مجله‌ی بررسی‌های آمار رسمی ایران
سال ۲۱، شماره‌ی ۲، پاییز و زمستان ۱۳۸۹، صص ۲۲۹- ۲۳۷

برازش یک مدل رگرسیون پواسونی بر نتایج فوتبال لیگ برتر ایران

سهیلا شعبانی، احسان بهرامی سامانی* و چنگیز اصلاحچی

دانشگاه شهید بهشتی

چکیده: امروزه پیش‌گویی نتایج مسابقات ورزشی به مساله‌ی مهمی تبدیل شده است. در این میان مسابقات فوتبال از اهمیت ویژه‌ای برخوردار است و در ایران نیز نیاز به یک مدل که بتواند علاوه بر نتیجه‌ی یک بازی، نتیجه‌ی کل یک لیگ را پیش‌گویی کند، احساس می‌شود. لذا در این مقاله سعی در ارایه‌ی مدلی آماری برای پاسخگویی به این نیاز با استفاده از روش‌های بیزی شده است. ابتدا مدل رگرسیون پواسون معرفی و برای به دست آوردن براورد پارامترهای مدل از روش بیزی و نرم‌افزار **Win BUGS** استفاده شده است. در نهایت مدل برای برازش بر داده‌های لیگ برتر ایران طی سال ۸۹- ۱۳۸۸ استفاده شده است.

واژگان کلیدی: مدل خطی تعمیم‌یافته؛ مدل رگرسیون پواسون؛ استنباط بیزی.

۱- مقدمه

امروزه ورزش فوتبال تبدیل به یک صنعت شده و پول زیادی در این زمینه هزینه می‌شود. پیش‌گویی و شرط‌بندی روی نتایج بازی‌ها بین هواداران وشبکه‌های تلویزیونی واینترنتی رایج شده است و مدیران ورزشی با چالش جدیدی روبرو شده اند. علم آمار می‌تواند به مدیران ورزشی برای تصمیم‌گیری درست در شرایط خاص و بحرانی کمک کند تا بتوانند تیم را برای رسیدن به نتیجه‌ی بهتر هدایت کنند. پیش‌گویی مسابقات فوتبال سخت به‌نظر می‌رسد. این پیش‌گویی شامل نتیجه‌ی بازی، برد و باخت، پیش‌گویی قهرمان لیگ، پیش‌گویی امتیاز تیم ها در آخر فصل و ... می‌شود. مدل‌های متنوعی در زمینه پیش‌گویی

* نویسنده‌ی عهده‌دار مکاتبات
دریافت: ۱۳۸۹/۷/۲۸، پذیرش: ۱۳۸۹/۱۱/۱

نتایج مسابقات ورزشی ارایه شده است که نوعی از مدل‌های خطی تعمیم‌یافته می‌باشند. می‌توان به مدل‌های خطی تعمیم‌یافته به‌عنوان بسط طبیعی مدل‌های رگرسیونی نرمال اشاره کرد. مدل‌های خطی به‌دلیل جامعیت و طیف وسیع کاربرد خیلی رایج و مشهور هستند.

یکی از مهم‌ترین این مدل‌ها، مدل رگرسیون پواسون می‌باشد که کاربرد زیادی در علوم مختلف از جمله علوم ورزشی و به‌خصوص فوتبال دارد. از جمله کسانی که مدل‌های مهم و پرکاربرد در زمینه‌ی علوم ورزشی مطرح نمودند، می‌توان به [۱]، [۲]، [۳]، [۷] و [۸] اشاره کرد. در این مقاله مدلی شبیه به مدل ماهر [۸] روی نتایج لیگ برتر فوتبال ایران برازش داده شده است و نتایج جالبی در خصوص پیش‌بینی بازی‌ها به دست آمده است.

ساختار مقاله به‌صورت زیر تنظیم شده است. در بخش ۲، مسئله‌ی مطرح‌شده تشریح گردیده است و مدل رگرسیون پواسن در بخش ۳، معرفی شده است، سپس در بخش ۴، استنباط بیزی توضیح داده شده است و در بخش ۵، مدل بر روی داده‌های لیگ ایران در طی سال ۸۹-۱۳۸۸ پیاده‌سازی شده و نتایج آن بیان شده است.

