APPLICATIONES MATHEMATICAE
28,1 (2001), pp. 55–72

Bolesław Kopociński (Wrocław)

COMPONENTS OF THE GAME RESULT IN A FOOTBALL LEAGUE

Abstract. We assume that the result of a football game depends upon the difference of the strengths of the teams, home-field advantage, random factors and also other components. We describe the goal outcome per game by independent Poisson random variables; we concentrate on expected values. The least squares estimators of the parameters are obtained. The study is illustrated by examples from the Italian and Polish leagues.

1. Introduction. The 1998 Soccer World Cup concentrated the attention of many people. The result of a football (soccer) game is in considerable degree a reflection of the "strengths" of the playing teams. Starting from this supposition and using famous rankings of teams, experts forecast the results of elimination groups and publish their predictions (on the Internet: Magne Aldrin, Norwegian Computing Center). In the description of a football league, important factors seem to be the home-field advantage and a random element. The purpose of the paper is to separate these components using diverse data: the results of a league season or, in the extreme case, the league table only. It is not our intention here to define the best strengths. It is reasonable to assume that the point table reflects the distribution of strengths. So we assume the knowledge of points in the final table of one season and use these as the strengths of the teams. The strength and home-field advantage were considered by Glickman and Stern [3] when analysing the differences of points scored in the American National Football League (NFL). The competitions of the Soccer World Cup have a limited element of home-field advantage, but the strengths and the random component remain.

2000 *Mathematics Subject Classification*: 62F10, 62P99.

Key words and phrases: football (soccer) league, strength of team, home-field advantage, style of game, retaliation, least squares estimation.

In the description of a football league we make considerable theoretical simplifications and assume a limited knowledge of outcomes of one season of competitions. We assume that the numbers of goals scored by teams in one game are independent Poisson random variables. In details we concentrate on the expected values of these variables. We assume that the strengths of teams are constant during each season. Note that Keller [4] assumed this for games of England, Ireland, Scotland and Wales in the wide 1883–1980 period. In the supplement we indicate additional components of the game result: the mutual dependence of a game–return pair and the question of game style (imposing the defensive or offensive style on the opponent).

Let us recall some football terminology and introduce the notation. Assume that in a league there are n teams, and their strengths are denoted by $m = (m_1, \ldots, m_n)$. The goal outcome of a game between team i and team j with strengths m_i and m_j is denoted by X_{ij}, Y_{ij} (the home team placed first). Define K to be the number of points gained by a winning team. In the past, $K = 2$ points were given to a winning team and 1 point to both teams in a tie game. Now there are usually $K = 3$ points for a win and 1 point for a tie. Denote by U_{ij}, V_{ij} the point outcome of a game between teams i, j and define

$$U_{ij} = K\,\mathbf{1}_{X_{ij}>Y_{ij}} + \mathbf{1}_{X_{ij}=Y_{ij}}, \qquad V_{ij} = K\,\mathbf{1}_{X_{ij}<Y_{ij}} + \mathbf{1}_{X_{ij}=Y_{ij}}. \tag{1}$$

The point outcome of team i in the whole season is

$$U_i = \sum_{j=1,\, j\neq i}^{n} (U_{ij} + V_{ji}), \qquad 1 \le i \le n. \tag{2}$$

The random variables (2) are mutually dependent. For example, if $K = 2$, then $V_{ij} = 2 - U_{ij}$ and we have

$$U_1 = U_{12} + U_{13} + \ldots + U_{1n} + V_{21} + V_{31} + \ldots + V_{n1},$$
$$U_2 = U_{21} + U_{23} + \ldots + U_{2n} + V_{12} + V_{32} + \ldots + V_{n2},$$

thus $\mathrm{Cov}(U_1, U_2) = -\mathrm{Var}(U_{12}) - \mathrm{Var}(U_{21})$ is negative.

The position of team i in the final table is given by the range of U_i in the sequence (2). The goals scored give the additional classification which is used if the numbers of points are equal. Here the additional classification is not essential but the difference of real and expected numbers of goals in the final table will illustrate the efficiency of the estimation.

We illustrate our considerations by examples from the Italian and Polish leagues. In particular, we consider past and present-day leagues in which the number of teams, the parameter K and the mean number of goals scored per game differ. The statistical data used in the paper come from [1] and [2]. Fragments of text concerning examples are enclosed within □ □.

We easily find analogies to the football problems in other sport disciplines and also in other areas. In the medical problem of carcinogenesis one considers the share of genetic and environmental components. In the psychological problem of the level of human intelligence the heredity and education components are considered. However the main components considered here are rather specific to sport.

2. Components of the game result. Let $f(k)$, $1 \le k \le n$, denote the point outcome for a team which occupied the kth place in the final table of a league season. If we assume that a high-placed team always beats a low-placed one, then the number of points in the final table is the linear function $f_A(k) = 2(n-k)K$. If a home team always wins, then the number of points in the final table is constant: $f_B(k) = (n-1)K$.

Now consider the result of each game to be purely random with any probabilities of a win by the home team, defeat and a tie game. Let $K = 2$. For every pair of teams i, j we introduce independent identically distributed random variables U_{ij} such that $P(U_{ij} = 2) = p > 0$, $P(U_{ij} = 1) = r \ge 0$, $p + r < 1$ $(1 \le i, j \le n,\ i \ne j)$. The expected number of points under random results was established by simulation for $p = 0.467$, $r = 0.327$ and $1 - p - r = 0.206$ (which corresponds to the Italian league in the 1954/55 season). The point scores in the final table are evidently unequal but the differences are smaller than in reality.

