Project Number: CJM-SC04

STASTISTIC IN SCOCCER
PREDICTION OF THE ATHENS 2004

An Interactive Qualifying Project Report

submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Bachelor of Science

By

_____

Cheuk Wai David To

_____

Yin Ming Wong

_____

Ying Fang Yu

Date: October 2004

Approval

_____

Professor Carlos J. Morales

Abstract

In this project we investigate the performance of a variation of the so-called Bradley-Terry (BT) model for paired comparisons in the context of the 2004 Athens Soccer Olympics and the English League. We explore the usefulness of the BT model both as a tool to rank sport teams and as a prediction tool. We find that for the English League, the BT performs adequately in terms of ranking, but it performs poorly for the Olympic since data is too sparse.

Contents

List of Tables

Chapter 1

Introduction

Football (Soccer) is undoubtedly the most favorite of sports in the world (except in North America). It is a popular game among kids and teenagers in schools; it is a demanding sport for professional players, and it is a great commercial tool for giant corporate to sponsor and advertise their brand. In some countries, it is even a popular gambling option. Millions of dollars were spent into the bets during the World cup 2002 held in Japan and Korea. It exists and is practiced almost everywhere in the world by all ages, and international tournaments of football is guaranteed to be found every year or two; Famous tournaments include the World Cup, DOHA Asian Games (Asian Olympics), English Premier League and the Olympics, which was recently held in August 2004 in Athens, Greece.

Football as the world's favorite sport shares the same most important question of every other competition events: Who is going to win? In our project, we would attempt to find a way to model football tournaments results as accurate as possible in order to make predictions. We will use statistical methods to model and predict football match outcomes by collecting data and analyzing these data.

The primary goal of our project is to evaluate prediction techniques of a (tie-allowable) tournament (e.g. soccer league in England, France, Germany, Italy and Spain, and Swiss-round of most soccer tournament). We collected two data sets 1) results from the Olympic Games in 2004 and 2) results from the 2004 England league Tournament. The data we collected in the aid of our project's goal was mostly base on the result of the Swiss-round of the Olympic tournament. In addition, we also collected all the match results and related information from all of the Olympia qualifying games such as the U-21 in Europe and other continental qualifying tournament. These additional data will be used to complement the 2004 Olympic man's soccer championship data. Since all the Olympics football players participated in all tournaments mentioned above, more matches results should give us a larger sample size which we needed in statistic for more accurate estimation.

In the project, we would use a statistical program, SAS, to aid our data analysis and calculations. The codes and the analysis data computed by SAS can be found in Chapter 5. In our project, we will explore the Bradley-Terry Model, which is commonly used to predict sport games results. We will determine if this model is appropriate for soccer games result prediction, and explore potential weakness of this Model. Many sports results have

been accurately predicted by the Bradley-Terry Model, for example NBA, NFL and baseball matches[1].

In Chapter 2, we briefly present the History of the Olympic Soccer, and show the Athens Olympic 2004 Soccer Game's Equipment Requirement and Rules. This is important since different equipment and rules could influence the result of the game.

In Chapter 3, we present tournament structure and qualifying requirements for the Athens 2004 Olympic Soccer Game.

In Chapter 4, the Problem and Method section, we will explain how the Bradley-Terry Model works, how to apply it, and how the model fit into our project. (We took the Athens Olympic Soccer Tournament game as an example to test the Bradley-Terry Model. We will try to rank the teams when they are in the half way through the whole tournament. First, we would use the Binomial Model to get the proportion of wins from each team, and then use the Bradley Model to predict the outcome of matches for teams which have not played with each other before. For example, imagine there are teams A, B and C, and Team A played Team B, and Team B played Team C, but Team A didn't play Team C yet. Under this

[1] See Gregory J. Matthews MQP cited in the Bibliography

condition, we can apply this model to Bradley-Terry method to get the probability of wins between Team A and Team C, and we can also get the ranking among these three teams. Data Analyses is discussed in Chapter 5. We analyses the data we collect in the past months, including the results and predictions we got from the Olympics Soccer and European tournament. Also, we would illustrate how we apply the data into SAS.

In Chapter 6, we present result of our prediction and the project. (The Athens Olympic Tournament Game is just a small group game; therefore, the Bradley-Terry Model might not generate accurate results. We have generate the result of the European League with Bradley-Terry Model)

Chapter 2

Soccer in Olympic

2.1 History

Football (U.S. Soccer) first appeared in the Olympic Game of 1908 as official Olympic sport, but it was first introduced as a demonstration sport in back in the 1896 Olympic in Athens; finally, women's football was included in the Olympic competition program a century later in the 1996 Olympic Games in Atlanta.

