

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

Mariana de Castro Pasqualini

2022-05-16

## 1. Introdução

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, 2020 e 2021, obtido em Gomide (2022) utilizando o software Stan e RStan para inferência bayesiana.

## 2. Modelos

Para o problema de contagem do número de gols em uma partida, o modelo mais comum é baseado na distribuição de Poisson. Essa distribuição é discreta, representando o número de eventos ocorridos em um intervalo de tempo. Uma das limitações dela é que sua média  $\lambda$  e variância são iguais, portanto se há uma superdispersão nos dados, o modelo pode não ser tão apropriado. Todos os modelos que serão apresentados aqui são baseados nessa distribuição e ajustados aos dados do Campeonato Brasileiro ou “Brasileirão” do ano de 2019.

### 2.1 Modelo 1

Baio (2010) sugere um modelo bayesiano hierárquico para os gols marcados em uma determinada partida. No modelo proposto, o número de gols realizados 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 e S. Coles (2007) e D. Karlis e 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, tem-se

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

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

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.

## Priori

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:

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

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  $\mu$  é definida uma  $Normal(0, 0.0001)$  tanto para o ataque quanto defesa, e  $Gamma(0.1, 0.1)$  para os  $\tau$  de ataque e defesa.

Nesse trabalho, o modelo foi implementado no Stan e uma adaptação foi necessária, considerando que a parametrização do software é diferente. A distribuição Normal é definida pela média e desvio padrão, então passamos a ter as prioris para os parâmetros:

$$\begin{aligned} att_t &\sim Normal(\mu_{att}, \sigma_{att}) \\ def_t &\sim Normal(\mu_{def}, \sigma_{def}) \end{aligned}$$

Além disso, é necessário definir também as distribuições a priori dos hiperparâmetros. Para as médias, como não há conhecimento de informações que podem ser agregadas à priori, a escolha são prioris pouco informativas:

- $\mu_{att} \sim Normal(0, 10)$
- $\mu_{def} \sim Normal(0, 10)$

Conforme demonstrado por Gelman (2008) e comentado em Almeida Inácio (n.d.), a priori não-informativa recomendada para o desvio padrão é uma Cauchy, portanto:

$$\begin{aligned} \sigma_{att} &\sim Cauchy(0, 2.5) \\ \sigma_{def} &\sim Cauchy(0, 2.5) \end{aligned}$$

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

$$\begin{aligned} \sum_{t=1}^T att_t &= 0 \\ \sum_{t=1}^T def_t &= 0 \end{aligned}$$

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, então, 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 estimativa 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. Os parâmetros do modelo usados para simulação são definidos como:

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

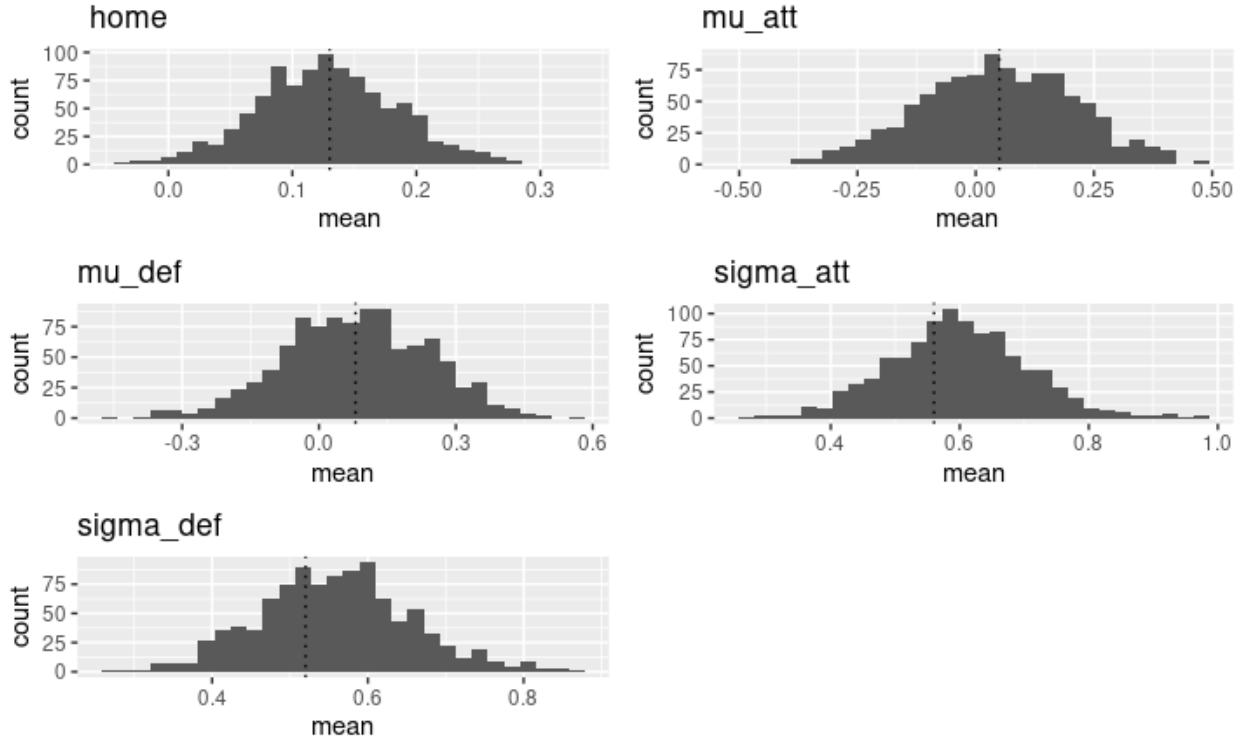
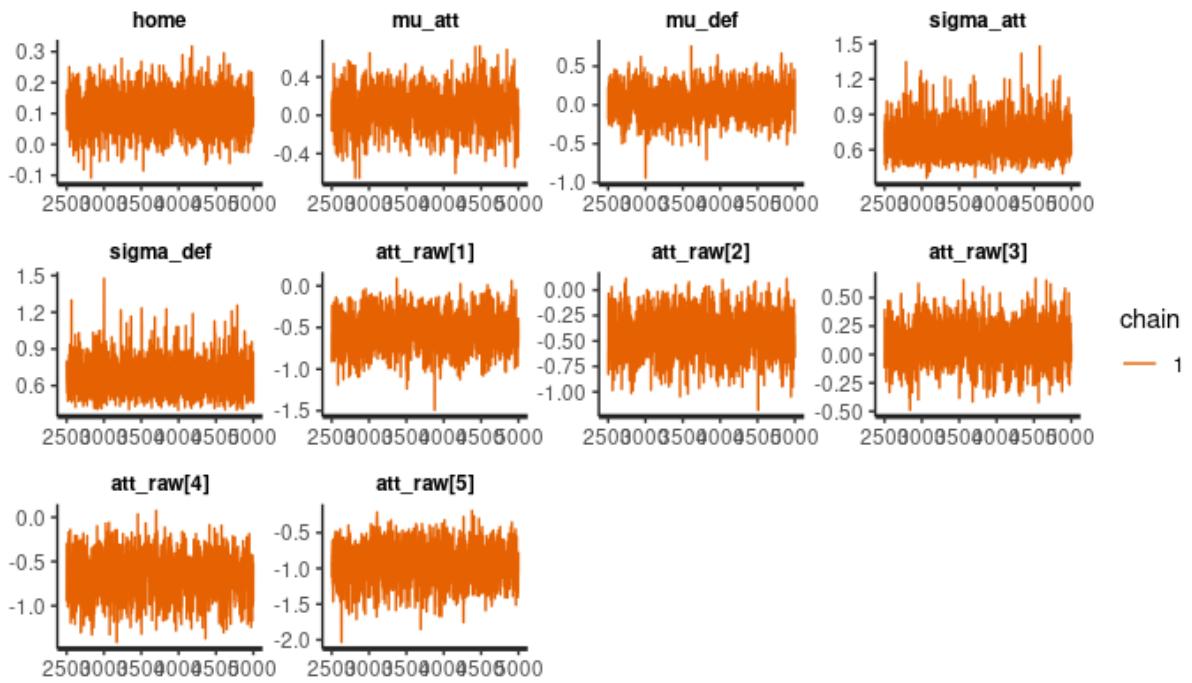


Figura 1: Simulação - Modelo 1

Observa-se que as distribuições da média da distribuição a posteriori 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.



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 do ano de 2019.

