AEK	55	**03**	**55.973**	**03**	**55.794**
05. Aigaleo	52	05	48.428	**06**	**48.152**
06. Panionios	47	06	48.320	**05**	**48.680**
07. Chalkidona	45	07	42.387	**08**	**42.049**
08. Iraklis	42	08	42.199	**07**	**42.205**
09. Ionikos	33	10	35.138	10	35.071
10. Xanthi	30	11	32.479	11	32.245
11. OFI	29	12	30.139	12	31.129
12. Kallithea	27	**09**	**37.654**	**09**	**37.681**
13. Aris	27	13	27.143	13	27.239
14. Akratitos	23	**16**	**24.259**	**16**	**23.976**
15. Panileiakos	21	**14**	**25.706**	**14**	**25.744**
16. Proodeftiki	20	**15**	**24.278**	**15**	**24.302**

Table 4.2.3: Expected ranking and points

It is clear that Olympiakos (2nd in rank) had a higher expected value of points than Panathinaikos (which won the championship) and Olympiakos is the best team with average difference of more than 2 points. This could easily be interpreted, since Olympiakos had better attack and defence, but lost "important" games against PAOK (3rd in rank) and AEK (4th in rank) as someone can see in Table 4.2.4 (Columns 4-7 show which percentage the model gave to each outcome and the possible score). It should be noted that the 2 games between these 2 challengers of the title ended without a winner. Also, AEK should had finished 3rd and should play in the place of PAOK at the Champion's League in the following year. Proodeftiki, Panileiakos and

Akratitos are indeed the three worst teams in the league, but they should probably finish in different ranking. This means that the 3rd from the end should be Panileiakos, which should be the one (and not Akratitos) with the right to fight with Ergotelis (the third strongest team of GNB) for a place in GNA for the following year. Finally, there are some other small differences, but the most impressive is that Kallithea should not live under constant strain and reach the 9th place instead of the 12th.

Match	**Score**	**Outcome**	**Probability of**			**Expected Score**
			1	**X**	**2**	
Panathinaikos-PAOK	3-0	*Win*	*0.63*	0.24	0.13	1.65-0.60
PAOK-*Panathinaikos*	1-2	*Win*	0.28	0.30	*0.42*	0.87-1.14
AEK-*Panathinaikos*	2-2	*Draw*	0.29	*0.27*	0.44	1.06-1.38
Panathinaikos-AEK	2-1	*Win*	*0.67*	0.20	0.13	1.99-0.73
Olympiakos- Panathinaikos	1-1	*Draw*	0.47	*0.29*	0.24	1.27-0.83
Panathinaikos -Olympiakos	2-2	*Draw*	0.43	*0.30*	0.27	1.21-0.88
PAOK-*Olympiakos*	0-2	*Win*	0.27	0.28	*0.45*	0.93-1.29
Olympiakos-PAOK	1-2	***Loss***	0.67	0.21	***0.12***	1.87-0.64
AEK-*Olympiakos*	0-1	*Win*	0.28	0.25	*0.47*	0.14-1.56
Olympiakos -AEK	0-1	***Loss***	0.71	0.18	***0.11***	2.25-0.78

Table 4.2.4: "Important" games for the 2 challengers of the title

From Table 4.2.3, we observe that the model performs well regarding the teams in the middle of the League Table, but there are deviations at the two edges of it. More specifically, according to the model, the top teams seem to have more points in reality that they should really get, while the opposite happens for the worst teams. This is shown clearly in Figure 4.2.1, where the observed points won by the teams are plotted against the expected points including the straight line for the perfect fit. The first teams are indeed above that line, while in the middle things seemed to be relatively satisfactory and the last teams go below the line. We have already said much about robustness theory and this is absolutely relevant to it, since we assume a model, but there is possibly some deviation from it. Also, in Figure 4.2.2 we make the graphical representation of the absolute values of the differences between these points (which give the residuals of the model, in a sense) and there is obviously a U-shape.

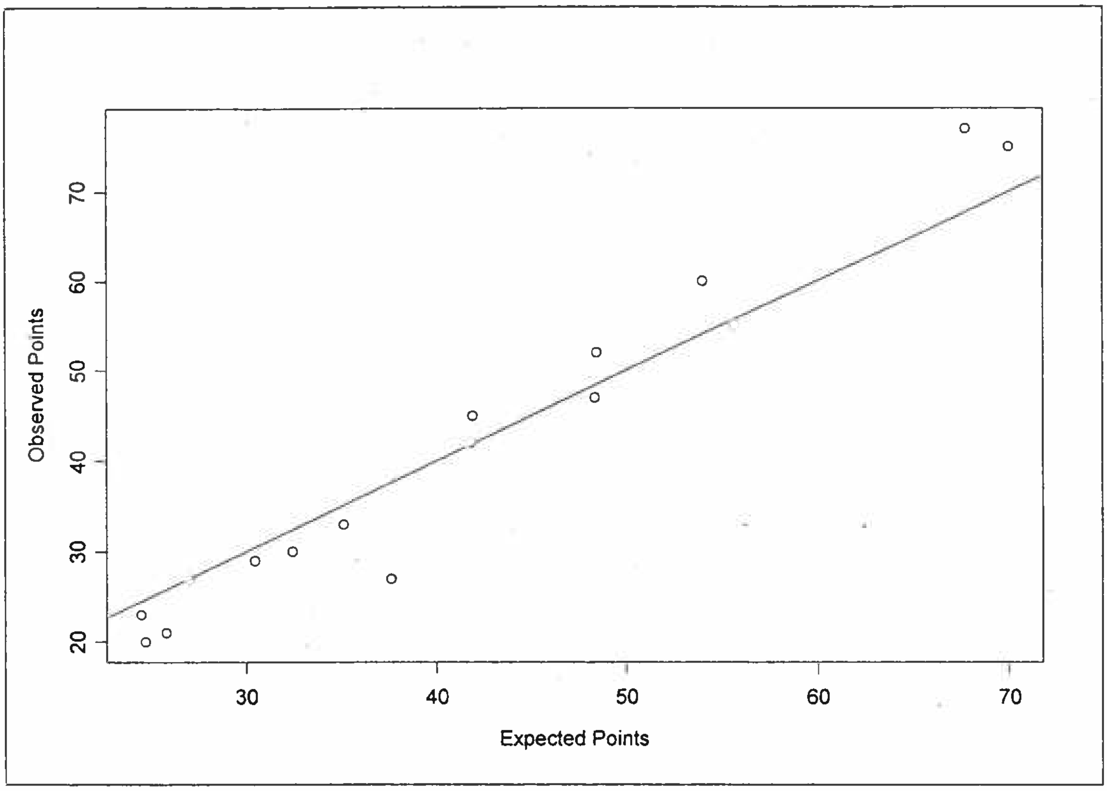


Figure 4.2.1: Observed Vs Expected Points

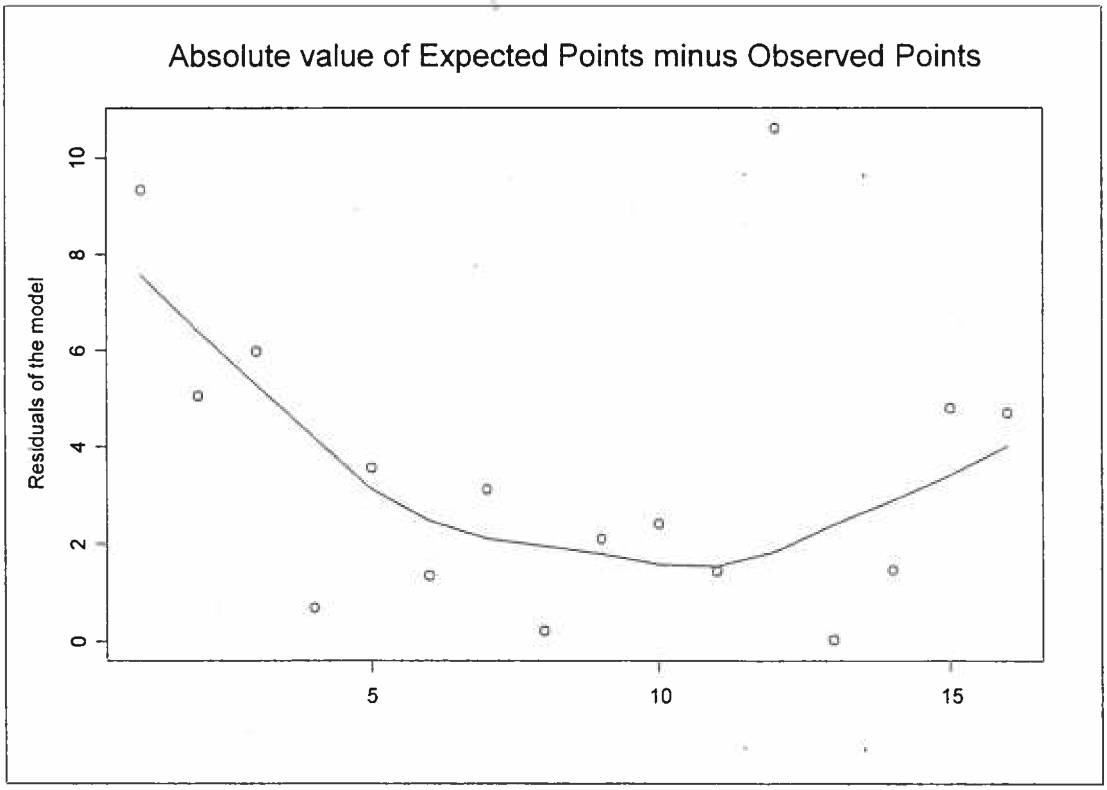


Figure 4.2.2: Existence of U-shape in the residuals of the model

As we said before, the model could predict correctly 106 out of 240 total played. It is very interesting to see how these predictions are distributed. The three possible outcomes are represented as 1 for a home win, X for a draw and 2 for an away win. The first two rows of Table 4.2.5 show how close was each predicted number of an outcome to what really happened. Our model predicted that there should have been 114 home wins, 55 draws and 71 away wins, but the true numbers were 115, 56 and 69 respectively. The point that is of great interest for most is the second part of Table 4.2.5; Rows 3-4 show that the model predicted correctly 53.91% of the total home wins (62 out of 115), 23.21% of the total draws (13 out of 56) and 44.93% of the total away wins (31 out of 69). This fact is a clear indication that this model (as most of the statistical models) lacks in the prediction of the draws, although it does a fairly good job with the wins (either home or away).

Outcomes	**1**	**X**	**2**
Observed number	115	56	69
Predicted number	114	55	71
Number of correct predictions	62	13	31
Percentage of correct predictions	53.91%	23.21%	44.93%

Table 4.2.5: Predictions of the model for the three outcomes

A very important issue is also the goodness-of-fit of the model. Discrete data allow calculating probabilities of single outcomes after the model has been estimated. For count data models, we can use the predictions to evaluate the goodness-of-fit (see also Winkelmann (2003)). A related procedure for the Poisson model is based on the Pearson statistic:

$$P = \sum_{i=1}^{n} \frac{\left(y_i - \hat{\lambda}_i\right)^2}{\hat{\lambda}_i},$$

where y_i are the observed values and $\hat{\lambda}_i$ the predicted ones. If the Poisson model is correctly specified, then it holds $E\left[(y_i - \lambda_i)^2 / \lambda_i\right] = 1$, so $E\left[\sum_{i=1}^{n} (y_i - \lambda_i)^2 / \lambda_i\right] = n$. In our case, we found $P = 385.4258 < 480 = \mathrm{n}$, so the Poisson assumption is valid.

Coming back to the results of our model, in the top half of Table 4.2.6 we can see the distribution of ranks after using all the data for the two edges of the League Table. According to this, Olympiakos should have been the champion, since it had higher probability to win the league than Panathinaikos. The case of any other team to win the championship is infinitesimal. The three last columns show the probability of ending up in the last place (1), in the last two places (2) or in the last three places (3). The probabilities of Columns 2 and 3 add up to 1, of Column 4 to 2 and of Column 5 to 3. Combining columns 3-5, someone can find Akratitos as the worst team, which is consistent with Table 4.2.3. Proodeftiki seems indeed to be the second weakest team, while Panileiakos fills the triad of the weakest links of GNA. Also, the bottom half of Table 4.2.6 gives the distribution of ranks for the model after using the data only from the 1st round. The results from these two halves of Table 4.2.6 are very similar.