Before 1984, only amateur players were allowed to participate in the Olympic football tournaments, and Eastern European countries had been dominating the competition until the inclusion of professional athletes at the Olympic Games. Participation rules, however, was a major issue, and it has been a subject of debate between the international Olympic committee and FIFA. Eventually, new regulations were established and specifying the age of 23 or younger could participant in the competition with the exception of three over-aged players. This new regulations had stopped the Euro domination, and it enabled other countries such as African to display their rich pool of Football talents.

2.2 Equipment

The Olympics Football tournament has it own equipment requirements (based on the fourth law of FIFA) which ensure the game is fair to all athletes from different countries. Players are not allowed to wear any equipment that is potentially dangerous to themselves or other players. The basic compulsory equipment of a player is: Ball, Uniform, Shoes, and Shin guards.

Ball

The Ball's requirement is very important because it is the major apparatus in the game, and a different ball can significantly affect the players' performance. In the 2004 Athens Olympic football game, FIFA requested the Ball to be made of leather or any other suitable synthetic material. As for the size, it must be 70 cm in diameter and 450 gram in weight.

Uniform

Each team member of the same team must wear identical uniform that has the same color. The goalkeepers, however, must wear different colors uniform to distinguish them from the other players. As for the referees and assistant referees, they also have to wear a different kind of uniform. Players are prohibited to wear jewelries which may cause injuries to themselves or other players.

Shinguards and Footwears

Players are allowed to wear some protective equipment, which is not dangerous to other players, to protect their leg and foot. The Shinguards must be made of suitable materials, such as rubber, plastic, or similar substance. The protective equipment is only allowed under the athletes socks.

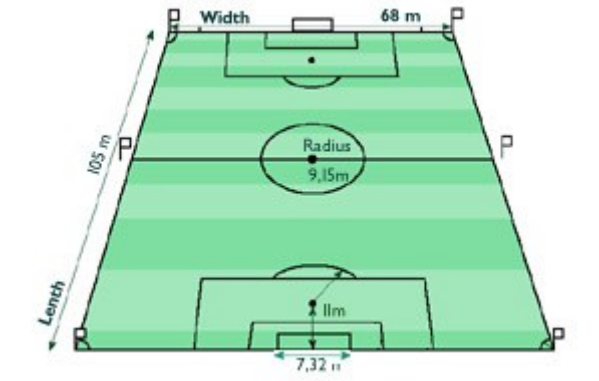
The players are allowed to wear the sport shoes with spikes on the soles, so athlete could get easier movement on the grass court. However, the spikes must be made of some sort of material which is not dangerous to other players.

Goalkeeper's Gloves

The goalkeepers could choose their gloves which they feel comfortable and protective. Special gloves that help a goalkeeper to grip the ball and goal post better are also allowed.

2.3 [2]Rule

Filed of play



A Football game is conducted on 105 x 68 m rectangular grass court marked by touch lines and goal lines. The court's centre is a circle with a diameter of 9.15 m. The area in front of the goalposts is divided into the goal area neighboring the goalpost and the penalty area extending towards the center. The goalpost has a length of 7.32 m and a height of 2.44 m. A white circular dot is marked at a distance of 11 m from the goalpost's centre for the penalty kicks. At the court's four corners are placed 1.5 m flag posts and a 90-degree arc marking the corner area.

[2] www.athens2004.com

The Contest

Each Football game is played by two teams of 11 players each, one of whom is the goalkeeper. During the game a team is allowed to make three substitutes from a group of seven players. No game can begin if a team has fewer than seven players. In Men's Olympic Football, each 18-member squad must include at least 15 athletes less than 23 years of age. There is no age limit for the remaining three.

A Football game lasts 90 minutes, two halves of 45 minutes each, with a 15-minute half-time break. The teams' aim is to score a goal without violating the rules of the sport. A goal is scored when the whole of the ball passes over the goal line between the opposite team's goalposts. The winner is the team to score the most goals.

A referee presides over a game and is in charge of implementing the rules. Two assistant referees moving along the two touch lines facilitate the referee's task. Before the beginning of the game, the referee draws lots. The winning team chooses a goalpost for the first half and the other team gets the ball at the referee's starting whistle.

