Comparação do desempenho de modelos propostos para o ajuste de dados de esportes coletivos

Mariana de Castro Pasqualini

2022-07-06

Modelos estatísticos podem ser aplicados em diferentes áreas do conhecimento. Uma delas, que tem crescido nos últimos anos, é a análise de dados de competições e eventos esportivos. O número de gols marcados, por exemplo, pode ser tratado como dados de contagem e representados por modelos discretos. Estes modelos são vastamente representados na literatura desde a década de 80, como em Pollard (1985) que utiliza a distribuição Binomial Negativa, enquanto Baxter (1988) apresentam as diferenças entre a Binomial Negativa e Poisson para modelar o placar de partidas de futebol. Tais modelos desconsideram uma estrutura de correlação entre os gols de cada oponente. Karlis (2003) sugere a distribuição Poisson bivariada, que permite uma correlação entre o número de gols marcados pelo mandante e visitante e, ainda, há uma proposta de um modelo bayesiano hierárquico com efeitos aleatórios como definido por Baio (2010). Neste trabalho, são implementados, ajustados e comparados modelos baseados na distribuição Poisson para dados do Campeonato Brasileiro de 2019, obtido em Gomide (2022) utilizando o software Stan e Rstan para inferência bayesiana.

O primeiro deles é o modelo bayesiano hierárquico proposto por Baio (2010), em que o número de gols de cada time é independente definido por $y_{gj}|\theta_{gj} \sim Poisson(\theta_{gj})$, em que $j = 1$ representa o time que jogou a g-ésima partida em casa e $j = 2$ indica o time que jogou a g-ésima partida como visitante. Os autores sugerem algumas distribuições a *priori* adequadas para o software WinBUGS, no qual é feita a implementação do modelo original. As distribuições a *priori* dos parâmetros de interesse foram adaptadas considerando as parametrizações do Stan e seu método de amostragem. As taxas de gols do time mandante e visitante são definidas pela função de ligação log, respectivamente $\log(\theta_{g1}) = home + att_{h(g)} + def_{a(g)}$ e $\log(\theta_{g2}) = att_{a(g)} + def_{h(g)}$. O efeito de jogar em casa é tido como constante ao longo do campeonato e tem uma distribuição a *priori* $home \sim Normal(0, 10)$. O ataque e a defesa dos dois times contribuem para a taxa de pontuação do time, tendo respectivamente efeitos aleatórios com as distribuições a *priori* $att_t \sim Normal(\mu_{att}, \sigma_{att})$ e $def_t \sim Normal(\mu_{def}, \sigma_{def})$. É incluído também distribuições a *priori* para os hiperparâmetros dos efeitos de cada habilidade, sendo que a média do ataque e defesa seguem, cada uma $Normal(0, 10)$ e o desvio-padrão segue $Cauchy(0, 2.5)$ como recomendação de priori pouco informativa para o desvio-padrão por Gelman (2008) e comentado por Almeida Inácio (n.d.).

Para verificar a qualidade da estimação do modelo, uma simulação foi realizada gerando 1000 amostras de tamanho 380, em que o tamanho amostral representa o número de jogos em um campeonato com 20 times. Os histogramas dos resultados obtidos na simulação mostram que as médias das distribuições dos parâmetros a posteriori se concentram em torno do valor real do parâmetro utilizado para simular os dados. Uma extensão do modelo 1 sugerida pelos autores consiste em um modelo de mistura com três componentes, identificando os times com um ótimo desempenho, que estariam no topo da tabela do campeonato, times com um desempenho ruim, no fim da tabela, e os times do meio da tabela, com um desempenho razoável.

O segundo modelo é definido utilizando a distribuição de Poisson bivariada, no qual são definidas três variáveis aleatórias X_1, X_2, X_3 independentes que seguem uma Poisson com seu respectivos parâmetros λ_1, λ_2 e λ_3. Daí, tem-se que $X = X_1 + X_3$ e $Y = X_2 + X_3$ e, conjuntamente X, Y tem uma distribuição Poisson bivariada $BP(\lambda_1, \lambda_2, \lambda_3)$ (Karlis 2003). O fato que, marginalmente, cada variável aleatória tem uma distribuição Poisson com $E(X) = \lambda_1 + \lambda_3$ e $E(Y) = \lambda_2 + \lambda_3$ foi utilizado para a definição do modelo no Stan, com os preditores lineares definidos como $log(\lambda_{1i}) = \mu + home + att_{h_i} + def_{a_i}$ e $log(\lambda_{2i}) = \mu + att_{a_i} + def_{h_i}$, em que i representa cada jogo. Para $\lambda_{3i} = \alpha_0 + \gamma_1\alpha_{h_i}^{home} + \gamma_2\alpha_{a_i}^{away}$ e tal definição permite quatro modelos

distintos através da variável *dummy* γ_i, que indica se o efeito do mandante e/ou visitante será incluído. Originalmente, os parâmetros foram estimados com o algoritmo EM, enquanto neste trabalho os parâmetros foram abordados como variáveis aleatórias e receberam as distribuições a *priori* $att_t \sim Normal(0, \sigma_{att})$ e $def_t \sim Normal(0, \sigma_{att})$, com desvio-padrão definido como no modelo anterior, $\mu \sim Normal(0, 10)$ e os efeitos de λ_3 como Normal(0, 1). Para o modelo mais simples, com $\gamma_1 = \gamma_2 = 0$, foi feita uma simulação de 1000 amostras de tamanho 380 e as médias das distribuições dos parâmetros se concentram em torno dos parâmetros utilizados para simular os dados.

O desempenho dos modelos para dados do Campeonato Brasileiro de 2019 são verificados por meio de uma comparação gráfica dos pontos acumulados ao longo da temporada. Além disso, também é calculado o erro quadrático médio entre a pontuação estimada pelo modelo e observada na competição.

Referências

Almeida Inácio, Marco Henrique de. n.d. "Introdução Ao Stan Como Ferramenta de Inferência Bayesiana." https://marcoinacio.com/stan.

Baio, Marta, Gianluca e Blangiardo. 2010. "Bayesian Hierarchical Model for the Prediction of Football Results." *Journal of Applied Statistics* 37 (2): 253–64. https://doi.org/10.1080/02664760802684177.

Baxter, Richard, Mike e Stevenson. 1988. "Discriminating Between the Poisson and Negative Binomial Distributions:an Application to Goal Scoring in Association Football." *Journal of Applied Statistics* 15 (3): 347–54. https://doi.org/10.1080/02664768800000045.

Gelman, Aleks e Pittau, Andrew e Jakulin. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *The Annals of Applied Statistics* 2 (4). https://doi.org/10.1214/08-AOAS191.

Gomide, Arnaldo, Henrique e Gualberto. 2022. *CaRtola: Extração de Dados Da API Do CartolaFC, Análise Exploratória Dos Dados e Modelos Preditivos Em r e Python*. https://github.com/henriquepgomide/caRtola.

Karlis, Ioannis, Dimitris e Ntzoufras. 2003. "Analysis of Sports Data by Using Bivariate Poisson Models." *Journal of the Royal Statistical Society: Series D (The Statistician)* 52 (3): 381–93. https://doi.org/10.1111/1467-9884.00366.

Pollard, Richard. 1985. "69.9 Goal-Scoring and the Negative Binomial Distribution." *The Mathematical Gazette* 69 (447): 45–47. https://doi.org/10.2307/3616453.