"Pure" models of a league generated only by the strength of teams, the home-field advantage and random effects give league tables of different forms. This permits the estimation of these components if we consider them together.

2.1. *Poisson model of league.* Assume that the strength vector m is known and that the goal outcome of a game depends upon: the difference of the strengths of the teams, the home-field advantage and a random factor. Formally, assume that the following formulas hold:

$$
\begin{aligned}
X_{ij} &= \Pi^{(1)}(a_1(r_{ij})) + \Pi^{(3)}(b(m_i)) + \Pi^{(4)}(c_1),\\
Y_{ij} &= \Pi^{(2)}(a_2(r_{ji})) + \Pi^{(5)}(c_2),
\end{aligned}
\tag{3}
$$

where $r_{ij} = \max(m_i - m_j, 0)$ is called the *strength index*; $\Pi(\lambda)$ denotes a Poisson random variable with parameter λ; the random variables $\Pi^{(1)}$ to $\Pi^{(5)}$ are mutually independent. Here a_1, a_2, b are known functions and c_1, c_2 are the model parameters. We will specify these functions later on.

We interpret the random variables $\Pi^{(1)}$, $\Pi^{(2)}$ as the components of the game result which are due to the difference in team strengths; $\Pi^{(3)}$ (strictly speaking its expected value) as home-field advantage; and $\Pi^{(4)}$, $\Pi^{(5)}$ as random factors. The properties of independent Poisson random variables

imply the existence of random variables Π_1, Π_2 such that

$$X_{ij} \stackrel{d}{=} \Pi_1(a_1(r_{ij}) + b(m_i) + c_1), \qquad Y_{ij} \stackrel{d}{=} \Pi_2(a_2(r_{ji}) + c_2). \tag{4}$$

In practice the model functions must be more precise. It is natural to assume that the expectations of $\Pi^{(1)}$, $\Pi^{(2)}$ are proportional to the strength index, the expectation of $\Pi^{(3)}$ is proportional to the strength of the home team, and the random elements have an identical expected value for each team. The data of one football season are not extensive, therefore a limitation of the number of parameters is desired. In what follows, in the Poisson model we make the following

ASSUMPTION. *The strength vector m is known, the model functions are as follows*:

$$a_1(r_{ij}) = ar_{ij}, \qquad a_2(r_{ji}) = ar_{ji}, \qquad b(m_i) = bm_i, \qquad c_1 = c_2 = c, \tag{5}$$

and the vector of the model parameters is $(a, b, c) = p$.

Under the assumption (5) the goal outcome of a game of teams i, j is a pair of independent Poisson random variables X_{ij}, Y_{ij} with expected values

$$\mathrm{E}(X_{ij}) = ar_{ij} + bm_i + c, \qquad \mathrm{E}(Y_{ij}) = ar_{ji} + c, \qquad 1 \le i, j \le n,\ i \ne j. \tag{6}$$

3. Problem of parameter estimation. We consider three variants of available data. First we consider the goal results for games of one league season. In the second case, besides goals, for each game we also consider the date of the match. In the third case we consider the final table of points only. Let $D = \{(x_{ij}, y_{ij}) : 1 \le i, j \le n,\ i \ne j\}$ denote the goal results of the games of one season. Obviously using (1) and (2) we can evaluate the league table.

The least squares method of estimation is applied for all variants of the problem. Note that other models and methods of analysis are possible. For example Lee [5] assumed that the strength index depends upon the available quotient of strengths. In this model the number of goals scored by team A in a game between team A and team B is Poisson distributed with parameter $\lambda(A, B) = (\text{constant})(\text{strength of team } A)(\text{strength of team } B)^{-1}$. Then log-linear models of mathematical statistics (see [6]) may be useful.

3.1. *Estimation of p from one season results.* Suppose that we estimate the components of a game result having D. Denote by $N = n(n-1)$ the number of games and let $1 \le k \le N$ index the games. For each k we introduce the following variables:

$m_1(k)$ — strength of the home team in the kth game,
$m_2(k)$ — strength of the guest,
$r_1(k)$ — strength index of the home team,

$r_2(k)$ — strength index of the guest,
$X(k)$ — goal outcome of the home team,
$Y(k)$ — number of goals lost by the home team.

If it does not cause confusion, we omit the index k and summation limits. Set $m_1 = m_1(k)$, $r = r(k) = m_1(k) - m_2(k)$, $r_1 = r_1(k) = \max(r(k), 0)$, $r_2 = r_2(k) = \max(0, -r(k))$, $X = X(k)$, $Y = Y(k)$. Let

$$A = \begin{pmatrix} \sum r^2 & \sum r_1 m_1 & \sum |r| \\ \sum r_1 m_1 & \sum m_1^2 & \sum m_1 \\ \sum |r| & \sum m_1 & 2N \end{pmatrix}, \tag{7}$$

$$B = (\textstyle\sum(r_1 X + r_2 Y) \quad \sum m_1 X \quad \sum(X + Y)). \tag{8}$$

THEOREM 1. *Under the assumption* (6) *the least squares method* (*LSM*) *estimator of* p *is given by*

$$\widehat{p} = BA^{-1}. \tag{9}$$

It is unbiased with covariance matrix

$$\operatorname{Cov}(\widehat{p}) = \sum_{k=1}^{N} (A^{-1} B_1(k)^T B_1(k) A^{-1} \operatorname{Var}(X(k)) + A^{-1} B_2(k) B_2(k)^T A^{-1} \operatorname{Var}(Y(k))), \tag{10}$$

where $B_1(k) = (r_1(k), m_1(k), 1)$, $B_2(k) = (r_2(k), 0, 1)$.