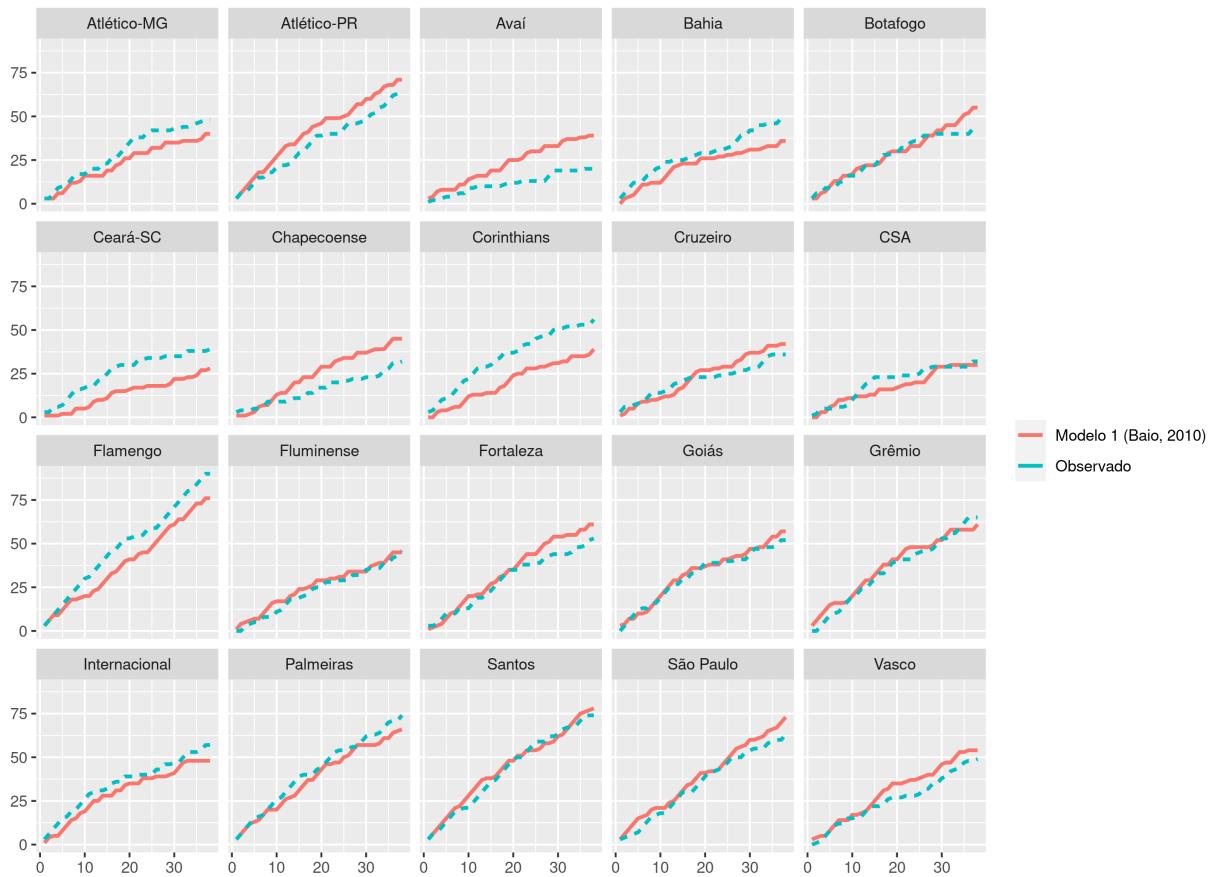
O dados foram disponibilizados por Gomide (2022) no Github, com o seguinte formato:

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:

## Campeonato Brasileiro 2019



Santos foi time com maior pontuação atribuída pelo modelo (78 pontos), estimando bem próximo da pontuação obtida e, assim, sendo o campeão segundo o modelo. Porém, o campeão de 2019 foi o Flamengo, que ficou em segundo lugar na pontuação estimada. Para o Fluminense, o erro foi de apenas 1 ponto. Em contrapartida, há estimativas subestimadas para Corinthians e Ceará.

Os efeitos de ataque e defesa são interpretados com base na referência, definida anteriormente como o último time pelo seu índice. A última equipe é o Fortaleza.

Flamengo é o time com o maior efeito de ataque em relação a linha de base Fortaleza, tendo uma média a posteriori de 0.546 e 90% de probabilidade da média a posteriori do ataque estar entre 0.30 e 0.794. Já o Cruzeiro apresenta um efeito de ataque negativo e seu intervalo de credibilidade de 90% é [-0.767; -0.124] e tem um menor efeito de fazer gols numa partida.

Para o efeito de defesa, os times com maior média apresentam maior propensão de conceder gols, tais como CSA e Avaí, dois dos piores do campeonato.

Conforme o esperado, o efeito de jogar em casa é positivo e grande, com intervalo de credibilidade de 90% entre 0.299 e 0.53.

## 2.2 Modelo 2

Conforme dito anteriormente, 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

Tabela 1: Efeitos de ataque estimados - Modelo 1

Time	Média	Mediana	Desvio-padrão	5%	95%
Flamengo	0.546	0.547	0.149	0.300	0.794
Botafogo	-0.341	-0.335	0.187	-0.675	-0.040
Corinthians	-0.104	-0.106	0.171	-0.383	0.179
Bahia	-0.062	-0.063	0.170	-0.335	0.228
Fluminense	-0.182	-0.182	0.179	-0.486	0.106
Vasco	-0.162	-0.156	0.182	-0.467	0.126
Palmeiras	0.218	0.217	0.159	-0.039	0.486
São Paulo	-0.169	-0.172	0.179	-0.472	0.122
Santos	0.210	0.212	0.162	-0.051	0.484
Atlético-MG	-0.036	-0.034	0.175	-0.328	0.250
Cruzeiro	-0.438	-0.434	0.195	-0.767	-0.124
Grêmio	0.269	0.270	0.160	0.005	0.530
Internacional	-0.063	-0.056	0.172	-0.355	0.215
Goiás	-0.011	-0.009	0.172	-0.296	0.277
Atlético-PR	0.061	0.062	0.164	-0.206	0.330
Avaí	-0.690	-0.683	0.215	-1.045	-0.359
Chapecoense	-0.339	-0.335	0.194	-0.664	-0.026
CSA	-0.513	-0.507	0.203	-0.863	-0.181
Ceará-SC	-0.226	-0.226	0.175	-0.514	0.061
Fortaleza	0.000	0.000	0.000	0.000	0.000

Tabela 2: Efeitos de defesa estimados - Modelo 1

Time	Média	Mediana	Desvio-padrão	5%	95%
21 Flamengo	-0.111	-0.105	0.158	-0.374	0.144
22 Botafogo	-0.038	-0.036	0.155	-0.295	0.212
23 Corinthians	-0.177	-0.174	0.158	-0.442	0.078
24 Bahia	-0.052	-0.051	0.156	-0.311	0.198
25 Fluminense	-0.020	-0.021	0.152	-0.263	0.228
26 Vasco	-0.034	-0.038	0.155	-0.281	0.222
27 Palmeiras	-0.196	-0.193	0.161	-0.469	0.064
28 São Paulo	-0.234	-0.226	0.167	-0.517	0.028
29 Santos	-0.178	-0.174	0.163	-0.451	0.088
30 Atlético-MG	0.019	0.020	0.157	-0.232	0.275
31 Cruzeiro	-0.026	-0.023	0.155	-0.284	0.228
32 Grêmio	-0.097	-0.097	0.158	-0.366	0.152
33 Internacional	-0.106	-0.105	0.155	-0.364	0.147
34 Goiás	0.195	0.193	0.159	-0.075	0.450
35 Atlético-PR	-0.199	-0.196	0.165	-0.478	0.066
36 Avaí	0.160	0.160	0.154	-0.100	0.409
37 Chapecoense	0.049	0.050	0.150	-0.203	0.291
38 CSA	0.115	0.114	0.153	-0.135	0.371
39 Ceará-SC	-0.086	-0.085	0.157	-0.351	0.172
40 Fortaleza	0.000	0.000	0.000	0.000	0.000

(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.

Tabela 3: Efeito de jogar em casa - Modelo 1

Parâmetro	Média	Mediana	Desvio-padrão	5%	95%
home	0.412	0.411	0.069	0.299	0.53

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  $\text{BP}(\lambda_1, \lambda_2, \lambda_3)$ .

Conforme mencionado anteriormente, tem-se 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  $\lambda_3$  representam as condições de jogo comuns aos dois times da partida, como ritmo do jogo e condições climáticas.

Contudo, tal modelagem tem uma limitação: levando em conta que a covariância entre  $X$  e  $Y$  também é o parâmetro da Poisson e o espaço paramétrico está definido em  $(0, +\infty)$ , a covariância também está limitada em  $(0, +\infty)$ . Isso significa que à medida que o número de gols de um dos times aumenta, o do outro time não tende a seguir a relação inversa e, por isso, a interpretação de condições favoráveis aos dois times simultaneamente. Porém, é razoável pensar que essa relação pode ser negativa, com o aumento do comportamento ofensivo de um time e a outra equipe sem muitas oportunidades de marcar gols.

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

$$\begin{aligned} X_i &\sim Poisson(\lambda_{1i}) \\ Y_i &\sim Poisson(\lambda_{2i}) \end{aligned}$$

e usando a função de ligação log para os preditores lineares, tem-se:

$$\begin{aligned} \log(\lambda_{1i}) &= \mu + home + att_{h_i} + def_{g_i} \\ \log(\lambda_{2i}) &= \mu + att_{g_i} + def_{h_i} \end{aligned}$$

Para a inclusão da covariância como  $\lambda_3$ , Karlis (2003) apresenta o preditor linear que permite combinar diferentes modelos:

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

No qual  $\gamma_j$  é uma variável *dummy*, indicando quais parâmetros serão incluídos no modelo de interesse. Para o modelo 2,  $\gamma_1 = \gamma_2 = 0$ , ou seja, tem-se apenas uma covariância constante.

No artigo original, ataque e defesa são tratados como efeitos fixos, portanto o número de parâmetros é o número de times multiplicado por dois mais 1, para o parâmetro que representa a covariância. Para os dados utilizados por Karlis do Campeonato Italiano de 1991-1992, são 37 parâmetros, enquanto para o Campeonato Brasileiro de 2019 seriam 41 parâmetros. Por isso, na adaptação do modelo, ataque e defesa foram abordados como efeitos aleatórios.

A restrição de identificabilidade dos efeitos de ataque e defesa é a mesma do modelo 1, com o efeito do último time definido como:

$$att_T = 0$$

$$def_T = 0$$

## Priori

A escolha das distribuições a priori deste modelo segue o mesmo princípio do modelo 1: prioris pouco informativas.

$$home \sim Normal(0, 10)$$

$$\begin{aligned}
\sigma_{att} &\sim Cauchy(0, 2.5) \\
\sigma_{def} &\sim Cauchy(0, 2.5) \\
\mu &\sim Normal(0, 10) \\
\alpha &\sim Normal(0, 1) \\
\alpha^{home} &\sim Normal(0, 1) \\
\alpha^{away} &\sim Normal(0, 1)
\end{aligned}$$

## Simulação

Com o objetivo de verificar a estimativa certa dos parâmetros, também foi feita uma simulação com 1000 réplicas de tamanho 380, representando o número de jogos de um campeonato com 20 times. Neste modelo, os parâmetros para simulação são definidos como:

- $home = 0.13$
- $\mu = 0.21$
- $\alpha = 0.20$
- $\sigma_{att} = 0.92$
- $\sigma_{def} = 0.80$

A partir dos resultados dos histogramas obtidos na simulação, tem-se que o modelo estima corretamente os parâmetros.