ALL THE DATA				
Teams	**Probability Of Winning The Champion**	**Probability of Relegation (1)**	**Probability of Relegation (2)**	**Probability of Relegation (3)**
Panathinaikos	0.3829			
Olympiakos	**0.5806**			
Other team	0.0325			
Kallithea		0.0059	0.0214	0.0516
Aris		0.1375	0.3076	0.4809
Akratitos		**0.2852**	**0.4991**	**0.6626**
Panileiakos		0.1899	0.3917	**0.5698**
Proodeftiki		0.2694	**0.4872**	**0.6625**

DATA ONLY FROM THE 1st ROUND				
Teams	**Probability of Winning The Champion**	**Probability of Relegation (1)**	**Probability of Relegation (2)**	**Probability of Relegation (3)**
Panathinaikos	0.3803			
Olympiakos	**0.5830**			
Other team	0.0367			
Kallithea		0.0061	0.0216	0.0514
Aris		0.1348	0.3001	0.4670
Akratitos		**0.2839**	**0.4986**	**0.6705**
Panileiakos		0.1951	0.3910	**0.5748**
Proodeftiki		0.2699	**0.4923**	**0.6668**

Table 4.2.6: Probability of ranks after using all the data and the data only of the 1st round (the three last columns differ to the number of the worst teams)

ALL THE DATA

Teams	Observed Rank	Probability of ending up in the Observed position	Expected Rank	Probability of ending up in the Expected position
Panathinaikos	01	0.3829	02	0.4788
Olympiakos	02	0.3408	01	0.5806
PAOK	03	0.2591	04	0.2802
AEK	04	0.2516	03	0.3618
Aigaleo	05	0.2143	05	0.2143
Panionios	06	0.1947	06	0.1947
Chalkidona	07	0.1839	07	0.1839
Iraklis	08	0.1797	08	0.1797
Ionikos	09	0.1552	10	0.1705
Xanthi	10	0.1416	11	0.1715
OFI	11	0.1409	12	0.1595
Kallithea	12	0.0749	09	0.1725
Aris	13	0.1620	13	0.1733
Akratitos	14	0.1635	16	0.2852
Panileiakos	15	0.2018	14	0.1781
Proodeftiki	16	0.2694	15	0.2178

Table 4.2.7: Probability of observed and expected ranks after using all the data

It would also be very interesting to see the probability for each team to get a specific position in the League Table. Column 3 of Table 4.2.7 gives the probabilities of getting the actual final positions. On the other hand, Column 5 of Table 4.2.7 gives the probabilities of getting the positions that the model predicted. For once more, we verify our former results. Panathinaikos should have finished in the 2nd place instead of the 1st (47.88%>38.29%). Olympiakos deserved to win the championship instead of ending up behind Panathinaikos (58.06%>34.08%). The team with the right to take part in the preliminary round of the next season's Champions League should have been AEK (36.18%>25.16%) and not PAOK (28.02%>25.91%). Kallithea had much bigger probability of finishing in the 9th place instead of the 12th (17.25%>7.49%) and Akratitos had greater chance to finish last in the League than 14th (28.52%>16.35%).

One of the big disadvantages of MLE is that it is completely model dependent. This means that there is strong dependence on the goals scored and the choice for ranking teams is basically made via attacking performances. Since we suspect that we should not have much faith in MLE, we changed by purpose <u>only 1</u> match event to see what would happen. We changed the easy win of Olympiakos against Akratitos; instead of the large score 7-0 we let Olympiakos win Akratitos only 1-0. Olympiakos would take again 3 points from the win and Akratitos none. Note that now the marginal differences (see Column 4 of Table 4.2.1) would be +45 for Olympiakos and +44 for Panathinaikos.

Teams	**Probability of Winning The Champion**	**Expected Ranking**	**Expected Points**
Panathinaikos	0.4730	2	67.736
Olympiakos	0.4772	1	67.739
Other team	0.0498		

Table 4.2.8: Results from MLE after changing only 1 observation

The results from Table 4.2.8 are impressive! Olympiakos is still expected to end up in the 1st place, but now we cannot be so sure. The gap between the two teams' expected points narrowed from (69.861-67.540=) 2.321 points to (67.739-67.736=) 0.003 points! Furthermore, Olympiakos cannot any longer be seen as the "certain" champion, since the probability of winning the championship reduced from 58.30% to 47.72%, while it increased for Panathinaikos from 38.03% to 47.72%, respectively.

Our model could predict what could happen in a specific match day, since we know and use all the former information. Table 4.2.9 shows a typical example. We use all the available data from the 1st round and we try to predict what will happen in the following day, which is the 1st day of the 2nd round. Of course, this can be done for every match day. If we use the biggest probabilities as the criterion to predict the outcome, then our model is still correct something less than 50%. Of course, we should not rely on these probabilities, because this is not the right way to decide. For example, the probability of Iraklis winning OFI is the biggest in their game and takes the value 39%. But the probability of Iraklis <u>not</u> winning OFI is 32%+29%=61%>39%. So, it is not very safe to predict based on the biggest probability. After all, most models have the disadvantage of not being able to predict a draw. Maybe the most predictable draw here could be the match between OFI and Iraklis, since all three probabilities are close each other.

Matches and observed scores	**Predicted Scores**	**P(*i* wins)**	**P(draw)**	**P(*j* wins)**	**Observed Outcome**
Kallithea: 0 Ionikos: 1	1.60 1.01	**0.51**	0.25	0.24	2
OFI: 2 Iraklis: 2	1.07 1.20	0.32	0.29	**0.39**	X
AEK: 2 Aigaleo: 2	1.47 0.83	**0.52**	0.27	0.21	X
Aris: 3 Akratitos: 0	1.67 1.09	**0.51**	0.24	0.25	**1**
Chalkidona: 2 Proodeftiki: 1	2.21 0.73	**0.71**	0.18	0.11	**1**
Olympiakos: 1 PAOK: 2	1.86 0.64	**0.67**	0.21	0.12	2
Panileiakos: 0 Panionios: 0	0.84 1.51	0.21	0.26	**0.53**	X
Xanthi: 0 Panathinaikos: 1	0.53 1.72	0.11	0.23	**0.66**	**2**

Table 4.2.9: Observed and predicted values for the 16th day

Finally, Figures 4.2.3 and 4.2.4 show how the 16 teams' coefficients change through the time and more precisely through the 2nd round after we have used the data only from the 1st round. The data were updated from match to match. These two figures give justify to our suspicions about the change in the teams' performances in the 2nd round. For example, Figure 4.2.3 show clearly the gradual improvement in Panathinaikos' offensive ability, while Olympiakos' attack generally improved, but this was not a stable improvement. Some kind of crisis in the attack of Olympiakos appeared around 21st and 24th match days. At the bottom of the League Table and for the 2nd round, Panileiakos and Proodeftiki had big problems with their attacks, but Akratitos; players could find the nets more easily. On the other hand, in Figure 4.2.4 we can see that Panathinaikos had a very reliable defence, Olympiakos was not again stable in his defensive performance this time and Aris defending performance was probably the reason, which kept the team away from relegation (recall that teams with good defences have small values of d_i).