Proof. Define

$$L(p \,|\, X(k), Y(k), 1 \le k \le N) = \sum_{k=1}^{N} ((ar_1(k) + bm(k) + c - X(k))^2 + (ar_2(k) + c - Y(k))^2).$$

The condition $L = \min$ yields the system of linear equations $Ap^T = B^T$, hence $p^T = A^{-1}B^T$. Since $A^T = A$, the solution of the estimation problem is (9). The estimator $\widehat{p}$ is unbiased: $\mathrm{E}(\widehat{p}^T) = A^{-1}\mathrm{E}(B^T) = A^{-1}Ap^T = p$.

Using the notations of Theorem 1 we obtain

$$\widehat{p} = BA^{-1} = \sum_{k=1}^{N} (B_1(k) A^{-1} X(k) + B_2(k) A^{-1} Y(k)).$$

This estimator is a linear combination of vectors with coefficients being independent random variables. This yields (10). ∎

3.2. *Expected number of goals.* We evaluate the efficiency of estimation taking into account the expected number of goals in the final table. Denote

by X_i the goal outcome and by Y_i the number of goals lost for team i during the season. We have

$$X_i = \sum_{j=1,\, j\neq i}^{n} (X_{ij} + Y_{ji}), \qquad Y_i = \sum_{j=1,\, j\neq i}^{n} (Y_{ij} + X_{ji}), \qquad 1 \leq i \leq n.$$

In applications we use the chi-square statistic as the measure of discrepancy of the observed and expected values but omit the statistical inference because the random variables used are dependent.

THEOREM 2. *Under the assumptions* (5) *we have*

$$\mathrm{E}(X_i) = 2a \sum_{j=1}^{n} r_{ij} + (n-1)bm_i + 2(n-1)c, \tag{11}$$

$$\mathrm{E}(Y_i) = 2a \sum_{j=1}^{n} r_{ji} + b(n\overline{m}_1 - m_i) + 2(n-1)c; \tag{12}$$

the expected number of goals per game in the whole season is

$$\mu = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} \mathrm{E}(X_{ij}) = (2a+b)\overline{m}_1 - \frac{4a\widetilde{m}_1}{n-1} + 2c, \tag{13}$$

where

$$\overline{m}_1 = \frac{1}{n} \sum_{i=1}^{n} m_i, \qquad \widetilde{m}_1 = \frac{1}{n} \sum_{i=1}^{n} (i-1)m_{i,n},$$

and $m_{1,n} \geq m_{2,n} \geq \ldots \geq m_{n,n}$ *is the ordered sequence* $m_1, \ldots, m_n$.

Proof. The formulas (11) and (12) can be easily verified. To prove (13), assume without loss of generality that $m_i \geq m_j$ for $i > j$. Then $r_{ij} = m_i - m_j$ for $i > j$ and $r_{ij} = 0$ otherwise. Hence from (11) we have

$$\mathrm{E}\Big(\sum_{i=1}^{n} X_i\Big) = a \sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij} + Nb\overline{m}_1 + 2Nc.$$

Under the above additional assumption we have $m_i = m_{i,n}$, $1 \leq i \leq n$. Hence

$$\sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij} = 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} (m_i - m_j) = 2 \sum_{i=1}^{n} (n - 2i + 1)m_i = 2N\overline{m}_1 - 4\widetilde{m}_1.$$

Substituting the definition of $\overline{m}_1$ we obtain (13). ∎

Table 1. The 1938 Polish football league (constant strengths)

	Points		Goals scored		Goals lost	
Ruch Chorzów	27	23.7	57	58.3	35	33.2
Warta Poznań	21	20.4	58	43.2	38	34.9
Wisła Kraków	20	19.7	41	40.9	36	35.3
Polonia Warszawa	19	19.0	40	38.8	38	36.0
Pogoń Lwów	19	19.0	23	38.8	26	36.0
AKS Chorzów	18	18.3	42	37.1	30	37.1
Cracovia	18	18.3	37	37.1	42	37.1
Warszawianka	15	15.8	34	33.1	46	41.5
ŁKS Łódź	12	13.3	25	29.8	45	46.5
Śmigły Wilno	11	12.5	29	28.9	50	48.4
Fitting	$\chi^2 = 0.83$		$\chi^2 = 13.00$		$\chi^2 = 5.88$	

Model parameters: $\widehat{a} = 0.0996$, $\widehat{b} = 0.0796$, $\widehat{c} = 1.167$. Components of the goal result: due to difference in strength—$\widehat{a}\overline{r} = \widehat{a}\frac{1}{n(n-1)}\sum r_{ij} = 0.261$; due to home-field advantage—$\widehat{b}\overline{m}_1 = 1.433$.