Basic Law of the Kick

Direct free kick: It is called against the team whose player kicks, holds, trips, pushes an opponent or attempts these actions. The same will apply when a player handles the ball deliberately. In these cases, a player of the opposite team may perform a direct kick

towards the opponent's goalpost from the spot where the violation was committed.

Penalty: It is called when the above violations take place within the goal or penalty area, regardless of the ball's position. The kick is made from a distance of 11 m from the team's goalpost.

Indirect free kick: It is called when, according to the referee's judgment, a player's conduct is hazardous or when a player obstructs an opponent or the goalkeeper. It is also called against a goalkeeper who holds the ball in the penalty area more than six seconds, drops and then picks up the ball without an other player coming in contact with it or touches the ball by hand after a pass from a team mate. The player making the kick can't shoot straight to the opponent team's goalpost, but has to first pass it on to a teammate.

Cards

Yellow: The referee shows a yellow card to a player who plays aggressively towards the opponent, reacts in an untoward manner in words or gestures, causes delays or exits/enters the court without the referee's permission.

Red: The referee shows a red card and expels an athlete from the court if he/she violates the code of sportsmanship, displays unfitting conduct, hits an opponent, uses a hand to block the ball and obstructs the opposite team or is shown a second yellow card at the same game.

Chapter 3

Tournament

Football as the one of most popular sport in the world, hundred and thousands valuable football athletes and fans look forward to the 2004 Athens Olympic. However, there are more than 200 countries in this world, and there are only 16 teams qualified to play in the Olympic. Each continent held its own tournament to decide which countries would represent the continent to play in Athens. Since the Olympic Football game is restricted to players that are 23 years old or under; a lot of player have never participate in any world-class tournament before. Therefore, it is very difficult to predict which team/country will win the Olympics football Champion.

During the qualifying round, 4 teams from Europe region, 2 teams from North America region, 2 teams from South America region, 3 teams from Asia region, 4 teams from Africa region and 1 teams from Oceania region were allowed to enter the final round.

3.1 Europe

The qualifying tournament of Europe is the European U-21 Championship (players must be of age 21 or under), a total 48 teams in Europe participated, and they are divided

into 10 groups to play in Swiss, 16 teams would advance to the second round and play a double elimination. Then, 8 teams would advance to the third round and are divided into 2 groups in Swiss. 4 teams would advance for the final single elimination; the 4 teams would all get a chance to play the 2004 Olympics.

3.2 South America

The qualifying Tournament for South America is held in Chili, 10 participating teams are divided into 2 groups. 2 teams would advance from each group to form another group. The top 2 teams in this group would qualify for the Olympic.

3.3 North and Central America

There are 20 teams in the qualifying Tournament for the Olympic in North and Central America Region. 2 rounds of Double elimination would eliminate 12 teams. The remaining 8 team would form 2 groups. The first and second standing teams of each group would advance for the final single elimination.

3.4 Africa

24 teams would fight to qualify in the African Region. 16 teams would advance to the second round and are divided into 4 groups. Each group's winner will automatically qualify for the Olympic tournament.

3.5 Asia

In Asia, 24 teams would enter the qualifying Tournament for Olympic. 2 rounds of double elimination for the first and second round and 12 teams would advance and are divided into 3 groups. The top standing of each group would automatically qualify for Olympic tournament.

3.6 Oceania

There are 10 teams in the qualifying Tournament for Olympics in the Oceania Region. 10 teams are divided into 2 groups, and the top standing of each group will play each other for the qualification of the Olympic tournament.

In real world competition, there are always too many players but there weren't enough time for each player to play with each other. Our project is base on this situation; try to find out a solution for ranking all the teams in the competition from the best to worst. We are using Olympic Soccer tournament as an example to show how a world wide tournament deals

with such a problem. And are there ways to improve the ranking system, so we could get the most accurate ranking.

During the 3 weeks of Olympic tournament, 28 sports over 300 matches were competing in Athens; in this short limit of time, not every sport have a chance to play with each opponent equally, especially the sports with score system, coz these sport may take an hour or 2 to finish each match, so find out a better way to improve the ranking system of these sport will greatly help how to determine whose the winner.

The ranking system which Olympic soccer tournament using now will clearly tell the first to the forth place, coz there are final match and a match for the third place, but how about the others, no one know. So that's our main concern, how to use statistic to rank teams from first place to the last, and use these data to estimate the winner of the tournament.

But the problem with our statistic analysis we don't have enough data fully analysis the strength of each team. Under the Olympia soccer tournament condition, only players under 23 or less were allowed to enter the tournament, most professional players were not allow to enter, this is one of the reason we lack of data to analysis. Another problem had mentioned before, lack of time for each team to play with each other, since every team only

play with a limited numbers of team, so we need to compare these by who they had defect and who defected it.