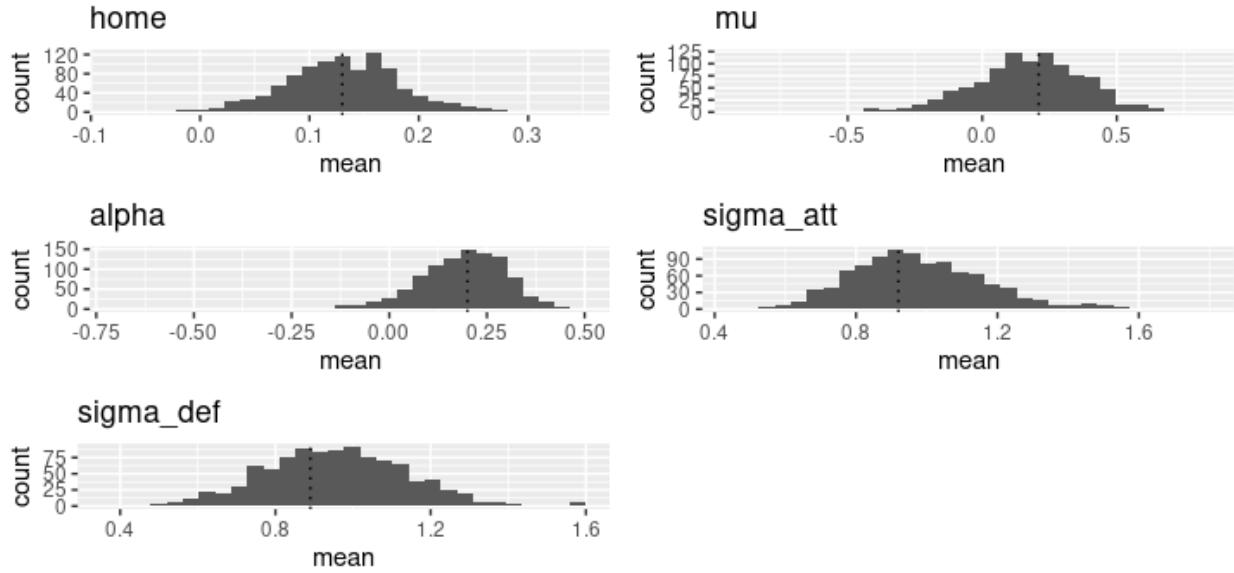


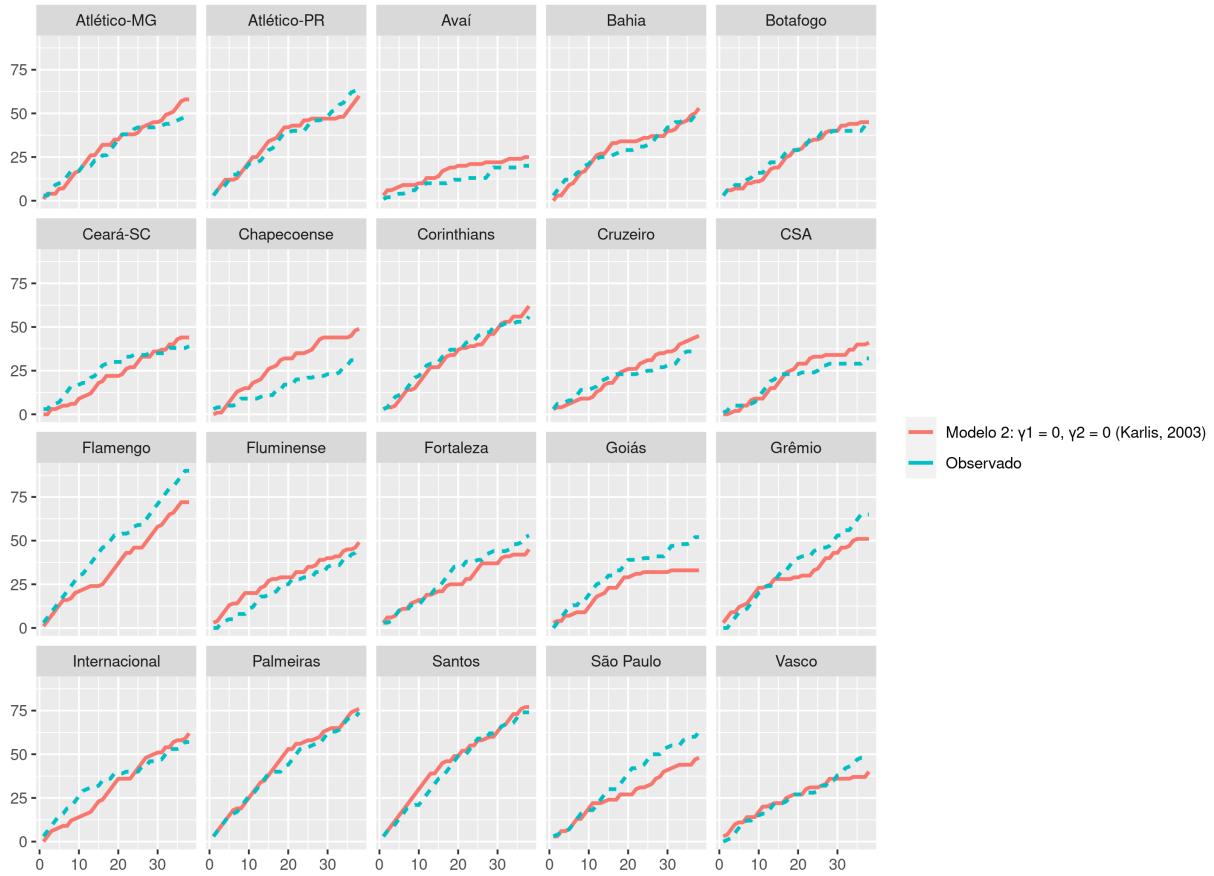
Figura 2: Simulação - Modelo 2

Com respeito às cadeias, a estatística  $\hat{R}$  para os parâmetros se mostrou próxima de 1, sendo o menor  $\hat{R} = 0.9995999$  e maior  $\hat{R} = 1.002963$ .

## Ajuste

Assim como o primeiro modelo, o modelo 2 foi ajustado para o Campeonato Brasileiro de 2019. Neste modelo,  $\gamma_1 = \gamma_2 = 0$ .

Campeonato Brasileiro 2019



Santos e Palmeiras tem curvas ótimas da pontuação acumulada no campeonato. Algumas curvas vão se afastando ao longo da competição, como nota-se o Flamengo, São Paulo e Chapecoense.

### 2.3 Modelo 3

O modelo três é uma extensão do modelo 2, no qual  $\gamma_1 = 1, \gamma_2 = 0$ . Apenas o efeito fixo e o efeito que depende da equipe mandante são considerados no parâmetro de correlação  $\lambda_3$ . Não foi feita uma simulação para o modelo por limitações de memória no computador.

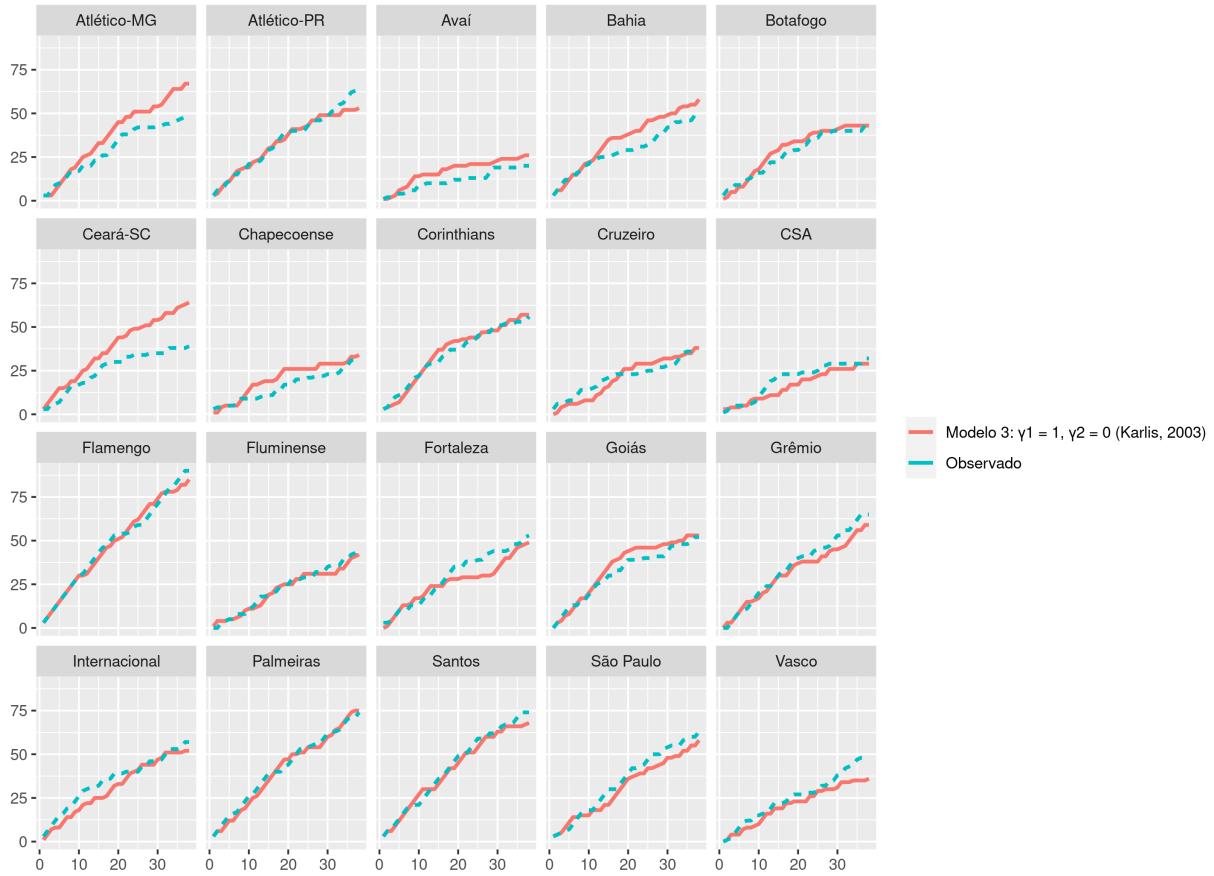
Tabela 4: Efeito de ataque de cada time, estimado pelo Modelo 2

