MATEMATYKA STOSOWANA 2, 2001

BOLESŁAW KOPOCIŃSKI (Wrocław)

Modelowanie ligi koszykówki

0. Modelowanie matematyczne w sporcie. Problematyka sportowa obfituje w interesujące przykłady modelowania matematycznego. Najczęściej przywołuje się w tym miejscu historyczne zagadnienie o podziale stawki w przerwanym turnieju, będące przykładem zastosowania teorii błądzenia losowego. Często modele są wspólne dla wielu dziedzin: techniki, ekonomii i przyrodoznawstwa. Znajdujemy więc wśród nich wielowymiarowy opis indywiduów przydatny przy wyborze zawodników do określonych dyscyplin, prawa wzrostu przydatne do analiz metod treningowych, optymalizację podejmowania decyzji deterministycznych i decyzji w warunkach niepewności.

Przykładem deterministycznej optymalizacji decyzji trenerskiej jest wybór składu drużyny na podstawie macierzy przydatności zawodników do określonych pozycji. Zagadnienie to oczywiście rozwiązuje się metodą programowania liniowego, zwaną przydziałem pracy. Uproszczenie rzeczywistości jest tu jednak daleko posunięte: badania szczegółowe wskazują na nieliniowość funkcji mocy drużyny w zależności od mocy zawodników.

Nieprofesjonalny miłośnik sportu dysponuje zazwyczaj bardzo ograniczonym zbiorem danych. Formułuje także proste pytania o wynik meczu, tabelę oczekiwaną turnieju oraz wpływ znanych mu werbalnie czynników determinujących wynik. Niektóre modele matematyczne dają odpowiedzi na te pytania. Problem szczupłości dostępnych danych jest wyzwaniem dla matematyków; tym bardziej interesujący może być wynik tych dociekań. I tak w zagadnieniu o podziale stawki w przerwanym turnieju nader skąpa informacja – dotychczasowy wynik – pozwala na daleko idący wniosek: umiarkowana przewaga jednego z zawodników daje mu znaczną przewagę w prognozowanym wyniku końcowym.

Podobnie ma się rzecz w turniejach gier zespołowych. Znane np. z rankingów moce drużyn, a także uwzględnienie atutu własnego boiska, pozwalają (przy zróżnicowaniu mocy) nieźle prognozować końcową tabelę. Glickman i Stern [2] przy użyciu tych pojęć analizują różnicę punktów osiąganych w pojedynkach ligi futbolu amerykańskiego NFL. Podobnie wielu autorów bada wyniki meczów piłki nożnej, np. Keller [3] modeluje rezultaty meczów Anglii, Walii, Szkocji i Irlandii (zob. także [6]). W [4] znaleziono oczekiwaną tabelę i prawdopodobnego mistrza w 1939 roku w polskiej lidze piłkarskiej, w sezonie przerwanym wybuchem wojny.

Podstawowym elementem modelowania matematycznego zagadnień realnego świata jest „rozsądne” uściślanie pojęć i estymacja parametrów modelu. Niżej Czytelnik znajdzie przykład realizacji tych przedsięwzięć.

1. Wprowadzenie. Przy modelowaniu rozgrywek ligi koszykówki zakładamy, że końcowy wynik meczu zależy od trzech czynników: mocy grających drużyn, atutu własnego boiska i przypadku. Te same czynniki były wzięte pod uwagę w [4], [5] do opisu sezonu ligi piłkarskiej. Tutaj pokażemy, że odpowiednio zdefiniowana moc drużyn koszykówki przenosi się na końcowy wynik inaczej niż w piłce nożnej, a ponadto występuje dodatkowa składowa wynikająca z częstej wymiany posiadania piłki. W koszykówce drużyny grają nie o możliwie dużą liczbę punktów, a o korzystną różnicę liczby punktów. Jeśli chodzi o moce drużyn, to obiecujące wydaje się rozważenie ich w dwóch składowych: w ataku i obronie. Celem pracy jest estymacja parametrów składowych wyniku meczu w lidze.