Chapter 4

Problem and Methods

We are considering how to get that Olympic Football game's result and predicting who would be the Gold Medal. In our project, we choose Bradley-Terry method to explore the paired comparison. And try to get the ranking of all the team with SAS.

4.1 Bradley-Terry Model

Bradley-Terry method was found by Bradley and Terry in 1952. It is a useful device for the scoring of the tournament for which the data are the result of contest between pairs of players.

The Bradley-Terry model is for paired comparisons. The Bradley-Terry model deals with a situation in which individuals or items are compared to one another in paired contests. The model assumes there are positive quantities $\gamma 1. \gamma 2 \ldots\ldots \gamma n,$ which can be assumed to sum to one, such that

If the competitions are assumed to be mutually independent, then the probability *Pij=P (T1 defeats T2)* satisfies the login model, but this model is definitely not apply in the game if it ties.

$$Log\,[P_{ij} / (1 - P_{ij})] = \gamma i - \gamma j + U$$

Pij is defined as the probability that the team *i* defeats the team *j* and *U* define the home field advantage.

This model can be fit to a particular set of data by setting up an appropriate design matrix and response vector for a binomial regression model. It is possible to construct functions that allow the data to be specified in a more convenient form.

$$P_{ij} / (1 - P_{ij}) = odds$$

As we know the P_{ij} is the wining probability between two teams. According to the binomial model, the parameters (n, p), when $0 \leqq p \leqq 1$, so that the range of the *Pij* would be [0, 1]. Odds of win, in the equation $P_{ij} / (1 - P_{ij}) = odds$. The factor $1 - P_{ij}$ is also equal to P_{ji} (Probability of team *i* defects team *j*). P_{ji} is a possibility that the range of P_{ji} should be [0, 1]. $P_{ij} / P_{ji} = P_{ij} / (1 - P_{ij}) = odds$, base on the variables we got above, we can calculated $0 \leqq odds < \infty$. When we taking the Log into the odds, the range will be expand to $-\infty < Log\,(odds) < \infty$.

The Log odds are then:

$$Log\,[\,Pij / (1 - Pij)] = \gamma i - \gamma j$$

From the formula above, we can got

$$Pij = e^{\gamma i - \gamma j} / (1 + e^{\gamma i - \gamma j})$$

4.2 Home Field Advantage

Home field advantage is one of the factors to determine the result. If two teams are in the same level of rank, when they match each other, the odd of the home field team would be better than the away team.

In the explanation above, we just shown the regular Bradley-Terry model, however, it is not 100% accurate. Now, we will add the home field advantage onto this model, and explore what the results would be.

The paired comparison is ordered and postulates that the probability of team *i* defeat team *j* depends on which individual is listed first. If the individuals are sports teams, this assumption leads to the home field advantage model.

$$P(i \text{ beats } j) = \theta\gamma_i / (\theta\gamma_i + \gamma_j) \quad \text{if } i \text{ is home}$$
$$P(i \text{ beats } j) = \gamma_i / (\gamma_i + \theta\gamma_j) \quad \text{if } j \text{ is home}$$

when θ is represent the home field advantage where the θ must be greater than zero. The equation above is shown the probability of team *i* beating team *j* when the team *i* is home and when the team *j* is home.

Tie would be one of the possibilities between two teams, if we want to compare two teams odd, we can use the following equations.

$$P(i \text{ beats } j) = \gamma_i / (\gamma_i + \theta\gamma_j)$$
$$P(j \text{ beats } i) = \gamma_i / (\theta\gamma_i + \gamma_j)$$

$$P(i \text{ ties } j) = (\theta^2 - 1)\,\gamma i\gamma j / [(\gamma i + \theta\gamma j)(\theta\gamma i + \gamma j)]$$

We should assume the worst result of it, so that we will use the equation that the teams play in the away game. And set the home field advantage θ is greater than 1 as "threshold" parameter.

Let's give the different adjustment to the Bradley-Terry model to account for ties, in which the probabilities are in the ratio.

$$P(i \text{ beats } j) : P(j \text{ beats } i) : P(i \text{ ties } j) = \gamma i : \gamma j : \theta\sqrt{(\gamma i \gamma j)}$$

The positive valued parameter θ in the equation above is the constant of proportionality if the probability of a tie is proportional to the geometric mean of the probabilities of a win by either individual.