□ The data concerning the Polish league in the 1938 season are taken from [1]. Table 1 shows the results of one season and the corresponding expected values following from the model. The estimates $\widehat{a}, \widehat{b}, \widehat{c}$ of the model are also given. As a final result we divide the goals scored in three parts: $\widehat{a}\overline{r}$, due to difference in strengths; $\widehat{a}\overline{m}_1$, due to home-field advantage; and $\widehat{c}$, due to random factors. Taking into account these components in the example we divide the mean number 2.861 of goals per game for a home team in the following way: effect of difference in strengths—0.261 goals; effect of home-field advantage—1.433 goals; effect of random factors—1.167 goals; and similarly we divide the number 1.428 of goals for a guest as the effect of difference in strengths—0.261 goals, and effect of random factors—1.167 goals. □

3.3. *Estimation of home-field advantage.* It is intuitively clear that the result of a game and the return enables the estimation of home-field advantage. Using notations of the Poisson model we now give two new estimators of b. The intuitive form is (15), and its improved version is (16). Let us introduce notations for moments of strengths:

$$\overline{m}_1 = \frac{1}{n}\sum_{i=1}^{n} m_i, \quad \overline{m}_2 = \frac{1}{n}\sum_{i=1}^{n} m_i^2. \tag{14}$$

THEOREM 3. *Let $D_{ij} = X_{ij} - Y_{ij}$ denote the difference of goals scored per game of teams i, j. Under the assumptions* (5)*, the following formulas give estimators of home-field advantage*:

$$\widehat{b}_1 = \frac{1}{N\overline{m}_1}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} D_{ij}, \tag{15}$$

$$\widehat{b}_2 = \frac{1}{c_0 n} \sum_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} c_{ij} D_{ij}, \tag{16}$$

where $c_0 = (n-2)^2 \overline{m}_2 + (3n^2 - 4n)\overline{m}_1^2$, $c_{ij} = (n-2)(m_i + m_j) + 2n\overline{m}_1$, $1 \le i, j \le n$. *The estimators are unbiased with variances*

$$\operatorname{Var}(\widehat{b}_1) = \left(\frac{1}{\overline{m}_1 N}\right)^2 \sum_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} \operatorname{Var}(D_{ij}), \tag{17}$$

$$\operatorname{Var}(\widehat{b}_2) = \left(\frac{1}{c_0 n}\right)^2 \sum_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} c_{ij}^2 \operatorname{Var}(D_{ij}), \tag{18}$$

where $\operatorname{Var}(D_{ij}) = \operatorname{Var}(X_{ij}) + \operatorname{Var}(Y_{ij}) = a|m_i - m_j| + bm_i + 2c$, $1 \le i, j \le n$.

Proof. From (6) we have $\mathrm{E}(D_{ij}) = ar_{ij} + bm_i - ar_{ji}$, and using the equality

$$\sum_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} r_{ij} = \sum_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} r_{ji}$$

we obtain

$$\mathrm{E}\Big(\sum_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} D_{ij}\Big) = bN\overline{m}_1.$$

This shows that (15) is an estimator for b, and also that it is unbiased.

Let $R_{ij} = D_{ij} + D_{ji}$ denote the sum of differences of goals in a game and in the return and let $R_i = \sum_{j=1,\, j\neq i}^{n} R_{ij}$. From (6) we have $\mathrm{E}(R_{ij}) = b(m_i + m_j)$, hence $\mathrm{E}(R_i) = b((n-2)m_i + n\overline{m}_1)$.

Then the LSM estimator of b based on $R_1, \ldots, R_n$ has the form

$$\widehat{b}_2 = \Big(\sum_{i=1}^{n}((n-2)m_i + n\overline{m}_1)^2\Big)^{-1} \sum_{i=1}^{n}((n-2)m_i + n\overline{m}_1)R_i = (nc_0)^{-1} \sum_{i=1}^{n} c_i R_i,$$

where $c_i = (n-1)m_i + n\overline{m}_1$. Hence

$$\widehat{b}_2 = (nc_0)^{-1} \sum_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} c_i(D_{ij} + D_{ji}) = (nc_0)^{-1} \sum_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} c_{ij} D_{ij}.$$

The estimators (15) and (16) are unbiased. They are linear combinations of independent random variables D_{ij}. This also gives the formulas for the variances. ∎

Table 2. Dispersions and covariance matrix of estimators for selected seasons of the Italian league

Season	Dispersion of estimator		
	Disp($\widehat{a}$)	Disp($\widehat{b}$)	Disp($\widehat{c}$)
1954/55	.0094	.0028	.0648
1969/70	.0087	.0031	.0602
1994/95	.0043	.0019	.0592
	Covariance matrix of $\widehat{p}$		
1954/55	1.000		
	−.078	1.000	
	−.415	−.547	1.000
1969/70	1.000		
	−.097	1.000	
	−.391	−.551	1.000
1994/95	1.000		
	−.095	1.000	
	−.404	−.514	1.000

□ Table 2 shows the dispersions of estimators (9) of a, b, c and their correlation matrix, calculated for selected seasons of the Italian league. We omit the results of estimation of home-field advantage from Theorem 3. While calculating the covariances the point strength estimators of a, b, c are used. The results support the opinion that the parameters of the Poisson model can be estimated from one season with usable precision. □

3.4. *League plays for points.* In the opinion of many football experts a league team plays for points (or for a place in the final table), not for goals. Note that in the football pools we often guess the point result of the game. Hence, for a game between teams i, j with strengths m_i, m_j it is interesting to anticipate the point result U_{ij}, V_{ij}. Consider the linear model

$$\mathrm{E}(U_{ij}) = \widetilde{a}r_{ij} + \widetilde{b}m_i + \widetilde{c}, \qquad \mathrm{E}(V_{ij}) = \widetilde{a}r_{ji} + \widetilde{c}.$$

It is easy to see that the LSM estimator of the parameter $\widetilde{p} = (\widetilde{a}, \widetilde{b}, \widetilde{c})$ has the form

$$\check{p} = \widetilde{B}A^{-1},$$

where A is of the form (7) and

$$\widetilde{B} = (\sum(r_{ij}U_{ij} + r_{ji}V_{ij}), \quad \sum m_i U_{ij} \quad \sum(U_{ij} + V_{ij})),$$

where the sum is over $1 \le i, j \le n$, $i \ne j$. We omit the covariance matrix of this estimator.