Time	Média	Mediana	Desvio-padrão	5%	95%
Flamengo	1.164	1.130	0.340	0.697	1.756
Botafogo	-0.523	-0.451	0.515	-1.332	0.108
Corinthians	-0.041	-0.028	0.388	-0.669	0.543
Bahia	-0.033	-0.002	0.401	-0.710	0.531
Fluminense	-0.521	-0.418	0.581	-1.445	0.169
Vasco	-0.346	-0.286	0.475	-1.202	0.281
Palmeiras	0.654	0.616	0.331	0.178	1.233
São Paulo	-0.229	-0.191	0.446	-0.937	0.386
Santos	0.693	0.646	0.347	0.203	1.333
Atlético-MG	0.110	0.114	0.343	-0.443	0.652
Cruzeiro	-0.907	-0.787	0.586	-1.945	-0.165
Grêmio	0.703	0.668	0.328	0.222	1.287
Internacional	0.063	0.070	0.383	-0.522	0.627
Goiás	0.170	0.161	0.365	-0.405	0.761
Atlético-PR	0.208	0.223	0.400	-0.354	0.752
Avaí	-1.231	-1.125	0.612	-2.286	-0.452
Chapecoense	-0.607	-0.542	0.520	-1.514	0.077
CSA	-0.830	-0.754	0.567	-1.810	-0.088
Ceará-SC	-0.313	-0.262	0.460	-1.058	0.276
Fortaleza	0.000	0.000	0.000	0.000	0.000

Tabela 5: Efeito de defesa de cada time, estimado pelo Modelo 2

Time	Média	Mediana	Desvio-padrão	5%	95%
21 Flamengo	-0.178	-0.150	0.281	-0.678	0.232
22 Botafogo	0.049	0.040	0.240	-0.335	0.441
23 Corinthians	-0.267	-0.238	0.282	-0.778	0.128
24 Bahia	-0.088	-0.059	0.259	-0.554	0.286
25 Fluminense	0.052	0.053	0.233	-0.326	0.435
26 Vasco	0.059	0.059	0.247	-0.344	0.452
27 Palmeiras	-0.238	-0.217	0.271	-0.722	0.164
28 São Paulo	-0.360	-0.324	0.303	-0.899	0.064
29 Santos	-0.307	-0.273	0.306	-0.863	0.126
30 Atlético-MG	0.048	0.047	0.247	-0.371	0.443
31 Cruzeiro	0.060	0.061	0.240	-0.338	0.456
32 Grêmio	-0.179	-0.148	0.281	-0.690	0.228
33 Internacional	-0.166	-0.140	0.261	-0.636	0.214
34 Goiás	0.534	0.518	0.258	0.133	0.966
35 Atlético-PR	-0.247	-0.225	0.263	-0.713	0.133
36 Avaí	0.411	0.390	0.242	0.035	0.842
37 Chapecoense	0.107	0.109	0.231	-0.277	0.470
38 CSA	0.246	0.238	0.240	-0.129	0.645
39 Ceará-SC	-0.033	-0.020	0.235	-0.424	0.335
40 Fortaleza	0.000	0.000	0.000	0.000	0.000

Campeonato Brasileiro 2019

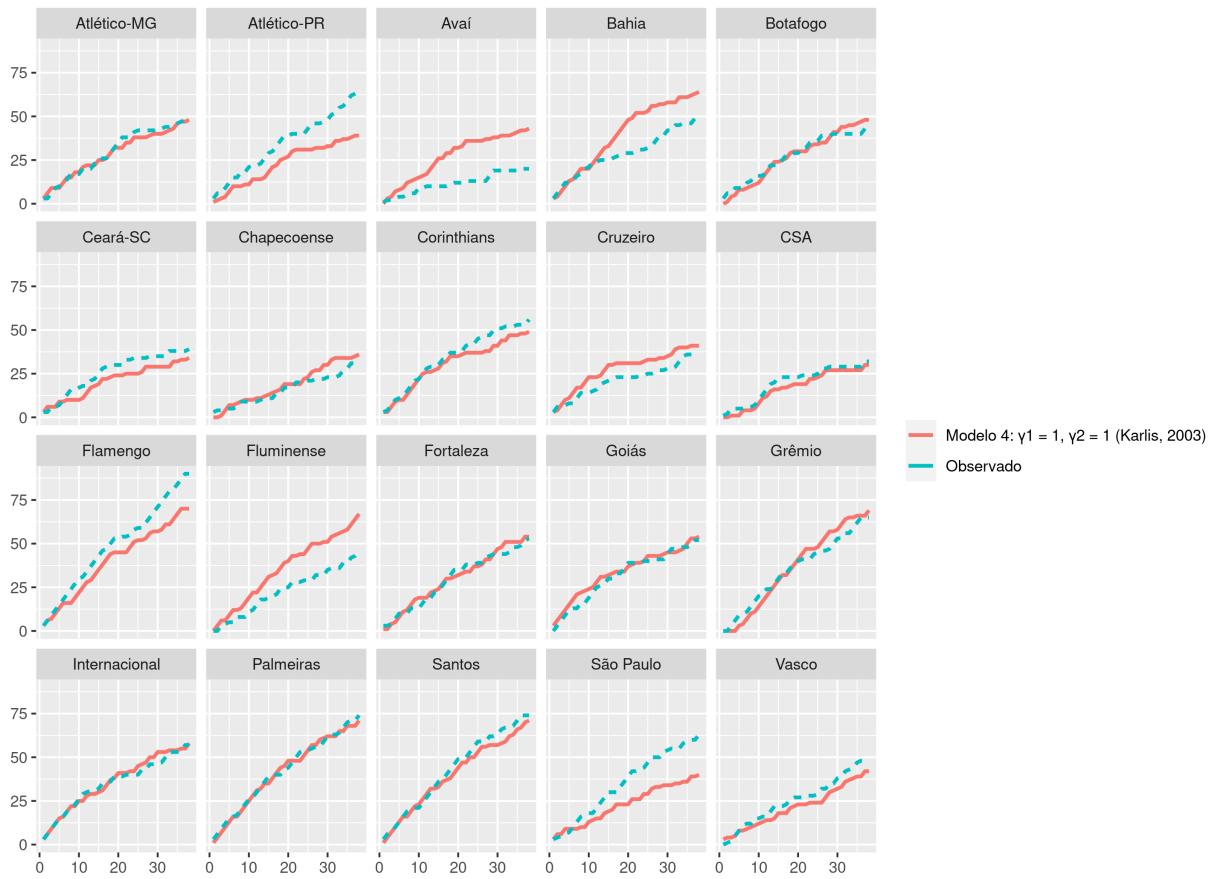


Diferentemente dos modelos anteriores, o terceiro modelo consegue acertar o time carioca como campeão, errando apenas por 5 pontos. As piores estimativas, em geral, foram para os times que não tiveram um desempenho tão notável negativa ou positivamente.

#### 2.4 Modelo 4

O modelo quatro é uma extensão do segundo modelo, no qual  $\gamma_1 = 1, \gamma_2 = 1$ . Nesse caso, a correlação entre os gols do mandante e visitante depende do efeito fixo e dos efeitos de cada time. É a variação mais completa do modelo 2. Contudo, talvez ela fizesse mais sentido se suportasse uma covariância negativa, representando as condições reais para os dois times em campo. Também não foi realizada uma simulação pela mesma razão do modelo 3.