۲- بیان مسئله

لیگ برتر ایران شامل هجده تیم می‌باشدکه بازی‌ها به‌صورت رفت و برگشت انجام می‌شود. در هر بازی یک تیم از امتیاز بازی خانگی برخوردار است .

اگر بازی به تساوی بیانجامد آن گاه امتیازها بین تیم‌ها تقسیم می‌شود و هر تیم یک امتیاز می‌گیرد. اگر بازی برنده داشته باشد آن‌گاه تیم برنده ۳ امتیاز و بازنده صفر امتیاز می‌گیرد. امتیازها با هم جمع می‌شوند و در نهایت تیمی که بیش‌ترین امتیاز را کسب کرده باشد به‌عنوان تیم قهرمان معرفی می‌شود. چهار تیم برتر لیگ امتیاز حضور در لیگ قهرمانان آسیا را نیز خواهند داشت. دو تیم آخر لیگ نیز به دسته‌ی پایین‌تر سقوط می‌کنند و دو تیم جدید جای آن‌ها را در لیگ می‌گیرند. در این جا مدل ساده شده‌ای با چهار عامل مؤثر بر روی تعداد گل‌های زده شده شامل یک پارامتر ثابت، اثر بازی خانگی، پارامتر دفاعی و پارامتر حمله‌ای در نظر می‌گیریم. از براورد این پارامترها برای پیش‌گویی نتایج بازی‌های آتی لیگ استفاده می‌کنیم. علاوه بر این رتبه‌ی تیم‌ها در لیگ و امتیاز تیم‌ها نیز براورد می‌شود.

۳- مدل رگرسیون پواسون

فرض کنید Y_i برای i امین آزمودنی دارای توزیع پواسون با پارامتر λ باشد که تابع جرم به‌صورت

$$P(Y_i = y_i) = \exp\{y_i \log(\lambda) - \lambda - \log(y_i !)\}.$$

همچنین Y_i متغیر تصادفی است که نشان دهنده‌ی تعداد رخ‌دادها در یک بازه‌ی زمانی یا مکانی معین است. مدل رگرسیون پواسون یک مدل خطی تعمیم‌یافته می‌باشد که به‌صورت

$$Y_i = \text{Poisson}(\lambda_i), \qquad \log(\lambda_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = X_i'\beta$$

معرفی می‌شود که در آن $\beta = (\beta_0, \beta_1, ..., \beta_p)'$ بردار پارامترهای مدل و X_i سطر i ام ماتریس طرح می‌باشند.

۴- توزیع پیشین و توزیع پسین در مدل رگرسیون پواسون

براورد و استنباط بیزی مزایایی در مدل‌بندی و تحلیل داده‌های آماری دارد. این روش، روشی برای به هنگام کردن داده‌ها بر اساس اطلاعات پیشین برای براورد پارامترها فراهم می‌کند. در مدل رگرسیون پواسون به دنبال براورد بردار پارامتر β هستیم. برای این منظور از روش استنباط بیزی استفاده می‌شود. ابتدا توزیع پیشین مناسبی را برای بردار پارامتر β در نظر می‌گیریم. توزیع پیشین به‌صورت زیر در نظر گرفته شده است:

$$\beta \sim \text{N}(\mu_\beta, \Sigma_\beta).$$

در یک مدل رگرسیون پواسون تابع درستنمایی به‌صورت زیر در نظر گرفته می‌شود:

$$L(y|\beta) = \exp\left\{-\sum_{i=1}^{n} \exp(X_i'\beta) + \sum_{i=1}^{n} y_i X_i'\beta - \sum_{i=1}^{n} \log(y_i !)\right\}$$

با استفاده از توزیع پیشین قسمت قبل توزیع پسین به‌صورت زیر به دست می‌آید:

$$f(\beta|y) \propto \exp\left\{-\sum_{i=1}^{n} \exp(X_i'\beta) + \sum_{i=1}^{n} y_i X_i'\beta - \sum_{i=1}^{n} \log(y_i !) \right.$$
$$\left. -\frac{1}{2}(\beta - \mu_\beta)^T \Sigma_\beta^{-1}(\beta - \mu_\beta)\right\}$$

توزیع پسین و براورد پارامترهایش فرم بسته‌ای ندارند، در نتیجه از روش‌های تقریبی مثل روش MCMC استفاده خواهیم کرد.