Table 3. The Italian football league (points, constant strengths)

Season 1954/55			Season 1969/70			Season 1994/95		
Milan	48	50.7	Cagliari	45	47.6	Juventus	73	78.4
Udinese	44	44.7	Inter	41	41.8	Lazio	63	63.1
Roma	41	40.5	Juventus	38	37.7	Parma	63	63.1
Bologna	40	39.2	Milan	36	35.1	Milan	60	58.9
Fiorentina	39	37.9	Fiorentina	36	35.1	Roma	59	57.6
Napoli	38	36.7	Napoli	31	29.5	Inter	52	48.9
Juventus	37	35.6	Torino	30	28.4	Napoli	51	47.7
Inter	36	34.5	Vicenza	29	27.5	Sampdoria	50	46.6
Sampdoria	34	32.6	Lazio	29	27.5	Cagliari	49	45.6
Torino	34	32.6	Bologna	28	26.7	Fiorentina	47	43.7
Genoa	31	30.1	Roma	28	26.7	Torino	45	41.9
Catania	30	29.4	Verona	26	25.4	Bari	44	41.1
Lazio	30	29.4	Sampdoria	24	24.3	Cremonese	41	38.9
Triestina	30	29.4	Brescia	20	22.4	Genoa	40	38.2
Atalanta	28	28.3	Palermo	20	22.4	Padova	40	38.2
Novara	28	28.3	Bari	19	22.1	Foggia	34	35.2
Spal	23	26.4				Reggiana	18	28.1
Pro Patria	21	25.7				Brescia	12	25.9
Fitting $\chi^2 = 1.86$			Fitting $\chi^2 = 1.65$			Fitting $\chi^2 = 13.51$		

The characteristics of the leagues are shown in Table 6.

□ Table 3 shows the result of estimations for selected seasons of the Italian league. The left columns show the points scored, and the right columns the expected ones. It may be observed that the main league teams scored fewer points than it follows from the model and there are overestimates in the lower area of the table. □

3.5. *Dynamics of strength.* The strength of teams varies during the season. At the beginning of a league season this is a topic of speculations, after each round it is reasonable to update it using the actual results.

Suppose that both the results and dates of all games of the season are known. Let us group the games in $N = 2(n-1)$ rounds consisting of $n/2$ games each (for simplicity we assume that n is even and that the league plays without changing the time table). Let $1 \le k \le 2(n-1)$ index rounds and let $1 \le l \le n/2$ index games in a round. We can assume that the strengths are random variables depending upon the number of the round. We arbitrarily assume that the initial strengths $m_i^{(1)}$ are known, and after each round we update them using the actual table of points $f_i^{(k)}$ in the following way:

$$m_i^{(k+1)} = m_i^{(k)} \frac{N-k-1}{N} + f_i^{(k)}, \quad 1 \le i \le n,\ 1 \le k \le N-1.$$

Let $(\boldsymbol{X}_k, \boldsymbol{Y}_k) = \{(X_{jl}, Y_{jl}) : 1 \le j \le k-1,\ 1 \le l \le n/2\}$ be the set of results of all games before the kth round. From (1) it follows that the

outcome table $f_i^{(k)}$, $1 \le i \le n$, as well as the strengths $m_i^{(k)}$ are random variables depending upon $(\boldsymbol{X}_k, \boldsymbol{Y}_k)$.

Assume that the game indexed by (k,l) is played by teams indexed by $i = i(k,l)$, $j = j(k,l)$ with strengths $m_i^{(k)}$, $m_j^{(k)}$ and the goal outcome is X_{kl}, Y_{kl}. For simplicity of notation we omit the indices and set

m_1 — strength of the home team in the game,
m_2 — strength of the guest,
$r_1 = \max(m_1 - m_2, 0)$ — strength index of the home team,
$r_2 = \max(m_2 - m_1, 0)$ — strength index of the guest,
X — goal outcome of the home team,
Y — number of goals lost by the home team.

THEOREM 4. *Assume that the random variables X, Y conditioned by $\boldsymbol{X}_k, \boldsymbol{Y}_k$ are mutually independent Poisson variables with* $\mathrm{E}(X \mid (\boldsymbol{X}_k, \boldsymbol{Y}_k)) = ar_1 + bm_1 + c$, $\mathrm{E}(Y \mid (\boldsymbol{X}_k, \boldsymbol{Y}_k)) = ar_2 + c$. *Then p has an estimator of the form* (9), *with covariance matrix* (10), *where A, B are given by* (7) *and* (8), *and the index k in* (10) *is replaced by* (k,l).

Table 4. The 1997/98 Polish football league (constant point strengths)

	Points		Goals scored		Goals lost	
ŁKS	66	60.8	52	55.8	23	33.6
Polonia	63	58.8	46	52.6	30	33.7
Wisła	61	57.4	50	50.5	30	34.0
Widzew	61	57.4	53	50.5	34	34.0
Legia	59	55.9	50	48.7	32	34.4
Ruch	55	53.0	48	45.1	39	35.5
Amica	50	49.2	38	41.0	31	37.1
Górnik	48	47.7	48	39.4	42	37.9
Odra	48	47.7	51	39.4	50	37.9
Lech	46	46.2	41	38.1	37	38.8
Stomil	45	45.4	38	37.5	45	39.3
GKS	43	43.9	37	36.3	33	40.5
Zagłębie	43	43.9	39	36.3	40	40.5
Pogoń	43	43.9	36	36.3	40	40.5
Petrochemia	38	40.1	28	34.3	54	44.2
Dyskobolia	29	33.6	30	31.0	55	51.2
KSZO	24	30.1	24	29.4	47	55.4
Raków	17	25.6	21	27.6	68	61.6
Fitting	$\chi^2 = 6.39$		$\chi^2 = 11.07$		$\chi^2 = 16.78$	

Model parameters: $\widehat{a} = 0.0258$, $\widehat{b} = 0.0123$, $\widehat{c} = 0.706$. Components of the goal result: due to difference in strength—$\widehat{a}\overline{r} = 0.199$; due to home-field advantage—$\widehat{a}\overline{m}_1 = 0.575$.