Campeonato Brasileiro 2019

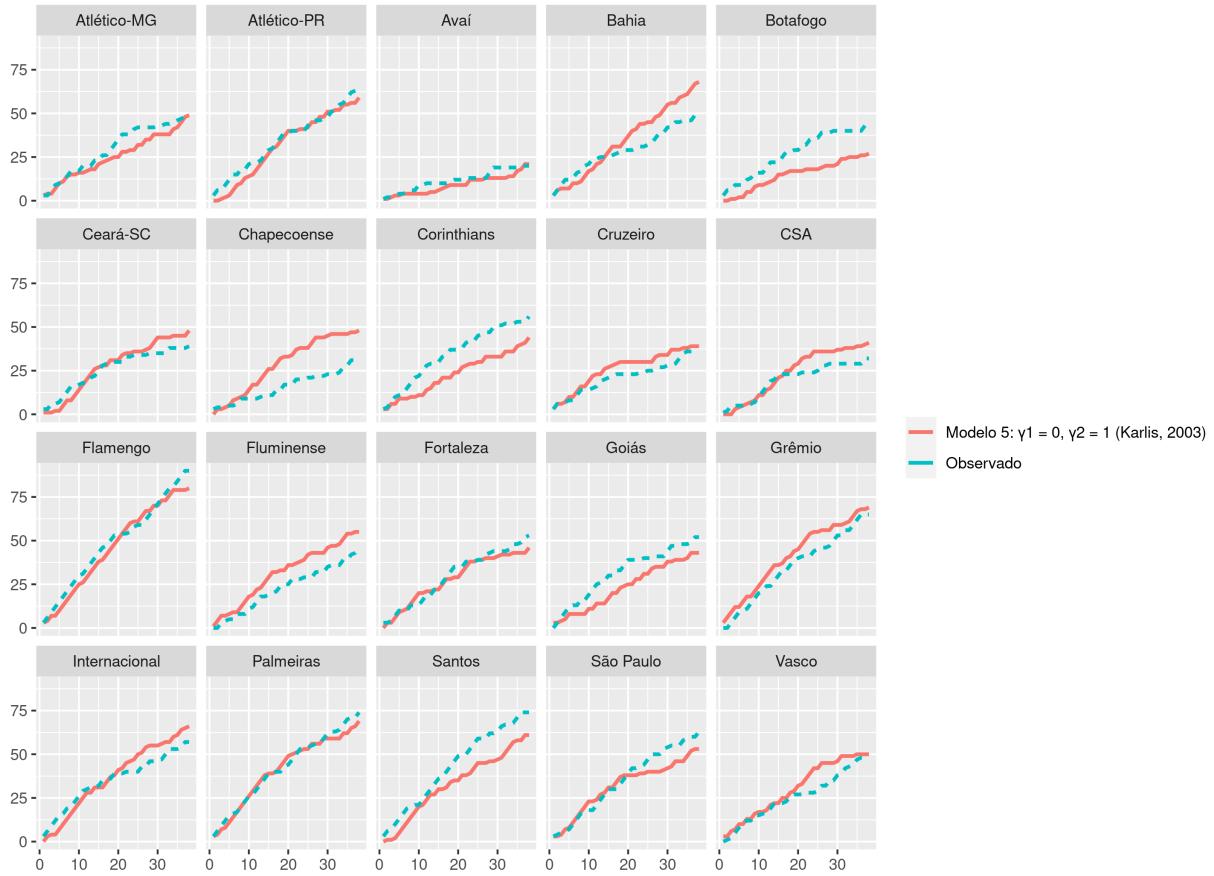


Observa-se um indicativo de o pior modelo ser o quarto. As curvas se afastam notavelmente para vários times, como Atlético Paranaense, Avaí e São Paulo. Ainda assim, o modelo foi capaz de acertar os times que estavam no *top 4* do Campeonato: Santos, Palmeiras, Flamengo e Grêmio.

## 2.5 Modelo 5

O modelo quatro é uma extensão do segundo modelo, que leva em consideração apenas o efeito fixo  $\alpha$  e o efeito dependente do time visitante, sendo  $\gamma_1 = 0, \gamma_2 = 1$ . Também não foi realizada uma simulação por limitações computacionais.

Campeonato Brasileiro 2019



O quinto modelo também não parece apresentar estimativas muito boas, como para o Botafogo, Bahia e Chapecoense. É compreensível que o modelo não tenha uma performance tão interessante, uma vez que ele desconsidera o efeito do mando de casa na dependência entre o número de gols, e há uma forte percepção do efeito do time da casa ser forte nos jogos competitivos.

## 2.6 Modelo 6

O sexto modelo é uma extensão do modelo 1, incluindo uma mistura de 3 componentes, indicando 3 categorias das habilidades do time. A partir disso, o efeito de ataque e defesa são definidos em função do grupo que a equipe pertence.

Neste modelo, o ataque e defesa seguem uma distribuição t com 4 graus de liberdade, **ponderados** pela probabilidade do time pertencer a um dos três grupos: (1) final da tabela, (2) meio da tabela e (3) topo da tabela (Baio (2010)).

$$att_t = \sum_{k=1}^3 \pi_{kt}^{att} \times t(\mu_k^{att}, \tau_k^{att}, \nu)$$

$$def_t = \sum_{k=1}^3 \pi_{kt}^{def} \times t(\mu_k^{def}, \tau_k^{def}, \nu)$$

## Priori

Novamente, define-se distribuições a priori para os parâmetros e hiperparâmetros do modelo. A probabilidade de pertencer ao grupo 1, 2 ou 3 é igual para todos os times, então define-se

- $\pi_{att} \sim Dirichlet([1, 1, 1])$
- $\pi_{def} \sim Dirichlet([1, 1, 1])$

Como é uma extensão do modelo 1, o efeito fixo de jogar em casa se mantém com a mesma priori  $home \sim Normal(0, 10)$ . Independentemente do grupo que o ataque e a defesa pertencem,  $\sigma_{att} \sim Cauchy(0, 2.5)$  e  $\sigma_{def} \sim Cauchy(0, 2.5)$ . As médias dos grupos que vão variar de acordo com qual cluster o efeito vêm:

Grupo 1

- $\mu_1^{att} \sim truncNormal(0, 10, -3, 0)$
- $\mu_1^{def} \sim truncNormal(0, 10, 0, 3)$

Grupo 2

- $\mu_2^{att} \sim truncNormal(0, 0.01)$
- $\mu_2^{def} \sim truncNormal(0, 0.01)$

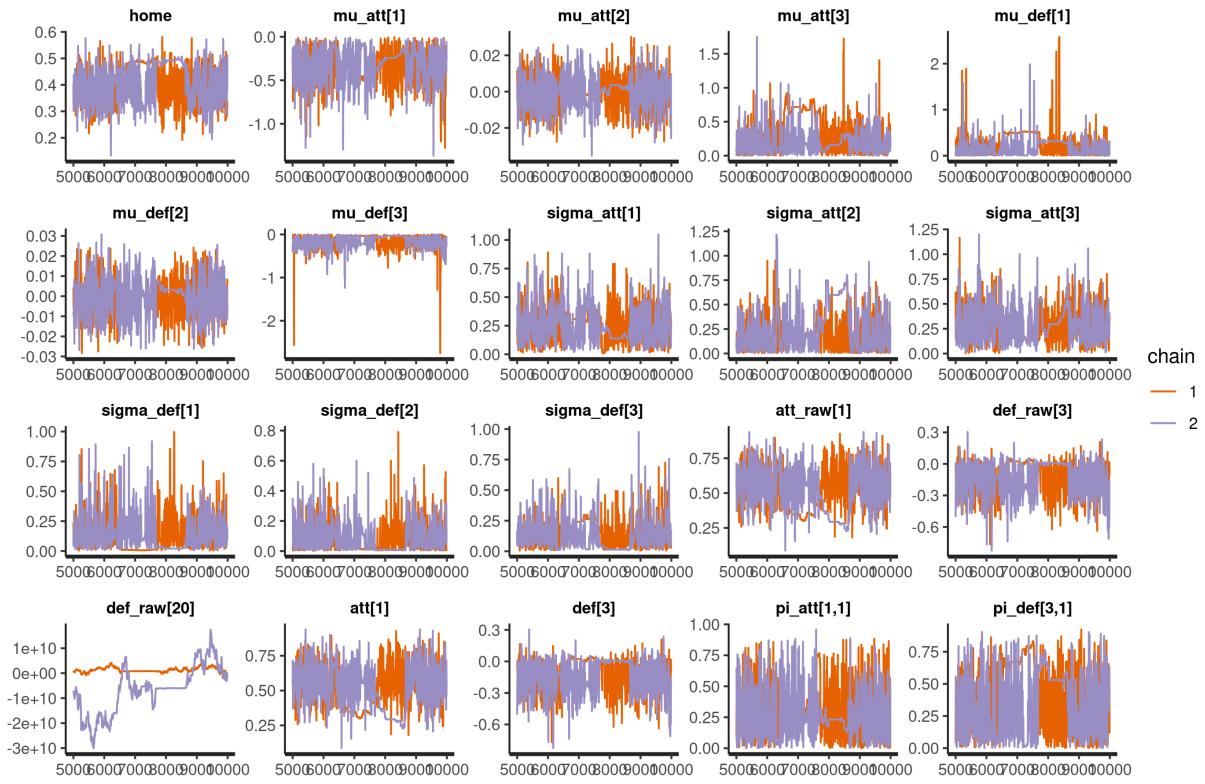
Grupo 3

- $\mu_3^{att} \sim truncNormal(0, 10, 0, 3)$
- $\mu_3^{def} \sim truncNormal(0, 10, -3, 0)$

Um time que performa mal no campeonato possivelmente apresenta uma tendência de pontuar pouco e uma propensão a conceder gols ao adversário. Esta é a motivação de usar a distribuição Normal truncada como priori para os efeitos das médias de ataque e defesa.