4.3 Bradley-Terry Model in application

When the Athens Olympic Tournament game was still processing, we were try to predict the small group qualify result. We took South America tournament Group A as an example. We were predicting the result when they are in the half way of their game. We computed the proportions of winning from each team. The table of proportion of wins is shown below.

TEAM	CHILE	BRAZIL	PARAGUAY	URAGUAY	VENEZUELA
PROPORTION OF WINS	3/4	1/2	1/2	0	0

Since the each team has to play with all of the rest of teams in the group, so we will set the home field advantage is 2. Apply for the Bradley-Terry Model.

P (Chile beats Brazil) = 2 x 3/4 / (2 x 3/4 + 1/2) = 3/4 (Chile home)

P (Chile beats Brazil) = 3/4 / (3/4 + 2 x 1/2) = 3/7 (Brazil home)

P (Brazil beats Chile) = 1/2 / (1/2 + 2 x 3/4) =1/4 (Chile home)

P(Chile Ties Brazil) = $(2^2 -1)$x 3/4x 1/2 / (3/4 + 2 x 1/2)(1/2 + 2 x 3/4) = 9/28 (Chile home)

From the calculation shown above, we can tell that how important of the home field advantage, the probability of the wins would be decrease when you are away.

P (Chile beats Brazil): P (Brazil beats Chile): P (Chile Ties Brazil)

=$\gamma i : \gamma j : \theta\sqrt{(\gamma i \gamma j)}$

=3/4 : 1/2 : 2√(3/4X1/2)

= 3 : 2 : 2√6

The ratio of the result, when the θ represents the home field advantage, γi represents the proportion of wins from Chile; and γj represents the proportion of wins from Brazil. We got the result that Chile would be the first place in the Group, and compare to the actual result, our prediction method is seems working for the teams that play with each other. Now, we would like to try to predict the result that all the teams which are not play with each others and looking for if any different between our prediction and the actual result.

Chapter 5

Data Analysis

5.1 Data collected from the 2004 Olympics Soccer Game

Here are the data from the Athens Olympic Soccer game's result and the final round of the European U-21 Championship result. Those are the basic data that we need to set up the for the SAS code. In both tables, the first column represents the number of wins from each team; the second column represents the number of ties from each team; the third column represent the number of lose from each team, and the fourth column represent the probability of winning of each team. In both tables, we can found out which team is the stronger from the data obviously. However, the percentage of wins also could help us to calculate the winning percentage of the pair that they didn't match as all. We can compute the result with the Bradley-Terry Model formula P (T1 defects T2) = $\gamma 1 / (\gamma 1 + \gamma 2)$ we shown in the last chapter.

TEAMS	GAMES	WINS	TIES	LOSE	PERCENTAGE OF WIN
GREECE	3	0	1	2	0
KOREA	4	1	2	1	1/4
MALI	4	1	2	1	1/4
MEXICO	3	1	1	1	1/3
PARAGUAY	6	5	0	1	5/6

JAPAN	3	2	0	1	2/3
GHANA	3	1	1	1	1/3
ITALY	6	3	1	2	1/2
ARGITINA	6	6	0	0	1
SERBIA&MONTEN	3	0	0	3	0
TUNISIA	3	1	1	1	1/3
AUSTRILIA	4	1	1	2	1/4
COSTA RICA	4	1	1	2	1/4
MOROCCO	3	1	1	1	1/4
IRAQ	6	3	0	3	1/2
PORTORGAL	3	0	0	3	0

Table 5.1: Table of Olympic Football Result

TEAMS	GAMES	WINS	TIES	LOSE	PERCENTAGE OF WIN
ITALY	5	4	0	1	4/5
SERBIA&MONTEN	5	2	1 a.e.t	2	2/5
BELARUS	3	1	1	1	1/3
CROATIS	3	0	0	3	0
SWEDEN	5	3	1 a.e.t	1	3/5
PORTUGAL	5	2	1	2	2/5
GERMANY	3	1	0	2	1/3
SWITZERLAND	3	0	1	2	0

Table 5.2: Table of U-21 Final Round Result

P.S: In the Game Sweden VS Serbia & Montenegro was tie at score 1:1 (A.E T at 5:6). So we consider this game is tie.

5.2 The English Premier League and SAS

SAS can be used to estimate the ranking among a large number of teams. We took the England league as an example. There are 20 teams in the league, so it is hard to rank them by hand, and so we can change their score onto the SAS code and run it with SAS program.

Since there are 20 teams, the SAS code file contains 380 lines; please see the England league detail SAS code in the index. Below is an example of the SAS code.

