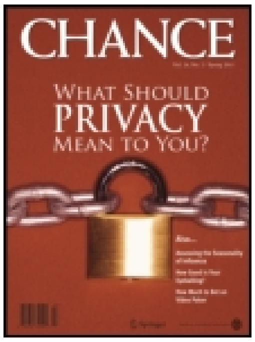

CHANCE

ISSN: 0933-2480 (Print) 1867-2280 (Online) Journal homepage: http://www.tandfonline.com/loi/ucha20

Modeling Scores in the Premier League: Is Manchester United Really the Best?

Alan J. Lee

Modeling Scores in the Premier League: Is Manchester United *Really* the Best?

Alan J. Lee

In the United Kingdom, Association football (soccer) is the major winter professional sport, and the Football Association is the equivalent of the National Football League in the United States. The competition is organized into divisions, with the Premier League comprising the best clubs. There are 20 teams in the league. In the course of the season, every team plays every other team exactly twice. Simple arithmetic shows that there are 380 = 20 × 19 games in the season. A win gets a team three points and a draw one point. In the 1995/1996 season, Manchester United won the competition with a total of 82 points. Did they deserve to win?

On one level, clearly Manchester United deserved to win because it played every team twice and got the most points. But some of the teams are very evenly matched, and some games are very close, with the outcome being essentially due to chance. A lucky goal or an unfortunate error may decide the game.

The situation is similar to a game of roulette. Suppose a player wins a bet on odds/evens. This event alone does not convince us that the player is more likely to win (is a better team) than the house. Rather, it is the long-run advantage expressed as a probability that is important, and this favors the house, not the player. In a similar way, the team that deserves to win the Premier League could be thought of as the team that has the highest probability of winning. This is not necessarily the same as the team that actually won.

How can we calculate the probability that a given team will win the Premier League? One way of doing this is to consider the likely outcome when two teams compete. For example, when Manchester United plays, what is the probability that it will win? That there will be a draw? Clearly these probabilities will depend on which team Manchester United is playing and also on whether the game is at home or away. (There are no doubt many other pertinent factors, but we shall ignore them.)

If we knew these probabilities for every possible pair of teams in the league, we could in principle calculate the probability that a given team will "top the table." This is an enormous calculation, however, if we want an exact result. A much simpler alternative is to use simulation to estimate this probability to any desired degree of accuracy. In essence, we can simulate as many seasons as we wish and estimate the "top the table" probability by the proportion of the simulated seasons that Manchester United wins. We can then rate the teams by ranking their estimated probabilities of winning the competition.

The Data

The first step in this program is to gather some data. The Internet is a good source of sports data in machine-readable form. The Web site http://dspace.dial.pipex.com/r-johnson/home.html has complete scores of all 380 games played in the 95/96 season, along with home and away information.

Modeling the Scores

Let's start by modeling the distribution of scores for two teams, say Manchester United playing Arsenal at home. We will assume that the number of goals scored by the home team (Manchester United) has a Poisson distribution with a mean λ_{HOME}. Similarly, we will assume that the number of goals scored by the away team (Arsenal) also has a Poisson distribution, but with a different mean λ_{AWAY}. Finally, we will assume that the two scores are independent so that the number of goals scored by the home team doesn't affect the distribution of the away team's score.

This last assumption might seem a bit far-fetched. If we cross-tabulate the home and away scores for all 380 games (not just games between Manchester U and Arsenal), however, we get the following table:

		Home team score				
		0	1	2	3	4+
	0	27	29	10	8	2
Away	1	59	53	14	12	4
team	2	28	32	14	12	4
score	3	19	14	7	4	1
	4+	7	8	10	2	0

A standard statistical test, the χ^2 test, shows that there is no evidence against the assumption of independence (χ^2 = 8.6993 on 16 df, p = .28). Accordingly, we will assume independence in our model.