## Diagnóstico das cadeias

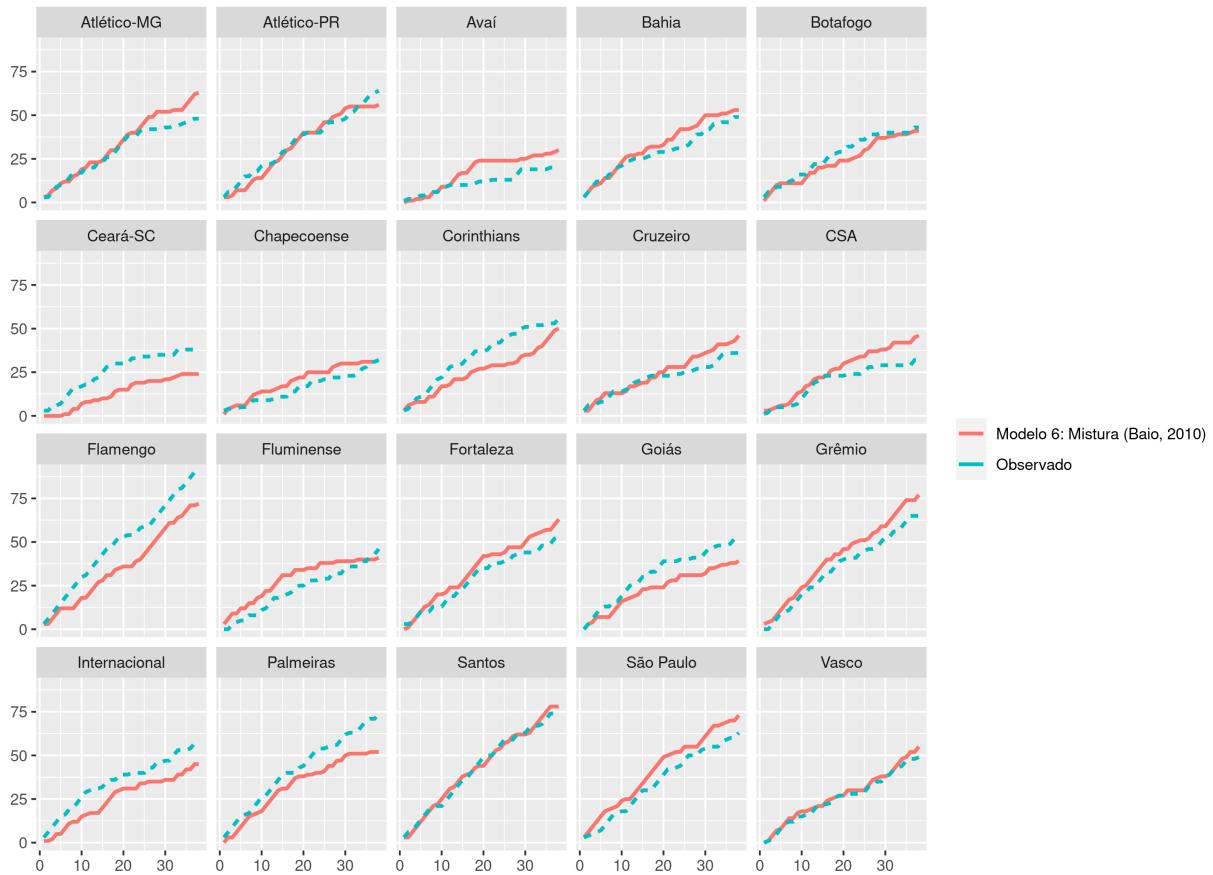
Para este modelo em específico, foram usadas duas cadeias para estimação das distribuições a posteriori, o parâmetro  $thin$  igual a 5, descartando tal valor de amostras, e 10000 iterações.



As cadeias apresentaram problemas de convergência do parâmetro `def_raw` do vigésimo time. Além disso, parece haver uma correlação entre os parâmetros já próximo do fim das iterações. Portanto, o modelo foi incluído mas a inferência pode estar comprometida.

## Ajuste

Campeonato Brasileiro 2019



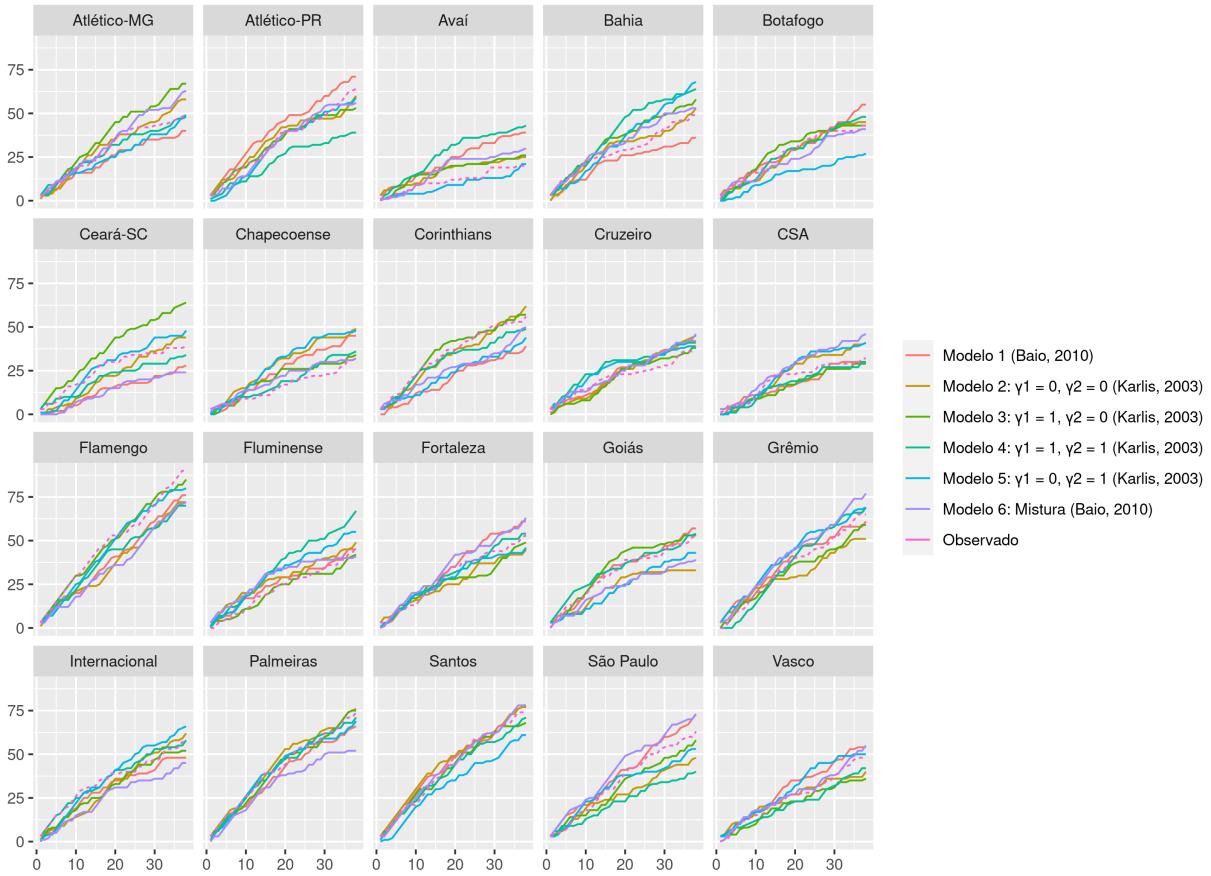
É importante considerar uma possível má especificação do modelo, por isso a análise deve ser realizada com cautela. Ele consegue estimar bem a pontuação de alguns times, por exemplo Santos, Botafogo e Chapecoense, apesar do leve afastamento em algumas rodadas da competição. Contudo, parece que as estimativas são otimistas para times que seriam considerados, popularmente, como topo da tabela, e pessimista para as equipes inferiores.

## 3. Comparação dos modelos

### 3.1 Erro quadrático médio

Um dos principais interesses em ajustar diferentes modelos é a comparação da qualidade de ajuste de cada um, com o objetivo de escolher o mais apropriado.

### Campeonato Brasileiro 2019



A pontuação acumulada de todos os modelos ajustados está no gráfico acima, com o tracejado representando a pontuação observada. Observa-se que alguns modelos subestimam o desempenho de alguns times e outros são superestimados. Para auxiliar a visualização, a tabela abaixo apresenta os valores observados e estimados para cada time, além do erro quadrático médio de cada modelo.

Os times em negrito são os melhores e os piores do campeonato. O modelo que chegou mais perto da pontuação do campeão Flamengo foi o modelo 3, que considera na correlação o efeito fixo e o do time mandante, enquanto as piores estimativas ficaram o segundo e o modelo de mistura.

### 3.2 Validação cruzada

Medidas de qualidade relativa dos modelos são fundamentais para seleção de modelos e avaliação da acurácia preditiva dos modelos.

Vehtari, Gelman, and Gabry (2015) propõe o cálculo do LOO-CV, que é feito baseado na esperança da densidade preditiva do modelo.

- $elpd_{loo} = \sum_{i=1}^n \log p(y_i|y_{-i})$ , onde  $p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i}d\theta)$  é a densidade preditiva
- $LOO_{ic} = -2 \times elpd_{loo}$
- Quanto menor, melhor

Os valores obtidos para todos os modelos estão a seguir:

A coluna `looic` indica o valor obtido para cada modelo. Já a coluna `elpd_diff` é a diferença entre os modelos, sendo o primeiro o modelo com o menor LOO, representando a diferença do melhor modelo e ele mesmo,

Time	Real	M1	M2	M3	M4	M5	M6
Atlético-MG	48	40	58	67	48	49	63
Atlético-PR	64	71	60	53	39	59	56
<b>Avaí</b>	<b>20</b>	<b>39</b>	<b>25</b>	<b>26</b>	<b>43</b>	<b>21</b>	<b>30</b>
Bahia	49	36	53	58	64	68	53
Botafogo	43	55	45	43	48	27	41
Ceará-SC	39	28	44	64	34	48	24
<b>Chapecoense</b>	<b>32</b>	<b>45</b>	<b>49</b>	<b>34</b>	<b>36</b>	<b>48</b>	<b>32</b>
Corinthians	56	39	62	57	49	44	50
<b>Cruzeiro</b>	<b>36</b>	<b>42</b>	<b>45</b>	<b>38</b>	<b>41</b>	<b>39</b>	<b>46</b>
<b>CSA</b>	<b>32</b>	<b>30</b>	<b>41</b>	<b>29</b>	<b>30</b>	<b>41</b>	<b>46</b>
<b>Flamengo</b>	<b>90</b>	<b>76</b>	<b>72</b>	<b>85</b>	<b>70</b>	<b>80</b>	<b>72</b>
Fluminense	46	45	49	42	67	55	41
Fortaleza	53	61	45	49	54	46	63
Goiás	52	57	33	53	54	43	39
<b>Grêmio</b>	<b>65</b>	<b>61</b>	<b>51</b>	<b>59</b>	<b>69</b>	<b>69</b>	<b>77</b>
Internacional	57	48	62	52	58	66	45
<b>Palmeiras</b>	<b>74</b>	<b>66</b>	<b>76</b>	<b>75</b>	<b>71</b>	<b>69</b>	<b>52</b>
<b>Santos</b>	<b>74</b>	<b>78</b>	<b>77</b>	<b>68</b>	<b>71</b>	<b>61</b>	<b>78</b>
São Paulo	63	73	48	58	40	53	73
Vasco	49	54	40	36	42	50	55
	99.7	98.6	79.6	149.1	95.9	125.4	

model	elpd_diff	se_diff	looic	se_looic
model1	0.000	0.000	2014.037	35.898
model6	-0.665	0.936	2015.367	35.467
model2	-1158.510	82.250	4331.057	195.930
model5	-2701.602	150.241	7417.241	331.722
model3	-2765.482	152.215	7545.002	335.825
model4	-4187.758	212.787	10389.553	456.748

assim em diante. É importante ressaltar novamente que, apesar da inclusão do modelo 6, como suas cadeias de Markov não estão adequadas, a inferência do modelo está comprometida.