Team A	Team B	Team C	Team D	Team E	Count
1	-1	0	0	0	2
1	0	-1	0	0	1
1	0	0	-1	0	0.5
1	0	0	0	-1	1
0	1	-1	0	0	1
0	1	0	-1	0	2
0	1	0	0	-1	2
0	0	1	0	-1	1
0	0	0	1	-1	0.5
-1	1	0	0	0	1
-1	0	1	0	0	2
-1	0	0	1	0	0.5
-1	0	0	0	1	2
0	-1	1	0	0	2
0	-1	0	1	0	1
0	-1	0	0	1	1
0	0	-1	0	1	2
0	0	0	-1	1	0.5

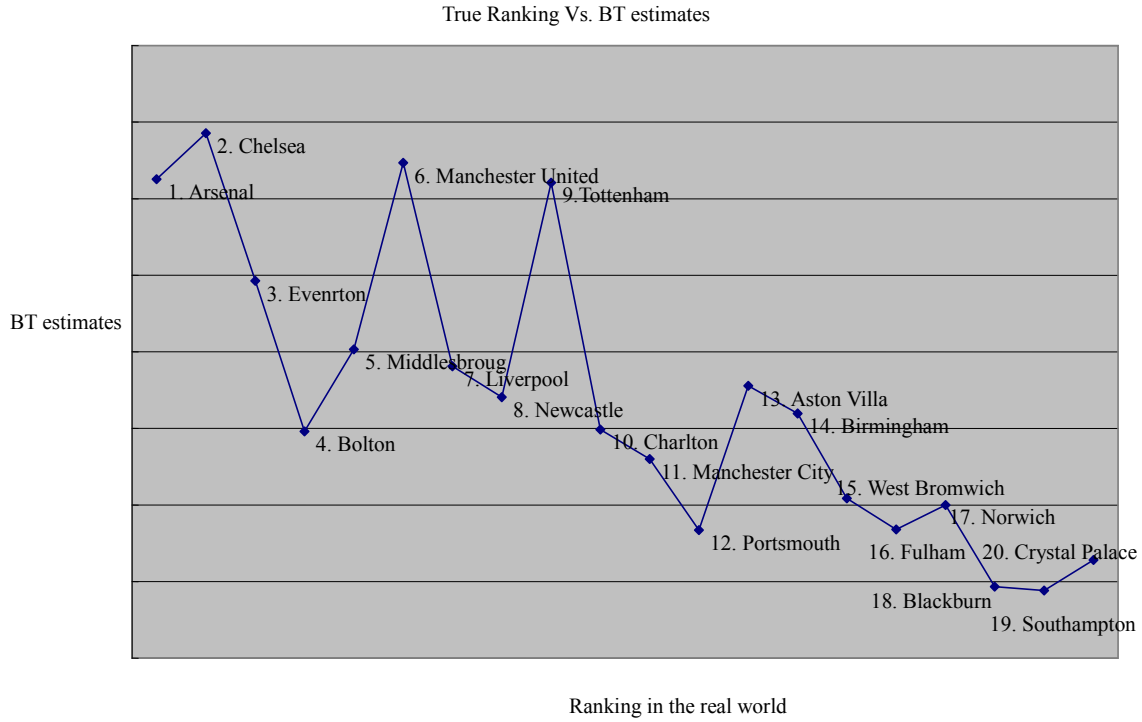
Table5.3: SAS Code Example

The first 5 columns shows which two team play each other. We are using -1 and 1 to represent the participating teams. The team with a number 1 is facing the team with number -1. In the sixth column, the numbers indicate that either the team wins, loses or both tie; with a 2 in the sixth column means the team is win, with a 1 means that team is

lose, and a 0.5 means they are tied. For the teams that do not get a chance to play with some other teams, there number of matches counter is filled with zeros.