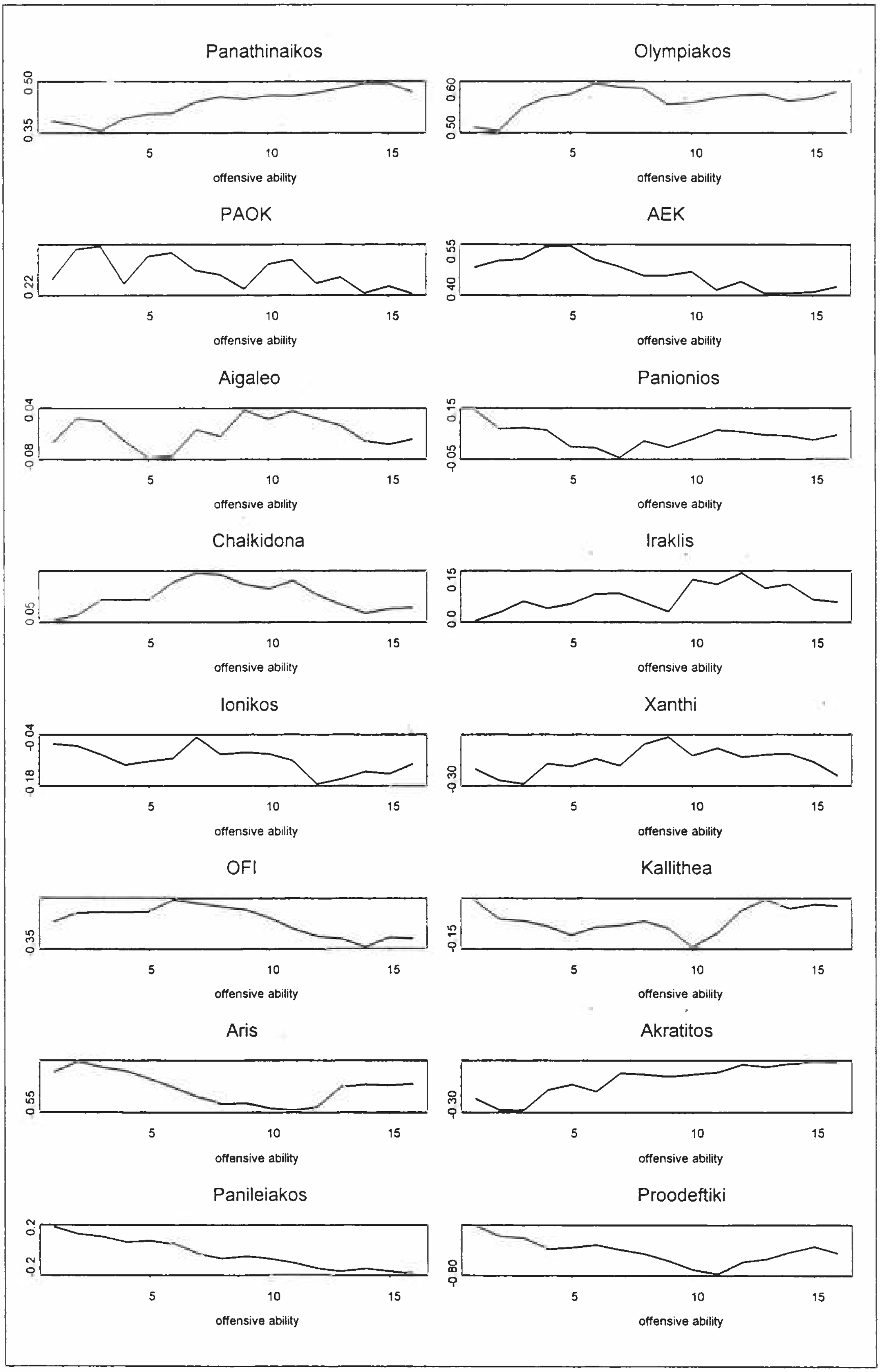


Figure 4.2.3: The offensive coefficients of all teams through the 2nd round

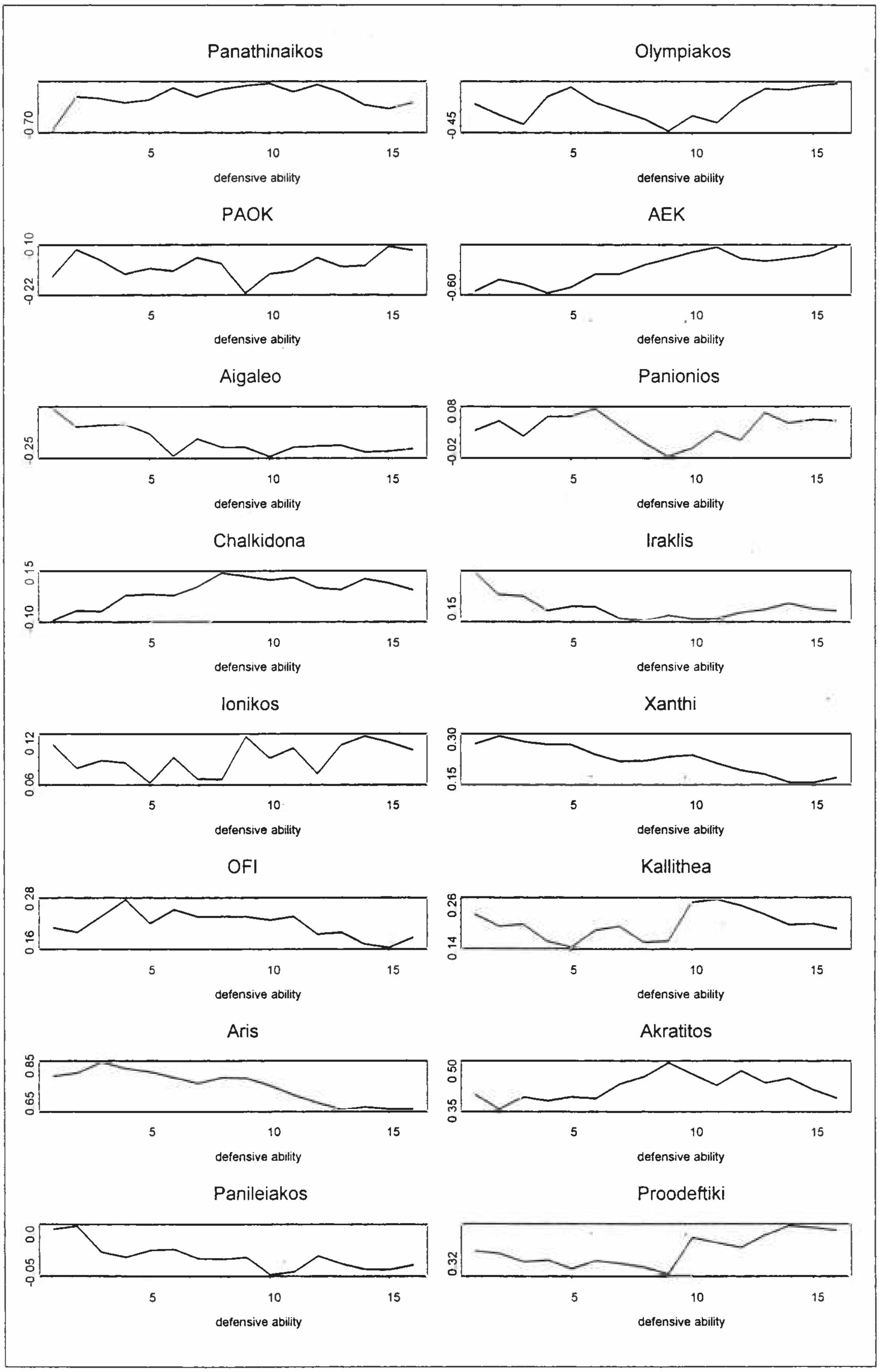


Figure 4.2.4: The defensive coefficients of all teams through the 2nd round

4.2.2 Estimating the parameters using WMLE

Just before we present two relevant methods, we shall make some important remarks about the weights that we shall use. Basu and Lindsay (1993b)'s IRLS is the nucleus of our application and the estimation for each parameter is given by:

$$\beta_{i+1} = \left[X^T W_{\beta_i} X\right]^{-1} X^T W_{\beta_i} Y .$$

The criterion for the algorithm to stop comes through the relationship:

$$\max\left|\beta_{i+1} - \beta_i\right| \leq 10^{-6} .$$

Function	Number Of Iterations	Estimated Value Of λ
MLE	2	1.310
MHD	21	1.291
NEYMAN	3100	1.128

Table 4.2.10: Estimated λ after using RAFs

According to Lindsay (1994), the estimating equations for λ have the form $\sum A(\delta(x))\nabla m_\beta(x) = 0$, where $A(\delta)$ is the RAF. The first thing to be solved is which RAF we should use. As we have said before, MHD has very good robust properties for outliers (while the NE for example is the best for inliers), so this could be a natural choice. Through the IRLS we compared the weights defined by the RAFs using MHD and Neyman's function. Our data set consisted of all the goals achieved in the league and we assumed that these were generated from a $Poisson(\lambda)$ without using any covariates at this time. The aim was to find which RAF could estimate λ faster and more precisely. We gave a starting value of 2 for λ in the algorithm and convergence was achieved, when the difference between two consecutive steps was less than or equal to 10^{-6}. Table 4.2.10 indicates that the MHD is more appropriate to work in the following, since it gives an estimation close to the MLE (which obviously converges in exactly 2 steps) and in a much more smaller number of iterations than Neyman's RAF. From the same table, we also make the assumption that the goals scored for the year 2003-2004 in the GNA came from a Poisson distribution with estimating parameter close to 1.3 or from a (possibly small) deviation of it.

At this point, we should make some remarks about the weights. These decline smoothly as the residuals depart from 0 to -1 or $+\infty$ and take the maximal value of 1 when the residuals are 0. An observation that is consistent with the assumed model receives a weight of approximately 1. On the other hand, a weight of approximately 0 indicates that the observation is highly inconsistent with the model. The final weights indicate which of the data points were downweighted in the final solution relative to the MLE. The corresponding to each observation weight depends on:

(i) The score itself

(ii) If the score belongs to a home or away team

(iii) The team itself

Team i	Team j	Score	**Weights**			
			Method I		**Method II**	
			w_{i_1}	w_{j_1}	w_{i_2}	w_{j_2}
Aigaleo	AEK	0-1	0.799	0.845	0.773	1.000
Akratitos	Aris	1-0	0.835	0.888	0.998	0.805
Chalkidona	Panileiakos	1-3	0.771	0.199	0.960	0.601
OFI	Panathinaikos	1-3	0.743	0.635	0.987	0.980
Olympiakos	Akratitos	7-0	0.515	0.964	0.977	0.953
Proodeftiki	Olympiakos	1-0	0.795	0.454	0.995	0.432
Olympiakos	Panathinaikos	1-1	0.845	0.838	1.000	1.000
Panathinaikos	Olympiakos	2-2	0.712	0.564	0.992	0.929

Table 4.2.11: Some weights for specific games

Consider now the model with covariates. Table 4.2.11 provides some examples at games and the associated weights for the observations. About (i), both matches between Olympiakos and Panathinaikos ended in a draw but with different scores. We see that all four weights are different, which means that scores themselves affect the weights. Regarding (ii), the first two games have the same scores but in "different" playing grounds. This means that the home ground is an important factor for determining the weights. For example, it was not so unexpected for Aigaleo to score a goal in home against AEK as it was for Aris to score a goal in Akratitos' home field. As for (iii), both Panathinaikos and Panileiakos succeeded to take an away win of 1-3, but the corresponding weights are totally different. The two games have the same scores in "similar" playing grounds, but they belong to different teams This happens, because Panileiakos is almost for sure unexpected to score 3 goals against

Chalkidona, but Panathinaikos had better chances of doing that against OFI. We also give the corresponding weights for Olympiakos' two "unexpected" results in home and away fields, which is assort of combination of the three above cases. According to Method I, Olympiakos should not have scored so many goals against Akratitos (despite the facts that Olympiakos had the better attack in the league and Akratitos was finally relegated). Similar to this, both methods indicate that Olympiakos should have scored one or more goals against Proodeftiki. It is very interesting to take a first look in the differences that appear in the weights for the two methods. These two kinds of weights depend on the residuals, which are defined right in the following.

Method I

One fundamental drawback in Lindsay (1994)'s approach for our application is that we cannot use the Pearson's residuals:

$$\delta(t)=\frac{d(t)}{m_\beta(t)}-1.$$

The problem arises at the numerator $d(t)$, which is the *proportion* of observations in the sample with value equal to t. We cannot find such frequencies, because each team has different o_i and d_j and for every combination of these covariates used in the model, there is only 1 observation. It is clear that such residuals are not suitable for every situation and mainly in the continuous case. Such problems are referred in the literature and many authors propose several methods and solutions. In order to surmount this obstacle, Basu and Lindsay (1994), formulated their thought of estimating $d(t)$ via a *kernel density estimator*, which could treat continuous data. This idea was widely used by many researchers, as Agostinelly and Markatou (1998), Markatou *et al.* (1998) and Markatou (2000).

Let us recall that the *kernel density estimator* has the general form:

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right),$$

where h is a smoothing parameter (called the *bandwidth*) and $K(\cdot)$ is a *kernel* function (the appropriate one here is the Poisson kernel). Since in our case it holds that $n=1$, $h=1$ and each point x represents the goals y_{ij}, one possible form of the residuals (after estimating the numerator by the number 1) could be:

$$\delta_\kappa = \frac{1}{P\left(y_\kappa \middle| \lambda_\kappa\right)} - 1,\ k = 1,2,\ldots,480.$$

The numerator implies that there exists exactly 1 observation from the specific distribution (in the denominator). In other words, we observe this specific observed score y_k, given that this number comes from a $Poisson(\lambda_k)$.

In Table 4.2.12 and Figure 4.2.5 we can briefly see how the residuals and the weights are distributed. There are indeed some outliers but most of the data points are (as it is logically expected) consistent with the assumed model. One score is totally unexpected (the 3 goals scored from Panileiakos against Panathinaikos) and three or four more observations have obviously smaller weights than the others. Table 4.2.13 presents these games.

	Min	**Max**	**Median**	**Mean**
Residuals 1	0.1959	1177.417	2.309	8.23
Weights 1	0.05741	0.9927	0.7973	0.7615

Table 4.2.12: Summary of residuals δ_k and weights (Method I)

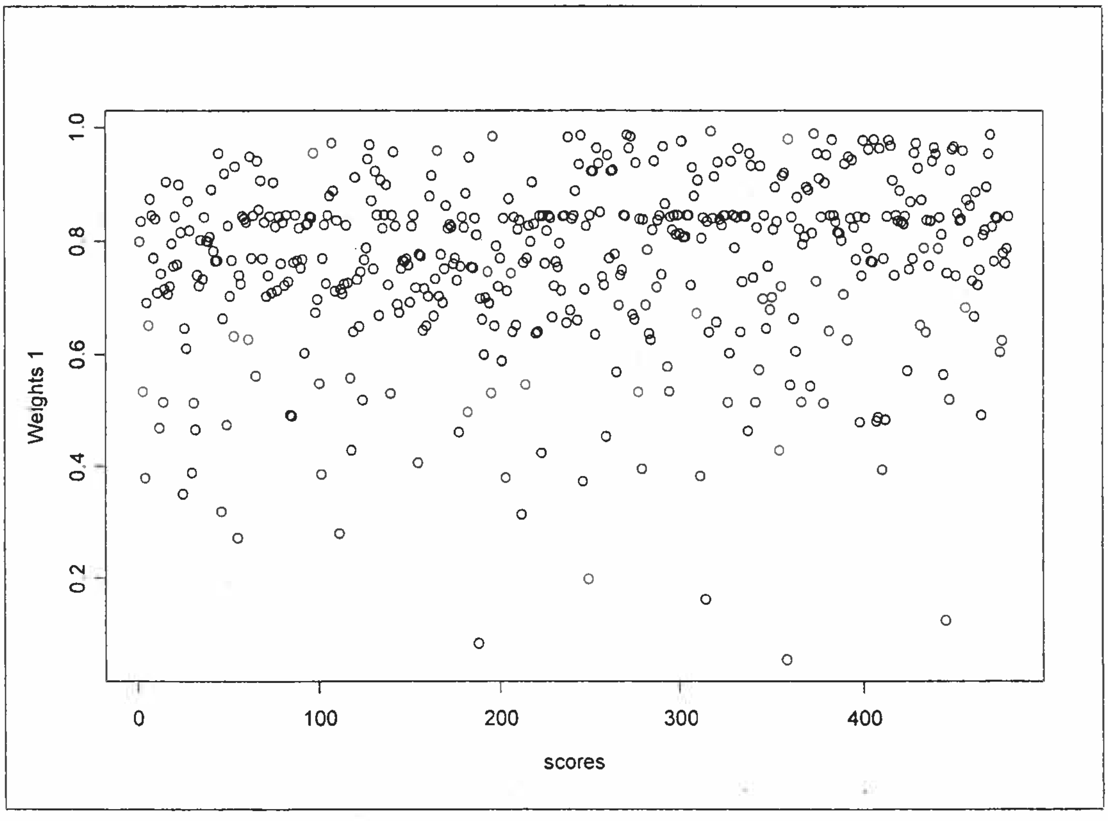


Figure 4.2.5: Weights' graphical representation (Method I)

Matches with the most unexpected results	Predicted Scores	Observed Scores	Residuals	Weights
Panathinaikos	3.18	6	15.746	0.429
Panileiakos	0.18	3	1177.417	**0.057**
Iraklis	1.31	6	519.872	**0.086**
Aris	0.48	0	0.609	0.955
Panileiakos	0.88	0	1.401	0.874
Kallithea	1.09	5	233.661	**0.126**
Akratitos	1.17	2	3.721	0.709
AEK	2.34	7	136.327	**0.163**
Chalkidona	1.99	1	2.685	0.771
Panileiakos	0.47	3	89.748	**0.199**

Table 4.2.13: Residuals and weights of the most outlying observations (Method I)

Teams	Using ALL the data		Data only from the 1st round	
	Offensive (o_i)	Defensive (d_i)	Offensive (o_i)	Defensive (d_i)
Intercept	*-0.173*		*-0.102*	
Home effect	*0.414*		*0.360*	
Panathinaikos	0.561	-0.889	0.459	-0.843
Olympiakos	0.709	-0.632	0.624	-0.752
PAOK	0.266	-0.332	0.287	-0.381
AEK	0.475	-0.057	0.503	-0.133
Aigaleo	-0.006	-0.326	-0.011	-0.583
Panionios	0.064	-0.235	0.135	-0.044
Chalkidona	0.096	0.062	0.043	0.027
Iraklis	-0.009	0.054	-0.007	-0.041
Ionikos	-0.132	0.206	-0.144	0.314
Xanthi	-0.267	0.131	-0.248	0.125
OFI	-0.355	0.182	-0.244	0.292
Kallithea	-0.052	0.214	0.024	0.244
Aris	-0.541	0.105	-0.490	0.223
Akratitos	-0.035	0.645	-0.211	0.781
Panileiakos	-0.454	0.408	0.013	0.450
Proodeftiki	-0.321	0.463	-0.731	0.326

Table 4.2.14: Model details (Method I)

The corresponding to MLE Table 4.2.2 for the Method I of the WMLE takes the form of Table 4.2.14. There are some obvious differences. The first one is that according to the model parameters, Olympiakos had indeed the best attack, but Panathinaikos had clearly the best defence of the League. On the other hand, during the two rounds, the three bottom teams had the worst defences, while there is some kind of deviance regarding the worst attacks; at the end of the 1st round the worst attacks came from Proodeftiki, Aris and Xanthi (or OFI), while at the end of the season Aris, Panileiakos and OFI had the worst attack. Finally, the value of the constant moves away from zero and the home effect is slightly bigger.

