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

Mariana de Castro Pasqualini

2022-05-16

1. Introdução

2. Modelos

Modelo 1

(Baio and Blangiardo 2010) sugerem um modelo bayesiano hierárquico para os gols marcados na partida. No modelo proposto, o número de gols marcados segue uma distribuição Poisson condicionalmente independentes, em que a correlação é incluída por meio dos hiperparâmetros. A distribuição Poisson é vastamente utilizada para problemas de contagem e amplamente aplicada à análises esportivas como sugerem M. Dixon and S. Coles (2007) e D. Karlis and I. Ntzoufras (2003), dentre outros autores.

O vetor $\mathbf{y} = (y_{g1}, y_{g2})$ como um vetor de contagens, podemos tomar

$$y_{gj}|\theta_{gj} \sim Poisson(\theta_{gj})$$

o vetor tendo uma distribuição Poisson condicional aos parâmetros $\theta = (\theta_{g1}, \theta_{g2})$, que representam a taxa de pontuação no g-ésimo jogo para o mandante, representado por $j = 1$ e o visitante $j = 2$.

Assumindo um modelo log-linear de efeitos aleatórios, temos

$$\log \theta_{g1} = home + att_{h(g)} + def_{a(g)} \log \theta_{g2} = att_{a(g)} + def_{h(g)}$$

em que o parâmetro *home* é um efeito fixo representando a vantagem de ter um jogo em casa e a taxa de pontuação considera o *ataque* e a *defesa* dos dois times que estão jogando. Os índices representam o time que da casa $h(g)$ e o time visitante $a(g)$ no g-ésimo jogo.

Considerando que o modelo proposto segue a abordagem bayesiana, os efeitos aleatórios são objetos aleatórios de interesse e é apropriado definir uma distribuição à priori para cada um deles. As prioris sugeridas pelos autores são:

$$home \sim Normal(0, 0.0001) att_t \sim Normal(\mu_{att}, \tau_{att}), def_t \sim Normal(\mu_{def}, \tau_{def})$$

Sendo t cada um dos times do campeonato. A Normal é definida pela média e precisão. O modelo original foi implementado no WinBUGS, que utiliza a mesma parametrização apresentada no artigo. Como priori para μ é definida uma $Normal(0, 0.0001)$ tanto para o ataque quanto defesa, e $Gamma(0.1, 0.1)$ para os τ de ataque e defesa.

Aqui, o modelo foi implementado no Stan e uma adaptação foi necessária, considerando que a parametrização do software é diferente, com a Normal definida pela média e desvio padrão. Passamos a ter:

$$att_t \sim Normal(\mu_{att}, \sigma_{att}), def_t \sim Normal(\mu_{def}, \sigma_{def})$$

Conforme apresentado em (Almeida Inácio, n.d.), a priori não-informativa recomendada é uma Cauchy, portanto:

$$\sigma_{att} \sim Cauchy(0, 2.5), \sigma_{def} \sim Cauchy(0, 2.5)$$

Para garantir a identificabilidade do modelo, os autores sugerem a seguinte restrição nos parâmetros específicos de cada time:

$$\sum_{t=1}^{T} att_t = 0, \sum_{t=1}^{T} def_t = 0$$

Ainda é proposto a restrição em que um dos times é definido como ataque e defesa iguais a 0, o que implica interpretar os parâmetros para os outros times utilizando como referência o time de base. A proposta foi implementada neste trabalho, portanto, a restrição de identificabilidade é:

$$att_T = 0, def_T = 0$$

Tal restrição foi fundamental para que as cadeias de Markov convergissem, além de ser um método mais rápido para a execução do código.

Simulação

Para checar a implementação dos modelos e estimação correta dos parâmetros, foi feita uma simulação com 1000 réplicas de tamanho 380, que é o número de jogos de um campeonato com 20 times, com dados gerados a partir de parâmetros conhecidos. Os parâmetros do modelo usados para simulação são definidos como:

- $home = 0.13$
- $\mu_{att} = 0.05$
- $\mu_{def} = 0.08$
- $\sigma_{att} = 0.56$
- $\sigma_{def} = 0.52$

Observa-se que as distribuições dos parâmetros estão centradas em torno dos valores "reais".

Diagnóstico de convergência da simulação As simulações foram realizadas com apenas 01 cadeia e 5000 interações. O gráfico traceplot mostra que a cadeia converge e consegue caminhar pelo espaço paramétrico.