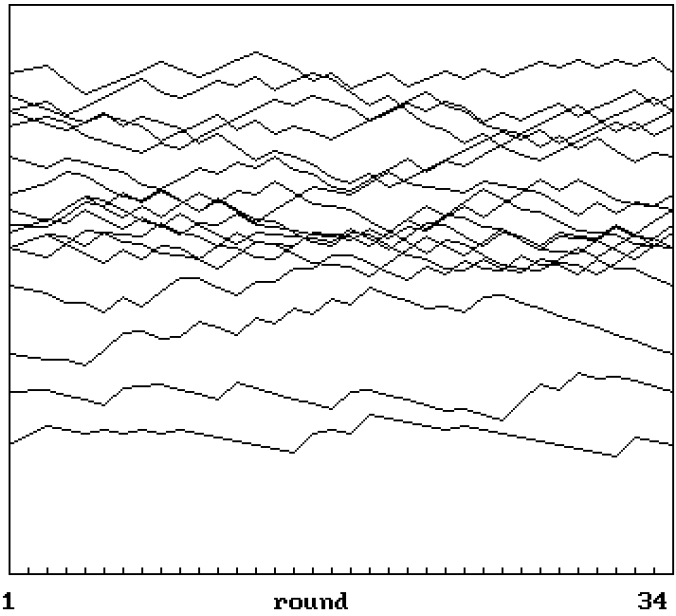

Fig. 1. The 1997/98 Polish football league. Variability of strengths during the season.

Table 5. The 1997/98 Polish football league (varied strengths)

	Points		Goals scored		Goals lost	
ŁKS	66	63.6	52	59.6	23	32.4
Polonia	63	58.6	46	52.0	30	33.4
Wisła	61	56.0	50	48.4	30	33.9
Widzew	61	59.2	53	52.5	34	32.9
Legia	59	58.4	50	51.2	32	33.1
Ruch	55	53.0	48	44.8	39	34.9
Amica	50	49.5	38	40.9	31	36.5
Górnik	48	47.5	48	39.0	42	37.7
Odra	48	46.4	51	38.1	50	38.4
Lech	46	44.8	41	36.8	37	39.2
Stomil	45	45.1	38	37.2	45	39.2
GKS	43	45.3	37	37.2	33	39.1
Zagłębie	43	43.2	39	35.8	40	41.0
Pogoń	43	44.1	36	36.3	40	39.9
Petrochemia	38	39.8	28	33.6	54	44.3
Dyskobolia	29	34.8	30	31.5	55	49.8
KSZO	24	27.8	24	28.3	47	58.7
Raków	17	23.7	21	26.8	68	65.5
Fitting	$\chi^2 = 4.71$		$\chi^2 = 12.40$		$\chi^2 = 15.95$	

Model parameters: $\widehat{a} = 0.0309$, $\widehat{b} = 0.0124$, $\widehat{c} = 0.677$. Components of the point result: due to difference in strength—$\widehat{a}\overline{r} = 0.231$; due to home-field advantage—$\widehat{a}\overline{m}_1 = 0.576$.

□ Table 4 shows the results for the Polish league of the 1997/98 season and the expected results given by the model under the assumption of constant point strengths. The model parameters and the expected components of the result in goals are also given.

Table 5 shows the result of estimation of the model parameters for the season considered in Table 4 under the assumption of changeable strength. The parameters and expected results are as above. Other modifications of the definition of strength are considered. A better forecast is given under the assumption of equal strengths for each team at the beginning of the season than assuming the strengths given by the table of the previous season. Figure 1 shows the variability of strength of teams in the whole season. In order to eliminate inessential drifts we assume that the initial and final strengths are equal. □

3.6. *Estimation of p from the table.* Suppose that we estimate the components of the game result under the limit data: having the final league table only. Let $p_k(\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}$, $k \geq 0$, $\lambda > 0$, denote the Poisson probabilities with parameter λ. In the Poisson model from (1), using the values (6) we get

$$\mathrm{E}(U_{ij}) = \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} p_k(r_{ij}a + m_i b + c)p_l(r_{ji}a + c)(K\mathbf{1}_{k>l} + \mathbf{1}_{k=l}).$$

Hence the expected point outcome for the ith team is a function u_i of p:

$$u_i(p) := \mathrm{E}(U_i) = \sum_{j=1,\, j\neq i}^{n} \mathrm{E}(U_{ij}).$$

We define

$$L^*(p \,|\, U_i, 1 \leq i \leq n) = \sum_{i=1}^{n} (u_i(p) - U_i)^2/u_i(p).$$

The condition $L^*(p \,|\, U_i, 1 \leq i \leq n) = \min$ gives the LSM estimator of p. The function L is very complex, so the calculations of the minimum are rather laborious. We get some simplification by assuming that the expected and empirical numbers of goals per game are equal.