Parameter	Estimate Rank	Standard Error	Wald Chi-Square	Pr>ChiSq	Current Rank
Chelsea	0.9708	0.9192	1.1155	0.2909	2
Manchester United	0.8938	0.8581	1.0849	0.2976	6
Arsenal	0.851	0.8543	0.9923	0.3192	1
Tottenham	0.8417	0.9901	0.7226	0.3953	9
Evenrton	0.5857	0.922	0.4036	0.5252	3
Middlesbroug	0.4065	0.944	0.1854	0.6667	5
Liverpool	0.3621	0.8068	0.2014	0.6536	7
Aston Villa	0.3113	1.1147	0.078	0.78	13
Newcastle	0.2815	1.0072	0.0781	0.7798	8
Birmingham	0.2392	1.0129	0.0558	0.8133	14
Charlton	0.1967	0.9572	0.0422	0.8372	10
Bolton	0.1924	0.9601	0.0402	0.8411	4
Manchester City	0.1198	0.9211	0.0169	0.8965	11
West Bromwich	0.0171	0.9936	0.0003	0.9863	15
Norwich	0	.	.	.	17
Fulham	-0.0641	0.9551	0.0045	0.9465	16
Portsmouth	-0.0657	0.9942	0.0044	0.9473	12
Crystal Palace	-0.1437	0.961	0.0223	0.8812	20
Blackburn	-0.2132	0.9989	0.0456	0.831	18
Southampton	-0.2235	0.979	0.0521	0.8194	19

Table 5.4: Analysis of Maximum Likelihood Estimates



This is the SAS result of the England League, the parameter column are the actually ranking from top from bottom when we collect our data. From our code, the program helps us rank the 20 teams by strength. The table above is the analysis of Maximum Likelihood Estimates; we can compare the differences from the point estimates column. We also found that the confidence intervals for the probability that one team defeats another team with ninety-five percent confidence.

According to SAS, the ranking by strength are Chelsea, Manchester united, Arsenal, Tottenham, Everton, Middleburg, Liverpool, Aston Villa, Newcastle, Birmingham,

Charlton, Bolton, Manchester city, west Bromwich, Norwich, Fulham, Portsmouth, crystal palace, Blackburn, and Southampton.

From the above we can obviously see that the top five team and the worst five team almost staying at their place, except that the teams in the middle are no in any order, this is due to all the teams have only play up to 7 teams, and some may face a stronger team, some may face a weak team. From the most recent ranking, Arsenal, Chelsea, Everton, Bolton, and Middlesburg are the top five team and Fulham, Norwich, Blackburn, Southampton and Crystal Palace are the worst team in the league. So although the data we analysis with Bradley-Terry model is not very accurate, but it could give us an over view on what is happening.

This is the SAS result of the England League, from our code, the program help us rank the 20 teams. The table above is the analysis of Maximum Likelihood Estimates; we can compare the differences from the point estimates column. We also found that the confidence intervals for the probability that one team defeats another team with ninety-five percent confidence.

5.3 Discussion of output from SAS

In Table 5.4, Analysis of Maximum Likelihood Estimates. In general, we could tell which the strongest team is, and which the worst team is by the estimate relative strength estimated.

The result we got from the SAS program is accurate, since each teams estimate with not much difference. Using Norwich as the base team, the SAS program compares all the other teams with Norwich. From the most recent result, it should that not all the teams follow the ranking, but some of the ranking are accurate.

In the three set of testing data, although there are large numbers of teams, they do not play against every other team in the set, but instead they only play against two to three team in the list. So the Bradley-Terry program can give you a overview of the strength of each team.

Chapter 6

Problem of the prediction

When people were still guessing who would be the gold medal of the Athens Soccer 2004, we were trying to use our statistical knowledge to predict the winner. In the previous Chapter, we learn how to estimate parameters for paired comparison data from teams. Note that the data from the Athens Soccer 2004 has the form required for an application of the Bradley-Terry Model, and use SAS to aid in getting results.

Here, we will use the same set up data as the one used for the England league data set. We set win to be 2, lose to be 1 and ties to be 0.5, if two teams which never play with each other will be set to 0.

Effect	Point Estimate	95% Wald confidence limit	
GRE	0.249	0.011	5.776
KOR	0.923	0.052	16.536
MLI	0.693	0.036	13.471
MEX	0.371	0.021	6.6900
PAR	2.297	0.255	20.709
JPN	1.637	0.135	19.807
GHA	1.939	0.126	29.934
ITA	1.925	0.231	16.039
ARG	1.776	0.242	13.020
SCG	0.737	0.057	9.469
TUN	1.144	0.078	16.828

AUS	1.603	0.134	19.143
CRC	1.099	0.200	6.0490

Table 6.1: Odds Ratio Estimates

Note that we obtained the above results using data only up-to the quarter finals. From this table, we can tell who would be the gold medal easily. The column with the header Point Estimate gives an estimated measure of the strength of that team relative to all other teams. Paraguay has the highest point estimate for strength, which means it has the highest chance to win the gold medal (in a round robin contest). An important point to make is for instance the case of Ghana. Ghana didn't qualify to the quarter final, but they have a high estimate. However, Ghana was grouped with Paraguay and Italy, both of which qualified to the semi-final round. In the first round, Ghana won against Paraguay and tied with Italy. Since Paraguay and Italy got a good score in the entire game, the BT model ranks Ghana very high.