2. Model. Zakładamy, że w meczu drużyn i, j o mocach m_i, m_j wynik meczu zapisany w liczbach X_{ij}, Y_{ij} uzyskanych punktów jest postaci

$$
\begin{aligned}
X_{ij} &= N^{(1)}(\mu_{ij}^{(1)} + c_1, \sigma_1^2) + N^{(2)}(\mu_{ij}^{(2)} + c_2, \sigma_2^2) + N^{(3)}(\mu_{ij}^{(3)}, \sigma_3^2), \\
Y_{ij} &= N^{(4)}(\mu_{ji}^{(1)} + c_1, \sigma_1^2) + N^{(3)}(\mu_{ij}^{(3)}, \sigma_3^2).
\end{aligned} \tag{1}
$$

Przez $N(\mu, \sigma^2)$ oznaczamy zmienną losową o rozkładzie normalnym ze średnią μ i wariancją σ^2; zakładamy, że $N^{(1)}$–$N^{(4)}$ są wzajemnie niezależne. Analogiczny model był wprowadzony w [4], [5] do opisu ligi piłki nożnej, ale wówczas użyto zmiennych losowych o rozkładzie Poissona.

Zmienne losowe $N^{(1)}$–$N^{(4)}$ nazywamy składowymi wyniku meczu. Zmienną $N^{(1)}$ ($N^{(4)}$) interpretujemy jako wkład własny drużyny, tzn. tę część, która pochodzi od mocy danej drużyny lub nadwyżki mocy nad przeciwnikiem. Zmienną $N^{(2)}$ interpretujemy jako przypisany gospodarzowi atut własnego boiska. Zamiana posiadania piłki po zaliczonym koszu i zazwyczaj duża skuteczność gry w ataku sprawiają, że w końcowym wyniku na koncie obu drużyn może pojawić się pewna liczba punktów, którą obie drużyny wypracowały wspólnie przez wymianę wzajemnych ataków. Z punktu widzenia rezultatu końcowego jest to bezproduktywna część meczu, ale uatrakcyjnia ona przebieg meczu. Wartości oczekiwane $\mu^{(1)}$–$\mu^{(3)}$, będące funkcjami mocy drużyn, oraz stałe c_1, c_2 opisują deterministyczne składniki wyniku meczu, a wariancje – czynnik losowy.

Zmienna $N^{(3)}$ w pewnym sensie charakteryzuje styl gry w lidze. Jeśli w danym meczu przyjmuje ona wartość powyżej średniej, znaczy to, że drużyny preferują atak, zaniedbując obronę; jeśli przeciwnie, to skupiają się one na obronie, zaniedbując atak. Gdy chodzi o zwycięstwo w meczu, $N^{(3)}$ nie ma znaczenia: o zwycięstwie decyduje różnica punktów $D_{ij} = X_{ij} - Y_{ij}$, która od $N^{(3)}$ nie zależy.

Znajomość mocy drużyn m, funkcji $\mu^{(1)}$–$\mu^{(3)}$ i stałych c_1, c_2, σ_1^2–σ_3^2 pozwala prognozować wynik każdego meczu w lidze, a więc także oczekiwaną tabelę ligi. Z (1) wynikają równości według rozkładu

$$\begin{aligned} X_{ij} &\overset{d}{=} N_1(\mu_{ij}^{(1)} + \mu_{ij}^{(2)} + \mu_{ij}^{(3)} + c_1 + c_2, \sigma_1^2 + \sigma_2^2 + \sigma_3^2), \\ Y_{ij} &\overset{d}{=} N_2(\mu_{ji}^{(1)} + \mu_{ij}^{(3)} + c_1, \sigma_1^2 + \sigma_3^2), \end{aligned} \tag{2}$$

przy czym N_1, N_2 są dwuwymiarowo normalne.

Weźmy pod uwagę przewagę gospodarza D_{ij}. Z (1) mamy

$$D_{ij} \overset{d}{=} N(\mu_{ij}, \sigma^2), \qquad \mu_{ij} = \mu_{ij}^{(1)} - \mu_{ji}^{(1)} + \mu_{ij}^{(2)} + c_2, \qquad \sigma^2 = 2\sigma_1^2 + \sigma_2^2. \tag{3}$$

Zasadniczą kwestią naszego modelowania jest znalezienie wartości oczekiwanych $\mu_{ij}^{(1)}$–$\mu_{ij}^{(3)}$ w zależności od mocy drużyn.

