Modeling group stages of football tournaments: using the Skellam’s distribution for the goal difference

Yuan He

1 Introduction

Football (or in American English, soccer) has always been one of the most popular sports worldwide and data-driven analysis of football games becomes increasingly popular in the recent years. Among football analytics, statistical predictive modeling is the jewel in the crown for its capability of predicting game outcomes. Accurate predictions, on the one hand, will allow football-related organizations to improve their performance and thus maximize the profit; on the other hand, it helps individuals in betting.

However, it is proved hard to produce highly accurate predictions because so many factors can possibly influence the outcome of a football game: team strengths, referee, crowds, weather, injuries and randomness itself. Despite the difficulties, statisticians have come up with various models over the years. The fundamental paper by Lee (1997) proposed to use the Poisson distribution to model the number of goals scored by each team. Brillinger (2006, 2009) modeled the Norwegian Premier League and the Chinese Super League, respectively, using ordinal-valued (win-draw-loss) responses along with the Poisson model. Karlis and Ntzoufras (2003) extended the bivariate Poisson model to incorporate some time-dependent effects. Their 2008 paper used the goal difference, instead of goals scored by each team, as the response. By modeling the goal difference using the Skellam’s distribution (or Poisson difference distribution), Karlis and Ntzoufras claimed that the effect of correlation between the two competing teams was removed. This paper will use the same model (based on Skellam’s distribution) to predict outcomes of group stage games of football tournaments.

Among various tournaments in European football, the UEFA Champions League (ECL) never fails to be the spotlight since late 1990s. Top teams from major European leagues qualify for this tournament and the group stage consists of 32 teams (ever since 1999-2000 season) divided into 8 groups. The 32 teams will be firstly divided into 4 seeding pots according to their UEFA club coefficients (calculated based on the club’s performance in recent European competitions), and then each group will have 4 teams, with one team from each seeding pot. In other words, each group from the ECL group stage will have one tier-1 team, one tier-2 team, one tier-3 team and one tier-4 team. Each team in the group will play against each one of the other teams, twice, home and away. So there will be a total of 12 games in each group.

In this paper, match results were recorded for each of the ECL group stages from season 2005/06 to season 2014/15 (a duration of ten years). The model for goal difference was fitted on these results and parameters of the model were estimated using the data. Simulations based on the model were then used to discuss some of the ranking rules of ECL group stages.

The structure of this paper is as follows: Section 2 describes the raw data and provides some exploratory analysis. Section 3 describes the model for goal difference, along with methods for parameter estimation and prediction algorithms. Section 4 assesses the model and produces simulations to discuss ranking rules. Finally section 5 discusses possible problems and future developments.

2 Data Description & Exploratory Analysis

2.1 Data Description

Data was collected from Wikipedia entries of each season’s (from 05/06 season to 14/15 season) ECL group stage. Game results were manually entered into a spreadsheet, and then imported into R for visualization and analysis (it was rather a pleasure, not torture, for a football fan to record and savor the results). A glimpse of the raw data can be shown in Figure 1:

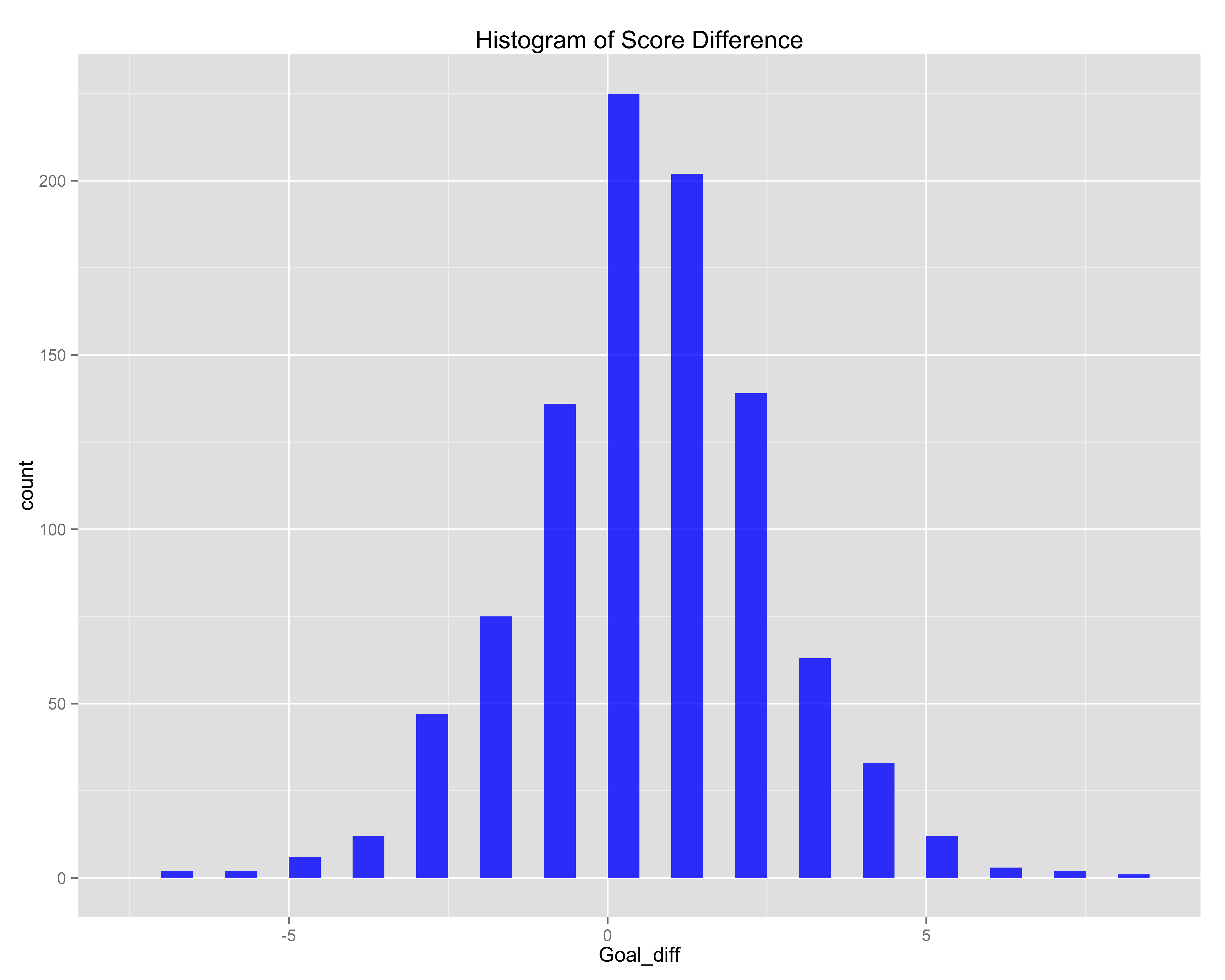


*Figure 1: A glimpse of raw data.*

The raw data has 960 rows with each row representing the information of one game. The fields (columns) of raw data are pretty straightforward: “Year” and “Group” record the season and group label of each game; “Home\_name” and “Home\_seed” record the club name and seeding pot of the home team, same for “Away\_name” and “Away\_seed”; “Home\_score” and “Away\_score” are the goals scored by the home team and away team, respectively; Lastly, “Score\_diff” is simply “Home\_score” minus “Away\_score” (a value of “2” means the home team win by 2 goals; “-1” means the home team lose by 1 goal and “0” indicates a draw).

2.2 Exploratory Analysis

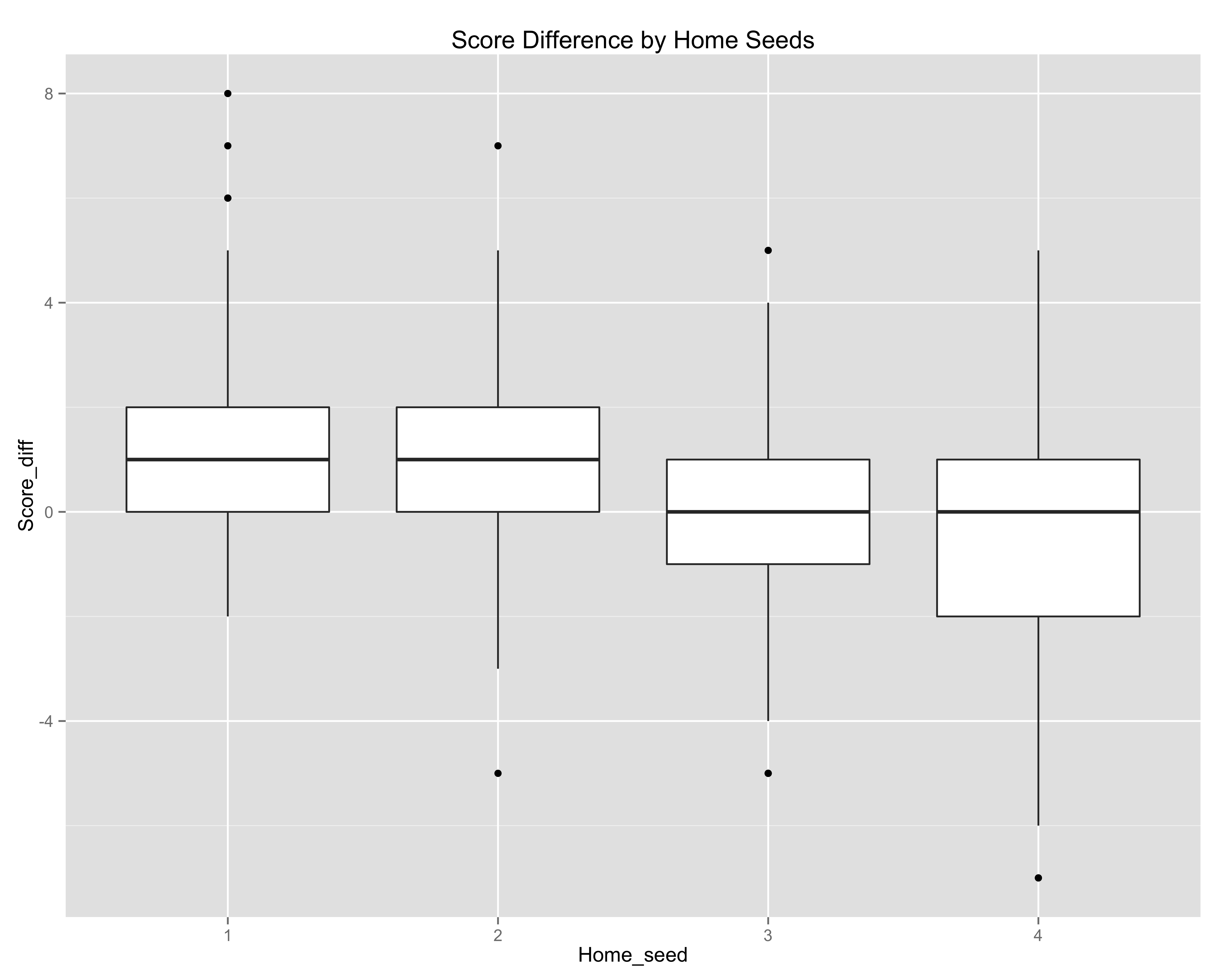
We are particularly interested in the last column of raw data since the response of our model is goal difference. A histogram visualization of the goal differences of these 960 games is shown in Figure 2:



*Figure 2: Histogram of goal difference of 960 ECL group stage matches. Skewness towards the right indicates home advantage.*

It can be seen from the histogram that for most matches, the goal difference centers around 0 and lies between the interval of -5 to 5. The distribution is significantly skewed towards the right (positive direction), indicating the presence of home advantage. In fact, the average of goal difference of these 960 games is 0.385. In other words, disregarding all the tier difference, the home team scores 0.385 more goals in general, a pretty revealing number addressing the home-court advantages.

What about the seedings? It is expected that teams from higher seeding pots (or higher tiers) will perform better than teams from lower seeds. And in fact this is the truth, as revealed in Figure 3.



*Figure 3: Boxplots goal difference by home seeds/tiers. Teams from tier-1 and tier-2 have goal differences way above zero.*

Moreover, Table 1 gives the average goal difference for each pair of match-up. It can be seen that tier-1 teams have advantages for each possible pair of match-up, no matter home or away. The biggest average goal difference appears when tier-1 teams play tier-4 teams at home, where clubs from seeding pot 1 win by 2 goals on average. The closest match-up seems to be tier-3 teams playing tier-2 teams at home, which is reasonable considering the closeness of two tiers and the home advantages.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Away  Home | 1 | 2 | 3 | 4 |
| 1 | - | 1.05 | 1.38 | 2.00 |
| 2 | -0.08 | - | 0.56 | 1.18 |
| 3 | -0.63 | -0.01 | - | 0.85 |
| 4 | -1.05 | -0.77 | 0.11 | - |

*Table 1: Pairwise average goal difference.*

3 The Skellam Model of goal difference

3.1 Model Description

As Karlis and Ntzoufras (2008) proposed, consider two independent Poisson random variables X and Y and their difference Z = X – Y. Then Z is a discrete random variable defined on the set of {…, -2,-1,0,1,2,…}. In our case, Z is the random variable describing goal differences of individual games.

The distribution of Z (called the Skellam’s distribution or Poisson difference distribution) was discussed by Skellam in 1946. We say that Z follows Skellam’s distribution with parameters and (these are Poisson rates of X and Y, respectively) if X and Y are independent Poisson random variables with different means (. In notation, with density function:

(3.1)

where and is the modified Bessel function of order r.

For each individual game i, we have the goal difference as

where i = 1,2,…,n and n is the number of games. For the Poisson parameters , the bivariate Poisson model could be used (see Lee, 1997; Karlis and Ntzoufras, 2003):

(3.2)

(3.3)

where *μ* is a fixed constant, *H* is the home advantage factor, and are attacking and defending coefficients of team K. For each game i, depends on the attacking coefficient of the home team and the defending coefficient of the away team; the order is reversed for .

For model interpretability, sum to zero constraints are made to and . In other words:

(3.4)

where *K* is the number of teams in the tournaments. In our case *K* = 4 because for a general ECL group, there will be 4 teams competing in it.

All the parameters in this model are easy to interpret and carry certain meanings. *H* reflects home advantage and is the expected goal difference if both teams have the same defense and offense abilities. and measure a team’s attacking and defending abilities comparing to average performance. It can be easily seen that positive indicates above-average offense and larger means better offense. On the other hand, positive indicates below-average defense and larger means poorer defense.

3.2 Parameter Inference

3.3 Predicting future games

4 Application: Discussing rules of ECL group stage

4.1 Model Assessment

4.2 Number of points per win

4.3 Choice of tiebreakers

5 Discussions

5.1 Short comes of model

Tails; group of seeds; lack of additional factors

5.2 Future developments

Bayesian approach with informative prior;

References