The next step is to model the distribution of the home team's score. This should depend on the following factors:

Using Poisson Regression to Model Team Scores

We will assume that the score X of a particular team in a particular game has a Poisson distribution so that

$$Pr[X = x] = \frac{e^{-\lambda}\lambda^x}{x!}$$

We want the mean λ of this distribution to reflect the strength of the team, the quality of the opposition, and the home advantage, if it applies. One way of doing this is to express the logarithm of each mean to be a linear combination of the factors. This neatly builds in the requirement that the mean of the Poisson has to be positive. Our equation for the logarithm of the mean of the home team is (say, when Manchester U plays Arsenal at home)

$$\log(\lambda_{HOME}) = \beta + \beta_{HOME} + \beta_{OFFENSE}(\text{Manchester U}) + \beta_{DEFENSE}(\text{Arsenal})$$

Similarly, to model the score of the away team, Arsenal, we assume the log of the mean is

$$\log(\lambda_{AWAY}) = \beta + \beta_{OFFENSE}(\text{Arsenal}) + \beta_{DEFENSE}(\text{Manchester U})$$

We have expressed these mean scores λ_{HOME} and λ_{AWAY} in terms of "parameters," which can be interpreted as follows. First, there is an overall constant β, which expresses the average score in a game, then a parameter β_{HOME}, which measures the home-team advantage. Next comes a series of parameters $\beta_{OFFENSE}$, one for each team, that measure the offensive power of the team. Finally, there is a set of parameters $\beta_{DEFENSE}$, again one for each team, that measures the strength of the defense.

The model just described is called a generalized linear model in the theory of statistics. Such models have been intensively studied in the statistical literature. We can estimate the values of these parameters, assuming independent Poisson distributions, by using the method of maximum likelihood. The actual calculations can be done using a standard statistical computer package. We used S-Plus for our calculations.

The parameters calculated by S-Plus are shown in Table 2, and they allow us to compute the distribution of the joint score for any combination of teams home and away. For example, if Manchester U plays Arsenal at home, the probability that Manchester scores h goals and Arsenal scores a goals is

$$\frac{e^{-\lambda_{HOME}}\lambda_{HOME}^h}{h!} \times \frac{e^{-\lambda_{AWAY}}\lambda_{AWAY}^a}{a!}$$

where λ_{HOME} and λ_{AWAY} are given by

$$\begin{aligned}\lambda_{HOME} &= \exp(\beta + \beta_{HOME} + \beta_{OFFENSE}(\text{Manchester U}) + \beta_{DEFENSE}(\text{Arsenal}))\\ &= \exp(.0165 + .3518 + .4041 - .4075)\\ &= \exp(.0165) \times \exp(.3518) \times \exp(.4041) \times \exp(-.4075)\\ &= 1.4405\end{aligned}$$

and

$$\begin{aligned}\lambda_{AWAY} &= \exp(\beta + \beta_{OFFENSE}(\text{Arsenal}) + \beta_{DEFENSE}(\text{Manchester U}))\\ &= \exp(.0165 + .0014 - .2921)\\ &= \exp(.0165) \times \exp(.0014) \times \exp(-.2921)\\ &= .7602\end{aligned}$$

Thus, if Manchester U played Arsenal at Manchester many times, on average Manchester U would score 1.44 goals and Arsenal .76 goals. To calculate the probability of a home-side win, we simply total the probabilities of all combination of scores (h,a) with $h > a$. Similarly, to calculate the probability of a draw, we just total all the probabilities of scores where $h = a$ and, for a loss, where $h < a$. A selection of these probabilities are shown in Table 3.

- How potent is the offense of the home team? We expect Manchester U to get more goals than Bolton Wanderers, at the bottom of the table.
- How good is the away team's defense? A good opponent will not allow the home team to score so many goals.
- How important is the home-ground advantage?

We can study how these factors contribute to a team's score against a particular opponent by fitting a statistical regression model, which includes an intercept to measure the average score across all teams, both home and away, a term to measure the offensive capability of the team, a term to measure the defensive capability of the opposition, and finally an indicator for home or away. A similar model is used for the mean score of the away team.

These models are Poisson regression models, which are special cases of *generalized linear models*. The Poisson regression model is described in more detail in the sidebar.

Data Analysis

Before we fit the Poisson regression model, let us calculate some averages that shed light on the home-ground advantage, the strength of the team, and the strength of the opposition. First, if we average the "home" scores in each of the 380 games, we get a mean of 1.53 goals per game. The corresponding figure for the "away" scores is 1.07, so the home-team advantage is about .46 goals per game—a significant advantage.

What about the offensive strength of each team? We can measure this in a crude way by calculating the average number of goals scored per game by each team. Admittedly, this takes no account of who played whom. Similarly, we can evaluate the defensive strength of each team by calculating the number of goals scored against each team. These values are given in Table 1. We see that Manchester United has the best offense, but Arsenal has the best defense.