3. Ogólny model liniowy. Przyjmijmy, że moc drużyny i jest s-wymiarowym wektorem $M_i = (m_i^{(1)}, \ldots, m_i^{(s)})^T$, gdzie ^{T} oznacza transpozycję wektora (macierzy). Niech w lidze gra n drużyn; biorąc pod uwagę mecz i rewanż, w sezonie mamy $N = n(n-1)$ meczów. Dalej mecze indeksujemy także pojedynczym indeksem k (kolejność meczów, a więc sposób liniowego uporządkowania zbioru $\{(i,j) : i \neq j,\ 1 \le i,\ j \le n\}$ jako $\{k : 1 \le k \le N\}$, nie ma tu znaczenia). Moce gospodarzy oznaczamy przez $M' = \{M'_k,\ 1 \le k \le N\}$, moce gości przez $M'' = \{M''_k,\ 1 \le k \le N\}$, punkty uzyskane przez gospodarza w meczu oznaczamy przez $X' = \{X'_k,\ 1 \le k \le N\}$, punkty gości przez $X'' = \{X''_k,\ 1 \le k \le N\}$. Niech $M = \{M', M''\}^T$, $X = \{X', X''\}^T$. Przez Y oznaczamy macierz $2N \times v$-wymiarową zmiennych opisujących drużyny, zawierającą jako kolumny moce M i określone funkcje mocy.

Zakładamy model liniowy dla liczby uzyskanych punktów:

$$X = YA + E, \tag{4}$$

gdzie $E = \{e_k, 1 \le k \le 2N\}^T$, e_k są niezależnymi zmiennymi losowymi normalnymi o wartości oczekiwanej zero i wariancji $\mathrm{Var}(e_k) = \sigma_1^2 + \sigma_2^2 + \sigma_3^2$ dla $1 \le k \le N$, $\mathrm{Var}(e_k) = \sigma_1^2 + \sigma_2^2$ dla $N+1 \le k \le 2N$; $A = \{a_1, \ldots, a_v\}^T$ jest v-wymiarowym wektorem parametrów modelu.

Niech $D = X' - X''$ oznacza wektor przewagi gospodarzy. Z (4) wynika model liniowy dla przewagi:

$$D = ZB + E^*, \tag{5}$$

gdzie $E^* = \{e_k^*, 1 \le k \le N\}^T$, e_k^* są niezależnymi zmiennymi losowymi $N(0, \sigma^2)$, $B = \{b_1, \ldots, b_w\}^T$ jest wektorem parametrów modelu przewagi, Z jest wykorzystującą moce $N \times w$-wymiarową macierzą zmiennych opisujących. Parametry wyestymowane w modelu przewagi mogą być wykorzystane przy estymacji w modelu ogólnym. Uzyskane w ten sposób rozwiązanie problemu estymacji traci ogólną optymalność, ale umożliwia jednolitą w obu podejściach interpretację wyników.

4. Estymacja parametrów i testowanie modelu. Do estymacji parametrów modeli wykorzystujemy wyniki meczów zasadniczej rundy jednego sezonu ligi. Naszym celem jest estymacja A, σ_1^2–σ_3^2 w (4) albo B, σ^2 w (5). Parametry A (B) estymujemy metodą najmniejszych kwadratów, natomiast σ_1^2–σ_3^2 estymujemy metodą momentów, wykorzystując reszty regresyjne, zaś σ^2 utożsamiamy z resztową wariancją regresyjną. W celu estymacji wariancji składowych wyniku meczu weźmy pod uwagę zmienne losowe

$$S_1^2 = \frac{1}{N}\sum_{i,j}(X_{ij} - \mathrm{E}(X_{ij}))^2, \qquad S_2^2 = \frac{1}{N}\sum_{i,j}(Y_{ij} - \mathrm{E}(Y_{ij}))^2,$$