۵- مدل رگرسیون پواسون بر روی داده‌های فوتبال

این مدل به‌وسیله‌ی ماهر [۸] بیان شد و توسط محققانی چون لی [۷]، نتزوفراس [۹]، کارلیس و نتزوفراس [۵ و ۶] و همچنین نتزوفراس و کارلیس [۴ و ۱۰] مورد استفاده قرار گرفته است. برای بیان این مدل ابتدا Y_{i1} و Y_{i2} به‌ترتیب برای تعداد گل‌های زده تیم میزبان (HT) و تیم مهمان (AT) در مسابقه‌ی i ام در نظر گرفته می‌شود. مدل مورد استفاده به‌صورت زیر خواهد بود.

$$Y_{ij} \sim \text{Poisson}(\lambda_{ij}) \qquad i = 1,...,n \quad j = 1,2$$
$$\log(\lambda_{i1}) = \mu + \text{home} + a_{HT_i} + d_{AT_i}$$
$$\log(\lambda_{i2}) = \mu + a_{AT_i} + d_{HT_i}$$

که در این مدل n تعداد بازی‌ها، μ یک مقدار ثابت (لگاریتم امید ریاضی تعداد گل‌ها در مسابقات)، home میزان اثر میزبانی، HT_i و AT_i اندیس‌ها مربوط به تیم‌های میزبان و مهمان مسابقه دهنده در مسابقه‌ی i ام، a_k و d_k به‌ترتیب نشان دهنده‌ی توانایی‌های هجومی و دفاعی تیم k، $(k = 1,...,K)$ و K تعداد تیم‌های مورد بررسی می‌باشند، که در مثال کاربردهای $K = 18$ است.

برای قابل براورد شدن پارامترهای مدل محدودیت‌هایی برای a_k و d_k $(k = 1,...,K)$ در نظر گرفته می‌شود. بنا بر این خواهیم داشت:

$$\sum_{k=1}^{K} a_k = \sum_{k=1}^{K} d_k = 0$$

مثبت بودن پارامتر هجومی یک تیم نشان دهنده‌ی این است که آن تیم توانایی هجومی خوبی نسبت به متوسط توانایی هجوم تیم‌ها در لیگ دارد. همچنین، منفی بودن پارامتر دفاعی یک تیم نشان دهنده‌ی این است که آن تیم توانایی دفاعی خوبی نسبت به متوسط توانایی دفاع تیم‌ها در لیگ دارد.

جدول ۱- میانگین و انحراف معیار پسین مربوط به پارامتر حمله‌ای و دفاعی

نام تیم	براورد حمله	انحراف معیار حمله	براورد دفاع	انحراف معیار دفاع
پرسپولیس	۰/۰۷۶۹	۰/۱۴۴	-۰/۰۵۶	۰/۱۵۲
مس کرمان	۰/۲۸۸	۰/۱۳۱	۰/۲۹۵	۰/۱۳۰
استقلال	۰/۱۳۲	۰/۱۴۱	-۰/۲۸۰	۰/۱۷۶
سپاهان	۰/۴۴۷	۰/۱۲۱	-۰/۳۱۴	۰/۱۷۷
مقاومت سپاسی	-۰/۲۲۳	۰/۱۶۴	۰/۰۴۱	۰/۱۴۷
پیکان قروین	-۰/۰۵۶۸	۰/۱۶۱	۰/۰۴۱	۰/۱۴۸
ابومسلم	-۰/۱۵۳	۰/۱۶۶	۰/۱۸۰	۰/۱۳۷
استقلال اهواز	-۰/۰۹۴	۰/۱۶۲	۰/۳۳۴	۰/۱۲۸
فولاد خوزستان	-۰/۳۲۶	۰/۱۷۸	-۰/۲۷۵	۰/۱۶۹
راه آهن	-۰/۱۶۶	۰/۱۶۱	۰/۰۲۳	۰/۱۴۶
استیل آذین	۰/۲۹۳	۰/۱۲۹	۰/۱۷۴	۰/۱۳۷
پاس همدان	-۰/۱۵۷	۰/۱۶۲	۰/۰۲۲	۰/۱۴۵
ملوان	-۰/۲۸۵	۰/۱۷۴	۰/۰۵۸	۰/۱۴۷
شاهین بوشهر	-۰/۱۵۰	۰/۱۵۹	-۰/۱۵۴	۰/۱۶۱
تراکتورسازی	-۰/۰۲۴	۰/۱۵۴	-۰/۰۳۳	۰/۱۵۷
صبای قم	۰/۲۰۹	۰/۱۳۷	۰/۰۷۰	۰/۱۴۴
سایپا	۰/۱۱۳	۰/۱۴۴	۰/۲۳۱۸	۰/۱۳٤
ذوب آهن	۰/۰۷۵	۰/۱۴۹	-۰/۳۶۰	۰/۱۸۰