## 4. Outros modelos

Um possível problema com modelos hierárquicos é um efeito de encolhimento, no qual observações extremas são arrastadas para a média global. Esse efeito faz com que equipes com um desempenho muito bom, que estão no topo da tabela, têm estimativas conservadoras e os times que estão nas últimas colocações são superestimados. O encolhimento é um possível problema do modelo 1 e Baio (2010) recomenda um modelo de mistura com três componentes para contornar esse efeito. Outra possibilidade é a Binomial Negativa, principalmente para o caso que temos uma variância maior que a média, o problema da superdispersão.

## 5. Conclusão

Foi observado o ajuste de 6 modelos para os dados do Campeonato Brasileiro 2019. O melhor ajustado, seguindo o critério de qualidade de ajuste LOO-CV, é o modelo hierárquico proposto por Baio (2010).

Há ainda a possibilidade de ajuste para outros anos, como 2020 e 2021, para ver o comportamento do modelo

diantre de novos times. Uma hipótese levantada e analisada por Benz (2020) é que durante a pandemia o efeito de jogar em casa sofreu uma mudança, considerando a falta de público nos jogos desse período. É possível também a aplicação dos modelos apresentados para outros esportes, como vôlei (Gabrio 2021a) e polo aquático (Karlis 2003).

## 6. Referências

## 7. Apêndice

### 1. Traceplot para diagnóstico das cadeias

Abaixo, os *traceplots* dos modelos ajustados com os dados do Campeonato Brasileiro. Como o número de parâmetros é muito grande, apenas alguns foram selecionados.