$$S_3^2 = \frac{1}{N}\sum_{i,j}(D_{ij} - \mathrm{E}(D_{ij}))^2.$$

Z (2)–(3) znajdujemy wartości oczekiwane $\mathrm{E}(S_1^2) = \sigma_1^2 + \sigma_2^2 + \sigma_3^2$, $\mathrm{E}(S_2^2) = \sigma_1^2 + \sigma_3^2$, $\mathrm{E}(S_3^2) = 2\sigma_1^2 + \sigma_2^2$. Zatem estymatorami momentowymi σ_1^2–σ_3^2 są

$$\widehat{\sigma}_1^2 = \frac{1}{2}(S_2^2 + S_3^2 - S_1^2), \qquad \widehat{\sigma}_2^2 = S_1^2 - S_2^2, \qquad \widehat{\sigma}_3^2 = \frac{1}{2}(S_1^2 + S_2^2 - S_3^2),$$

natomiast estymatorem momentowym $\sigma^2 = 2\sigma_1^2 + \sigma_2^2$ jest

$$\widehat{\sigma}^2 = S_3^2.$$

4.1. *Tabela oczekiwana.* Liczby uzyskanych i straconych punktów dla drużyny i w sezonie ligi są następujące:

$$X_i = \sum_{j \ne i}(X_{ij} + Y_{ji}), \qquad Y_i = \sum_{j \ne i}(Y_{ij} + X_{ji}).$$

Niech $P_{ij} = \mathbf{1}_{X_{ij} > Y_{ij}}$. W meczu drużyn i, j zmienna losowa $1 + P_{ij}$ jest, zgodnie z regulaminem ligi, liczbą punktów meczowych gospodarza, zaś $1 + Q_{ij} = 2 - P_{ij}$ jest liczbą punktów meczowych gości. W sezonie liczba punktów drużyny i wynosi

$$P_i = 2(n-1) + \sum_{j \ne i}(P_{ij} + Q_{ji}).$$

4.2. *Testowanie modelu.* Zauważmy, że

$$\mathrm{Var}(X_i) = \mathrm{Var}(Y_i) = (n-1)\sigma_0^2, \quad \sigma_0^2 = 2\sigma_1^2 + \sigma_2^2 + 2\sigma_3^2,$$
$$\mathrm{Var}(P_i) = \sum_{j\neq i}(\mathrm{E}(P_{ij})\mathrm{E}(Q_{ij}) + \mathrm{E}(P_{ji})\mathrm{E}(Q_{ji})).$$

Zmienne losowe X_i, $1 \le i \le n$, są wzajemnie niezależne. Dla drużyny i niech p_i, x_i, y_i oznaczają odpowiednio liczbę punktów meczowych, punktów uzyskanych i straconych w grze. Jako miarę zgodności liczby uzyskanych punktów w tabeli przyjmujemy statystykę

$$\chi^2 = \sum_i (x_i - \mathrm{E}(X_i))^2/((n-1)\sigma_0^2). \tag{6}$$

Analogicznie definiujemy miarę zgodności dla punktów straconych.

Zmienne losowe P_i, $1 \le i \le n$, są wzajemnie zależne, dlatego do testowania zgodności odpowiednik (6) w tym przypadku nie może być użyty. Wprowadźmy zmienne losowe

$$\widetilde{P}_i = n - 1 + \sum_{j\neq i} P_{ij}$$

określone jako liczba punktów meczowych uzyskanych przez i na własnym boisku. Odnotujmy, że $\widetilde{P}_i$, $1 \le i \le n$, są wzajemnie niezależne. Dla tej liczby punktów przyjmujemy miarę zgodności

$$\chi^2 = \sum_i (\widetilde{p}_i - \mathrm{E}(\widetilde{P}_i))^2/\mathrm{Var}(\widetilde{P}_i), \tag{7}$$

gdzie $\mathrm{Var}(\widetilde{P}_i) = \sum_{j\neq i}\mathrm{E}(P_{ij})\mathrm{E}(Q_{ij})$, natomiast $\widetilde{p}_i$ oznacza empiryczną liczbę punktów meczowych uzyskanych przez i na własnym boisku.