We simulated 10000 leagues using the above estimates and the predicted results are shown in Table 4.2.15. In this case, the model came up to the striking correct prediction of 124 out of 240 or 51.67% of the total games played!

Teams		Model using all the data		Model using data only from the 1^{st} round	
	Observed Points	Predicted Ranking	Predicted Points	Predicted Ranking	Predicted Points
01. Panathinaikos	77	**02**	**70.141**	**02**	**70.550**
02. Olympiakos	75	**01**	**70.761**	**01**	**70.906**
03. PAOK	60	**04**	**55.248**	**04**	**54.741**
04. AEK	55	**03**	**55.485**	**03**	**55.531**
05. Aigaleo	52	**06**	**47.380**	**06**	**47.622**
06. Panionios	47	**05**	**48.123**	**05**	**47.942**
07. Chalkidona	45	07	42.851	07	42.574
08. Iraklis	42	08	40.310	08	39.980
09. Ionikos	33	10	34.083	10	33.900
10. Xanthi	30	12	32.571	11	32.804
11. OFI	29	11	29.696	12	29.989
12. Kallithea	27	**09**	**35.545**	**09**	**35.376**
13. Aris	27	13	27.398	13	27.748
14. Akratitos	23	14	25.737	14	25.584
15. Panileiakos	21	**16**	**22.918**	**16**	**23.310**
16. Proodeftiki	20	**15**	**24.207**	**15**	**24.612**

Table 4.2.15: Expected ranking and points (Method I)

The results are similar to those of Table 4.2.3, but there are again some differences. The only and probably imperceptible disadvantage in Table 4.2.15 is that the confusion from the 1st round for places 5-6 there still exists at the end of the season, while this is solved in Table 4.2.3. We could say that some kind of improvement appears in the first two places. Olympiakos keeps the higher expected value of points than Panathinaikos and Olympiakos is the best team but with average difference of only 0.62 points. Similar results hold for AEK and PAOK. However, the big advantage is that the 3rd worst team (Akratitos) takes the place that it "deserves". Thus, Ergotelis should face Akratitos for a place in GNA (as it really happened). Panileiakos and Proodeftiki are the two worst teams in the league, but the difference between them is now more visible. The rest are very alike (including Kallithea).

ALL THE DATA				
Teams	**Probability of Winning The Champion**	**Probability of Relegation (1)**	**Probability of Relegation (2)**	**Probability of Relegation (3)**
Panathinaikos	0.4729			
Olympiakos	**0.5132**			
Other team	0.0139			
Kallithea		0.0108	0.0353	0.0812
Aris		0.1054	0.2108	0.3772
Akratitos		0.1303	0.2906	**0.4485**
Panileiakos		**0.3892**	**0.6135**	**0.7578**
Proodeftiki		0.2305	**0.4490**	**0.6260**

Table 4.2.16: Probability of ranks after using all the data (Method I)

Table 4.2.16 gives the distribution of ranks for the model after using all the data for the two edges of the League Table. According to this, Olympiakos should have been the champion, since it had higher probability to win the league than Panathinaikos, whose probability is increased in comparison to Table 4.2.5. There is also a very small chance for some other team to win the league. From columns 3-5, Panileiakos appears as the worst team, which agrees with Table 4.2.15. Proodeftiki is the second weakest team, while Akratitos fills the triad of the weakest links of GNA. Also, Table 4.2.17 gives the distribution of ranks for the model after using the data only from the 1st round. Column 2 shows for once more, that Panathinaikos improved its performance in the 2nd round.

DATA ONLY FROM THE 1st ROUND				
Teams	**Probability of Winning The Champion**	**Probability of Relegation (1)**	**Probability of Relegation (2)**	**Probability of Relegation (3)**
Panathinaikos	0.4786			
Olympiakos	**0.5071**			
Other team	0.0143			
Kallithea		0.0099	0.0372	0.0798
Aris		0.1151	0.2615	0.4271
Akratitos		0.1223	0.2824	**0.4497**
Panileiakos		**0.3873**	**0.6100**	**0.7601**
Proodeftiki		0.2326	**0.4464**	**0.6170**

Table 4.2.17: Probability of ranks using the data only of the 1st round (Method I)

The updated probabilities of the three possible outcomes for the 16th day are shown in Table 4.2.18. As someone can see, there are not big differences. We could only mention that the two obvious expected wins of Chalkidona and Panathinaikos against Proodeftiki and Xanthi respectively are even more apparent and that the match between OFI and Iraklis becomes more in the balance.

Matches and observed scores	Predicted Scores	P(i wins)	P(draw)	P(j wins)	Observed Outcome
Kallithea: 0 Ionikos: 1	1.48 0.91	**0.51**	0.26	0.23	2
OFI: 2 Iraklis: 2	0.94 1.00	0.33	0.31	**0.36**	X
AEK: 2 Aigaleo: 2	1.47 0.79	**0.53**	0.27	0.20	X
Aris: 3 Akratitos: 0	1.41 0.90	**0.49**	0.27	0.24	**1**
Chalkidona: 2 Proodeftiki: 1	2.22 0.65	**0.74**	0.17	0.09	**1**
Olympiakos: 1 PAOK: 2	1.85 0.58	**0.68**	0.21	0.11	2
Panileiakos: 0 Panionios: 0	0.64 1.35	0.18	0.28	**0.54**	X
Xanthi: 0 Panathinaikos: 1	0.40 1.68	0.08	0.23	**0.69**	**2**

Table 4.2.18: Observed and predicted values for the 16th day (Method I)

Method II

In this second method, we propose an alternative estimator of the numerator. We think of how the kernel works in practice and we give an "extra" weight exactly *at* the observation, while we smooth this weight in an area around the observation. Hence, the residuals could be further updated as:

$$\delta_k = \frac{P(y_\kappa|y_\kappa)}{P(y_\kappa|\lambda_\kappa)} - 1,\ k = 1,2,\dots,480.$$

	Min	Max	Median	Mean
Residuals 2	0.0000002809	65.7683	0.3283	1.137
Weights 2	0.2297853	1.000	0.9825	0.9309

Table 4.2.19: Summary of residuals δ_k and weights (Method II)

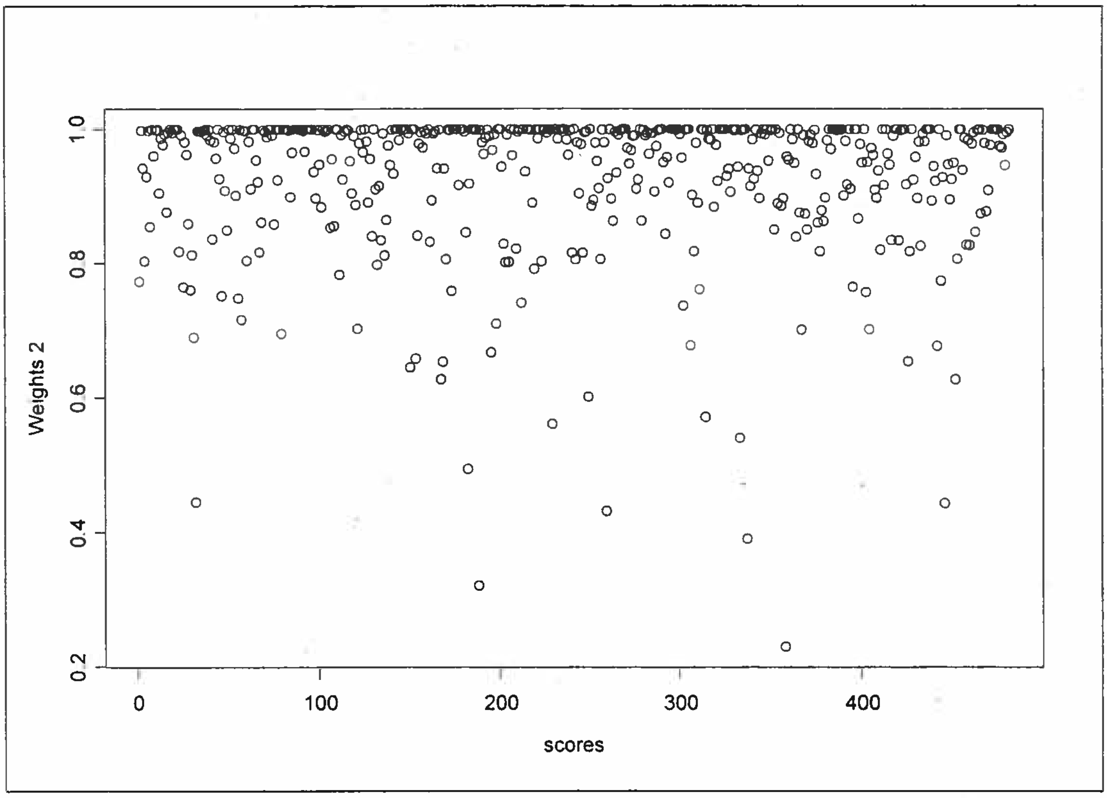


Figure 4.2.6: Weights' graphical representation (Method II)

Matches with the most unexpected results	Predicted Scores	Observed Scores	Residuals	Weights
Panathinaikos	3.49	6	1.098	0.904
Panileiakos	0.30	3	65.768	**0.230**
Iraklis	1.62	6	31.379	**0.321**
Aris	0.64	0	0.901	0.925
Akratitos	0.58	0	0.788	0.936
Panathinaikos	3.03	0	19.739	**0.391**
Proodeftiki	0.55	1	0.163	0.995
Olympiakos	2.80	0	15.498	**0.432**
Panileiakos	1.18	0	2.255	0.801
Kallithea	1.41	5	14.467	**0.444**
PAOK	2.73	0	14.327	**0.446**
Proodeftiki	0.51	0	0.672	0.949
Olympiakos	2.48	0	10.944	**0.494**
AEK	0.81	1	0.021	1.000

Table 4.2.20: Residuals and weights of the most outlying observations (Method II)

In Table 4.2.19 and Figure 4.2.6 we can briefly see how the residuals and the weights are distributed. All the descriptive statistics for the weights take larger values than in Method I and furthermore they can take the maximum admissible value of 1.

There are still some outliers and the bulk of the data points are (as it is logically expected) consistent with the assumed model. For once more, the 3 goals scored from Panileiakos against Panathinaikos is a totally unexpected number and there is a group of probably six more observations with obviously larger residuals than the others. In Table 4.2.20 the reader can see these games and the differences with Table 4.2.11 are not only in the values of the residuals and the weights but also in the games themselves.

Table 4.2.21 shows the updated model parameters. The results are similar to those of Table 4.2.15 and we can just notice that the parameters of the top teams seem to be slightly downweighted towards 0. In Table 4.2.22 there the predicted results after the simulation of 10000 leagues using the above The model predicted correctly 120 out of 240 or 50% of the total games played. The confusion for places 1-2, 3-4 and 5-6 appears again and from the data of the 1st round the three worst teams are almost equivalent.