Table 1—Average Goals for and Against

Team	Average goals for	Average goals against	Team record (W	L	D)	Competition points
Arsenal	1.29	.84	17	9	12	63
Aston Villa	1.37	.92	18	11	9	63
Blackburn R.	1.61	1.24	18	13	7	61
Bolton Wan.	1.03	1.87	18	25	5	29
Chelsea	1.21	1.16	12	12	14	50
Coventry C.	1.11	1.58	8	16	14	38
Everton	1.68	1.16	17	11	10	61
Leeds U.	1.05	1.50	12	19	7	43
Liverpool	1.84	.89	20	7	11	71
Man. City	.87	1.53	9	18	11	38
Man. U.	1.92	.92	25	6	7	82
Middlesbro	.92	1.32	11	17	10	43
Newcastle U.	1.74	.97	24	8	6	78
Nottm. Forest	1.32	1.42	15	10	13	58
QPR	1.00	1.50	9	23	6	33
Sheff. Wed.	1.26	1.61	10	18	10	40
Southampton	.89	1.37	9	18	11	38
Tottenham H.	1.32	1.00	16	9	13	61
West Ham. U.	1.13	1.37	14	15	9	51
Wimbledon	1.45	1.84	10	17	11	41

Table 2—Team and Opposition Parameters From Fitting the Generalized Linear Model

Team	Offensive parameter	Offensive multiplier	Defensive parameter	Defensive multiplier
Arsenal	.00	1.00	-.41	.67
Aston Villa	.06	1.07	-.31	.73
Blackburn R.	.24	1.27	-.01	.99
Bolton Wan.	-.19	.83	.38	1.46
Chelsea	-.05	.95	-.09	.91
Coventry C.	-.12	.88	.22	1.24
Everton	.28	1.33	-.07	.93
Leeds U.	-.18	.84	.16	1.18
Liverpool	.36	1.43	-.32	.72
Man. City	-.37	.69	.17	1.19
Man. U.	.40	1.50	-.29	.75
Middlesbro	-.32	.73	.02	1.03
Newcastle U.	.31	1.36	-.24	.78
Nottm. Forest	.05	1.05	.12	1.13
QPR	-.23	.80	.16	1.17
Sheff. Wed.	.01	1.01	.24	1.27
Southampton	-.34	.71	.06	1.07
Tottenham H.	.03	1.03	-.23	.79
West Ham. U.	-.11	.90	.07	1.08
Wimbledon	.16	1.17	.38	1.47

Now we "fit the model" and estimate the parameters. The intercept is .0165, and the home-team advantage parameter is .3518. The first value means that a "typical" away team will score 1.0166 (= $e^{.0165}$) goals, and the second means that, on average, the home team can expect to score 100 × $e^{.3518}$ = 142% of the goals scored by their opposition. This agrees with the preceding crude estimate; 1.5263 is 142% of 1.0737.

Next we come to the offensive and defensive parameters. The estimates of these are contained in Table 2. We see that Manchester United has the largest offensive parameter (.4041) and Arsenal the smallest defensive parameter (− .4075), which is consistent with the preceding preliminary analysis. To get the expected score for a team, we multiply the "typical away team" score (1.0166) by the offensive multiplier and by the defensive multiplier. In addition, if the team is playing at home, we multiply by 1.4216 (= $e^{.3518}$). Note that these parameters are relative rather than absolute: The average of the offensive and defensive parameters has been arbitrarily set to 0 and the "typical team" parameter adjusted accordingly.

What do we get from this more complicated analysis that we didn't get from the simple calculation of means? First, the model neatly accounts for the offensive and defensive strengths of both the home team and the opposition. In addition, using the model, we can calculate the chance of getting any particular score for any pair of teams. In particular, the model gives us the probability of a win, a loss, or a draw.

The results in Tables 1 and 2 are in agreement, giving the same orderings for offense and defense. This is a consequence of every team playing every other team the same number of times.

If we perform the calculations described in the sidebar on page 18, we can calculate the probability of win, lose, and draw for any pair of teams, home and away. For example, Table 3 gives these probabilities for the top few teams. To continue our example, we see from these tables that when Manchester United plays Arsenal at Manchester, they will win with probability .53, draw with probability .27, and lose with probability .20.