Dla testowania modelu przewagi zastosujemy rozważania subtelniejsze. Mamy $D_i = \sum_{j\neq i}(D_{ij} + D_{ji})$, zatem $\mathrm{Var}(D_i) = 2(n-1)\sigma^2$, $\mathrm{Cov}(D_i, D_k) = \mathrm{Var}(D_{ik}) + \mathrm{Var}(D_{ki}) = 4\sigma^2$.

Rozważmy macierz korelacji $R = (\varrho_{ij})$: $\varrho_{ii} = 1$, $\varrho_{ij} = 1/(n-1)$. Macierzą odwrotną jest $R^{-1} = (\varrho_{ij}^{(-1)})$: $\varrho_{ii}^{(-1)} = 1+\lambda$, $\varrho_{ij} = -\lambda$, gdzie $\lambda = 1/(2(n-1))$. Z [1] (zob. twierdzenie 3.3.3) wynika, że statystyka

$$2(n-1)\sigma^2\chi^2 = D^T R^{-1} D = (1+\lambda)\sum_{i=1}^n D_i^2 - \lambda \sum_{i=1}^n \sum_{j\neq i}^n D_i D_j$$

ma rozkład chi-kwadrat z n stopniami swobody.

4.4. *Test niezależności.* Z danych empirycznych wynika, że korelacja liczby punktów uzyskanych i straconych w meczach sezonu ligi jest dodatnia. Tego faktu można się spodziewać, gdyż jest tak zazwyczaj dla mieszanych zmiennych losowych. W naszym modelu zakładamy niezależność składowych wyniku meczu. W celu testowania tego założenia weźmy pod uwagę zmienną losową

$$C = \frac{1}{N} \sum_{j \neq i} (X_{ij} - \mathrm{E}(X_{ij}))(Y_{ij} - \mathrm{E}(Y_{ij})). \tag{8}$$

Mamy $\mathrm{E}(C) = \mathrm{Cov}(X_{ij}, Y_{ij}) = \sigma_3^2$. Niech więc $\widehat{\varrho} = (S_1^2 + S_2^2 - S_3^2)/(2S_1^2S_2^2)$ będzie estymatorem korelacji $\varrho = \sigma_3^2((\sigma_1^2 + \sigma_2^2)(\sigma_1^2 + \sigma_2^2 + \sigma_3^2))^{-1/2}$.

5. Przykład. Weźmy pod uwagę rozgrywki polskiej ligi koszykówki w sezonie 1998/99. Podajemy kilka liczb opisujących ten sezon. Średnia liczba uzyskanych punktów przez jedną drużynę w meczu wynosi 75.58, dyspersja – 12.61; dla gospodarzy ta średnia wynosi 78.46, dyspersja – 12.56; dla gości – odpowiednio 72.69 i 12.10. Przewaga gospodarza w meczu wynosi średnio 5.77, dyspersja przewagi – 14.57. Dodajmy, że gospodarze wygrywali 163 razy, a przegrywali 77 razy.

Subtelności problemu modelowania sezonu ligi dotyczą zmiennych opisujących wyniki meczów. Przypomnijmy, że mocami $M_i = (m_i^{(1)}, m_i^{(2)}, m_i^{(3)})$ o $s = 3$ składowych drużyny i są punkty meczowe oraz średnie (w jednym meczu sezonu) liczby uzyskanych i straconych punktów. Zauważmy dla porządku, że duża liczba punktów straconych w meczu niekorzystnie świadczy o obronie, jednakże niezbyt fortunny termin „moc" pozostawimy. Rezultaty meczów z dogrywką redukujemy proporcjonalnie do regulaminowego czasu gry. Tabela sezonu jest w tablicy 2.

W pierwszym podejściu, w modelu (1), przyjęto analogicznie do [4], [5] moc jednowymiarową $m_i^{(1)}$, wkład własny zależny od indeksu mocy $r_{ij}^{(l)} = \max(m_i^{(l)} - m_j^{(l)}, 0)$, atut własnego boiska – od mocy gospodarza, pominięto natomiast składową wymiany. Estymacja parametrów i test zgodności dla tabeli końcowej nie potwierdziły tego prostego modelu.