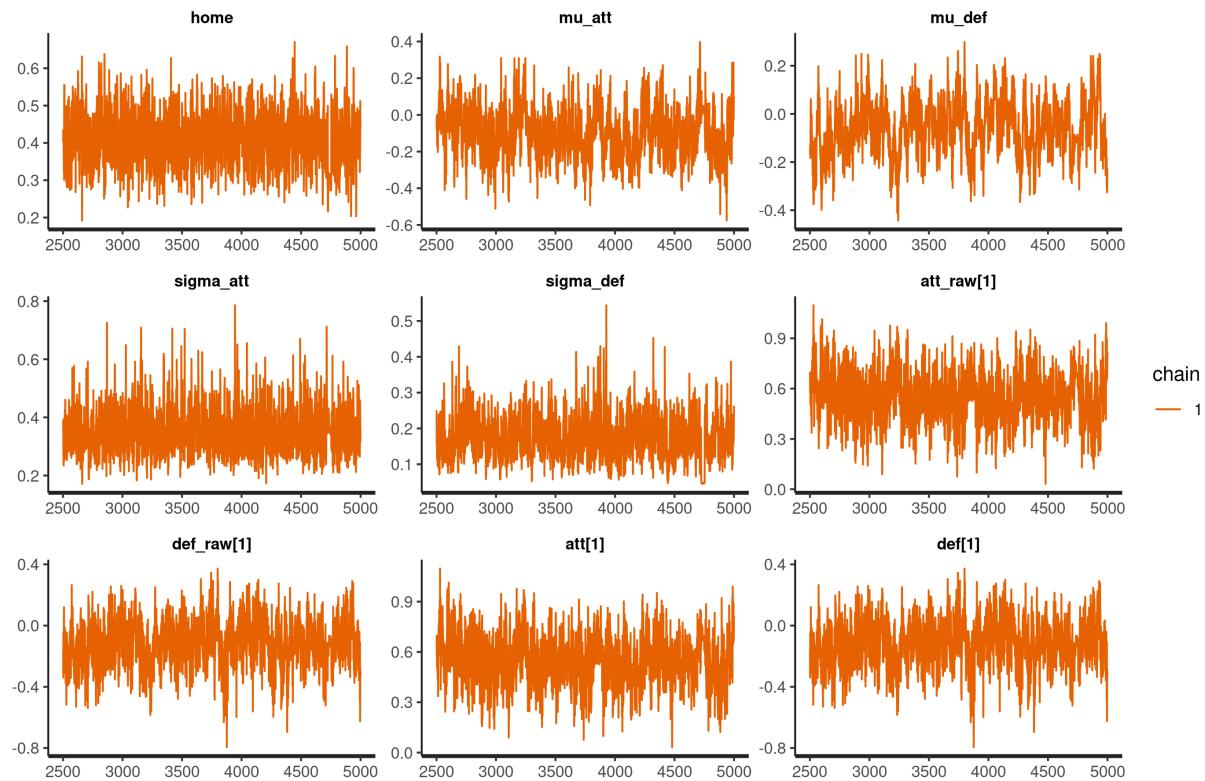


Figura 3: Convergência - Modelo 1

### 1.1 Modelo 1

### 1.2 Modelo 2

### 1.3 Modelo 3

### 1.4 Modelo 4

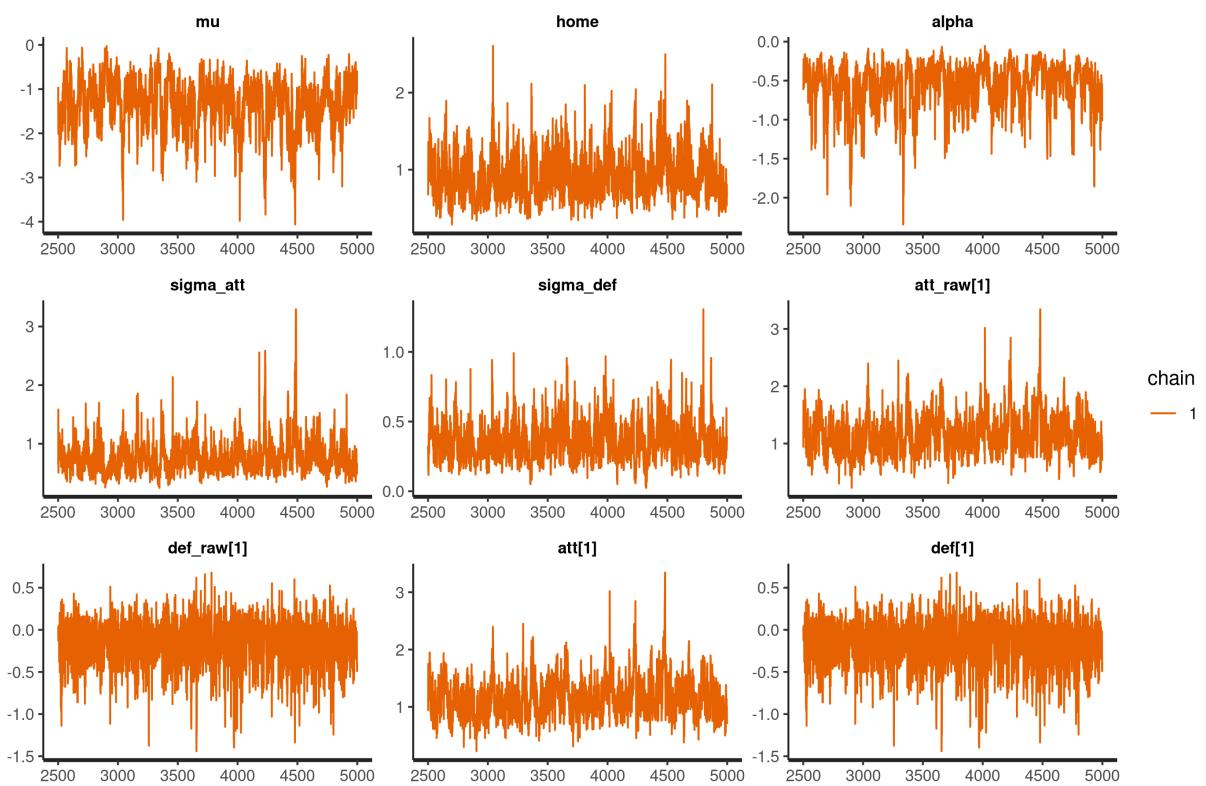


Figura 4: Convergência - Modelo 2

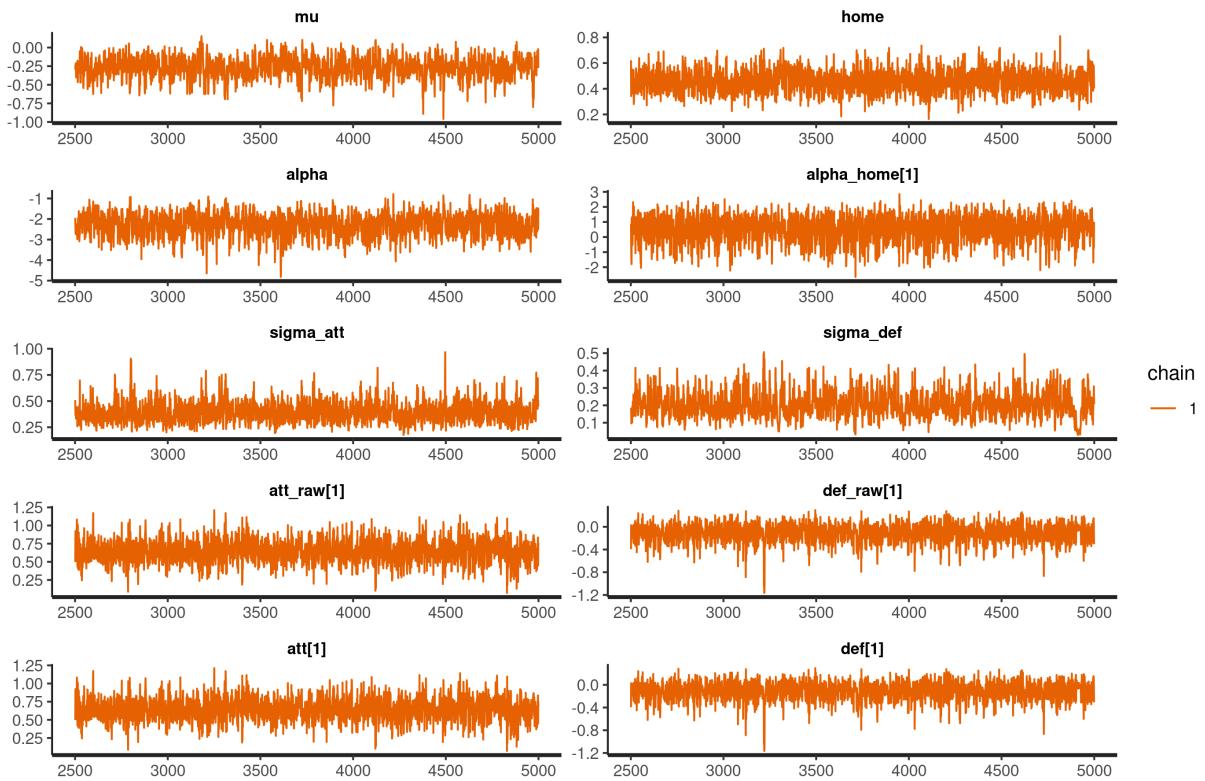


Figura 5: Convergência - Modelo 3

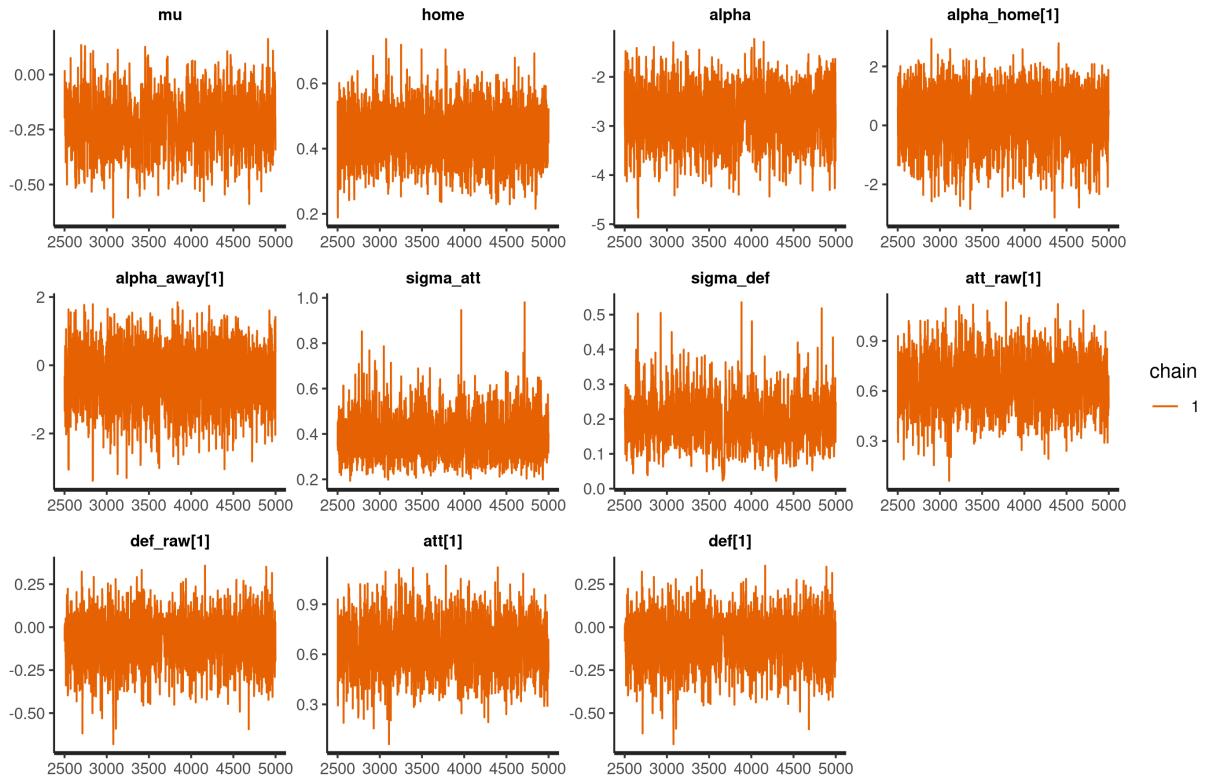


Figura 6: Convergência - Modelo 4

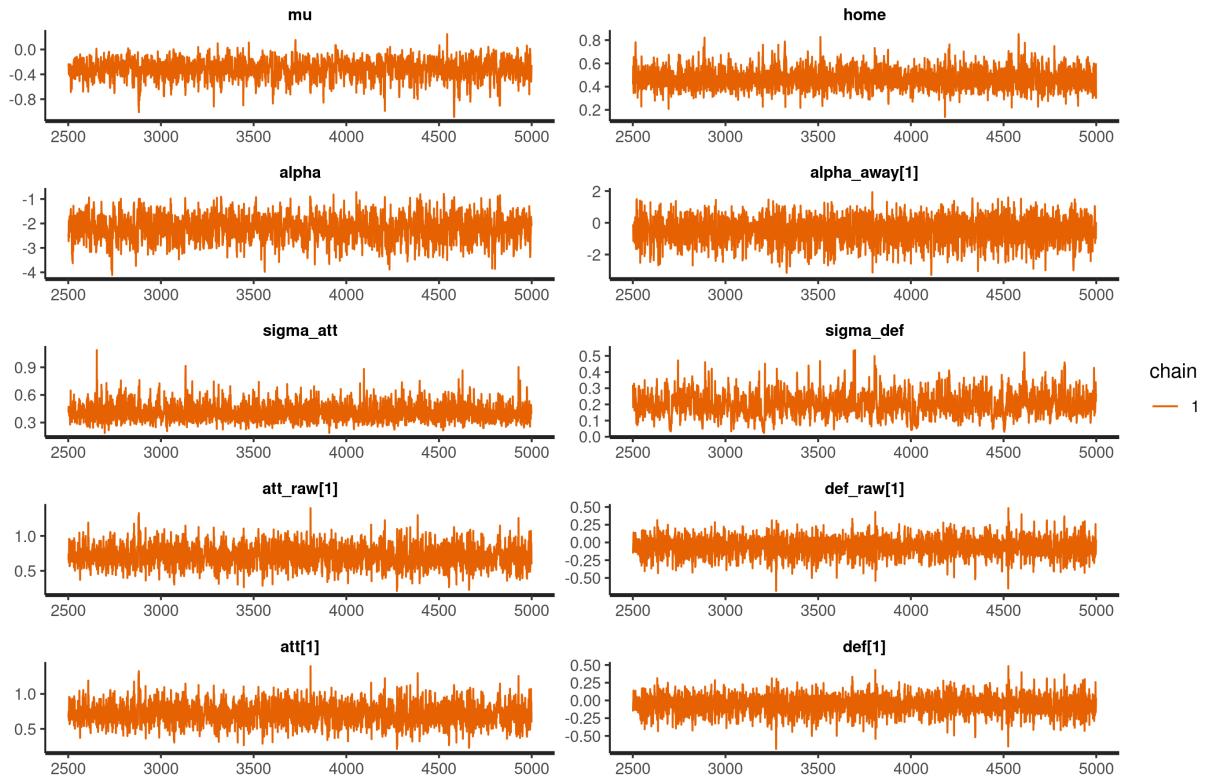


Figura 7: Convergência - Modelo 5

- 1.5 Modelo 5** 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>.
- Benz, Michael J., Luke S. e Lopez. 2020. “Estimating the Change in Soccer’s Home Advantage During the Covid-19 Pandemic Using Bivariate Poisson Regression.” <https://doi.org/10.48550/ARXIV.2012.14949>.
- Gabrio, Andrea. 2021a. “Bayesian Hierarchical Models for the Prediction of Volleyball Results.” *Journal of Applied Statistics* 48 (2): 301–21. <https://doi.org/10.1080/02664763.2020.1723506>.
- . 2021b. “Bayesian Hierarchical Models for the Prediction of Volleyball Results.” *Journal of Applied Statistics* 48 (2): 301–21. <https://doi.org/10.1080/02664763.2020.1723506>.
- 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/10.111/1467-9884.00366>.
- McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 2nd ed. CRC Texts in Statistical Science. Boca Raton: Taylor; Francis, CRC Press.
- 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>.
- R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org>.
- Stan Development Team. 2022. “Stan Modeling Language Users Guide and Reference Manual, Version 2.21.0.” <http://mc-stan.org/>.
- Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2015. “Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC.” <https://doi.org/10.48550/ARXIV.1507.04544>.