Teams	**Using ALL the data**			**Data only from the 1st round**		
		Offensive (o_i)	**Defensive** (d_i)		**Offensive** (o_i)	**Defensive** (d_i)
Intercept	*-0.034*			*0.017*		
Home effect	*0.372*			*0.331*		
Panathinaikos		0.493	-0.815		0.442	-0.731
Olympiakos		0.635	-0.613		0.561	-0.716
PAOK		0.236	-0.301		0.267	-0.350
AEK		0.434	-0.064		0.480	-0.152
Aigaleo		-0.017	-0.331		-0.013	-0.578
Panionios		0.078	-0.207		0.156	-0.031
Chalkidona		0.059	0.059		0.018	0.034
Iraklis		0.020	0.054		-0.017	-0.080
Ionikos		-0.094	0.160		-0.059	0.252
Xanthi		-0.276	0.107		-0.263	0.114
OFI		-0.317	0.150		-0.230	0.274
Kallithea		-0.040	0.179		-0.019	0.186
Aris		-0.462	0.123		-0.392	0.226
Akratitos		-0.065	0.650		-0.232	0.789
Panileiakos		-0.352	0.418		0.062	0.455
Proodeftiki		-0.331	0.430		-0.760	0.308

Table 4.2.21: Model details (Method II)

Teams		Model using all the data		Model using data only from the 1st round	
	Observed Points	Predicted Ranking	Predicted Points	Predicted Ranking	Predicted Points
01. Panathinaikos	77	**02**	**69.698**	**02**	**69.790**
02. Olympiakos	75	**01**	**70.522**	**01**	**70.536**
03. PAOK	60	**04**	**54.792**	**04**	**54.458**
04. AEK	55	**03**	**55.460**	**03**	**55.403**
05. Aigaleo	52	**06**	**48.192**	**06**	**48.366**
06. Panionios	47	**05**	**48.933**	**05**	**48.400**
07. Chalkidona	45	07	41.702	07	41.950
08. Iraklis	42	08	41.348	08	41.388
09. Ionikos	33	10	35.906	10	35.620
10. Xanthi	30	11	33.103	11	32.468
11. OFI	29	12	31.066	12	30.646
12. Kallithea	27	**09**	**36.278**	**09**	**36.622**
13. Aris	27	13	28.374	13	28.548
14. Akratitos	23	14	23.987	**15**	**24.126**
15. Panileiakos	21	**16**	**23.590**	**16**	**24.113**
16. Proodeftiki	20	**15**	**23.744**	**14**	**24.223**

Table 4.2.22: Expected ranking and points (Method II)

ALL THE DATA				
Teams	Probability of Winning The Champion	Probability of Relegation (1)	Probability of Relegation (2)	Probability of Relegation (3)
Panathinaikos	0.4502			
Olympiakos	**0.5261**			
Other team	0.0237			
Kallithea		0.0084	0.0261	0.0656
Aris		0.0879	0.2188	0.3752
Akratitos		0.2693	0.4989	**0.6796**
Panileiakos		**0.2781**	**0.5065**	**0.6877**
Proodeftiki		0.2744	**0.5027**	**0.6829**

Table 4.2.23: Probability of ranks after using all the data (Method II)

For Method II, Table 4.2.23 gives the distribution of ranks for the model after using all the data for the two edges of the League Table. It agrees with the belief that Olympiakos should have been the champion and that Panileiakos was the worst team, Proodeftiki the second weakest team and Akratitos the third weakest team of GNA. However, the "weird" thing is that the data only from the 1st round (Table 4.2.24) indicate Akratitos as the second worst team of GNA. This is something that we had seen in the MLE. Panileiakos was the worst team and Proodeftiki the third weakest team of GNA.

DATA ONLY FROM THE 1st ROUND				
Teams	**Probability of Winning The Champion**	**Probability of Relegation (1)**	**Probability of Relegation (2)**	**Probability of Relegation (3)**
Panathinaikos	0.4438			
Olympiakos	**0.5348**			
Other team	0.0214			
Kallithea		0.0075	0.0291	0.0637
Aris		0.0895	0.2149	0.3730
Akratitos		0.2705	**0.5008**	**0.6850**
Panileiakos		**0.2796**	**0.5094**	**0.6896**
Proodeftiki		0.2668	0.4912	**0.6729**

Table 4.2.24: Probability of ranks using the data only of the 1st round (Method II)

The new updated probabilities of the three possible outcomes for the 16th day are shown in Table 4.2.25. The differences for once more are small.

Matches and observed scores	**Predicted Scores**	**P(i wins)**	**P(draw)**	**P(j wins)**	**Observed Outcome**
Kallithea: 0 Ionikos: 1	1.58 1.05	**0.50**	0.25	0.25	2
OFI: 2 Iraklis: 2	1.08 1.15	0.34	0.29	**0.37**	X
AEK: 2 Aigaleo: 2	1.56 0.89	**0.53**	0.26	0.21	X
Aris: 3 Akratitos: 0	1.69 1.02	**0.53**	0.24	0.23	**1**
Chalkidona: 2 Proodeftiki: 1	2.29 0.74	**0.73**	0.17	0.10	**1**
Olympiakos: 1 PAOK: 2	1.96 0.66	**0.68**	0.20	0.12	2
Panileiakos: 0 Panionios: 0	0.80 1.59	0.19	0.25	**0.56**	X
Xanthi: 0 Panathinaikos: 1	0.47 1.76	0.09	0.22	**0.69**	**2**

Table 4.2.25: Observed and predicted values for the 16th day (Method II)

Method I Vs Method II

First of all, the two methods are very similar and produce alike results. If we compare at start Figure 4.2.5 with Figure 4.2.6, we can see that both methods find easily the two most outlying observations. But after that, Method I identifies maybe four more obvious outliers, while Method II shows a group of at least six (or even more than ten) data points with large residuals. From these figures, we can see that the bulk of the data points are (as it is logically expected) consistent with the assumed models. From Tables 4.2.18 and 4.2.25 the two methods produce almost identical probabilities for a typical match day, which agree a lot with those given by the bookmakers (see Table 4.2.28).

League Table		**Model using all the data** **Predicted (Ranking) & Points**					
Observed Ranking	**Observed Points**	**MLE**		**WMLE**			
				Method I		**Method II**	
(01) Panathinaikos	77	**(02)**	**67.540**	**(02)**	**70.141**	**(02)**	**69.698**
(02) Olympiakos	75	**(01)**	**69.861**	**(01)**	**70.761**	**(01)**	**70.522**
(03) PAOK	60	**(04)**	**53.702**	**(04)**	**55.248**	**(04)**	**54.792**
(04) AEK	55	**(03)**	**55.973**	**(03)**	**55.485**	**(03)**	**55.460**
(05) Aigaleo	52	(05)	48.428	**(06)**	**47.380**	**(06)**	**48.192**
(06) Panionios	47	(06)	48.320	**(05)**	**48.123**	**(05)**	**48.933**
(07) Chalkidona	45	(07)	42.387	(07)	42.851	(07)	41.702
(08) Iraklis	42	(08)	42.199	(08)	40.310	(08)	41.348
(09) Ionikos	33	(10)	35.138	(10)	34.083	(10)	35.906
(10) Xanthi	30	(11)	32.479	(12)	32.571	(11)	33.103
(11) OFI	29	(12)	30.139	(11)	29.696	(12)	31.066
(12) Kallithea	27	**(09)**	**37.654**	**(09)**	**35.545**	**(09)**	**36.278**
(13) Aris	27	(13)	27.143	(13)	27.398	(13)	28.374
(14) Akratitos	23	**(16)**	**24.259**	(14)	25.737	(14)	23.987
(15) Panileiakos	21	**(14)**	**25.706**	**(16)**	**22.918**	**(16)**	**23.590**
(16) Proodeftiki	20	**(15)**	**24.278**	**(15)**	**24.207**	**(15)**	**23.744**

Table 4.2.26: Comparison of MLE and WMLE (including both methods)

However, there are some few and possibly critical differences, which make us believe that Method I is slightly better than Method II. The most obvious one is the total number or the percentage of the correct predicted games. Also, Akratitos in Table 4.2.15 goes safely away from the last two places, but in Table 4.2.22 this distance is not so big and the data of the 1st round imply that Akratitos should be in the 15th place. From the same tables, Olympiakos is better team than Panathinaikos with average difference of 0.62 points in Method I and 0.824 points in Method II. We should not forget though, that Panathinaikos was the true champion. This means that we would like to find that Panathinaikos is the best team; if not, it would be desirable the difference from Olympiakos to be as small as possible. In the same logic for the two major challengers of the title, the probabilities in Table 4.2.13 are more acceptable than those of Table 4.2.23. A last disadvantage of Method II is that the data of the 1st round indicate Akratitos as the worst team (see Table 4.2.24), while Panileiakos is steadily the worst team for Method I (see Tables 4.2.16 and 4.2.17).

All the above are briefly summarised in Table 4.2.26. For once more it is sown that the model performs well in the middle of the League Table, but there are deviations at the two edges of it. According to the model, the top teams in GNA seem to have more points in reality that they should really get, while the opposite happens for the worst teams (see Figure 4.2.1). Also, the graphical representation of the residuals of the model (i.e. the absolute values of the observed points minus the expected points) shows an obvious U-shape (see also Figure 4.2.2). This is a clear indication that there must exist a deviation of the model we assumed and that is why we used robustness methods.

4.2.3 Challenging the bookmakers

We must mention that our model did not come up for betting purposes, but as it appears, the bookmakers use similar models. Many bookmakers in the website gave quite the same estimated probabilities to ours. Everyone who deals with the betting market sees the inconsistency of soccer bets, since outcomes with small probabilities give small returns. The challenge is to find "good" bets, in which the bettor suspects that the probability of occurrence of an outcome is bigger than the probability determined by the bookmaker's odds, giving a positive expected return. According to many persons' belief, the realistic case is the one in which someone should try not to gain as more money as he can, but to limit his loss. Only few, highly skilled and selective bettors, who can reach a level of 55% or maximum 60% accuracy against the bookmakers, achieve a successful strategy for a possible small or no profit.

Day 16th	**Bookmakers' odds**								
Matches	Bookmaker **GB**			Bookmaker **IW**			Bookmaker **SB**		
	H	**D**	**A**	**H**	**D**	**A**	**H**	**D**	**A**
Kallithea Vs Ionikos	2.00	3.00	3.35	1.90	3.00	3.30	2.05	3.10	3.20
OFI Vs Iraklis	2.30	3.05	2.75	2.30	3.00	2.60	2.25	3.20	2.75
AEK Vs Aigaleo	1.50	3.45	5.75	1.55	3.30	4.80	1.44	3.75	6.00
Aris Vs Akratitos	1.50	3.45	5.75	1.45	3.40	5.40	1.40	3.75	6.50
Chalkidona Vs Proodeftiki	1.60	3.30	5.00	1.50	3.40	5.00	1.53	3.50	5.25
Olympiakos Vs PAOK	1.40	3.75	7.00	1.40	3.60	6.00	1.44	3.60	6.50
Panileiakos Vs Panionios	2.50	3.00	2.50	2.45	2.90	2.45	2.50	3.10	2.50
Xanthi Vs Panathinaikos	5.50	3.35	1.55	5.40	3.50	1.45	5.50	3.40	1.53

Table 4.2.27: Odds given by different bookmakers in the website for the 16th day

Finally, we quote the predictions given by three different known bookmakers in the website for the 16th day. Table 4.2.28 shows what everyone can see in a usual betting coupon. GB stands for the URL address **http://www.globet.com**, IW for **http://www.interwetten.com** and SB for **http://www.sportingbet.com**. Someone could use such odds and then construct a coupon of his own with probabilities. We linked the two coupons through the following relationships: Let O_H, O_D and O_A be the three odds for a home-win, a draw and an away-win respectively, then the corresponding probabilities are given by:

$$P_H = \frac{1}{1+\frac{O_H}{O_D}+\frac{O_H}{O_A}},\ P_D = \frac{1}{1+\frac{O_D}{O_H}+\frac{O_D}{O_A}} \text{ and } P_A = \frac{1}{1+\frac{O_A}{O_D}+\frac{O_A}{O_H}},$$

under the assumption, of course, that the game is fair.