4. Strength in the return match. Football experts devote much attention to "historical" remarks. In particular it is supposed that the return game has an element of retaliation: a defeat implies additional strength in the return. We describe the role of retaliation using the residuals of the LSM estimation.

For the game and return of teams i, j, consider the residuals $X_{ij} - \mathrm{E}(X_{ij})$, $Y_{ij} - \mathrm{E}(Y_{ij})$ and the difference Δ_{ij} of residuals:

$$\begin{aligned}\Delta_{ij} &= D_{ij} - a(r_{ij} - r_{ij}) - bm_i \\ &= D_{ij} - a(m_i - m_j) - bm_i, \qquad 1 \le i < j \le n,\end{aligned} \tag{19}$$

and also the difference $\Delta'_{ij} = \Delta_{ji}$ of the residuals in the return:

$$\Delta'_{ij} = D_{ji} - a(m_j - m_i) - bm_j.$$

Under the assumption (5) these random variables are centered and mutually independent. We say that the retaliation hypothesis is confirmed if the covariance of the residuals is positive. For the correlation test we define the following statistic:

$$\varrho = \frac{1}{2N} \sum_{i=1}^{n} \sum_{j=1, j<i}^{n} \Delta_{ij}\Delta'_{ij} (\operatorname{Var}(D_{ij}) \operatorname{Var}(D_{ji}))^{-1/2}.$$

□ A few seasons of the Italian league do not confirm the retaliation hypothesis. Taking into account the goal results of the whole season, assuming point strengths m_i and the LSM estimators of the parameters, we calculate ϱ for residuals (19). We get $\varrho = 0.012$ in 1954/55, $\varrho = 0.122$ in 1969/70, $\varrho = 0.040$ in 1994/95. These correlations are positive, but the differences from zero are not significant. □

5. Stable strength vector. The definition of team or player orderings, for example in tennis, chess and so on, is an interesting problem for experts and mathematicians. Obviously the result of competition is expected to be consistent with the defined order. Hence the proposed strengths in football should simulate the point table.

For the strength vector m and the data D let $\widehat{p} = P(m, D)$ denote the estimator of p. Having the strengths and parameters, using the model, we can anticipate each outcome, and from (1) and (2) we can create the expected final point table m'. It may play the role of the new strength vector. The vector m' is a function of m and $\widehat{p}$. We denote it by T:

$$m' = T(P(m, D), m).$$

We say that a strength vector m^* is *stable* for D if it satisfies the equation

$$m^* = T(P(m^*, D), m^*).$$

The equation does not guarantee the uniqueness of a stable strength. Some examples can be calculated using the possible convergence of the sequence

$$m_{t+1} = T(P(m_t, D), m_t), \qquad t \ge 0, \tag{20}$$

where m_0 is the initial strength, proposed in practice by experts.

□ Starting from the point strength for the Italian league the sequence (20) converges for the 1954/55 and 1969/70 seasons, while for the 1994/95 season

it is periodic. The stable (periodic) strengths do not differ much from the point strengths. We omit the numerical details. □

6. Supplements and generalizations. The assumption that home-field advantage is a linear function of the strength of the home team may be attractive. A natural modification of the model may be the differentiation of the mean number of random goals for the home and guest teams. Under the linear dependence in (6) this idea is included in the modification of home-field advantage given below.

6.1. *Modified home-field advantage.* Suppose that in the Poisson model the goal outcome of the game of teams i, j with strengths m_i, m_j equals

$$X_{ij} = \Pi_1(ar_{ij} + bm_i + b_0 + c), \qquad Y_{ij} = \Pi_2(ar_{ji} + c). \tag{21}$$

Comparing this with (4) we see here the additional parameter b_0.

Having the strength vector m and the data D we can estimate a, b, b_0, c using the least squares method. The Italian league shows that the assumption $b_0 = 0$, $b \neq 0$ describes the league better than the assumption $b_0 \neq 0$, $b = 0$. In general, one season data do not permit us to reject the hypothesis $b_0 = 0$ under $b \neq 0$.

It is obvious that the number of points scored in one season depends upon the number of teams. In the history of many leagues the numbers of teams are different, hence also the model parameters based on point strength are not comparable. This disadvantage may be eliminated by a standardization of strengths.

Consider a league with strength vector m and model (21) with parameters a, b, b_0, c. If m is linearly transformed to

$$m_i^* = Am_i + B, \qquad 1 \leq i \leq n,$$

then the model parameters are

$$a^* = a/A, \qquad b^* = b/A, \qquad b_0^* = b_0 + B/A, \qquad c^* = c.$$

In the standard league we take $n = 18$ and $K = 3$. If $n \neq 18$ then we use $A = 17 \cdot 18/(n(n-1))$. When $K \neq 2$, then we use $A = E(U_3)/\mathrm{E}(U_K)$ independently of the previous one, where $\mathrm{E}(U_K)$ is the mean number of points scored per game in the whole season, $\mathrm{E}(U_2) = 2$. We omit the details of examples.

6.2. *Style of game.* Experts often emphasize the importance of imposing one's own style of playing on one's opponent. In great simplification suppose that a style can be defensive or offensive. In the first case the team puts its whole effort to self-defense, in the other case the team responds to each attack by an attack. The Poisson model considered before is now

understood as being defensive. For the offensive style we assume that in the game between teams i, j with strengths m_i, m_j the goal outcome is

$$X_{ij} = \Pi_1(ar_{ij} + bm_i + c) + \Pi_3(d), \qquad Y_{ij} = \Pi_2(ar_{ji} + c) + \Pi_3(d),$$

where Π_1, Π_2 have the interpretation as in (4), and $\Pi_3(d)$ is a Poisson random variable with parameter d, independent of Π_1 and Π_2. We interpret Π_3 as the result of reciprocal attacks.