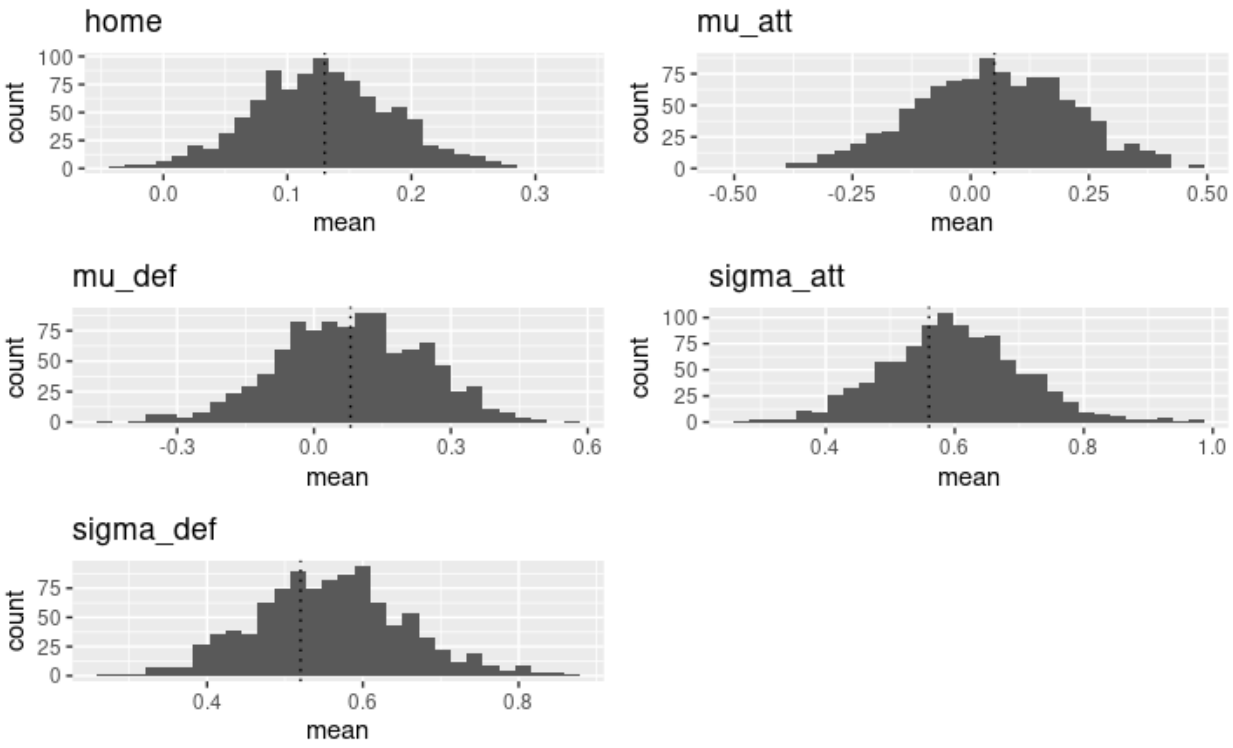
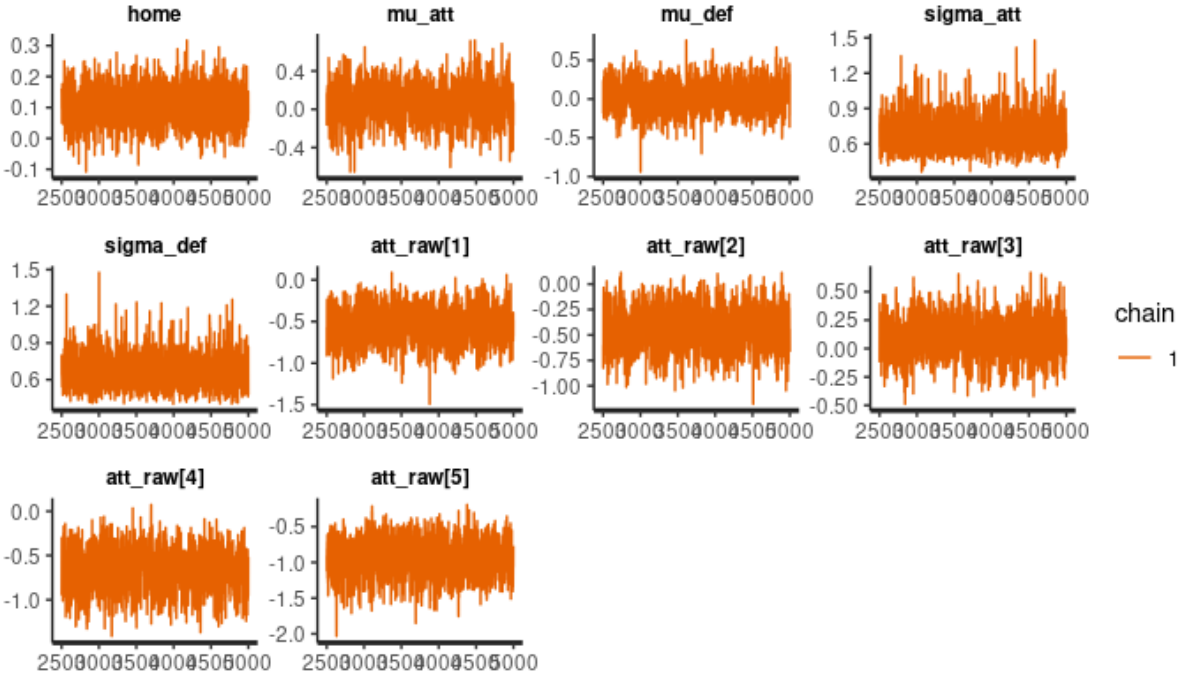