مدل رگرسیون پواسون با در نظر گرفتن دو فرض اساسی و تعیین کننده در مورد هر تیم به کار گرفته می‌شود: آ) استقلال بین تعداد گل‌های میزبان ومهمان، و ب) برابری بین میانگین‌ها و واریانس‌ها.

توزیع‌های پیشین برای پارامترهای a_k و d_k $(k = ۱,...,K)$، home و μ توزیع نرمال با میانگین صفر و واریانس ۱۰۰۰۰ استفاده شده است. حال با استفاده از نرم‌افزار **Win BUGS** نتایج مربوط به براورد پارامترهای مدل را به دست می آوریم.

۱-۵ نتایج

براورد پارامترهای مدل رگرسیون پواسون در جدول ۱ ارایه شده است.

جدول ۲- براورد اثر خانه و پارامتر ثابت

پارامتر	براورد	انحراف معیار
اثر بازی تیم‌های لیگ در خانه‌ی خودشان	۰/۲۸۳	۰/۰۷۱
پارامتر ثابت	۰/۰۲۴	۰/۰۵۵

جدول ۳- میانگین تعداد گل‌های پیش‌بینی‌شده

تعداد گل‌های پیش‌بینی‌شده‌ی نیم‌فصل دوم	براورد	انحراف معیار	مقدار واقعی
ذوب‌آهن (میزبان)-پیکان	۰/۹۷۲-۲/۵۸۹	۱/۰۱۴-۱/۳۰۵	۱-۳
ملوان (میزبان)-استقلال	۱/۲۴۷-۰/۵۷۵	۱/۱۴۲-۰/۷۷۸۳	۱-۰

همان‌طور که در پارامترهای براورد شده‌ی مدل دیده می‌شود، سپاهان بالاترین قدرت هجومی (۰/۴۴۷) را دارد، در حالی که ذوب آهن بالاترین قدرت دفاعی (۰/۳۶۰-) را دارد. در بازی‌هایی که در آن دو تیم به‌طور متوسط قدرت یکسانی دارند، تعداد مورد انتظار گل‌ها برای تیم میزبان معادل ۱/۳۶ و برای تیم مهمان ۱/۰۲ است که با احتمال۳۲٪ منجر به افزایش میانگین امتیاز تیم میزبان می‌شود. در این‌جا پیش‌بینی دو بازی آینده انجام شده است. رویکرد مورد استفاده می‌تواند به‌راحتی برای بازی‌های دیگر تعمیم داده شود. خلاصه ای از گل‌های پیش‌بینی‌شده در جدول ۳ ارایه شده است.

با توجه به جدول ۳ براورد به دست آمده برای تعداد گل‌های زده شده برای هر تیم به مقدار واقعی آن‌ها نزدیک است به‌طوری که تیم ذوب‌آهن در بازیی که در خانه‌ی خود مقابل پیکان داشت با نتیجه‌ی ۳ بر ۱ به پیروزی رسید و نتیجه‌ی براورد (۲/۵۸۹-۰/۹۷۲) همچنین تیم ملوان در خانه‌ی خود مقابل استقلال ۰ بر ۱ شکست خورد و نتیجه‌ی براورد (۱/۲۴۷-۰/۵۷۵) است.

علاقه‌مندیم تا با استفاده از مدل رگرسیون پواسون امتیازهای تیم‌ها را در آخر فصل پیش‌بینی کنیم. نتایج به دست آمده از این مدل در جدول ۴ گردآوری شده است که تقریباً با واقعیت مطابقت دارد.