In fact, the gold medal winner is Argentina, which has a point estimate of 1.776 for its strength. According to this SAS output, this ranking seems reasonable. However, Argentina comes in this ranking behind Ghana. How can this be? It seems that the prediction for the Olympics is not as good as the prediction for the case of the English League.

In the SAS output (Appendix B), base on the model, Paraguay has the highest ranked in the tournament, Greece has the lowest ranked in the tournament, and Portugal has the zero estimates in the maximum likelihood estimate. This output seems to be accurate. We also found ninety-five percent of confidence interval for the probability of one team defect another team, and the interval are all from 0 to 1. The reason of the failure is because our teams had never played with all the teams in the tournament, and they only play with each other zero time or one times, so we don't have enough information to connect two teams which never play with each other in the tournament.

Base on the result we got from SAS, and the actual result from the Athens Soccer 2004. We found out there is a big difference in our prediction, and it is not even close. This findings suggest that the Bradley-Terry Model is not appropriate for this data set since there aren't enough connections among teams; that is, teams do not play each other often enough, and they have very few direct opponents.

Chapter 7
Conclusion

Due to all the failure we encounter when we using Bradley-Terry model to simulate the result, we concluded that Bradley-Terry model is not suitable to use in tournaments that all the teams doesn't have much connection, which means either they didn't face each others at least once or they don't even have a chance to face each other. As the data for the Olympic 2004 men's soccer shown, all the teams only match up for one or two times only (for those team that match up 2 times, the would be in the same group and both of them advance to face each other). Another important issue concern is that Bradley-Terry model doesn't support scoring system with ties (in our case, we use 0.5 points as a tire instead of a winning score of 2 or lost for 1).

As a result Bradley-Terry model will work the best when there is an environment that each team gets the chance to compete a large number of times with all competitors, in order to get an accurate result. In this case, Bradley-Terry model will work better with football or baseball, when team played each other many times (i.e. at least two times in NFL, and as many as 19 times in baseball).

Another use of the Bradley-Terry model is to help setting up a tournament table. Since you can collect a lot of data for each of the teams, we can use the Bradley-Terry to get a ranking for each team. This way you can prevent seed teams or player match up on the first round. Note the ranking is not absolute, but the high ranked teams will have a higher chance to win the tournament.

The Bradley-Terry model is a very simple ranking and data analysis system for sport, although it needs a large amount of data to get an accurate result. Yet it still very useful for most sports whose matches do not result in ties.

In conclusion, the Bradley-Terry model is a useful ranking tool for most sports, and it can be used to predict the winner at the end of the season.

Chapter 8

Appendix

8.1 SAS Code

```
data w; input
TeamA TeamB TeamC Wins;
y=1; cards;
1 -1 0 2
0 1 -1 0.5
-1 1 0 1
0 -1 1 0.5
;
proc logistic;
        freq Wins;
model y = TeamA TeamB TeamC / noint covb;
        output out = out p = p;
run;
proc print;
run;
```

8.2 SAS Data

A. England League 2004

B. Athens 2004 Soccer Game

Bibliography

Gilbert W. Bassett, Jr. (1997). Robust Sport Rating Based on Least Absolute Errors. In The American Statistician (Vol.51, No. 2, 99-105.)

David R. Hunter. (2004). MM Algorithms For Generalized Bradley-Terry Model. In The Annals of Statistics (Vol. 32, No.1 384-406)

Leonhard Knorr-Held. (1997). Dynamic Rating of Sports Teams. Germany: Ludwigster 33 D-80539 Munchen.

Gregory J. Matthews (2003) Paired Comparison Logistic regression Models Used to rank Competitors in Sports. Major Qualifying Project at WPI.

Ross Sheldon. (2002). A First Course in Probability(6th edition). NJ: Prentice Hall.

Joseph D. Petruccelli, Balgobin Nandram, & Minghui Chen. (1999). Applied Statistics For Engineers and Scientists. NJ: Prentice Hall.

The official website of the ATHENS 2004 Olympic Games - Games of the XXVIII Olympiad. (2004). Retrieved August, 2004 from the internet: http://www.athens2004.com/en/OlympicGamesIndex/olympichome

England League (1997- 2004) http://www.soccerstats.com/latest.asp?league=england