Rozsądne wydaje się założyć, że udział własny drużyny zależy od mocy i indeksów mocy, zaś atut własnego boiska zależy od mocy gospodarza (ewentualnie jest stały). Składową wymiany uzależniamy od symetrycznych funkcji mocy, w tym sumy mocy, ta bowiem cecha drużyn sprzyja wymianie posiadania piłki, a także od modułu różnicy mocy. Dopuszczamy myśl, że drużyny o zbliżonych mocach preferują wymianę ataków (to przypuszczenie nie znajduje potwierdzenia w dalszej analizie). W obliczeniach estymatorów metodą najmniejszych kwadratów wprowadzaliśmy funkcje wagowe dla odchyleń zależne od przewidywanych wariancji zmiennych.

Przyjmijmy oznaczenia:

$m_i^{(1)}$ – liczba punktów meczowych drużyny i w tabeli sezonu,

$m_i^{(2)}$ – średnia liczba punktów uzyskanych w jednym meczu w sezonie,

$m_i^{(3)}$ – średnia liczba punktów straconych w jednym meczu w sezonie,

$r_{ij}^{(l)} = \max(m_i^{(l)} - m_j^{(l)}, 0)$ – indeks mocy l drużyn, $l = 1, 2, 3$, $1 \leq i, j \leq n$.

W modelu ogólnym zakładamy, że macierz Y ma $v = 16$ następujących kolumn:

$y_k^{(1)} = 1$, $y_{k+N}^{(1)} = 0$ – indykator gospodarza meczu,
$y_k^{(l+1)} = m_i^{(l)}$, $y_{k+N}^{(l+1)} = m_j^{(l)}$ – moc l drużyny,
$y_k^{(l+4)} = r_{ij}^{(l)}$, $y_{k+N}^{(l+4)} = r_{ji}^{(l)}$ – indeks mocy l drużyn,
$y_k^{(l+7)} = m_i^{(l)}$, $y_{k+N}^{(l+7)} = 0$ – atut l własnego boiska,
$y_k^{(l+10)} = y_{k+N}^{(l+10)} = m_i^{(l)} + m_j^{(l)}$ – suma mocy l drużyn,
$y_k^{(l+13)} = y_{k+N}^{(l+13)} = |m_i^{(l)} - m_j^{(l)}|$ – moduł różnicy mocy l drużyn, $l = 1, 2, 3$, $1 \leq i, j \leq n$, $1 \leq k \leq N$.

W modelu przewagi zakładamy, że macierz Z ma jako $w = 9$ kolumn moce i różnice mocy:

$y_k^{(l)} = m_i^{(l)}$ – moc l gospodarza,
$y_k^{(l+3)} = m_j^{(l)}$ – moc l gości,
$y_k^{(l+6)} = m_i^{(l)} - m_j^{(l)}$ – różnica mocy l drużyn, $l = 1, 2, 3$, $1 \leq i, j \leq n$, $1 \leq k \leq N$.

5.1. *Parametry.* Wyestymowane parametry modelu dla badanego sezonu polskiej ligi koszykówki znajdujemy w tablicy 1. Są to wartości tych parametrów, które (na poziomie istotności 0.05) różnią się od zera. W tablicy są także wariancje estymatorów. Odnotujmy, że obliczona z (8) kowariancja wynosi $C = 0.42$ i nie różni się od oczekiwanej.