Table 3—Probabilities of a Win, Draw, or Loss for Selected Match-ups

Home team	Away team	Prob. of win	Prob. of draw	Prob. of loss
Man. U.	Liverpool	.48	.25	.26
Liverpool	Man. U.	.47	.25	.27
Man. U.	Newcastle U.	.53	.24	.23
Newcastle U.	Man. U.	.44	.26	.30
Newcastle U.	Liverpool	.43	.26	.31
Liverpool	Newcastle	.52	.25	.23
Man. U.	Arsenal	.53	.27	.20
Arsenal	Man. U.	.37	.30	.33
Arsenal	Liverpool	.37	.30	.33
Liverpool	Arsenal	.52	.28	.20
Arsenal	Newcastle U.	.41	.30	.29
Newcastle U.	Arsenal	.49	.29	.22

Table 4—Results From Simulating the Season

Team	Actual points 95/96	Poisson model expected points	Simulated mean points	Simulated std. dev. points	Proportion at top of table
Man. U.	82	75.7	75.5	7.1	.38
Newcastle U.	78	70.7	70.5	7.8	.16
Liverpool	71	74.9	74.9	7.5	.33
Arsenal	63	63.8	63.6	7.7	.03
Aston Villa	63	63.7	63.6	7.4	.03
Blackburn R.	61	61.2	61.4	7.4	.03
Everton	61	64.9	65.0	7.5	.04
Tottenham H.	61	60.2	60.8	7.5	.01
Nottm. Forest	58	50.0	49.5	7.4	.00
West Ham. U.	51	46.3	46.1	7.7	.00
Chelsea	50	53.4	53.5	7.4	.00
Leeds U.	43	41.4	41.4	7.4	.00
Middlesbro	43	41.5	41.8	7.4	.00
Wimbledon	41	44.7	44.7	7.6	.00
Sheff. Wed.	40	44.8	44.9	7.2	.00
Coventry C.	38	41.2	41.4	7.6	.00
Man. City	38	35.7	35.4	6.9	.00
Southampton	38	39.6	39.5	7.0	.00
QPR	33	39.9	40.1	7.3	.00
Bolton Wan.	29	33.9	34.0	7.2	.00

Simulating the Season

Now we can approach the problem of whether or not Manchester United *was* lucky to top the table in the 95/96 season. As we noted previously, the Poisson regression approach allows us to calculate the chance of a win, loss, or draw for a game between any pair of teams. In principle, this allows us to calculate exactly the chance a given team will top the table. The calculation is too large to be practical, however, so we resort instead to simulation.

For each of the 380 games played, we can simulate the outcome of each game. Essentially, for each game, we throw a three-sided die (conceptually only) whose faces are win, lose, and draw. The probabilities of these three outcomes are similar to those given in the preceding tables. From these 380 simulated games, we can calculate the points table for the season, awarding three points for a win and one for a draw, and see which team topped the table.

In fact we used a computer program to simulate the 95/96 season 1,000 times. We can calculate the mean and standard deviation of 1,000 simulated points totals for each team and also the expected number of points under the Poisson model described previously. We can also count the proportion of times each team topped the table in the 1,000 simulated seasons, which gives an estimate of the probability of topping the table. Table 4 gives this information.

Manchester seems to have been a little lucky, but it still has the highest average score. Liverpool was definitely unlucky and according to our model is really a better team than Newcastle United, who actually came second.

Of course, our approach to modeling the scores is a little simplistic. We have taken no account of the fact that teams differ from game to game due to injuries, trades, and suspensions. In addition, we are assuming that our model leads to reasonable probabilities for winning/losing/drawing games. Teams that tend to "run up the score" against weak opponents may be overrated by a model that looks only at scores, and teams that settle into a "defensive shell" once they have got the lead may be underrated. Still, our results do seem to correspond fairly well to the historical result of the 95/96 season.

References and Further Reading

Groeneveld, R. A. (1990), "Ranking Teams in a League With Two Divisions of *t* Teams," *The American Statistician*, 44, 277–281.

Hill, I. D. (1974), "Association Football and Statistical Inference," *Applied Statistics*, 23, 203–208.

Keller, J. B. (1994), "A Characterization of the Poisson Distribution and the Probability of Winning a Game," *The American Statistician*, 48, 294–298.

McCullagh, P., and Nelder, J. A. (1989), *Generalised Linear Models*, London: Chapman and Hall.

Schwertman N. C., McCready, T. A., and Howard, L. (1991), "Probability Models for the NCAA Basketball Tournaments," *The American Statistician*, 45,179–183.

Stern, H. S. (1995), "Who's Number 1 in College Football? . . . And How Might We Decide?," *Chance*, 8(3), 7–14.