جدول ۴- امتیاز براورد شده برای تیم‌ها توسط مدل

تیم‌ها	مقدار واقعی	براورد	انحراف معیار
سپاهان	۶۷	۶۹/۵۹۰	۸/۷۶۰
ذوب‌آهن اصفهان	۶۱	۵۹/۵۲۰	۹/۲۹۶
استقلال	۵۹	۵۸/۱۲۰	۹/۲۲۱
صبای قم	۴۸	۵۱/۲۰۰	۹/۵۱۷
استیل آذین	۵۲	۵۰/۹۵۰	۹/۵۲۳
پرسپولیس	۵۳	۵۰/۶۳۰	۹/۴۵۶
مس کرمان	۴۲	۴۷/۰۰۰	۹/۴۷۰
تراکتورسازی	۴۷	۴۶/۸۹۰	۹/۴۵۸
شاهین بوشهر	۳۹	۴۶/۲۸۰	۹/۲۸۷
فولاد خوزستان	۴۲	۴۴/۲۸۰	۹/۲۱۸
پیکان قزوین	۴۱	۴۳/۸۲۰	۹/۳۴۳
سایپا	۴۶	۴۳/۵۵۰	۹/۳۸۱
راه آهن	۳۸	۴۱/۳۰۰	۹/۳۲۲
پاس همدان	۳۸	۴۱/۰۷۰	۹/۱۴۱
مقاومت سپاسی	۳۷	۳۸/۸۱۰	۹/۰۸۴
ابومسلم	۳۲	۳۷/۰۸۰	۸/۹۷۴
ملوان	۴۱	۳۵/۵۰۰	۸/۸۳۰
استقلال اهواز	۳۰	۳۴/۰۶۰	۸/۸۶۱

۶- بحث و نتیجه‌گیری

نتایج به دست آمده تقریباً با واقعیت مطابقت دارد. در صورت وجود اختلاف بین مدل و واقعیت به‌راحتی می‌توان آن‌ها را توجیه کرد. برای بهبود مدل و به دست آوردن پارامترهای مؤثر بر روی نتیجه‌ی بازی‌ها کارهایی انجام شده است ولی هنوز مدل کاملی که بتواند نتایج را بهتر و با خطای کم‌تری به دست آورد ارایه نشده است.

مرجع‌ها

[1] Baxter, M. and Stevenson, R. (1988). Discriminating between the Poisson and negative binomial distributions: An application to goal scoring in association football. *Journal of Applied Statistics,* **15**, 347-438.

[2] Dixon, M. and Coles, S. (1997). Modelling association football scored and inefficiencies in football betting market. *Journal of the Royal Statistical Society*, **46**, 265-280.

[3] Fahrmeir, L. and Tutz, G. (1994). Dynamic stochastic models for time-dependent ordered paired comparison system. *Journal of the American Statistical Association,* **89**, 1438-1449.

[4] Karlis, D. and Ntzoufras, I. (1998). Statistical Modeling for Soccer Games. *In Proceedings of Hellenic European Conference on Computer Mathematics and its Application*.

[5] Karlis, D. and Ntzoufras, I. (2003) Analysis of sports data using bivariate Poisson models. *J. R. Stat. Soc.Seris*, **52**, 381-393.

[6] Karlis, D. and Ntzoufras, I. (2006) Bayesian analysis of the differences of count data. *Stat. Med.*, **25**, 1885-1905.

[7] Lee, A. (1997). Modeling scores in the premier league: Is Manchester United really the best? *Chance,* **10**, 15-19.

[8] Maher, M. (1982). Modelling association football scores. *Statistica Neerlandica,* **36**, 109-118.

[9] Ntzoufras, I. (2009) *Bayesian Modeling Using WinBUGS*. John Wiley & Sons, INC.

[10] Ntzoufras, I. and Karlis, D. (2008). Bayesian modeling of football outcomes: using the Skellam's distribution for the goal difference, *IMA Journal of Management Mathematics.* **20**, 133-145..

سهیلا شعبانی
کارشناس ارشد ریاضی
تهران، اوین، دانشگاه شهید بهشتی، دانشکده‌ی علوم ریاضی، گروه ریاضی.
رایانشانی: soheila.shabani@gmail.com

احسان بهرامی سامانی
دکتری آمار
تهران، اوین، دانشگاه شهید بهشتی، دانشکده‌ی علوم ریاضی، گروه آمار.
رایانشانی: e_bahrami@sbu.ac.ir

چنگیز اصلاحچی
دکتری ریاضی
تهران، اوین، دانشگاه شهید بهشتی، دانشکده‌ی علوم ریاضی، گروه ریاضی.
رایانشانی: ch-eslahchi@sbu.ac.ir