Tablica 1. Parametry modeli

Model przewagi				
j		b_j		$\sigma(b_j)$
1	Atut własnego boiska	5.824		0.72
2	Różnica mocy ataku	0.946		0.10
3	Różnica mocy obrony	−0.935		0.11
Ogólny model liniowy				
j		a_j	$\widetilde{a}_j$	$\sigma(a_j)$
1	Wyraz wolny	−83.994	−79.381	13.68
2	Atut własnego boiska	5.828	5.824	0.95
3	Moc ataku	1.004	0.946	0.10
4	Moc obrony	−0.933	−0.935	0.14
5	Suma mocy obrony	1.001	1.001	0.10

Oznaczenia w tablicy 1: j – indeks zmiennej opisującej, b_j – parametry w modelu przewagi, a_j – parametry w ogólnym modelu liniowym, $\widetilde{a}_j$ – parametry w modelu ogólnym przy wykorzystaniu parametrów z modelu przewagi, $\sigma(a_j)$, $\sigma(b_j)$ – błędy standardowe parametru. Wariancje: $\sigma_1^2 = 57.57$, $\sigma_2^2 = 7.93$, $\sigma_3^2 = 45.30$, $\sigma^2 = 123.07$.

Tablica 2. Polska liga koszykówki w sezonie 1998/99: analiza przewagi

i		p_i	$\mathrm{E}(P_i)$	d_i	$\mathrm{E}(D_i)$
1	Wrocław	57	55.3	417	417.2
2	Włocławek	54	54.4	367	369.7
3	Bytom	51	52.4	277	279.3
4	Ruda Śl.	50	47.5	86	87.3
5	Pruszków	48	47.9	99	100.0
6	Tarnów	48	47.4	79	82.9
7	Toruń	46	44.2	−18	−31.3
8	Stargard Sz.	45	45.7	23	20.7
9	Ostrów	42	43.1	−71	−72.7
10	Sosnowiec	42	41.4	−133	−134.1
11	Sopot	41	40.6	−163	−165.0
12	Szczecin	40	43.5	−61	−59.9
13	Lublin	40	41.6	−129	−128.8
14	Bydgoszcz	40	38.6	−240	−240.6
15	Przemyśl	39	39.3	−228	−218.3
16	Zielona Góra	37	37.1	−304	−306.7
	χ^2	11.18		6.63	

W tablicach 2–3 oznaczamy: i – indeks drużyny, $p_i = m_i^{(1)}$ – liczba punktów meczowych (moce drużyn), $x_i = 30m_i^{(2)}$ – liczba punktów uzyskanych ($m_i^{(2)}$ – moce w ataku), $y_i = 30m_i^{(3)}$ – liczba punktów straconych ($m_i^{(3)}$ – moce w obronie), $d_i = x_i - y_i$ – różnica punktów, $\mathrm{E}(P_i)$ – oczekiwana liczba punktów meczowych, $\mathrm{E}(X_i)$ – oczekiwana liczba punktów uzyskanych, $\mathrm{E}(Y_i)$ – oczekiwana liczba punktów straconych, $\mathrm{E}(D_i)$ – oczekiwana różnica punktów; * – miara zgodności (7) dla punktów uzyskanych na własnym boisku.

Wykorzystując parametry z tablicy 1, obliczamy wartości oczekiwane liczby punktów meczowych, punktów uzyskanych, straconych i różnicę punktów w tabeli sezonu. W obliczeniach dotyczących modelu ogólnego ograniczymy się do wykorzystania parametrów estymowanych bez ograniczeń sugerowanych przez model przewagi. Wyniki znajdujemy w tablicach 2–3, tamże znajdzie Czytelnik wartość testu zgodności. Testy chi-kwadrat potwierdzają zgodność. Błędy standardowe estymatorów (zob. Silvey [7]) dają odpowiedź na pytanie, jak dokładnie jeden sezon ligowy pozwala estymować parametry. Dla śledzenia stabilności parametrów z sezonu na sezon obliczenia powtórzyliśmy dla sezonu 1999/2000. Wyniki różnią się nieistotnie.