Figure 1: Simulação - Modelo 1



Outra estatística útil é o $\hat{R}$, que próximo de 1 é condição para convergência. Todos os parâmetros apresentaram $\hat{R}$ próximo de 1, sendo o menor $\hat{R} = 0.9995999$ e maior $\hat{R} = 1.002963$.

Ajuste

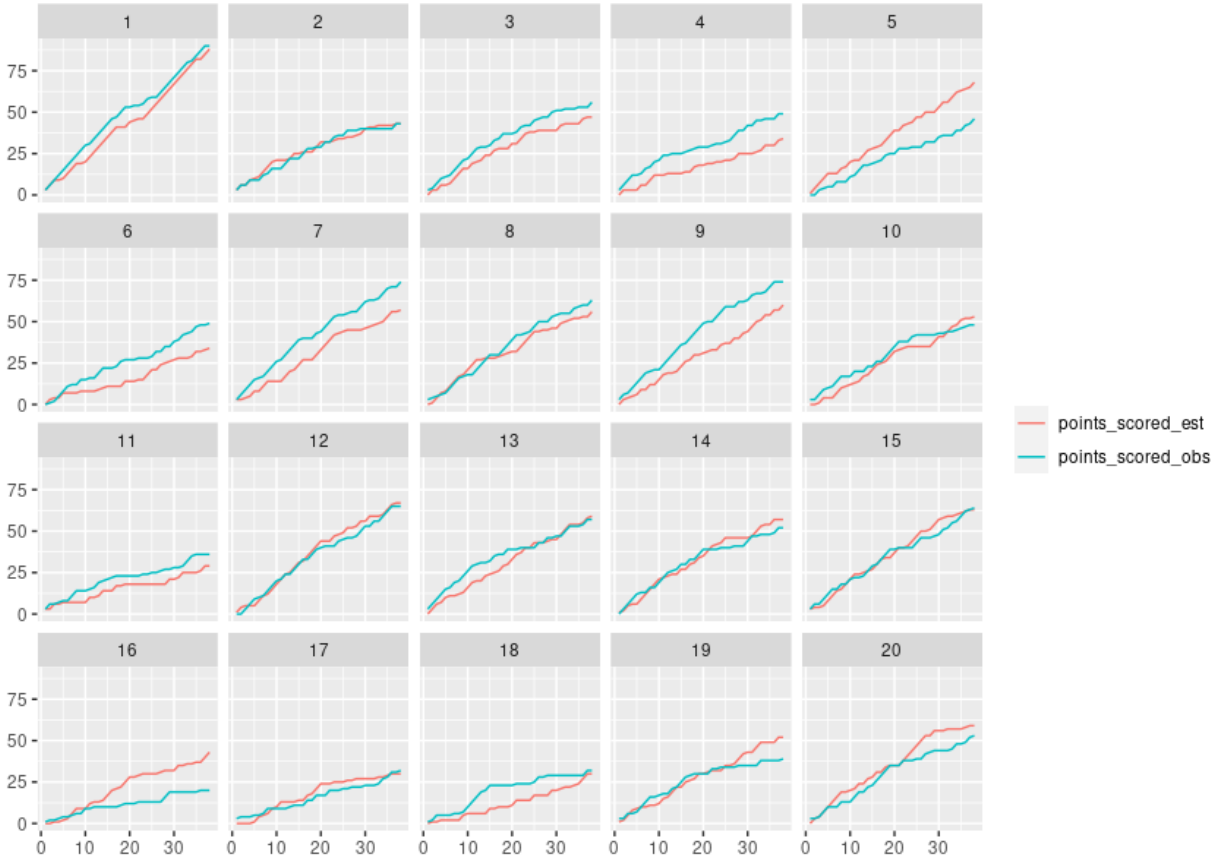
Para verificar o comportamento do modelo com um conjunto de dados reais, assim como no artigo original o modelo é ajustado para dados do campeonato italiano, aqui ele será testado com dados do Campeonato Brasileiro "Brasileirão" ano de 2019.