Day 16th	**Bookmakers' probabilities**								
Matches	Bookmaker **GB**			Bookmaker **IW**			Bookmaker **SB**		
	H	**D**	**A**	**H**	**D**	**A**	**H**	**D**	**A**
Kallithea Vs Ionikos	0.44	0.30	0.26	0.45	0.29	0.26	0.43	0.29	0.28
OFI Vs Iraklis	0.39	0.29	0.32	0.38	0.29	0.33	0.40	0.28	0.32
AEK Vs Aigaleo	0.59	0.26	0.15	0.56	0.26	0.18	0.61	0.24	0.15
Aris Vs Akratitos	0.59	0.26	0.15	0.59	0.25	0.16	0.63	0.23	0.14
Chalkidona Vs Proodeftiki	0.55	0.27	0.18	0.55	0.27	0.18	0.58	0.25	0.17
Olympiakos Vs PAOK	0.63	0.24	0.13	0.62	0.24	0.14	0.62	0.25	0.13
Panileiakos Vs Panionios	0.353	0.294	0.353	0.352	0.296	0.352	0.355	0.28	0.355
Xanthi Vs Panathinaikos	0.16	0.27	0.57	0.16	0.25	0.59	0.16	0.26	0.58

Table 4.2.28: Corresponding probabilities for the bookmakers' odds for the 16th day

From the odds of Table 4.2.27 we are driven to Table 4.2.28, which gives the probabilities after using the three above equalities. These probabilities give a good proof of how well our model works or actually of how close it is to the bookmakers' belief. Against our model, all these bookmakers predicted that Panileiakos and Panionios seemed to be equivalent for the win of the game. However, for the other match in the balance between OFI and Iraklis, the bookmakers gave the biggest probability to the home team, while our model gave it to the hosted team. Hence, these could be some hints to suspect a possible draw, since this is the most difficult outcome of the three to predict, as we have said before.

In the end, we mention that we programmed several routines in order to extract numerical results, which can come up at the reader's disposal. We chose to work in the R "environment".

CHAPTER 5

Conclusions

Although at the beginning of its history the statistical community did not give much attention to soccer, later on the statisticians found this area, and sports generally, as a blooming field for applying statistical methodologies and developing methods for dealing with athletic data. Section 2 shows clearly that the number of publications scaled up, perhaps due to the growing popularity of soccer or even to the birth of powerful computers, which permit the calculation of extremely complicated models.

A first issue that could be debatable is which distribution we should use, because different distributions could fit soccer data sets well. Moroney (1956) declared that the Poisson and even better an allied distribution, the Negative Binomial distribution (NBD), are the most appropriate to find the probability of winning a game. A major differentiation raised by Maher (1982), who observed that each match had a different fitted Poisson distribution. The choice of either the NBD or the Poisson distribution concerned Baxter and Stevenson (1988) and other statisticians, who concluded that after 1970 both distributions are adequate for soccer data. Hence, someone might prefer the Poisson distribution, because it appears less complicated.

Our approach assumes that the goals scored by each team are independent. This sounds paradoxical, since the two teams act and compete together. Practice has shown that usually the correlation is relatively small (and statistically not significant). Karlis and Ntzoufras (2003) used a bivariate Poisson distribution with its extensions and defined more general models in order to cope with excess of draws and correlation observed in certain championships. Their models could indeed predict more precisely the draws and allowed for better fit of soccer data, because they could handle both correlation and over-dispersion. Here, we do not extend to so complicated models and we restrict to the simple Poisson distribution.

In practice, it was found that the Poisson and the NBD were very close to each other. We prefer to fit the Poisson for our soccer data, since among other advantages this distribution has a formal theoretical basis and is naturally used for events that occur randomly at a constant rate to the observed time period. In addition, as shown by goodness-of-fit test based on Pearson's residuals, our model fitted well the data and the need for a more complicated model, as the NBD, was non-existent. Nonetheless, we have the suspicion that the underlying distribution is not exactly this one, but a (probably small) deviation of it. Using a robust estimation method, such small deviation can be handled successfully.

		Poisson			**NBD**			**Over-dispersion**
λ_i	λ_j	P_W	P_D	P_L	P_W	P_D	P_L	
1	0.109	0.593	0.367	0.040	0.593	0.367	0.040	1.01
1	0.605	0.441	0.342	0.217	0.443	0.341	0.216	1.05
1	1.201	0.305	0.289	0.406	0.311	0.289	0.400	1.10
1	1.345	0.278	0.275	0.447	0.288	0.275	0.437	1.15
1	2.406	0.140	0.176	0.684	0.153	0.182	0.665	1.20
1	1.250	0.296	0.420	0.284	0.311	0.405	0.284	1.25

Table 5.1: Poisson Vs NBD; probabilities of outcomes in a hypothetical match

In Table 5.1 we give an example, where the two distributions are compared. We suppose that team i plays against team j and the parameter λ of the Poisson distribution is supposed to be equal to 1 for team i, while the other parameter for team j takes several values. Firstly, we fitted the Poisson distribution and we found the three probabilities P_W, P_D and P_L for the three possible outcomes of the game, which are the win, the draw and the loss respectively for team i. Next, we fitted the NBD with the same parameters and allowing a small over-dispersion. We can clearly see, that the new probabilities are quite the same to those that came from the Poisson distribution. Some differences begin to appear, only when the over-dispersion becomes larger, but a value like 1.25 or more is not usually true for soccer data. Hence, although our data might be slightly over-dispersed relative to the Poisson assumption, the differences are small. For all the above reasons we chose to work with the Poisson distribution.

A second issue is the structure of the model. Since we deal with count data, we used a generalized linear model with the Poisson distribution underlying, of course. There seems to exist a "home effect", while the presence of the constant parameter and the interpretation of all the coefficients depend on the constraint. We take into account the different abilities of both teams in a match and we separate each team's ability into "offensive" and "defensive". The insuperable difficulty is that no model can fully embody several other aspects, such as psychological factors or the referee effect.

Under Douglas' (1994) thought and our suspicion, we apply robust methods using the Poisson distribution. The need for such an approach arose, since the theories of parametric models were only approximations to reality and it was not clarified, whereas certain assumptions were fulfilled. The two desirable features of robustness is that, it deals firstly with model deviation and secondly with data contamination. One solution is the "correction" of surprising observations by downweighting data points with large Pearson residuals. During 1964-1981 Huber's researches constituted the pattern in robust theory. Many years later, Lindsay's (1994) paper engaged in a very important function, called "residual adjustment function" (RAF) in order to find the key structural element that links ML and other distances and disparity measures such as the "minimum Hellinger" distance (MHD). His goal was to measure the robustness properties of MHD estimation. Many interesting and important articles followed. In our work we focused on the "Iterative Reweighting Least Squares" (IRLS) algorithm used by Basu and Lindsay (1993b). See also an improved algorithm in Basu and Lindsay's (2004) paper.

In this report we applied a log linear model with the inkling of a possible existence of outlying observations. We did not rely on MLE, which could give very misleading results. Indeed, we gave an illuminating example where after changing only 1 observation (by arbitrarily downweighting it), the results from MLE changed completely. Nevertheless, we presented several results (such as estimated parameters and expected ranking and total points) through MLE just to have a measure for comparison. Also, we showed in two figures how all the teams' coefficients were changing through the 2nd round of the champion, after having used the data only from the 1st round.

Hence, we were based on more robust methods by using WMLE. At first, with the contribution of several RAFs, we estimated the value of the parameter λ for the underlying Poisson distribution. We chose to work with MHD, but we had problem in using Pearson's residuals as Lindsay (1994) did. The basic drawback was the estimation of the numerator for the Pearson's residual. Several authors have mentioned relevant problems, which become clearer in the continuous case. There are several methods, like the kernel density estimators, which treat continuous data, but the scope of our work was not to resort to such methods. Instead, we proposed two different methods (which were grounded on IRLS and IREE) by just thinking of how the kernel density estimators work. Their only difference was in the definition of the residuals. A first sign that we were in the right way came from the fact, that the bulk of the data points were consistent with the assumed model(s). The two methods produced similar results and mush better than those of MLE, as expected. Both methods locate easily the most outlying observations. However, there are some few and possibly critical differences, which indicate that the first method is slightly better than the second one.

We shall now discuss some issues that can be studied more extensively and we suggest refinements for further improvement. For example, we assumed that the Poisson distribution is adequate for describing soccer data by exploiting the offensive and defensive abilities of the competing teams. The Poisson distribution is very simply to be applied via standard statistical software, while the NBD demands more intensive computations and special programming for estimating the model parameters. The choice of Poisson distribution against the NBD still concerns some statisticians. Up to 1970 there were strong grounds for preferring the NBD and thereafter the Poisson seemed adequate. For some (known or unknown) reasons this could change again. According to many fans of the sport, the gap between the strong and the weakest teams is narrowing. The year of 2004 is a fine example, since an outsider, Porto, won the Champions League (against Monaco) and the Greek National Football Team conquered the Euro Cup. The point is that a distributions which fitted the data well for the last 30 years or more, might not be appropriate for the future.

Another point for further research in our work is the definition of the residuals. The residuals from a robust fit show outliers and the proper random variability of the "good" data much clearer than from least squares, which tend to smear the effect of outliers over many data points and makes their detection more difficult. We proposed two methods, where we gave two estimations of the numerator for the Person's residuals used by Lindsay (1994), which could be refined. Someone could find different estimates of the numerator either by a thought like ours or by using similar estimators as for continuous data, such as kernel density estimators. Both ways appear quite challenging and they could probably improve our results.

We mention that our model did not come up for betting purposes, but as it appears, the bookmakers use similar models and their estimated probabilities for the three possible outcomes of soccer games are quite the same to ours. Everyone who is knowledgeable about betting market sees the inconsistency of soccer bets, since outcomes with small probabilities give small returns. The challenge is to find "good" bets, in which the bettor suspects that the probability of occurrence of an outcome is bigger than the probability determined by the bookmaker's odds, giving a positive expected return. Thus, we challenge the interested reader and anyone who deals with such issues to develop refined models, which could be very useful devices for this purpose.

Finally, we underline the importance and the more general use of WMLE. A lot can be said about robustness in GLMs and the treatment of other distributions except for Poisson. Here, we have just seen an application to soccer data, but the methodology stands for every occasion in which "strange" observations can distort the results by using MLE for example. It is well known that the MLE of the unknown parameters are explicitly model-dependent and as such are heavily criticised as being non-robust. The WMLE method is a fine alternative and someone could use the WMLE for more general data, such as in Economics, Biometrics, Psychology and many other areas of scientific research.

Appendix I

Web resources

This page will give to the interested reader some web links to soccer pages.

Soccer data and general information

- The FIFA official web site:
  http://www.fifa.com
- The UEFA official web site:
  http://www.uefa.org
- The results of all European cups:
  http://www.europeancups.bravepages.com
- The results of the National championship of several countries
  http://www.rsssf.com/ec/
- Several data and statistics:
  http://www.soccerstats.com
- Other data about soccer:
  http://www.soccerway.com

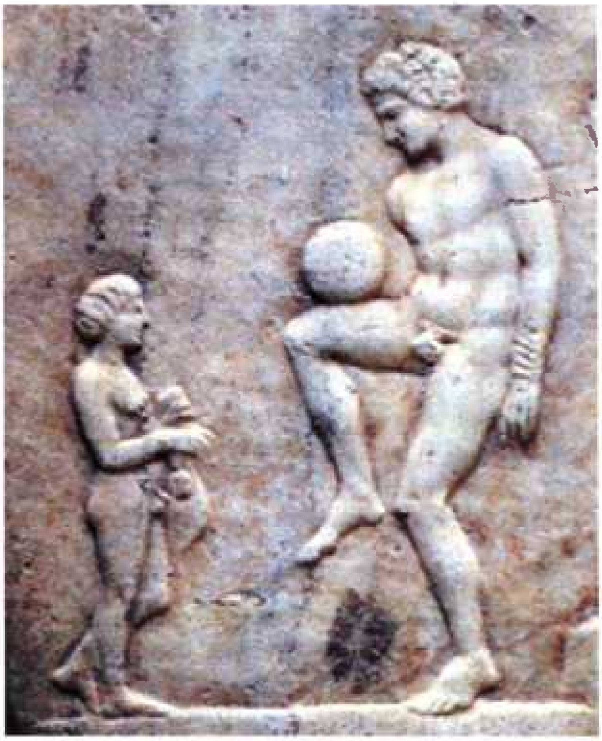
Statistics in soccer and betting

- Statistical community and journals:
  http://www.statsci.org
- Software on-line in order to perform prediction:
  http://users.aol.com/soccerslot/forthdim.html
- Betting odds:
  http://www.bettinggenius.com
- A site permitting to bet on-line:
  http://www.betandwin.com

Appendix II

The contribution of Ancient Greece in history of soccer

This page shows a marble relief from the National Museum of Archaeology in Athens; a Greek athlete balances a ball on his thigh, supposedly demonstrating a training technique to a boy. This very same image is nowadays featured on the European Cup trophy.



Episcyros played by the ancient Greeks and *Harpastum* played later on by the Romans can count themselves as the ancestors of the modern form of soccer.