Tablica 3. Polska liga koszykówki w sezonie 1998/99: analiza liczby punktów

i		p_i	$E(P_i)$	x_i	$E(X_i)$	y_i	$E(Y_i)$	d_i	$E(D_i)$
1	Wrocław	57	56.0	2363	2363.1	1946	1940.9	417	422.2
2	Włocławek	54	55.4	2597	2601.6	2232	2212.6	365	389.1
3	Bytom	51	53.4	2544	2547.3	2266	2251.2	278	296.1
4	Ruda Śl.	50	47.9	2327	2328.8	2239	2239.5	88	89.3
5	Pruszków	48	48.4	2379	2379.4	2280	2272.4	99	107.0
6	Tarnów	48	47.7	2346	2348.2	2268	2262.1	78	86.1
7	Toruń	46	44.3	2356	2348.7	2374	2377.1	−18	−28.4
8	Stargard Sz.	45	45.4	2074	2072.7	2051	2064.8	23	7.9
9	Ostrów	42	42.7	2110	2110.3	2180	2193.6	−70	−83.3
10	Sosnowiec	42	41.0	2179	2178.8	2312	2318.8	−133	−140.0
11	Sopot	41	40.2	2219	2216.2	2382	2385.3	−163	−169.1
12	Szczecin	40	43.2	2215	2214.6	2276	2279.1	−61	−64.5
13	Lublin	40	41.3	2265	2265.0	2394	2393.6	−129	−128.6
14	Bydgoszcz	40	38.0	2169	2168.6	2409	2415.0	−240	−246.4
15	Przemyśl	39	38.5	2076	2080.0	2304	2312.0	−228	−232.0
16	Zielona Góra	37	36.6	2300	2303.8	2606	2609.2	−306	−305.3
	χ^2	8.11*		0.10		0.95		0.64	

6. Wnioski. Zgodnie z oczekiwaniem moce drużyn w koszykówce są wielowymiarowe. Zwracamy uwagę na to, że o randze drużyn świadczą nie punkty meczowe, a punkty uzyskane w meczu. Przewaga gospodarza zależy jedynie od różnicy mocy, co sugeruje, że atut własnego boiska nie zależy od mocy gospodarza. Wymiana jest istotnym elementem składowym wyniku. Zaskakująca jest zależność oczekiwanej wartości wymiany od sumy mocy w obronie.

Duży wpływ przypadku na wynik na boisku nie daje perspektyw na ścisłą prognozę wyniku pojedynczego meczu. Dla ilustracji prognozujemy wynik meczu wicemistrza z mistrzem sezonu. Da to bowiem odpowiedź na pytanie, czy w tym przypadku większa moc drużyny przyjezdnej niweluje atut własnego boiska.

Moce w ataku i obronie mistrza ($i = 1$) i wicemistrza ($i = 2$) są następujące:

$$m_1^{(2)} = 78.8, \quad m_2^{(2)} = 86.0, \quad m_1^{(3)} = 64.9, \quad m_2^{(3)} = 73.8.$$

Wykorzystując parametry z tablicy 1, znajdujemy dla przewagi:

$$\mathrm{E}(D_{21}) = 4.29, \quad \mathrm{Var}(D_{21}) = \sigma^2 = 11.1, \quad P(D_{21} > 0) = 0.69.$$

Dla modelu ogólnego znajdujemy: $\mathrm{E}(X_{21}) = 73.1$, $\mathrm{E}(Y_{21}) = 77.9$. Na boisku padł wynik 88 : 89.

Prace cytowane

[1] T. W. Anderson, *Introduction to Multivariate Statistical Analysis*, Wiley, New York, 1958.
[2] M. E. Glickman and H. S. Stern, *A state-space model for National Football League scores*, J. Amer. Statist. Assoc. 93 (1998), 25–35.
[3] J. B. Keller, *A characterization of the Poisson distribution and the probability of winning a game*, Amer. Statistician 48 (1994), 294–298.
[4] B. Kopociński, *Modelowanie matematyczne wyników ligi piłki nożnej*, Human Movement 3 (1) (2001), 130–135.
[5] —, *Components of the game result in a football league*, Appl. Math. (Warsaw) 28 (2001), 1–18.
[6] A. Lee, *Modeling scores in the Premier League: Is Manchester United really the best?*, Chance 10 (1997), 15–19.
[7] S. D. Silvey, *Wnioskowanie statystyczne*, PWN, Warszawa, 1978.

Instytut Matematyczny
Uniwersytet Wrocławski
Pl. Grunwaldzki 2/4, 50-384 Wrocław
E-mail: ibk@math.uni.wroc.pl