We reduce testing the hypothesis on imposing the offensive style to testing the parametric hypothesis H_0: $d = 0$ against H_1: $d > 0$.

□ The data concerning the Italian league, for example in the 1994/95 season, do not permit the rejection of the hypothesis H_0 under the alternative H_1. One may suppose that the element of style also appears in other games, for example in basketball, but this is beyond the scope of this note. □

7. Final remarks. The components of sport results described in this note are rather evident for experts, but the examples give usable conclusions, thanks to quantitative expressions of the effects considered.

Note that our data are always limited: we have the full results of the league round after round, composite results of games of one season, results of an incomplete season, or at least, the table of the league only. The model which uses strengths, home-field advantage and a random component gives in my opinion a quite good description of the expected league table. For the Italian and Polish leagues the established components are comparable. Note that the random component is large. This does not permit forecasting the result of one particular game, but it may be used to anticipate the result of group eliminations.

The Poisson model sometimes gives interesting details. If we estimate the components from the table, the point outcome of a champion is smaller than expected; the point outcome in the critical area of the table is larger than expected. We observe large deviations in the home-field advantage for the leagues considered.

The estimation of components which determine the outcome is not equivalent to the problem of forecasting a result. Having the expected number of goals per game, assuming Poisson variables X_{ij}, Y_{ij} for goals scored, we can calculate the probability of win $P(X_{ij} > Y_{ij})$, tie $P(X_{ij} = Y_{ij})$ and defeat $P(X_{ij} < Y_{ij})$. For many leagues we have $1 < \mathrm{E}(X_{ij}) < 2$, $1 < \mathrm{E}(Y_{ij}) < 2$, which implies that the most probable result is 1:1.

□ According to the assumptions (3) and (5) of the Poisson model in the game between teams i, j with strengths m_i, m_j the expected score is as follows: $\widehat{a}r_{ij}$ is the expected number of goals following from the strength index, $\widehat{b}m_i$ is the expected number goals following from home-field advantage and $\widehat{c}$ is

the expected number of incidental goals. Hence $a\overline{r}_1$ is the expected number of goals in the season due to the strength index, and $b\overline{m}_1$ is the expected number of goals due to home-field advantage.

Table 6. Model parameters in Polish and Italian leagues

Season	n	K	Goals	Points	$\widehat{a}$	$\widehat{b}$	$\widehat{c}$	$\widehat{a\overline{r}}$	$\widehat{b\overline{m}}_1$
LSM model for goals: estimation from league season									
I 1954/55	18	2	2.72	2	.0467	.0155	.9028	.193	.528
I 1969/70	16	2	1.94	2	.0441	.0134	.5727	.195	.402
I 1994/95	18	3	2.53	2.75	.0221	.0129	.7747	.188	.601
P 1938	10	2	4.29	2	.0996	.0796	1.167	.261	1.43
P 1997/98	18	3	2.39	2.75	.0258	.0123	.706	.199	.575
LSM model for points: estimation from league season									
I 1954/55	18	2			.0365	.0142	.6074		
I 1969/70	16	2			.0409	.0150	.5946	.181	.450
I 1994/95	18	3		2.75	.0366	.0172	.6593	.312	.806
Estimation from league table									
I 1969/70	16	2			.0473	.0206	.4495		
I 1994/95	18	3			.0216	.0244	.5101		
I 1996/97	18	3			.0269	.0170	.5890		

I – Italian league, P – Polish league, n – number of teams, K – number of points for winning team; goals – mean number of goals in the season; points – mean point outcome in the season; $\widehat{a}, \widehat{b}, \widehat{c}$ – estimated values of the model parameters; $\widehat{a\overline{r}}$ – expected number of goals per game following from the strength index; $\widehat{b\overline{m}}_1$ – expected number of goals per game following from home-field advantage. The parameters are not reduced to the standard league.

Table 6 gathers the results concerning the Italian league in 1954/55–1994/95 and the Polish league in 1938 and 1997/98. The number of teams in the league, the number of teams dropping to the lower class after the season and the parameter K vary. Also tendencies in football art change (see the mean number of goals scored per game). This improves the variation of the parameters in Table 6. The unexpected conclusion from this analysis is a relatively large expected number of goals due to the strength index, compared to the number of goals due to home-field advantage and to the expected number of random goals. □

References

[1] J. Jeleń *et al.*, *Liga gra po pięćdziesiątce*, Wyd. II, Wydawnictwo Sport i Turystyka, Warszawa, 1987 (in Polish).

[2] Encyklopedia Fuji, Rocznik 98-99, Tom 22, Wydawnictwo GiA, Katowice 1998 (in Polish).

[3] M. E. Glickman and H. S. Stern, *A state-space model for National Football League scores*, J. Amer. Statist. Assoc. 93 (1998), 25–35.
[4] J. B. Keller, *A characterization of the Poisson distribution and the probability of winning a game*, Amer. Statistician 48 (1994), 294–298.
[5] A. Lee, *Modeling scores in the Premier League*: *Is Manchester United really the best?*, Chance 10 (1997), 15–19.
[6] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd. ed., Chapman & Hall, 1989.

Mathematical Institute
University of Wrocław
Pl. Grunwaldzki 2/4
50-384 Wrocław, Poland
E-mail: ibk@math.uni.wroc.pl

Received on 15.2.2000;
revised version on 26.6.2000 (1528)