Appendix III

The reaction of World's media for the Greek success in the Euro Cup (2004)

This page shows what reported some of the media in the entire world about the Greek National Football Team after winning the Euro Cup 2004 (Portugal).

- *Greece's improbable but deserved victory at UEFA EURO 2004™ has stunned the whole of Europe. It's not a dream. Greece are the champions! Charisteas's goal made eleven million Greeks the happiest nation in the world. The 4th of July must be Greece's new national holiday. Otto's warriors marched on to a surprising but well deserved triumph. Eusébio was among the spectators and he would probably have needed to come on to the pitch if his compatriots were to ever get past the resilient Greek defence. That defence was the best at EURO 2004 and no one could beat them. Sensational.*
  (Bulgaria - Meridian Match)

- *Greece are European champions! Believe it or not! Sensational, a real football miracle, unbelivable. Otto Rehhagel set the unbreakable defence and Greece deserved this unbelievable triumph. Like before at EURO 2004, Greece played with very high discipline, and with a strong and solid defence they didn't allow Portugal to get near their goal and produce any serious threat. Figo and company were weak, short of ideas and creativity. The Greeks were like football gods at the Luz last night.*
  (Croatia - Sportske novosti)

- *Charisteas sent shockwaves through the 62,000 in Estádio da Luz in Lisbon, as he headed the ball powerfully past the Portuguese keeper Ricardo. The Greek way of playing may and will be criticised but, the fact is, this underestimated Greece team are European champions."*
  (Denmark - Ekstra Bladet)

☺ *A tear flowed down the face of the beautiful game last night. An evening that was heaven for Hellas will be remembered as a triumph of tactics over instinct, of set-piece preparation that saw Angelos Charisteas head the Greeks, the 100-1 outsiders as EURO 2004 opened, to the most improbable of victories. If Greece's moment in the international sun was a success for organisation, good coaching and supreme fitness, last night's showdown of European football will trigger too many laments.*
(England - Daily Telegraph)

☺ *Greece, heroes of a modern mythology, caused one of the biggest surprises ever in football history at the Luz stadium. Winning the EURO 2004 final against the arch-favourites of Portugal, Otto Rehhagel's players achieved their Olympus. Their success can be compared with Uruguay's victory against Brazil in the 1950 World Cup or Germany's win over Hungary in 1954 or even to Denmark's achievement at EURO 92.*
(France - Le Parisien/Aujourd'hui en France)

☺ *50 years after the 'Wunder von Bern', Otto Rehhagel led Greece to the biggest sensation in the history of the European Championship. A goal from Bremen striker Angelos Charisteas on 57 minutes gave the Hellenic outsiders a surprising 1-0 win over favourites Portugal. The 15,000 Greek fans among the 62,865 spectators in the sold-out Estádio da Luz celebrated their side's first-ever trophy in a major tournament as Portugal failed to avenge their 2-1 defeat in the opening match.*
(Germany - Frankfurter Allgemeine Zeitung)

☺ *Just think - had Russia converted a single chance of the many they created against Greece the future winners would have been stopped in their tracks. But there is no sense in recalling that now. Let us think about another thing - that 65-year-old Otto Rehagel is the main hero of the championship. He did not possess the strongest team in the competition, but it proved not to be decisive. Because this team had the wisest and the most cunning head coach, a head coach who outplayed everyone.*
(Russia - Sport-Express)

☺ *Greece are champions of Europe. Believe me, it is not a dream. It is true, it is a reality. The captain, Theodoros Zagorakis, the best player in the stadium, raised the cup to the sky in Lisbon. These celebrations may be the first and only ever time for the Greeks. Angelos Charisteas was already the hero against Spain and France and, from yesterday, he is the seventh Greek god. Zeus, Apollo, Hermes, Ares, Poseidón, Ifestos, now Charisteas.*
(Spain - Marca)

☺ *Greece are the top team. The genius behind Greece is already royalty in Germany. It is King Otto, who has for a long time demonstrated his ability of saving teams that appear to have lost before the start of the game. He was unpopular in Greece for a long time but by making them Europe's top football nation last night, King Otto has become royalty in another football kingdom. He is the man who achieves the impossible.*
(Sweden - Aftonbladet)

☺ *The sensation is complete: football minnows Greece have beaten Portugal 1-0 in the European Championship final. Angelos Charisteas scored the golden goal. Exactly 50 years to the day after the 'Miracle of Berne', when Germany beat Hungary 3-2 in the World Cup final, football has written a new fairy tale - this time the 'Miracle of Lisbon'. It is even more remarkable than the 1992 European Championship (which had only eight teams), when Denmark became European champions after coming straight from their holidays because Yugoslavia were unable to take part.*
(Switzerland - Blick)

References

Agostinelli, C. and Markatou, M. (1998). A one-step robust estimator for regression based on the weighted likelihood reweighting scheme, *Statistics and Probability Letters*, 37, 341-350

Agostinelli, C. (2002). Robust model selection in regression via weighted likelihood methodology, *Statistics and Probability Letters*, 56, 289-300

Ali, A. H. (1988). A statistical analysis of tactical movement patterns in soccer, In *Science and Football* (eds. Reilly, T., Lees, A., Davids, K. and Murphy, W. J.), 302-308, Spon, London

Anderson, J. and Siddiqui, M. (1994). The Sampling Distribution of the Index of Dispersion, *Communication in Statistics: Theory and Methods*, 23, 897-911

Barnett, V. (1982). *Comparative statistical inference*, John Wiley & Sons, New York, 2nd edition

Bassett, G. W. Jr. (1997). Robust Sports Rating Based On Least Absolute Errors, *The American Statistician*, 51, 99-105

Barnett, V. and Hilditch, S. (1993). The effect of an Artificial Pitch Surface on Home Team Performance in Football (Soccer), *Journal of the Royal Statistical Society*, A 156, 39-50

Basu, A. and Lindsay, B. G. (1993b). The iteratively reweighted estimating equation in minimum distance problems, *Technical report*, Univ. Texas at Austin

Basu, A. and Sarkar, S. (1994). The trade-off between robustness and efficiency and the effect of model smoothing in minimum disparity inference, *Journal of Statistical Computation and Simulation*, 50, 173-185

Basu, A., Markatou, M. and Lindsay, B. G. (1995). Weighted likelihood estimating equations: The continuous case, Department of Statistics, Columbia University, New York

Basu, A., Sarkar, S. and Vidyashankar, A. N. (1997). Minimum negative exponential disparity estimation in parametric models, *Journal of Statistical Planning and Inference*, 58, 349-370

Basu, A. and Lindsay, B. G. (2004). The iteratively reweighted estimating equation in minimum distance problems, *Computational Statistics & Data Analysis*, 45, 105-124

Baxter, M. and Stevenson R. (1988). Discriminating between the Poisson and negative binomial distributions: an application to goal scoring in association football, *Journal of Applied Statistics*, 15, 347-354

Beaton, A. E. and Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band spectroscopic data, *Technometrics*, 16, 147-185

Becker, M. P., Yang, I. and Lange, K. (1997). EM Algorithms Without Missing Data, *Statistical Methods in Medical Research*, 6, 38-54

Beran, R. J. (1977). Minimum Hellinger distance estimates for parametric models, *Annals of Statistics*, 5, 445-463

Birch, J. B. (1980). Some convergence properties of iterated least squares in the location model, *Communications in Statistics*, B 9, 359-369

Bland, N. D. (1995). A mathematical analysis of football, *Mathematics Project*, Pimlico School, London

Bland, N. D. and Bland, J. M. (1996). Comment on 'Home ground advantage of individual clubs in English soccer' (44, pp. 509-521), *The Statistician*, 45, 381-383

Bõhning, D. (1998). Zero-inflated Poisson models and C.A.MAN: A tutorial collection of evidence, *Biometrical Journal*, 427, 157-162

Bõhning, D. and Hoffmann, K. - H. (1982). Numerical Techniques for Estimating Probabilities, *Journal of Statistical Computation and Simulation*, 14, 283-293

Boulier, B. L. and Stekler, H. O. (2003). Predicting the outcomes of National Football League games, *International Journal of Forecasting*, 19, 257-270

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs-I. The method of paired comparisons, *Biometrica*, 39, 324-345

Brown, M. C. (1982). Administrative succession and organizational performance: The succession effect, *Administrative Science Quarterly*, 27, 1-6

Byrd, R. H. and Pyne, D. A. (1979). Some results on the convergence of the iteratively reweighted least squares, *ASA Proceedings on Statistical Computation*

Cantoni E. and Ronchetti, E. (2001). Robust Inference for Generalized Linear Models, *Journal of the American Statistical Association*, 96, 1022-1030

Church, S. R. and Hughes, M. (1987). A computerised approach to soccer notation analysis, *Abstract of the 1st World Congress of Science and Football, Liverpool,* p.20, Liverpool Polytechnic, Liverpool

Clarke, S. R. and Norman, J. M. (1995). Home ground advantage of individual clubs in English soccer, *The Statistician*, 44, 509-521

Courneya, K. S. and Carron, A. V. (1992). The home advantage in sport competitions: a literature review. *Journal of Sport and Exercise Psychology*, 14, 13-27

Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society,* 46, 440-464

Croucher, J. S. (1984). The effect of changing competition points in the English football league, *Teaching Statistics*, 6, 39-42

Croucher, J. S. (1995). Scoring Patterns in Rugby League, *Teaching Statistics*, 17, 47-50

Crowder, M., Dixon, M., Ledford, A. and Robinson M. (2002). Dynamic modelling and prediction of English Football League matches for betting, *The Statistician*, **51**, 157-168

Del Pino, G. E. (1989). The unifying role of the iterative generalized least squares in statistical algorithms (with discussions), *Statistical Science*, 4, 394-408

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood From Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, B 39, 1-38

De Pierro, A. R. (1995). A Modified Expectation Maximization Algorithm for Penalized Likelihood Estimation in Emission Tomography, *IEEE Transactions on Medical Imaging*, 14, 132-137

Dewenter, R. (2003). Raising the Scores? Empirical Evidence on the Introduction of the Three-Point Rule in Portuguese Football, *Discussion Paper*, 22, Department of Economics, University of the Federal Armed Forces

Dixon, M. J. and Coles S. G. (1997). Modelling association Football Scores and Inefficiencies in the Football Betting Market, *Applied Statistics*, 46, 265-280

Dixon, M. J. and Robinson M. E. (1998). A birth process model for association football matches, *The Statistician*, 47, 523-538

Dobson, S. M. and Gerrard, W. (1997). Testing for rent-sharing in Football Transfer fees: Evidence from the English Football League, *Leeds University Business School Working Paper*, E97/03

Dobson, S. M. and Goddard, J. A. (1995). The demand for professional league football in England and Wales, 1925-92, *The Statistician*, 44, 259-277

Dobson, S. M. and Goddard, J. A. (2003). Persistence in sequences of football match results: A Monte Carlo Analysis, *European Journal of Operational Research*, 148, 247-256

Douglas, J. B. (1994). Empirical fitting of Discrete Distributions, *Biometrics*, 50, 576-579

Dowie, J. (1982). Why Spain should win the World Cup, *New Scientist*, 94, 693-695

Emonet, B. (2000). Revisiting Statistical Applications in Soccer, *Technical Report*, Department of Mathematics, Swiss Federal Institute of Technology, Lausanne

Fahrmeir L. and Tutz, G. (1994). Dynamic Stochastic Models for Time-Dependent Ordered Paired Comparison Systems, *Journal of the American Statistical Association*, 89, 1438-1449

Field, C. and Smith, B. (1994). Robust Estimation - A Weighted Maximum Likelihood Approach, *International Statistical Review*, 62, 405-424

Forrest, D. and Simmons, R. (2000). Making up the results: the work of the Football Pools Panel, 1963-1997, *The Statistician*, 49, 253-260

Forrest, D. and Simmons, R. (2002). Outcome uncertainty and attendance demand in sport: the case of English soccer, *The Statistician*, 51, 229-241

Franks, I. M. (1988). Analysis of association football, *Soccer Journal*, 33, 35-43

Gandar, J. M., Zuber, R. A., O'Brien, T. and Russo, B. (1988). Testing rationality in the point spread betting market, *Journal of Finance*, 43, 995-1008

Gelman, A. (2000). Discussion article, In *Optimization Transfer Using Surrogate Objective Functions* (eds. Lange, K., Hunter, D. R. and Yang, I.), *Journal of Computational and Graphical Statistics*, 9, 49-51

Gill, P. S. (2000). Late-Game Reversals in Professional Basketball, Football and Hockey, *The American Statistician*, 54, 94-99

Goddard, J. and Asimakopoulos, I. (2004). Forecasting Football Results and the Efficiency of Fixed-odds Betting, *Journal of Forecasting*, 23, 51-66

Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities, *Biometrika*, 58, 1-38

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some more robust and resistant alternatives (with discussions), *Journal of the Royal Statistical Society*, B 46, 149-192

Groenen, P. J. F. and Heiser W. J. (2000). Discussion article, In *Optimization Transfer Using Surrogate Objective Functions* (eds. Lange, K., Hunter, D. R. and Yang, I.), *Journal of Computational and Graphical Statistics*, 9, 44-48

Grunert, V. and Fieller, N. R. J. (1995). Data contamination versus model deviation, *Research Report*, 457, Department of Probability and Statistics, University of Sheffield

Hampel, F. R., (1968). *Contributions to the theory of robust estimation*, Ph.D. thesis, University of California, Berkeley

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics*, John Wiley & Sons, New York

Harris, I. R. and Basu A. (1994). Hellinger distance as a penalized Log Likelihood, *Communications in Statistics*, B 23, 1097-1113

Hill, I. D. (1974). Association football and statistical inference, *Applied Statistics*, 23, 203-208

Holland, P. W. and Welsch, R. E. (1977). Robust regression using reweighted least squares, *Communications in Statistics*, A 6, 813-827

Huber, P. J. (1964). Robust estimation of a location parameter, *Annals of Mathematical Statistics*, 35, 73-101

Huber, P. J. (1965). A robust version of the probability ratio test, *Annals of Mathematical Statistics*, 36, 1753-1758

Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol.1, 221-233

Huber, P. J. (1968). Robust confidence limits, *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 10, 269-278

Huber, P. J. (1972). Robust statistics: A review, *Annals of Mathematical Statistics*, 43, 1041-1067

Huber, P. J. (1981). *Robust Statistics*, John Wiley, New York

Hunter, D. R. and Lange, K. (2000). Rejoinder, Discussion article in *Optimization Transfer Using Surrogate Objective Functions* (eds. Lange, K., Hunter, D. R. and Yang, I.), *Journal of Computational and Graphical Statistics*, 9, 52-59

Hunter, D. R. and Lange, K. (2004). A Tutorial on MM Algorithms, *The American Statistician*, 58, 30-37

Jackson, D. A. (1994). Index betting on sports, *The Statistician*, 43, 309-315

Karlis, D. and Ntzoufras, I. (1998). Statistical modelling for soccer games: The Greek League, *Proceedings of HERCMA 98, Athens*, 541-548

Karlis, D. and Ntzoufras, I. (2000a). On Modelling Soccer Data, *Student,* 3, 229-245

Karlis, D. and Ntzoufras, I. (2000b). Distributions Based on Poison Differences with Applications in Sports, *Technical Report,* 101, Department of Statistics, Athens University of Economics and Business

Karlis, D. and Ntzoufras, I. (2003). Analysis of Sports Data Using Bivariate Poisson Models, *Journal of the Royal Statistical Society*, 52, 381-393

Karlis, D. and Xekalaki, E. (1998). Minimum Hellinger distance estimation for Poisson mixtures, *Computational Statistics and Data Analysis*, 29, 81-103

Karlis, D. and Xekalaki, E. (2001). Robust inference for finite Poisson mixtures, *Journal of Statistical Planning and Inference*, 93, 93-115

Keller, J. B. (1994). A characterization of the Poisson distribution and the probability of winning a game, *The American Statistician*, 48, 294-298

Knorr-Held, L. (2000). Dynamic rating of sports teams, *The Statistician*, 49, 261-276

Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate Discrete Distributions*, Marcel and Decker, NY

Koning, R. H. (2000). Balance in competition in Dutch soccer, *The Statistician*, 49, 419-431

Koning, R. H. (2003). An econometric evaluation of the effect of firing a coach on team performance, *Applied Economics*, 35, 555-564

Koning, R. H., Koolhaas, M., Renes, G. and Ridder, G. (2003). A simulation model for football championships, *European Journal of Operational Research*, 148, 268-276

Kuk, A. Y. C. (1995). Modelling paired comparison data with large number of draws and large variability of draw percentages among players, *The Statistician*, 44, 523-528

Kuonen, D. (1996). Modelling the Success of Football Teams in the European Championship (in French), *Technical Report,* 96.1, Department of Mathematics, Swiss Federal Institute of Technology, Lausanne

Kuonen, D. (1997). Statistical models for knock-out tournaments, *Technical Report*, 97.3, Department of Mathematics, Swiss Federal Institute of Technology, Lausanne

Kuonen, D., Chavez-Demoulin, V. Roehrl A. S. A. and Chavez, E. (1999). La France, championne du monde de football: qui l'eût cru?, *Technical Report*, 99.1, Department of Mathematics, Swiss Federal Institute of Technology, Lausanne

Kuonen, D. and Roehrl A. S. A. (2000). Was France's World Cup win pure chance?, *Student*, 3, 153-166

Lange, K., Hunter, D. R. and Yang, I. (2000). Optimization Transfer Using Surrogate Objective Functions, *Journal of Computational and Graphical Statistics*, 9, 1-20

Lee, A. J. (1997). Modelling scores in the Premier League: Is Manchester United Really the Best?, *Chance*, 10, 15-19

Lee, A. (1999). Modelling rugby league data via bivariate negative binomial regression, *Australian & New Zealand Journal of Statistics*, 41, 141-152

Leeuw, J., and Michailidis, G. (2000). Discussion article, In *Optimization Transfer Using Surrogate Objective Functions* (eds. Lange, K., Hunter, D. R. and Yang, I.), *Journal of Computational and Graphical Statistics*, 9, 26-31

Leroux, B. G. and Puterman, M. L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models, *Biometrics*, 48, 545-558

Li, C. S., Lu, J.C., Park, J., Kim, K. and Peterson J. (1999). Multivariate zero-inflated Poisson models and their applications, *Technometrics*, 41, 29-38

Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods, *The Annals of Statistics*, 22, 1081-1114

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, New York, Wiley

Lu, Z., Hui, Y.V. and Lee, A. H. (2003). Minimum Hellinger Distance Estimation for Finite Mixture of Poisson Regression Models and Its Applications, *Biometrics*, 59, 1016-1026

Maher, M. J. (1982). Modelling association football scores, *Statistica Neerlandica*, 36, 109-118

Marchand, E. (2002). On the comparison between standard and random knockout tournaments, *The Statistician*, 51, 169-178

Markatou, M. (1999a). A Closer Look at Weighted Likelihood in the Context of Mixtures, *Festschrift in Honor of T. Cacoullos*, Chapman & Hall

Markatou, M. (1999b). Weighting Games in Robust Linear Regression, *Journal of Multivariate Analysis*, 70, 118-135

Markatou, M. (2000). Mixture models, robustness and the weighted likelihood methodology, *Biometrics*, 56, 483-486

Markatou, M., Basu, A., and Lindsay, B. G. (1997). Weighted likelihood estimating equations: The discrete case with applications to logistic regression, *Journal of Statistical Planning and Inference*, 57, 215-232

Markatou, M., Basu, A., and Lindsay, B. G. (1998). Weighted Likelihood Equations With Bootstrap Root Search, *Journal of the American Statistical Association*, 93, 740-750

Marples, M. (1954). *A History of Football*, Secker and Warburg, London

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition, Chapman & Hall, London

McGarry, T. and Schutz, R. W. (1994). Analysis of the 1986 and 1994 World Cup Soccer Tournament, *ASA Proceedings from the 1994 Joint Statistical Meeting in Toronto*, *Statistics in Sports*, 61-65

McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm and Extensions*, New York, Wiley

Meng, X.-L. and Rubin, D. B. (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework, *Biometrika*, 800, 899-909

Meng, X.-L. (2000). Discussion article, In *Optimization Transfer Using Surrogate Objective Functions* (eds. Lange, K., Hunter, D. R. and Yang, I.), *Journal of Computational and Graphical Statistics*, 9, 35-43

Moroney, M. (1956). *Facts from figures*, 3rd edition, Penguin, London

Morris, D. (1981). *The Soccer Tribe*, Jonathan Cabe, London

Norman, J. M. (1998). Soccer, In *Statistics in Sport* (ed. Bennet, J.), 105-120, Arnold, New York

Olsen, E. (1988). An analysis of goalscoring strategies in the World Championship in Mexico, 1986, In *Science and Football* (eds. Reilly, T., Lees, A., Davids, K. and Murphy, W. J.), 259-277, Spon, London

Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative Solutions of Nonlinear Equations in Several Variables*, New York, Academic, 253-255

Pak, R. J. and Basu, A. (1998). Minimum disparity estimation in linear regression models: Distribution and efficiency, *Annals of the Institute of Statistical Mathematics*, 50, 503-521

Paukku, T. (1994). And it's another goal, *New Scientist*, 30-32

Peel, D. A. and Thomas, D. A. (1988). Outcome uncertainty and the demand for football, *Scottish Journal of Political Economy*, 35, 242-249

Peel, D. A. and Thomas, D. A. (1992). The demand for football: some evidence on outcome uncertainty, *Empirical Economics*, 17, 323-331

Pollard, R. (1986). Home advantage in soccer: a retrospective analysis, *Journal of Sports Sciences*, 4, 237-248

Pollard, R. (1995). Do long shots pay off?, *Soccer Journal*, 40, 41-43

Pollard, R. and Reep, C., (1997). Measuring the effectiveness of playing strategies at soccer, *The Statistician*, 46, 541-550

Pollard, R., Reep, C. and Hartley, S. (1988). The quantitative comparison of playing styles in soccer, In *Science and Football* (eds. Reilly, T., Lees, A., Davids, K. and Murphy, W. J.), 259-277, Spon, London

Read, T. R. C. and Cressie, N. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer, New York

Reep, C. and Benjamin, B. (1968). Skill and chance in association football, *Journal of the Royal Statistical Society A*, 131, 581-585

Reep, C., Pollard, R. and Benjamin, B. (1971). Skill and Chance in Ball Games, *Journal of the Royal Statistical Society A*, 134, 623-629

Ridder, G., Cramer, J. S. and Hopstaken, P. (1994). Down to ten: Estimating the Effect of a Red Card in Soccer, *Journal of the American Statistical Association*, 87, 1124-1127

Rotshtein, A., Posner, M. and Rakytyanska, H. (2003). Prediction of Football Games Results Based on Fuzzy Model with Genetic and Neuro Tuning, *Eastern European Journal of Enterprise Technologies*, 2, 10-18

Rousseeuw, P. J. and Leroy, A. (1987). *Robust Regression and Outlier Detection*, Wiley, New York

Rue, H. and Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league, *The Statistician*, 49, 399-418

Scully, G. W. (1995). *The Market Structure of Sports*, University of Chicago Press, Chicago

Simpson, D. G. (1987). Minimum Hellinger distance estimation for analysis of counted data, *Journal of the American Statistical Association*, 82, 802-807

Snyder, S., Subramanian, N. and Sun, J. (1998). Discrimination and Clustering – Can we learn from this college football data set?, Section in Statistics in Professional Sports, JSM, Dallas

Stefani, R. T. (1980). Improved least squares football, basketball and soccer predictions, *IEEE Transactions on Systems, Man and Cybernetics*, 10, 116-123

Stefani, R. T. (1983). Observed Betting Tendencies and Suggested Betting Strategies for European Football Pools, *The Statistician*, 32, 319-329

Stefani, R. T. (1987). Applications of statistical methods to American football, *Journal of Applied Statistics*, 14, 61-73

Stern, H. S. (1994). A Brownian Motion Model for the Progress of Sports Scores, *Journal of the American Statistical Association*, 89, 1128-1134

Szymanski, S. and Smith, R. (1997). The English football industry: profit, performance and industrial structure, *International Review of Applied Economics*, 11, 135-154

Titterington, D. M., Smith, A. F. and Markov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York

Van Dalen, H.P. (1994). Loont het om een voetbaltrainer te ontslaan?, *Economisch Statistische Berichten*, 79, 1089-1092

Winkelmann, R. (2003). *Econometric Analysis of Count Data*, 4th edition, Spinger, Berlin

Wright, D. B. (1997). Football standings and measurement levels, *The Statistician*, 46, 105-110

Wu, Y. N. (2000). Discussion article, In *Optimization Transfer Using Surrogate Objective Functions* (eds. Lange, K., Hunter, D. R. and Yang, I.), *Journal of Computational and Graphical Statistics*, 9, 32-34

Zavoina, R. and McElvey, W. (1975). A statistical model for the analyses of ordinal level variables, *The Journal of Mathematical Sociology*, 2, 103-120

Zermelo, E. (1929). Die Berechnung der Turnierergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung, *Mathematische Zesschrift*, 29, 436-460

Δωρεά