Os dados foram disponibilizados por (Gomide and Gualberto 2022) no Github, com o seguinte formato:

home_team	away_team	home_score	away_score	home_team_index	away_team_index
282	314	2	1	10	16
315	285	2	0	17	13
262	283	3	1	1	11
276	263	2	0	8	2
293	267	4	1	15	6
265	264	3	2	4	3

As colunas *home_team_index* e *away_team_index* foram criadas atribuindo um valor inteiro ordinal para cada time, seguindo a notação do modelo.

Comparando a pontuação **acumulada** ao longo do campeonato observada e a pontuação estimada pelo modelo, tem-se o seguinte comportamento para cada time:



O desempenho do time 5 foi superestimado pelo modelo, enquanto o contrário aconteceu para o time 9, seu desempenho foi subestimado pelo modelo hierárquico. Os times 2, 12, 15 e 14 apresentam a pontuação acumulada mais próxima entre o estimado e o observado.

Diagnóstico de convergência do ajuste

Modelo 2

A distribuição Poisson é um dos modelos mais utilizados na literatura para análises do número de gols marcados em uma partida de futebol. As variáveis-resposta são usualmente modeladas como duas Poisson independentes, considerando que o número de gols de um time não afeta o número de gols do outro time. Tal suposição não é muito razoável, considerando, por exemplo, que a força de defesa de um time interfere nas oportunidades para a marcação de gols do oponente. A partir disso, (Karlis and Ntzoufras 2003) sugerem a modelagem do número de gols a partir de uma Poisson bivariada, que permite a inclusão de uma covariância positiva que faz o papel da dependência entre as duas variáveis Poisson que, marginalmente, são independentes.

Sendo $X = X_1 + X_3$ e $Y = X_2 + X_3$, duas variáveis aleatórias com $X_i \sim Poisson(\lambda_i)$, então X e Y seguem conjuntamente uma Poisson bivariada $\mathbf{BP}(\lambda_1, \lambda_2, \lambda_3)$.

Conforme mencionado anteriormente, temos duas Poisson independentes marginalmente com $E(X) = \lambda_1 + \lambda_3$ e $Y = \lambda_2 + \lambda_3$. Além disso, $cov(X, Y) = \lambda_3$. Se $\lambda_3 = 0$, então temos simplesmente duas Poisson independentes. Os autores sugerem que o parâmetro λ_3 representam as condições de jogo comuns aos dois times da partida, como ritmo do jogo e condições climáticas.

Definindo diretamente o modelo aplicado à futebol, temos que para cada jogo i

$$X_i \sim Poisson(\lambda_{1i}), Y_i \sim Poisson(\lambda_{2i})$$

e

$$\log(\lambda_{1i}) = \mu + home + att_{h_i} + def_{g_i}, \log(\lambda_{2i}) = \mu + att_{g_i} + def_{h_i}.$$

Para a inclusão da covariância como λ_3, Karlis (2003) apresenta o preditor linear que permite combinar diferentes modelos:

$\log(\lambda_{3i}) = \beta^{con} + \gamma_1 \beta_{h_i}^{home} + \gamma_2 \beta_{g_i}^{away}$

No qual γ_j varia de acordo com o modelo de interesse: quando assume valor 0, temos que a covariância é constante.

Modelo 3

Abordagem com uma distribuição Poisson inflada de zeros.

Modelo 4

Abordagem com uma distribuição Binomial Negativa.

Modelo 5

Modelo 6

3. Implementação e resultados

4. Análise de convergência

5. Referências

(R Core Team 2019)

Exemplo Tabela

R Markdown

```
data_frame(
   parameter = c(
    "$\\lambda_1$ (old normal)",
    "$N$ (total days)",
    "$d_2$ (time to new normal)"
  ),
  value = c(
    300, 400, 12
  )
) %>%
  kable(
    escape = FALSE, booktabs = TRUE,
    caption = "Simulation parameters"
  ) %>%
  # extra formatting with kableExtra:
  kable_styling(latex_options = "hold_position") #%>%
```

```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

Table 1: Simulation parameters

parameter	value
λ_1 (old normal)	300
N (total days)	400
d_2 (time to new normal)	12

```
  #group_rows(index = c("Rates (unknown)" = 3,
  #                     "Other parameters" = 5))
```


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

Baio, Gianluca, and Marta 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.

Gomide, Henrique, and Arnaldo 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, Dimitris, and Ioannis 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.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.